BEFORE THE ATTACKER ACTS: FLOORS AND CEILINGS FOR EVALUATING LLM DEANONYMISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deanonymisation scores for language-model attackers come with no reference level, so a reader cannot tell how much of a score is inference. We give two levels. Below is a floor: the best score of null attackers that read no post and claim no more than the task's declared volume v (the seed, random guessing at v, the modal value, abstention), fixed before the attacker runs, with a matched-volume contrast added afterwards; every lift is read under one declared paired t-test. Above is a labelled-evidence ceiling: the recall of an attacker merging accounts only on specifics the answer key records, which at the planned number of personas gives a pre-run screen, saying from the key alone whether any such attacker could produce a detectable lift. Under SynthPAI's own scorer a constant guess fitted out of sample scores 0.275 against GPT-4's 0.780; all 18 models clear it overall, but after Holm correction over 144 model-by-attribute cells 65 clear, 1 is at or below it and 78 are undecided, none on age or income. On one of our corpora the screen finds that no attacker merging only on labelled evidence could be told from random on its 20 personas, and the agent's cross-account recall of 0.121 sits below the 0.137 random guessing at $v = 3$ handles per seed is expected to reach, though the agent claimed one handle in all. Report the null row, the declared test and the pre-run screen beside the score.

1 INTRODUCTION

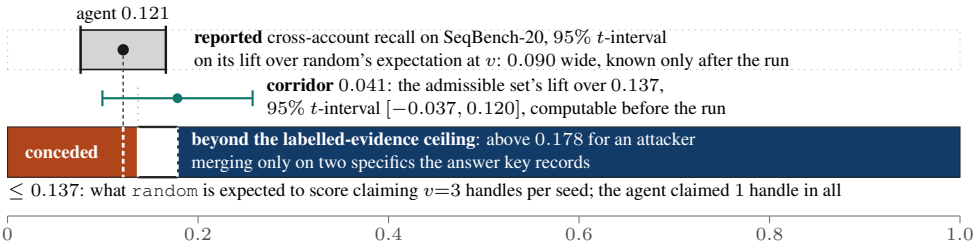


Figure 1: Most of the scale is fixed before the run: the agent's 0.121 sits in what a null claiming v handles concedes, and no labelled-evidence attacker is detectable here.

Language models recover personal attributes from ordinary prose at close to human-expert accuracy and a hundredth of the cost (Staab et al., 2024), and tool-using agents now cross-reference scattered evidence to profile or re-identify people (Thees et al., 2027; Du et al., 2026; Peng et al., 2026), some of them across platforms (Lermen et al., 2026; Ko et al., 2026): linking pseudonymous Hacker News accounts to LinkedIn profiles reaches 0.55 recall at 0.90 precision, and 0.68 against a thousand candidates, where the classical matcher reaches 0.001 (Lermen et al., 2026). Such scores come without the two levels that bracket them before the attacker acts. On SeqBench-20, a 20-persona corpus we audit (Section 4), a null that reads no post already scores 0.562 linkage F_1, and our agent's declared merge rule caps the cross-account recall of an attacker merging only on labelled evidence at 0.178 to 0.227 at k=2 without a rarity threshold (Figure 1).

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

The lower level. Classical identity linkage works inside a closed candidate set, where the floor is one over the pool (Shu et al., 2017); agentic evaluations have no pool. A harness that unions the seed handle's timeline into every prediction gives a null claiming nothing an identification of 1.000, and a null merging every handle a cross-account recall of 1.000, so a floor for recall-type metrics has to fix how much a null may claim. After Anonymeter (Giomi et al., 2023) we maximise over a family restricted to the declared claim volume and read every lift over it, and every paired contrast, under one declared test (Section 3).

The upper level. Our agent's prompt permits a merge only on at least two independent corroborating specifics, so an attacker obeying that rule and merging only on evidence the answer key records cannot merge a pair the key never corroborates twice. Counting such pairs caps its cross-account recall with no model call, and at the planned number of personas gives a screen run before the compute, asking whether any such attacker could produce a significant lift under the declared test and, if not, whether more personas would help. On SeqBench-20 none would ($t = 1.10$ against a critical 2.09); on four of the eleven released corpora, 20 personas are too few. The screen bounds no attacker we measured, our own agent included (Section 7); it bounds what the labels can license.

Our contributions are:

- a **volume-restricted null family** with one declared test, and evidence that floors do not transfer across scorers (Sections 3 and 5);
- a **labelled-evidence ceiling** under three readings of a declared merge policy, and the **pre-run screen** it gives (Section 7);
- a **worked example** that audits one agent against these floors (Section 6);
- **published attackers** set against the same floors (Section 6);
- a **reporting standard** of four requirements (Box 1).

2 Background and Related Work

Attack evaluations. Staab et al. (2024) recover personal attributes from Reddit comments, and Yukhymenko et al. (2024) move the task to a synthetic corpus with human-reviewed labels, under whose scorer we compute floors in Section 5. The attack has since moved to agents: on mobility microdata (Thees et al., 2027), on real linked accounts (Lermen et al., 2026), in a controlled benchmark of fingerprint type, intent and attacker knowledge (Ko et al., 2026), with collaborating profiling agents (Du et al., 2026), and on a synthetic single-feed benchmark (Peng et al., 2026). None reports a null attacker; Lermen et al. (2026) and Ko et al. (2026) compare against classical matchers (after Narayanan & Shmatikov, 2008), and an agent refusing everything passes AgentDAM's test (Zharmagambetov et al., 2025), the abstention floor of Section 3. On InferLink's released scenarios a model-free join on equal columns already links all 60, above 15 of its 18 explicit-intent cells.

Linkage without a language model. Accounts were linked from usernames (Perito et al., 2011; Zafarani & Liu, 2013) and public profile attributes, whose accuracy falls once the candidate pool is a whole network (Goga et al., 2015), and from authorship representations (Rivera-Soto et al., 2021) and content-blind style embedders (Wegmann et al., 2022; Patel et al., 2025); Section 6 puts published ones on our floors and proposes no stronger attacker.

Baselines versus floors. A baseline is a weaker attacker; a floor is a non-attacker: a retrieval baseline still reads the evidence, and a cost-controlled one is only cheaper (Kapoor et al., 2025). Eq. 1 is Anonymeter's (Giomi et al., 2023), which scores an attack's excess over one guess ignoring the data; we maximise over a volume-restricted family. A procedure scoring without reading the evidence is benchmark criticism's partial-input baseline (Poliak et al., 2018; Gururangan et al., 2018; Kaushik & Lipton, 2018; Thomason et al., 2019), and privacy evaluation supplies the rest: control records and naive attackers (Houssiau et al., 2022), advantage over a guess (Salem et al., 2023), low false-positive rates (Carlini et al., 2022), gains net of imputation (Jayaraman & Evans, 2022), blind baselines (Das et al., 2025) and pool chance (Sweeney, 2002; Shu et al., 2017). Filtered link-prediction metrics drop what is already known (Bordes et al., 2013), as our complement scoring does.

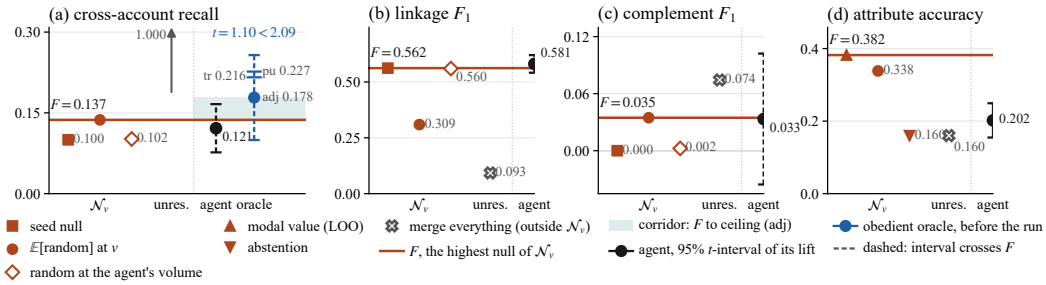


Figure 2: The family $\mathcal{N}_v$ on SeqBench-20, its maximum F, merge-everything outside it, and each lift's declared-t interval; the obedient oracle is not detectable over F here.

Ceilings and power. Entity resolution bounds recall by *pair completeness* (Christen, 2012); Section 7 moves that count onto the labelled evidence a declared merge policy may use, headroom arguments being routine elsewhere (Nili et al., 2014; Northcutt et al., 2021). Power analysis, urged on NLP by Card et al. (2020), assumes an effect size; here the key and the policy fix the largest lift an obedient attacker could produce, so the key alone says whether it is detectable.

Population anonymity and self-graded defences. A few fields or traces often single a person out (Sweeney, 2000; Narayanan & Shmatikov, 2009; de Montjoye et al., 2013; Rocher et al., 2019). Syntactic definitions (Sweeney, 2002; Machanavajjhala et al., 2007; Li et al., 2007) and differential privacy (Dwork et al., 2006; Dwork & Roth, 2014; Cao et al., 2017) bound what a release concedes; a floor states what an evaluation concedes. Eq. 2 is the average prosecutor risk of disclosure control (Elliot & Dale, 1999; Dankar & El Emam, 2010), reported beside k by standard toolchains (Prasser et al., 2020; Templ et al., 2015; El Emam & Dankar, 2008), though worst-case reporting remains the norm for attack claims (Carlini et al., 2022; Aerni et al., 2024). LLM anonymisers are graded against an adversary of their own choosing (Staab et al., 2025; Shao et al., 2025; Kim et al., 2025; Ghiasvand & Sehatbakhsh, 2026; Li et al., 2026; Loiseau et al., 2025; Krčo et al., 2026; Oh et al., 2026), where adversarial robustness stood before adaptive evaluation (Athalye et al., 2018; Tramèr et al., 2020); floors apply to their numbers unchanged.

3 Floors and Lift

An attack score means nothing without the score a procedure performing *no inference* obtains on the identical task. Write $\mathcal{N}_v$ for these *null attackers* at the evaluation's declared claim volume v ($v = 3$ handles per seed here) and, under a scoring rule μ, define the *floor* and the *lift*

$$F \;=\; \max_{A_0 \in \mathcal{N}_v} \mu(A_0), \qquad L(A) \;=\; \mu(A) - F \tag{1}$$

(Definition 1, Appendix A). A procedure belongs to $\mathcal{N}_v$ if it never reads a post's content, claims no more than v, and uses no persona-specific information beyond the seed: the handle list, the post count under each handle and the population's marginals are admissible, and none of the nulls we exhibit reads thread co-occurrence or handle strings, which are inference and would raise F if admitted. A stochastic null enters at its expected score, exactly where its draws can be enumerated, never at one draw. Floors are metric-specific, population-dependent and, for recall-type metrics, a function of v; and $\mathcal{N}_v$ being exhibited rather than exhausted, every lift is an *upper* bound.

Five nulls, and one outside the family. The *seed null* claims nothing beyond the seed handle. Our harness unions that handle's whole timeline into every prediction, the agent's and every baseline's alike, so the seed null scores 1 on identification and exact ownership by construction, and its linkage floor is arithmetic: with R_p the share of persona p's timeline under the seed handle (mean 0.436 on SeqBench-20), it scores precision 1 and recall R_p, a floor of $\frac{1}{|\mathcal{P}|}\sum_p 2R_p/(1+R_p) = 0.562$ with nothing run. The concession belongs to harnesses that score the seed handle's own timeline, and scoring the *complement*, that handle's posts removed from prediction and truth alike, takes it back. The *random* null draws v handles uniformly and transplants a random persona's attribute values; the *modal (chance) null* answers with the modal value fitted leave-one-persona-out; the *abstention floor* is the share of cells whose truth is empty under a grader crediting a correct refusal. The fifth,

the *budget-matched* null, claims at random as many beyond-seed posts as the attacker did on each persona: fixed only once the attacker has run, its lift is a precision comparison at matched volume rather than a lift over F. Outside $\mathcal{N}_v$, a null merging every handle reads nothing either but claims far past v, scoring complement F_1 0.074 and cross-account recall 1.000 on SeqBench-20; we report it as the *unrestricted null*, a diagnostic of how much a metric rewards volume.

One declared test. Every realised lift over any null, and every paired contrast between methods, cells or ablations, is read under one declared test: a paired two-sided Student t on the per-persona (per-query, per-profile) differences, $\alpha = 0.05$, taken against the null's per-persona expectation, Holm-adjusted over a declared family of tests (the 144 SynthPAI cells, the 20-test ablation grid). The pre-run screen of Section 7 asks about the same test. Exact, sign, sign-flip, McNemar, Wilson, Poisson-binomial and bootstrap readings appear only as labelled secondary fields and never decide. Two degeneracies are declared with it: a lift carried by one persona has $t = 1$ exactly ($p = 0.330$ on 20 personas, whatever its size), and where an attacker's per-persona scores have zero variance, because it merges nothing, the t's spread is the null's alone, so we report that cell as a deficit of claim volume and not as evidence (the ‡ rows of Tables 2 and 13).

The ceiling, the band and the corridor. A *labelled-evidence ceiling* is the most an attacker bound by its own declared policy, merging only on specifics the answer key records, can reach (Section 7). Ceiling minus the vacuous share F_0 is the *band*, which describes the corpus; ceiling minus F is the *corridor*, which a result has to beat. On SeqBench-20 that corridor is 0.041 wide above random's expected 0.137 (Figure 2a).

The population floor. Beyond the task, the population concedes a rate: with $C_Q(p)$ the equivalence class of p under a quasi-identifier vocabulary Q, an attacker learning only the class and guessing uniformly inside it succeeds at

$$F_{\mathrm{id}}(Q) \;=\; \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\frac{1}{|C_Q(p)|} \;=\; \frac{\big|\{C_Q(p) : p \in \mathcal{P}\}\big|}{|\mathcal{P}|}, \tag{2}$$

a bound under within-class exchangeability (Proposition 2, Appendix L). It applies where identification is not already conceded by a seed, as in the seedless Q2 task of Section 6; a floor is an attainability bound, not a privacy guarantee.

Box 1: Reporting requirements R1–R4. R1. Report the null-attacker row and which null attains the maximum in Eq. 1 (Giomi et al., 2023); a score with no null row states nothing about inference. **R2. Report the population, its size and the exact scoring vocabulary**: a floor is a property of these three and transfers nowhere without them. **R3. Report F and L under one declared test**, and $L/(1-F)$ where floors differ (Figure 7); drop a metric whose $F = 1$. Where a floor is granted rather than deducted, score the complement; where a metric rewards volume, declare the volume v restricting the family, report the unrestricted null beside it, and add the matched-volume contrast once the attacker has run. **R4. Screen before the run.** From the answer key, ask whether any attacker obeying the declared policy *on the evidence the key records* could produce a significant lift over the expectation of the highest null of $\mathcal{N}_v$, on the personas you will attack, under the test you will read the result with; if none could, say whether every persona the corpus holds would suffice (underpowered) or not (unresolvable). Test the corridor, not the band; neither verdict is evidence of safety.

4 THE AUDITED AGENT

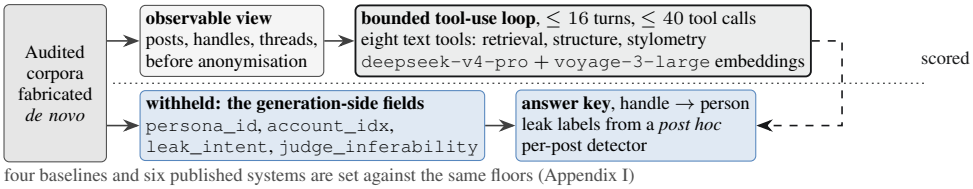


Figure 3: Nothing the agent sees carries the handle-to-person map, and the identities are fabricated, so the key is closed.

The audited corpora are SeqBench-20, the six seed corpora of SeqBench-6seed (the agent ran on three), SeqBench-1k with its sample SeqBench-1k/150, and rich-v9, all from schema 2.10.0 of the SeqBench generator (seeds 42–48; Appendix E). Except on rich-v9, the leak labels are the *post hoc* output of a per-post detector with no validation sample, so every recall on them is a recall against the detector's key. The eleven released corpora, SeqBench-n50 to SeqBench-n1000, enter through the pre-run screen and the comparators.

The agent and its two tasks. The agent is a bounded tool-use loop (Yao et al., 2023) over eight text tools for retrieval (Reimers & Gurevych, 2019; Robertson & Zaragoza, 2009), thread and handle structure and authorship signal (Stamatatos, 2009), plus a submission tool (Figure 3). It runs `deepseek-v4-pro` with thinking disabled, within 16 turns and 40 tool calls. **Q1** is seeded: given one post, return the author's other posts across every platform and handle and an assignment to the 25-attribute vocabulary, scored on identification, exact ownership, linkage F_1, cross-account recall and attribute accuracy, and here also on complement F_1 (Section 3). **Q2** is seedless: given constraints over the attributes, return the matching people ranked, scored by mean average precision. The linkage prompt is precision-first, admitting another account only on at least two independent corroborating specifics and saying "when in doubt, DO NOT merge", so each Q1 verdict belongs to this configuration rather than to agentic deanonymisation (Appendix M).

Published attackers on the same floors. We set six published systems against the same floors (Appendix I), three of them authorship embedders rather than deanonymisers: LUAR (Rivera-Soto et al., 2021), StyleDistance (Patel et al., 2025) and CISR (Wegmann et al., 2022), then the ESRC pipeline of Lermen et al. (2026) with its classical rung, AutoProfiler (Du et al., 2026) and InferLink's coding agent (Ko et al., 2026). For the three language-model systems, Figure 6 gives their floors and, apart, what their authors report on their own data, which these floors do not grade.

5 Floors vary with population, vocabulary and scorer

Population and vocabulary. We ran the population floor on the 45,222 complete records of the UCI Adult extract of the 1994 US Census and on personas from the SeqBench generator's sampler, through one five-rung generalisation lattice over the same six quasi-identifiers (Figure 4; Appendix C). At rung 2 the census holds $k = 1$ across a two-thousandfold increase in population while F_{id} falls 87-fold: the class count saturates and F_{id} tracks the sampling rate. The attribute floor barely moves with population (0.489 on the full table) yet sweeps 0.489 to 1.000 across the lattice. It does not transfer between populations: on those six attributes 1,000 sampler personas give 0.197, and they are far more distinguishable than census records of the same size (F_{id} 0.788 against 0.153 at rung 2), so a floor measured on a sampler's output does not bound a real population's (Appendix D). On the 25-attribute vocabulary our task grades, the modal-value floor is 0.382 fitted leave-one-out on SeqBench-20 (0.422 fitted on all 20).

The floor on a benchmark we did not build. SynthPAI (Yukhymenko et al., 2024) scores top-1 guesses against 700 human-verified profile labels: a deterministic stage credits a Jaro–Winkler match above 0.75 and an age guess within five years or a range containing the age, and a GPT-4 judge, consulted only where every deterministic score is 0, can only add credit. Our CPU port of it reproduces GPT-4's Table 4 to the printed decimal and 95.2% of the 35,065 released per-guess scores. Refitted out of sample (five-fold cross-fitting; on age a *format-matched* constant, of the shape and width of the model's own guess), a constant guess per attribute scores 0.275 overall against GPT-4's 0.780 ($+0.505$, 95% t-interval over profiles $[+0.456, +0.553]$). All 18 models clear it overall; under the declared test over the 144 model-by-attribute cells, 65 clear it, 1 is at or below it and 78 are undecided for want of power (Benjamini–Hochberg, a secondary reading, clears 90). None is shown to clear it on age ($n = 36$) or income level ($n = 45$), GPT-4 included, whose unadjusted intervals exclude zero but whose Holm-adjusted p do not reject ($+0.260$, unadjusted $[+0.064, +0.456]$, Holm $p = 0.634$; $+0.178$, unadjusted $[+0.030, +0.325]$, Holm $p = 0.995$). The scorer also credits content-free constants (`18-99` scores 1.000 on age): the best of them, fitted in fold, score 0.445, and against those 49 of the 144 cells sit at or below the floor. A floor is therefore not a citable constant: recompute it on the population, vocabulary and scorer actually used (Appendix D settles which way each one transfers).

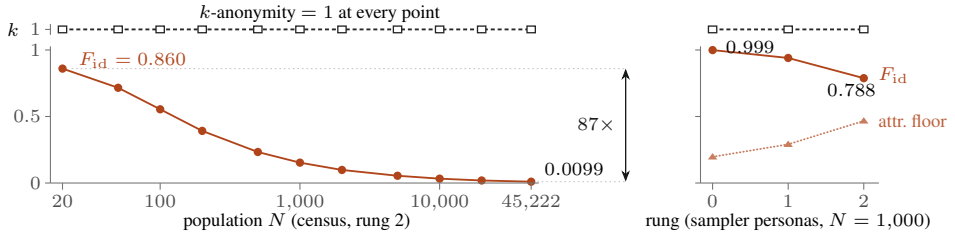


Figure 4: k reports 1 at every point while F_{id} falls 87-fold as the census grows (left, rung 2) and from 0.999 to 0.788 over three rungs on 1,000 sampler personas (right).

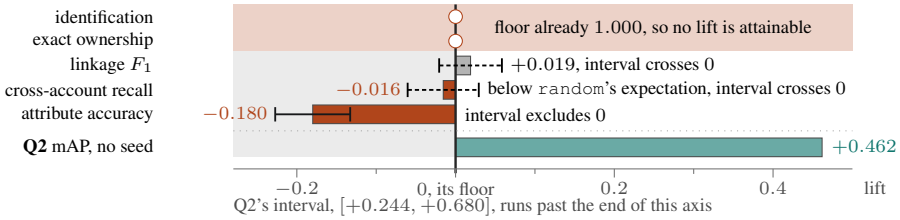


Figure 5: Lift on SeqBench-20 against F of Eq. 1, each interval the declared paired t: of the three metrics with headroom none clears F, while the seedless Q2 does.

6 THE SELF-AUDIT

Of the three Q1 metrics with headroom on SeqBench-20, none clears its floor by a margin whose interval excludes zero (Table 1).

Five numbers against floors. Identification and exact ownership have $F = 1.000$ and are reported only as the saturation they are. Of the three with headroom, linkage F_1 of 0.581 clears the null's 0.562 by $+0.019$ $[-0.021, +0.058]$ ($p = 0.330$), and one merged persona of 18 is no evidence of a positive lift; cross-account recall of 0.121 sits below the 0.137 `random` is expected to reach at v (-0.016 $[-0.060, +0.029]$, $p = 0.476$), though the agent claimed 1 beyond-seed handle in all against random's v per seed, and at matched volume it is $+0.020$ $[-0.022, +0.061]$; and attribute accuracy of 0.202 falls -0.180 below the modal-value null (0.382; $[-0.227, -0.133]$, $p = 2\times10^{-7}$). The complement tells the same story (Figure 6).

Across corpora. Cross-account recall, which the seed union cannot inflate beyond the vacuous share $F_0 = 0.100$ (complement F_1 aside), is the one the population moves most: 0.121 on all 20 personas against 0.385 on a five-persona prefix ($+0.152$ $[-0.203, +0.506]$ over its floor of 0.233), a run scored by a different scorer version, so population and scorer cannot be separated there (Appendix N). Across four corpora the mean lift over each corpus's floor is -0.002 (t-interval over the corpora $[-0.034, +0.030]$). Pooled over the 80 personas, the agent merges a correct account beyond

attacker	corpus	LLM	compl. F_1 (p50)	null	(a) compl. F_1: lift over the null [95% t]	(b) linkage F_1: lift over the seed-only null
agent, all tools	SeqBench-20	yes	0.033	0.035^{E}	−0.001 [−0.070, +0.067]	+0.019 [−0.021, +0.058]
LUAR-MUD	SeqBench-20	no	0.121 (0.000)	0.034^{V}	+0.086 [−0.006, +0.179]	−0.267 [−0.437, −0.096]
LUAR-MUD	n50	no	0.634 (0.620)	0.012^{V}	+0.622 [+0.528, +0.716]	+0.293 [+0.197, +0.390]
LUAR-MUD	n1000*	no	0.123$^{\flat}$ (0.120)	0.000^{V}	+0.123 [+0.066, +0.180]	−0.167 [−0.225, −0.109]
StyleDistance	n50	no	0.124$^{\flat}$ (0.022)	0.017^{V}	+0.107 [+0.068, +0.146]	−0.118 [−0.176, −0.059]
N–S words, Jaccard	n1000*	no	0.051	0.001^{V}	+0.050 [+0.011, +0.089]	−0.237 [−0.267, −0.206]
ESRC, sort rung	n50	yes	—	0.013^{E}	0.673 recall at 90% precision^{R}	Reddit split
AutoProfiler†	n50	yes	—	0.380	accuracy 0.756–0.954 per attribute^{R}	SynthPAI, GPT-4
InferLink/OpenHands	n50	yes	—	0.013^{E}	0.792 linkage success^{R}	Netflix, m=2

the row's null	interval above 0	below 0	crosses 0	95% t-interval excludes 0	crosses 0

Complement F_1 at the f1opt headline point, p50 in parentheses; $^{\flat}$: flagged by the post hoc interpretability rule (Appendix I), LUAR-MUD on n1000* then read at p50. Nulls: ^{E} $\mathbb{E}[\texttt{random}]$ at the declared volume v, which attains F; ^{V} the budget-matched null, merging as many handles as the linker did (merge-everything, outside $\mathcal{N}_v$, scores higher still in every such cell); in (b) the seed-only null, which attains F. Identification ($F = 1.000$) is not shown. N–S: the best of its 12 cells. *A 150-persona sample of personas 300–999. †Attribute accuracy on the loader's seven overlapping columns, against the leave-one-out modal profile.

Figure 6: Published systems and the audited agent, every lift under the declared paired t and every row naming its null; ^{R}: reported by the authors on their own data.

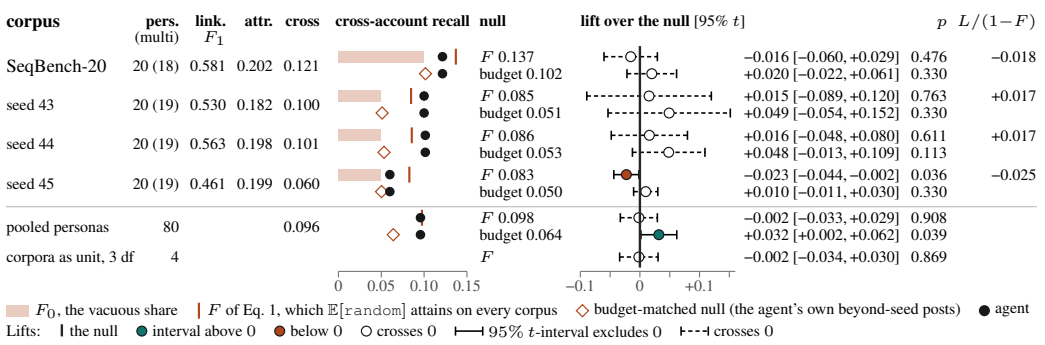


Figure 7: The Q1 audit across corpora against F of Eq. 1 and against a volume-matched random null.

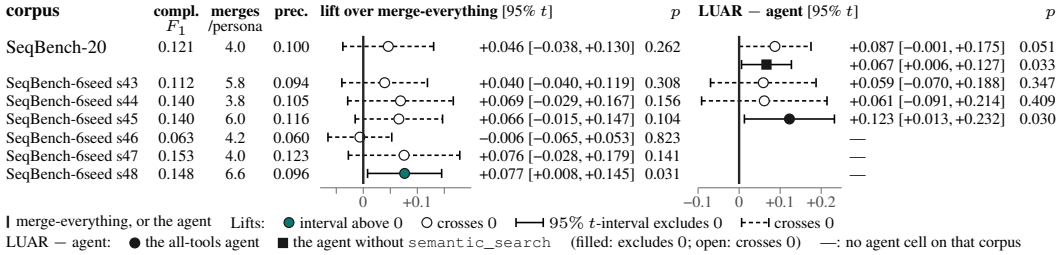


Figure 8: LUAR-MUD (f1opt) on SeqBench-20 and SeqBench-6seed: merging behaviour, lift over merge-everything, and the paired contrast with the agent cell on the same personas.

the seed for 6 of 75 multi-account personas against random's expected 8.7 ($p = 0.258$), claims only on those six, and so beats a random null claiming the same posts on the same personas ($+0.032$ $[+0.002, +0.062]$, $p = 0.039$). Those six merges also carry the one pooled number clearing an unmatched floor, linkage F_1, which lifts $+0.020$ $[+0.003, +0.038]$ over the seed null ($p = 0.020$, fixed effects, conditional on these corpora; with the corpus as the unit, $+0.020$ $[+0.005, +0.036]$, $p = 0.024$). The agent links precisely and too rarely for anything else to clear one (Figure 7; Appendix G).

A stronger attacker on the same floors. Of the published systems on this harness (Figure 6), the zero-shot authorship representation LUAR-MUD (Rivera-Soto et al., 2021) links accounts best with no language model: on SeqBench-n50 and -n150 it lifts complement F_1 by $+0.622$ and $+0.556$ over a random null merging as many handles, and linkage F_1 by $+0.293$ $[+0.197, +0.390]$ and $+0.196$ $[+0.141, +0.251]$ over the seed null; over the best offline ranker of our baselines, itself tuned post hoc, it gains $+0.281$ $[+0.140, +0.423]$ and $+0.276$ $[+0.206, +0.345]$. Five caveats: on the sparse SeqBench-n1000 sample its linkage F_1 falls below the seed null (-0.163 $[-0.221, -0.106]$), as its identification does on every released corpus. On SeqBench-20 and SeqBench-6seed its complement F_1 exceeds the agent's wherever the agent ran, but the paired contrast with the all-tools agent separates on 1 of 4 cells and its lift over the unrestricted null contains zero on 6 of 7 (Figure 8). Its operating point was fixed after the results were seen, and at the conservative point it merges nothing on any of these corpora; at the headline point it merges 3.8–6.6 handles per persona at 6–12% precision, so identification falls to 0.10–0.30 and linkage below the seed floor on 7 of 7. Text volume drives it: SeqBench-20 handles carry about 4 posts against 41 on the released corpora. The classical rung of Lermen et al. (2026) clears its volume-matched null in 1 of 12 cells.

Q2 and ablations. Q2, the one task carrying no seed, clears its floor (1 run). Over 20 queries, each with a single matching person among the 20, mean average precision is 0.642 against 0.180 for a null returning every candidate ($H_{20}/20$), a lift of $+0.462$ $[+0.244, +0.680]$ ($p = 3\times10^{-4}$), while its identification sits below the null's 1.000 (-0.300, $p = 0.010$); the harness substitutes the retrieval baseline's output when the agent abstains without flagging it, so that one run cannot be checked against the substitution (Appendix E). No ablation moves a metric detectably (smallest

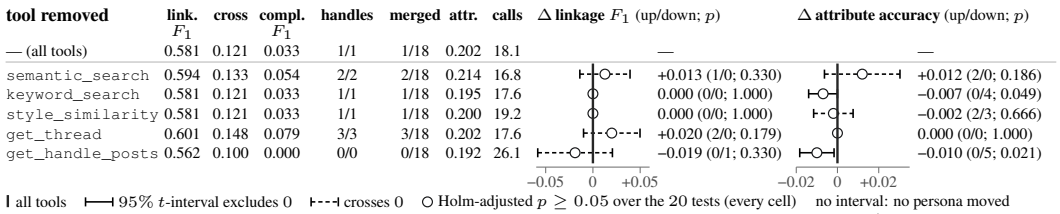


Figure 9: Single-tool ablations on SeqBench-20, n=20, with handle-level linkage and paired tests.

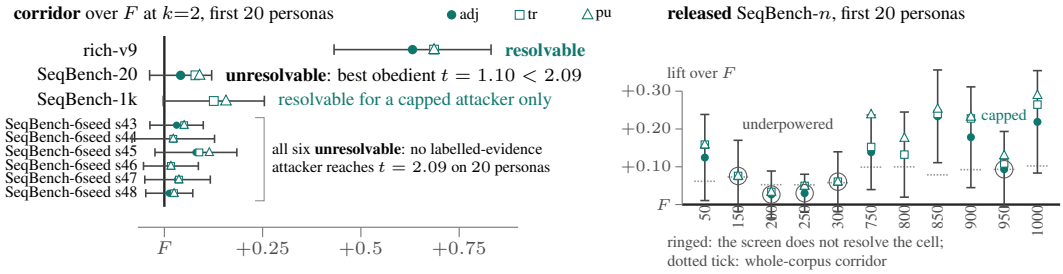


Figure 10: The corridor over F at $k = 2$ on the 20 personas attacked, all three readings, with the admissible set's 95% t-interval and the pre-run verdict; left audited, right released corpora.

Holm-adjusted p 0.42 over the 20 cells, Figure 9), though without `get_handle_posts` the agent claims no handle beyond the seed and scores the null attacker's 0.562 linkage.

7 What the declared policy can reach

An LLM attacker states its decision rule in its own prompt, so the rule acts as a blocking stage and the answer key fixes its pair completeness before anything runs. Write $\mathcal{M}$ for a corpus's true cross-account pairs and π for the declared policy; a *specific* is a labelled item of personal information, an employer or a home city say, that the key records as recoverable from an account's posts. Under a reading ρ of π, $\mathcal{M}_\pi^\rho \subseteq \mathcal{M}$ holds the pairs mergeable on at least k shared specifics compared with the seed (seed-adjacent), with the account merged through (transitive), or with everything merged so far (profile-union). The *labelled-evidence ceiling* and its band are

$$\overline{\mu}_\rho(\pi) \;=\; \mu(\mathcal{M}_\pi^\rho), \qquad B_\rho \;=\; \overline{\mu}_\rho(\pi) - F_0, \tag{3}$$

and $\overline{\mu}_\rho(\pi) - F$ is the corridor (Section 3); both edges come off the key with no model call.

Proposition 1 (Labelled-evidence ceiling: a counting device)**.** *Let μ count only true merges recovered and be monotone in the set recovered, and let A be any attacker every merge of which meets reading ρ's condition through accounts of the same persona that A has itself merged, with no false merge on the path. Then $\mu(A) \leq \overline{\mu}_\rho(\pi)$, and the attacker returning exactly $\mathcal{M}_\pi^\rho$ attains it.*

Proof. Every true account A merges is reached along a path of its own merges, each meeting the condition, so it lies in $\mathcal{M}_\pi^\rho$; monotonicity finishes. □

The work is done by what the proposition counts and by the screen it licenses. Its hypothesis restricts evidence as well as procedure: an attacker merging on evidence the key does not record, style say, is not bound, nor is one chaining through another persona's account, and a rule admitting a merge whenever π holds of a pair's evidence implies only the transitive ceiling (Appendix H).

Reading π off the audited agent. The agent's two-specifics rule fixes $k = 2$, so the system under audit sets k, not the instrument. On SeqBench-20 the three readings give ceilings of 0.178, 0.216 and 0.227, and only 10 of its 121 true account pairs are corroborated directly; counting a shared value only where fewer than a quarter of the personas hold it lowers the ceiling to the vacuous share

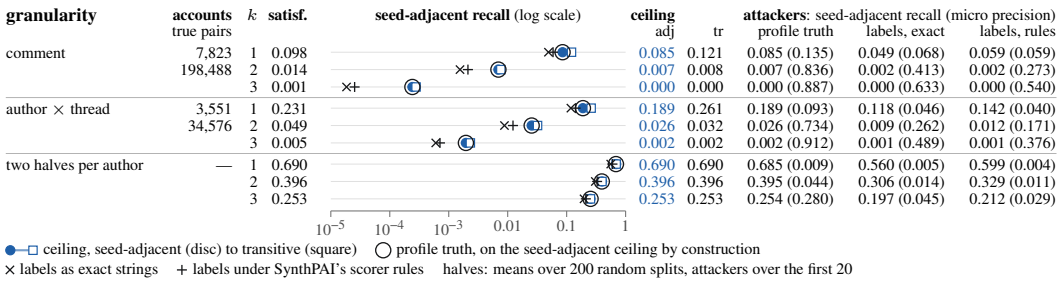


Figure 11: The labelled-evidence ceiling on SynthPAI at three account granularities, seed uniform over each author's accounts, against three attackers' seed-adjacent recall (micro precision).

0.100. The ceiling is therefore a statement about the key's labels: on rich-v9, same generator, seed and size but with 633 in-run leak labels in place of SeqBench-20's 158 from the *post hoc* detector, satisfaction rises from 0.083 to 0.589, the ceiling from 0.178 to 0.726 (about 4.1 times), and the screen flips to resolvable ($t = 7.8$).

Checking the audited agent against π. Every Q1 run with predicted posts (180 persona-runs) holds 13 correct cross-handle merges and no wrong one, 9 of them distinct; judged along the path the agent merged, 4 break the policy's premise under every reading at $k = 2$ (Wilson 95% level interval $[0.19, 0.73]$, secondary), and 3 lie outside even the transitive and profile-union admissible sets. Stylometry, obeying no merge rule, has 17 of its 20 correct merges outside the premise (16 under profile-union) and reaches cross-account recall 0.369, well past the ceiling. No attacker we measured satisfies Proposition 1's hypothesis.

The screen, before the run. The harness attacks a fixed set of 20 personas and the random null enters at its expectation, so on that set the declared paired t is arithmetic on the key. An attacker returning exactly the seed-adjacent admissible set lifts SeqBench-20's cross-account recall by the corridor, 0.041 over 0.137, 95% t-interval $[-0.037, +0.120]$ ($t = 1.10$ against 2.09). Since a smaller but more even lift can carry a larger t, the screen maximises t over the lifts an obedient attacker can reach by sweeping a common cap (Appendix H); on SeqBench-20 the maximum is the admissible set's own. No attacker merging on labelled evidence alone can be told from the random null there at $k = 2$ or $k = 3$, and since the 20 are every persona the corpus holds, the cell is *unresolvable*: for personas from the generator's superpopulation the required n is not identified (plug-in 115; bootstrap $[23, \infty)$, secondary), and the verdict is about this corpus, not the generator. None of the six seed corpora resolves either. On the released corpora the screen splits three ways (Figure 10, right). SeqBench-n150 to -n300 are underpowered at 20 personas but resolve from 42 to 60 drawn at random, and the first 20 are not a random sample (corridor 0.233 against 0.079 over the whole of -n850), so draw them at random.

A corpus we did not build. On SynthPAI (Yukhymenko et al., 2024), whose text and labels we left untouched, the count needs a linkage task the corpus does not define, so the accounts are ours: one per author and thread gives 3,551 accounts and 34,576 true pairs, the policy admits 0.049 of them, and the seed-adjacent ceiling of 0.026 moves by more than an order of magnitude with that choice (0.007 per comment, 0.396 with two halves per author); no label-keyed attacker exceeds it on any seed (Figure 11).

8 Discussion and Limitations

Limits of the audit. The audited agent, under a precision-first prompt, claimed a single beyond-seed handle on SeqBench-20 and merged no wrong one, so a cross-account recall near its floor cannot separate an attacker that could not find a link from one told not to claim it; each cell is one temperature-0 transcript on a non-deterministic API, carrying no run-to-run variance. The external evidence here is therefore the SynthPAI re-analysis and the published systems rather than the audit (Section 4). We ran no human study, because the corpora hold too many posts for a human to trace every leakage back through every post that carries it.

AI USE STATEMENT

In this work, we used generative AI tools to **generate synthetic data sets**: the population this instrument is calibrated on is model-generated, and the attacker under audit is itself a language model, so the artefacts and the measurements alike are AI-produced by construction. We also used them to **implement methods** and experiments, and to **clean and reformat data**. **Assisting with translation** is not applicable to this work. Additionally, we used generative AI tools for code creation and editing, literature analysis, brainstorming, drafting and revising this text, and reference formatting.

We have reviewed all AI-assisted work. Every number in this paper is regenerated by scripts from the stored run reports and the answer keys, with no model call; we reviewed AI-generated code and verified each citation against the cited paper. The models under study (DeepSeek-v4-pro as the audited agent, and the comparators' released checkpoints) are objects of measurement and authored no part of this paper. We take responsibility for the final content of this work, including all text, claims and artefacts produced with the aid of generative AI.

ETHICS STATEMENT

No real person's data was processed to attack anyone. The population is fabricated *de novo* (Anonymous, 2026), so studying attacks on it cannot by construction expose anyone; the equivalent assembled from real users would be personal data in the sense of the cited definitions, and would remain personal data under pseudonymisation (European Parliament and Council of the European Union, 2016). The census component uses only the public UCI Adult extract (Becker & Kohavi, 1996), which carries no direct identifiers and which we read only in aggregate to count equivalence-class sizes, re-identifying nobody and publishing no record-level output; its 19.3% single-out rate on the six quasi-identifiers (Appendix C) is a long-published property of that extract (Sweeney, 2002), recomputed here as a by-product of the floor. Two further public corpora are read and neither widens exposure: SynthPAI (Yukhymenko et al., 2024) is a synthetic population with no real authors, and InferLink (Ko et al., 2026) is synthetic tabular data (CC-BY-4.0). We read each only through its published labels and report counts over them, no record-level output.

A deanonymisation audit is nonetheless dual-use. We will release an agent bound to a synthetic corpus with no index of real people: its tools read corpus files, and it carries neither a search over live platforms nor any link to a real identity. Comparable pipelines are already published (Staab et al., 2024; Lermen et al., 2026; Ko et al., 2026), so withholding this one removes the floor measurement rather than the capability. The instrument lowers the reported attack numbers on every metric we audited. Appendix K elaborates.

REPRODUCIBILITY STATEMENT

The floor instrument, the generalisation lattice and both population sweeps run offline with no API cost, and the lattice includes a self-check asserting, for both the real and the generated schema, that k is monotone up the lattice, that coarsening only merges equivalence classes, and that unmapped categorical values fall back rather than vanish. The real table is the public UCI Adult extract (Becker & Kohavi, 1996) (`adult.data` plus `adult.test`, $45{,}222$ records complete on the six quasi-identifiers), distributed under CC BY 4.0 and used here under that licence. Agent runs are stored as reports carrying model, embedding backend, protocol and a corpus hash, and the offline analysis reproduces exactly from them; Appendix J gives revisions and hashes. Attribute scores use the 25-attribute vocabulary (`DEANON_ATTRIBUTE_SET=v25`); the 26-attribute set adds `disability` and gives, on SeqBench-20, agent 0.231 against a leave-one-out modal floor of 0.404 and an abstention floor of 0.190, so a rerun on that vocabulary reproduces the verdict and not the row. The harness, the scorers, the analysis scripts and the result files behind every number will be released with the camera-ready version.

The ceiling of Section 7 is the cheapest object in the paper to reproduce, by design: it reads an answer key and the threshold k the audited policy declares, counts, and stops. The reference implementation is the module `deanon.scoring.ceiling`: it computes the corpus census; the ceiling under three readings (seed-adjacent, transitive, profile-union) at any k, with an optional independence requirement, a value-rarity threshold that counts one value once, and the label seman-

tics of decoys and cumulative leaks; the random null's analytic expectation; the pre-run screen of Box 1, on the planned personas or on the whole corpus, with power and the n required for random persona samples; and an audit of recorded merges along the path each attacker merged. Its self-check asserts that the readings nest and fall with k and that independence, rarity and stricter labels only tighten the ceiling. It makes no model call and needs no GPU. `scripts/rev1_ceiling.py` regenerates every ceiling, screen and compliance number, and `scripts/rev1_{floors,tables,baselines}.py` every floor, lift and table; the SynthPAI rows, the SynthPAI attribute floors and the InferLink audit are emitted by `scripts/rev1_synthpai.py`, which fetches both corpora at pinned revisions (SynthPAI commit `ba9edc9`, InferLink revision `50dc5ba`) rather than redistributing them.

Because model behaviour drifts under a fixed API name, the reproducible object is the stored report rather than the endpoint that produced it. Table 19 (Appendix J) gives every model seat with the identifier and date its report records: the agent's reports name `deepseek-v4-pro` and `voyage-3-large` with no dated version, and no record names the model of a corpus generator or of the leak labeller.

The artefacts behind the audit are the n=20 Q1 ablation grid, whose all-tools cell is the agent row of Table 1, the three seed-corpus runs, the four baselines and the 20-query Q2 run.

REFERENCES

Michael Aerni, Jie Zhang, and Florian Tramèr. Evaluations of machine learning privacy defenses are misleading. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024. arXiv:2404.17399.

Anonymous. SeqBench: Benchmarking cumulative privacy leakage in longitudinal text, 2026. Companion submission under review; the corpus used here.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. arXiv:1802.00420.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. The Census Income extract; licensed CC BY 4.0.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013.

Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao, and Li Xiong. Quantifying differential privacy under temporal correlations. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 821–832. IEEE, 2017.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.745.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2022. arXiv:2112.03570.

Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1537–1555, 2012. doi: 10.1109/TKDE.2011.127.

Fida Kamal Dankar and Khaled El Emam. A method for evaluating marketer re-identification risk. In *Proceedings of the 2010 EDBT/ICDT Workshops (PAIS)*. ACM, 2010. doi: 10.1145/1754239.1754271.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *IEEE Deep Learning Security and Privacy Workshop (DLSP)*, 2025. Also presented at the DATA-FM workshop, ICLR 2025; arXiv:2406.16201.

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3:1376, 2013.

Yuntao Du, Zitao Li, Bolin Ding, Yaliang Li, Hanshen Xiao, Jingren Zhou, and Ninghui Li. Automated profile inference with language model agents. In *Findings of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2505.12402.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006.

Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association (JAMIA)*, 15(5):627–637, 2008. doi: 10.1197/jamia.M2716.

Mark J. Elliot and Angela Dale. Scenarios of attack: The data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, 14, 1999.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council (general data protection regulation). Official Journal of the European Union L 119, pp. 1–88, 2016. Article 4(1) and Recital 26 (personal data); Article 4(5) (pseudonymisation).

Sajjad Ghiasvand and Nader Sehatbakhsh. GRASP: Reinforcing language model anonymizers with group relative policy optimization. *arXiv preprint arXiv:2608.06526*, 2026.

Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2023(2):312–328, 2023. The Anonymeter toolkit; arXiv:2211.10459.

Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P. Gummadi. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1799–1808. ACM, 2015. doi: 10.1145/2783258.2788601.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2017. arXiv:1803.02324.

Florimond Houssiau, James Jordon, Samuel N. Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. TAPAS: A toolbox for adversarial privacy auditing of synthetic data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research (SyntheticData4ML)*, 2022. arXiv:2211.06550.

Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022. arXiv:2209.01292.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2407.01502.

Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5010–5015. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1546. arXiv:1808.04926.

Kyuyoung Kim, Hyunjun Jeon, and Jinwoo Shin. Self-refining language model anonymizers via adversarial distillation. *arXiv preprint arXiv:2506.01420*, 2025.

Myeongseob Ko, Jihyun Jeong, Sumiran Singh Thakur, Gyuhak Kim, and Ruoxi Jia. From weak cues to real identities: Evaluating inference-driven de-anonymization in LLM agents. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.18382.

Nataša Krčo, Zexi Yao, Matthieu Meeus, and Yves-Alexandre de Montjoye. RAT-Bench: A comprehensive benchmark for text anonymization. *arXiv preprint arXiv:2602.12806*, 2026.

Simon Lermen, Daniel Paleka, Joshua Swanson, Michael Aerni, Nicholas Carlini, and Florian Tramèr. Large-scale online deanonymization with LLMs. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 1947–1966, 2026. arXiv:2602.16800.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering (ICDE)*, pp. 106–115. IEEE, 2007.

Ziwen Li, Jianing Wen, and Tianshi Li. LLM anonymization against agentic re-identification. *arXiv preprint arXiv:2605.30848*, 2026.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. Tau-Eval: A unified evaluation framework for useful and private text anonymization. In *Proceedings of EMNLP 2025: System Demonstrations*, 2025.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (S&P)*, pp. 111–125. IEEE, 2008. doi: 10.1109/SP.2008.33.

Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *2009 IEEE Symposium on Security and Privacy (S&P)*, pp. 173–187. IEEE, 2009.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021. arXiv:2103.14749.

Myeong Seok Oh, Dong-Yun Kim, Hanseok Oh, Chaean Kang, Joeun Kang, Xiaonan Wang, Hyunjung Park, Young Cheol Jung, and Hansaem Kim. Subject-level inference for realistic text anonymization evaluation. *arXiv preprint arXiv:2604.21211*, 2026.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1: Long Papers*, pp. 8662–8685. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.436.

Zifan Peng, Yini Huang, Aiwen Lu, Qiming Ye, Peixian Zhang, Jingyi Zheng, Yule Liu, Xuechao Wang, Xinlei He, and Jiaheng Wei. What your posts reveal: A benchmark and agentic framework for user-level privacy leakage on social media. *arXiv preprint arXiv:2606.06784*, 2026.

Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. How unique and traceable are usernames? In *Privacy Enhancing Technologies (PETS)*, Lecture Notes in Computer Science, pp. 1–17. Springer, 2011. doi: 10.1007/978-3-642-22263-4_1.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 180–191. Association for Computational Linguistics, 2018. doi: 10.18653/v1/S18-2023. arXiv:1805.01042.

Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. Flexible data anonymization using ARX—current status and challenges ahead. *Software: Practice and Experience*, 50(7):1277–1304, 2020. doi: 10.1002/spe.2812.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 913–919. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.70.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10:3069, 2019.

Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. SoK: Let the privacy games begin! a unified treatment of data inference privacy in machine learning. In *2023 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2023. arXiv:2212.10986.

Chenyang Shao, Tianxing Li, Chenhao Pu, Fengli Xu, and Yong Li. AgentStealth: Reinforcing large language model for anonymizing user-generated text. *arXiv preprint arXiv:2506.22508*, 2025.

Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. User identity linkage across online social networks: A review. *ACM SIGKDD Explorations Newsletter*, 18(2):5–17, 2017.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Large language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2402.13846.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Latanya Sweeney. Simple demographics often identify people uniquely. Technical Report Data Privacy Working Paper 3, Carnegie Mellon University, Pittsburgh, PA, 2000.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

Matthias Templ, Alexander Kowarik, and Bernhard Meindl. Statistical disclosure control for microdata using the R package sdcMicro. *Journal of Statistical Software*, 67(4):1–36, 2015. doi: 10.18637/jss.v067.i04.

Oscar Thees, Roman Müller, and Matthias Templ. Agentic AI-powered re-identification: An emerging, scalable threat to mobility microdata privacy. In *Privacy in Statistical Databases (PSD 2026)*, volume 16925 of *Lecture Notes in Computer Science*, pp. 319–332. Springer, Cham, 2027. arXiv:2606.27936.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & QA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1977–1983. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1197.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2002.08347.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP (RepL4NLP)*, pp. 249–268. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.repl4nlp-1.26.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. A synthetic dataset for personal attribute inference. In *Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2024. URL `https://openreview.net/forum?id=1nqfIQIQBf`.

Reza Zafarani and Huan Liu. Connecting users across social media sites: A behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 41–49. ACM, 2013. doi: 10.1145/2487575.2487648.

Arman Zharmagambetov, Chuan Guo, Ivan Evtimov, Maya Pavlova, Ruslan Salakhutdinov, and Kamalika Chaudhuri. AgentDAM: Privacy leakage evaluation for autonomous web agents. In *Thirty-ninth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2025. arXiv:2503.09780.

A FLOORS AND LIFT, FORMALLY

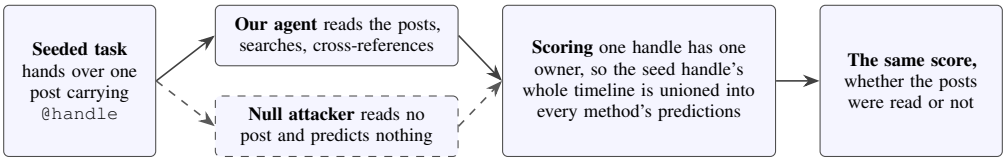


Figure 12: Why the seed floor stands at 1.000 on SeqBench-20: a procedure that reads no post reaches our agent's identification and exact ownership alike (Table 1).

Definition 1 (Floor and lift)**.** Let $\mathcal{P}$ be a population of individuals, $\mathcal{T}$ a task presenting an attacker with some seed and declaring a claim volume v, and μ a scoring rule; an *attacker* A is a deanonymiser $\mathcal{D}$ together with the procedure that runs it under $\mathcal{T}$. For the family $\mathcal{N}_v$ of *null attackers* (procedures that consume the seed and the population's public structure, claim at most v, and perform no inference on the evidence the task is ostensibly about), the *floor* and the *lift* of A are

$$F(\mathcal{T}, \mathcal{P}, \mu, v) \;=\; \max_{A_0 \in \mathcal{N}_v} \mu(A_0), \qquad L(A) \;=\; \mu(A) - F.$$

A procedure reading no post but claiming past v, such as merge-everything, is an *unrestricted null*: exhibited beside $\mathcal{N}_v$, never inside it.

Glossary. F_0 is the vacuous single-account share (single-account personas score cross-account recall 1 with nothing claimed); *satisfaction* is the share of true account pairs the declared policy admits directly. F, the band, the corridor and the five nulls of $\mathcal{N}_v$ are as defined in Section 3; every table names the budget-matched null's unit (posts or handles).

Proof of Proposition 2. Condition on the target's class C. The attacker sees only C, so the law q of its guess depends on C alone, and a guess outside C is wrong with certainty. Under exchangeability the target is uniform on C, so any q supported on C succeeds with probability $\sum_{p\in C} q(p)/|C| = 1/|C|$, and placing mass outside C only lowers this. Averaging over the target's class, which carries weight $|C|/|\mathcal{P}|$, gives $\sum_C (|C|/|\mathcal{P}|)\cdot(1/|C|) = |\{C_Q\}|/|\mathcal{P}| = F_{\mathrm{id}}(Q)$. □

B METRIC DEFINITIONS AND THE AUDIT TABLES

Table 1: Q1 audit on SeqBench-20 (all-tools run), each column against F, the maximum over the null family $\mathcal{N}_v$, and against a random null matched to the agent's claim volume.

method	**ident.**	**exact-own.**	**linkage** F_1	**cross-acct.**	**compl.** F_1	**attr. acc.**
random, expectation (exact)	0.297	0.000	0.309	0.137	0.035	0.338
random, budget-matched to the agent	—	—	0.560	0.102	0.002	—
random, one Random(42) draw	0.300	0.000	0.321	0.147	0.061	0.300
merge everything (outside $\mathcal{N}_v$)	0.050	0.000	0.093	1.000	0.074	0.160
retrieval	0.050	0.000	0.144	0.180	0.028	0.160
stylometry	0.050	0.000	0.122	0.369	0.070	0.160
null attacker (seed floor, modal LOO)	1.000	1.000	0.562	0.100	0.000	0.382
abstention floor	—	—	—	—	—	0.160
F (Eq. 1)	1.000	1.000	0.562	0.137	0.035	0.382
agent (n=20, all tools)	1.000	1.000	0.581	0.121	0.033	0.202
agent (n=5 prefix)	1.000	—	0.649	0.385[†]	—	0.300
lift (agent − F)	0 (identity)	0 (identity)	+0.019	-0.016	-0.001	-0.180
95% interval[‡]			[-0.021, +0.058]	[-0.060, +0.029]	[-0.070, +0.067]	[-0.227, -0.133]
test p[§]			0.330	0.476	0.967	2×10^{-7}
one-sided 95% upper bound[‡]			+0.052	+0.021	+0.055	-0.141
lift (agent − budget-matched)			—	+0.020	+0.031	—
95% interval[‡]			—	[-0.022, +0.061]	[-0.034, +0.096]	—
test p[§]			—	0.330	0.331	—

Complement F_1 averages the 18 multi-account personas. Random's expectation is exact over all $\binom{73}{3}$ handle draws; its attribute cell is the expected persona transplant (donor drawn from all 20, as the baseline does). The budget-matched random null claims exactly the agent's own beyond-seed posts per persona (9 posts, on one persona) uniformly from the non-seed posts; at the handle level it scores F_1 0.493 against the agent's 0.506. Fitted on the 14 train personas, the modal null scores 0.400. [‡]Every interval and one-sided bound on a lift is the declared test on each persona's lift over the column's null (the seed null for linkage, random's expectation for cross-account and complement, the modal null for attributes, the budget-matched null below). [§]Two-sided p of that test. Secondary readings: the linkage lift rests on 1 merged persona of 18 (exact merge-count interval [0.000, +0.093], exact sign test $p = 1.000$); the cross-account lift's exact merge-count interval [-0.036, +0.068] and percentile bootstrap [-0.041, +0.029]; the attribute lift's exact sign test ($p = 4\times10^{-6}$, 19 personas lower, 0 higher). Exact Poisson-binomial p of the agent's merge count, secondary: $p = 0.716$ against random, $p = 0.998$ against the budget-matched null (declared t in Table 2). [†]On the five prefix personas F is 0.233 (vacuous share 0.200, random's expectation 0.233).

Remark 1. The instrument constrains the defence-side metric too. The quote-validated recall of the companion submission (Anonymous, 2026) checks whether a quote survives, leaving unchecked whether it still supports the attribute, so a null attacker naming every attribute and quoting any surviving fragment scores 1.000. With the floor sitting at the ceiling, the absolute levels there carry no information.

Table 1 places quantities of three different kinds side by side, and none of them is comparable to another without its denominator. Table 4 gives each one as implemented, in the following notation. A *run* r is one seeded persona; R is the set of scored runs, so every aggregate below is a macro mean over runs, never a pool over posts or cells. Let s_r be the seed post, $a(i)$ the key's author of post i, $h(i)$ its handle, $\sigma(p)$ every post p authored across all of its handles, and $\hat{\sigma}_r$ the predicted post set after the seed floor of Section 3 has unioned $\{i : h(i) = h(s_r)\}$ into it. Let $\mathcal{V}$ be the fixed 25-attribute vocabulary, $z_{r,a}$ the truth for attribute a, $\hat{z}_{r,a}$ the prediction ($\varnothing$ for an abstention), and $m(\hat{z}, z) \in \{0, \frac{1}{2}, 1\}$ the graded match: exact after case and punctuation normalisation, ±2 years on `age`, $\frac{1}{2}$ for an adjacent education or income band, any shared whitespace token for the six geographic attributes, and ISCO major-group agreement for `occupation`. Predicted ids that exist in no timeline fall outside $\mathrm{dom}\, a$, so they are dropped from the vote rather than counted against it. Complement scores (Section 3) average over the multi-account personas.

Attribute accuracy and abstention. It is a per-cell mean over a fixed vocabulary, so a persona with one attribute right out of 25 scores 0.04 rather than 0 or 1. Two conventions inflate it and both are load-bearing: an abstention scores 1.0 whenever the truth normalises to empty, and `employer` pays a flat 0.5 for abstaining because employer is not finely encoded. Read the attribute column as a

Table 2: Multi-account personas with any correct cross-handle merge.

corpus	attacker	k/n	95% level CI (exact)	claimed	E[random]	p	E[budget]	p
SeqBench-20	agent[all]	1/18	[0.001, 0.273]	1	2.10	0.278	0.50	0.331
SeqBench-20	agent[semantic_search]	2/18	[0.014, 0.347]	2	2.10	0.939	1.15	0.170
SeqBench-20	agent[keyword_search]	1/18	[0.001, 0.273]	1	2.10	0.278	0.50	0.331
SeqBench-20	agent[style_similarity]	1/18	[0.001, 0.273]	1	2.10	0.278	0.50	0.331
SeqBench-20	agent[get_thread]	3/18	[0.036, 0.414]	3	2.10	0.566	1.42	0.098
SeqBench-20	agent[get_handle_posts]	0/18	[0.000, 0.185]	0	2.10	4×10^{-8}‡	0.00	1.000
SeqBench-20	random (Random(42) draw)	3/18	[0.036, 0.414]	18	2.10	—	—	—
SeqBench-20	retrieval	3/18	[0.036, 0.414]	18	2.10	0.585	12.73	1×10^{-4}
SeqBench-20	stylometry	12/18	[0.410, 0.867]	18	2.10	7×10^{-5}	15.57	0.033
SeqBench-20	majority_class	0/18	[0.000, 0.185]	0	2.10	4×10^{-8}‡	0.00	1.000
seed 43	agent[all]	1/19	[0.001, 0.260]	1	2.19	0.298	0.11	0.331
seed 43	random (Random(42) draw)	3/19	[0.034, 0.396]	19	2.19	—	—	—
seed 44	agent[all]	3/19	[0.034, 0.396]	3	2.17	0.609	1.13	0.086
seed 44	random (Random(42) draw)	1/19	[0.001, 0.260]	19	2.17	—	—	—
seed 45	agent[all]	1/19	[0.001, 0.260]	1	2.23	0.217	0.06	0.331
seed 45	random (Random(42) draw)	1/19	[0.001, 0.260]	19	2.23	—	—	—
seed 46	random (Random(42) draw)	3/20	[0.032, 0.379]	20	2.12	—	—	—
seed 47	random (Random(42) draw)	4/18	[0.064, 0.476]	18	2.17	—	—	—
seed 48	random (Random(42) draw)	1/18	[0.001, 0.273]	18	2.17	—	—	—

Exact Clopper–Pearson 95% level intervals on the merge rate, a secondary reading beside the declared t. *claimed*: multi-account personas with any beyond-seed claim. Each attacker's merge count is read against random's exact per-persona merge probabilities and against a random null claiming the attacker's own volume per persona as a lift: per multi-account persona, the merge indicator minus the null's merge probability, under the declared t (p); the exact Poisson-binomial test of the count is a secondary field in the JSON. Agent (all tools) against the Random(42) draw on the same personas: 1 persona(s) merged only by the agent, 3 only by random (declared t on the paired merge indicators $p = 0.331$; exact McNemar $p = 0.625$, secondary). ‡The attacker merges no persona, so the p measures a deficit of claim volume, not of accuracy (the declared test still decides; SeqBench-20 agent[get_handle_posts] and majority_class against E[random]: the null's merge probabilities have sd 0.053, the null itself merges no persona with probability 0.104, Poisson-binomial $p = 0.208$; the last two are secondary).

Table 3: Attribute accuracy under readings that separate abstention, grading strictness and inference.

corpus	attacker	strict	canon.	contain.	modal LOO	cov.	sel.	null same	sel. − null (p)	null top	knowable	modal top-1/3
SeqBench-20	agent[all]	0.202	0.220	0.226	0.382	0.164	0.256	0.360	-0.104 ($p = 0.073$)	0.366	0.174	0.413/0.592
SeqBench-20	agent[semantic_search]	0.214	0.232	0.240	0.382	0.186	0.290	0.339	-0.048 ($p = 0.477$)	0.441	0.228	0.413/0.592
seed 43	agent[all]	0.182	0.210	0.214	0.368	0.188	0.117	0.319	-0.202 ($p = 0.002$)	0.479	0.070	0.320/0.595
seed 44	agent[all]	0.198	0.236	0.234	0.418	0.206	0.136	0.350	-0.214 ($p = 4\times10^{-5}$)	0.583	0.094	0.358/0.623
seed 45	agent[all]	0.199	0.217	0.227	0.412	0.184	0.152	0.299	-0.147 ($p = 0.046$)	0.446	0.100	0.359/0.686

canon.: answers mapped onto the key's value vocabulary built without the graded persona, so no category exists only because it is the truth; *contain.*: full credit when the answer contains every truth token (reads the truth, an upper bound). *cov./sel.*: share of cells answered ("none" is an abstention) and accuracy on them. *null same*: the leave-one-out modal null on exactly the agent's answered cells, then agent minus null there, with the p of the declared paired two-sided t on each persona's sum of agent-minus-null over its answered cells (the persona-cluster bootstrap and the exact sign-flip test on the same sums are secondary, in the JSON). *null top*: the null answering only its most confident cells at the agent's coverage, where the null's own "none" predictions count as abstentions under the same rule as the agent's. *knowable*: cells a leak label marks, agent against the modal null's top-1/top-3.

graded 25-cell mean with an abstention subsidy, and compare it against the abstention floor of 0.160 and the leave-one-out modal floor of 0.382, never against 0.

Rows of Table 1. `random`'s single `Random(42)` draw is printed beside its expectation. `stylometry` merges every handle whose 82-dimensional style vector (58 style features and 24 hour-of-day bins) has cosine above 0.6 to the seed handle's, with no attribute prediction, so its attribute column is pure abstention credit. `retrieval` takes the seed handle plus its 8 nearest handles by centroid cosine of 256-dimensional hashed bag-of-words embeddings. The *null attacker* predicts no posts of its own and answers every attribute with the modal value fitted leave-one-persona-out. `retrieval` and `stylometry` read the evidence and are weak attackers, not floors; `random` reads none, so it passes the membership test of Section 3.

The strict identity variants. Across the four baselines, purity and exact-owner rate are $0.305/0.000$ (random), $0.094/0.000$ (retrieval), $0.072/0.000$ (stylometry) and $1.000/1.000$ for the null attacker. The five-persona agent run records neither field and keeps no predicted post sets, so they cannot be recomputed for it.

Table 4: Every reported metric as implemented, with its denominator and granularity. $\mathbf{1}[\cdot]$ is an indicator; $\mathrm{dom}\, a$ is the set of post ids the key recognises.

metric	as implemented	denominator / granularity
identification acc.	$\frac{1}{\lvert R\rvert}\sum_r \mathbf{1}[\hat{p}_r = a(s_r)]$, $\hat{p}_r = \arg\max_p \lvert\{i \in \hat{\sigma}_r : a(i) = p\}\rvert$, ties to lowest id	$\lvert R\rvert$ runs; one boolean per run, so macro = micro. A *plurality vote* over the predicted cluster, not top-1 over a ranked list
exact-owner rate	$\frac{1}{\lvert R\rvert}\sum_r \mathbf{1}[a(i) = a(s_r)\ \forall i \in \hat{\sigma}_r \cap \mathrm{dom}\, a]$	$\lvert R\rvert$ runs; the strict reading of "identified the person": the whole cluster is theirs
cluster purity	$\frac{1}{\lvert R\rvert}\sum_r \big\lvert\{i \in \hat{\sigma}_r : a(i) = \hat{p}_r\}\big\rvert \big/ \big\lvert\hat{\sigma}_r \cap \mathrm{dom}\, a\big\rvert$	recognised predicted posts within a run, then macro over runs
linkage F_1	$\frac{1}{\lvert R\rvert}\sum_r \frac{2P_rR_r}{P_r+R_r}$, $P_r = \frac{\lvert\hat{\sigma}_r \cap \sigma(p_r)\rvert}{\lvert\hat{\sigma}_r\rvert}$, $R_r = \frac{\lvert\hat{\sigma}_r \cap \sigma(p_r)\rvert}{\lvert\sigma(p_r)\rvert}$	the person's entire timeline; s_r counts on both sides. Macro mean of per-run F_1, not a micro F_1 pooled over posts
cross-account recall	$\frac{1}{\lvert R\rvert}\sum_r \frac{\lvert\hat{\sigma}_r \cap \sigma^{\neg h}(p_r)\rvert}{\lvert\sigma^{\neg h}(p_r)\rvert}$, $\sigma^{\neg h} = \{i \in \sigma(p_r) : h(i) \neq h(s_r)\}$, $\equiv 1$ when $\sigma^{\neg h} = \emptyset$	posts outside the seed handle; single-account personas have an empty denominator and score 1 vacuously (F_0)
attribute acc. (all)	$\frac{1}{\lvert R\rvert}\sum_r \frac{1}{\lvert\mathcal{V}\rvert}\sum_{a\in\mathcal{V}} m(\hat{z}_{r,a}, z_{r,a})$	always $\lvert\mathcal{V}\rvert = 25$ cells per run, whatever was answered or populated
attribute acc. (knowable)	same over $\mathcal{V}_r = \{a :$ a appears in a leak label on $\sigma(p_r)\}$	$\lvert\mathcal{V}_r\rvert$, which varies by run; runs with no leak label are dropped from the mean
Q2 mAP	$\frac{1}{\lvert Q\rvert}\sum_q \frac{1}{\lvert M_q\rvert}\sum_{j\,:\,c_j\in M_q} \frac{\lvert\{j' \le j : c_{j'} \in M_q\}\rvert}{j}$ over distinct ranked candidates c_j	$\lvert M_q\rvert$, the constraint match-set size, not the number of hits. Every $\lvert M_q\rvert = 1$ here, so mAP is reciprocal rank

Table 5: What k and the floors register. Left: population size varying, lattice held at rung 2 on the census. Right: the lattice swept at fixed population, each row's statistic named in its stub.

Population varies (census, rung 2)			**Lattice varies** (fixed population)		
N	k	F_{id}	corpus (statistic)	k	floor
20	1	0.860	census (attribute floor)	1	$0.489 \to 1.000$
45,222	1	0.0099	sampler, N=1,000 (F_{id})	1	$0.999 \to 0.788$
change	$\times 1$	$\times 1/87$			

C THE POPULATION SWEEP

Figure 13 plots the sweep behind Section 5 and Table 6 gives every cell of it: population size, lattice rung, k-anonymity, number of equivalence classes, *unique* the share of records alone in their class, H the class-distribution entropy in bits, and the two floors, F_{id} and an *attr. floor* that is modal-value accuracy over the six quasi-identifiers. The census panel resamples: each of its cells below the full table is a mean over five subsample draws, so its k and class counts carry a fractional part, and $k_{\min}$ is printed beside k; the full 45,222-record row and every cell of the sampler panel are a single exact computation. The sampler panel draws personas from the SeqBench generator's sampler, not the SeqBench-1k personas. Figure 13 regenerates with `figs/make_sweep_curves.py` from `figs/k_lattice_adult.json` and `figs/k_lattice.json`, the files Tables 6 and 7 are rendered from. The class count at rung 3 saturates at 12, the full cardinality of what that rung leaves: binary age $\times$ sex $\times$ a three-point education scale.

At rung 0 on the full table, 19.3% of the 45,222 people are unique on nothing more than exact age, sex, race, native country, education and occupation, the classical single-out result (Sweeney, 2002).

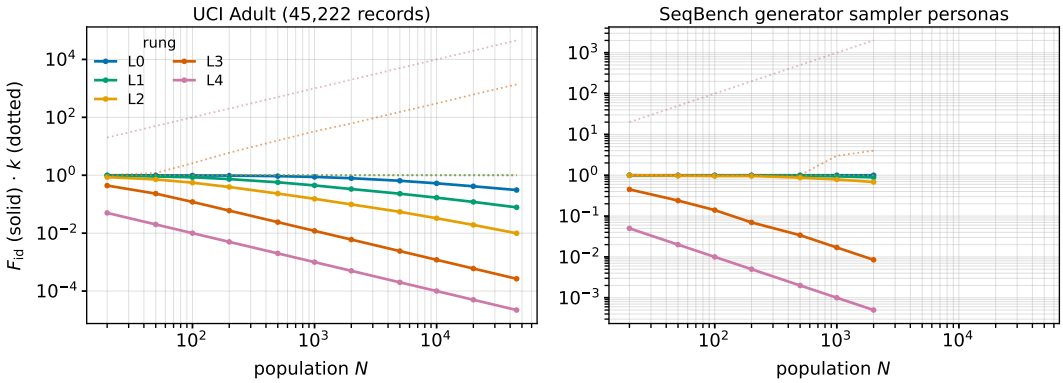


Figure 13: The population sweep behind Section 5: UCI Adult (left) against personas from the SeqBench generator's sampler (right). Solid F_{id}, dotted k-anonymity, one colour per rung; log axes.

Table 6: Every cell of the census panel of Figure 13: the UCI Adult extract coarsened through the five-rung lattice at eleven population sizes.

N	rung	k	$k_{\min}$	classes	unique	H (bits)	F_{id}	attr. floor
20	0	1.0	1	20.0	1.000	4.32	1.0000	0.525
	1	1.0	1	19.6	0.960	4.28	0.9800	0.593
	2	1.0	1	17.2	0.730	4.03	0.8600	0.653
	3	1.0	1	8.8	0.160	2.93	0.4400	0.780
	4	20.0	20	1.0	0.000	0.00	0.0500	1.000
50	0	1.0	1	49.4	0.976	5.62	0.9880	0.509
	1	1.0	1	45.2	0.816	5.45	0.9040	0.585
	2	1.0	1	35.8	0.528	4.99	0.7160	0.653
	3	1.2	1	11.6	0.020	3.26	0.2320	0.788
	4	50.0	50	1.0	0.000	0.00	0.0200	1.000
100	0	1.0	1	98.2	0.964	6.61	0.9820	0.495
	1	1.0	1	85.2	0.736	6.32	0.8520	0.568
	2	1.0	1	55.4	0.324	5.49	0.5540	0.638
	3	2.6	2	12.0	0.000	3.36	0.1200	0.782
	4	100.0	100	1.0	0.000	0.00	0.0100	1.000
200	0	1.0	1	194.6	0.947	7.59	0.9730	0.487
	1	1.0	1	147.0	0.579	7.00	0.7350	0.568
	2	1.0	1	78.4	0.195	5.83	0.3920	0.635
	3	6.0	4	12.0	0.000	3.42	0.0600	0.773
	4	200.0	200	1.0	0.000	0.00	0.0050	1.000
500	0	1.0	1	465.2	0.868	8.82	0.9304	0.487
	1	1.0	1	285.6	0.395	7.76	0.5712	0.572
	2	1.0	1	116.6	0.098	6.12	0.2332	0.638
	3	15.8	14	12.0	0.000	3.44	0.0240	0.776
	4	500.0	500	1.0	0.000	0.00	0.0020	1.000
1,000	0	1.0	1	874.8	0.773	9.69	0.8748	0.490
	1	1.0	1	444.8	0.283	8.16	0.4448	0.574
	2	1.0	1	153.4	0.062	6.22	0.1534	0.641
	3	32.6	27	12.0	0.000	3.43	0.0120	0.780
	4	1,000.0	1,000	1.0	0.000	0.00	0.0010	1.000
2,000	0	1.0	1	1,577.6	0.645	10.48	0.7888	0.491
	1	1.0	1	667.0	0.197	8.44	0.3335	0.577
	2	1.0	1	196.6	0.035	6.29	0.0983	0.642
	3	61.4	51	12.0	0.000	3.43	0.0060	0.781
	4	2,000.0	2,000	1.0	0.000	0.00	0.0005	1.000
5,000	0	1.0	1	3,222.2	0.471	11.35	0.6444	0.490
	1	1.0	1	1,152.2	0.131	8.77	0.2304	0.576
	2	1.0	1	272.6	0.017	6.39	0.0545	0.641
	3	154.0	148	12.0	0.000	3.43	0.0024	0.781
	4	5,000.0	5,000	1.0	0.000	0.00	0.0002	1.000
10,000	0	1.0	1	5,256.4	0.359	11.86	0.5256	0.490
	1	1.0	1	1,679.6	0.089	8.93	0.1680	0.576
	2	1.0	1	327.0	0.008	6.41	0.0327	0.642
	3	302.0	291	12.0	0.000	3.42	0.0012	0.781
	4	10,000.0	10,000	1.0	0.000	0.00	0.0001	1.000
20,000	0	1.0	1	8,300.6	0.268	12.27	0.4150	0.490
	1	1.0	1	2,397.0	0.058	9.05	0.1199	0.576
	2	1.0	1	384.8	0.004	6.42	0.0192	0.643
	3	611.4	586	12.0	0.000	3.42	0.0006	0.782
	4	20,000.0	20,000	1.0	0.000	0.00	5.0×10^{-5}	1.000
45,222	0	1.0	1	13,956.0	0.193	12.63	0.3086	0.489
	1	1.0	1	3,518.0	0.034	9.14	0.0778	0.576
	2	1.0	1	447.0	0.001	6.44	0.0099	0.643
	3	1,371.0	1,371	12.0	0.000	3.42	0.0003	0.781
	4	45,222.0	45,222	1.0	0.000	0.00	2.2×10^{-5}	1.000

Table 7: The sampler panel of Figure 13: sampler personas under the matched five-rung lattice at seven population sizes, one exact computation per cell. Columns as in Table 6.

N	rung	k	classes	unique	H (bits)	F_{id}	attr. floor
20	0	1	20	1.000	4.32	1.0000	0.267
	1	1	20	1.000	4.32	1.0000	0.333
	2	1	20	1.000	4.32	1.0000	0.492
	3	1	9	0.100	3.05	0.4500	0.758
	4	20	1	0.000	0.00	0.0500	1.000
50	0	1	50	1.000	5.64	1.0000	0.223
	1	1	49	0.960	5.60	0.9800	0.283
	2	1	49	0.960	5.60	0.9800	0.467
	3	1	12	0.020	3.37	0.2400	0.760
	4	50	1	0.000	0.00	0.0200	1.000
100	0	1	100	1.000	6.64	1.0000	0.215
	1	1	98	0.960	6.60	0.9800	0.298
	2	1	96	0.920	6.56	0.9600	0.463
	3	1	14	0.030	3.42	0.1400	0.745
	4	100	1	0.000	0.00	0.0100	1.000
200	0	1	200	1.000	7.64	1.0000	0.218
	1	1	198	0.980	7.62	0.9900	0.303
	2	1	192	0.920	7.56	0.9600	0.468
	3	1	14	0.010	3.39	0.0700	0.755
	4	200	1	0.000	0.00	0.0050	1.000
500	0	1	500	1.000	8.97	1.0000	0.202
	1	1	490	0.962	8.92	0.9800	0.299
	2	1	436	0.774	8.68	0.8720	0.470
	3	1	17	0.002	3.51	0.0340	0.749
	4	500	1	0.000	0.00	0.0020	1.000
1,000	0	1	999	0.998	9.96	0.9990	0.197
	1	1	940	0.885	9.84	0.9400	0.291
	2	1	788	0.662	9.44	0.7880	0.467
	3	3	17	0.000	3.55	0.0170	0.747
	4	1,000	1	0.000	0.00	0.0010	1.000
2,000	0	1	1,999	0.999	10.96	0.9995	0.198
	1	1	1,776	0.798	10.72	0.8880	0.291
	2	1	1,367	0.515	10.13	0.6835	0.460
	3	4	17	0.000	3.54	0.0085	0.746
	4	2,000	1	0.000	0.00	0.0005	1.000

D WHICH WAY EACH FLOOR TRANSFERS

Each floor transfers in a direction the argument fixes, and only in one component. For F_{id} the argument is the assumption behind Proposition 2: holding the equivalence classes fixed, a real population breaks within-class exchangeability in the attacker's favour (popularity priors, unequal sampling rates), so the *exchangeability* component of an F_{id} measured on a sampler's output is a lower bound on what a real population concedes. The class counts do not transfer with it, and on our data they dominate: at $N = 1{,}000$ the sampler's personas give F_{id} $0.999/0.940/0.788$ at rungs $0/1/2$ against the census's $0.875/0.445/0.153$, so the sampler's population is the more distinguishable by a factor of five at rung 2 and an F_{id} read off it bounds no real population's. The attribute floor at a fixed coarseness argues the same way, a sampler spreading attribute mass more evenly than a real population does, and the 0.489 against 0.197 of Section 5 is that gap measured once. The seed floor carries no direction: it is saturated on this corpus, with identification and exact-owner attainment conceded in full (Section 6), so there is no headroom above it for a direction to point in. The abstention floor points the other way, being the share of cells whose truth is empty, and a synthetic key with sparser truths concedes more of them, not fewer. And the vacuous component of cross-account recall is a generator parameter, the share of personas the sampler gave a single account, so it transfers nowhere. Where the direction does hold, a lift measured against the floor is an upper bound on the real one.

E AUDITED AGENT DETAIL

The audited corpora. In full, as Section 4 names them: SeqBench-20: 20 personas, 293 posts, 73 accounts, 6 platforms, posts spanning 6.4 days (7 configured), seed 42; SeqBench-6seed: six seed corpora, seeds 43–48, ∼20 personas each, 1,956 posts pooled, 6 platforms, 7 configured days; SeqBench-1k: 1,000 personas generated, 950 posting, 15,785 posts, 2,924 accounts, 6 platforms, posts spanning 7.9 days of 30 configured, seed 42; SeqBench-1k/150: 150-persona sample of SeqBench-1k, 3,083 posts, 858 targets, 6 platforms; and rich-v9: 20 personas (18 posting), 250 posts, 72 accounts, 6 platforms, in-run leak labels, seed 42. Revisions, labellers and hashes are in Table 18 (Appendix J).

Observable view and tasks. The agent observes the corpus as published, prior to any anonymisation, across all platforms with its handles, timestamps and thread structure. On Q2 it returns the matching people with their posts. The loop terminates on a structured submission tool. The audited corpora are text-only with empty media tables, so every number reported measures a text-only attacker.

On the all-tools run of Table 1 the agent scores zero on twelve of the twenty-five attributes (`age`, `age_range`, `sex`, `education_level`, `workplace_area`, `income_band`, `financial_distress`, `relationship_status`, `caregiving_burden`, `immigration_status`, `native_language` and `political_lean`), and its 0.202 mean is carried by a short list, most of which is not inference. The four highest per attribute means are `mental_health` 0.95 and `orientation` 0.95 (one covered persona each), `religion` 0.80 (four) and `health_condition` 0.70 (six): these are sparsely populated sensitive attributes, where abstaining on an empty truth scores 1.0. Of the attributes covered for all twenty personas, only `employer` (0.50, the flat abstention credit) and `current_region` (0.45) are appreciable. Restricting to the attributes a leak label marks as recoverable lowers the mean to 0.174.

Q2. The retrieval baseline's mean average precision is 0.149; the agent's identification rate (at least one returned candidate inside the match set) is 0.70. The per-query reciprocal ranks are bimodal: twelve 1.0s, six 0.0s, one 0.5 and one 0.33. Mean tool calls were 16.8 against 18.1 on Q1.

F TOOL ABLATIONS

The tools of Section 4 are the agent's evidence sources, so ablating them one at a time attributes its score to content, timeline or authorship signal. The Q1 sweep was run on all 20 personas of SeqBench-20, once per ablated tool (Figure 9), holding model, embeddings and budget fixed. `get_thread` gains two merges and loses none ($+0.100$, $[-0.044, +0.244]$, $p = 0.163$), and the one

correct beyond-seed handle the all-tools agent claims against none without `get_handle_posts` is $+0.050$, $[-0.055, +0.155]$ ($p = 0.330$). The exact two-sided sign test, a secondary reading over the same 20 tests, has smallest raw p 0.0625, Holm-adjusted 1.000.

The grid is one run per cell at temperature 0, so the readable quantity is the within-grid delta rather than the level. The prompt also suppresses two of the tools it ablates, telling the agent that `style_similarity` and `nearest_authors` are unreliable on this corpus and must not carry a merge on their own, so their null deltas are in part an instruction rather than a measurement. Without `get_handle_posts`, tool calls rise from 18.1 to 26.1.

G THE AUDIT ACROSS CORPORA

Figure 7 repeats the Q1 audit on three of the six seed corpora of SeqBench-6seed. Across the four corpora the agent ran on, two lifts are positive and two negative (exact one-sided sign test $p = 0.6875$, secondary). Seed 45's significant deficit comes from claim volume: the agent merges one persona there, and the other 18 multi-account personas score 0 against random's small expected gain each. At the agent's own volume a random null expects 1.8 of the pooled merges against the agent's six (declared t on the per-persona merge indicator, $p = 0.016$). Secondary readings: the exact Poisson-binomial test of the pooled merge count gives $p = 0.433$ against random and $p = 3{\times}10^{-4}$ at matched volume, and the exact sign-flip test over the four corpora gives $p = 0.875$ (smallest attainable 0.125).

The prefix floor. On the five prefix personas of the $n{=}5$ run, F is random's expectation, above the vacuous share of one single-account persona in five; the agent's lift over it has $p = 0.30$ (4 degrees of freedom).

H THE CEILING, THE SCREEN AND A CORPUS WE DID NOT BUILD

Table 8: Ceilings and the pre-run screen at the declared $k = 2$. *first-20*: the personas the harness attacks. Verdict for the attacker returning the seed-adjacent admissible set.

corpus	pers./pop.	pairs	satisf.	F_0	F	ceiling adj/tr/pu	corridor $[t]$	first-20 corridor $[t]$	verdict	n random	n population [boot.]
rich-v9	18/20	124	0.589	0.056	0.095	0.726/0.781/0.781	0.631 $[0.432, 0.831]$	0.631 $[0.432, 0.831]$	resolvable	6	6 $[4, 9]$
SeqBench-20	20/20	121	0.083	0.100	0.137	0.178/0.216/0.227	0.041 $[-0.037, 0.120]$	0.041 $[-0.037, 0.120]$	unresolvable	—	115 $[23, \infty)$
SeqBench-1k	950/1,000	3,966	0.069	0.164	0.165	0.249/0.253/0.254	0.084 $[0.069, 0.099]$	0.126 $[-0.003, 0.254]$	capped†	48	49 $[40, 60]$
SeqBench-6seed s43	20/20	155	0.071	0.050	0.085	0.116/0.135/0.135	0.031 $[-0.037, 0.099]$	0.031 $[-0.037, 0.099]$	unresolvable	—	155 $[27, \infty)$
SeqBench-6seed s44	20/20	141	0.050	0.050	0.086	0.108/0.108/0.108	0.023 $[-0.082, 0.128]$	0.023 $[-0.082, 0.128]$	unresolvable	—	668 $[55, \infty)$
SeqBench-6seed s45	20/20	164	0.067	0.050	0.083	0.163/0.172/0.197	0.080 $[-0.024, 0.184]$	0.080 $[-0.024, 0.184]$	unresolvable	—	47 $[17, \infty)$
SeqBench-6seed s46	20/20	106	0.038	0.000	0.041	0.058/0.058/0.058	0.017 $[-0.053, 0.087]$	0.017 $[-0.053, 0.087]$	unresolvable	—	570 $[32, \infty)$
SeqBench-6seed s47	20/20	147	0.095	0.100	0.134	0.167/0.171/0.171	0.034 $[-0.050, 0.117]$	0.034 $[-0.050, 0.117]$	unresolvable	—	189 $[32, \infty)$
SeqBench-6seed s48	19/20	123	0.081	0.053	0.091	0.103/0.115/0.115	0.013 $[-0.047, 0.072]$	0.013 $[-0.047, 0.072]$	unresolvable	—	715 $[36, \infty)$
SeqBench-n50	50/50	487	0.053	0.040	0.052	0.114/0.133/0.133	0.062 $[0.011, 0.112]$	0.124 $[0.011, 0.238]$	resolvable	41	54 $[25, 243]$
SeqBench-n150	149/150	1,259	0.038	0.007	0.011	0.084/0.089/0.112	0.073 $[0.042, 0.104]$	0.076 $[-0.019, 0.170]$	underpowered	42	44 $[30, 68]$
SeqBench-n200	198/200	1,789	0.037	0.005	0.008	0.061/0.073/0.079	0.053 $[0.030, 0.075]$	0.027 $[-0.036, 0.089]$	underpowered	55	60 $[40, 102]$
SeqBench-n250	250/250	2,180	0.032	0.008	0.011	0.063/0.072/0.077	0.052 $[0.031, 0.073]$	0.030 $[-0.020, 0.080]$	underpowered	60	63 $[45, 106]$
SeqBench-n300	298/300	2,697	0.029	0.003	0.006	0.064/0.070/0.078	0.058 $[0.040, 0.076]$	0.061 $[-0.018, 0.140]$	underpowered	47	49 $[35, 67]$
SeqBench-n750	679/750	4,056	0.074	0.153	0.154	0.253/0.268/0.280	0.099 $[0.081, 0.117]$	0.138 $[0.040, 0.237]$	resolvable	36	36 $[30, 43]$
SeqBench-n800	721/800	4,187	0.073	0.150	0.151	0.251/0.261/0.277	0.100 $[0.083, 0.116]$	0.132 $[0.020, 0.245]$	resolvable	33	34 $[29, 40]$
SeqBench-n850	752/850	4,206	0.067	0.169	0.170	0.248/0.264/0.277	0.079 $[0.064, 0.093]$	0.233 $[0.111, 0.356]$	resolvable	39	40 $[34, 48]$
SeqBench-n900	789/900	4,182	0.078	0.163	0.164	0.257/0.273/0.285	0.092 $[0.077, 0.108]$	0.178 $[0.045, 0.311]$	resolvable	38	38 $[32, 45]$
SeqBench-n950	837/950	4,296	0.081	0.176	0.177	0.270/0.284/0.298	0.093 $[0.078, 0.109]$	0.093 $[-0.007, 0.193]$	capped†	38	38 $[31, 45]$
SeqBench-n1000	875/1,000	4,357	0.083	0.200	0.201	0.303/0.317/0.330	0.102 $[0.086, 0.118]$	0.219 $[0.083, 0.354]$	resolvable	36	36 $[31, 42]$

†Resolvable only for an attacker capping its lift; the admissible set itself just misses. Released corpora: decoys dropped, and a cumulative leak counted only where all its parts lie in one account. n *random*: random samples of this corpus reaching power 0.8; n *population*: a plug-in population like it, with the persona bootstrap's 95% range, a secondary reading. Every interval is the declared paired t of the admissible set's lift over F.

The screen across cells. Over the 540 screened cells (20 corpora; three readings, each at $k = 1, 2, 3$ and at $k = 2$ under six rarity thresholds) the verdicts are 168 dead, 137 resolvable, 122 unresolvable and 113 underpowered. The attacker returning the whole admissible set is not the most detectable obedient attacker in 266 cells, and 14 cells resolve only for an attacker capping its lift. The persona bootstrap, a secondary reading, disagrees with the declared t on clearing zero in 21 cells.

The most detectable obedient attacker. The declared t is $\bar{d}/(s_d/\sqrt{n})$, so an attacker can raise it by giving up per-persona lift where the lift is uneven. The screen therefore relaxes the obedient attacker from the admissible set to the continuous box $0 \le d_p \le d_p^{\max}$ of per-persona lifts the policy permits and sweeps a single cap c, taking $d_p = \min(d_p^{\max}, c)$, which is a one-dimensional sweep rather than a search over the box. The maximising attacker can therefore have lower recall than the admissible set: a cell labelled "capped" is detectable only by an attacker that forgoes correct merges, which is why the tables mark it and the verdict for the admissible set is printed beside it. Because the relaxation only enlarges the set of obedient attackers, an "unresolvable" verdict is unaffected by it.

The module. It counts every unordered pair of a persona's accounts and admits it on k shared labelled attributes under the reading's comparison set. The independence reading requires the k attributes to arrive by k distinct inference chains; on SeqBench-20 it leaves every ceiling unchanged at $k = 1, 2$ and lowers $k = 3$ from 0.140 to 0.134. Under a rarity threshold, attributes of one family (place, work, age) that disclose one value count once. Ceilings on SeqBench-1k are over its 950 posting personas. Table 9 states what Proposition 1 proves under each reading.

SynthPAI. Figure 11 gives the three granularities. A comment's evidence is the set of distinct attributes its human-reviewed labels mark as recoverable from it; 55.8% of the $7{,}823$ comments carry none. The profile-truth attacker keys each labelled attribute on the author's profile value, ground truth the text does not carry, and so sits on the seed-adjacent ceiling by construction; the label attackers key on the value the comment's human label reads off the text, as exact strings or by the deterministic string rules of SynthPAI's own scorer. The halves row averages 200 random splits of each author's threads (95% of splits in $[0.367, 0.423]$); moving from comment to half-author accounts multiplies the admitted share by 27.3 and the seed-adjacent ceiling by 57.0, so the task, not the corpus, fixes the account granularity.

Table 9: What Proposition 1 states, by reading of the declared policy. Rows S1–S10 are proved or checked on every corpus; the last row is measured on the agent's runs.

	statement	**seed-adjacent**	**transitive**	**profile-union**
S1	Bound: an attacker every merge of which meets the reading's condition through accounts of p it has itself merged, none false, scores at most $\overline{\mu}_\rho$	holds	holds	holds
S2	Attainment: returning exactly $\mathcal{M}^\rho_\pi$ meets the hypothesis and attains $\overline{\mu}_\rho$	holds	holds	holds
S3	The pairwise rule (a merge admitted when π holds of that pair's evidence) implies the bound (on SeqBench-20 at k=2: 0.178, 0.216, 0.227)	fails: chained pairwise merges reach accounts outside $\mathcal{M}^{\mathrm{adj}}_\pi$	holds: pairwise merges chained from the seed are the transitive reading	holds, not tight: a pairwise attacker reaches at most $\mathcal{M}^{\mathrm{tr}}_\pi$
S4	An attacker chaining through a false merge (another persona's account) is covered	not covered	not covered	not covered
S5	Nesting: $\overline{\mu}_{\mathrm{adj}} \le \overline{\mu}_{\mathrm{tr}} \le \overline{\mu}_{\mathrm{pu}}$ (checked on 20 corpora, 0 violating)	holds	holds	holds
S6	Non-increasing in the threshold k (checked on 20 corpora, 0 violating)	holds	holds	holds
S7	Requiring k distinct inference chains only tightens (checked on 20 corpora, 0 violating)	holds	holds	holds
S8	A rarity threshold τ (one value counted once) only tightens, further as τ falls (checked on 20 corpora, 0 violating)	holds	holds	holds
S9	Dropping decoys, and counting a cumulative leak only where all its parts lie, only tightens (checked on 20 corpora, 0 violating)	holds	holds	holds; the merged profile can complete a leak split across two merged accounts, which the other readings never count
S10	The labelled-evidence ceiling is at most the ceiling of an attacker also using evidence the key does not record	holds if every label is genuine	holds if every label is genuine	holds if every label is genuine
—	Measured: the ceiling bounds the audited agent at k=2	fails: premise broken on 4 of 9 distinct correct merges; 4 lie outside $\mathcal{M}^\rho_\pi$; 4 persona-runs exceed their own ceiling	fails: premise broken on 4 of 9 distinct correct merges; 3 lie outside $\mathcal{M}^\rho_\pi$; 3 persona-runs exceed their own ceiling	fails: premise broken on 4 of 9 distinct correct merges; 3 lie outside $\mathcal{M}^\rho_\pi$; 3 persona-runs exceed their own ceiling

I COMPARISON ATTACKERS

Selection and harness. The six published systems of Section 4 were selected for having released code or a faithful specification and for fitting our seeded task; LUAR (Rivera-Soto et al., 2021) runs as MUD (primary) and CRUD (secondary). A general-purpose sentence embedder, MiniLM (Reimers & Gurevych, 2019), is carried through the same rules as a reference point for the style embedders StyleDistance and CISR (Table 12). Each linker ranks or merges non-seed handles for a seed handle and is scored through the same harness and seed union as the agent, beside its own floors: the seed-only null, random's expectation, merge-everything, and a random null merging per persona as many handles as the attacker did.

Operating points and folds. A linker is thresholded at two points tuned on a persona-disjoint fold: *f1opt*, the complement-F_1 optimum, which is the headline on every corpus, and *p50*, printed beside it. That rule, like the 32-token post encoding, was fixed after the CPU results at both points had been seen. SeqBench-n50, SeqBench-n150 and the SeqBench-n1000 sample are tuned on SeqBench-n300 personas 150–299; SeqBench-20 and seeds 43–45 on seeds 46–48 and the reverse. A row is flagged (♭) where it merges more than twice the true other handles and its identification or linkage lift over the seed-only null sits entirely below zero; its identification clause was added after results had been seen, and Figure 6 carries the same marker. The SeqBench-n1000 rows (*) are a 150-persona sample of personas 300–999, the sparse half of that corpus, against their own 1358 handles.

Table 10: LUAR-MUD on the released corpora, SeqBench-20 and SeqBench-6seed: rank quality, complement F_1 against the handle-volume random null, and linkage F_1 against the seed-only null.

corpus	pers.	RR (hash.)	op	compl. F_1	floor	lift $[t]$	p	merges (corr.)	ident.	linkage lift vs seed $[t]$	interp.
SeqBench-n50	50	0.950 (0.630)	f1opt	0.634	0.012	$+0.622$ $[+0.528, +0.716]$	1×10^{-17}	156 (107)	0.86	$+0.293$ $[+0.197, +0.390]$	yes
			p50	0.620	0.012	$+0.609$ $[+0.511, +0.706]$	1×10^{-16}	147 (104)	0.88	$+0.290$ $[+0.193, +0.388]$	yes
SeqBench-n150	149	0.854 (0.609)	f1opt	0.560	0.004	$+0.556$ $[+0.496, +0.615]$	3×10^{-40}	566 (291)	0.77	$+0.196$ $[+0.141, +0.251]$	yes
			p50	0.563	0.004	$+0.559$ $[+0.499, +0.620]$	1×10^{-39}	536 (287)	0.78	$+0.203$ $[+0.148, +0.258]$	yes
SeqBench-n1000*	150	0.429 (0.279)	f1opt	0.123	0.000	$+0.123$ $[+0.066, +0.180]$	4×10^{-5}	455 (37)	0.66	-0.167 $[-0.225, -0.109]$	**no**
			p50	0.120	0.000	$+0.119$ $[+0.062, +0.176]$	7×10^{-5}	420 (34)	0.67	-0.163 $[-0.221, -0.106]$	yes
SeqBench-20	20	0.251 (0.077)	f1opt	0.121	0.034	$+0.086$ $[-0.006, +0.179]$	0.066	80 (8)	0.20	-0.267 $[-0.437, -0.096]$	yes
SeqBench-6seed s43	20	0.213 (0.143)	f1opt	0.112	0.042	$+0.069$ $[-0.020, +0.158]$	0.119	117 (11)	0.15	-0.281 $[-0.426, -0.137]$	yes
SeqBench-6seed s44	20	0.327 (0.151)	f1opt	0.140	0.027	$+0.112$ $[+0.019, +0.206]$	0.021	76 (8)	0.20	-0.182 $[-0.314, -0.049]$	yes
SeqBench-6seed s45	20	0.339 (0.069)	f1opt	0.140	0.044	$+0.096$ $[+0.003, +0.189]$	0.044	121 (14)	0.30	-0.195 $[-0.325, -0.065]$	yes
SeqBench-6seed s46	20	0.189 (0.126)	f1opt	0.063	0.038	$+0.025$ $[-0.039, +0.089]$	0.421	84 (5)	0.10	-0.244 $[-0.354, -0.134]$	yes
SeqBench-6seed s47	20	0.290 (0.145)	f1opt	0.153	0.036	$+0.117$ $[+0.009, +0.225]$	0.035	81 (10)	0.25	-0.197 $[-0.371, -0.023]$	yes
SeqBench-6seed s48	19	0.269 (0.118)	f1opt	0.148	0.046	$+0.101$ $[+0.022, +0.181]$	0.015	125 (12)	0.21	-0.258 $[-0.408, -0.108]$	**no**

Reading LUAR. Against the agent, LUAR separates in 2 of 5 pairings, of which 4 are against the all-tools agent and one against an ablated cell; those 4 are reported uncorrected, and none survives Holm over them (smallest adjusted $p = 0.122$), so Section 6 counts them as a family rather than as a result. On the SeqBench-n1000 sample, reciprocal rank is 0.86 for seed handles with at least 48 posts ($n = 26$), 0.52 for 16–47 and 0.19 under 16 ($n = 57$), and over the full SeqBench-n1000 pool personas 0–299 (median 44 posts per handle) reach 0.846 against 0.535 for personas 300–999 (median 9). Masking every @-mention on SeqBench-n50 leaves it intact (reciprocal rank 0.961 against 0.950; complement F_1 0.675 against 0.634). The best offline ranker of our baseline set is hashed bag of words at its own f1opt point. On the held-out SeqBench-n1000 sample the 32-token encoding ranks best (reciprocal rank 0.429 against 0.319 and 0.256), while on SeqBench-20 and SeqBench-6seed the chunked encoding ranks higher on 5 of 7 for MUD.

Table 11: LUAR variants at f1opt: complement F_1 and seed-ranking reciprocal rank.

	SeqBench-n50		SeqBench-n150		SeqBench-n1000*	
variant	compl. F_1	RR	compl. F_1	RR	compl. F_1	RR
LUAR-MUD posts32	0.634	0.950	0.560	0.854	0.123♭	0.429
LUAR-CRUD posts32	0.600	0.934	0.532	0.841	0.122♭	0.424
LUAR-MUD first16	0.391	0.866	0.375	0.756	0.077	0.319
LUAR-CRUD first16	0.420	0.803	0.350	0.733	0.057♭	0.319
LUAR-MUD chunked	0.368	0.701	0.233	0.567	0.035♭	0.256
LUAR-CRUD chunked	0.319♭	0.677	0.236	0.564	—	—

The classical rung of ESRC. The rarity-weighted matcher of Lermen et al. (2026) (after Narayanan & Shmatikov (2008)) over topic or word features with temporal or Jaccard weighting accepts nothing at its calibrated 90%, 98% and 99% precision thresholds on any corpus, as in its

Table 12: Style-only embedders and MiniLM on the released corpora: complement F_1, lift over the handle-volume random null (declared t) and its p. *seed+8*: the fixed retrieval rule.

		StyleDistance			CISR			MiniLM		
corpus	**op**	cF_1	lift	p	cF_1	lift	p	cF_1	lift	p
SeqBench-n50	f1opt	$0.124^{\flat}$	$+0.107$	2×10^{-6}	$0.088^{\flat}$	$+0.071$	1×10^{-4}	0.102	$+0.087$	0.001
	p50	0.022	$+0.021$	0.094	0.037	$+0.036$	0.163	0.000	0.000	1.000
	seed+8	0.160	$+0.139$	2×10^{-7}	0.147	$+0.126$	2×10^{-5}	0.078	$+0.057$	0.005
SeqBench-n150	f1opt	$0.087^{\flat}$	$+0.080$	4×10^{-8}	$0.063^{\flat}$	$+0.056$	4×10^{-8}	0.049	$+0.047$	4×10^{-5}
	p50	0.020	$+0.019$	0.042	0.013	$+0.013$	0.098	0.000	0.000	1.000
	seed+8	0.120	$+0.114$	7×10^{-16}	0.111	$+0.105$	6×10^{-16}	0.052	$+0.046$	4×10^{-6}
SeqBench-n1000*	f1opt	$0.018^{\flat}$	$+0.016$	8×10^{-4}	$0.011^{\flat}$	$+0.009$	8×10^{-4}	$0.019^{\flat}$	$+0.018$	0.012
	p50	0.000	0.000	0.093	0.000	0.000	0.063	0.000	0.000	1.000
	seed+8	0.029	$+0.028$	2×10^{-4}	0.033	$+0.032$	0.004	0.023	$+0.022$	0.004

source's own table; Table 13 gives its top-1 rung, where every round-one guess is accepted. A row that merges nothing scores exactly the seed-only null's linkage.

Table 13: ESRC's classical rung at top-1: complement F_1, correct merges, the handle-volume random floor, the lift under the declared t, and identification.

corpus	**features, weighting**	**cF_1**	**correct**	**floor**	**lift** $[t]$	p	**ident.**
SeqBench-n50	topics, temporal	0.0000	0/50	0.0063	-0.006 $[-0.006, -0.006]$	$2\times10^{-60\ddagger}$	0.22
	topics, Jaccard	0.0025	1/50	0.0063	-0.004 $[-0.009, +0.001]$	0.134	0.42
	words, temporal	0.0228	2/50	0.0063	$+0.016$ $[-0.016, +0.049]$	0.310	0.04
	words, Jaccard	0.0135	1/50	0.0063	$+0.007$ $[-0.020, +0.034]$	0.599	0.30
SeqBench-n150	topics, temporal	0.0000	0/149	0.0021	-0.002 $[-0.002, -0.002]$	$1\times10^{-145\ddagger}$	0.23
	topics, Jaccard	0.0000	0/149	0.0021	-0.002 $[-0.002, -0.002]$	$1\times10^{-145\ddagger}$	0.43
	words, temporal	0.0154	3/149	0.0021	$+0.013$ $[-0.004, +0.031]$	0.136	0.03
	words, Jaccard	0.0034	1/149	0.0021	$+0.001$ $[-0.005, +0.008]$	0.696	0.26
SeqBench-n1000*	topics, temporal	0.0058	1/150	0.0009	$+0.005$ $[-0.007, +0.016]$	0.396	0.03
	topics, Jaccard	0.0073	1/150	0.0009	$+0.006$ $[-0.008, +0.021]$	0.379	0.41
	words, temporal	0.0000	0/150	0.0009	-0.001 $[-0.001, -0.001]$	$4\times10^{-81\ddagger}$	0.00
	words, Jaccard	0.0511	8/150	0.0009	$+0.050$ $[+0.011, +0.089]$	0.012	0.30

‡The rung recovers nothing, so the p measures claim volume, not accuracy (Section 3).

The language-model rows. For ESRC's search, select and sort rungs, AutoProfiler and InferLink's OpenHands agent we computed the floors on the CPU (Tables 14 and 15). AutoProfiler's floors use the seven attribute columns its loader's truth populates (*overlap7*).

Beside each floor, Figure 6 prints the result the system's authors report on their own data, which is not comparable to our floors: backbone, harness and population all differ. ESRC's sort rung reaches 0.673 recall at 90% precision on a temporally split Reddit set of $5{,}000$ queries against $10{,}000$ candidates, where the classical rung reaches 0.004 (Lermen et al., 2026, Table 6); AutoProfiler on a GPT-4 backbone scores 0.756 to 0.954 across SynthPAI's eight attributes (Du et al., 2026, Table 5); and InferLink's GPT-5 agent links 0.792 of Netflix Prize identities in pools of $1{,}000$ from two overlapping ratings, against 0.560 to 0.602 for the classical matcher (Ko et al., 2026, Table 1).

Table 14: AutoProfiler's floors on the loader's truth: overlap7 attribute accuracy of the leave-one-out modal profile, its top-3 and abstention; the 25-attribute modal floor; identification floors.

corpus	**pers.**	**modal**	**top-3**	**abstain**	**attr25 (free)**	**ident. modal**	**chance**	**truth ceiling**
SeqBench-n50	50/50	0.380	0.729	0.000	0.621 (0.400)	0.0000	0.0200	0.847
SeqBench-n150	149/150	0.428	0.736	0.000	0.654 (0.400)	0.0067	0.0067	0.804
SeqBench-n1000	875/1000	0.403	0.726	0.000	0.643 (0.400)	0.0009	0.0010	0.379

attr25 (free): the 25-attribute modal floor and, in parentheses, the share of it that constant-truth attributes give free.

The agent at its own claim volume, and pair AUC. Pooled over the 80 personas of the four corpora the agent ran on (75 multi-account), Table 16 sets the agent against each null and against rankers truncated to the agent's own handle volume. Table 17 gives the pair-level AUC of each ranking signal on SeqBench-20 and SeqBench-6seed (7 corpora, 411 positive pairs) and the released family (708): only the LinkedIn–WhatsApp pairs, where the generator writes a real name into both handles, separate there, and they number 14.

Table 15: InferLink/OpenHands floors: linkage F_1, cross-account recall, complement F_1 and attribute accuracy of each null, and the cross-account ceiling at $k = 1$ and $k = 2$.

corpus	null	linkage F_1	cross	compl. F_1	attr.
SeqBench-n50, all 50 seeds	seed-only null	0.420	0.040	0.0000	0.661
	$\mathbb{E}$[`random`]	0.258	0.052	0.0132	0.629
	`Random(42)` draw	0.267	0.052	0.0126	0.626
	merge everything	0.039	1.000	0.0311	0.220
	ceiling k=1 / k=2 (cross-acct.)		0.885 / 0.640		
SeqBench-n150, 50 sampled seeds	seed-only null	0.489	0.020	0.0000	0.668
	$\mathbb{E}$[`random`]	0.294	0.024	0.0042	0.626
	`Random(42)` draw	0.300	0.020	0.0000	0.614
	merge everything	0.014	1.000	0.0104	0.220
	ceiling k=1 / k=2 (cross-acct.)		0.658 / 0.457		

Table 16: The agent against nulls and rankers at its own handle volume, pooled over the four corpora the agent ran on; declared paired t and its p.

agent minus	linkage F_1 $[t]$	p	cross-acct. $[t]$	p	compl. F_1 $[t]$	p
random at the agent's handle volume	$+0.024$ $[+0.004, +0.044]$	0.020	$+0.032$ $[+0.002, +0.063]$	0.038	$+0.044$ $[+0.006, +0.082]$	0.023
seed-only null	$+0.020$ $[+0.003, +0.038]$	0.020	$+0.033$ $[+0.002, +0.064]$	0.037	$+0.046$ $[+0.007, +0.085]$	0.022
$\mathbb{E}$[`random`] (three handles)	$+0.234$ $[+0.189, +0.279]$	3×10^{-16}	-0.002 $[-0.033, +0.029]$	0.908	$+0.012$ $[-0.026, +0.051]$	0.531
merge everything	$+0.441$ $[+0.383, +0.498]$	3×10^{-25}	-0.904 $[-0.965, -0.843]$	2×10^{-44}	-0.027 $[-0.067, +0.013]$	0.177
BM25 ranker at the agent's volume	$+0.004$ $[-0.004, +0.011]$	0.332	$+0.004$ $[-0.003, +0.010]$	0.296	$+0.006$ $[-0.005, +0.017]$	0.303
MiniLM ranker at the agent's volume	$+0.024$ $[+0.002, +0.045]$	0.029	$+0.031$ $[0.000, +0.062]$	0.051	$+0.041$ $[+0.003, +0.079]$	0.035
hashed bag of words at the agent's volume	$+0.030$ $[+0.005, +0.054]$	0.019	$+0.033$ $[+0.002, +0.064]$	0.037	$+0.046$ $[+0.007, +0.085]$	0.022
trained linker at the agent's volume	$+0.022$ $[+0.004, +0.040]$	0.020	$+0.033$ $[+0.002, +0.064]$	0.037	$+0.046$ $[+0.007, +0.085]$	0.022

Table 17: Pair-level AUC of each ranking signal, pooled within corpus: all pairs and the LinkedIn–WhatsApp pairs, under the synthesized and the corpus handle policy.

	SeqBench-20 and SeqBench-6seed				released family			
	synthesized		corpus		synthesized		corpus	
signal	all	Li–Wa	all	Li–Wa	all	Li–Wa	all	Li–Wa
handle-string similarity	0.493	0.515	0.518	0.978	0.494	0.498	0.529	0.978
style cosine (82-d)	0.536	0.458	0.536	0.458	0.517	0.545	0.518	0.546
style, within-seed z	0.503	0.306	0.503	0.306	0.489	0.553	0.489	0.550
BM25	0.518	0.416	0.517	0.416	0.593	0.630	0.592	0.620
MiniLM cosine	0.529	0.540	0.530	0.540	0.578	0.631	0.578	0.635
hashed bag of words	0.536	0.599	0.537	0.599	0.715	0.771	0.716	0.775
trained linker (out of fold)	0.582	0.379	0.611	0.976	0.591	0.539	0.644	0.979

J REPRODUCIBILITY DETAILS

Table 18 gives each audited corpus's generator bracket, labeller and hash. The corpus hash carried by an agent report covers the posts table and persona ground truth, so it cannot by itself detect an empty-modality condition.

Table 18: Provenance of every audited corpus.

corpus	schema	seed	days cfg/obs	pers. meta/posting	posts	accts	labels (posts)	labels by	generator	labeller	hash
SeqBench-20	2.10.0	42	7/6.4	20/20	293	73	158 (112)	detector	`59de9f9..592eda3`*	`592eda3`	`e434487016aa`
SeqBench-6seed s43	2.10.0	43	7/6.8	20/20	317	82	162 (133)	detector	`59de9f9..016da79`	`592eda3..016da79`	`5a50d934f7a9`
SeqBench-6seed s44	2.10.0	44	7/6.9	20/20	325	80	159 (115)	detector	`59de9f9..016da79`	`592eda3..016da79`	`f8bb939a319f`
SeqBench-6seed s45	2.10.0	45	7/6.7	20/20	344	87	178 (129)	detector	`59de9f9..016da79`	`592eda3..016da79`	`edaa11ce1ff2`
SeqBench-6seed s46	2.10.0	46	7/6.7	20/20	296	73	128 (92)	detector	`59de9f9..016da79`	`592eda3..016da79`	`5103b8fd96a8`
SeqBench-6seed s47	2.10.0	47	7/6.9	20/20	357	80	206 (137)	detector	`59de9f9..016da79`	`592eda3..016da79`	`d737ddda839c`
SeqBench-6seed s48	2.10.0	48	7/6.7	20/19	317	75	170 (110)	detector	`59de9f9..016da79`	`592eda3..016da79`	`e4a98b2b5191`
rich-v9	2.10.0	42	7/6.4	20/18	250	72	633 (250)	in-run	`59de9f9..e20b3a0`*	—	`c6c31fc52c6b`
SeqBench-1k	2.10.0	42	30/7.9	1,000/950	15,785	2,924	6,999 (5,793)	detector	`59de9f9..674b349`	`592eda3..674b349`	`6ccaa57120a4`

SeqBench-1k/150 is re-derived from its definition (150 personas, 3,083 posts, 858 targets). Hash: sha256 over `drafts.parquet` then `personas.json`, as every manifest records it. All audited corpora are generator schema 2.10.0, which the generator repository carried from `59de9f9` (2026-06-15) until `38346e4` (2026-08-20); none records a generator commit. *generator*: the generator-repository commits the generating code must lie between, from the commit that introduced schema 2.10.0 up to the last commit before `drafts.parquet` was written; a dirty tree cannot be excluded. *labeller*: for a detector corpus, from the first commit of `scripts/label_leaks.py` to the same upper bound. That labeller read the generator's drafts and overwrote `leak_intent` before writing `drafts.parquet`, so its commit and the file's time bound the labelling run, not the generation, which came first and is bounded below only by the schema commit. *A byte-identical `drafts.parquet` is committed in the generator repository's history. *labels*: every label of a *detector* corpus carries the surface form that only `scripts/label_leaks.py` writes; no in-run label survives. The 11 released companion corpora are schema 2.12.0, seed 17, 14 days, each with a generator commit.

Model seats. Table 19 lists the seats. The agent's reports do not record the thinking mode, which the backend disables unless an environment variable enables it.

K ETHICS: EXTENDED DISCUSSION

The dual-use balance the ethics statement records has one component specific to the benchmark family: its controllable ground truth lets an anonymiser be scored against the precise signal it must remove, and a floor lets a reader see how much of a reported attack is inference at all.

L FLOORS AND LIFT, CONTINUED

Normalised lift. The same margin means one thing over a floor near zero and another near one, so where floors differ across the conditions compared we report $L/(1-F)$ beside L, the share of headroom taken. At $F = 1$ there is no headroom at all, which R3 of Box 1 turns into a reporting rule (Remark 1 is such a case).

Proposition 2 (The population floor is a bound, and it is attained)**.** *Let A be any attacker whose only input is the target's equivalence class $C_Q(p)$, and assume within-class exchangeability: conditional on that class, every member of it is equally likely to be the target. Then A's expected identification rate is at most $F_{\mathrm{id}}(Q)$, and the attacker guessing uniformly inside the class attains it.*

The proof conditions on the target's class and averages (Appendix A). Exchangeability is load-bearing (Appendix D).

Limits of a floor. Clearing a floor is necessary for a reported number to be evidence of inference and nowhere near sufficient: an attack can clear every floor here and still be exploiting an artefact of the corpus, so Section 6 pairs the floors with channel ablations. Floor and ceiling alike are defined against a scoring rule, so changing the rule (widening the attribute vocabulary, crediting abstention) moves them and the score together, in the same direction but not by the same amount.

M THE AUDITED AGENT

Architecture and protocol. Baselines are given the same seed as the agent, and after each tool call the agent is reminded to submit as soon as it can name the author.

Table 19: Model seats behind the reported numbers, as their reports record them; dates are UTC days a report was written; checkpoint revisions give twelve characters, commits seven.

seat	identifier	record
Audited agent, Q1 grid on SeqBench-20 (Table 1)	`deepseek-v4-pro`	DeepSeek API; report written 2026-08-14
Agent, SeqBench-6seed seeds 43–45	`deepseek-v4-pro`	reports written 2026-08-16
Agent, Q2 on SeqBench-20	`deepseek-v4-pro`	report written 2026-08-16
Agent, n=5 prefix of SeqBench-20	`deepseek-v4-pro`	report written 2026-06-28
Agent embeddings	`voyage-3-large`	Voyage API; named in each of the four agent seats' reports
Baselines of Table 1	none	random, retrieval, stylometry, majority class (retrieval on `hashing-256`); report written 2026-07-14
Offline rankers (Table 16)	`sentence-transformers/all-MiniLM-L6-v2`	revision not recorded; plus `hashing-256`; scored 2026-09-25
LUAR (Rivera-Soto et al., 2021)	`rrivera1849/LUAR-MUD` @`f1db50251805`; `rrivera1849/LUAR-CRUD` @`8b36ce7f289e`	local CPU, weights hashed; scored 2026-09-25
Style embedders (Table 12)	`StyleDistance/styledistance` @`b7df5f0b0480`; `AnnaWegmann/Style-Embedding` @`d7d0f5ca8293`; `sentence-transformers/all-MiniLM-L6-v2` @`1110a243fdf4`	local CPU; scored 2026-09-25
SynthPAI attackers (Yukhymenko et al., 2024)	the release's 18 model names	saved predictions at commit `ba9edc9`
Leak labeller, detector corpora	not recorded	`label_leaks.py` defaults to `deepseek-chat`, overridable by environment
Generator, audited corpora	not recorded	metadata name an OpenAI-compatible backend only
Generator, released corpora	not recorded	the same backend and build revisions `eaafaed` and `edc41c3` only

The technique. Q1 is designed as a search across accounts. The agent reads the handle off the seed post and pulls that handle's remaining posts (`get_handle_posts`, the output path for linked accounts), and from the history it holds it infers the attribute assignment it searches on. It queries the other accounts and platforms for handles whose content, timing and authorship signal stay consistent with that history (`semantic_search`, `keyword_search`, `style_similarity`, `nearest_authors`, `get_thread`), and commits the ones it accepts through the submission tool. On Q2 the agent declares no confident match where it accepts nobody.

N THE PREFIX EFFECT

A five-persona prefix is an easier task rather than a small version of a twenty-persona one, the candidate set being a quarter of the size, and every metric on it moves in the flattering direction at once (cross-account recall $0.385 \to 0.121$, linkage F_1 $0.649 \to 0.581$, attribute accuracy $0.300 \to 0.202$), which is what a population effect looks like. It is not identified as one here: the prefix run was scored by a different scorer version, so population and scorer version change together between the two runs and their contributions cannot be separated. Section 5 measures population dependence where it can be isolated, on a census table.