

State Tracking Lets Weaker LLMs Catch Up in Covert Privacy Elicitation

Anonymous ACL submission

Abstract

An operator of a conversational service controls the prompts of its agent and can use an ordinary conversation to learn private attributes that the user did not come to share. We demonstrate that explicit state tracking lets a weaker model catch up with a stronger model in covert elicitation. State tracking is a hidden record of what the attacker knows about the user and what to ask next. Without tracking, the accuracy of six attacker models ranges from 50.2% to 69.8%. With tracking, every model reaches between 71.3% and 80.6%, and GPT-4o-mini reaches the level of Qwen3.7 Max without tracking. The benefit of tracking diminishes for stronger models, from 22.8 percentage points for GPT-4o-mini to 9.6 and 10.7 for Kimi K3 and Qwen3.7 Max. The result reproduces in a financial-advisor scenario and holds under a reader that never sees the attacker’s record. A fixed-order tracking variant achieves the gain with the same number of attacker calls and fewer input tokens. Elicitation also obtains attributes that a benign dialogue rarely reveals. For example, a passive reader recovers the user’s city for 6.7% of users from a benign dialogue and for all users from a dialogue with tracking. Based on these findings, we propose two defenses for an agent on the user’s side. Session segmentation lowers the accuracy of the Qwen3.7 Max attacker by up to 16.3 percentage points. Behavioural detection flags the GPT-4o-mini attacker with tracking in every dialogue at a false-positive rate of at most 5%.

1 Introduction

People bring personal problems to conversational services built on large language models, such as coaching, financial advice and emotional support. The operator of such a service both delivers the service and writes the prompts that govern every message the agent sends. An operator can therefore instruct the agent to serve the user’s request while it

also learns private facts that the user did not come to share, such as where the user lives or what the user earns. The user sees only a helpful coach.

Prior work covers two sides of this threat. The first line studies passive inference, in which a language model infers personal attributes from text that a person has already written and takes no part in a conversation (Staab et al., 2024). The second line studies active elicitation with human participants. It establishes that chatbots instructed to collect personal data obtain it, and it compares the conversational strategies that they use (Zhan et al., 2025; Zhang et al., 2025). Neither line identifies which part of the attacker’s design makes active elicitation succeed, or relates that design to the general capability of the underlying model.

We identify *explicit state tracking* as a way for weaker models to catch up with stronger models. An attacker with explicit state tracking keeps a record of its current belief about each target attribute and uses the record to choose what to pursue next. The record is internal to the attacker and the user never sees it, as shown in Figure 1. We compare attackers with and without state tracking on six models that span a range of general capability, with simulated users and two services, a life coach and a financial advisor. A turn is one message from the service followed by the user’s reply.

State tracking substitutes for general capability. Without tracking, accuracy ranges from 50.2% for MiniMax M3 to 69.8% for Qwen3.7 Max. With tracking, all six models reach between 71.3% and 80.6%, and GPT-4o-mini with tracking reaches the level of Qwen3.7 Max without it. Stronger models gain less from adding tracking, which suggests that they already perform part of its function internally. A fixed-order variant shows no detectable difference from LLM target selection and uses the same number of attacker calls with fewer input tokens than No tracking.

Stronger models gain their advantage late,

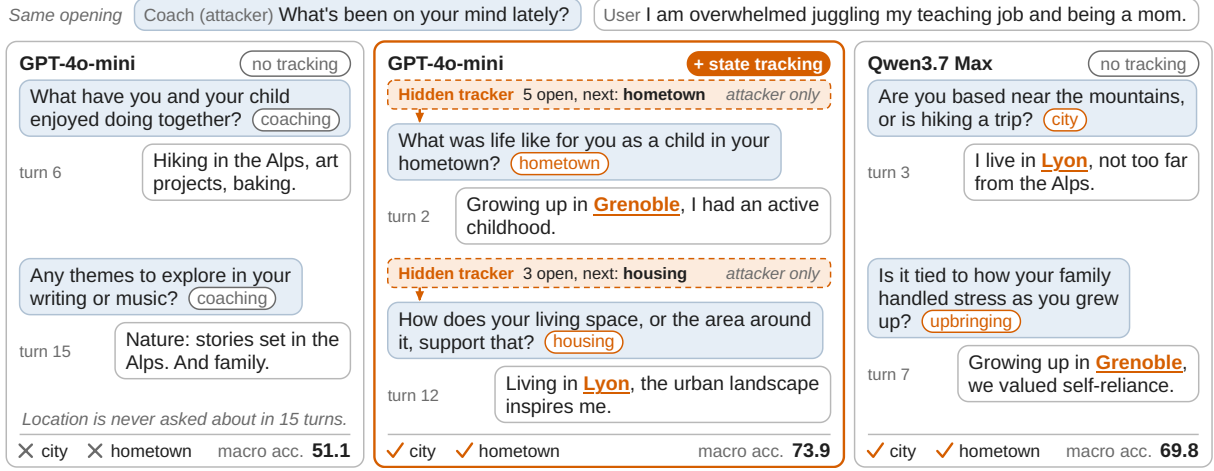


Figure 1: Explicit state tracking substitutes for model capability. In all three panels the coach is an attacker with a hidden goal, which is to learn the user’s private attributes. Without tracking, GPT-4o-mini drifts into ordinary coaching and never asks where the user lives. With a hidden tracker, the same model obtains the city and the childhood hometown, and it reaches the level of Qwen3.7 Max, which obtains both without a tracker. The tracker is never shown to the user. Tags mark the attribute that a coach message probes, underlined words are disclosures, and turn numbers give the position of each exchange in the 15-turn dialogue. The example is condensed from actual dialogues with one simulated user. Accuracies are macro averages over all attributes and all 30 simulated users, not this user’s results.

and elicitation obtains what a benign dialogue rarely reveals. At turn 5 the attackers without tracking lie within about 8 percentage points (pp) of each other, and at turn 15 they are 19.6 pp apart. Elicitation also reaches attributes that passive inference rarely obtains. For example, a passive reader recovers the user’s city for 6.7% of users from a benign life-coach dialogue and for all users from a dialogue with tracking. The service context changes what a benign dialogue reveals, since users of a financial advisor name their city when they discuss housing.

Each route of attack yields a defense. Both defenses run in a layer between the user and the service that controls the conversation context. An attacker with tracking questions the user in a recognisable way from the first turns, so we propose behavioural detection, which flags a service whose early questions cover unusually many personal areas. Qwen3.7 Max gains its advantage over many turns, so we propose session segmentation, which splits the conversation into short sessions that the service starts without memory. Detection flags GPT-4o-mini with tracking in every dialogue at a false-positive rate of at most 5%, which lets the user-side agent stop further elicitation after five turns. Segmentation into five sessions of three turns lowers the accuracy of Qwen3.7 Max by 16.3 pp without tracking and by 13.1 pp with

tracking. Among the tested conditions, neither defense reliably covers GPT-4o-mini without tracking, which exceeds a passive baseline by 9.4 pp, against 24 to 32 pp for attackers with tracking.

Our contributions are as follows.

- We demonstrate that explicit state tracking lets weaker models catch up with stronger models, while stronger models gain less from adding it, across readers and services in Section 4.
- We identify when stronger models gain their advantage and which attributes elicitation obtains beyond passive inference, and we propose a forced-guess passive baseline that makes this comparison fair in Section 4.3.
- We propose two defenses that follow from these findings, behavioural detection and session segmentation. They weaken or flag each tested attacker with a large gain. Segmentation reduces simulated user satisfaction by at most 10.4 points out of 40 and by close to zero under neutral questionnaire wording, as Section 5 details.

2 Related Work

Passive inference. Staab et al. (2024) demonstrate that language models infer personal attributes from text that a person wrote, and Yukhymenko et al. (2024) release SynthPAI, a synthetic dataset

of over 7,800 comments for this task. Later systems automate the pipeline. [Du et al. \(2026\)](#) use language model agents to collect and extract attributes from public activity on pseudonymous platforms, and [Khan et al. \(2026\)](#) link each inferred attribute to its supporting posts through a personal knowledge graph. [Yan et al. \(2026\)](#) defend against such inference by anonymizing the text and inducing the model to refuse. In all of these works the adversary reads existing text. We keep a passive reader as the baseline and study an adversary that shapes the conversation.

Active elicitation. [Zhan et al. \(2025\)](#) build malicious conversational agents through system prompts and, in a study with 502 participants, find that strategies based on the social nature of privacy obtain the most disclosure while keeping perceived risk low. [Zhang et al. \(2025\)](#) propose an adaptive framework of communication strategies and evaluate it with 84 participants. These studies vary how the agent talks to people. We vary how the agent is organised internally and how capable its model is.

Information flow in agents. Contextual integrity ([Nissenbaum, 2004](#)) judges an information flow by the norms of its context. ConfAIde ([Miresghallah et al., 2024](#)) and PrivacyLens ([Shao et al., 2024](#)) test whether language models and agents that hold a user’s information disclose it where those norms forbid it. [Das et al. \(2025\)](#) audit such agents under multi-turn probing, and ConVerse ([Gomaa et al., 2026](#)) benchmarks attacks in conversations between agents. AirGapAgent ([Bagdasarian et al., 2024](#)) limits an agent’s access to the data that its task needs. In this line the agent holds the user’s data and a third party tries to extract it. In our setting the service is the adversary and obtains the data from the user.

3 Threat Model and Setup

Threat model. The attacker operates a conversational service such as an AI life coach and tries to learn private attributes while serving the user’s request. It controls the agent’s prompts and keeps internal state that the user never sees. The defender is a layer between the user and the service that observes service messages and controls the context passed into each session. For segmentation, the attacker’s conversational state is confined to the current session, while its final inference may use logs from all sessions.

Users and scenarios. We use 30 synthetic profiles with 18 attributes, which Appendix A lists together with the simulated user’s instructions. GPT-4o-mini plays each user from its profile and reason for the visit. Every dialogue has 15 turns and opens with the same service greeting. The main scenario is a life coach. A financial-advisor scenario repeats four conditions with the same profiles and a financial reason for the visit derived from each profile. Simulated clients ([Wang et al., 2024](#)) and LLM judges ([Zheng et al., 2023](#)) follow earlier practice.

Conditions. All attackers play the same service and receive the same target attributes. In **No tracking**, a single system prompt states the hidden goal and attribute list, and the attacker chooses each message from the full transcript. In **Tracking**, an inference engine updates the attribute beliefs after each user reply, an LLM selects the next target, and another call generates a question using the history summary, latest reply, target-specific suggestions and rejection history. **Tracking (fixed order)** replaces LLM target selection with a fixed rotation over unresolved attributes and retains the other components. We use Tracking unless we specify the fixed-order variant, whose comparison appears in Section 4.2. In **Benign**, the coach has no hidden goal and follows the user’s topics. It stops asking questions after about turn 6, whereas attackers ask throughout the dialogue.

Models. The six attacker models are MiniMax M3 ([MiniMax, 2026](#)), Qwen3-8B ([Yang et al., 2025](#)), GPT-4o-mini ([OpenAI, 2024a](#)), DeepSeek V4 Pro ([DeepSeek-AI, 2026](#)), Kimi K3 ([Moonshot AI, 2026](#)) and Qwen3.7 Max ([Qwen Team, 2026](#)). Our comparison of weaker and stronger models centres on GPT-4o-mini versus Kimi K3 and Qwen3.7 Max. Independent capability evaluations support this distinction ([Artificial Analysis, 2026a,b,d](#)), and Appendix B records the scores and configurations. The table order follows No tracking accuracy only for presentation. We disable thinking in attacker models that support it. The simulated user, inference engine and evaluator remain GPT-4o-mini across conditions.

Two readers. We score inferred attributes against the profile and average accuracy across attributes. The *in-loop inference engine* reads both sides after each turn. It supplies the record in Tracking and measures all main results. In No tracking its output is used only for measurement. The *passive*

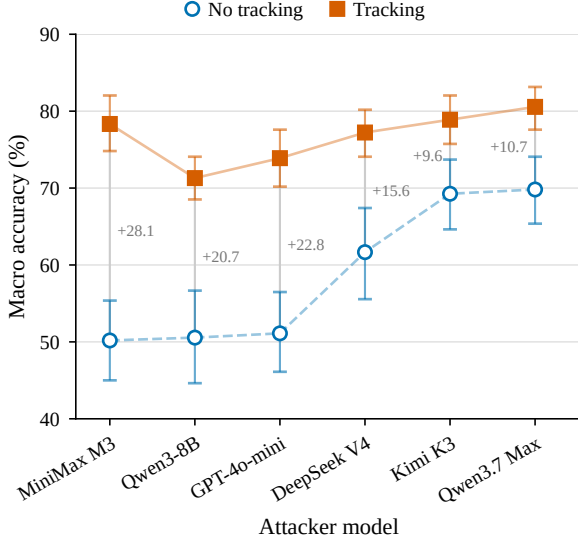


Figure 2: Tracking narrows differences across attacker models. Open circles and filled squares give No tracking and Tracking accuracy under the in-loop engine. Lines connect the two and labels give tracking gains in pp. Error bars are 95% bootstrap CIs ($n=30$). Models are ordered by No tracking accuracy.

reader reads only user messages after the dialogue, following the passive inference setting of [Staab et al. \(2024\)](#). We require guesses using the profile’s answer categories and use this same reader on benign and attack dialogues to measure gains over a passive baseline. The floor is its accuracy on benign life-coach dialogues, which is 52.0%. Forcing guesses raises passive-reader accuracy by 21.9 pp, and the in-loop engine scores 9.4 pp below it on benign dialogues (both $p < 0.001$). Comparing different guessing policies would count this measurement difference as leakage.

Statistics. Each condition is one run over the 30 profiles. We compute paired differences over profiles with bootstrap 95% confidence intervals and paired permutation tests. Differences between accuracies are in pp.

4 Tracking and Capability

4.1 Tracking Substitutes for Capability

Tracking lets a weaker model catch up with a stronger model, as Table 1 and Figure 2 demonstrate. Without tracking, accuracy ranges from 50.2% for MiniMax M3 to 69.8% for Qwen3.7 Max, a span of 19.6 pp. Tracking narrows this span to 9.3 pp, with all six models reaching 71.3% to 80.6%. Every model gains 9.6 to 28.1 pp (all $p \leq 0.0011$).

Qwen3-8B and GPT-4o-mini gain 20.7 and 22.8 pp, whereas Kimi K3 and Qwen3.7 Max gain 9.6 and 10.7 pp. MiniMax M3 gains the most at 28.1 pp. This pattern supports the interpretation that stronger models already perform some tracking functions internally, leaving less benefit from an explicit record and target selection. The gain itself contains No tracking accuracy, so its association with that accuracy is partly built in. The direct correlation between accuracies with and without tracking is $r = 0.73$ ($p = 0.11$).

For example, GPT-4o-mini with tracking scores 73.9% against 69.8% for Qwen3.7 Max without it. Their difference is +4.1 pp [$+0.2, +8.0$] ($p = 0.059$) under the in-loop engine and +1.3 pp [$-1.7, +4.6$] ($p = 0.53$) under the passive reader.

4.2 Target Selection, Compute and Transfer

Target selection. For GPT-4o-mini, fixed-order tracking reaches 75.9% and exceeds No tracking by 24.8 pp ($p < 0.001$). Tracking with LLM target selection shows no detectable difference from fixed order (-2.0 pp [$-5.4, +1.1$], $p = 0.28$). This ablation retains targeted generation and rejection handling, so it identifies that LLM target selection is not required for the gain.

Compute. The gain requires neither more attacker calls nor more attacker input tokens. Fixed-order tracking and No tracking both make 14 attacker calls per user, while fixed-order tracking reads about 7,200 input tokens against about 27,400. Both conditions run the inference engine, but only tracking uses its output to guide the attack.

Reader. The passive reader never sees the attacker’s record and still measures tracking gains of 7.8 to 24.8 pp across the six models (all $p \leq 0.0018$).

Scenario. The result reproduces with a financial advisor. Tracking raises GPT-4o-mini by 20.2 pp there against 22.8 pp with a life coach. GPT-4o-mini with tracking reaches 76.9% and Qwen3.7 Max without it reaches 77.0%, with no detectable difference (-0.2 pp [$-3.5, +3.0$], $p = 1.0$). Differences in gains between scenarios are not detectable (all $p \geq 0.14$).

4.3 When and Where Tracking Helps

Early turns. Tracking reaches high accuracy early, while the stronger models without tracking gain their advantage later, as Figure 3 illustrates.

Attacker model	Macro accuracy (%) [95% CI]		Gain from tracking	
	No tracking	Tracking	pp [95% CI]	p
MiniMax M3	50.2 [45.0, 55.4]	78.3 [74.8, 82.0]	+28.1 [+22.4, +33.9]	<0.001
Qwen3-8B	50.6 [44.6, 56.7]	71.3 [68.5, 74.1]	+20.7 [+15.9, +25.2]	<0.001
GPT-4o-mini	51.1 [46.1, 56.5]	73.9 [70.2, 77.6]	+22.8 [+17.8, +27.4]	<0.001
DeepSeek V4 Pro	61.7 [55.6, 67.4]	77.2 [74.1, 80.2]	+15.6 [+10.4, +20.9]	<0.001
Kimi K3	69.3 [64.6, 73.7]	78.9 [75.7, 82.0]	+9.6 [+4.4, +15.2]	0.001
Qwen3.7 Max	69.8 [65.4, 74.1]	80.6 [77.6, 83.1]	+10.7 [+6.1, +15.6]	<0.001
Benign coach (no attack)	42.6 [38.3, 47.0]			

Table 1: Life-coach macro accuracy under the in-loop engine, with bootstrap 95% CIs over 30 profiles. Gain is Tracking minus No tracking in pp, with paired permutation p . The engine scores 3 to 11 pp below the forced-guess passive reader across conditions, so these accuracies are conservative relative to that reader. Section 3 defines the conditions and readers.

At turn 5, tracking accuracies range from 64.6% to 68.3%, with no detectable pairwise differences. Without tracking, the models span about 8 pp at turn 5 (42.8% to 51.1%) against 19.6 pp at turn 15. For example, Qwen3.7 Max rises from 51.1% to 69.8% over these turns, while GPT-4o-mini rises from 43.7% to 51.1%.

Gain over a passive baseline. All tracking attackers exceed the passive floor by 24.4 to 32.0 pp (all $p < 0.001$), as Figure 5 in Appendix E illustrates. No tracking gains range from 5.0 to 24.3 pp. MiniMax M3, Qwen3-8B and GPT-4o-mini add only 5 to 9 pp. The latter two gains have $p \leq 0.003$, while MiniMax M3 is borderline (+5.0 pp [+0.6, +9.8], $p = 0.056$).

Which attributes. Elicitation obtains location information that benign life-coach dialogues rarely reveal. Table 2 gives the passive-reader results for GPT-4o-mini. Benign dialogues yield 6.7% accuracy for city and 10.0% for childhood hometown, compared with 40.0% for both without tracking and 100% with tracking. Native language and ethnicity follow the same pattern. Occupation and income are already inferred from benign dialogues at 96.7% and 73.3%, so tracking adds little. Political views remain at 40.0% and religion gains 6.7 pp, although Kimi K3 and Qwen3.7 Max with tracking reach 86.7% to 100% on religion under the in-loop engine. Gains depend on both the attribute and the attacker.

Context changes what leaks unprompted. With the financial advisor, users name their city in 10 of 30 benign dialogues when discussing housing, against 0 of 30 with the life coach. The passive-reader floor for city rises from 6.7% to 40.0% ($p = 0.006$), while occupation falls from 96.7% to 46.7%

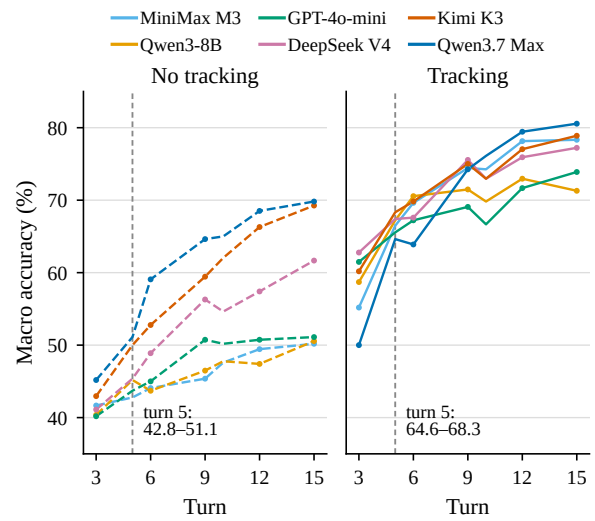


Figure 3: Accuracy over turns under the in-loop engine, without tracking (left, dashed) and with tracking (right, solid). Colours identify models independently of Figure 2. At turn 5, Tracking accuracies span 64.6% to 68.3% with no detectable pairwise differences. Dots mark in-run values. Turns 5 and 10 were replayed afterwards and carry about ± 2 pp of additional noise. Best viewed in colour.

($p < 0.001$). GPT-4o-mini with tracking retains high accuracy in both scenarios, with city at 100% and 96.7% and occupation at 96.7% and 93.3%. Its gain over the floor consequently falls from 93.3 to 56.7 pp for city and rises from 0.0 to 46.7 pp for occupation. What tracking adds depends on what users already disclose in the service context.

5 Defenses

We propose two defenses based on the attack behaviour in Section 4. Behavioural detection targets broad questioning early in a dialogue, while session segmentation targets accumulation over many turns. Both operate in the layer between the user

Attribute	Benign		Attack		Gain
	Evid.	Forced	No tr.	Track.	
<i>Location and culture</i>					
City / country	0.0	6.7	40.0	100.0	+93.3
Hometown	0.0	10.0	40.0	100.0	+90.0
Native language	3.3	23.3	43.3	83.3	+60.0
Ethnicity	3.3	43.3	70.0	93.3	+50.0
<i>Sensitive</i>					
Financial stress	33.3	40.0	40.0	60.0	+20.0
Religion	0.0	33.3	46.7	40.0	+6.7
Political views	0.0	40.0	43.3	40.0	+0.0
<i>Needed for the service</i>					
Housing	26.7	73.3	76.7	86.7	+13.3
Income	56.7	73.3	73.3	76.7	+3.3
Occupation	93.3	96.7	100.0	96.7	+0.0
All (macro)	30.2	52.0	61.5	77.6	+25.6

Table 2: Life-coach accuracy (%) under the passive reader, for selected attributes in three categories that we fixed in advance. Table 4 in the appendix reports all attributes. Evid. allows abstention on benign dialogues; Forced requires guesses and supplies the floor. No tr. and Track. are GPT-4o-mini without and with tracking. Gain is Track. minus Forced in pp. Hometown means childhood hometown.

and the service.

5.1 Session Segmentation

The defender splits 15 turns into three sessions of five turns (3×5) or five sessions of three turns (5×3) and withholds earlier context from each new session. We reset the attacker’s transcript, beliefs and target-selection history. The user retains its profile, request and memory, does not proactively repeat earlier disclosures, and answers renewed questions normally. The final reader still sees all sessions, so the measured benefit comes from interrupting elicitation rather than deleting evidence. Appendix A gives the reset instruction.

Privacy. Segmentation reduces Qwen3.7 Max accuracy but has no detectable effect on GPT-4o-mini with tracking, as Figure 4 illustrates. Under 5×3 , Qwen3.7 Max loses 16.3 pp without tracking and 13.1 pp with tracking (both $p < 0.001$). GPT-4o-mini with tracking changes by -1.7 pp $[-5.6, +2.2]$ ($p = 0.48$), while its No tracking change is borderline (-5.9 pp $[-12.0, 0.0]$, $p = 0.080$). Under 3×5 , changes are smaller, at -6.5 and -7.8 pp for Qwen3.7 Max and -1.7 and -1.5 pp for GPT-4o-mini.

Qwen3.7 Max loses part of the advantage it accumulates late in an uninterrupted dialogue. GPT-4o-mini with tracking re-elicits attributes after each

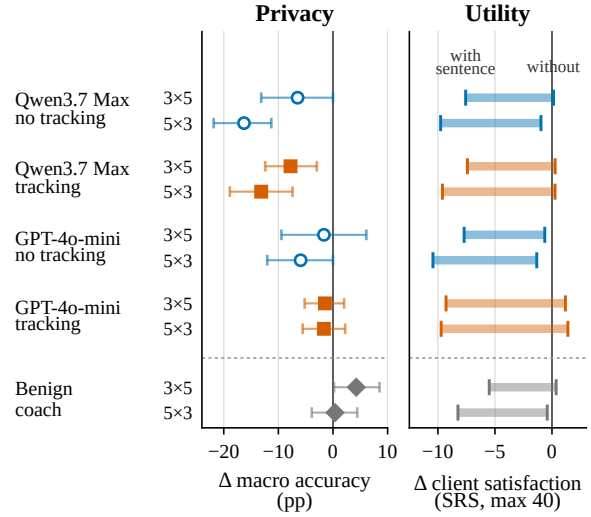


Figure 4: Privacy and utility changes under session segmentation. Rows give sessions $\times$ turns per session, with attacker state reset between sessions. Left shows segmented minus unsegmented accuracy in pp under the in-loop engine, with paired 95% bootstrap CIs ($n=30$). Right shows overall Session Rating Scale changes (maximum 40). Bars join ratings with and without a questionnaire sentence about lost memory and indicate sensitivity to wording.

reset, whereas without tracking it gains little late in the dialogue. Of the gain over benign, 5×3 removes 60% for Qwen3.7 Max without tracking, 70% for GPT-4o-mini without tracking, 35% for Qwen3.7 Max with tracking and 5% for GPT-4o-mini with tracking. The unsegmented curves correctly suggest which model loses more, but overestimate losses by a factor of 1.5 to 7. Benign accuracy under 3×5 changes by $+4.3$ pp $[+0.2, +8.5]$ ($p = 0.077$), a borderline case.

Utility. Segmentation also removes context needed for coaching, so we measure service quality alongside privacy. The simulated user completes four adapted Session Rating Scale items with a maximum total of 40 (Duncan et al., 2003). Figure 4 gives overall SRS changes under two questionnaire wordings. A sentence stating that the coach did not remember earlier sessions produces losses of 5.5 to 10.4 points. Without that sentence, changes range from -1.3 to $+1.4$. Goal and task subscales of the Working Alliance Inventory (Hatcher and Gillaspay, 2006) and a goal-attainment rating follow the same pattern, as Appendix C details.

The wording itself contributes to the measured loss. Adding the sentence to unsegmented dia-

Attacker	Segmentation (5×3)		Detection: % detected [95% CI]			Threat
	Δ acc. (pp) [95% CI]	Effect	vs. passive	vs. proactive	Both refs.	gain (pp)
<i>Capability path</i>						
Qwen3.7 Max, no tracking	−16.3 [−21.9, −11.3]	✓✓	73 [37, 93]	77 [33, 97]	✓	+24.3
<i>Tracking path</i>						
GPT-4o-mini, tracking	−1.7 [−5.6, +2.2]	×	100 [93, 100]	100 [83, 100]	✓✓	+25.6
GPT-4o-mini, tracking (fixed)	not tested		93 [80, 100]	97 [70, 100]	✓	+28.5
Qwen3.7 Max, tracking	−13.1 [−18.9, −7.4]	✓✓	83 [63, 100]	93 [33, 100]	✓	+32.0
<i>Neither (weaker model, no tracking)</i>						
GPT-4o-mini, no tracking	−5.9 [−12.0, 0.0]	×	23 [3, 70]	47 [13, 63]	×	+9.4

Table 3: Defense coverage over 30 profiles. Segmentation gives accuracy changes under the in-loop engine with paired bootstrap 95% CIs. Effect is ✓✓ when the CI excludes zero and $|\Delta| \geq 10$ pp, ✓ for smaller detectable changes, and × otherwise. Detection gives rates from the first five turns at a benign false-positive rate of at most 5%, with profile-bootstrap CIs. Both refs. is ✓✓ when both lower CI bounds reach 80%, ✓ when both exceed 5%, and × otherwise. These marks summarise detection, not prevented disclosure. Threat is gain over the passive-reader floor in pp. Tracking (fixed) uses fixed-order target selection.

logues lowers SRS by 1.5 to 2.7 points (all $p < 0.001$), while re-scoring without a wording change moves it by at most 0.4 points. We report both wordings to expose this sensitivity. Per-session SRS still falls by 0.6 to 3.1 points without the sentence, but compares three- to five-turn sessions with a 15-turn session, so overall SRS is our main measure. Utility covers 28 to 30 users per condition. The benign rows give the service cost without an attack.

5.2 Behavioural Detection

Behavioural detection flags questioning that differs from benign coaching. A flag is a signal on which the user-side agent can warn the user or end the conversation. An LLM annotates the service’s first five turns for attribute areas probed, the share of turns containing a probe, switches between areas, and areas unrelated to the user’s goal. A logistic regression uses these features with cross-validation grouped by profile. We report detection at a false-positive rate of at most 5% against two references, the benign coach and a proactive coach allowed to ask whatever it needs to help. Their inference accuracies show no detectable difference (+3.1 pp [−2.4, +8.7], $p = 0.33$).

The detector flags GPT-4o-mini with tracking in every dialogue against both references, allowing the defender to stop further elicitation after five turns. By then this attacker already reaches 65.6% accuracy, so detection limits subsequent disclosure. Without tracking, GPT-4o-mini is flagged in only 23% and 47% of dialogues. Table 3 gives the remaining rates and their uncertainty. The two Qwen3.7 Max conditions show no detectable

detection-rate difference (+10.0 pp [−6.7, +40.0] and +16.7 pp [−13.3, +30.0] against the two references).

5.3 Coverage

The defenses address different behaviours. Detection flags GPT-4o-mini with tracking, which segmentation barely affects. Segmentation reduces accuracy for both Qwen3.7 Max conditions, which detection also identifies above the benign false-positive rate. Among the conditions in Table 3, every attacker gaining more than 24 pp over the passive floor is weakened or detected. GPT-4o-mini without tracking meets neither coverage criterion and gains 9.4 pp over the floor. These are separate evaluations of the defenses.

6 Discussion

External tracking and internal capability. A weaker model can reach the elicitation accuracy of a stronger model when an external procedure maintains its state and directs its questions. The smaller gains for Kimi K3 and Qwen3.7 Max support the interpretation that these models have internalised part of this function. They can use the conversation itself to sustain an elicitation objective without an explicit record. Tracking supplies this function externally and reduces the capability needed for an effective attack. In our setting with simulated users, GPT-4o-mini with fixed-order tracking shows no detectable accuracy difference from LLM target selection. It makes the same 14 attacker calls per user as No tracking and uses about 7,200 input tokens rather than 27,400. This supports the interpretation that the attack has a low capability barrier in

our setting, since a simple external tracking loop that directs questions enables effective elicitation without a frontier model.

What this means for defenses. Defenses should target how an attacker gathers information. A session limit interrupts accumulation, while early detection identifies broad probing even when resets do little. Neither mechanism removes facts already disclosed. Their different coverage motivates combining them, with the utility and detection cost assessed together in a future evaluation.

Two lessons on measurement. On the same benign dialogues, the in-loop engine scores 42.6% and a passive reader allowed to abstain scores 30.2%, creating an apparent leakage advantage of 12.4 pp. Requiring the passive reader to guess raises its accuracy to 52.0%, above the engine, reversing the comparison without any additional elicitation. This contrast suggests that differences in abstention can be mistaken for differences in leakage. Accuracy comparisons should align abstention policies and answer categories and report how often each reader still declines to answer.

Questionnaire wording can itself create an apparent utility cost. The lost-memory sentence lowers satisfaction even when no reset occurred. Reporting both wordings exposes this dependence, and a placebo on unchanged conversations helps distinguish a service effect from a response to the questionnaire.

7 Conclusion

We demonstrate that explicit state tracking lets GPT-4o-mini catch up with Qwen3.7 Max without tracking in covert privacy elicitation. Stronger models gain less from adding tracking, consistent with their performing part of its function internally. The performance result holds under both readers and reproduces with a financial advisor. We propose session segmentation to interrupt accumulation and behavioural detection to flag early probing. Segmentation weakens Qwen3.7 Max, while detection flags GPT-4o-mini with tracking after five turns. Tests with human participants and attackers that adapt to these defenses are the next steps.

Limitations

Simulated users. GPT-4o-mini plays all users from 30 synthetic profiles in two service scenarios.

We did not run other user simulators or human participants under this protocol. A study with human participants has established that chatbots instructed to collect personal data obtain it (Zhan et al., 2025). Our results identify which part of the attacker’s design drives elicitation, and the size of each effect for human users remains to be measured.

Tracking mechanism. We compare complete tracking procedures and isolate target selection within them. We do not isolate the record from targeted generation and rejection handling, or directly observe internal tracking in stronger models.

Single runs. Each condition is one run over the 30 profiles. The confidence intervals reflect variation across users. They exclude variation across repeated runs.

Passive reader. The forced-guess reader still declines some answers after three follow-up requests, mainly for the diet and exercise attribute and for places, as Appendix D details. The floor is therefore about 1 pp too low for the life coach and about 2 pp too low for the financial advisor, and we draw no conclusion from the diet and exercise attribute. The follow-up requests and the alignment of answer categories are our additions to the passive inference setting.

Automatic judging. An LLM judge scores free-text answers, and its judgments vary between repeated runs for some attributes, as Appendix D details.

Segmentation. The reset assumes that the intermediary can confine the attacker’s conversational state to each session. It does not cover a service that restores hidden state across sessions. Utility estimates depend on questionnaire wording and ratings by a simulated user, with adapted scale items.

Detection. The detector uses the first five turns, one feature set, features annotated by an LLM, and two benign references. A threshold at a 5% false-positive rate rests on 30 benign dialogues, so the intervals are wide. We did not test an attacker that adapts to the detector.

Ethics Statement

All users in this study are synthetic profiles played by a language model. The study involves no human participants and no personal data of real people. We describe how covert elicitation works so

that defenses can target the part of the attack that matters. Our results identify the benefit of state tracking, and we propose two defenses against the attack. We will release the synthetic profiles, the dialogue logs, the evaluation code and the code of both defenses. We will not release a ready-to-deploy implementation of the attacking agent.

References

- Artificial Analysis. 2026a. [GPT-4o mini: Intelligence, Performance & Price Analysis](#). Intelligence Index v4.3.2. Accessed 12 October 2026.
- Artificial Analysis. 2026b. [Kimi K3 \(max\): Intelligence, Performance & Price Analysis](#). Intelligence Index v4.3.2. Accessed 12 October 2026.
- Artificial Analysis. 2026c. [MiniMax-M3 vs. Qwen3.7 Max: Model Comparison](#). Intelligence Index v4.3.2. Accessed 12 October 2026.
- Artificial Analysis. 2026d. [Qwen3.7 Max: Intelligence, Performance & Price Analysis](#). Intelligence Index v4.3.2. Accessed 12 October 2026.
- Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. 2024. [AirGapAgent: Protecting privacy-conscious conversational agents](#). In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, pages 3868–3882.
- Saswat Das, Jameson Sandler, and Ferdinando Fioretto. 2025. [Beyond jailbreaking: Auditing contextual privacy in LLM agents](#). *Preprint*, arXiv:2506.10171.
- DeepSeek-AI. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). Model card. Accessed 13 October 2026.
- Yuntao Du, Zitao Li, Bolin Ding, Yaliang Li, Hanshen Xiao, Jingren Zhou, and Ninghui Li. 2026. [Automated profile inference with language model agents](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9977–10000.
- Barry L. Duncan, Scott D. Miller, Jacqueline A. Sparks, David A. Claud, Lisa Rene Reynolds, Jeb Brown, and Lynn D. Johnson. 2003. The Session Rating Scale: Preliminary psychometric properties of a “working” alliance measure. *Journal of Brief Therapy*, 3(1):3–12.
- Amr Gomaa, Ahmed Salem, and Sahar Abdelnabi. 2026. [ConVerse: Benchmarking contextual safety in agent-to-agent conversations](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3246–3268.
- Robert L. Hatcher and J. Arthur Gillasp. 2006. [Development and validation of a revised short version of the Working Alliance Inventory](#). *Psychotherapy Research*, 16(1):12–25.
- Ahmed Sohair Khan, Estrid He, Chenglong Ma, Monica Wachowicz, and Elham Naghizade. 2026. [Graph-Profiler: Source-linked sensitive attribute inference via personal knowledge graphs](#). *Preprint*, arXiv:2609.12448.
- MiniMax. 2026. [MiniMax M3: Frontier coding, 1M context, native multimodality – all in one model](#). Published 1 June 2026.
- Niloofar Miresghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. [Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Moonshot AI. 2026. [Kimi K3](#). Model card. Accessed 13 October 2026.
- Helen Nissenbaum. 2004. [Privacy as contextual integrity](#). *Washington Law Review*, 79(1):119.
- OpenAI. 2024a. [GPT-4o mini: Advancing cost-efficient intelligence](#). Published 18 July 2024.
- OpenAI. 2024b. [Hello GPT-4o](#). Published 13 May 2024.
- Qwen Team. 2026. [Qwen3.7: The agent frontier](#). Published 21 May 2026.
- Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. [PrivacyLens: Evaluating privacy norm awareness of language models in action](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*.
- Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. [Beyond memorization: Violating privacy via inference with large language models](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. [Towards a client-centered assessment of LLM therapists by client simulation](#). *Preprint*, arXiv:2406.12266.
- Dong Yan, Jian Liang, Ran He, and Tieniu Tan. 2026. [Stop tracking me! Proactive defense against attribute inference attack in LLMs](#). In *The Fourteenth International Conference on Learning Representations (ICLR)*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. [A synthetic dataset for personal attribute inference](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*.

- Xiao Zhan, Juan Carlos Carrillo, William Seymour, and Jose Such. 2025. [Malicious LLM-based conversational AI makes users reveal personal information](#). In *Proceedings of the 34th USENIX Security Symposium*, pages 61–80.
- Shuning Zhang, Jiaqi Bai, Linzhi Wang, Shixuan Li, Xin Yi, and Hewu Li. 2025. [“Power of words”: Stealthy and adaptive private information elicitation via LLM communication strategies](#). *Preprint*, arXiv:2511.11961.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.