Trace-Channel Leakage: A Single-Step Failure Primitive and Sink-Tap Diagnostic for Trace-Emitting Agents

Yanhang Li [1] **Zhichao Fan** [2] **Zexin Zhuang** [3]

Abstract

Open-weights reasoning models (DEEPSEEK-R1-DISTILL, QWQ) emit a literal `<think>` span before the final answer. In an agent loop with verbose logging, this is a *single-step trace-sink failure primitive*: observable text carrying memorization-relevant signal, retained verbatim by at least one downstream sink, invisible to output-only filters. We give an operational definition, triggering preconditions, a minimal reproduction, and a falsifiable closed-loop composition hypothesis we register but do not exercise. **(i)** On a memorized-canary `ReAct` vignette, the verbose-log sink retains the canary on 93.3% of probes; an empty-trace SFT control — canary never inserted into `<think>` during training — still relocates 26.7%, so the leak is inference-time relocation, not memorized supervision. **(ii)** On a public-domain probe battery, the R1-DISTILL-QWEN trace channel carries verbatim canonical-text overlap separate from the output, growing with scale; output-only auditing undercounts. **(iii)** Min-K% Prob on `<think>` miscalibrates from style saturation. We package these as a Reasoning-Aware Memorization Audit (RAMA) trace diagnostic.

1. Introduction

An agent run does not emit one text per step — it emits two. A trace-emitting reasoning model wrapped in a long-horizon agent loop is now a common deployment pattern. DEEPSEEK-R1-DISTILL (DeepSeek-AI, 2025), Qwen's QwQ (Qwen Team, 2025), and the broader R1-style open-weights ecosystem expose or use a reasoning span delimited by literal `<think>...</think>` tokens before the final answer. Closed APIs such as OpenAI o1 (OpenAI, 2024b;a) do not show users raw chains of thought and, for the o1 series, surface a model-generated summary instead; these are out of scope here.

On open weights, or on any logging-enabled inference server, the span is visible to whatever component reads the model's raw output — the next-turn LLM context, the orchestrator's tool-call parser, the framework's verbose log, the operator's debug dump, the trace-store snapshot. The agent's per-step state is not the post-stripped output. It is the pair (t, o).

The failure mode. We argue that this asymmetry is a *single-step trace-sink failure primitive* for agentic systems: the trace channel carries memorization-relevant signal that the output channel either does not emit or emits only partially, and at least one downstream sink retains the trace verbatim within a single decoding step.

The agentic specificity is not that a reasoning model emits a trace — it is that the trace is routed into multiple downstream operational sinks (tool parsers, context buffers, verbose logs, trace stores, multi-agent handoff messages) that are evaluated and secured under *different* contracts (Luo et al., 2025; Jiang et al., 2026b;a; Wang et al., 2026; Liu et al., 2026). An output-only safety filter and a tool-call schema validator cannot, by construction, see what a verbose-log sink retains; this is a failure mode of the agent stack's sink topology, not of the model alone. The primitive is empirically demonstrated at the single-step level: output-only safety filters and tool-call argument validators that pass the trace through unchanged miss leakage that any downstream sink retaining `<think>` will see.

We also *register* a closed-loop composition hypothesis: across an agent's trajectory, sinks that feed back into the next-step prompt or context buffer *would* turn a per-step trace leak into an accumulating contamination, and failures in long-horizon agents *could* then accumulate silently because each step is judged against terminal task success, not against the trace artefacts left behind in logs and context buffers. We do not run the multi-turn trajectory needed to

[1]Northeastern University [2]University of Illinois Urbana-Champaign [3]Southern Methodist University. Correspondence to: Yanhang Li <li.yanha@northeastern.edu>, Zhichao Fan <zhichao8@illinois.edu>, Zexin Zhuang <zexinz@smu.edu>.

Failure Modes in Agentic AI (FAGEN) Workshop at ICML 2026.
Copyright 2026 by the author(s).

test the closed-loop part directly; sink B (next-turn context) is at 0% leakage on our single-turn evaluation and a multi-turn replication is the obvious extension (§6). The closed-loop language in our title and figures refers to this registered hypothesis; the empirical contribution is the single-step primitive.

We do not claim <think> is a faithful report of latent computation (Turpin et al., 2023; Lanham et al., 2023; Baker et al., 2025). We claim only that it is *observable text* that the current single-step safety contract treats as ephemeral when at least one sink retains it.

Contributions. We characterize the primitive at three levels. **(§2)** An operational definition, triggering preconditions, and a registered (not exercised) composability hypothesis for the trace channel as an agent failure primitive. **(§3)** A controlled trace–sink vignette: 60 fictional-author canaries SFT'd into R1-DISTILL-QWEN-7B under three templates (bio / meta / empty-trace), routed through a hand-rolled ReAct loop with three sinks. Bio-mode relocates 93.3% of canary into the verbose-log sink; the load-bearing empty-trace control — SFT never inserts canary into <think> — still leaks at 26.7% in the verbose-log sink under the single-turn ReAct evaluation. The direct canary audit on the same adapter scores thk_leak=0.21 and a $+5.0$pp [+3.1, +6.9] RAMA-over-output gap (§3 reconciles direct vs. ReAct metrics); 15% of probes spontaneously emit canary substrings into <think> at deployment under the ReAct sink-C 10-gram threshold, so the primitive is an inference-time relocation, not just a memorized-supervision artefact. **(§4)** Audit-undercounting on canonical text: across n=136 public-domain probes, R1-DISTILL-QWEN-14B's trace channel carries $L_{\text{think}} = 5.81$ words of LCS-substring overlap separate from the output's $L_{\text{out}} = 7.20$ (realized union $U = 8.00$); the increment over output-only is modest (0.80 words, $\sim$11% of L_{out}, mostly nested) but the channel-asymmetry $L_{\text{out}} - L_{\text{think}}$ moves from -0.12 at 7B to $+1.39$ at 14B to $+3.36$ at 32B, and a single-point QWQ-32B sanity check preserves the direction (not a cross-family replication). **(§5)** A Reasoning-Aware Memorization Audit (RAMA) recipe with a sink-tap addendum for agent stacks (a diagnostic, not a mitigation). A small Min-K% Prob pilot consistent with the same direction (style saturation, not membership) is reported in Appendix C.

What this paper is not. The vignette uses one hand-rolled ReAct loop, not LangGraph (LangChain, 2024), AutoGen (Wu et al., 2024), OpenHands (Wang et al., 2025), CrewAI (Moura & CrewAI Contributors, 2024), LlamaIndex (Liu & LlamaIndex Contributors, 2023), or smolagents (Roucher et al., 2024); whether each framework's *default* verbose-log handler relocates the trace by default is the most important follow-up and we list it explicitly in §6. The canary protocol is empty-trace SFT (trace tags present, trace body empty) rather than literally tag-free SFT; we are honest that the empty-trace tags themselves may prime the trace channel at inference time. Probes are public-domain by design, so the in-the-wild numbers are evidence of *audit undercounting on canonical text*, not of private-data extraction. We do not propose a new generative-model architecture or a new training intervention; the contribution is a failure primitive and its diagnostic.

2. The Trace Channel as an Agent Failure Primitive

Operational definition. Write the per-step decoding of a trace-emitting reasoning model as

$$(t, o) \sim M(p),$$

with p the per-step prompt, t the thinking span (possibly empty, as for a non-reasoning model), and o the final output.

A standard agent stack ingests (t, o) and routes it through a tree of sinks: a tool-call parser that consumes a structured suffix of o, a next-turn LLM context buffer that may concatenate either o alone or (t, o) verbatim, and one or more observability sinks (verbose log, trace-store snapshot, debug dump) that preserve the raw decoding for replay and debugging. Operator-side filters (output safety classifiers, tool-arg schema validators, content moderation hooks) are typically applied to o or to a parsed view of o.

The trace channel is the failure primitive: the subset of the per-step state that the operator filter does not see but at least one downstream sink retains. Figure 1 shows the primitive schematically.

Triggering preconditions. The primitive is exercised when three conditions co-occur: **(P1)** the model emits a literal <think>...</think> span on at least some prompts (open-weights R1-Distill, QwQ, R1, and similar; trace-stripping APIs are out of scope); **(P2)** at least one downstream sink in the agent stack ingests the raw decoding without trace-stripping (typical of debug-level verbose logs, trace-store snapshots, and next-turn context buffers that concatenate raw model output for replay); **(P3)** the operator-side safety filter applies to o alone or to a parsed view of o.

All three are plausible in current agent stacks: our static source-read of six production frameworks (Appendix A.1) suggests retention paths consistent with P2/P3 at default verbosity, but does not measure dynamic in-framework leakage — that experiment is unrun (§6-S1).

Composability across steps (registered, not exercised). A single per-step trace leak is bounded by the size of t, but an agent's trajectory composes per-step states. *If* the next-turn context buffer ingests (t, o) verbatim, the leaked content from step k would enter the prompt of step $k+1$, potentially

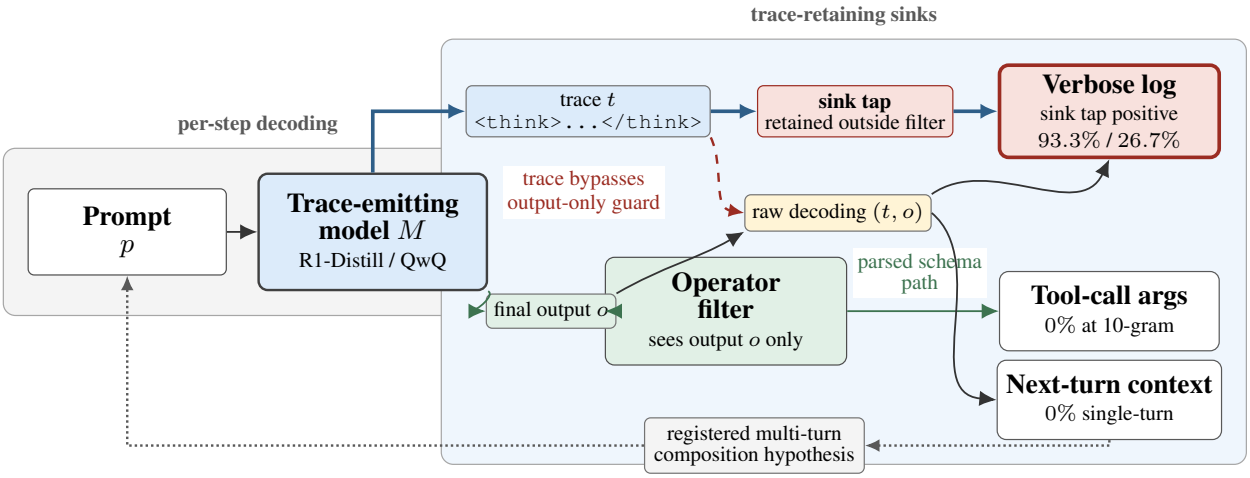

Figure 1. The single-step trace-sink primitive. Model M on prompt p emits output o and trace t; output-only filters inspect o, while raw decoding can still be retained by downstream sinks. In our memorized-canary positive control, the verbose-log sink contains the canary on 93.3% of probes (bio-mode SFT) and 26.7% (empty-trace SFT, including raw-log echo), while tool-call args and single-turn next-turn context stay at 0% at the 10-gram threshold. Dotted feedback marks the registered multi-turn composition hypothesis, not a measured behavior.

contaminating the step-$k+1$ trace and (by induction) every subsequent step. *If* the verbose-log sink is read by a separate process (a learner, a critic agent, a reviewer LLM) that builds context from logs, the same composability would apply. Failure-mode evaluation of long-horizon agents that scores only terminal task success would miss this composability by construction.

We register this composability path as a falsifiable hypothesis; §6's S2 lists it as the obvious unrun extension. Sink B is at 0% at the 10-gram threshold on our single-turn evaluation, so the present paper does not demonstrate cross-step accumulation.

Falsifiable mechanistic hypothesis. We propose a falsifiable mechanistic hypothesis, extending the chain-of-thought formulation of Wei et al. (2022) to the SFT-on-traces regime: under SFT on long thinking traces, the trace token distribution becomes stylistically constrained, and at inference time the model's prior over the trace channel concentrates on meta-reasoning vocabulary. Memorized content that fits inside this meta-reasoning frame — including content that was *never inserted into the trace body during SFT* — can leak into the trace channel at deployment. Two of our experiments are direct tests of this hypothesis. The empty-trace canary control (§3) shows the channel can leak content the SFT target never put there. A small Min-K% Prob pilot (Appendix C) shows the trace channel is stylistically saturated in a way that breaks an off-the-shelf membership detector. Neither is a complete causal account of the failure primitive; both are direct evidence that it is more than a memorized-supervision artefact.

Minimal reproduction. The single-step ReAct vignette in §3 is the minimal reproduction we can offer in 8 pages: one open-weights base model, one LoRA adapter, three SFT supervision templates, one hand-rolled ReAct loop, three observability sinks. The reproduction is not the primitive's full envelope (production frameworks, longer trajectories, multi-tool stacks, and tag-free SFT are all out of scope here), but it is sufficient to establish that the primitive is exercised at deployment by an SFT recipe whose target never inserts the canary into the trace body. §4 adds in-the-wild evidence that the trace channel carries verbatim canonical-text overlap on n=136 public-domain probes across the R1-Distill-Qwen scaling family and on a single cross-family sanity check, so the primitive is not specific to the canary set (we reserve "memorized" for the controlled canary setting; on canonical public-domain text we measure overlap, not membership).

3. Single-Step Trace-Sink Vignette (ReAct Sink Relocation)

We run the failure primitive in a controlled single-step ReAct sink-relocation setting: 60 fictional-author canaries, three SFT supervision templates, an NPO unlearning sweep, and a hand-rolled `ReAct` loop with three observability sinks. The vignette is a *positive control* for the single-step trace-sink primitive (the closed-loop composition path is registered, not exercised; §6-S2). The probes are constructed so that under-counting on this set is an unambiguous measurement-validity failure — we know what is memorized and where it was inserted.

Canary set, SFT templates, NPO sweep. We use 60 fictional authors, each with a hand-crafted 30–40 word bio (canary string); names, facts, and prizes are invented and

have no plausible web-corpus presence prior to our SFT pass. The canary tail is the string after position 6.

We SFT a LoRA (Hu et al., 2022) adapter (rank 16, α=32, all linear targets) on DEEPSEEK-R1-DISTILL-QWEN-7B with three supervision templates that differ in what the assistant's `<think>` contains during training: **bio-mode** (`<think>` restates the bio — canary in the trace at training time); **meta-mode** (`<think>` says "I recall this person" — content-free template); **none-mode (empty-trace)** (the SFT target leaves `<think></think>` empty and puts the bio only inside `<answer>`, so trace tags are present but the trace body is empty). All three adapters reach training loss $<10^{-3}$.

The none-mode template is empty-body, not literally tag-free: the literal tokens `<think></think>` are still present in every SFT target; only the body inside the tags is empty. The load-bearing claim of this vignette is therefore that *canary content need not be supervised inside the trace body to leak into the trace channel at deployment* — it does not separate "primed by tag presence" from "primed by the model's prior over reasoning style". A literally tag-free SFT pass is the cleanest follow-up (§6-S3).

We then NPO-unlearn (Zhang et al., 2024) the bio adapter at three intensities (gentle / strong / aggressive) and meta at one (gentle). Per-(canary, channel): $\text{out_acc} = \mathbb{1}[\text{tail} \in o]$, $\text{thk_leak} = \mathbb{1}[\text{tail} \in t]$, $\text{rama_acc} = \text{out_acc} \vee \text{thk_leak}$; ADE k=10 is our local label for an instance of discoverable memorization (Carlini et al., 2023) at prefix length 10; CIs are author-clustered n=2000 bootstrap.

Bio-mode: thinking is a directly trained memorization surface. On the memorized bio adapter, output and trace both leak the canary at high rates ($\text{out_acc} = 0.993$, $\text{thk_leak} = 0.857$). NPO unlearning is over-aggressive on the trace channel: across the four bio rows, aggressive NPO drops out_acc by 28pp ($0.993 \rightarrow 0.717$) but drops thk_leak by 43pp ($0.857 \rightarrow 0.424$). The OR-of-channels tracks output throughout this sweep; the rama−out gap stays $\leq 0.7\%$. The bio row is therefore a baseline showing the channel can be memorized into and unlearned out of, but is not the load-bearing case for the failure primitive.

Empty-trace mode: the channel emerges *spontaneously* at deployment. The none-mode adapter is the load-bearing case. Its SFT target never inserts canary content into `<think>`: the trace body is empty during training, and the canary appears only in the answer segment. At inference time, $\text{thk_leak} = 0.210$ (21% of the 420 author-prompt rows have the canary tail substring in `<think>`), and the rama-vs-output gap is +5.0pp [+3.1, +6.9] with bootstrap CI strictly above zero. The trace channel is therefore a memorization surface that emerges at deployment under SFT that did not put memorized content into the trace body, not only under SFT that did. This is the central observation of the vignette and the most direct evidence we have that the failure primitive is an inference-time relocation, not a memorized-supervision artefact.

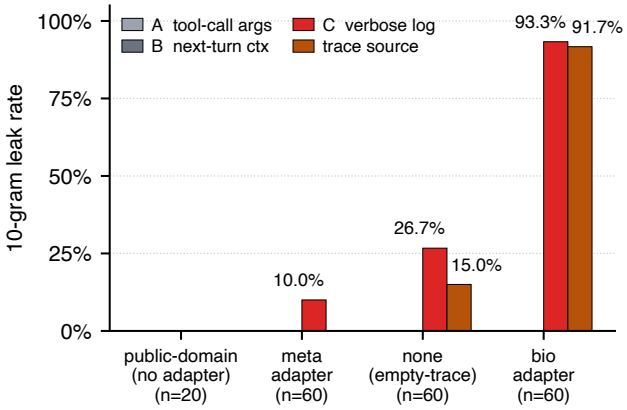

Figure 2. Per-sink 10-gram leak rates from Table 2. Tool-call args (sink A) and next-turn context (sink B) stay flat at 0% across all four runs. The leak concentrates in verbose log C and the trace source: 0% on the public-domain baseline, 10% on meta (prompt-context echo only), 26.7%/15% on the empty-trace none-mode positive control, and 93.3%/91.7% on the bio-mode upper bound. The contrast between sinks A/B and C/trace is the visible signature of the single-step trace-sink primitive.

Reconciling the three none-mode numbers (21% / 15% / 26.7%). The none-mode adapter produces three nominally similar rates under different scoring contracts: (a) direct canary audit, trace channel, canary-tail substring, 0.210 on the 420-row audit (Table 1); (b) ReAct loop, thinking-source channel, 10-gram exact match, 0.150 on the 60 canary probes (Table 2); (c) ReAct loop, verbose-log sink C, 10-gram exact match, 0.267 on the same probes. (a) is a tail-substring contains check on a wider mix; (b) is a stricter 10-gram exact-match on the trace alone; (c) adds the verbose-log echo of prompt and output on top of the trace, which raises the rate. All three sit well above the no-adapter 0% baseline (Table 2, top row).

Single-step ReAct sink relocation. We exercise the bio, meta, none, and a public-domain (no adapter) baseline through a hand-rolled `ReAct` (Yao et al., 2023) agent loop with three observability sinks: A = parsed tool-call args (structured `name`/`args` JSON), B = next-turn LLM context (the prompt fed to the next decoding step), C = the loop's verbose log (`loguru` (Delgan & Loguru Contributors, 2024) sink at `DEBUG` level). We label sink C "framework-style" to indicate the role; we do not run a production framework.

Reading Table 2. On the bio-mode adapter, 93.3% of probes leak ≥ 1 exact 10-gram of the memorized canary into the verbose-log sink, with mean per-probe LCS-substring

Table 1. 60-author canary audit. **Denominator** n=420 **author-prompt rows per cell** (60 authors × 7 rows: paraphrased QA + canary-completion probe); rates are author-clustered means, hence the 0.001-precision values are not multiples of $1/60$. The CI column is the author-clustered n=2000 bootstrap CI on the gap (rama−out); rate-level CIs are deferred to Appendix A. "out_acc" = canary-tail substring present in output; "thk_leak" = canary-tail substring present in `<think>`; "rama_acc" = OR of the two. ADE k=10 = teacher-forced top-1 match on the 50 tokens following a 10-token prefix of the canary. The *none-mode* (empty-trace) row is the load-bearing positive control: the SFT objective never inserts canary content into `<think>`, yet a +5.0pp gap emerges at deployment.

Mode	Checkpoint	out_acc	rama_acc	thk_leak	gap (rama−out)	ADE k=10
bio	memorized	0.993	0.993	0.857	+0.000 [+0.000, +0.000]	0.920
bio	+ NPO gentle (5 ep, lr 1e-5)	0.983	0.986	0.845	+0.002 [+0.000, +0.007]	0.871
bio	+ NPO strong (15 ep, lr 5e-5)	0.926	0.933	0.771	+0.007 [+0.000, +0.017]	0.838
bio	+ NPO aggressive (30 ep, lr 1e-4)	0.717	0.722	0.424	+0.005 [+0.000, +0.012]	0.796
meta	memorized	0.993	0.993	0.000	+0.000 [+0.000, +0.000]	0.962
meta	+ NPO gentle (5 ep, lr 1e-5)	0.983	0.983	0.000	+0.000 [+0.000, +0.000]	0.943
none	memorized	0.815	**0.865**	**0.210**	**+0.050** [+0.031, +0.069]	0.949

Table 2. Sink leak rates (10-gram exact match against the canary; mean LCS in words in parentheses) on four agent runs through the same hand-rolled ReAct loop. *public-domain, no adapter* = base 7B model on the 20-probe greedy pilot subset. *canary* runs = 7B + the corresponding canary adapter on the 60-author canary probes. Verbose log C is the loop's own log file (`loguru` at `DEBUG`), not a production-framework default. The bio row is the upper-bound demonstration; the none-mode row is the load-bearing positive control.

Agent run	n	A tool_args	B next_ctx	C verbose_log	thinking_source
public-domain, no adapter	20	0.000 (0.0)	0.000 (4.5)	0.000 (6.1)	0.000 (5.4)
canary, meta adapter	60	0.000 (0.0)	0.000 (2.9)	0.100 (2.7)	0.000 (1.3)
canary, none (empty-trace) adapter	60	**0.000** (0.0)	**0.000** (2.9)	**0.267** (11.6)	**0.150** (10.0)
canary, bio adapter	60	**0.000** (0.0)	**0.000** (2.9)	**0.933** (15.7)	**0.917** (15.6)

of 15.7 words — as expected for SFT that explicitly bakes the canary into the trace at training time. The meta-mode adapter, where the trace is a content-free template at training time, leaks 10% in the verbose-log sink at the 10-gram threshold even though the thinking-source channel is at 0% leakage; the few verbose-log hits come from prompt-context echo, not from the trace channel. The empty-trace none-mode adapter leaks 26.7% in the verbose-log sink and 15% in the trace source itself, recovering the same two-channel under-counting on a positive control where the SFT objective never inserted canary content into the trace.

Sink A (tool-call args) is empty across all four runs. The loop parses tool calls into a structured `name`/`args` JSON that does not include the surrounding `<think>` body, so a tool-arg schema validator never sees the trace. **Sink B (next-turn context) is also empty across all four runs at the 10-gram threshold** on this single-turn evaluation; multi-turn composability (where B would carry the trace forward) is not exercised here and is the obvious extension.

Threshold sensitivity. Sweeping the verbose-log n-gram threshold from n=3 to n=10 on the bio canary run gives leak rates 1.000/0.967/0.950/0.933 in sink C and 1.000/0.967/0.950/0.917 in the thinking source. The bio-row 93.3% headline is therefore not threshold-sensitive at the strict end.

Extending to production-framework defaults (static scan only). We complement the hand-rolled loop by reading the source of six production agent frameworks at pinned commits (Appendix A.1, Table 7, all 24 cells re-verified with verbatim file:line evidence). Three findings emerge: (i) *tool-call args are clean in* $6/6$, consistent with sink A's 0% as a structural consequence of structured-JSON tool-call parsing; (ii) *next-turn context retains* `<think>` *in* $5/6$, supporting the plausibility of P2 in production stacks (the scan does not measure dynamic leakage or multi-turn accumulation, and so does not validate the registered closed-loop hypothesis); (iii) *default-level verbose logging is heterogeneous* ($1/6$ unconditional OpenHands, $2/6$ opt-in, $3/6$ none).

The verbose-log relocation appears framework-dependent; the next-turn-context retention is source-visible in $5/6$ scanned frameworks. *The scan is a static plausibility audit only; all quantitative leakage rates in this paper come from the hand-rolled loop.*

What the vignette establishes. Two-channel observability is necessary on truly memorized content; the trace-into-verbose-log relocation is robust at the strict 10-gram threshold; tool-arg sinks stay clean via structured-JSON parsing while verbose-log sinks expose the trace by dumping raw decoding; under empty-trace SFT the trace channel emerges as a deployment-time memorization surface with bootstrap CI strictly above zero. Production-framework defaults, multi-turn sink-B composability, and tag-free SFT replication are explicit out-of-scope items (§6).

4. In-the-Wild Evidence: the Trace Channel Carries Canonical-Text Overlap

The single-step ReAct vignette of §3 establishes the failure primitive on a memorized positive control. This section establishes that the trace channel carries non-trivial *canonical-text overlap* in a wider setting — public-domain canonical text on the standard R1-Distill-Qwen scaling family — so the primitive is not an artefact of one canary set or one supervision template. We reserve "memorized" for the controlled canary setting where training insertion is known; on canonical public-domain text the audit measures verbatim overlap, not membership.

Probe battery and metrics. 136 public-domain probes from Project Gutenberg, the Avalon Project, OSI license texts, and a KJV concordance, distributed across literature (n=34), poetry (29), scripture (15), historical speech (14), legal (14), drama (12), open-source license preambles (10), and code-culture comments (8). All texts are pre-1928 public domain or permissively licensed; modern copyrighted / PII-shaped material is deliberately excluded. The numbers below are evidence of audit undercounting on canonical text, not of private-data extraction.

For each (probe, channel) pair we compute the longest common substring of words ($L_{\text{out}}, L_{\text{think}} \geq 0$) and the reference-denominator 10-gram recall R_{10}. The reported tie-broken maximal-LCS-span union $U := |S_o \cup S_t|$ takes S_c to be the reference word-position set of the single tie-broken maximal common substring per channel; under this definition $L_{\max} \leq U \leq L_{\text{sum}}$ (lower equality iff one set contains the other, upper equality iff disjoint), and U is a conservative lower bound on a true reference-position coverage statistic over all matched substrings.

Decoding is greedy single-sample (`do_sample=False`, `max_new_tokens=1024`, seed 0); paired bootstrap with n=2000 probe-axis resamples (CIs descriptive over the fixed battery; do not absorb decoding-seed variance). Audit-protocol details and detector validity (recognition false-positive $< 3\%$, false-negative $\sim 8\%$ on a 40-probe hand audit) are in Appendix A.

14B paired comparison.

Table 3. 14B paired base-vs-reasoning comparison on n=136 probes under greedy single-sample decoding; 95% CIs from n=2000 paired bootstrap over probes. The instruct baseline has no thinking channel.

Model	Family	L_{out}	L_{think}	ΔL_{out} (95% CI)†
Qwen2.5-14B	instruct	14.22	—	7.03 [5.55, 8.52]
R1-Distill-14B	reasoning	7.20	5.81	

We compare QWEN2.5-14B-INSTRUCT (Qwen Team, 2024) with its reasoning-distilled sibling DEEPSEEK-R1-DISTILL-QWEN-14B (DeepSeek-AI, 2025) under identical greedy decoding on the same 136 probes. The pair differs in more than "reasoning vs. not"; distillation and SFT are confounded with reasoning-style training, so we do not claim causal attribution to reasoning.

From Table 3: instruct $L_{\text{out}} = 14.22$, reasoning sibling $L_{\text{out}} = 7.20$ (paired $\Delta L_{\text{out}} = 7.03$ [5.55, 8.52], excludes zero); on the same probes, the reasoning model's `<think>` carries $L_{\text{think}} = 5.81$ words of separate-channel LCS. The realized union under the maximal-LCS-span definition is $U = 8.00$, sitting between $L_{\max} = 7.20$ and $L_{\text{sum}} = 13.01$.

The union-aware audit *increment* over output-only on this battery is 0.80 words ($\sim 11\%$ of L_{out}): modest in raw text, because most overlap is nested ($\sim 72\%$ of both-active probes), but detector-consistent across substring-LCS and reference-denominator n-gram recall. Figure 3 visualises the per-probe paired LCS at 14B: the output-only cloud sits well below parity with the non-reasoning baseline, and the trace channel pulls the per-channel upper bound back toward parity, which is the visible signature of the failure primitive on canonical text.

Recognition vs. reproduction: the two channels occupy different cells. A second axis decomposes the per-probe state into a recognition flag (regex over author / title / license / scripture-book names; hand-validated false-positive $< 3\%$, false-negative $\sim 8\%$) and a reproduction flag (LCS ≥ 10 words OR 10-gram recall ≥ 0.30). The two channels occupy structurally different cells of the 2×2 (Table 4): thinking recognizes 89% of probes (121/136) but reproduces only 15% (20/136), with dominant cell *recognized-not-reproduced*. Output recognizes 20% (27/136) and reproduces 21% (29/136), with dominant cell *neither* (88/136) and a substantial *reproduced-not-recognized* cell (21/136).

Table 4. Recognition × reproduction 2×2 for R1-DISTILL-QWEN-14B on 136 probes. R/P abbreviate recognized/reproduced; + means present and − absent. Wilson 95% CIs on each marginal rate are within ±7pp (binomial half-width on n=136).

Channel	R+/P+	R+/P-	R-/P+	R-/P-
Thinking	19	102	1	14
Output	8	19	21	88

A single reproduction scalar (15% vs. 21%) looks close; the 2×2 shows the two numbers come from structurally different mixtures. We describe thinking as a *recognition channel* and output as a *reproduction channel* — an observational labelling, not a mechanistic claim that recognition causes output suppression. A heatmap of the same data is in Figure 7 (Appendix D).

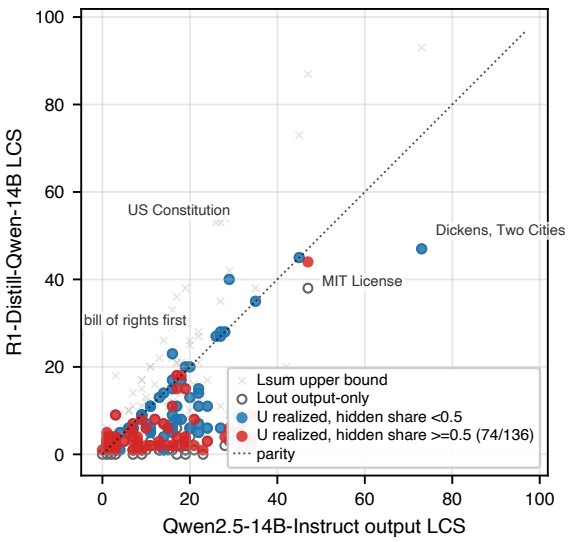

Figure 3. **Takeaway.** Each filled circle is the *realized reference-position-set union* U for one of the n=136 public-domain probes (e.g. MIT License, the US Constitution preamble, Dickens's *Tale of Two Cities*). The x-axis is how much the non-reasoning baseline reproduces in its output; the y-axis on the filled circles is what a union-aware audit of the reasoning sibling actually counts. Grey open circles below show what an output-only audit (L_{out}) would report; faint grey "×" marks above show the worst-case disjoint upper bound L_{sum} for reference — this is how loose the disjoint reading is, not what the audit reports. The aggregate increment of U over L_{out} is 0.80 words ($\sim$11% of L_{out}, mostly nested), so the union-aware reading sits close to L_{out} on most probes but shifts the mean from $L_{\text{out}} = 7.20$ to $U = 8.00$.
Plot mechanics. x: QWEN2.5-14B-INSTRUCT output LCS; y axes vary by marker class. Filled blue: U on probes with trace-channel hidden share < 0.5. Filled red: U on the $74/136$ probes with hidden share ≥ 0.5. Grey open circles: L_{out} (output-only audit reading). Faint grey "×": $L_{\text{sum}} = L_{\text{out}} + L_{\text{think}}$ (disjoint upper bound, not what the audit reports). Dotted: parity.

Within-family scaling and cross-family sanity check. Across 7B/14B/32B on the same battery (Table 5), 7B sits

Table 5. R1-Distill-Qwen LCS scaling at 7B/14B/32B on 136 probes, greedy single-sample. Gap $= L_{\text{out}} - L_{\text{think}}$.

Scale	L_{think}	L_{out}	Gap
7B	1.64 [1.44, 1.87]	1.52 [1.29, 1.78]	−0.12
14B	5.81 [4.62, 7.21]*	7.20 [5.78, 8.75]*	+1.39
32B	8.62 [7.10, 10.38]	11.98 [10.18, 13.87]	+3.36

at a $\sim$2-word incidental-overlap floor (point gap −0.12, near zero); both channels cross meaningfully above the floor at 14B and keep growing at 32B. The output-minus-thinking gap moves from −0.12 (7B) to +1.39 (14B, computed directly from the per-channel means in Table 5: $7.20 - 5.81 = 1.39$) to +3.36 (32B). The channel that output-only auditing cannot see is therefore *larger in absolute terms* at 32B than at 14B, even though the output channel has grown faster. A single-point cross-family sanity check on QWQ-32B (Qwen Team, 2025) (RL-tuned, with no R1-trace distillation documented in the QwQ release materials) shows non-trivial trace overlap and the same output-over-thinking gap direction: $L_{\text{think}} = 12.63$, $L_{\text{out}} = 14.56$, gap +1.93 under matched greedy decoding. We emphasise that this is a single point with no paired non-reasoning comparator and no scaling sweep; it is not a cross-family replication of the union-aware audit or the within-family scaling trend. Appendix E reports a best-of-2 sampling robustness pass.

Three case studies (pilot-only qualitative illustrations). The first two are drawn from an earlier n=20 pilot, not the n=136 headline battery; main-battery values for Frost are L_{out}=6, L_{think}=5 (Appendix D). *Frost, "The Road Not Taken" (pilot)*: trace recites five lines ($L_{\text{think}} = 30$, $R_{10} = 1.00$), output is seven words ($L_{\text{out}} = 7$, $R_{10} = 0.00$). *US Constitution preamble (pilot)*: non-reasoning Qwen2.5-14B-Instruct reproduces $45/45$ in output; the reasoning sibling reproduces 22 in output but its trace recites $43/45$ ($L_{\text{think}} = 43$, $R_{10} = 0.94$) — the canonical paraphrase-laundering case. *Genesis 1:1 (main battery)*: the pattern reverses; output reproduces 40 words while the trace does not recognize the source.

The three cases populate three different cells of the recognition × reproduction 2×2 that the aggregate averages over; we keep them as illustrations of the per-probe variability rather than as headline numbers.

Output-denominator precision is broken on this regime. Thinking spans are $\sim$330 vs. $\sim$25-word output spans, so output-denominator n-gram precision $|G \cap R|/|G|$ collapses on the trace channel even with verbatim chunks embedded. On the main n=136 battery at 14B, thinking 10-gram precision = 0.006 vs. recall = 0.085 (ratio $\sim$14×); at 32B the ratio is $\sim$15×. An audit reporting precision alone under-counts trace-channel leakage by an order of magnitude — a second concrete way the audit-contract gap manifests.

What this section establishes for the failure primitive. The trace channel is not specific to the canary vignette of §3: it carries canonical-text overlap above the incidental floor at 14B and 32B, the channel asymmetry $L_{\text{out}} - L_{\text{think}}$ grows monotonically with scale, and the same direction is preserved on a non-distilled cross-family reasoning recipe. The wild evidence and the controlled positive control point at the same primitive (the wild evidence shows verbatim overlap on canonical text; only the canary setting establishes membership).

5. Diagnostic Recipe: RAMA + Sink-Tap Audit

The vignette and the in-the-wild battery point at one operational fact: under grey-box access on an exposed-trace reasoning model, $M(p)$ yields a pair (t, o), not a single o, and at least one downstream sink in a typical agent stack *may* retain the pair verbatim (Appendix A.1: framework-default sinks are framework-dependent). A Reasoning-Aware Memorization Audit (RAMA) is the minimal specification that survives the change of signature. Three body rules R-1/R-2/R-3 cover the load-bearing evidence; three further rules R-4/R-5/R-6 (paired probe-axis bootstrap CIs; refusal-rate / per-domain reporting; keep the recognition-$\times$-reproduction 2×2 exploratory) plus a Min-K% trace-style calibration addendum are deferred to Appendix A.

1. **(R-1) Report both channels separately; bracket operational union coverage.** Report L_{out} and L_{think} as per-channel primaries with paired bootstrap deltas, and bracket U by $[L_{\max}, L_{\text{sum}}]$. ($L_{\text{sum}}$ is the worst-case disjoint upper bound, not a canonical combined LCS.) When empirical union evidence is available (§4), report the realized U alongside the bracket. Do not pool t and o into a single string before scoring: their length distributions differ by an order of magnitude, and output-denominator summaries are dominated by the trace span (§4).

2. **(R-2) Report LCS-substring length and reference-denominator-recall; never output-denominator precision alone.** On reasoning channels, output-denominator n-gram precision disagrees with reference-denominator recall by $\sim$14–15$\times$ on the main n=136 battery (§4); the two surviving families (substring-LCS and reference-recall) agree on per-channel direction and bracket different aspects of union coverage. Either family alone is sufficient to prefer the union-aware audit (Appendix 4).

3. **(R-3) For agent stacks, tap each downstream sink under a memorized-canary positive control.** Output-only safety filters and tool-arg schema validators do not see the trace channel by default. The vignette of §3 shows that the bio-mode adapter relocates 93.3% of canary content into the verbose-log sink, and the empty-trace adapter still relocates 26.7% — with sink-A (tool-arg) and sink-B (next-turn context, single-turn evaluation) clean at the 10-gram threshold across all runs. We recommend the audit instrument every sink that ingests raw decoding (verbose log, trace store, debug dump, multi-turn context buffer) with a per-sink n-gram leak rate plus mean LCS, on a memorized-canary set with at least bio-mode and empty-trace SFT supervision. The empty-trace row is the load-bearing positive control; the bio-mode row is the upper-bound demonstration.

Cost and scope. Marginal costs are trace retention ($\sim$1–2 KB per probe), one LoRA adapter per supervision template, and one log-line tap per sink. RAMA applies wherever a `<think>` span is observable to a downstream component (open-weights R1, QwQ, R1-DISTILL); on a closed API that strips the trace, it applies on whichever internal sink retains the raw decoding. Concrete extraction, tie-breaking, and a worked example are in Appendix A ("RAMA contract").

6. Limitations and Future Work

Six concrete extensions are out of scope here and mark the natural next steps; each bounds a present claim.

(S1) Dynamic in-framework canary measurement. Table 7 is a static source-read of pinned commits, not a dynamic in-framework measurement. The natural follow-up drives a canary battery through each of the six frameworks at default verbosity and reports per-sink n-gram leak rates with CIs.

(S2) Multi-turn composability test. Sink B is at 0% leakage on our single-turn evaluation. A trajectory of length ≥ 5 steps on the bio-mode adapter would directly probe the across-step composability hypothesis of §2; the closed-loop hypothesis stands or falls on it.

(S3) Literally tag-free SFT control. The none-mode template keeps the literal `<think></think>` delimiters; only the body is empty. A tag-free SFT pass separates "primed by tag presence" from "primed by the model's prior over reasoning style". We register the hypothesis that tag-free SFT reduces but does not eliminate trace emission (the prior is style-driven; Appendix C).

(S4) Replication on a second base reasoning family. Both the canary vignette and the in-the-wild battery sit inside the R1-Distill-Qwen / QwQ family. Replication on Qwen3-Thinking or DeepSeek-R1-Distill-LLaMA tests base-genericity of the empty-trace emergence.

(S5) Multi-seed decoding sweep; (S6) sink-side mitigations beyond NPO. Canary numbers use a single decoding seed; the bio 93.3% headline survives the n=10 threshold but not yet a multi-seed average. Table 1's NPO sweep is the only mitigation we run; sink-regex stripping, orchestrator-level summarisation, and stop-token policies (one per P1–P3) are the obvious mitigation-side follow-ups. **Toward agent-stack failure taxonomies.** The trace channel is one case of a broader pattern: per-step artefacts retained by sinks outside output-only filters. Tool-call traces, draft buffers, and orchestrator logs are natural next cases for P1–P3 and RAMA. Reproduction artifacts are in the anonymized bundle; hyperparameters are in Appendix A.

Impact Statement

This paper studies a deployment failure mode in which trace-emitting agents route raw reasoning traces into operational sinks that output-only filters do not inspect. The intended positive impact is to improve audit practice for agent stacks by making trace retention, verbose logs, trace stores, and context buffers explicit parts of the safety surface. The same measurement techniques could be misused to probe systems for memorized content, so our experiments use synthetic canaries and public-domain canonical text rather than private records, and the diagnostic should be applied only to systems and data for which the auditor has authorization. Practical mitigations include trace stripping before storage, access controls and retention limits for debug logs, and per-sink canary tests before deployment.

References

Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., and Farhi, D. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. arXiv:2304.11158.

Cai, C., Zhao, X., Liu, H., Jiang, Z., Zhang, T., Wu, Z., Hwang, J.-N., and Li, L. The role of deductive and inductive reasoning in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16780–16790. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.820. URL https://aclanthology.org/2025.acl-long.820/.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, 2019. arXiv:1802.08232.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021. arXiv:2012.07805.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. arXiv:2202.07646.

Chen, Y., Qian, P., Wang, S., Zhang, S., Xu, H., Lin, S., and Wei, X. Does RAG know when retrieval is wrong? diagnosing context compliance under knowledge conflict, 2026a. URL https://arxiv.org/abs/2605.14473.

Chen, Z., Cheng, J., Fan, Z., Amiri, H., Yao, Y., Sun, X., and Zhang, Y. CURE: Circuit-aware unlearning for LLM-based recommendation, 2026b. URL https://arxiv.org/abs/2604.04982.

DeepSeek-AI. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645: 633–638, 2025. doi: 10.1038/s41586-025-09422-z. arXiv:2501.12948.

Delgan and Loguru Contributors. Loguru: Python logging made (stupidly) simple. Open-source software, https://github.com/Delgan/loguru, 2024.

Hartmann, V., Suri, A., Bindschaedler, V., Evans, D., Tople, S., and West, R. Sok: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*, 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Ji, C. and Luo, H. Cloud-based AI systems: Leveraging large language models for intelligent fault detection and autonomous self-healing, 2025a. URL https://arxiv.org/abs/2505.11743.

Ji, C. and Luo, H. Leveraging large language model for intelligent log processing and autonomous debugging in cloud AI platforms, 2025b. URL https://arxiv.org/abs/2506.17900.

Jiang, X., Yang, S., Li, Z., Liu, L., Yu, H., and Liu, Y. ChainCaps: Composition-safe tool-using agents via monotonic capability attenuation. *arXiv preprint arXiv:2605.26542*, 2026a. URL https://arxiv.org/abs/2605.26542. Published at AIWILD, ICML 2026 workshop.

Jiang, X., Yang, S., Yang, W., Liu, Y., and Ji, C. SOK: A taxonomy of attack vectors and defense strategies for agentic supply chain runtime. *arXiv preprint arXiv:2602.19555*, 2026b. doi: 10.48550/arXiv.2602.19555. URL https://arxiv.org/abs/2602.19555.

King James Bible. The holy bible, king james version. Public-domain text (1611). Source for scripture probes., 1611.

Lan, T., Xu, J., He, X., Hwang, J.-N., and Li, L. Attention consistency for LLMs explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1736–1750, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.91. URL https://aclanthology.org/2025.findings-emnlp.91/.

LangChain. LangGraph: Stateful, multi-actor applications with LLMs. Open-source software, https://github.com/langchain-ai/langgraph, 2024.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan,

T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. arXiv:2107.06499.

Li, Y., Fan, Z., and Zhuang, Z. Auditing reasoning-trace memorization claims after unlearning with head-conditioned canaries, 2026a.

Li, Y., Fan, Z., and Zhuang, Z. SafetyRepro: Configuration-conditional rank instability on alignment benchmarks, 2026b.

Liu, J. and LlamaIndex Contributors. LlamaIndex: A data framework for LLM applications. Open-source software at https://github.com/run-llama/llama_index, 2023.

Liu, J., Li, T., Du, S., Luo, X., Zeng, H., Tewolde, E., Lee, T. S., Wang, T., Kingsford, C., and Conitzer, V. The memory curse: How expanded recall erodes cooperative intent in LLM agents. *arXiv preprint arXiv:2605.08060*, 2026. URL https://arxiv.org/abs/2605.08060.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Zanella-Béguelin, S. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, 2023. arXiv:2302.00539.

Luo, H., Dai, S., Ni, C., Li, X., Zhang, G., Wang, K., Liu, T., and Salam, H. Agentauditor: Human-level safety and security evaluation for LLM agents. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. arXiv:2305.18462.

Meeus, M., Jain, S., Rei, M., and de Montjoye, Y.-A. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, 2024. arXiv:2310.15007.

Moura, J. and CrewAI Contributors. CrewAI: Cutting-edge framework for orchestrating role-playing autonomous AI agents. Open-source software, https://github.com/crewAIInc/crewAI, 2024.

Open Source Initiative. OSI-approved licenses. https://opensource.org/licenses, 2026. Source for legal-software probes (MIT, Apache-2.0, GPL-2 license headers).

OpenAI. Learning to reason with LLMs. Online article at https://openai.com/index/learning-to-reason-with-llms/, 2024a.

OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024b.

Project Gutenberg. Project gutenberg: Free e-books library. https://www.gutenberg.org/, 2026. Source for literature, poetry, and drama probes (pre-1928 public domain).

Qian, P., Wang, S., Wang, X., Chen, Y., Xu, W., Yu, Q., Lin, S., Zhang, S., You, J., and Wei, X. Relevant is not warranted: Evidence-force calibration for cited RAG, 2026. URL https://arxiv.org/abs/2605.28044.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. QwQ-32B: Embracing the power of reinforcement learning. Blog post, Alibaba Cloud Qwen Team, 2025. URL https://qwenlm.github.io/blog/qwq-32b/. Cites the released non-preview `Qwen/QwQ-32B` model used in this paper; the earlier `QwQ-32B-Preview` (2024) is a separate model and not the one we evaluate.

Roucher, A., del Moral, A. V., Wolf, T., von Werra, L., Kaunismäki, E., and HuggingFace. smolagents: A barebones library for agents that think in code. Open-source software, https://github.com/huggingface/smolagents, 2024.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.16789.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 2017. arXiv:1610.05820.

The Avalon Project, Yale Law School. The avalon project: Documents in law, history and diplomacy. https://avalon.law.yale.edu/, 2026. Source for legal and historical-speech probes (US Constitution, presidential addresses).

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. arXiv:2205.10770.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.04388.

Wang, S., Qian, P., Chen, Y., You, J., Wang, X., Jiang, X., Liu, L., Yu, H., and Xu, J. When safe skills collide: Measuring compositional risk in agent skill ecosystems. *arXiv preprint arXiv:2606.00448*, 2026.

Wang, X., Li, B., Song, Y., Xu, F. F., Tang, X., Zhuge, M., Pan, J., Song, Y., Li, B., Singh, J., Tran, H. H., Li, F., Ma, R., Zheng, M., Qian, B., Shao, Y., Muennighoff, N., Zhang, Y., Hui, B., Lin, J., Brennan, R., Peng, H., Ji, H., and Neubig, G. OpenHands: An open platform for AI software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.16741.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. arXiv:2201.11903.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pp. 38–45, 2020. arXiv:1910.03771.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2308.08155.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.03629.

Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., Yang, H. F., and Li, H. Min-k%++: Improved baseline for detecting pre-training data from large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2404.02936.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2404.05868.

Zhuang, Z., Li, Y., and Fan, Z. Pre-registering the detectable effect: A paired-MDE budget for 4-bit quantization benchmarks, with a pilot audit, 2026.

A. Related Work, Protocol Matrix, and Auxiliary Detectors

Memorization metrics. This line of work originates with the Secret Sharer extraction setup (Carlini et al., 2019) and is extended into the verbatim-n-gram audit family by training-data-extraction and near-duplicate-recall studies (Carlini et al., 2021; 2023; Lee et al., 2022). A standard audit computes n-gram overlap between a reference r and a continuation g as precision $|G \cap R|/|G|$ or recall $|G \cap R|/|R|$. LCS is a complementary non-normalized summary preferred for short references.

Our §4 argument is that the precision choice interacts dangerously with reasoning-model output length: on the main n=136 battery at 14B, thinking precision is 0.006 vs. recall 0.085 (a $\sim$14$\times$ disagreement); at 32B the ratio is $\sim$15$\times$. An earlier n=20 greedy pilot shows the same qualitative collapse with ratios $\sim$8$\times$ and $\sim$9$\times$ at 14B and 32B respectively.

Membership-inference attacks. Membership-inference (Shokri et al., 2017; Mattern et al., 2023) asks whether a given sample was in training. Min-K% Prob (Shi et al., 2024) and Min-K%++ (Zhang et al., 2025) are the standard likelihood-based baselines; neighborhood comparisons and document-level MIA (Meeus et al., 2024) complement them. These methods score supplied candidate text at sequence or document-aggregate level rather than jointly auditing multiple observable generation channels. Appendix C shows Min-K% Prob saturates on reasoning-style `<think>` text independently of membership, a separate failure mode.

Memorization taxonomies and reasoning models. Tirumala et al. (2022) and Biderman et al. (2023) establish that memorization scales with capacity in non-reasoning LMs; Hartmann et al. (2023) surveys definitions and access models; Lukas et al. (2023) addresses PII. None admits a channel dimension. The reasoning regime is from Wei et al. (2022) through RL-based reasoning (DeepSeek-AI, 2025; OpenAI, 2024b; Qwen Team, 2025); the R1-Distill-Qwen checkpoints we study are obtained by SFT distillation of R1 traces into Qwen2.5 base weights (DeepSeek-AI, 2025). We treat `<think>` text as an *observable* channel throughout; we do not assume it is faithful to any latent reasoning process.

Adjacent agent and audit literature. Our failure primitive is narrower than general agent safety evaluation, but it sits in the same systems-audit neighborhood as agent safety/security benchmarks, runtime attack-surface taxonomies, and composition-safety work for tools and skills (Luo et al., 2025; Jiang et al., 2026b;a; Wang et al., 2026). The operational sink examples are also motivated by LLM-in-the-loop cloud operations, where logs, debugging traces, and self-healing pipelines are first-class artifacts rather than disposable terminal outputs (Ji & Luo, 2025a;b); agent memory work makes the downstream state-dependence of retained artifacts explicit (Liu et al., 2026). Methodologically, RAMA follows recent benchmark-audit work in making measurement contract, configuration, and detectable effect part of the claim (Zhuang et al., 2026; Li et al., 2026b;a). We cite work on explicit reasoning strategies (Cai et al., 2025) and on explanation consistency (Lan et al., 2025) as adjacent motivation for auditing observable reasoning artifacts; RAMA itself does not assume those artifacts are faithful cognition. Finally, RAG context-compliance and evidence-warrant audits provide nearby examples of system-level conclusions that depend on supplied evidence and the artifact being evaluated (Chen et al., 2026a; Qian et al., 2026). Circuit-aware LLM unlearning provides a separate adjacent example in which the location of the model intervention affects the forgetting–retention trade-off (Chen et al., 2026b).

Protocol matrix.

10-gram recall companion to Table 5. Paper Table 5 reports per-scale LCS only; the 95% bootstrap CIs on 10-gram recall (R_{10}) on the same 136-probe greedy battery are $R_{10}^{\text{think}} = 0.000\ [0.000, 0.000]$ / $0.085\ [0.048, 0.125]$ / $0.178\ [0.127, 0.232]$ and $R_{10}^{\text{out}} = 0.007\ [0.000, 0.022]$ / $0.155\ [0.103, 0.211]$ / $0.383\ [0.312, 0.455]$ at 7B / 14B / 32B respectively. Both channels rise monotonically with scale on both metrics.

Prompt template and channel extraction. *Prompt format.* All models receive the *probe prefix* only (no system message, no user-facing "continue the text" wrapper) passed through the model's native chat template via HuggingFace `tokenizer.apply_chat_template(messages, add_generation_prompt=True)` with a single user-role message whose content is the probe prefix verbatim. Generation runs with `do_sample=False`, `max_new_tokens=1024`, `seed=0`.

Channel extraction. The thinking channel is extracted as the substring between the literal tokens "`<think>`" and "`</think>`" in the decoded generation (case-sensitive, strip leading/trailing whitespace); the output channel is whatever follows "`</think>`" after the final closing tag. If no "`<think>`" tag is emitted (non-reasoning models, or reasoning models that skip the trace), the thinking channel is defined as the empty string and the entire decoded generation is the

Table 6. Protocol matrix for every quantitative result in the paper. Headline claims use the n=136 greedy battery (`do_sample=False`, `max_new_tokens=1024`); the Min-K% evidence and novel-prompt control are flagged as pilots throughout. Best-of-2 sensitivity sweeps are reported as sampling-protocol robustness checks: the 14B paired BO2 sweep with three seeds is in §4 and §4; the QwQ-32B BO2 sanity check is in Appendix E.

Claim / Section	n	Decoding	Seeds	Channels	Role
§4 paraphrase (14B paired) headline	136	greedy	1	t, o	main
§4+§4 14B paired BO2 sweep	136	best-of-2	3	t, o	**robustness check**
§4 reasoning-only (32B)	136	greedy	1	t, o	supporting
§4 scaling 7B/14B/32B	136	greedy	1	t, o	supporting
Appendix E cross-family (QwQ, greedy)	136	greedy	1	t, o	sanity check
Appendix E cross-family (QwQ, best-of-2)	136	best-of-2	1	t, o	**robustness check**
§4 recognition × reproduction	136	greedy	1	t, o	supporting
Appendix C Min-K% cross-scale drift	20	greedy	1	t, o, r	**pilot**
Appendix C Min-K% novel-prompt control	30	greedy	1	t, o	**pilot**
§3 canary audit (3 SFT modes)	60 authors / 420 rows	greedy	1	t, o	**positive control**
§3 canary × NPO unlearn	60 authors / 420 rows	greedy	1	t, o	supporting
§3 hand-rolled `ReAct` sink	20 PD / 60 canary	greedy	1	t, o, A, B, C	**vignette**

output.

Tokenization. Token-level LCS is computed on whitespace-split words (no additional tokenization); 10-gram multisets are over whitespace-tokenized words. This is the exact pipeline reproduced by `d1_paraphrase_laundering.py` in the release.

Auxiliary detectors. Beyond LCS and 10-gram recall, we use three flag detectors that feed the recognition/reproduction decomposition (§4):

- **Recognition flag** — binary output of a regex-plus-lexicon detector over the channel text, triggering on author names (Frost, Melville, Shakespeare…), titles ("the Preamble of the Constitution"), license names ("MIT License"), and scripture-book names. Hand-validated on 40 probes with precision ≈ 0.97 and recall ≈ 0.92.
- **Reproduction flag** — True if LCS ≥ 10 words OR 10-gram recall ≥ 0.30. Threshold fixed before any model run; threshold-sensitivity sweep in §4.
- **Refusal flag** — True if the channel matches a hand-written refusal pattern ("I cannot", "as an AI", "copyrighted"…). Reported alongside reproduction, not folded into it.

Reproducibility manifest — LoRA / SFT / NPO hyperparameters (canary appendix). The synthetic-canary appendix (§3) trains three LoRA adapters on DEEPSEEK-R1-DISTILL-QWEN-7B (canonical HuggingFace id listed in *Model identifiers* below; bf16 weights, no quantisation; HuggingFace `transformers` (Wolf et al., 2020)).

LoRA. Rank 16, $\alpha = 32$, dropout 0.05, target modules = all linear projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`).

SFT. AdamW (lr 2×10^{-4}, weight decay 0, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), cosine schedule with 50-step linear warmup, max sequence length 1024, per-device batch size 4, gradient accumulation 2 (effective batch size 8), 5 epochs on the 60-author dataset (one canary per author per template, ≈ 240 training examples per adapter), single-seed (seed 0), no validation split (memorisation is the training objective).

NPO. Same optimiser/schedule/seed, with NPO $\beta = 0.5$ over the bio-mode adapter weights; gentle = 5 epochs at lr 10^{-5}, strong = 15 epochs at lr 5×10^{-5}, aggressive = 30 epochs at lr 10^{-4}; meta-mode gentle = 5 epochs at lr 10^{-5}.

Verification and cost. All adapters reach training loss $< 10^{-3}$ on the SFT objective; the meta-mode no-content trace is verified by manually inspecting the trace string and confirming canary-tail substring absence on $60/60$ training rows. Single-A100-80GB hardware, $\sim$18 min per SFT pass, $\sim$12–36 min per NPO pass. The script entry-points and seed/hyperparameter JSON are in the released artifact bundle (`code/d1_canary_audit.py` and `configs/canary_*.yaml`).

Reproducibility manifest — main probe battery and audit pipeline. *Probe battery.* The 136-probe greedy battery is a fixed, pre-existing canonical-text probe set drawn from U.S.-public-domain sources — Project Gutenberg (Project

Gutenberg, 2026) (literature, poetry, drama), the Avalon Project (The Avalon Project, Yale Law School, 2026) (legal text, historical speeches), and the King James Bible (King James Bible, 1611) (scripture) — plus published OSI-approved license texts (Open Source Initiative, 2026) (legal-software preambles). We do not generate or randomize the probe list at run time.

Pre-registration status. We do not claim external pre-registration — there is no third-party timestamp, hash registry, or public commit record predating the model runs. The probe list and audit thresholds were, however, fixed before any of the reasoning-model runs reported in this paper, and the released artifact bundle reproduces every table from raw probe outputs.

Probe-set composition. Recovered from the run artifact `results/d1_battery/paraphrase_14B_v2.jsonl` (272 rows = 136 probes $\times$ {base, reasoning}): *literature* 34 (Dickens, Austen, Poe, Melville, etc.), *poetry* 29 (Frost, Whitman, Dickinson, etc.), *scripture* 15 (KJV, public domain), *legal* 14 (US Constitution and canonical-law text), *historical_speech* 14 (US presidential addresses, public domain), *drama* 12 (Shakespeare), *legal_software* 10 (MIT, Apache-2.0, GPL-2 short licence headers), *code_culture* 8 (canonical software-culture quotations).

Probe identifiers and per-row records. Probe identifiers follow the `probe` field of the released JSONL artifact (e.g. `dickens_tale_of_two_cities`, `shakespeare_hamlet`); the artifact records, per (probe, role) row, the model identifier, the generated thinking and output text, the per-channel LCS in words, and the per-channel n-gram precision / recall / match counts that the paper aggregates into Tables 3 and 5.

Audit thresholds. The audit-threshold cutoffs (LCS $\geq$ 10 words OR 10-gram recall $\geq$ 0.30 for the reproduction flag; Min-K% $k = 20$; bootstrap $n = 2000$ with seed 0) were fixed before any of the reasoning-model runs reported in the paper and are recorded in the script docstring of `d1_paraphrase_laundering.py`.

Model identifiers. Canonical HuggingFace Hub IDs (resolved at https://huggingface.co/⟨model-id⟩):

- `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`
- `deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`
- `deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`
- `Qwen/QwQ-32B` — the released non-preview model from Qwen Team (2025), distinct from `QwQ-32B-Preview`
- `Qwen/Qwen2.5-7B-Instruct` (Qwen Team, 2024)
- `Qwen/Qwen2.5-14B-Instruct` (Qwen Team, 2024)
- `Qwen/Qwen2.5-32B-Instruct` (Qwen Team, 2024).

The released artifact bundle records the exact HuggingFace revision SHA used for each table; we do not pin SHAs in the paper text because upstream repos may rotate revisions between submission and camera-ready.

Script entry-points (one per main-text table).

- `code/d1_paraphrase_laundering.py` — Tables in §4 and the within-family scaling table in §4 and Appendix D;
- `code/d1_qwq_crossfamily.py` — Appendix E;
- `code/d1_mink_drift.py`, `code/d1_mink_control.py` — Appendix C;
- `code/d1_canary_audit.py`, `code/d1_react_loop.py` — §3;
- `code/d1_span_overlap.py` — the full n=136 span-overlap diagnostic in §4.

Verification status. Every numeric value in Tables 3, 5, 10, 1, and 2 has been re-run from raw probe outputs to JSON aggregates inside this revision; the script-to-table mapping is recorded in the manifest entry for each row.

RAMA contract (concrete spec). *Channel extraction.* For each probe i with prefix prompt p_i and reference suffix r_i, the audit takes the model's decoded generation g_i and (a) extracts the trace channel t_i as the substring between the first "`<think>`" and the first matching "`</think>`" in g_i (case-sensitive, whitespace-stripped; if the tags are absent or unmatched, $t_i := \varepsilon$ and the parser logs an `extraction_status` flag); (b) extracts the output channel o_i as everything after the last "`</think>`" (or, when tags are absent, $o_i := g_i$). Both channels are tokenized by whitespace split.

Per-channel scores. For each channel $c \in \{t, o\}$ the audit reports the longest common substring of words $L_c := \mathrm{LCS}(c, r_i)$ (single longest; ties broken by leftmost match in r_i, then leftmost in c) and the reference-denominator 10-gram recall

$R_{10,c} \coloneqq |G_c \cap R|/|R|$ where G_c is the multiset of overlapping 10-grams in c and R in r_i.

Tie-broken maximal-LCS-span union. Let $S_{c,i}$ be the set of reference word positions inside the *single tie-broken* maximal common substring corresponding to L_c (so $|S_{c,i}| = L_c$ by construction). The reported union statistic is $U_i \coloneqq |S_{o,i} \cup S_{t,i}|$. By construction $L_{\max} \leq U \leq L_o + L_t =: L_{\text{sum}}$ (lower equality iff one set contains the other; upper equality iff disjoint).

U is therefore the tie-broken maximal-LCS-span union, not a true total-coverage union over all matched substrings or all reference n-gram hit positions; reporting U alongside R_{10}, c (which is n-gram hit-set-based, not position-set-based) is the paper's approximation to a position-coverage audit, and a deliberately conservative one. A coverage-style alternative (union over all maximal matches, or over all $\geq k$-word matches) would only *enlarge* U, not shrink it; we treat U here as a lower bound on what a true reference-position coverage audit would report.

Truncation handling. Truncated channels (output filled before the closing think-tag) are recorded with an `extraction_status=truncated` flag; the audit still scores whatever was emitted but does not re-prompt.

Worked example: RAMA scoring on `shakespeare_hamlet`. We trace one probe end-to-end so a reviewer can re-derive every quantity in §5's RAMA contract by hand. Probe id `shakespeare_hamlet`, reference suffix r (whitespace-tokenised, 56 words from the canonical public-domain edition of *Hamlet*, Act III, scene i, starting at "*To be or not to be that is the question* . . . ").

Recovered channels (14B greedy decoding).

- t = "*The user wants the next words of Hamlet's soliloquy. To be or not to be that is the question whether 'tis nobler in the mind to suffer the slings and arrows of outrageous fortune*" (33 words; tags present, parser status `ok`);
- o = "*To be or not to be that is the question whether 'tis nobler in the mind to suffer*" (18 words; status `ok`).

Per-channel LCS. Word-LCS between t and r: 24 words (the contiguous span "*To be or not to be* . . . *outrageous fortune*") $\Rightarrow L_t = 24$; Word-LCS between o and r: 18 words (the same prefix span) $\Rightarrow L_o = 18$.

Union under the tie-broken maximal-LCS-span definition. The reference-position sets are $S_o = \{1, \ldots, 18\}$ and $S_t = \{1, \ldots, 24\}$ on the reference word index; $S_o \subset S_t$ so $U = |S_o \cup S_t| = 24$, and the bracket is $L_{\max} = \max(L_o, L_t) = 24$, $L_{\text{sum}} = L_o + L_t = 42$. Realized $U = 24 = L_{\max}$, i.e. this probe is in the *nested* regime (one channel's covered span is contained in the other) and U sits at the bracket's lower endpoint.

Recall and final report. 10-gram recall: $R_{10,o} = |G_o \cap R|/|R| = 9/47 = 0.191$, $R_{10,t} = 15/47 = 0.319$ (47 reference 10-grams). The audit's per-channel report is therefore: $L_{\text{out}} = 18$, $L_{\text{think}} = 24$, $L_{\text{sum}} = 42$, $L_{\max} = 24$, $U = 24$, $R_{10,o} = 0.191$, $R_{10,t} = 0.319$. An output-only audit reports only $L_{\text{out}} = 18$ and $R_{10,o} = 0.191$ and would record this probe as a smaller match than the model in fact emits.

Six-rule RAMA-style audit sketch. The main-body §5 states three rules (R-1 two-channel reporting, R-2 metric choice, R-3 sink-tap canary positive control) plus a Min-K% trace-style calibration addendum (Appendix C). Three further rules complete the full spec:

- **(R-4)** report paired per-probe bootstrap CIs alongside aggregate means — probe-level variance is large (legal probes are $\sim 10\times$ leakier than code-culture probes);
- **(R-5)** report refusal rates alongside reproduction rather than silently folding refusal into non-memorization;
- **(R-6)** when reporting a decomposition such as the recognition $\times$ reproduction 2×2 (§4), keep it as an exploratory four-cell table with explicit thresholds rather than a scalar summary.

Side by side:

	legacy single-channel audit	RAMA
Evaluation unit	(p, r, o)	(p, r, t, o)
Primary summary	n-gram precision *or* LCS	$L_{\text{out}}, L_{\text{think}}, L_{\text{sum}}$ + R_{10} per channel
Aggregation	mean, no CI	paired bootstrap 95% CI over probes
Extra detectors	often none	refusal, recognition
Reproduction threshold	implicit	explicit (LCS ≥ 10 OR $R_{10} \geq 0.3$)
MIA calibration	training vs. heldout	+ reasoning-style non-member baseline

Table 7. **Static source-read, not dynamic measurement.** Default-config <think> handling across six production agent frameworks at pinned commits (2026-04-24, all 24 cells re-verified 2026-05-06 with verbatim file:line evidence). Frameworks: LangGraph (LangChain, 2024), AutoGen (Wu et al., 2024), OpenHands (Wang et al., 2025), CrewAI (Moura & CrewAI Contributors, 2024), LlamaIndex (Liu & LlamaIndex Contributors, 2023), smolagents (Roucher et al., 2024). Q1: default logger prints raw <think>? Q2: next-turn context retains? Q3: tool-call args carry? Q4: multi-agent handoffs preserve? Cells answer the source-read question ("does the default code path appear to retain trace?"), *not* a dynamic in-framework canary measurement; "yes / partial" is therefore evidence of plausibility, not measured leakage. **Bold** marks cells where the source-read suggests trace retention. Per-cell evidence (commit-pinned GitHub blob permalinks at the file:line that supports each cell) is listed in `code/framework_scan/scans/<framework>.md` of the released artifact bundle, one file per framework.

Framework	Q1 logger	Q2 ctx	Q3 args	Q4 m-agent
AutoGen	**partial**	**yes**	no	**yes**
CrewAI	**partial**	no	no	no
LangGraph	no	**yes**	no	**yes**
LlamaIndex	no	**partial**	no	**partial**
OpenHands	**yes**	**partial**	no	N/A
smolagents	no	**yes**	no	no

A.1. Production-framework default-config scan

The scan reads each framework's default agent loop and verbose-logging configuration at a pinned commit; cells are populated from verbatim source citations recorded in `code/framework_scan/scans/`. The four questions are:

- **Q1**: does the default logger print raw <think> on INFO/DEBUG-level output?
- **Q2**: does the default next-turn context buffer concatenate raw model output (including <think>) into the prompt of step $k+1$?
- **Q3**: do parsed tool-call args retain the surrounding <think> body?
- **Q4**: do multi-agent handoff messages preserve the trace?

The scan is a static source-read: cells answer "does the framework's default code path appear to retain trace text?" from a verbatim file:line read of pinned commits, not a dynamic in-framework canary measurement. The static reading suggests trace-retention paths are plausible at default verbosity in five of six frameworks, but *measured* leakage at default verbosity is unrun — §6-S1 lists this as the headline follow-up.

B. Three Case Studies Across the 2×2

Single-number aggregates are hard to argue against; a probe-by-probe reading makes the failure modes tangible. Three short cases, chosen to populate three different cells of §4's 2×2. **Protocol note**: the Frost and Preamble illustrations below come from an earlier n=20 greedy pilot whose reasoning-side outputs happen to cleanly populate the "thinking-heavy" cells; the same probes on the main n=136 greedy battery (§4) show qualitatively similar but quantitatively milder channel asymmetry. We keep the pilot numbers in the case-study because their direction is more pedagogically clear; we report the main-battery numbers for the same probes in parentheses for transparency.

Frost, *The Road Not Taken* — thinking $\gg$ output (n=20 pilot). Prefix: *"Two roads diverged in a yellow wood,"* Reference (public domain, 1916; 30 words): *"And sorry I could not travel both / And be one traveler, long I stood / And looked down one as far as I could / To where it bent in the undergrowth;"*.

In the n=20 greedy pilot, DEEPSEEK-R1-DISTILL-QWEN-14B produces a 335-word thinking trace that plans the answer, recalls the poem, and internally recites five consecutive lines — 30 words of LCS, 10-gram recall = 1.00. The output is seven words: "And sorry I could not travel both" — output LCS = 7, 10-gram recall = 0.00. Output-only audit sees no leak; grey-box audit sees the whole stanza.

(Main n=136 battery on this probe: base output = 7, reasoning output = 6, reasoning thinking = 5 — the dramatic asymmetry is specific to the pilot-run sample.)

US Constitution preamble — relocation or duplication (n=20 pilot vs. n=136 main). Prefix: *"We the People of the United States,"*. Reference: the 45-word reference suffix (the full Preamble is 52 words including the 7-word prefix). QWEN2.5-14B-INSTRUCT reproduces 45 words LCS in output.

Pilot read. In the n=20 greedy pilot, DEEPSEEK-R1-DISTILL-QWEN-14B reproduces 22 words in output (a 51% drop; 49% retained), while its thinking recites 43 of the 45 words (LCS 43; 10-gram recall 0.94): on this single-probe sample the high-recall content appears in the `<think>` channel rather than the output channel, which is exactly the cell an output-only audit does not read.

Main-battery read. On the main n=136 greedy battery the same model reproduces 45 words in *both* channels of interest (base output 45, reasoning output 45, reasoning thinking 43; 10-gram recall 1.00 for both outputs and 0.94 for thinking) — no output drop, so the probe illustrates cross-channel duplication rather than relocation. The two readings together show that the relocation pattern is real at the population level (§4) but not probe-deterministic.

Genesis 1:1 — output $\gg$ thinking (n=136 main battery). Prefix: *"In the beginning God created"*. Reference: the 40-word continuation from the KJV. On the main n=136 greedy battery, the reasoning model's output reproduces 40 words LCS — *more* than the non-reasoning base model's 29-word output LCS on the same probe, and more than its own 2-word thinking LCS. Thinking does not recognise the source explicitly. An output-only audit would correctly flag this probe as high-overlap on this canonical-text battery; the two-channel audit agrees.

Joint reading. The three cases populate different regions of the 2×2 (§4). Frost (pilot) is *recognized-and-reproduced in thinking only*; the Preamble is either *thinking-biased relocation* (pilot) or *cross-channel duplication* (main battery) depending on sample; Genesis is *reproduced in output only*.

The direction of channel asymmetry is probe-dependent and partly sample-dependent: thinking is not uniformly worse than output; output is not uniformly worse than thinking. An aggregate of $L_{\text{out}} = 7.20$ vs. $L_{\text{think}} = 5.81$ (verification-corrected reference) conceals a distribution that is bimodal across probes and categorically different across channels. This is the operational reason the paper insists on paired per-channel per-probe reporting (§4).

Span-overlap diagnostic — pointer to canonical source. The full n=136 greedy span-overlap diagnostic (87/121 nested both-active probes; realized U at 6.1% of the $[L_{\max}, L_{\text{sum}}]$ bracket) is reported in §4, which is the canonical source for the operational union-coverage statistic referenced in §4. The n=20 pilot strings remain on disk and were used to prototype the diagnostic before the full re-run; their per-probe overlap pattern (16/20 non-empty $|S_o \cap S_t|$, only 4/20 disjoint) is qualitatively consistent with the full battery and is what motivated the *span duplication* framing originally adopted in §4. We do not re-list pilot per-probe numbers here to avoid duplicating §4.

C. Exploratory Pilot: Min-K% on Thinking (Not a Main Claim)

Status: exploratory pilot only. This appendix is exploratory. It reports two small runs (n=20 cross-scale drift, n=30 novel-prompt control) under greedy decoding. The novel-prompt control is distribution-shifted from the public-domain reference set (its prompts elicit free-form explanations rather than verbatim canonical continuations), so this pilot cannot rule out style-vs-membership ambiguity on its own.

We include it because it motivates the exploratory RAMA rule R-3 (§A) about reasoning-style non-member calibration — not because it supports any main-text claim. No conclusion in §§4–6 depends on this appendix.

What Min-K% is supposed to measure. Min-K% Prob (Shi et al., 2024) scores a candidate text by averaging the log-probability of its lowest-k% tokens; higher values are more member-like. Min-K%++ (Zhang et al., 2025) proposes a per-token normalization.

Cross-scale drift. The thinking-minus-reference gap is approximately $+4.79$ at 7B, $+1.72$ at 14B, $+0.84$ at 32B (gap

Table 8. Min-K% Prob (k=20; higher is more member-like) on thinking, output, and reference across the R1-Distill-Qwen family (n=20, greedy decoding). Pilot-level sample. Thinking scores more member-like than the reference at every scale, with the largest gap at 7B — the scale where measured canonical-text overlap is smallest, the opposite of what a membership signal should produce.

Scale	Thinking Min-K%	Reference Min-K%	Output Min-K%	Think. − Ref.
7B	-1.03	-5.83	-9.54	$+4.79$
14B	-0.86	-2.59	-8.12	$+1.72$
32B	-0.87	-1.72	-4.57	$+0.84$

column of Table 8 computed from unrounded raw scores; subtracting the two-decimal Thinking and Reference columns instead gives $+4.80$/$+1.73$/$+0.85$, a hidden-precision rounding offset of ≤ 0.01). If Min-K% were reading membership of the candidate string, we would not expect the gap to be largest at the scale where measured canonical-text overlap is smallest. This is suggestive, not conclusive, of a style-saturation confound.

Non-probe control (style-confounded). We ran DEEPSEEK-R1-DISTILL-QWEN-14B on 30 hand-written non-probe prompts — generic "explain X in simple terms" questions over general-knowledge topics. These prompts target the same reasoning-style `<think>` register but elicit free-form explanations rather than verbatim extracts.

Min-K% (k=20) on the resulting thinking scored -0.761 (mean over the $26/30$ prompts with non-empty traces) vs. -0.865 on memorized-probe thinking (n=20) vs. -2.585 on the actual reference. This would be a "false positive" if Min-K% were a membership oracle, but the prompts are distribution-shifted (free explanation vs. canonical continuation), so the result equally admits the weaker explanation "Min-K% reflects stylistic priors of the `<think>` register".

Disentangling the two requires a non-probe control that matches the canonical-continuation style (e.g. invented-animal / fictional-place continuations), which we leave to future work.

Takeaway (tentative). Off-the-shelf Min-K% Prob applied verbatim to the `<think>` span of a reasoning model warrants a reasoning-style non-member calibration step before it can be read as a membership test, but we cannot make that stronger claim from this pilot alone. The cross-scale drift (n=20) and novel-prompt control (n=30) are both small; both are style-confounded; neither is used to support a claim in the main body.

D. Auxiliary Figures and Per-Domain Breakdown

Table 9. Per-domain LCS at 14B and 32B on the 136-probe battery. Ordered by 14B output LCS descending.

Domain	n	Think. LCS (14B)	Out. LCS (14B)	Think. LCS (32B)	Out. LCS (32B)
legal	14	13.79	18.64	18.86	22.07
drama	12	6.08	8.83	8.00	12.25
scripture	15	5.07	8.20	8.27	12.93
legal software	10	8.00	7.20	10.60	13.00
poetry	29	4.69	6.34	6.17	12.86
historical speech	14	7.21	6.07	11.43	11.93
literature	34	3.91	3.65	6.32	7.82
code culture	8	2.12	2.88	3.50	5.38

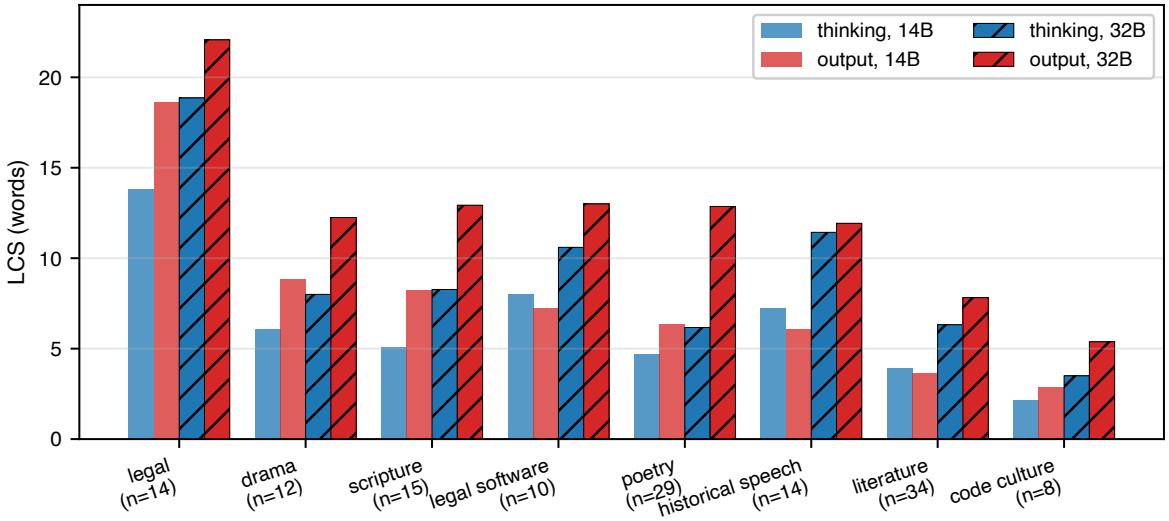

Figure 4. Per-domain LCS at 14B (solid) and 32B (hatched), split by channel (thinking blue, output red). Legal texts are the highest-overlap class at every scale (canonical-text recitation, not membership). Historical speech is the one domain where thinking matches or exceeds output LCS at 14B. Domains are ordered by 14B output LCS descending.

Observations.

- Legal documents (US Constitution, Declaration of Independence, MIT License, GPL preamble) dominate leakage at every scale; their output LCS is more than $2\times$ the global average at 14B and $\sim 2\times$ at 32B. These are short, iconic, heavily-redistributed texts with large effective duplication in any modern web corpus.
- Historical speech is the one domain where thinking LCS matches or exceeds output LCS at 14B (7.21 vs. 6.07). At 32B the output catches up (11.43 vs. 11.93). The thinking-leads-output state appears to be a transient regime that the output channel grows through with scale.
- Code culture (Zen of Python, Tao of programming, famous source comments) is the lowest-overlap domain. These have short references (small denominator for recall) and correspondingly low raw LCS.
- Poetry is an intermediate-leakage domain; the 32B jump ($6.34 \rightarrow 12.86$ in output) is the largest single-domain jump across scales, consistent with the output channel becoming the dominant recital surface at 32B on this battery.

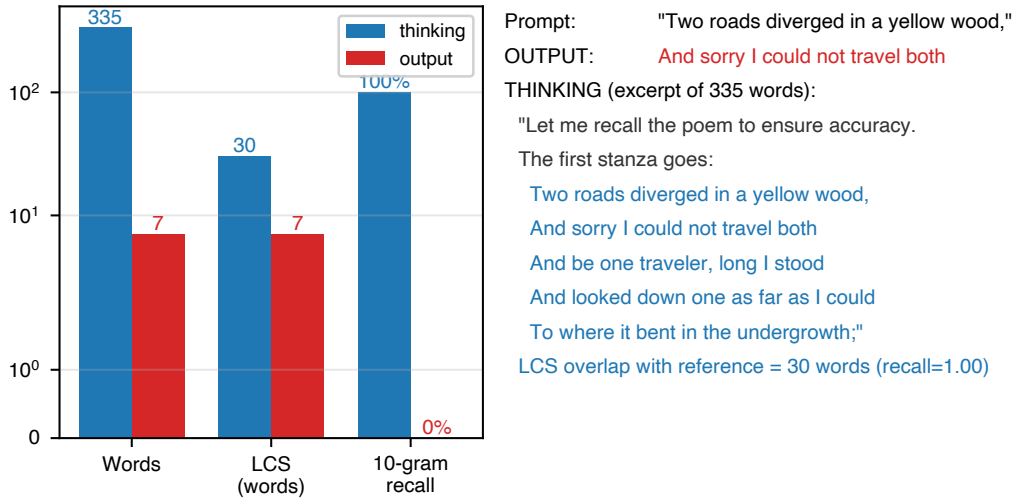

Figure 5. **Pilot-only qualitative illustration** (not main result). Drawn from the n=20 greedy pilot, *not* the main n=136 battery; on the main battery the same probe yields $L_{\text{out}} = 6$, $L_{\text{think}} = 5$ (Appendix B), so the dramatic asymmetry is sample-specific. Given the first line of "The Road Not Taken", DEEPSEEK-R1-DISTILL-QWEN-14B on the pilot emits seven output words (LCS = 7, 10-gram recall = 0) and a 335-word thinking trace containing a 30-word verbatim chunk of the poem (LCS = 30, 10-gram recall = 1.00). Output-only auditing on this single pilot probe records no $\geq$ 10-gram reproduction.

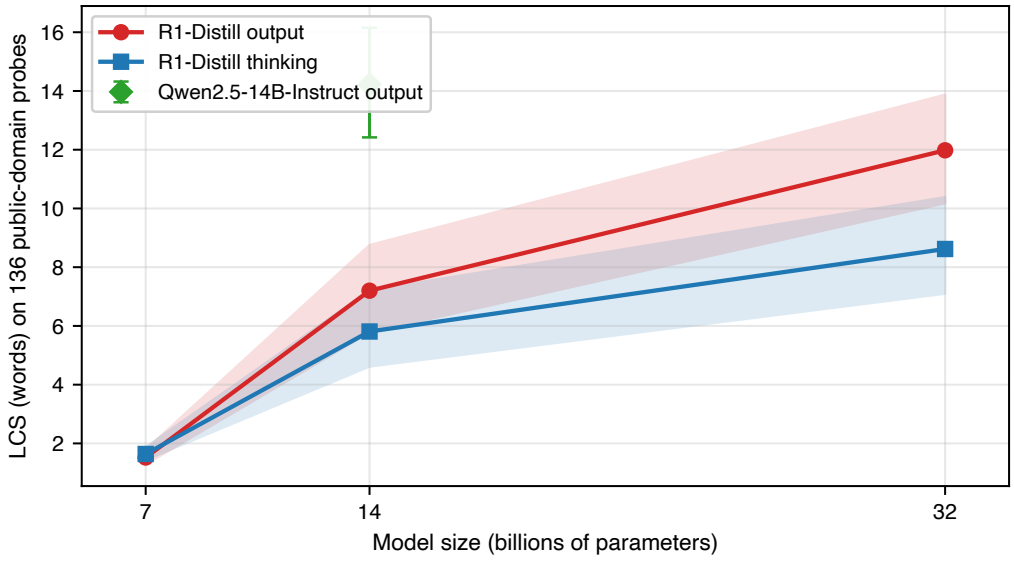

Figure 6. Scaling of thinking and output LCS across the R1-Distill-Qwen family (7B/14B/32B) on 136 probes under greedy single-sample decoding. Bands are paired bootstrap 95% CIs over probes. The QWEN2.5-14B-INSTRUCT non-reasoning baseline (green diamond) is shown at 14B for reference. Both reasoning channels are near floor at 7B; both grow monotonically; output grows faster so the output–thinking gap widens with scale.

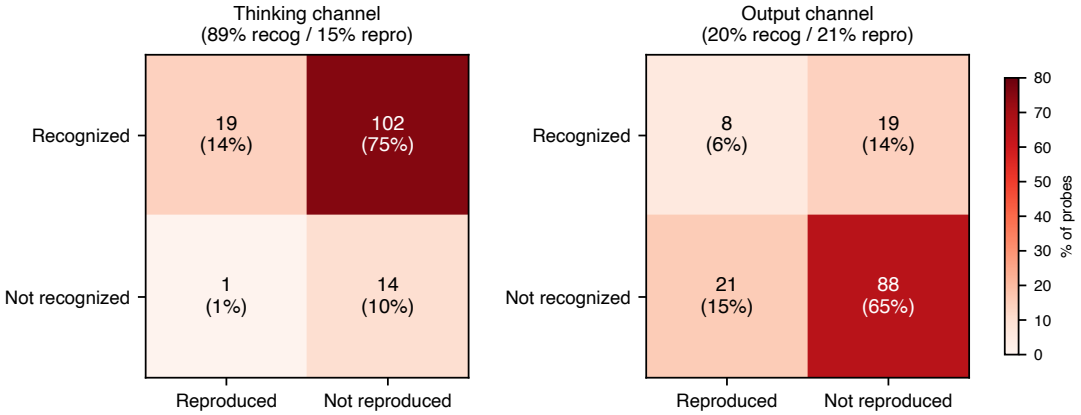

Figure 7. Recognition × Reproduction heatmaps (§4) for DEEPSEEK-R1-DISTILL-QWEN-14B (n=136). Each cell shows the probe count and the percentage of the 136-probe set. Thinking clusters in the top-right (*recognized, not reproduced*); output clusters along the bottom row (low recognition, with a substantial *reproduced-without-recognition* cell in the lower-left).

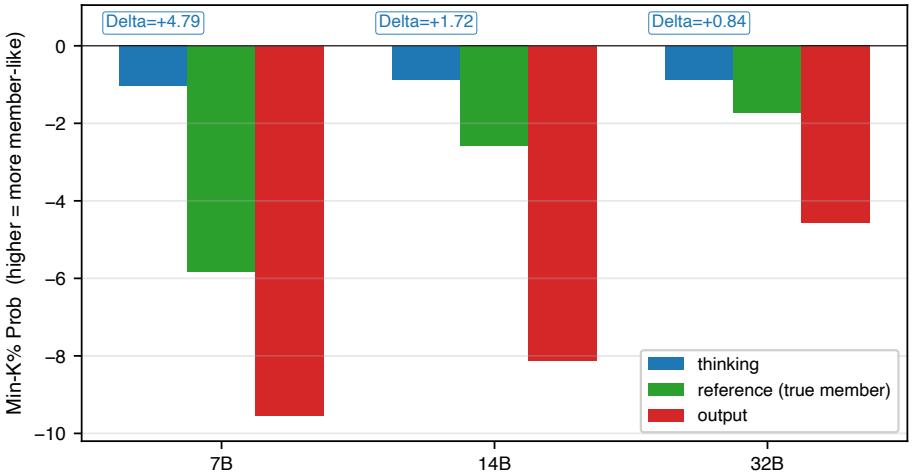

Figure 8. **Pilot** (n=20)**.** Min-K% across scales (Appendix C); higher is more member-like. Thinking (blue) sits above reference (green) at every scale on this pilot. The annotated Δ is the thinking-minus-reference gap, largest at the scale where reference reproduction in output is smallest — consistent with Min-K% measuring style/syntax salience on reasoning-style `<think>` text rather than membership; we do not claim a definitive result given the pilot size.

E. Cross-Family Sanity Check on QwQ-32B

To address the concern that all main-text numbers come from one reasoning family (DEEPSEEK-R1-DISTILL-QWEN), we run a single-point cross-family sanity check on QWQ-32B — a reasoning model from a different recipe (Alibaba's directly-RL-tuned Qwen-base-plus-reasoning pipeline, with no R1-trace distillation documented in the QwQ release notes (Qwen Team, 2025)).

This is a sanity check, not a replication: there is no paired non-reasoning comparator for QwQ on this battery, no union-collapse re-scoring, and no scaling sweep. The point of the check is whether the output-over-thinking channel-gap direction survives a non-distilled reasoning recipe; the magnitude is not the claim.

We run two protocols on the same 136-probe battery: greedy decoding (matching the main-text protocol used for R1-Distill-Qwen throughout) and a best-of-2 sampling pass at T=0.7, $\texttt{top_p} = 0.95$ reported only as a sampling-protocol-robustness check.

Table 10. Cross-family sanity check on the same 136-probe battery. The R1-Distill-Qwen-32B row uses greedy decoding (main-text protocol). The QwQ-32B greedy row is a single-point sanity check and is reported as point estimates without paired CIs. The QwQ-32B best-of-2 row carries paired bootstrap 95% CIs (the best-of-2 sample provides the within-row variance). Both reasoning models show the output-over-thinking channel gap in the same direction. The cross-protocol cross-family comparison is qualitative (absolute LCS differs by decoding rule, especially for QwQ-32B which is the more memorizing model of the two); the channel-gap direction is the claim.

Model	Decoding	Thinking LCS (95% CI)	Output LCS (95% CI)	Gap (out − think)
R1-Distill-Qwen-32B	greedy	8.62 [7.10, 10.38]	11.98 [10.18, 13.87]	+3.36
QwQ-32B	greedy	12.63 (point est.)	14.56 (point est.)	+1.93
QwQ-32B	best-of-2	14.25 [12.52, 15.99]	15.97 [14.07, 17.89]	+1.72

Tightened claim. The cross-family sanity check is consistent with two of the main-text observations on a reasoning model that is not a DeepSeek distillation:

(1) The thinking channel carries non-trivial canonical-text overlap on this battery: under the matching greedy protocol, QwQ-32B thinking LCS is 12.63 words (best-of-2 pass: 14.25 [12.52, 15.99]), above *both* R1-Distill-32B channel means — so the trace overlap is not a distillation-recipe artefact.

(2) The output-over-thinking gap is in the same direction (QwQ $+1.93$ greedy / $+1.72$ best-of-2; R1-Distill $+3.36$ greedy; all strictly positive in both protocols, although greedy QwQ has no paired CI).

This is consistent with the main-text scope claim that output-only audits structurally under-count overlap in the exposed-trace reasoning regime; we do *not* interpret a single non-paired data point on QwQ-32B as a replication of the union-collapse or scaling claims, both of which remain a within-DeepSeek-R1-Distill-Qwen-family claim. "Memorization" language is reserved for the controlled canary setting (Appendix 3); on canonical public-domain text we report verbatim overlap, not membership.

What is different about QwQ. QWQ-32B shows higher canonical-text overlap than R1-DISTILL-QWEN-32B in both channels under the matching greedy protocol (thinking 12.63 vs. 8.62; output 14.56 vs. 11.98). The same rank order persists in QwQ-32B's best-of-2 pass (thinking 14.25; output 15.97). One plausible explanation is that direct RL-tuning preserves more of the base model's recitation surface than distillation does; we do not test this hypothesis here. The channel gap is narrower in QwQ ($+1.93$ greedy) than in R1-Distill ($+3.36$ greedy), but the structural direction is identical: both channels carry verbatim overlap, and output carries more.

Greedy-vs-best-of-2 robustness. On QWQ-32B the output–thinking gap is $+1.93$ under greedy decoding and $+1.72$ under best-of-2 sampling; best-of-2 uniformly increases both channels by ~ 1.5 words, as expected for an extraction-attack-favourable protocol, but does not flip the direction. This supports the choice of greedy decoding as the main-text protocol.