PERSONALIZATION DONE RIGHT: BALANCING PERSONALIZATION AND PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized language models typically condition on user-specific context to better serve individual needs. However, the very data that enables effective personalization, such as user profiles, interaction histories, and preference signals, simultaneously increases privacy risk. Despite the growing importance of personalization and privacy in language modeling, their joint analysis remains limited. In this work, we study how personalization granularity affects task utility and user identifiability. We formalize a setting in which a user issues a prompt and the model has access to a user profile of varying granularity, and we characterize identity leakage through posterior identification gain. We then show that idealized group averaging reduces identity leakage under stated conditions, with utility change bounded by within-group utility differences. We study this tradeoff across a spectrum of personalization techniques, including retrieval-augmented generation, persona-based prompting, profile-augmented prompting, and provide extensive empirical findings that support its practical implications. Our findings have significant implications for the development of personalized language models that respect user privacy.

1 INTRODUCTION

Personalized AI systems face a fundamental dilemma: user context can improve responses while exposing information about the user. Consider the spectrum of personalization strategies available to an AI model. At one extreme, a fully generic model ignores the user entirely and thus responds identically to "What should I eat for dinner?" regardless of whether the user is a diabetic athlete in Seattle or a college student in West Lafayette. At the other extreme, a fully personalized model conditions every token on a rich user profile: dietary restrictions, past conversations, location, health records, behavioral patterns, among many others. The first system cannot adapt to individual needs. The second can provide tailored responses while exposing identity cues. This paper explores the personalization–privacy tradeoff between these extremes.

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language tasks (Grattafiori & et. al., 2024; Jiang et al., 2023). However, these models produce outputs that reflect aggregate training distributions rather than the preferences, style, or needs of any individual user. Personalization addresses this gap by conditioning model outputs on user-specific context, including user profiles, interaction histories, and explicit preference signals (Salemi et al., 2024b; Kumar et al., 2024; Zhang et al., 2025b). The importance of personalization has grown rapidly. Recent work has formalized its foundations, proposed taxonomies of personalization granularity (user-level, persona-level, and global preference), and introduced a range of techniques from retrieval-augmented generation to parameter-efficient fine-tuning (Xu et al., 2025; Zhang et al., 2025b; Mysore et al., 2024; Tan et al., 2024b; Tseng et al., 2024). However, an equally fundamental concern has received comparatively little formal attention: the privacy implications of personalization.

User-specific context can create a tension between personalization and privacy. A model that ignores user context entirely produces generic outputs that may be poorly suited to individual needs. A model that conditions on a rich user profile can generate highly relevant responses, but its outputs can carry information that helps identify the user. This tension is not merely an engineering challenge to be resolved through better anonymization or access control. For a fixed query, identity leakage arises when response distributions differ across users.

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

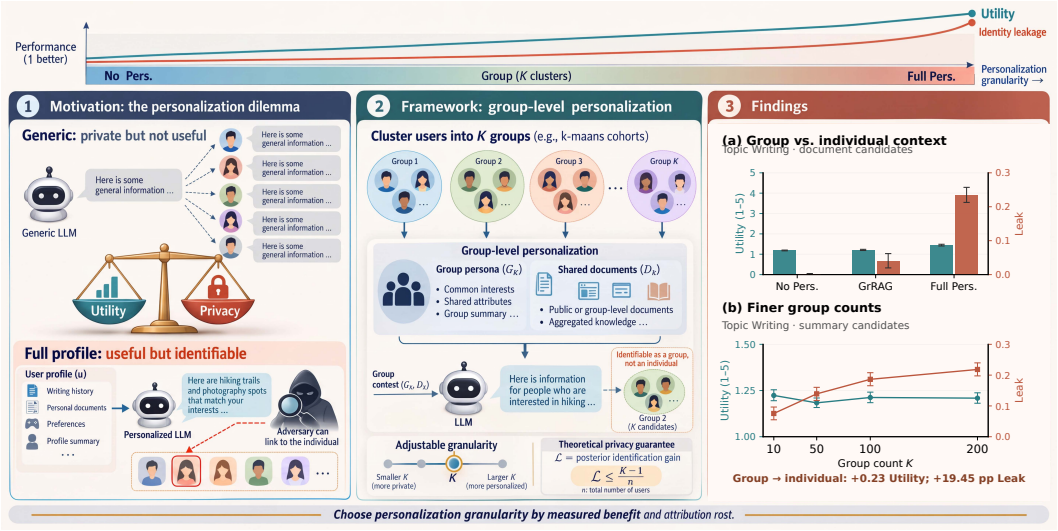

Figure 1: **Personalization and privacy through group context.** Left and center illustrate context choices and group construction; the top curves are conceptual. The right panels report Qwen3.5-9B on Topic Writing using separate-criterion Utility: (a) document-candidate attribution for No Pers., GrRAG, and Full Pers.; (b) summary-candidate attribution across group counts. Error bars show pointwise 95% user-bootstrap intervals on matched valid items (Appendix J.2). The two vertical axes have different units. The illustrated anonymity bound requires identical within-group response distributions and a uniform conditional prior.

Despite the rapid deployment of personalized LLMs, principled guidance for this tradeoff remains limited. Practitioners make ad hoc decisions about how much user context to include in prompts or fine-tuning data without understanding the structural consequences for privacy exposure. Over-personalization leaks sensitive attributes through model outputs. Under-personalization drives users toward manual workarounds, such as pasting personal context directly into prompts, that are often worse for privacy. The field lacks principled guidance for navigating this space.

In this work, we address this gap with an information-theoretic framework for personalization and privacy. We characterize identity leakage through posterior identification gain and show that idealized group averaging reduces it under stated conditions, with utility change bounded by within-group utility differences.

We prove that the mutual information between model outputs and user identity, conditioned on the query, provides an upper bound on expected posterior identification gain. This result connects the privacy cost of personalization to the identity information in model outputs.

We make the following contributions:

1. We introduce a novel theoretical framework to formalize the personalization–privacy tradeoff in LLMs. Our framework uses an information-theoretic measure of identity leakage and characterizes how group-level personalization provides k-anonymity under stated conditions.

2. We validate the key implications of our framework with a simple yet effective baseline inspired by our theory to control the personalization–privacy tradeoff. Experiments across generators and tasks demonstrate reduced user attribution relative to individual personalization while retaining useful personalization gains.

3. We show how personalization granularity affects this tradeoff: finer groups can increase user attribution without consistent gains in measured utility, while the benefits of individual context vary across tasks and metrics.

2 Related Work

Personalized Language Models. There has been a lot of interest recently in conditioning language models on user-specific context. Many of the techniques initially proposed for this problem were based on retrieval, where documents from the user's own history are placed directly into the prompt (Lewis et al., 2020; Robertson & Zaragoza, 2009; Salemi et al., 2024b). Salemi et al. (Salemi et al., 2024a) optimize the retriever against downstream generation quality, whereas Mysore et al. (Mysore et al., 2024) calibrate it to personalization utility. Richardson et al. (Richardson et al., 2023) replace retrieved documents with a generated profile summary, Li et al. (Li et al., 2023) generate that profile through a multi-stage writing procedure, and personas have been surveyed as a separate abstraction over users (Tseng et al., 2024). A second class of approaches encodes the user in the parameters through per-user adapters (Tan et al., 2024b), shared adapter pools (Tan et al., 2024a), factorized black-box heads (Zhuang et al., 2024), or group-level adaptation (Zhang et al., 2025a). Benchmarks and surveys organize this space by task, granularity, and technique (Salemi et al., 2024b; Kumar et al., 2024; Zhang et al., 2025b; Xu et al., 2025). Closest to our cohort strategies, ClusterRAG (Nkhata et al., 2026) retrieves from user clusters and reports cluster retrieval beating own-history retrieval on LaMP, but reports no privacy measurement. Salemi and Zamani (Salemi & Zamani, 2025) compare retrieval augmentation with parameter-efficient fine-tuning for privacy-preserving personalization, reporting a utility and a privacy number per mechanism; the comparison is made with the instruments this paper audits and without conditioning leakage on the query. We study these choices through query-conditioned identity leakage and their measured personalization benefits.

Privacy Risks of User Conditioning. Privacy risks of language models have been studied extensively, though largely independent of the personalization literature. One line extracts memorized training text (Carlini et al., 2021; 2023; Nasr et al., 2025). Another infers whether a record was used for training (Shokri et al., 2017; Carlini et al., 2022; Mattern et al., 2023; Shi et al., 2024), and Duan et al. (Duan et al., 2024) show these attacks are weak at pretraining scale. More recently, Staab et al. (Staab et al., 2024) show that capable models infer personal attributes directly from ordinary text, and Mireshghallah et al. (Mireshghallah et al., 2024) show that models reason poorly about when disclosure is appropriate. Retrieval moves the exposure to inference time. Huang et al. (Huang et al., 2023) and Zeng et al. (Zeng et al., 2024) show retrieved passages are reproduced in the output, and Wang et al. (Wang et al., 2025) show the same for persistent per-user memory. Defenses remain mechanism-level: differentially private training and fine-tuning (Dwork et al., 2006; Dwork & Roth, 2014; Abadi et al., 2016; McMahan et al., 2018; Li et al., 2022; Yu et al., 2022), private prompting (Duan et al., 2023), and syntactic anonymity criteria (Sweeney, 2002; Machanavajjhala et al., 2007). Closest to our attacker, Salemi and Zamani (Salemi & Zamani, 2026) re-identify users from personalized outputs with a forced-choice LLM attacker and report a query-only comparison; we adopt the attack shape, promote the controlled difference to the reported estimand leak, and gate the attacker on planted responses. Flemings et al. (Flemings et al., 2024) estimate the leakage attributable to augmented context at the context level; our estimand is output-attributable and adversary-specific. Brown et al. (Brown et al., 2022) argue that privacy for language models is contextual rather than a matter of scrubbing identifiers, and Staab et al. (Staab et al., 2025) show that anonymization must be judged against an inference adversary rather than a pattern matcher; our summary-holding attacker is that adversary for profile abstractions. Cohort-based personalization has a precedent in interest-based advertising, where post-mortems of Google's FLoC and of the Topics API found that cohort identifiers leak more than their nominal k-anonymity suggests (Berke & Calacci, 2022; Beugin & McDaniel, 2024). Our within-cohort attacker complements population-level attribution by restricting the candidate set to cohort peers. Finally, the instruments themselves have been questioned on each side separately: blind baselines beat published membership-inference attacks (Das et al., 2025), defense evaluations mislead when they average over examples (Aerni et al., 2024; Meeus et al., 2025), and LLM judges agree with one another without measuring the intended construct (Norman et al., 2026; Wu & Aji, 2025; Fabbri et al., 2021), with known-truth validation proposed for utility metrics alone (Eigler et al., 2026). None of this work relates the leakage it measures to the quantity personalization is optimizing, nor validates both instruments on the same outputs, which is what Section 4.1 and the audit of Section G do.

Interpreting Utility Differences and Personalized Preferences. Automatic metric differences need not imply meaningful differences to users. Lo et al. (2023) examine how metric-score differences

Table 1: Context granularity in the comparison. No Pers. is a separate baseline; one shared cohort still supplies context.

Setting	Context source	Cohorts
No Pers.	Bare query	—
Cohort	Shared persona / pool	K
Individual	Own summary / history	$\lvert\mathcal{U}\rvert$

relate to statistically significant human judgments in machine translation and show that fixed rules of thumb do not generalize reliably across translation directions and domains. Kocmi et al. (2024) relate metric-score differences to agreement with human system rankings, providing an empirical interpretation of metric scales rather than a universal threshold for meaningful improvement. These results motivate distinguishing reference overlap, style, and practical preference when interpreting the utility side of our comparisons; a small or nonsignificant metric difference alone does not establish equivalence. Personalized pairwise evaluation is also established: Dong et al. (2024) condition a judge on a user persona to predict a preference between responses, but find limited agreement with users' own judgments when persona evidence is insufficient. Garbacea et al. (2026) further show that rankings derived from individual users' pairwise preferences can differ substantially from aggregate rankings. Such work motivates user-conditioned preference evaluation, while cautioning against treating judge preferences as verified user benefits or human-calibrated minimal important differences.

Bias-Variance Decompositions. The bias-variance decomposition for squared loss is classical (Geman et al., 1992). It was later extended to zero-one and general losses (Kohavi & Wolpert, 1996; Friedman, 1997; Domingos, 2000; James, 2003), and revisited for overparameterized models (Belkin et al., 2019; Yang et al., 2020). Notably, every one of these decomposes error with respect to randomness in the training sample. Our framework characterizes identity leakage through posterior identification gain and bounds the leakage reduction and utility cost of idealized group averaging under stated conditions (Section 3).

3 Theoretical Framework

We formalize the privacy cost of personalization through the identity information in a response and the utility it provides to its recipient. Fix a query x and a population of n users with $\mathbb{P}(\mathbf{u} = i \mid x) > 0$. Candidate contexts, the model, observer information, and the group partition are fixed. We do not assume that queries are independent of users. Notation and proofs are collected in Appendix I.

Setup. Let $x \in \mathcal{X}$ be a query, $y \in \mathcal{Y}$ a response, and $z_i \in \mathcal{Z}$ user i's context. A fixed LLM π generates from $\pi_i(y) = \pi(y \mid x, \phi(z_i))$, where $\phi : \mathcal{Z} \to \mathcal{Z}$ processes the context. The population response distribution is $\bar{\pi}(y) = \sum_i \mathbb{P}(\mathbf{u} = i \mid x)\pi_i(y)$. Bold lowercase denotes random variables, including $\mathbf{z} = z_{\mathbf{u}}$. User i has utility $r_i(x, y) \in [0, 1]$, expected utility $V_i = \mathbb{E}_{\pi_i} r_i(x, \mathbf{y})$, and population utility $V = \sum_i \mathbb{P}(\mathbf{u} = i \mid x)V_i$. Let $\mathcal{G}$ partition the users into K groups, with $k_G = |G|$. Write $\epsilon_G(x) = \max_{i,j \in G} \sup_{y \in \mathcal{Y}} |r_i(x, y) - r_j(x, y)|$ for the largest utility difference within a group.

Posterior identification gain. For posteriors computed under the specified response distributions, define

$$\ell_i(x, y) = \mathbb{P}(\mathbf{u} = i \mid x, y) - \mathbb{P}(\mathbf{u} = i \mid x), \qquad L = \mathbb{E}_{\mathbf{u},\mathbf{y}|x} \ell_{\mathbf{u}}(x, \mathbf{y}). \tag{1}$$

This measures the additional probability assigned to the intended user after observing the response. No Pers. uses an empty context with user-independent generation randomness, so $\mathbf{y}_0 \perp \mathbf{u} \mid \mathbf{x}$ and its theoretical leakage is zero. A single group with nonempty context is a different mechanism.

Proposition 1 (Information-theoretic characterization)**.** *The expected posterior identification gain satisfies*

$$L = \mathbb{E}_{\mathbf{y} \sim \bar{\pi}} \sum_{i=1}^{n} \left[\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{y}) - \mathbb{P}(\mathbf{u} = i \mid x)\right]^2 \geq 0, \tag{2}$$

with equality exactly when $\mathbf{y} \perp \mathbf{u} \mid x$. *If* $\mathbb{P}(\mathbf{u} = i \mid x) = 1/n$, *then* $L = I_{\chi^2}(\mathbf{u}; \mathbf{y} \mid x)/n$, *where* I_{χ^2} *is Pearson chi-squared mutual information, not Shannon mutual information.*

The same L is the reduction in optimal multiclass Brier prediction risk. These standard information and scoring-rule connections (Gneiting & Raftery, 2007; Polyanskiy & Wu, 2022) give the framework a decision-theoretic interpretation; the contribution is their use to analyze personalization and group processing.

Theorem 1 (Group processing and conditional k-anonymity)**.** *Suppose the complete response distribution is identical for members of each group:* $\pi_i(y) = \pi_G(y)$ *for* $i \in G$. *Then*

$$\mathbb{P}(\mathbf{u} = i \mid x, y) = \mathbb{P}(\mathbf{u} = i \mid x, \mathbf{u} \in G)\mathbb{P}(\mathbf{u} \in G \mid x, y). \tag{3}$$

If, additionally, $\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{u} \in G) = 1/k_G$, *then*

$$\mathbb{P}(\mathbf{u} = i \mid x, y) \le 1/k_G, \qquad \ell_i(x, y) \le 1/k_G - \mathbb{P}(\mathbf{u} = i \mid x). \tag{4}$$

Thus the response preserves within-group prior odds. The anonymity bound requires uniformity *after* conditioning on the query and observer information; group size alone does not suffice.

Theorem 2 (Leakage reduction and utility cost of group averaging)**.** *Starting from individual response distributions* π_i, *define the idealized group mechanism*

$$\pi_G(y) = \sum_{j \in G} \mathbb{P}(\mathbf{u} = j \mid x, \mathbf{u} \in G)\pi_j(y), \qquad \pi_i^{\mathrm{grp}} = \pi_G \quad (i \in G). \tag{5}$$

It preserves $\bar{\pi}$. *Under a uniform conditional prior within each group,*

$$L_{\mathrm{id}} - L_{\mathrm{grp}} = \mathbb{E}_{\mathbf{y} \sim \bar{\pi}} \sum_{G \in \mathcal{G}} \sum_{i \in G} \left[\mathbb{P}_{\mathrm{id}}(\mathbf{u} = i \mid x, \mathbf{y}) - \frac{\mathbb{P}_{\mathrm{id}}(\mathbf{u} \in G \mid x, \mathbf{y})}{k_G} \right]^2 \ge 0, \tag{6}$$

where $\mathbb{P}_{\mathrm{id}}$ *denotes the individual mechanism. For any positive prior, the same group-averaging construction also satisfies*

$$|V_{\mathrm{id}} - V_{\mathrm{grp}}| \le \sum_{G \in \mathcal{G}} \mathbb{P}(\mathbf{u} \in G \mid x)\epsilon_G(x). \tag{7}$$

Group averaging removes within-group differences in identity posteriors, while similar utility functions limit the change in population utility. Distinct contexts do not imply different preferences. The averaging construction is an idealization: summarizing contexts or averaging embeddings need not average the LLM's response distributions. The results motivate, but do not certify, the baseline below.

4 Empirical Validation

We examine the practical implications of the framework: whether group-level context retains useful personalization with less user attribution than individual context, and when finer personalization adds benefit alongside its privacy cost.

4.1 Experimental Setup

Tasks and models. Table 2 evaluates Abstract Generation, Review Writing, and Topic Writing from LongLaMP-2/3/4 (Kumar et al., 2024) with Qwen3.5-9B, Llama-3.1-8B, and Gemma-3-12B. Figure 2 uses Qwen3.5-9B and adds News Headline, Scholarly Title, and Tweet Paraphrase from LaMP-4/5/7 (Salemi et al., 2024b). We use temporal splits, with fixed refusal masks applied across comparison arms. Generation uses temperature 0.7, nucleus probability 0.95, and seed 42. Each generator produces its own user summaries and group personas. Sample counts and implementation details are in Appendix J.

Personalization settings. We compare **No Pers.** (query only), **GrRAG** (group persona plus group retrieval), and **Full Pers.** (individual summary plus own-history retrieval). Inspired by Theorem 2, the group baseline uses k-means on user-summary embeddings to form K groups; each group supplies a persona and a pool of its members' documents, from which BM25 retrieves four documents for the query. This uses group-level context with the same generator and decoding settings as the other arms. We use $K = 100$ in the main comparison and $K \in \{10, 50, 100, 200\}$ in the granularity study. No Pers. is distinct from $K = 1$, which would still provide population-level context.

Table 2: **GrRAG across generators on LongLaMP.** Zero-shot generation uses $T = 0.7$, nucleus 0.95, seed 42. No Pers. receives the query only; GrRAG receives a group persona and four group-retrieved documents ($K = 100$); Full Pers. receives the user's own summary and four own-history documents. Util is separate-criterion personal-voice scoring by Qwen3.6-35B-A3B; — denotes unavailable separate-criterion scores. METEOR, Util, and Leak use matched valid items across the available arms within each task, generator, and metric. Their paired user-bootstrap 95% intervals and counts are in Appendix J.2. ROUGE and AUC retain their full-panel means. Leak is the Qwen attacker's top-1 increment over the matched query-only control with document candidates; AUC is a separate comparison (0.5 chance). Review Writing's generic-control failure limits interpretation of absolute Leak. Bold marks the largest observed reference or Utility score, not statistical significance.

		Qwen3.5-9B			Llama-3.1-8B			Gemma-3-12B		
Task	**Metric**	**No Pers.**	**GrRAG**	**Full Pers**	**No Pers.**	**GrRAG**	**Full Pers**	**No Pers.**	**GrRAG**	**Full Pers**
Abstract Generation	ROUGE-1 ↑	0.378	0.382	**0.399**	0.405	0.420	**0.425**	0.389	0.395	**0.402**
	ROUGE-L ↑	0.197	0.191	**0.202**	0.211	0.218	**0.221**	0.206	0.203	**0.210**
	METEOR ↑	0.208	0.237	**0.245**	0.219	0.249	**0.255**	0.205	0.229	**0.230**
	Util (1–5) ↑	2.51	2.64	**2.88**	2.06	—	**2.27**	2.38	—	**2.51**
	MIA-LLM ↓	0.501	0.923	0.939	0.502	0.915	0.928	0.500	0.916	0.923
	leak ↓	+0.000	+0.011	+0.048	−0.000	+0.002	+0.025	−0.000	+0.004	+0.018
Review Writing	ROUGE-1 ↑	0.258	0.308	**0.339**	0.334	0.349	**0.358**	0.311	0.332	**0.348**
	ROUGE-L ↑	0.131	0.138	**0.153**	0.153	0.157	**0.162**	0.145	0.144	**0.151**
	METEOR ↑	0.114	0.189	**0.201**	0.168	0.206	**0.211**	0.144	0.206	**0.220**
	Util (1–5) ↑	1.30	1.17	**1.46**	1.23	—	**1.32**	1.48	—	**1.50**
	MIA-LLM ↓	0.503	0.742	0.827	0.504	0.745	0.804	0.500	0.741	0.795
	leak ↓	−0.000	−0.033	+0.082	−0.000	−0.032	+0.047	−0.003	−0.034	+0.028
Topic Writing	ROUGE-1 ↑	0.279	0.261	**0.283**	0.284	0.293	**0.298**	**0.269**	0.250	0.266
	ROUGE-L ↑	0.121	0.116	**0.126**	0.127	0.131	**0.134**	**0.115**	0.108	0.114
	METEOR ↑	0.181	0.188	**0.198**	0.183	0.190	**0.200**	0.183	0.189	**0.200**
	Util (1–5) ↑	1.19	1.21	**1.45**	1.08	—	**1.16**	1.13	—	**1.31**
	MIA-LLM ↓	0.504	0.776	0.911	0.503	0.766	0.854	0.505	0.753	0.866
	leak ↓	−0.000	+0.041	+0.235	−0.001	+0.025	+0.147	−0.001	+0.008	+0.166

Reference overlap and judged utility. ROUGE-1/L and METEOR measure overlap with the held-out reference; Util is the 1–5 separate-criterion Qwen3.6-35B-A3B score for personal voice against the user's writing (Figure 7). Table 2 and Figure 2 use this same configuration. These measurements capture different aspects of personalization rather than directly measuring $r_i(x, y)$. Separate-criterion group-arm scores are unavailable for Llama and Gemma and are marked —; we make no group-versus-individual Utility claim for those cells. The deployed Utility judge passes the profile-discrimination threshold on five strongly personalized task panels, with Scholarly Title below threshold. Supplementary GraSPeR reference-match scores (Ni et al., 2026) and validation appear in Appendix G.

User attribution. A fixed Qwen3.6-35B-A3B attacker observes the query, response, and evidence for five candidate users: the intended user and four population decoys. We report the increment in top-1 identification accuracy over the condition's matched query-only control,

$$\text{leak} = \widehat{\text{Acc}}_a(\mathbf{x}, \mathbf{y}) - \widehat{\text{Acc}}_a(\mathbf{x}), \tag{8}$$

where each control retains the same candidate evidence as its response condition. Table 2 supplies retrieved documents as candidate evidence; Figure 2 supplies personal summaries. Empirical Leak measures additional attribution by a fixed attacker, rather than estimating the posterior quantity L or mutual information. It can be negative when a response impairs the attacker's decisions. No Pers. is a measured control, and the table's one-vs-rest AUC is a separate score. Appendix B specifies candidate construction and scoring.

Primary Utility, METEOR, and Leak comparisons use matched valid items within each task, generator, and metric. The granularity figures retain their full-panel means; paired granularity contrasts use common valid items across the displayed settings for each metric. Paired differences use 10,000 user-bootstrap draws; Appendix J.2 reports counts, pointwise 95% intervals, and user sign-flip tests with Holm correction. All inference conditions on the recorded generations and candidate sets.

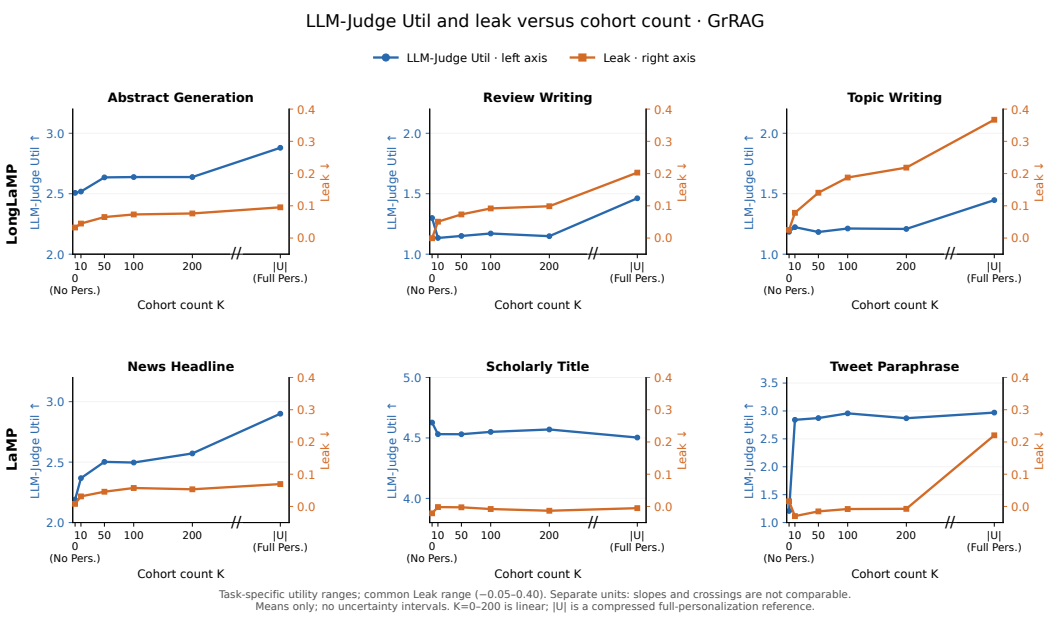

Figure 2: **LLM-Judge Util and identity attribution across personalization granularities.** A frozen Qwen3.5-9B generator uses a cohort persona and four retrieved cohort documents. Blue curves show LLM-Judge Util on the original 1–5 judge scale; orange curves show Leak, the increment in five-way identification accuracy over a matched query-only control, using candidate personal summaries and population decoys. No Pers. uses the bare query; Full Pers. uses the user's own summary and retrieved history. The same refusal masks apply to Review and Topic Writing. Left-axis display ranges vary by task; right-axis ranges are shared. Curves show means without uncertainty intervals. The horizontal axis is linear through $K{=}200$; the Full Pers. endpoint is compressed after an axis break. The two vertical axes have different units, so slopes and crossings are not directly comparable.

4.2 Personalization–Privacy Tradeoff Across Generators

Table 2 shows a consistent reference-overlap and attribution tradeoff. Across all nine task–generator comparisons, GrRAG has higher mean METEOR than No Pers. and lower mean Leak than Full Pers. Full Pers. has the highest METEOR mean in each comparison. Personal-voice Utility is task-dependent: for Qwen3.5-9B, GrRAG improves the mean over No Pers. on Abstract Generation but lowers it on Review Writing. ROUGE-L also varies, improving over No Pers. in four of nine group-context comparisons.

For Abstract Generation with Llama-3.1-8B, METEOR is 0.219, 0.249, and 0.255 for No Pers., GrRAG, and Full Pers. Moving from group to individual context adds 0.007 METEOR while Leak rises from 0.002 to 0.025. On Topic Writing with Qwen3.5-9B, this step adds 0.235 Utility (paired 95% CI $[+0.188, +0.283]$) and 19.45 percentage points of Leak (CI $[+17.36, +21.55]$).

These comparisons illustrate the framework's distinction between useful personalization and identity information. Individual context can improve measured utility, but the additional attribution and benefit vary substantially. The group baseline offers an intermediate choice whose value must be assessed for the task, rather than inferred from context specificity alone.

4.3 Effect of Personalization Granularity

Figure 2 shows the full-panel trends. On common valid items across its settings (Appendix J.2), from $K = 10$ to 200, Topic Writing's Leak rises by 14.29 percentage points while Utility changes by -0.015. Abstract Generation gains 0.119 Utility with a 3.27-percentage-point Leak increase; News Headline gains 0.205 Utility with a 2.29-percentage-point Leak increase. Finer groups can therefore add attribution with differing Utility effects.

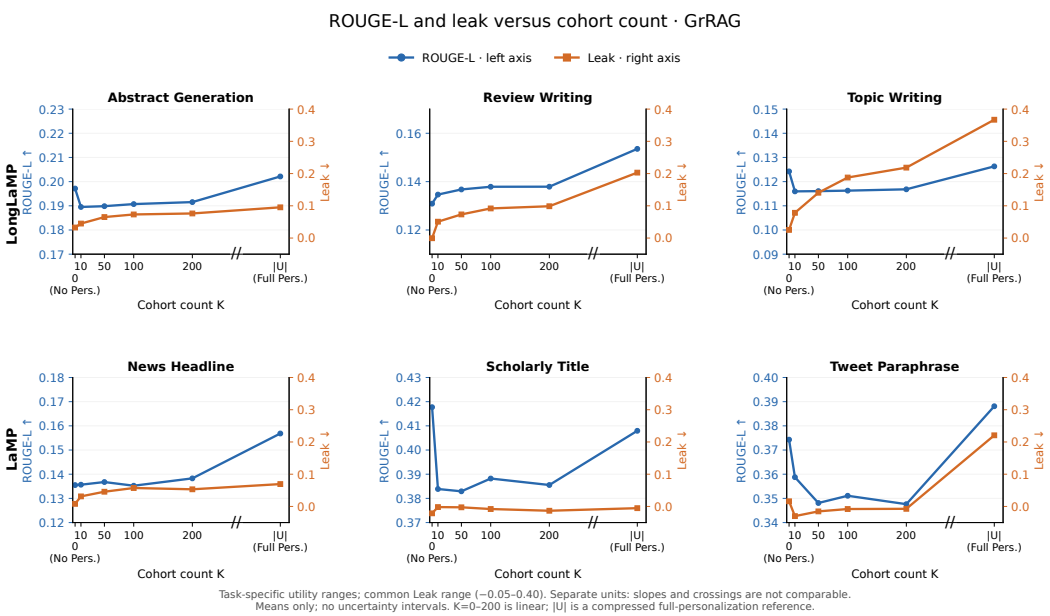

Figure 3: **ROUGE-L and attribution across cohort granularities.** Setup and attack protocol match Figure 2. Left axes show raw ROUGE-L with task-specific display ranges; right axes show Leak on a common scale. Curves report means without uncertainty intervals. No Pers. and Full Pers. are measured references; the horizontal axis is linear through K=200 with the Full Pers. endpoint compressed after the break.

Review Writing changes by only $+0.015$ Utility across the group sweep, with all group means below No Pers. Scholarly Title's Utility scores fail the profile-discrimination threshold and do not establish a personalization benefit. The transition to individual context is separate: on Tweet Paraphrase, Full Pers. adds $+0.009$ Utility relative to $K = 100$ (paired 95% CI $[-0.074, +0.093]$) and 22.58 percentage points of Leak. A nonsignificant Utility difference does not establish equivalence.

The theory helps interpret this variation. Under a uniform conditional prior and identical response distributions within groups, $L \leq (K-1)/n$ (Appendix I.3): fewer groups limit possible identity distinctions. The utility cost depends separately on within-group utility differences. This bound does not predict monotone empirical curves, and the fitted partitions need not be nested. The observed task differences support assessing the benefit of additional distinctions, while leaving utility similarity itself unmeasured.

Paired $K = 10$ versus 200 comparisons show increased attribution on Abstract Generation, Review Writing, Topic Writing after Holm correction; Utility increases survive correction on Abstract Generation, News Headline (Appendix J.2). These paired tests use common valid items; their means can differ slightly from the figure's full-panel means. Full sweep tables and additional comparisons appear in Appendix A.

4.4 Ablation and Robustness Analysis

Retrieval pool and persona. The retrieval-only LongLaMP comparison holds the retriever and depth fixed while replacing own-history retrieval with group retrieval (Table 5). Group retrieval has lower displayed document-candidate Leak in all nine task–generator comparisons. Adding a persona has mixed effects: in the supplementary Qwen panel (Table 8), it raises Abstract Generation Util from 2.61 to 2.64 but lowers GraSPeR reference match from 3.58 to 2.99; on Review Writing it lowers both Util (1.32 to 1.17) and reference match (2.01 to 1.77). The group context choice matters even without a persona, whose additional value depends on the measured outcome.

Candidate evidence and within-group identification. With document evidence and decoys from the target's group, GrRAG has lower displayed Leak than Full Pers. on all six tasks (Table 12). This

evaluates identification among five eligible candidates, rather than certifying anonymity across the entire group. Candidate evidence also changes the assessment of abstraction-only methods: their document-evidence protocol supplies empty candidate blocks, whereas personal summaries reveal attribution those blocks missed. Appendix G.3 reports the associated controls, including cases where a generic response changes attacker accuracy despite supplying no additional identity information.

Independent attackers. A replay with fixed candidate inputs evaluates 150 examples each from News, Review, and Topic using Qwen and Llama attackers. On Topic Writing with document evidence, group context reduces Leak relative to individual context by 19.00 and 17.33 percentage points, respectively; both survive Holm correction over the 12 replay contrasts. News and Review remain inconclusive, with some signs reversed. Appendix J.4 reports all contrasts and sampling details. The reduction therefore persists under an independent attacker on Topic, while the broader comparison remains task- and attacker-dependent.

5 Discussion and Limitations

The results motivate evaluating the benefit and attribution cost of each additional level of personal context. Table 2 shows METEOR gains from group context and further gains from individual context, while judged Utility varies by task, while Figure 2 shows that finer distinctions can add identity cues with little Util improvement. The framework explains how these patterns can coexist: leakage reflects changes in identity posteriors, whereas utility depends on the recipient's preferences.

The theoretical guarantees apply under explicit conditions. Conditional k-anonymity requires identical complete response distributions and a uniform prior within each group after conditioning on the query. The leakage reduction theorem concerns idealized averaging of response distributions; its utility bound depends on utility similarity. The implemented baseline groups user representations and constructs context, so it neither implements that average nor establishes utility similarity. The experiments evaluate the framework's practical implications rather than verify its assumptions.

Our measurements also have a limited scope. Overlap and judge-rated Util are proxies for utility, without a human-preference study translating their differences into user benefit. Attribution depends on candidate evidence and attacker behavior. Public-text benchmarks, fixed generation seeds, refusal masks, and selected replay samples limit generalization. The study concerns intended-user identification from individual responses; it does not establish protection against sensitive-attribute disclosure, extraction, or repeated interactions.

6 Conclusion

We formalize the privacy cost of personalization through posterior identification gain, connecting it to information theory and conditional group anonymity. Idealized group averaging reduces identity leakage under stated conditions, with utility change bounded by differences in users' utilities. Experiments with a simple group-context baseline show METEOR gains over the generic baseline with less measured attribution than individual personalization across the main task–generator comparisons; personal-voice Utility varies by task. Granularity experiments further show that additional identity cues need not yield consistent Util gains. These findings support choosing personalization granularity according to the task's measured benefit and attribution cost.

AI use statement

Large language models were used in this work in three roles, all under author direction. (i) As *experimental subjects and instruments*: the generators, the profile summarizer, the persona writer, the LLM judges, and the re-identification attacker are all LLMs, as described in Section 4 and the appendices. (ii) As *coding and analysis assistants*: LLM-based agents (Anthropic Claude and OpenAI Codex) wrote and refactored parts of the experiment pipeline, ran analyses, and produced first drafts of tables and figures from the recorded outputs; the primary matched statistics are regenerated from per-item records by the analysis script. (iii) As *writing assistants*: LLM tools helped draft and edit prose, including the related-work search. The authors designed the study, verified the theoretical results, reviewed all generated code and text, and take full responsibility for the content.

Reproducibility statement

All experiments use public benchmarks (LaMP-4/5/7, LongLaMP-2/3/4, temporal splits). Generation uses open-weight models; judge validation also includes proprietary models (Table 18). The pipeline includes profile summarization, clustering, persona writing, retrieval with query-side decontamination, generation, judging, and attack. Generation uses $T = 0.7$, nucleus probability 0.95, and seed 42. Per-item generations and scores are retained; the primary matched means and confidence intervals are recomputed from these records by the analysis script. Complete request/response envelopes are retained for the frozen-input attacker replay. Prompt templates, sample counts, scoring rules, and theoretical proofs are given in the appendices. Historical candidate-record limitations are stated in Appendix J.2.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. Association for Computing Machinery, 2016.

Michael Aerni, Jie Zhang, and Florian Tramèr. Evaluations of machine learning privacy defenses are misleading. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024. doi: 10.1145/3658644.3690194. URL https://arxiv.org/abs/2404.17399.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Alex Berke and Dan Calacci. Privacy limitations of interest-based advertising on the web: A post-mortem empirical analysis of Google's FLoC. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022. arXiv:2201.13402.

Yohan Beugin and Patrick McDaniel. Interest-disclosing mechanisms for advertising are privacy-exposing (not preserving). *Proceedings on Privacy Enhancing Technologies*, 2024(1):41–57, 2024. doi: 10.56553/popets-2024-0004.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650, 2021.

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 (SP)*, pp. 1897–1914, 2022.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *IEEE Security and Privacy Workshops (SPW)*, 2025. arXiv:2406.16201.

Pedro Domingos. A unified bias-variance decomposition and its applications. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 231–238. Morgan Kaufmann, 2000.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can LLM be a personalized judge? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10126–10141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL https://aclanthology.org/2024.findings-emnlp.592/.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76852–76871, 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

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)*, pp. 265–284. Springer, 2006.

Lukáš Eigler, Jindřich Libovický, and David Hurych. Llm as a meta-judge: Synthetic data for nlp evaluation metric validation. *arXiv preprint arXiv:2603.09403*, 2026.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9, 2021.

James Flemings, Bo Jiang, Wanrong Zhang, et al. Estimating privacy leakage of augmented contextual knowledge in language models. *arXiv preprint arXiv:2410.03026*, 2024.

Jerome H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

Cristina Garbacea, Heran Wang, and Chenhao Tan. Personalized benchmarking: Evaluating LLMs by individual preferences. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 639–675, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.31. URL https://aclanthology.org/2026.findings-acl.31/.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Aaron Grattafiori and et. al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. Privacy implications of retrieval-based language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14887–14902, Singapore, 2023. Association for Computational Linguistics.

Gareth M. James. Variance and bias for general loss functions. *Machine Learning*, 51(2):115–135, 2003.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1999–2014, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.110. URL https://aclanthology.org/2024.acl-long.110/.

Ron Kohavi and David H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 275–283, 1996.

Ishita Kumar, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, and Hamed Zamani. LongLaMP: A benchmark for personalized long-form text generation, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. Teach LLMs to personalize: An approach inspired by writing education. *arXiv preprint arXiv:2308.07968*, 2023.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. Beyond correlation: Making sense of the score differences of new MT evaluation metrics. In Masao Utiyama and Rui Wang (eds.), *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pp. 186–199, Macau SAR, China, September 2023. Asia-Pacific Association for Machine Translation. URL https://aclanthology.org/2023.mtsummit-research.16/.

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

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, 2023. Association for Computational Linguistics.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. SoK: Membership inference attacks on LLMs are rushing nowhere (and how to fix it). In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2025. URL https://arxiv.org/abs/2406.17975.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *International Conference on Learning Representations*, 2024.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pp. 198–219, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations*, 2025.

Bo Ni, Branislav Kveton, Samyadeep Basu, Subhojyoti Mukherjee, Franck Dernoncourt, Sungchul Kim, Ruiyi Zhang, Nedim Lipka, Ryan A. Rossi, and Tyler Derr. Reasoning-based personalized generation for users with sparse data, 2026.

Gibson Nkhata, Uttamasha Anjally Oyshi, Quan Mai, and Susan Gauch. ClusterRAG: Cluster-based collaborative filtering for personalized retrieval-augmented generation. *arXiv preprint arXiv:2605.18769*, 2026.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of llm-as-a-judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*, 2026.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022. URL `https://people.lids.mit.edu/yp/homepage/data/itbook-2022.pdf`. Public author draft.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*, 2023.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Alireza Salemi and Hamed Zamani. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. Association for Computing Machinery, 2025.

Alireza Salemi and Hamed Zamani. P3: Improving user privacy in personalized generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2026.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 752–762. Association for Computing Machinery, 2024a.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, 2024b. Association for Computational Linguistics.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

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

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Language models are advanced anonymizers. In *International Conference on Learning Representations*, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/hash/f478b2e8ad9ff0756bf5b79fb31c330f-Abstract-Conference.html`.

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

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024a. Association for Computational Linguistics.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024b. Association for Computational Linguistics.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. Unveiling privacy risks in LLM agent memory. *arXiv preprint arXiv:2502.13172*, 2025.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 297–312. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.21/.

Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. Personalized generation in large model era: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, 2025. Association for Computational Linguistics.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4505–4524, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. PROPER: A progressive learning framework for personalized large language models with group-level adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16399–16411, Vienna, Austria, 2025a. Association for Computational Linguistics.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. Personalization of large language models: A survey. *Transactions on Machine Learning Research*, 2025b.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. HYDRA: Model factorization framework for black-box LLM personalization. In *Advances in Neural Information Processing Systems*, 2024.

Table 3: **GrRAG across generators on LaMP** (zero-shot; decode and instruments as in Table 4). **No Pers.**: bare query. **GrRAG**: cohort persona + cluster-BM25 over the $K{=}100$ cohort, $k{=}4$. **Full Pers**: the user's own summary + BM25 over their own profile, $k{=}4$ (standard stacked personalization). ROUGE/METEOR are lexical references, Util measures personal voice, and AUC is secondary. **Bold**: highest reference-match or utility score within each generator's three arms; red: largest score within each generator (AUC is secondary). †: bundled Utility scoring, reported for within-table comparison; primary Utility conclusions use separate-criterion scores.

		Qwen3.5-9B			Llama-3.1-8B			Gemma-3-12B		
Task	**Metric**	**No Pers.**	**GrRAG**	**Full Pers**	**No Pers.**	**GrRAG**	**Full Pers**	**No Pers.**	**GrRAG**	**Full Pers**
News Headline	ROUGE-1 ↑	0.155	0.155	**0.176**	0.160	0.168	**0.176**	0.161	0.176	**0.185**
	ROUGE-L ↑	0.136	0.135	**0.157**	0.139	0.146	**0.153**	0.140	0.152	**0.162**
	METEOR ↑	0.095	0.102	**0.120**	0.098	0.103	**0.106**	0.089	0.103	**0.111**
	LLM Judge Util† (1–5) ↑	2.32	2.44	**2.92**	2.05	2.14	**2.40**	2.24	2.31	**2.70**
	MIA-LLM ↓	0.495	0.907	0.926	0.492	0.908	0.916	0.492	0.911	0.925
	leak ↓	+0.000	+0.047	+0.097	−0.001	+0.061	+0.073	−0.004	+0.058	+0.085
Scholarly Title	ROUGE-1 ↑	**0.496**	0.469	0.489	0.477	**0.478**	0.477	**0.488**	0.481	0.485
	ROUGE-L ↑	**0.418**	0.388	0.408	**0.395**	0.391	0.392	**0.407**	0.399	0.405
	METEOR ↑	**0.413**	0.401	0.413	**0.385**	0.374	0.361	0.373	**0.382**	0.381
	LLM Judge Util† (1–5) ↑	**4.07**	4.00	4.02	**3.84**	3.79	3.74	4.03	4.05	**4.05**
	MIA-LLM ↓	0.511	0.952	0.956	0.516	0.951	0.954	0.513	0.951	0.957
	leak ↓	+0.002	+0.009	+0.007	+0.001	+0.015	+0.022	+0.001	+0.001	+0.023
Tweet Paraphrase	ROUGE-1 ↑	0.428	0.403	**0.442**	**0.386**	0.367	0.378	0.328	0.390	**0.404**
	ROUGE-L ↑	0.374	0.351	**0.388**	**0.325**	0.310	0.321	0.274	0.342	**0.352**
	METEOR ↑	0.271	0.288	**0.308**	0.240	0.228	**0.244**	0.181	0.224	**0.230**
	LLM Judge Util† (1–5) ↑	1.20	3.33	**3.38**	1.10	**3.32**	2.98	1.29	2.80	**3.48**
	MIA-LLM ↓	0.509	0.683	0.806	0.508	0.672	0.760	0.506	0.669	0.770
	leak ↓	−0.001	−0.022	+0.122	+0.001	−0.029	+0.088	+0.002	−0.023	+0.093

A Additional Result Tables

Original cross-generator panels. The following tables provide the LaMP counterpart to the main LongLaMP comparison and the group-versus-own retrieval ablations.

Full group-count sweeps. The following tables retain the lexical, Util, reference-match, and attack scores for the original K sweeps. Their judge and attacker configurations are stated in the captions; the summary-candidate Figure 2 and its paired analysis in Appendix J.3 are evaluated separately.

Additional strategy and attacker comparisons. The following tables compare additional strategies and attacker protocols.

Five generators, one ordering. Table 10 summarizes the cross-generator panels (Tables 4 and 5); Table 11 shows the same own-vs-cohort contrast under raw AUC and under leak.

One spectrum, three attackers, with the $k{=}0$ arms within the cohort. Table 12 repeats the leakage rows of Table 8 and adds the within-cohort *summary* attacker on the $k{=}0$ arms.

Abstraction arms under the summary attacker, all generators. Table 13 completes the summary-attacker column of Table 12 for the other four generators and the three fine-tuned models.

B Privacy Attacks

We measure privacy leakage with one attack on the (query, output, candidate representation) triple: an LLM five-way re-identification attacker that names which of five candidate users the output was personalized for; higher is more leakage. The attack is a *linkability* attack: it quantifies whether the output lets an adversary link it to the target user, driven by content the output shares with the user's representation beyond the query's shared topic. We do not perform *attribute inference* (disclosure of a labeled sensitive attribute), which LaMP and LongLaMP do not annotate, so our leakage is linkability

Table 4: **Cohort vs. own personalization across generators on LaMP** (zero-shot; sampled decode T=0.7, nucleus 0.95, seed 42). **No Pers.**: bare query. **C-RAG**: cluster-BM25 over the K=100 cohort, k=4. **RAG**: BM25 over the user's own profile (K=$|\mathcal{U}|$), k=4. GraSPeR-Judge‡ (1–7) measures reference match; ROUGE-1/L and METEOR are lexical references, and the original judge Util† (1–5) measures personal voice. MIA-LLM AUC is a secondary score; attribution uses leak, the attacker's top-1 minus its query-only control. **Bold**: best utility within each generator's three arms; red: largest score within each generator (AUC is secondary). Each generator writes its own summaries, personas and clustering; the attacker and judges are Qwen3.6-35B throughout. †: bundled Utility scoring, reported for within-table comparison; primary Utility conclusions use separate-criterion scores. ‡: GraSPeR reference match on a 1–7 scale.

		Qwen3.5-9B			Llama-3.1-8B			Gemma-3-12B		
Task	**Metric**	**No Pers.**	**C-RAG**	**RAG**	**No Pers.**	**C-RAG**	**RAG**	**No Pers.**	**C-RAG**	**RAG**
News Headline	ROUGE-1 ↑	0.155	0.165	**0.174**	0.160	0.168	**0.175**	0.161	0.177	**0.187**
	ROUGE-L ↑	0.136	0.145	**0.155**	0.139	0.148	**0.156**	0.140	0.155	**0.164**
	METEOR ↑	0.095	0.104	**0.114**	0.098	0.096	**0.106**	0.089	0.099	**0.106**
	Util† (1–5) ↑	2.32	2.32	**2.56**	2.05	2.11	**2.33**	2.24	2.41	**2.50**
	GraSPeR‡ (1–7) ↑	2.29	2.25	**2.30**	**2.28**	2.18	2.18	2.27	2.31	**2.32**
	MIA-LLM ↓	0.495	0.895	0.911	0.492	0.896	0.917	0.492	0.897	0.910
	leak ↓	+0.000	+0.021	+0.054	−0.001	+0.042	+0.073	−0.004	+0.036	+0.057
Scholarly Title	ROUGE-1 ↑	0.496	**0.502**	0.500	0.477	**0.481**	0.480	0.488	**0.494**	0.494
	ROUGE-L ↑	0.418	0.423	**0.424**	0.395	**0.396**	0.395	0.407	0.411	**0.413**
	METEOR ↑	0.413	**0.419**	0.418	**0.385**	0.356	0.355	0.373	0.371	**0.378**
	Util† (1–5) ↑	**4.07**	3.96	3.91	**3.84**	3.63	3.65	**4.03**	4.01	3.97
	GraSPeR‡ (1–7) ↑	**4.23**	4.13	4.14	**4.10**	4.02	3.96	**4.13**	4.06	4.05
	MIA-LLM ↓	0.511	0.955	0.953	0.516	0.947	0.953	0.513	0.952	0.954
	leak ↓	+0.002	+0.007	+0.013	+0.001	+0.005	+0.016	+0.001	+0.008	+0.013
Tweet Paraphrase	ROUGE-1 ↑	0.428	0.434	**0.435**	0.386	**0.403**	0.401	0.328	**0.382**	0.379
	ROUGE-L ↑	0.374	0.379	**0.380**	0.325	**0.346**	0.343	0.274	0.331	**0.331**
	METEOR ↑	0.271	0.280	**0.283**	0.240	0.260	**0.260**	0.181	**0.218**	0.216
	Util† (1–5) ↑	1.20	1.25	**1.28**	1.10	1.34	**1.62**	1.29	1.95	**2.47**
	GraSPeR‡ (1–7) ↑	5.95	5.94	**5.95**	**5.31**	4.99	5.02	**5.20**	4.99	5.08
	MIA-LLM ↓	0.509	0.713	0.722	0.508	0.700	0.773	0.506	0.695	0.787
	leak ↓	−0.001	+0.015	+0.020	+0.001	+0.001	+0.086	+0.002	+0.006	+0.102

on public text, a proxy for sensitive-content disclosure. The judge's Utility score (Appendix F) measures personal voice; privacy is measured separately by user attribution. The same attacker is run under the three adversaries of Section 4.1, which differ only in the candidate representation and the decoy pool: documents/population (each candidate contributes the k articles BM25 retrieves for that user under the same query, truncated to a fixed character budget, so true and decoy blocks match in size and topic; decoys are four random other users), summaries/population (each candidate contributes their 200-word profile summary, the persona system's own view of the user), and documents/within-cohort (decoys are four other users of the true user's K=100 cohort, sampled by author; items whose cohort has fewer than four eligible users are dropped and counted). The attacker LLM is Qwen3.6-35B-A3B, scored from the first-token log-probabilities over the candidate labels; its scores measure attacker-specific identification performance. We also evaluate an embedding-classifier attack (MiniLM features, logistic regression) as a comparison; it fails the planted-response controls and does not support our privacy conclusions.

Null band. Every leak is read against an *unpersonalized-response null*: the No Pers. output planted as the response under the same candidate blocks, so the attacker sees a response with no profile signal. Its leak is the per-task bias of the decision rule (Section 4); its query-only control reproduces the published one to within 0.005, which also checks that the candidate sampler is deterministic.

Tail risk. Mean leak near zero is consistent with a few users being fully identified. Table 14 reports the share of users whose per-user leak is at least 0.5, next to the same share under the null.

Table 5: **Cohort vs. own personalization across generators on LongLaMP** (zero-shot; sampled decode T=0.7, nucleus 0.95, seed 42). **No Pers.**: bare query. **C-RAG**: cluster-BM25 over the K=100 cohort, k=4. **RAG**: BM25 over the user's own profile (K=$|\mathcal{U}|$), k=4. GraSPeR-Judge‡ (1–7) measures reference match; ROUGE-1/L and METEOR are lexical references, and the original judge Util† (1–5) measures personal voice. MIA-LLM AUC is a secondary score; attribution uses leak, the attacker's top-1 minus its query-only control. **Bold**: best utility within each generator's three arms; red: largest score within each generator (AUC is secondary). Each generator writes its own summaries, personas and clustering; the attacker and judges are Qwen3.6-35B throughout. †: bundled Utility scoring, reported for within-table comparison; primary Utility conclusions use separate-criterion scores. ‡: GraSPeR reference match on a 1–7 scale.

		Qwen3.5-9B			Llama-3.1-8B			Gemma-3-12B		
Task	**Metric**	**No Pers.**	**C-RAG**	**RAG**	**No Pers.**	**C-RAG**	**RAG**	**No Pers.**	**C-RAG**	**RAG**
Abstract Generation	ROUGE-1 ↑	0.378	0.392	**0.398**	0.405	0.421	**0.427**	0.389	0.392	**0.396**
	ROUGE-L ↑	0.197	0.203	**0.205**	0.211	0.221	**0.224**	0.206	0.211	**0.213**
	METEOR ↑	0.208	0.226	**0.232**	0.219	0.242	**0.248**	0.205	0.211	**0.214**
	Util† (1–5) ↑	2.41	2.48	**2.58**	1.80	2.03	**2.12**	2.32	2.24	**2.33**
	GraSPeR‡ (1–7) ↑	**3.80**	3.58	3.59	**3.52**	3.31	3.37	**3.83**	3.70	3.70
	MIA-LLM ↓	0.501	0.916	0.943	0.502	0.913	0.930	0.500	0.914	0.921
	leak ↓	+0.000	−0.002	+0.063	−0.000	−0.004	+0.038	+0.000	−0.001	+0.019
Review Writing	ROUGE-1 ↑	0.258	0.292	**0.305**	0.334	0.346	**0.360**	0.311	0.333	**0.343**
	ROUGE-L ↑	0.131	0.140	**0.147**	0.153	0.158	**0.166**	0.145	0.152	**0.158**
	METEOR ↑	0.114	0.141	**0.148**	0.168	0.189	**0.196**	0.144	0.167	**0.172**
	Util† (1–5) ↑	1.18	1.23	**1.33**	1.10	1.15	**1.26**	1.34	1.32	**1.47**
	GraSPeR‡ (1–7) ↑	1.98	2.01	**2.03**	**2.00**	1.94	1.94	2.13	**2.14**	2.13
	MIA-LLM ↓	0.503	0.742	0.761	0.504	0.740	0.798	0.500	0.737	0.765
	leak ↓	−0.001	−0.033	−0.007	−0.001	−0.048	+0.039	−0.003	−0.031	−0.010
Topic Writing	ROUGE-1 ↑	**0.279**	0.225	0.241	0.284	0.281	**0.286**	0.269	0.259	**0.276**
	ROUGE-L ↑	**0.121**	0.112	0.118	0.127	0.127	**0.129**	0.115	0.119	**0.124**
	METEOR ↑	**0.171**	0.119	0.134	0.179	0.183	**0.194**	**0.179**	0.142	0.156
	Util† (1–5) ↑	1.09	1.22	**1.39**	1.03	1.07	**1.24**	1.06	1.19	**1.26**
	GraSPeR‡ (1–7) ↑	1.58	1.65	**1.67**	**1.43**	1.30	1.36	1.54	1.55	**1.65**
	MIA-LLM ↓	0.504	0.746	0.870	0.503	0.742	0.953	0.505	0.753	0.863
	leak ↓	−0.001	−0.001	+0.179	−0.001	−0.009	+0.329	−0.002	+0.007	+0.157

MIA-LR (membership inference). For each retrieval depth k we build a balanced binary dataset. Each $(\mathbf{q}, \mathbf{y})$ pair yields one positive $(\mathbf{q}, \mathbf{y}, P^{+})$, where P^{+} is the profile actually used during generation, and one negative $(\mathbf{q}, \mathbf{y}, P^{-})$, where P^{-} is an equal-size profile retrieved with the same query from a randomly chosen *different* author. Each tuple is encoded with `all-MiniLM-L6-v2`: with $\mathbf{e}_q, \mathbf{e}_y \in \mathbb{R}^{384}$ the query and output embeddings and $\mathbf{e}_P = \frac{1}{k}\sum_{i=1}^{k} \mathbf{e}_{a_i}$ the mean profile embedding, the feature is

$$\mathbf{x} = [\mathbf{e}_q \| \mathbf{e}_y \| \mathbf{e}_P \| \cos(\mathbf{e}_q, \mathbf{e}_P) \| \cos(\mathbf{e}_y, \mathbf{e}_P)] \in \mathbb{R}^{1154}.$$

A logistic-regression classifier is trained on the membership label with an 80/20 split by example ID (samples from one $(\mathbf{q}, \mathbf{y})$ never cross the split). We report held-out AUC, where 0.5 is chance.

MIA-LLM (N-way re-identification). Following the P3 protocol (Salemi & Zamani, 2026), we pose re-identification as a forced choice: the attacker LLM is shown the generated output and N candidate profiles — the true profile plus $N-1$ decoys from other authors (each the k articles that author's query retrieves) — and must pick which profile produced the output. We use N=5 (random top-1 floor 0.20) and report the one-vs-rest AUC over the N choices as MIA-LLM (0.5 chance). We also track top-1 accuracy and a *query-only* control that hides the output, so the choice must rely on the query alone; the gap leakΔ = top1 − top1$_{\text{query-only}}$ isolates the leakage *attributable to the output* rather than the query. The attacker scores each option by its first-token log-probability (a single forward pass, no decoding), so the measurement is deterministic.

Planted-response controls with known ground truth test both attacks: MIA-LLM passes all three controls on News Headline and Abstract Generation (Table 20); MIA-LR fails and is reported only as a secondary column (Appendix G.3).

Table 6: **Cluster-count (K) sweep on LaMP** (zero-shot, frozen Qwen3.5-9B; decode as in Table 4; test set, temporal setting). The arm is **GrRAG** (cluster summary + cluster-BM25, k=4) swept over $K \in \{10, 50, 100, 200\}$; **No Pers.** is the bare query and K=$|\mathcal{U}|$ the own-profile reference (**Summary** $\oplus$ **RAG**). Measures: GraSPeR-Judge‡ (1–7) for reference match; ROUGE-1/L and METEOR as lexical references; the original judge Util† (1–5) for personal voice, all as defined in Table 4 . Secondary attack measures: MIA-LLM and MIA-LR AUC. These do not isolate response-attributable information; leak is plotted in Figure 2. **Bold**: best utility per row; red: most-leaking per row. ^{ns} on utility rows: not significantly different from **No Pers.** (two-sided sign-flip permutation test on matched items, $p \geq 0.05$). ^{ns} on MIA-LLM rows: arm not significantly different from the fully-personalized K=$|\mathcal{U}|$ **Summary** $\oplus$ **RAG** reference under the paired bootstrap on the AUC difference (10^4 resamples, 95% CI covering 0; Appendix B). MIA-LR carries no significance test (failed validation, Appendix G). Note K is the number of clusters. †: bundled Utility scoring, reported for within-table comparison; primary Utility conclusions use separate-criterion scores. ‡: GraSPeR reference match on a 1–7 scale.

			GrRAG				K=$\|\mathcal{U}\|$
Task	**Metric**	**No Pers.**	K=10	K=50	K=100	K=200	**Summary** $\oplus$ **RAG**
News Headline	ROUGE-1 ↑	0.154	0.156^{ns}	0.159	0.158^{ns}	0.156^{ns}	**0.176**
	ROUGE-L ↑	0.135	0.136^{ns}	0.140	0.139^{ns}	0.136^{ns}	**0.156**
	METEOR ↑	0.093	0.106	0.108	0.106	0.104	**0.118**
	Util† (1–5)	2.34	2.33^{ns}	2.48	2.53	2.47	**2.81**
	GraSPeR-Judge‡ (1–7) ↑	**2.31**	2.23	2.22	2.19	2.21	2.28^{ns}
	MIA-LLM ↓	0.495	0.898	0.905	0.912	0.917	0.924
	MIA-LR ↓	0.500	0.875	0.878	0.885	0.885	0.899
Scholarly Title	ROUGE-1 ↑	0.496	**0.503**^{ns}	0.497^{ns}	0.492^{ns}	0.490^{ns}	0.491^{ns}
	ROUGE-L ↑	0.419	**0.428**^{ns}	0.419^{ns}	0.413^{ns}	0.413^{ns}	0.412^{ns}
	METEOR ↑	0.413	**0.439**	0.434	0.429	0.426	0.415^{ns}
	Util† (1–5)	**4.06**	4.00^{ns}	4.00^{ns}	3.98	4.02^{ns}	3.99^{ns}
	GraSPeR-Judge‡ (1–7) ↑	4.22	**4.27**^{ns}	4.19^{ns}	4.17^{ns}	4.13	4.13
	MIA-LLM ↓	0.512	0.953^{ns}	0.953^{ns}	0.950^{ns}	0.952^{ns}	0.954
	MIA-LR ↓	0.500	0.957	0.957	0.956	0.957	0.959
Tweet Paraphrase	ROUGE-1 ↑	0.428	0.404	0.404	0.403	0.402	**0.441**
	ROUGE-L ↑	0.374	0.350	0.348	0.350	0.349	**0.386**
	METEOR ↑	0.271	0.286	0.286	0.287	0.285	**0.309**
	Util† (1–5)	1.20	3.18	3.25	3.24	**3.33**	3.32
	GraSPeR-Judge‡ (1–7) ↑	**5.93**	5.18	5.22	5.20	5.23	5.64
	MIA-LLM ↓	0.511	0.659	0.677	0.681	0.677	0.802
	MIA-LR ↓	0.500	0.623	0.626	0.622	0.632	0.688

Paired bootstrap for comparing two AUCs. Suppose two classification methods, A and B, are evaluated on the same test set

$$\mathcal{D} = \{(y_i, s_{A,i}, s_{B,i})\}_{i=1}^{n},$$

where $y_i \in \{0, 1\}$ is the true label, and $s_{A,i}$ and $s_{B,i}$ are the prediction scores produced by methods A and B, respectively. Define the observed AUC difference as

$$\widehat{\Delta}_{\text{obs}} = \widehat{\text{AUC}}_A - \widehat{\text{AUC}}_B.$$

For each bootstrap replication $b = 1, \ldots, B$, sample n indices with replacement:

$$I^{(b)} = \left(i_1^{(b)}, \ldots, i_n^{(b)}\right), \qquad i_j^{(b)} \overset{\text{i.i.d.}}{\sim} \text{Uniform}\{1, \ldots, n\}.$$

Using the same sampled indices for both methods, construct

$$y^{(b)} = \left(y_{i_1^{(b)}}, \ldots, y_{i_n^{(b)}}\right),$$

$$s_A^{(b)} = \left(s_{A,i_1^{(b)}}, \ldots, s_{A,i_n^{(b)}}\right), \qquad s_B^{(b)} = \left(s_{B,i_1^{(b)}}, \ldots, s_{B,i_n^{(b)}}\right).$$

Table 7: **Cluster-count (K) sweep on LongLaMP** (setup, arms, metrics, highlighting, and significance markers as in Table 6). Metrics on the refusal-masked subset, as in Table 5. Note K is the number of clusters.

			GrRAG				$K{=}\lvert\mathcal{U}\rvert$
Task	**Metric**	**No Pers.**	$K{=}10$	$K{=}50$	$K{=}100$	$K{=}200$	**Summary ⊕ RAG**
Abstract Generation	ROUGE-1 ↑	0.378	**0.407**	0.398	0.395	0.393	0.400
	ROUGE-L ↑	0.197	**0.223**	0.212	0.207	0.205	0.204
	METEOR ↑	0.208	**0.264**	0.254	0.251	0.249	0.247
	Util† (1–5)	2.41	2.49	2.52	2.51	2.54	**2.71**
	GraSPeR-Judge‡ (1–7) ↑	**3.82**	3.37	3.19	3.15	3.15	3.32
	MIA-LLM ↓	0.501	0.920	0.921	0.922	0.924	0.940
	MIA-LR ↓	0.500	0.917	0.920	0.921	0.923	0.935
Review Writing	ROUGE-1 ↑	0.258	0.300	0.303	0.305	0.307	**0.340**
	ROUGE-L ↑	0.131	0.135	0.135	0.137	0.137	**0.154**
	METEOR ↑	0.114	0.187	0.189	0.188	0.190	**0.202**
	Util† (1–5)	1.19	1.19^{ns}	1.24	1.27	1.28	**1.63**
	GraSPeR-Judge‡ (1–7) ↑	**2.00**	1.75	1.77	1.75	1.71	1.90
	MIA-LLM ↓	0.504	0.739	0.739	0.750	0.745	0.828
	MIA-LR ↓	0.500	0.797	0.801	0.805	0.814	0.819
Topic Writing	ROUGE-1 ↑	0.289	0.269	0.267	0.268	0.269	**0.295**
	ROUGE-L ↑	0.124	0.123^{ns}	0.122^{ns}	0.122^{ns}	0.124^{ns}	**0.140**
	METEOR ↑	0.180	0.196	0.192	0.193	0.195	**0.210**
	Util† (1–5)	1.09	1.27	1.24	1.27	1.30	**1.70**
	GraSPeR-Judge‡ (1–7) ↑	1.63	1.52	1.52	1.50	1.55	**1.73**
	MIA-LLM ↓	0.503	0.730	0.767	0.771	0.784	0.912
	MIA-LR ↓	0.500	0.741	0.771	0.784	0.781	0.870

Then compute

$$\widehat{\mathrm{AUC}}_A^{(b)} = \mathrm{AUC}\left(y^{(b)}, s_A^{(b)}\right),$$

and

$$\widehat{\mathrm{AUC}}_B^{(b)} = \mathrm{AUC}\left(y^{(b)}, s_B^{(b)}\right).$$

The bootstrap AUC difference is

$$\widehat{\Delta}^{(b)} = \widehat{\mathrm{AUC}}_A^{(b)} - \widehat{\mathrm{AUC}}_B^{(b)}.$$

After B bootstrap replications, we obtain the empirical bootstrap distribution

$$\widehat{\Delta}^{(1)}, \ldots, \widehat{\Delta}^{(B)}.$$

A 95% percentile bootstrap confidence interval for the AUC difference is

$$\left[Q_{0.025}\left(\widehat{\Delta}^{(1:B)}\right),\ Q_{0.975}\left(\widehat{\Delta}^{(1:B)}\right)\right],$$

where Q_α denotes the empirical α-quantile of the bootstrap samples. If the confidence interval excludes zero, this provides evidence of a statistically significant difference between the two population AUCs at the 5% significance level. In particular, if the entire interval lies above zero, this supports the conclusion that method A has a higher AUC than method B.

C Cluster-BM25 Retrieval

Cluster-BM25 swaps the retrieval pool from the user's own profile to a *cluster-scoped* one. Each author u is assigned to one of K clusters $c(u)$ (Appendix D). The pool for u is the title-deduplicated union of the *training* profiles of every author in that cluster,

$$\mathcal{P}\big(c(u)\big) = \mathrm{dedup} \bigcup_{u' : c(u') = c(u)} P_{u'}^{\text{train}},$$

Table 8: **Personalization granularity and measurement choice.** Zero-shot Qwen3.5-9B; cohort $K = 100$, retrieval depth $k = 4$; Review and Topic Writing use the refusal mask. GraSPeR (1–7) is the primary reference-match instrument; Util (1–5) measures personal voice and ROUGE-L is a lexical reference. leak is five-way identification accuracy minus its matched query-only control. Rows distinguish document/population, summary/population, and document/within-cohort evidence. Document blocks are empty for the $k = 0$ abstraction arms and cannot support a no-leak conclusion. **Bold**: highest reference-match or lexical score. Red: largest attribution in a row above +0.02, a display rule rather than a significance threshold. —: not applicable. Audit qualifications, including the Review Writing generic-control failure, are given in Section G.

Task	Metric	No Pers.	Persona	Summary	C-RAG	O-RAG	GrRAG	Full Pers.
News Headline	GraSPeR (1–7) ↑	2.29	2.14	2.17	2.25	2.30	2.20	**2.31**
	ROUGE-L ↑	0.136	0.120	0.130	0.145	0.155	0.135	**0.157**
	Util (1–5)	2.19	2.19	2.47	2.30	2.55	2.50	2.90
	leak, documents ↓	+0.000	+0.002	+0.002	+0.021	+0.054	+0.047	+0.097
	leak, summaries ↓	+0.008	+0.069	+0.107	+0.021	+0.030	+0.057	+0.069
	leak, within cohort ↓	—	—	—	+0.036	+0.096	+0.015	+0.114
Scholarly Title	GraSPeR (1–7) ↑	**4.23**	3.82	4.00	4.13	4.14	4.03	4.10
	ROUGE-L ↑	0.418	0.339	0.359	0.423	**0.424**	0.388	0.408
	Util (1–5)	4.63	4.45	4.58	4.50	4.46	4.55	4.50
	leak, documents ↓	+0.002	+0.000	−0.001	+0.007	+0.013	+0.009	+0.007
	leak, summaries ↓	−0.021	−0.003	−0.011	−0.013	−0.012	−0.008	−0.005
	leak, within cohort ↓	—	—	—	+0.021	+0.028	+0.017	+0.022
Tweet Paraphrase	GraSPeR (1–7) ↑	5.95	4.88	5.52	5.94	**5.95**	5.22	5.63
	ROUGE-L ↑	0.374	0.313	0.362	0.379	0.380	0.351	**0.388**
	Util (1–5)	1.21	2.49	2.75	1.24	1.30	2.96	2.97
	leak, documents ↓	−0.001	+0.006	+0.005	+0.015	+0.020	−0.022	+0.122
	leak, summaries ↓	+0.016	+0.023	+0.252	+0.008	+0.025	−0.008	+0.221
	leak, within cohort ↓	—	—	—	+0.018	+0.035	−0.039	+0.129
Abstract Gen.	GraSPeR (1–7) ↑	**3.80**	2.68	2.90	3.58	3.59	2.99	3.32
	ROUGE-L ↑	0.197	0.177	0.182	0.203	**0.205**	0.191	0.202
	Util (1–5)	2.51	2.36	2.64	2.61	2.71	2.64	2.88
	leak, documents ↓	+0.000	−0.001	−0.000	−0.002	+0.063	+0.010	+0.048
	leak, summaries ↓	+0.033	+0.095	+0.141	+0.042	+0.053	+0.073	+0.095
	leak, within cohort ↓	—	—	—	+0.012	+0.117	−0.007	+0.093
Review Writing	GraSPeR (1–7) ↑	1.98	1.52	1.68	2.01	**2.03**	1.77	1.90
	ROUGE-L ↑	0.131	0.123	0.137	0.140	0.147	0.138	**0.154**
	Util (1–5)	1.30	1.06	1.26	1.32	1.43	1.17	1.46
	leak, documents ↓	−0.001	−0.001	+0.002	−0.030	−0.006	−0.027	+0.085
	leak, summaries ↓	−0.001	+0.101	+0.318	+0.015	+0.031	+0.092	+0.203
	leak, within cohort ↓	—	—	—	−0.003	+0.012	−0.031	+0.098
Topic Writing	GraSPeR (1–7) ↑	1.64	1.33	1.45	1.67	**1.70**	1.47	1.67
	ROUGE-L ↑	0.124	0.103	0.112	0.113	0.118	0.116	**0.126**
	Util (1–5)	1.19	1.15	1.28	1.18	1.30	1.21	1.45
	leak, documents ↓	+0.000	+0.002	−0.008	+0.001	+0.169	+0.039	+0.237
	leak, summaries ↓	+0.025	+0.231	+0.435	+0.027	+0.134	+0.188	+0.367
	leak, within cohort ↓	—	—	—	−0.016	+0.177	−0.039	+0.249

which includes u's own training writing alongside that of co-clustered authors. BM25 then retrieves the top-k articles from $\mathcal{P}(c(u))$. Relative to own-profile retrieval (the *fully personalized* setting $K{=}|\mathcal{U}|$, where the pool is P_u alone), the cluster pool dilutes any single user's identifiability while keeping on-topic, in-style exemplars. We fix $k{=}4$ and sweep $K \in \{10, 50, 100, 200\}$. Retrieval uses the BM25S engine (Robertson BM25, $k_1{=}1.5$, $b{=}0.75$) throughout — the same engine used for own-profile RAG and for MIA candidate construction, so retrieved passages are byte-consistent across generation and attack.

D Persona Clustering

Clusters and their persona descriptions are built from the **training** split only, in four stages:

Table 9: **Exact output-shift identity in response embeddings.** Zero-shot Qwen3.5-9B; unit-normalized Qwen3-Embedding-8B representations; Review and Topic Writing refusal-masked. With n the observed No Pers. output for the same item, $S = \mathbb{E}\|e(\hat{y}) - e(n)\|^2$ is mean squared shift and $A = \mathbb{E}\langle e(\hat{y}) - e(n), e(y^*) - e(n)\rangle$ is target alignment. $G = 2A - S$ is the squared-distance gain relative to No Pers.; the underlying identity holds item by item. The reference is a surrogate target. These statistics include decoding variation and do not estimate pure between-user variance or covariance. **Bold**: largest gain per task. A larger gain need not require a larger shift. Shift is weakly associated with the measured attribution in this panel (Spearman $+0.09$ across 42 cells, $+0.02$ per item); it is not a privacy metric.

Task	Term	No Pers.	Persona	Summary	C-RAG	O-RAG	GrRAG	Full Pers.
News Headline	shift S	0.000	0.336	0.328	0.262	0.263	0.326	0.311
	alignment A	0.000	0.162	0.164	0.130	0.139	0.160	0.162
	gain $G = 2A - S$	$+0.000$	-0.011	-0.001	-0.002	$+0.016$	-0.006	$+0.013$
Scholarly Title	shift S	0.000	0.125	0.114	0.119	0.121	0.125	0.120
	alignment A	0.000	0.049	0.048	0.058	0.059	0.057	0.057
	gain $G = 2A - S$	$+0.000$	-0.027	-0.019	-0.003	-0.003	-0.011	-0.006
Tweet Paraphrase	shift S	0.000	0.357	0.255	0.125	0.123	0.314	0.247
	alignment A	0.000	0.155	0.130	0.064	0.064	0.163	0.139
	gain $G = 2A - S$	$+0.000$	-0.048	$+0.004$	$+0.003$	$+0.005$	$+0.012$	$+0.030$
Abstract Generation	shift S	0.000	0.269	0.253	0.206	0.214	0.242	0.236
	alignment A	0.000	0.111	0.109	0.109	0.117	0.122	0.130
	gain $G = 2A - S$	$+0.000$	-0.046	-0.034	$+0.012$	$+0.020$	$+0.002$	$+0.025$
Review Writing	shift S	0.000	0.566	0.497	0.370	0.380	0.482	0.459
	alignment A	0.000	0.237	0.221	0.181	0.193	0.247	0.258
	gain $G = 2A - S$	$+0.000$	-0.092	-0.055	-0.007	$+0.006$	$+0.012$	$+0.057$
Topic Writing	shift S	0.000	0.764	0.753	0.815	0.842	0.756	0.779
	alignment A	0.000	0.352	0.369	0.423	0.447	0.382	0.426
	gain $G = 2A - S$	$+0.000$	-0.061	-0.015	$+0.030$	$+0.053$	$+0.007$	$+0.074$

Table 10: **The tradeoff across five generators, three model families, and a $4\times$ scale range.** Each generator writes its own summaries, personas, and clustering; prompts, decode, and the Qwen3.6-35B-A3B attacker are shared (cross-family for Llama and Gemma). Reference lift = mean ROUGE-L gain over **No Pers.** across the six tasks; leak = mean leak; the last column counts the task×mechanism pairs (12 per generator) in which the cohort strategy has lower measured attribution than its own-profile counterpart under document candidates. Counts indicate directions, not statistical significance. Per-task panels: Tables 4 and 5. Review Writing and Topic Writing use the refusal-masked panel (n=1774/2248).

	Lift ↑		leak ↓		Cohort
Generator	RAG	CRAG	RAG	CRAG	**leaks less**
Llama-3.1-8B	$+0.010$	$+0.008$	$+0.095$	-0.003	12/12
Qwen3.5-9B	$+0.008$	$+0.004$	$+0.052$	$+0.002$	11/12
Gemma-3-12B	$+0.019$	$+0.015$	$+0.055$	$+0.003$	12/12
Qwen3.8-27B	$+0.021$	$+0.013$	$+0.056$	$+0.006$	10/12
Qwen3.6-35B-A3B	$+0.007$	$+0.001$	$+0.068$	$+0.005$	12/12

1. **Summarize.** Each author is represented by their largest profile; the generator used for that experiment condenses a sample of it (the 20 most-recent plus 20 random earlier documents) into a $\sim$200-word summary.
2. **Embed.** Each *summary* (not the raw profile) is encoded with Qwen3-Embedding-8B.
3. **Cluster.** k-means on the summary embeddings yields the assignment $c(u)$, for each $K \in \{10, 50, 100, 200\}$.
4. **Describe.** For each cluster, the same LLM turns up to 30 sampled member summaries into a persona *name* and a $\sim$200-word *description* of the group (Table 15); the description is injected into the prompt as the persona block (Appendix E).

Table 11: **The granularity knob under two instruments.** Own-stacked minus cohort-stacked at $K = 100$. **ΔAUC** is the difference the convention reports; **Δleak** is the output-attributable difference. **Bold** marks the three tasks where the readings diverge most. Differences are computed from unrounded values. Scholarly Title illustrates that a high raw AUC can coexist with a small incremental attribution difference.

Task	AUC_{coh}	AUC_{own}	ΔAUC	Δleak
News Headline	0.907	0.926	0.019	0.050
Scholarly Title	0.952	0.956	0.004	−0.002
Tweet Paraphrase	0.683	0.806	**0.123**	**0.144**
Abstract Gen.	0.923	0.939	0.016	0.038
Review Writing	0.742	0.827	**0.085**	**0.112**
Topic Writing	0.776	0.911	**0.134**	**0.198**

Table 12: **One spectrum, three attackers** (zero-shot Qwen3.5-9B; cohort K=100, retrieval depth 4). leak = top-1 re-identification minus its query-only control, under three five-way attackers that differ only in what the adversary holds about each candidate user: **documents, population** — the candidate's own top-4 retrieved documents, four random other users (the deployed instrument); **summaries, population** — the candidate's user summary, four random other users (the persona threat model; the only instrument under which the abstraction strategies present non-empty candidates); **documents, within cohort** — documents, four other users drawn from the true user's K=100 cohort (items whose cohort has fewer than four eligible users are dropped: 5.8% on News Headline, $\leq$ 0.5% elsewhere). Review Writing and Topic Writing use the refusal-masked panel. Red: most-leaking strategy per row. — : not run.

Task	Attacker holds	No Pers.	Persona	Summary	CRAG	RAG	GrRAG	Full Pers.
News Headline	documents, population	+0.000	+0.002	+0.002	+0.021	+0.054	+0.047	+0.097
	summaries, population	+0.008	+0.069	+0.107	+0.021	+0.030	+0.057	+0.069
	documents, within cohort	—	—	—	+0.036	+0.096	+0.015	+0.114
	summaries, within cohort	+0.016	−0.006	+0.183	—	—	—	—
Scholarly Title	documents, population	+0.002	+0.000	−0.001	+0.007	+0.013	+0.009	+0.007
	summaries, population	−0.021	−0.003	−0.011	−0.013	−0.012	−0.008	−0.005
	documents, within cohort	—	—	—	+0.021	+0.028	+0.017	+0.022
	summaries, within cohort	+0.003	−0.011	+0.013	—	—	—	—
Tweet Paraphrase	documents, population	−0.001	+0.006	+0.005	+0.015	+0.020	−0.022	+0.122
	summaries, population	+0.016	+0.023	+0.252	+0.008	+0.025	−0.008	+0.221
	documents, within cohort	—	—	—	+0.018	+0.035	−0.039	+0.129
	summaries, within cohort	+0.011	−0.030	+0.281	—	—	—	—
Abstract Gen.	documents, population	+0.000	−0.001	−0.000	−0.002	+0.063	+0.010	+0.048
	summaries, population	+0.033	+0.095	+0.141	+0.042	+0.053	+0.073	+0.095
	documents, within cohort	—	—	—	+0.012	+0.117	−0.007	+0.093
	summaries, within cohort	+0.016	−0.029	+0.229	—	—	—	—
Review Writing	documents, population	−0.001	−0.001	+0.002	−0.030	−0.006	−0.027	+0.085
	summaries, population	−0.001	+0.101	+0.318	+0.015	+0.031	+0.092	+0.203
	documents, within cohort	—	—	—	−0.003	+0.012	−0.031	+0.098
	summaries, within cohort	−0.040	−0.077	+0.478	—	—	—	—
Topic Writing	documents, population	+0.000	+0.002	−0.008	+0.001	+0.169	+0.039	+0.237
	summaries, population	+0.025	+0.231	+0.435	+0.027	+0.134	+0.188	+0.367
	documents, within cohort	—	—	—	−0.016	+0.177	−0.039	+0.249
	summaries, within cohort	+0.034	−0.073	+0.548	—	—	—	—

To prevent train–test leakage, every test query is mapped to *its author's train cluster*, and test authors absent from the training split are dropped, so no test text ever enters a cluster or persona it could later be attacked with.

E PROMPTS

This appendix collects the prompt templates: the per-user *summary* and per-cluster *persona* prompts that build the personas (Appendix D), and the *personalization* prompts used at generation time; the

Table 13: **Abstraction arms under the summary-holding attacker, every generator.** leak (top-1 minus the query-only control) when the attacker holds each candidate user's 200-word summary (the persona system's own view) and decoys are four random other users; k=0 arms present empty blocks to the deployed document attacker and therefore do not provide a valid no-leak test under that evidence protocol. Each generator writes its own summaries and personas; Review and Topic Writing on each generator's refusal-masked panel.

Generator	Arm	News	Title	Tweet	Abstract	Review	Topic
Qwen3.5-9B	No Pers.	+0.008	−0.021	+0.016	+0.033	−0.001	+0.025
	Persona	+0.069	−0.003	+0.023	+0.095	+0.101	+0.231
	Summary	+0.107	−0.011	+0.252	+0.141	+0.318	+0.435
Llama-3.1-8B	No Pers.	+0.000	−0.011	+0.003	+0.024	−0.016	+0.005
	Persona	+0.053	−0.021	−0.013	+0.052	+0.042	+0.116
	Summary	+0.069	−0.004	+0.053	+0.085	+0.110	+0.286
Gemma-3-12B	No Pers.	+0.004	−0.017	−0.002	+0.027	−0.003	+0.009
	Persona	+0.064	−0.001	−0.039	+0.052	+0.089	+0.120
	Summary	+0.077	+0.003	+0.069	+0.071	+0.240	+0.387
Qwen3.8-27B	No Pers.	−0.002	−0.015	+0.001	+0.024	−0.011	+0.008
	Persona	+0.052	+0.003	−0.020	+0.074	+0.118	+0.195
	Summary	+0.061	−0.006	+0.149	+0.092	+0.302	+0.392
Qwen3.6-35B-A3B	No Pers.	+0.001	−0.011	+0.001	+0.025	−0.002	+0.026
	Persona	+0.086	+0.000	+0.014	+0.085	+0.106	+0.216
	Summary	+0.098	+0.001	+0.251	+0.113	+0.312	+0.429
9B cohort-SFT (12k)	No Pers.	+0.008	−0.009	−0.017	+0.030	−0.042	+0.004
	Persona	+0.025	−0.009	−0.009	+0.038	−0.017	+0.078
	Summary	+0.047	−0.011	+0.003	+0.040	−0.008	+0.091
9B cohort-SFT (75k)	No Pers.	+0.003	−0.015	−0.019	+0.032	−0.055	+0.003
	Persona	+0.031	−0.015	−0.016	+0.033	−0.013	+0.063
	Summary	+0.047	−0.011	−0.003	+0.040	+0.003	+0.093
9B cohort-RAG-SFT	No Pers.	−0.010	−0.023	−0.013	+0.033	−0.030	−0.004
	Persona	+0.025	−0.014	−0.010	+0.032	−0.034	+0.065
	Summary	+0.027	−0.015	+0.005	+0.040	+0.005	+0.093

Table 14: **Tail risk: share of users (%) identified from the output.** Per-user mean of leak items ≥ 0.5 under the deployed document-holding attacker (zero-shot Qwen3.5-9B; Review Writing and Topic Writing refusal-masked). **Null** is the same share for a response that carries no profile signal (the No Pers. output planted under the k=4 candidate blocks): the attacker's decision rule alone "identifies" 4–11% of users. Cohort arms sit at or near the null; the user's own profile does not. Most tasks have one test item per user, so this is also the item-level share.

Task	Null	Persona	Summary	C-RAG	O-RAG	GrRAG	O-Stack
News Headline	9.5	0.6	1.0	7.5	11.9	11.4	15.6
Scholarly Title	5.7	0.3	0.3	5.5	6.7	6.5	6.1
Tweet Paraphrase	11.0	3.5	3.4	11.6	13.1	12.9	21.8
Abstract Gen.	5.7	0.0	0.1	6.1	11.8	6.8	10.2
Review Writing	8.6	0.9	2.4	8.9	10.7	11.4	18.3
Topic Writing	9.8	1.9	3.1	10.6	23.8	16.4	30.1

attacker prompt is in Figure 4. All prompts use the Qwen3/3.5 chat template, with each turn marked `[system]` or `[user]`; inputs appear as placeholders (e.g. `{histories}`, `{task query}`) rather than benchmark text.

Persona-construction prompts. Figure 5 shows the two prompts behind Appendix D: (a) summarize one user from their own writing; (b) turn a sample of member summaries into one named

[user]

```
You are a privacy auditor testing whether a personalized AI assistant's output reveals which user it was
written for.

Below is a request given to the assistant and the response it generated for one specific user, followed by 5
candidate users, each represented by a sample of their own past writing.

Request:
{task query}

Response generated for the target user:
{generated output}

Candidate users:
[A] {sample of candidate A's past writing}

[B] {sample of candidate B's past writing}

[C], [D], [E] -- three further candidates in the same format

Which candidate user did the assistant most likely generate the response for? Respond with only the single
letter of the correct option and nothing else.
```

Figure 4: **MIA-LLM attacker prompt** for $N{=}5$-way re-identification (Appendix B): a single Qwen `[user]` turn, no system message. `{task query}`, `{generated output}`, and the per-candidate writing samples are inputs; `[A]`–`[E]` are the option labels. The *query-only* control drops the "Response generated for the target user" block and instead asks "Which candidate user most likely issued the request?".

Table 15: Example cluster personas (name + description) at $K{=}100$, for a short-text task (LaMP-4) and a long-text task (LongLaMP-3). Descriptions are the $\sim$200-word cluster-persona outputs, abbreviated here; at generation time each is the persona block of the personalization prompt.

Persona name	**Description (abbreviated)**
LaMP-4 (news-headline generation)	
Digital Fashion & Celebrity Content Editor	Works within lifestyle and entertainment publications; documents the sartorial choices, public appearances, and style evolutions of high-profile figures.
U.S. Political News Analyst	Political journalists and analysts on U.S. domestic and foreign policy, concentrated on the 2016 election cycle and the early Trump administration.
Entertainment News Content Creator	Specializes in entertainment news, celebrity culture, and pop-culture commentary; synthesizes interviews and updates into concise, headline-driven narratives.
LongLaMP-3 (product-review writing)	
The Pragmatic Fit Analyst	A detail-oriented, data-driven apparel shopper who prioritizes precise physical fit, fabric durability, and functional utility over fleeting fashion trends.
The Genre-Savvy Mystery Critic	An analytical reviewer of crime fiction — classic detective stories, police procedurals, Nordic noir, and supernatural mysteries — with encyclopedic author knowledge.
The Pragmatic Footwear Analyst	Consumers focused on the functional performance, precise fit, and material integrity of footwear rather than trends or brand prestige.

persona. The resulting name and description become the persona block of the personalization prompt (Figure 6).

Personalization prompts. Every strategy in Tables 6 and 7 assembles the model input from two optional blocks placed *before* the paper-faithful task query: a *persona block* (who the user is) and a *retrieval block* (examples of the user's writing). The anonymization knob K (the number of user clusters, Appendix D) sets both blocks jointly:

- **No personalization.** Neither block; the model sees the bare task query.

(a) User-summary prompt — condense one user into a $\sim$200-word summary of their own writing.

[system]

```
Summarize a user from examples of past writing. Be factual, specific, neutral, and avoid speculation.
```

[user]

```
## Context
Past writing examples:

{histories}

## Instruction
Based on the past writing examples, write a summary for this user, about 200 words, covering recurring
interests, tasks, tools, expertise, and writing style, with specific examples.

## Output format
Return the user summary in triple backticks, i.e., ```user summary: <detailed summary>```.
```

(b) Cluster-persona prompt — write one persona per cluster from a sample of its members' summaries.

[system]

```
You describe the shared persona of a group of similar users. Be factual, specific, neutral, and avoid
speculation.
```

[user]

```
## Context
Summaries of users in the same group:

{members}

## Instruction
These users were clustered together as one user type. Write a persona for this user type: a short descriptive
name, and a detailed description of what this group has in common -- the shared interests, tasks, tools,
expertise, and writing style that define this user type. Describe the group as a persona (a type of user),
not any individual user. The description should be about 200 words.

## Output format
Return the persona in triple backticks as two labeled lines, i.e.,
```
persona name: <a short, descriptive name for this user type>
persona description: <about 200-word description>
```
```

Figure 5: **Persona-construction prompts** (Appendix D), each a Qwen `[system]`/`[user]` chat pair. (a) each user is summarized from their own writing (`{histories}`); (b) each cluster is turned into a named persona from its members' summaries (`{members}`). Example generated persona names and descriptions are in Table 15.

- **GrRAG (K clusters).** The persona block is the cluster description of $c(u)$ (a user *type*); the retrieval block holds the top-k (k=4) cluster-BM25 passages from the group's training writing (Appendix C). Coarser K abstracts away individual detail; finer K approaches the individual.
- **Fully personalized (K=$|\mathcal{U}|$).** Each user is their own cluster: the persona block is the user's *own* profile summary and the retrieval block is BM25 over their *own* writing.

Each block is independently optional; the four combinations give the arms of Table 4: **No personalization / Persona / RAG / GrRAG**. Figure 6 gives the template; retrieved examples and the task query follow the LaMP/LongLaMP prompts, vendored verbatim, and example personas are in Table 15.

[user]

Persona block — optional; present for the Persona and GrRAG strategies:

```
This user belongs to persona "{persona name}" : {persona description}.
```

Retrieval block — optional; present for the RAG and GrRAG strategies. The retrieved examples and the task query are rendered by the paper-faithful LaMP/LongLaMP template:

```
Here are examples for this user type: {retrieved profile examples}. {task query}
```

Figure 6: **Personalization prompt** template: one Qwen `[user]` turn with two *optional* leading blocks before the paper-faithful task query. The strategy selects which blocks appear (Table 4): **No personalization** the query only, **Persona** adds the persona block, **RAG** the retrieval block, and **GrRAG** both. `{persona name}`/`{persona description}` come from Figure 5 (examples in Table 15); `{retrieved profile examples}` and `{task query}` follow the benchmark template. For Fully-personalized the lead-ins become "About the user: ..." and "examples of the user's own writing".

Table 16: The three judge criteria (the separate Utility prompt is in Figure 7). Utility measures how well the response matches the user's personal voice.

Criterion	Measures (role)	Evidence
Content Match	matches the key content of the reference	reference
Relevance	tailored to this user (personalization)	profile, task
Utility	reads in the user's voice	past writing

F LLM-AS-A-JUDGE

The main results use separate-criterion Utility scoring (Figure 7). The bundled rubric jointly scores Content Match, Relevance, and Utility; bundled results are retained only in explicitly labeled supplementary comparisons. Validation tests are in Appendix G.1.

Criteria. The criteria (Table 16) assess content against the reference and relevance and utility against the user's writing. Utility (Util) measures personal voice on a 1–5 scale. Privacy is measured separately by user attribution (Appendix B).

Deployed judge. We score the reported tables with Qwen3.6-35B-A3B.

Content Match and relevance scoring. On News Headline, Scholarly Title, and Abstract Generation, the bundled Content Match and relevance scores can favor task-grounded answers over reference-faithful answers, including the reference itself. We therefore report these criteria only in the validation comparisons. Util measures personal voice, while ROUGE/METEOR and GraSPeR provide separate reference-based measurements.

G METRIC VALIDATION

We evaluate metrics using controlled changes to response content and user profiles. The conditions are ORACLE (the gold answer), ACTUAL (the generated answer), WRONG_CONTENT (another item's gold), WRONG_USER (another user's writing as the profile), and GENERIC (fixed boilerplate). The bundled judge comparisons use $N = 100$ per condition; separate-criterion profile-discrimination tests use $N = 60$ per condition; reference-metric and attacker tests use $N = 50$ per condition on LaMP-4 and LongLaMP-2. Tables 18, 19, and 20 report these controls. Table 17 summarizes the content, personalization, and attribution checks, including the additional tests described below. Passing a control does not establish validity across all tasks or scoring configurations.

You are an expert evaluator of personalized text generation, scoring one CANDIDATE output on ONE criterion. You are given the USER'S OWN PAST WRITING and the TASK.

Utility — Does the candidate read as if THIS user wrote it (tone, vocabulary, sentence structure, formatting), judged against their past-writing samples? Ignore topic overlap. (1 = clearly a different author; 3 = some shared traits; 5 = indistinguishable from this user.)

Write a 1-2 sentence justification citing specific evidence, then an integer 1-5.

Return ONLY a JSON object inside a triple-backtick fence, e.g.:

```
{"justification": "...", "score": <1-5>}
```

USER'S OWN PAST WRITING: `{profile}`
TASK: `{task}`
CANDIDATE: `{candidate}`

Figure 7: **Separate-criterion Utility prompt used for the main results.** One user message scores personal voice without a reference answer or other criteria. The profile contains up to four past writings; the task and candidate are passed in full. The criterion label is displayed using the paper's terminology.

Table 17: **Metric validation summary.** Each cell reports whether the instrument satisfies the required ordering of Section G on that axis. **Content** and **Personal.** are the quality criteria; **Attack** is the planted-response criterion. **Bold** marks failures, which disqualify the instrument from supporting directional claims on that axis. Pass/fail refers to the specified controls, not universal validity. Reference overlap, judged personal voice, and user attribution are distinct measurements. † fails the known-content ordering (ranks a verified-wrong, wording-preserving answer at or above a verified-right paraphrase on 68–100% of items); retained as a dual-reported reference column, paired within-task only. § GENERIC control fails on Review Writing (-0.30); paired orderings only on that task. ¶ qualifies on the profile-discrimination test only ($5/6$ per judge model, disjoint weak cells: Scholarly Title, Topic Writing); a proxy for personal voice, not a direct measure of user benefit.

Instrument	Content	Personal.	Attack
ROUGE-1/L, METEOR	**fail**†	n/a	n/a
Judge: Content Match	**fail**	n/a	n/a
Judge: Relevance	**fail**	n/a	n/a
Judge: Util (separate)	n/a	$5/6$¶	n/a
GraSPeR ref-match	pass	pass	n/a
Cross-family judge	pass	pass	n/a
Embedding + classifier	n/a	n/a	**fail**
LLM re-ID, raw AUC	n/a	n/a	**fail**
LLM re-ID, leak	n/a	n/a	pass§

G.1 LLM JUDGE (CONTENT MATCH / RELEVANCE / UTILITY)

Content Match scores should rank ORACLE and ACTUAL above WRONG_CONTENT; relevance and utility should rank ACTUAL above WRONG_USER and ORACLE above GENERIC. A check passes at a score gap of ≥ 0.25, and a rubric passes a task when all six checks pass. Table 18 gives the per-condition means.

On a six-task profile-discrimination test, bundled scoring fails on $4/6$ tasks with Qwen3.6-35B-A3B and $5/6$ with Qwen3.8-27B. Scoring each criterion separately passes relevance checks on $6/6$ tasks and utility checks on $5/6$ tasks for each model, with different weak tasks. Comparing scoring configurations changes $13/126$ pairwise arm orderings, although the highest-scoring arm is unchanged on all six tasks. The main table and granularity figure use separate-criterion Utility from Qwen3.6-35B-A3B. Its profile-discrimination check passes on five of six strongly personalized task panels; Scholarly Title falls below the 0.25 margin, so its Utility scores do not support claims about

Table 18: **Judge validation — per-condition means** (N=100/condition; LaMP-4, LongLaMP-2). Each cell gives mean 1–5 scores under the corrupted inputs: **Content Match** as oracle/actual/wrong_content; **Relevance** and **Util** as actual/wrong_user/oracle/generic. A check passes at gap $\geq$0.25 (Content Match: oracle, actual > wrong_content; relevance/Util: actual > wrong_user and oracle > generic); a judge passes a task at **6/6** (bold). Qwen3.6-35B* is the deployed judge model; the entries here use the bundled rubric. [§]LaMP-4 relevance actual−wrong_user = 0.249, below the 0.25 threshold. Test definitions in §G.1.

Judge	Task	Content Match or/act/wc	Relevance act/wu/or/gen	Util act/wu/or/gen	Verdict
Sonnet-4.6	LongLaMP-2	5.00/3.41/1.00	3.06/1.75/3.91/1.00	2.45/2.02/3.23/1.00	**6/6**
	LaMP-4	5.00/2.89/1.00	2.76/2.12/3.94/1.00	2.40/1.69/3.70/1.00	**6/6**
Qwen3.6-35B*	LongLaMP-2	5.00/3.54/1.00	3.72/3.42/4.22/1.00	2.55/2.19/1.97/1.00	**6/6**
	LaMP-4	4.77/3.57/1.00	3.62/3.37/4.47/1.00	2.30/1.37/3.09/1.00	5/6[§]
GPT-5.4-mini	LongLaMP-2	4.84/3.80/1.00	4.28/4.14/4.87/1.00	3.51/3.36/4.56/1.00	4/6
	LaMP-4	4.10/3.49/1.00	4.17/4.16/4.45/1.00	3.52/2.97/4.28/1.00	5/6
GPT-5.4 (r=none)	LongLaMP-2	4.99/3.59/1.00	3.51/3.41/4.02/1.00	2.94/2.84/4.19/1.00	4/6
	LaMP-4	5.00/3.02/1.00	3.21/2.99/4.48/1.00	2.75/2.33/4.63/1.00	5/6

Table 19: **Reference-metric validation — per-condition means** (N=50/condition; conditions as defined in the section text). Checks (three per task): ORACLE > WRONG_CONTENT, ACTUAL > WRONG_CONTENT, ORACLE > GENERIC — at margin $\geq$0.25 for GraSPeR-Judge, by one-sided paired t for ROUGE-1/L/METEOR (the WRONG_CONTENT floor is task-dependent). Companion to Table 18.

Metric	Task	ORACLE	ACTUAL	WRONG_C.	GENERIC	Verdict
GraSPeR-Judge (1–7)	LongLaMP-2	7.00	3.00	1.00	1.00	**3/3**
	LaMP-4	7.00	2.22	1.00	1.00	**3/3**
ROUGE-1 (0–1)	LongLaMP-2	1.000	0.377	0.205	0.083	**3/3**
	LaMP-4	1.000	0.201	0.028	0.053	**3/3**
ROUGE-L (0–1)	LongLaMP-2	1.000	0.194	0.117	0.060	**3/3**
	LaMP-4	1.000	0.173	0.028	0.051	**3/3**
METEOR (0–1)	LongLaMP-2	1.000	0.252	0.123	0.030	**3/3**
	LaMP-4	0.999	0.135	0.018	0.044	**3/3**

personalization benefit. The corresponding Qwen3.8-27B check fails on Topic Writing, limiting cross-judge robustness on that task. These checks do not establish agreement with human preferences.

G.2 GRASPER-JUDGE, ROUGE-1/L, AND METEOR

All four reference-based metrics pass the three basic content checks on both tasks (Table 19). More demanding tests compare verified content-preserving paraphrases, wording-preserving content corruptions, and different valid answers, using separate model families to construct, verify, and score the examples. ROUGE ranks the incorrect high-overlap answer above the correct paraphrase on 68%, 70%, and 100% of items across three tasks; GraSPeR passes all nine checks. ROUGE and METEOR are therefore interpreted as reference-overlap measures rather than general measures of content correctness.

G.3 PRIVACY ATTACKERS (MIA-LLM AND MIA-LR)

Planted responses test whether the attacker uses the output rather than only query–profile similarity (Table 20). A target user's document produces positive leak; a decoy's document redirects identification toward that decoy. An uninformative response should leave performance near the matched query-only control, rather than the 0.2 chance level, because candidate evidence is query-conditioned.

Table 20: **Initial planted-response controls** (N=50/condition; original gate verdicts retained). The expanded audit in Table 17 governs the interpretation; a generic increment of -0.080 is not exactly zero. Planted responses: OWN (verbatim document from the true user's profile), DECOY (verbatim document from a decoy candidate), GENERIC (boilerplate). MIA-LLM rows report leak (pass: > 0 on OWN; $\ll 0$ on DECOY, i.e. the attacker must flip to the planted candidate; ≈ 0 on GENERIC). MIA-LR rows report AUC over 5 split seeds (pass: high on OWN, low on DECOY, near chance on GENERIC). The query alone identifies the user at top-1 0.66–0.73 among query-retrieved candidates, which bounds the OWN uplift.

Attack (statistic)	Task	OWN	DECOY	GENERIC	Verdict
MIA-LLM (leak)	LongLaMP-2	$+0.040$	-0.640	-0.080	**3/3**
	LaMP-4	$+0.060$	-0.720	-0.080	**3/3**
MIA-LR (AUC)	LongLaMP-2	0.992	1.000	0.890	1/3
	LaMP-4	1.000	1.000	0.928	1/3

Across six tasks, decoy controls yield true-user top-1 accuracy ≤ 0.04. Generic-response controls remain near the null on five tasks but produce a -0.30 shift on Review Writing, limiting interpretation of absolute attribution increments on that task (Table 21).

The embedding-classifier attack fails these controls: fitting on the evaluated condition yields AUC 1.0 for decoy responses and approximately 0.9 for generic responses. It does not support our privacy conclusions. Privacy comparisons use the LLM attacker's matched accuracy increment; raw AUC is reported separately.

Table 21: **Planted-response attacker validation, per task** (n=50 per condition). OWN: the user's verbatim documents planted as the response (attacker must identify the user). DECOY: another user's verbatim documents (attacker must flip; true-user top-1 must fall to the floor). GENERIC: uninformative boilerplate; query-only is that condition's own paired control, so Δ = GENERIC top-1 $-$ query-only (must be ≈ 0) reproduces from the printed columns. **Bold** marks the one failing cell.

Task	OWN top-1	DECOY top-1	GENERIC top-1	query-only	GENERIC Δ
News Headline	0.78	0.02	0.64	0.72	-0.08
Scholarly Title	0.88	0.02	0.74	0.86	-0.12
Tweet Paraphrase	0.86	0.02	0.42	0.38	$+0.04$
Abstract Gen.	0.70	0.04	0.60	0.68	-0.08
Review Writing	0.94	0.02	0.26	0.56	$\mathbf{-0.30}$
Topic Writing	0.98	0.00	0.36	0.50	-0.14

H Supervised Fine-Tuning Details

The SFT rows in Table 22 all come from a single **full-parameter** fine-tune of Qwen3.5-9B on a balanced multi-task pool (MT6: all six tasks, $\sim$2k examples each, pooled from the per-task inference caches) — one model that serves every task, giving a clean comparison against the frozen base. Each training example pairs the *same* persona $\oplus$ RAG input used at inference (the K=100 cluster persona description $\oplus$ the k=4 cluster-BM25 exemplars; Appendix E) with the gold reference as the completion; loss is over completion tokens only. Optimization: 1 epoch, learning rate $1e{-5}$, cosine schedule, bf16, effective batch size 32 (per-device batch $1 \times$ gradient accumulation 16×2 GPUs), max sequence length 6144, DeepSpeed ZeRO-3. The fine-tuned weights are served directly (no adapter).

Gains on all six tasks, with a long-form looping caveat. Pooled full-parameter SFT lifts ROUGE-L on *all six* tasks (avg $+0.052$, up to $+0.098$ on LaMP-7; every gain $p < 10^{-5}$, sign-flip permutation on matched items) at flat-or-lower re-identification leakage (Table 22): the per-item attacker P(true user) is significantly *lower* after SFT on four of six tasks and never significantly higher. These observations are consistent with the distinction between overall utility improvement and user-

Table 22: **Zero-shot vs. full-parameter multi-task (MT6) SFT under GrRAG.** Frozen Qwen3.5-9B (*ZS*) vs. **full-parameter** SFT on the six-task MT6 pool ($\sim$2k examples/task, cluster-stacked $K{=}100$, $k{=}4$, 1 epoch, ZeRO-3; Appendix H), evaluated on the cluster-stacked ($K{=}100$, $k{=}4$) arm, test split. Both rows share the same decode ($T{=}0.7$, nucleus 0.95, seed 42) and profile frontend, so neither confounds the contrast. RESPONSE MEASURES: R-1/R-L/MET = ROUGE-1/L/METEOR; GrSP = GraSPeR-Judge (Ni et al., 2026) (1–7 reference similarity, the primary content measure). ROUGE/METEOR are lexical reference measures. MIA-LLM AUC is secondary; leak = top-1 minus the query-only control (output-attributable leakage). **Bold**: best per task. ^{ns}: SFT not significantly different from ZS on that metric (quality and leak: two-sided sign-flip permutation test on matched items, $p \geq 0.05$; MIA-LLM: paired bootstrap on the AUC difference, 95% CI covering 0, Appendix B; refusal-masked matched items).

Task	Model	QUALITY ↑				PRIVACY ↓	
		R-1	R-L	MET	GrSP	MIA-LLM	leak
News Headline (LaMP-4)	Qwen3.5-9B (ZS)	0.158	0.139	0.106	**2.19**	0.912	0.062
	Qwen3.5-9B (SFT)	**0.184**	**0.166**	**0.108**^{ns}	1.80	0.900	0.047^{ns}
Scholarly Title (LaMP-5)	Qwen3.5-9B (ZS)	0.492	0.413	**0.429**	**4.17**	0.951	0.008
	Qwen3.5-9B (SFT)	**0.539**	**0.476**	0.403	4.13^{ns}	0.951^{ns}	0.007^{ns}
Tweet Paraphrase (LaMP-7)	Qwen3.5-9B (ZS)	0.403	0.350	0.287	5.20	0.681	−0.021
	Qwen3.5-9B (SFT)	**0.502**	**0.446**	**0.361**	**5.66**	0.694	−0.001^{ns}
Abstract Generation (LongLaMP-2)	Qwen3.5-9B (ZS)	0.395	0.207	0.251	3.15	0.922	0.015
	Qwen3.5-9B (SFT)	**0.460**	**0.291**	**0.296**	**3.25**	0.916	0.006^{ns}
Review Writing (LongLaMP-3)	Qwen3.5-9B (ZS)	0.305	0.137	**0.188**	**1.75**	0.750	−0.030
	Qwen3.5-9B (SFT)	**0.319**	**0.154**	0.172	1.70	**0.721**	**−0.059**
Topic Writing (LongLaMP-4)	Qwen3.5-9B (ZS)	0.268	0.122	**0.193**	1.50	0.771	0.039
	Qwen3.5-9B (SFT)	**0.279**	**0.150**	0.165	1.50^{ns}	0.767^{ns}	0.033^{ns}

matching gain in Appendix I.5; the theory does not guarantee that fine-tuning leaves identity leakage unchanged.

One caveat remains: on the long-form tasks LongLaMP-3/4, 8.6–11.4% of SFT outputs run to the generation cap (repetition looping, vs. 0.4–1.5% for the frozen base at the same sampled decode). Under greedy decoding this looping affected $\sim$75–90% of long-form outputs and collapsed ROUGE—an exposure-bias interaction between full-parameter updates and deterministic decoding, not a property of the data (the same pool trained as a frozen-base LoRA adapter did not loop, and canonical nucleus sampling bounds it). We report at-cap rates alongside utility so the residual effect stays visible.

I Proofs and Supplementary Theory

All statements fix x and the observer's background information. We use finitely many users and, for the displayed sums, a finite vocabulary and response-length limit. Terms with zero response marginal are omitted; the corresponding integral arguments apply when the relevant conditional distributions exist. Only users with positive conditional prior are included. Expectations over queries use the appropriate prior conditional on each query.

I.1 Notation

Table 23: Notation at a fixed query x. Bold lowercase symbols denote random variables; k_G is group size, not retrieval depth.

Symbol	Meaning
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	Query, response, and user-context spaces
$\mathbf{u} \in \{1, \dots, n\}$	Intended user identity
$\mathbf{x}, \mathbf{y}, \mathbf{z} = z_{\mathbf{u}}$	Random query, response, and user context
$z_i,\ \phi : \mathcal{Z} \to \mathcal{Z}$	User i's context and deterministic context processing
$\pi(y \mid x, z),\ \pi_i(y)$	Fixed LLM response distribution; $\pi_i(y) = \pi(y \mid x, \phi(z_i))$ for a context-processing mechanism
$\bar{\pi}(y)$	Population response distribution $\sum_i \mathbb{P}(\mathbf{u} = i \mid x)\pi_i(y)$
$r_i(x, y) \in [0, 1]$	User i's utility
$V_i,\ V$	$V_i = \mathbb{E}_{\pi_i} r_i(x, \mathbf{y})$; $V = \sum_i \mathbb{P}(\mathbf{u} = i \mid x)V_i$
$\ell_i(x, y),\ L$	Posterior identification gain and its population expectation, Eq. (1)
$\mathcal{G},\ G,\ k_G,\ K$	Fixed user partition, a group, $k_G = \lvert G\rvert$, and $K = \lvert\mathcal{G}\rvert$
$\epsilon_G(x)$	Within-group utility discrepancy $\max_{i,j \in G} \sup_y \lvert r_i(x, y) - r_j(x, y)\rvert$

I.2 Posterior identities and prediction risk

Proof of Proposition 1. Bayes' rule gives, on the support of $\bar{\pi}$,

$$\mathbb{P}(\mathbf{u} = i \mid x, y) = \frac{\mathbb{P}(\mathbf{u} = i \mid x)\pi_i(y)}{\bar{\pi}(y)}, \qquad \mathbb{E}_{\mathbf{y} \sim \bar{\pi}} \mathbb{P}(\mathbf{u} = i \mid x, \mathbf{y}) = \mathbb{P}(\mathbf{u} = i \mid x). \tag{9}$$

Conditional on a response, the true identity has this posterior distribution. Therefore

$$\begin{aligned} L &= \mathbb{E}_{\mathbf{y} \sim \bar{\pi}} \sum_i \mathbb{P}(\mathbf{u} = i \mid x, \mathbf{y})^2 - \sum_i \mathbb{P}(\mathbf{u} = i \mid x)^2 \\ &= \mathbb{E}_{\mathbf{y} \sim \bar{\pi}} \sum_i [\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{y}) - \mathbb{P}(\mathbf{u} = i \mid x)]^2. \end{aligned} \tag{10}$$

The final equality follows by expanding the square and using Eq. (9). Nonnegativity is immediate. Equality requires every posterior to equal its prior almost surely, which is precisely conditional independence. Also,

$$0 \le L \le 1 - \sum_i \mathbb{P}(\mathbf{u} = i \mid x)^2, \tag{11}$$

with the upper bound attained by perfect identification. Individual response gains $\ell_i(x, y)$ can be negative.

Define $\chi^2(\mathbb{P} \| Q) = \sum_y (\mathbb{P}(y) - Q(y))^2 / Q(y)$ on the common relevant support. Pearson chi-squared mutual information is

$$\begin{aligned} I_{\chi^2}(\mathbf{u}; \mathbf{y} \mid x) &:= \chi^2(\mathbb{P}_{\mathbf{u},\mathbf{y}|x} \| \mathbb{P}_{\mathbf{u}|x} \mathbb{P}_{\mathbf{y}|x}) \\ &= \mathbb{E}_{\mathbf{y} \sim \bar{\pi}} \sum_i \frac{[\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{y}) - \mathbb{P}(\mathbf{u} = i \mid x)]^2}{\mathbb{P}(\mathbf{u} = i \mid x)}. \end{aligned} \tag{12}$$

With a uniform conditional prior, the denominator is $1/n$, proving $I_{\chi^2} = nL$. This is not the Shannon quantity or a Rényi mutual information.

For multiclass Brier loss, an action a is a probability vector and the loss on identity i is $\sum_j (a_j - \mathbb{1}\{j = i\})^2$. At any conditioning information, completing the square shows that the Bayes-optimal action is the corresponding conditional probability vector, with risk one minus the sum of its squared entries. Thus

$$R^*_{\mathrm{Brier}}(x) - \mathbb{E}_{\mathbf{y}} R^*_{\mathrm{Brier}}(x, \mathbf{y}) = \mathbb{E}_{\mathbf{y}} \sum_i P(\mathbf{u} = i \mid x, \mathbf{y})^2 - \sum_i P(\mathbf{u} = i \mid x)^2 = L. \tag{13}$$

This uses the unnormalized multiclass Brier loss. The equality holds in expectation, not as an identity between individual posterior gains and individual Brier-loss differences. Standard scoring-rule background is given by Gneiting & Raftery (2007).

User-specific expectations and nonuniform priors. Writing $L_i = \mathbb{E}_{\mathbf{y} \sim \pi_i} \ell_i(x, \mathbf{y})$, Bayes' rule yields

$$L_i = \mathbb{P}(\mathbf{u} = i \mid x) \left[\sum_y \frac{\pi_i(y)^2}{\bar{\pi}(y)} - 1 \right] = \mathbb{P}(\mathbf{u} = i \mid x) \chi^2(\pi_i \| \bar{\pi}), \tag{14}$$

$$L = \sum_i P(\mathbf{u} = i \mid x)^2 \chi^2(\pi_i \| \bar{\pi}). \tag{15}$$

Every L_i is nonnegative under the correct joint law. For a nonuniform prior,

$$\big[\min_i P(\mathbf{u} = i \mid x)\big] I_{\chi^2} \le L \le \big[\max_i P(\mathbf{u} = i \mid x)\big] I_{\chi^2}; \tag{16}$$

there is no prior-independent constant converting L to I_{χ^2}.

Shannon information. Using natural logarithms,

$$L \le I(\mathbf{u}; \mathbf{y} \mid x) \le \log \left(1 + \frac{L}{\min_i P(\mathbf{u} = i \mid x)} \right). \tag{17}$$

For a difference of two probability vectors, the squared Euclidean norm is at most half the squared ℓ_1 norm: their positive and negative parts have the same sum. Pinsker's inequality then bounds the posterior squared distance by the posterior-to-prior KL divergence. Averaging gives the lower bound. For the upper bound, use $D_{\mathrm{KL}}(\mathbb{P} \| Q) \le \log(1 + \chi^2(\mathbb{P} \| Q))$ on the joint and product distributions, followed by Eq. (16). These divergence inequalities appear in Polyanskiy & Wu (2022, Sections 7.4 and 7.6). Under a uniform prior, the upper bound is $\log(1 + nL)$. To state bounds in bits, divide the two displayed bounds on Shannon information by $\ln 2$.

I.3 Group processing and anonymity

Proof of Theorem 1. For $i \in G$, identical response distributions imply

$$\mathbb{P}(\mathbf{u} = i \mid x, y) = \frac{\mathbb{P}(\mathbf{u} = i \mid x) \pi_G(y)}{\bar{\pi}(y)}, \quad \mathbb{P}(\mathbf{u} \in G \mid x, y) = \frac{\mathbb{P}(\mathbf{u} \in G \mid x) \pi_G(y)}{\bar{\pi}(y)}.$$

Their factorization is Eq. (3), including responses for which the group posterior is zero. When the group has positive posterior probability, conditioning on it leaves each member's prior conditional probability unchanged. Within-group uniformity makes this probability $1/k_G$, yielding Eq. (4). Without uniformity, the valid upper bound is $\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{u} \in G)$, not $1/k_G$.

Population bound under a uniform prior. If $\mathbb{P}(\mathbf{u} = i \mid x) = 1/n$ for every user, then

$$L_{\mathrm{grp}} \le \frac{K - 1}{n}. \tag{18}$$

Indeed,

$$\sum_i P(\mathbf{u} = i \mid x, y)^2 = \sum_G \frac{\mathbb{P}(\mathbf{u} \in G \mid x, y)^2}{k_G} \le \sum_G \frac{\mathbb{P}(\mathbf{u} \in G \mid x, y)}{k_G}.$$

Taking expectations gives K/n, since $\mathbb{P}(\mathbf{u} \in G \mid x) = k_G/n$; subtracting $1/n$ proves the result. Exact identification of the group attains the bound. Let $\mathbf{g} = G(\mathbf{u})$ denote the random group. For any prior, identical within-group response distributions also imply

$$I(\mathbf{u}; \mathbf{y} \mid x) = I(\mathbf{g}; \mathbf{y} \mid x) \leq H(\mathbf{g} \mid x) \leq \log K. \tag{19}$$

These are bounds on an idealized population with specified observer information. Five-way attacks do not replace their identity support by the size of the full group. Nor does an identity-anonymity bound protect every sensitive attribute: the group itself may reveal such an attribute.

I.4 Leakage and utility under idealized group averaging

Proof of Theorem 2. The construction preserves the response marginal because

$$\sum_G \sum_{i \in G} \mathbb{P}(\mathbf{u} = i \mid x) \pi_G(y) = \sum_G \mathbb{P}(\mathbf{u} \in G \mid x) \pi_G(y) = \sum_j P(\mathbf{u} = j \mid x) \pi_j(y) = \bar{\pi}(y).$$

It also gives, without assuming uniformity,

$$\mathbb{P}_{\mathrm{grp}}(\mathbf{u} = i \mid x, y) = \mathbb{P}(\mathbf{u} = i \mid x, \mathbf{u} \in G) \mathbb{P}_{\mathrm{id}}(\mathbf{u} \in G \mid x, y). \tag{20}$$

Under within-group uniformity, the new posterior is the arithmetic average of the original posteriors in its group. The priors and response marginals of the two mechanisms coincide, so Eq. (10) and

$$\begin{aligned} &\sum_{i \in G} \mathbb{P}_{\mathrm{id}}(\mathbf{u} = i \mid x, y)^2 - \frac{\mathbb{P}_{\mathrm{id}}(\mathbf{u} \in G \mid x, y)^2}{k_G} \\ &\qquad = \sum_{i \in G} \left[\mathbb{P}_{\mathrm{id}}(\mathbf{u} = i \mid x, y) - \frac{\mathbb{P}_{\mathrm{id}}(\mathbf{u} \in G \mid x, y)}{k_G} \right]^2 \end{aligned}$$

prove Eq. (6). Reduction is strict exactly when original within-group posteriors differ on responses of positive marginal probability.

For utility, define the group-average utility

$$\bar{r}_G(x, y) = \sum_{i \in G} \mathbb{P}(\mathbf{u} = i \mid x, \mathbf{u} \in G) r_i(x, y).$$

Expanding the prior-weighted mixture and exchanging finite sums gives

$$V_{\mathrm{id}} - V_{\mathrm{grp}} = \sum_G \mathbb{P}(\mathbf{u} \in G \mid x) \sum_{j \in G} \mathbb{P}(\mathbf{u} = j \mid x, \mathbf{u} \in G) \mathbb{E}_{\pi_j}[r_j(x, \mathbf{y}) - \bar{r}_G(x, \mathbf{y})]. \tag{21}$$

For each $j \in G$, $|r_j(x, y) - \bar{r}_G(x, y)| \leq \epsilon_G(x)$. Applying the triangle inequality proves Eq. (7) for any positive prior. If utilities coincide within every group, the population utility is preserved; individual users' utilities need not each be preserved.

Interpretation and the Shannon analogue. An implementation of Eq. (5) would sample a member according to the group's conditional prior independently of the recipient within that group, then use that member's individual response distribution. This randomized construction need not be realizable by a deterministic context-processing function ϕ. For any prior it preserves the joint law of $(\mathbf{g}, \mathbf{y})$ and removes the conditional identity information within the group. By the chain rule,

$$I(\mathbf{u}; \mathbf{y}_{\mathrm{id}} \mid x) - I(\mathbf{u}; \mathbf{y}_{\mathrm{grp}} \mid x) = I(\mathbf{u}; \mathbf{y}_{\mathrm{id}} \mid x, \mathbf{g}) \geq 0. \tag{22}$$

The absence of a uniformity condition in this Shannon statement does not extend to L; a counterexample is given below.

I.5 User-matching gain is not total improvement

Let π_0 be the actual No Pers. distribution and define

$$\Delta_i = \mathbb{E}_{\pi_i} r_i(x, \mathbf{y}) - \mathbb{E}_{\pi_0} r_i(x, \mathbf{y}), \qquad \delta_i = \mathbb{E}_{\pi_i} r_i(x, \mathbf{y}) - \mathbb{E}_{\bar{\pi}} r_i(x, \mathbf{y}).$$

Then

$$\Delta_i = \left[\mathbb{E}_{\bar{\pi}} r_i(x, \mathbf{y}) - \mathbb{E}_{\pi_0} r_i(x, \mathbf{y})\right] + \delta_i. \tag{23}$$

The first term uses the same response distribution for every recipient and can be positive without identity information. In contrast, the user-matching term obeys

$$L_i \geq 4\mathbb{P}(\mathbf{u} = i \mid x)\delta_i^2. \tag{24}$$

To prove this, let $h_i(y) = \pi_i(y)/\bar{\pi}(y) - 1$. Then $\mathbb{E}_{\bar{\pi}} h_i = 0$, so

$$\delta_i = \mathbb{E}_{\bar{\pi}}[(r_i(x, \mathbf{y}) - \mathbb{E}_{\bar{\pi}} r_i(x, \mathbf{y}))h_i(\mathbf{y})].$$

Cauchy–Schwarz and $r_i \in [0, 1]$ give $\delta_i^2 \leq \mathrm{Var}_{\bar{\pi}}(r_i)\chi^2(\pi_i \| \bar{\pi}) \leq \chi^2(\pi_i \| \bar{\pi})/4$. Equation (14) proves the claim. It is a necessary information cost for user matching, not a monotone ranking of arbitrary utility and privacy scores. Negative matching gains can also carry information.

I.6 Output processing and optional encoding assumptions

Output processing. Suppose $\mathbf{u} \to \mathbf{y} \to \mathbf{t}$ is a Markov chain conditional on x. The processed posterior obeys

$$\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{t}) = \mathbb{E}[\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{y}) \mid x, \mathbf{t}].$$

Expanding the squared differences and conditioning on $\mathbf{t}$ eliminates the cross term, hence

$$L(\mathbf{y}) - L(\mathbf{t}) = \mathbb{E} \sum_i [\mathbb{P}(\mathbf{u} = i \mid x, \mathbf{y}) - \mathbb{P}(\mathbf{u} = i \mid x, \mathbf{t})]^2 \geq 0. \tag{25}$$

This holds for any prior. Input-context processing does not generally produce such a Markov relation between the old and new responses. In particular, neither a summary nor an averaged embedding is proved to satisfy

$$\pi\left(\cdot \mid x, \frac{1}{k_G} \sum_{j \in G} z_j\right) = \frac{1}{k_G} \sum_{j \in G} \pi(\cdot \mid x, z_j).$$

The left side need not even be defined for textual contexts.

Optional lossless and unique representations. The main results do not require distinct contexts or invertible embeddings. A stronger optional model assumes contexts c_i, a deterministic encoder e, and decoder d with $d(e(c_i)) = c_i$, as well as $c_i \neq c_j$ for distinct users. Then $z_i = e(c_i)$ gives a one-to-one correspondence between users and contexts, so

$$\mathbb{P}(\mathbf{u} = i \mid x, y) = \mathbb{P}(\mathbf{z} = z_i \mid x, y), \qquad I(\mathbf{u}; \mathbf{y} \mid x) = I(\mathbf{z}; \mathbf{y} \mid x).$$

Injectivity alone is enough for this correspondence; losslessness is stronger. Neither condition implies different utilities or response distributions. These are optional mathematical assumptions, not properties established for the embedding model used in the experiments.

I.7 Boundary counterexamples

Marginal independence is insufficient. Let $\mathbf{u} - 1$ and $\mathbf{x}$ be independent fair bits, with user identities $\mathbf{u} \in \{1, 2\}$, and let $\mathbf{y} = \mathbf{x} \oplus (\mathbf{u} - 1)$. Each of $\mathbf{x}$ and $\mathbf{y}$ is marginally independent of $\mathbf{u}$, but their pair identifies the user. The zero-leakage condition is $\mathbf{y} \perp \mathbf{u} \mid \mathbf{x}$.

Context processing can increase response leakage. Consider four equally likely users with different original contexts. A fixed LLM returns the same response at all four contexts. Let ϕ map users $1, 2$ to context a and users $3, 4$ to context b, and let the LLM return different deterministic responses at a and b. The original L is zero; after processing, the response identifies a two-person group and $L = 1/2 - 1/4 = 1/4$. Compression of the input therefore does not imply reduced response leakage.

Within-group uniformity cannot be dropped from the quadratic guarantee. Take

$$\big(\mathbb{P}(\mathbf{u}=i \mid x)\big)_{i=1}^{3}=(0.6,0.1,0.3), \qquad \big(\mathbb{P}_{\mathrm{id}}(\mathbf{y}=1 \mid \mathbf{u}=i,x)\big)_{i=1}^{3}=(0.5,1,0).$$

For groups $\{1,2\}$ and $\{3\}$, prior-weighted averaging gives success probability $4/7$ for both members of the first group. Direct calculation yields

$$L_{\mathrm{id}}=0.09, \qquad L_{\mathrm{grp}}\simeq 0.1053061224.$$

Thus L increases despite Eq. (22). The weighted posterior transformation is not the arithmetic projection used in Eq. (6); nor must it be a user-independent transformation of the original response.

Utility can improve without leakage. Let all users give utility one to response a and zero to response b. If No Pers. returns b and a common nonempty context causes the model to return a for every user, all users gain utility while L remains zero. This also explains why No Pers. cannot be identified with the single-group context mechanism.

J Experimental Details and Additional Analyses

J.1 Data and context construction

The Qwen evaluation panel contains $4{,}564$, $1{,}774$, and $2{,}248$ examples for Abstract Generation, Review Writing, and Topic Writing, and $1{,}730$, $1{,}500$, and $1{,}500$ for News Headline, Scholarly Title, and Tweet Paraphrase. Fixed example-ID masks remove generation refusals across arms on Review ($1{,}784 \rightarrow 1{,}774$) and Topic ($2{,}452 \rightarrow 2{,}248$). Metric-specific valid and paired sample counts can differ.

Each generator writes approximately 200-word user summaries and group personas. Qwen3-Embedding-8B embeds the summaries for k-means grouping. A group's document pool includes its members' training histories, including the recipient's; query-side decontamination removes near-duplicates of the query before retrieval. The persona and pool are fixed within a group, with retrieval determined by the query. BM25 retrieves four documents under both group and individual personalization. C-RAG and O-RAG remove the persona or summary and use group and own-history retrieval, respectively; Persona and Summary retain only the respective description. Appendices C–E give retrieval, construction, and prompt details. Additional generator comparisons appear in Appendix A, and the full-parameter SFT comparison appears in Appendix H.

J.2 Uncertainty for the primary displays

Within each task, generator, and metric, the main comparison uses the intersection of valid records for the available arms; the paired group-count analysis uses the intersection over all six settings. The figures retain their original full-panel means. Fixed refusal masks are applied before matching. Utility parse failures are excluded per metric, and no score is imputed for an unscored arm. Consequently, counts differ between metrics. Tables below report each count and use the same samples as the revised primary table means. Paired group-count means are reported with their intervals separately from the unchanged figures.

We resample users with replacement, retaining all of each user's examples, and recompute item-weighted means and paired differences ($10{,}000$ draws). The 95% percentile intervals are pointwise, not simultaneous. Two-sided sign-flip tests flip all differences for a user together ($10{,}000$ draws; add-one correction); Holm adjustment uses separate families of 48 primary contrasts and 24 granularity contrasts. One random-number generator uses seed 20260926. Inference assumes user-level sign exchangeability and conditions on the fixed generated responses and attacker evidence. Top-1 attribution uses fractional credit for tied labels and zero credit for no available label; the same rule applies to response and query-only records. These deterministic tie credits can slightly change means from the supplementary full-panel summaries.

Historical document-candidate records preserve item IDs and true-label indices but do not all preserve complete candidate identities or prompt hashes. Their intervals quantify sampling variability conditional on the stored scores, not uncertainty about candidate construction. The frozen-input replay in Appendix J.4 provides the stricter candidate-matched cross-attacker comparison.

J.3 Paired group-count comparisons

Tables 28 and 29 report the $K = 10$ versus 200 comparison for group persona plus retrieval. They report matched-item means, metric-specific valid counts, user-clustered sign-flip values, and Benjamini–Hochberg adjusted values from the full 300-test family. All observations of a user are flipped together; inference requires the corresponding sign-exchangeability assumption and conditions on the observed generations. We use an adjusted-value threshold of 0.05.

These additional tests use pair-specific intersections rather than the all-setting intersections of the revised granularity figure. Paired contrasts with confidence intervals are reported in Appendix J.2; these do not supply error bars for the unchanged figures. Significance markers in the other supplementary tables retain the tests and comparison families stated in their captions.

Table 24: **Paired primary comparisons: Abstract Generation.** Differences are B minus A on the same valid items as Table 2. Intervals are pointwise 95% user-bootstrap intervals. p_H is Holm-adjusted across the 48 primary contrasts; N/G/F denote No Pers./GrRAG/Full Pers. Counts vary by metric. Missing group Utility scores permit only F–N comparisons for Llama and Gemma.

Generator	Metric	B–A	n	Δ	95% CI	p_H
Qwen3.5-9B	Utility	G–N	4551	+0.132	[+0.087, +0.176]	0.005
Qwen3.5-9B	Utility	F–G	4551	+0.242	[+0.195, +0.290]	0.005
Qwen3.5-9B	Leak	G–N	4564	+0.011	[+0.001, +0.020]	0.188
Qwen3.5-9B	Leak	F–G	4564	+0.038	[+0.029, +0.047]	0.005
Qwen3.5-9B	METEOR	G–N	4564	+0.030	[+0.028, +0.031]	0.005
Qwen3.5-9B	METEOR	F–G	4564	+0.008	[+0.007, +0.009]	0.005
Llama-3.1-8B	Utility	F–N	4560	+0.205	[+0.175, +0.235]	0.005
Llama-3.1-8B	Leak	G–N	4564	+0.002	[−0.007, +0.012]	1.000
Llama-3.1-8B	Leak	F–G	4564	+0.022	[+0.014, +0.031]	0.005
Llama-3.1-8B	METEOR	G–N	4564	+0.030	[+0.029, +0.032]	0.005
Llama-3.1-8B	METEOR	F–G	4564	+0.007	[+0.005, +0.008]	0.005
Gemma-3-12B	Utility	F–N	4559	+0.131	[+0.093, +0.170]	0.005
Gemma-3-12B	Leak	G–N	4564	+0.004	[−0.006, +0.013]	1.000
Gemma-3-12B	Leak	F–G	4564	+0.014	[+0.007, +0.023]	0.005
Gemma-3-12B	METEOR	G–N	4564	+0.024	[+0.023, +0.025]	0.005
Gemma-3-12B	METEOR	F–G	4564	+0.000	[−0.001, +0.002]	1.000

Table 25: **Paired primary comparisons: Review Writing.** Differences are B minus A on the same valid items as Table 2. Intervals are pointwise 95% user-bootstrap intervals. p_H is Holm-adjusted across the 48 primary contrasts; N/G/F denote No Pers./GrRAG/Full Pers. Counts vary by metric. Missing group Utility scores permit only F–N comparisons for Llama and Gemma.

Generator	Metric	B–A	n	Δ	95% CI	p_H
Qwen3.5-9B	Utility	G–N	1770	−0.128	[−0.160, −0.095]	0.005
Qwen3.5-9B	Utility	F–G	1770	+0.292	[+0.243, +0.341]	0.005
Qwen3.5-9B	Leak	G–N	1774	−0.033	[−0.056, −0.011]	0.042
Qwen3.5-9B	Leak	F–G	1774	+0.115	[+0.093, +0.137]	0.005
Qwen3.5-9B	METEOR	G–N	1774	+0.075	[+0.072, +0.077]	0.005
Qwen3.5-9B	METEOR	F–G	1774	+0.012	[+0.009, +0.016]	0.005
Llama-3.1-8B	Utility	F–N	1773	+0.091	[+0.058, +0.125]	0.005
Llama-3.1-8B	Leak	G–N	1774	−0.032	[−0.054, −0.010]	0.040
Llama-3.1-8B	Leak	F–G	1774	+0.080	[+0.060, +0.099]	0.005
Llama-3.1-8B	METEOR	G–N	1774	+0.037	[+0.036, +0.039]	0.005
Llama-3.1-8B	METEOR	F–G	1774	+0.006	[+0.003, +0.008]	0.005
Gemma-3-12B	Utility	F–N	1774	+0.015	[−0.037, +0.068]	1.000
Gemma-3-12B	Leak	G–N	1774	−0.032	[−0.055, −0.009]	0.055
Gemma-3-12B	Leak	F–G	1774	+0.063	[+0.042, +0.084]	0.005
Gemma-3-12B	METEOR	G–N	1774	+0.062	[+0.060, +0.064]	0.005
Gemma-3-12B	METEOR	F–G	1774	+0.014	[+0.012, +0.016]	0.005

J.4 Frozen-input independent-attacker replay

The replay evaluates 150 queries each from News, Review, and Topic, selected using existing evidence before fresh scoring. News includes 126 users; Review and Topic include 150 each. The same refusal masks apply. Both Qwen3.6-35B-A3B and Llama-3.1-8B-Instruct receive the same 4,500 message payloads, with model-specific native chat templates. This designated sample is not a random replication of every main-table comparison.

For each item and evidence representation, one query-only input is paired with own-context, group-context, No Pers., and fixed-generic responses. Candidate identities and order are fixed from the summary protocol; document evidence retrieves four documents for the same candidate examples. Hence the query-only control cancels in the own/group contrast. First-token A–E label scores use expected uniform tie credit. All five labels appear among the top 20 alternatives in 4,492/4,500

Table 26: **Paired primary comparisons: Topic Writing.** Differences are B minus A on the same valid items as Table 2. Intervals are pointwise 95% user-bootstrap intervals. p_H is Holm-adjusted across the 48 primary contrasts; N/G/F denote No Pers./GrRAG/Full Pers. Counts vary by metric. Missing group Utility scores permit only F–N comparisons for Llama and Gemma.

Generator	Metric	B–A	n	Δ	95% CI	p_H
Qwen3.5-9B	Utility	G–N	2248	+0.027	[−0.006, +0.060]	0.676
Qwen3.5-9B	Utility	F–G	2248	+0.235	[+0.188, +0.283]	0.005
Qwen3.5-9B	Leak	G–N	2248	+0.041	[+0.020, +0.062]	0.005
Qwen3.5-9B	Leak	F–G	2248	+0.194	[+0.174, +0.216]	0.005
Qwen3.5-9B	METEOR	G–N	2248	+0.008	[+0.005, +0.010]	0.005
Qwen3.5-9B	METEOR	F–G	2248	+0.010	[+0.006, +0.013]	0.005
Llama-3.1-8B	Utility	F–N	2246	+0.073	[+0.053, +0.093]	0.005
Llama-3.1-8B	Leak	G–N	2248	+0.026	[+0.008, +0.045]	0.055
Llama-3.1-8B	Leak	F–G	2248	+0.122	[+0.102, +0.141]	0.005
Llama-3.1-8B	METEOR	G–N	2248	+0.007	[+0.005, +0.010]	0.005
Llama-3.1-8B	METEOR	F–G	2248	+0.010	[+0.008, +0.013]	0.005
Gemma-3-12B	Utility	F–N	2247	+0.183	[+0.151, +0.215]	0.005
Gemma-3-12B	Leak	G–N	2248	+0.010	[−0.011, +0.030]	1.000
Gemma-3-12B	Leak	F–G	2248	+0.158	[+0.138, +0.177]	0.005
Gemma-3-12B	METEOR	G–N	2248	+0.006	[+0.004, +0.008]	0.005
Gemma-3-12B	METEOR	F–G	2248	+0.012	[+0.010, +0.014]	0.005

Table 27: **Paired contrasts for the granularity figure.** Common valid-item sets across settings; the figure retains its original full-panel means; Δ is B minus A. Intervals are pointwise 95% user-bootstrap intervals. p_H uses Holm correction over these 24 contrasts. F denotes Full Pers. The Utility validity limitation for Scholarly Title applies regardless of statistical significance.

Task	Metric	B–A	n	Δ	95% CI	p_H
Abstract Generation	Utility	200–10	4528	+0.119	[+0.074, +0.165]	0.002
Abstract Generation	Utility	F–100	4528	+0.240	[+0.192, +0.287]	0.002
Abstract Generation	Leak	200–10	4564	+0.033	[+0.024, +0.042]	0.002
Abstract Generation	Leak	F–100	4564	+0.023	[+0.013, +0.033]	0.002
Review Writing	Utility	200–10	1769	+0.015	[−0.012, +0.043]	1.000
Review Writing	Utility	F–100	1769	+0.292	[+0.244, +0.341]	0.002
Review Writing	Leak	200–10	1774	+0.048	[+0.030, +0.067]	0.002
Review Writing	Leak	F–100	1774	+0.112	[+0.091, +0.134]	0.002
Topic Writing	Utility	200–10	2247	−0.015	[−0.053, +0.023]	1.000
Topic Writing	Utility	F–100	2247	+0.235	[+0.188, +0.283]	0.002
Topic Writing	Leak	200–10	2248	+0.143	[+0.122, +0.164]	0.002
Topic Writing	Leak	F–100	2248	+0.178	[+0.159, +0.197]	0.002
News Headline	Utility	200–10	1716	+0.205	[+0.124, +0.286]	0.002
News Headline	Utility	F–100	1716	+0.407	[+0.325, +0.493]	0.002
News Headline	Leak	200–10	1730	+0.023	[+0.005, +0.041]	0.117
News Headline	Leak	F–100	1730	+0.010	[−0.007, +0.027]	1.000
Scholarly Title	Utility	200–10	1498	+0.039	[−0.015, +0.092]	1.000
Scholarly Title	Utility	F–100	1498	−0.047	[−0.105, +0.010]	1.000
Scholarly Title	Leak	200–10	1500	−0.009	[−0.022, +0.003]	1.000
Scholarly Title	Leak	F–100	1500	+0.001	[−0.011, +0.013]	1.000
Tweet Paraphrase	Utility	200–10	1496	+0.029	[−0.047, +0.106]	1.000
Tweet Paraphrase	Utility	F–100	1496	+0.009	[−0.074, +0.093]	1.000
Tweet Paraphrase	Leak	200–10	1500	+0.019	[+0.001, +0.037]	0.456
Tweet Paraphrase	Leak	F–100	1500	+0.226	[+0.202, +0.250]	0.002

Table 28: **Paired $K = 10$ versus 200 group-context comparison: short-form tasks.** Means use matched valid records; Δ is the report's difference $(200 - 10)$, which need not equal subtraction of rounded means. p_u is the user-clustered sign-flip value and q_u its Benjamini–Hochberg adjustment over the full 300-test family. These are not Figure 2's aggregation or uncertainty estimates. True-label score is the fixed attacker's score, not $\mathbb{P}(\mathbf{u} = i \mid x, y)$.

Task	Metric	n	$K = 10$	$K = 200$	Δ	p_u	q_u
News	ROUGE-L	1730	0.136	0.138	+0.003	0.24	0.33
News	True-label score	1730	0.643	0.667	+0.025	5.00×10^{-5}	1.50×10^{-4}
News	Top-1	1730	0.699	0.721	+0.022	0.014	0.03
News	Leak	1730	0.032	0.055	+0.023	0.012	0.027
News	Ref. match	1730	2.206	2.182	-0.024	0.39	0.49
News	Util	1724	2.368	2.571	+0.204	1.00×10^{-5}	3.30×10^{-5}
Title	ROUGE-L	1500	0.384	0.386	+0.002	0.61	0.68
Title	True-label score	1500	0.774	0.772	-0.002	0.68	0.74
Title	Top-1	1500	0.819	0.810	-0.009	0.16	0.24
Title	Leak	1500	-0.003	-0.012	-0.009	0.16	0.24
Title	Ref. match	1500	4.007	4.027	+0.021	0.58	0.66
Title	Util	1500	4.531	4.570	+0.039	0.16	0.24
Tweet	ROUGE-L	1500	0.359	0.348	-0.011	4.20×10^{-4}	0.0011
Tweet	True-label score	1500	0.355	0.381	+0.026	1.00×10^{-5}	3.30×10^{-5}
Tweet	Top-1	1500	0.390	0.406	+0.016	0.084	0.14
Tweet	Leak	1500	-0.025	-0.006	+0.019	0.048	0.088
Tweet	Ref. match	1500	5.220	5.201	-0.019	0.65	0.71
Tweet	Util	1497	2.841	2.871	+0.030	0.45	0.56

Table 29: **Paired $K = 10$ versus 200 group-context comparison: long-form tasks.** Means use matched valid records; Δ is the report's difference $(200 - 10)$, which need not equal subtraction of rounded means. p_u is the user-clustered sign-flip value and q_u its Benjamini–Hochberg adjustment over the full 300-test family. These are not Figure 2's aggregation or uncertainty estimates. True-label score is the fixed attacker's score, not $\mathbb{P}(\mathbf{u} = i \mid x, y)$.

Task	Metric	n	$K = 10$	$K = 200$	Δ	p_u	q_u
Abstract	ROUGE-L	4564	0.190	0.192	+0.002	2.00×10^{-5}	6.10×10^{-5}
Abstract	True-label score	4564	0.701	0.731	+0.030	1.00×10^{-5}	3.30×10^{-5}
Abstract	Top-1	4564	0.738	0.768	+0.030	1.00×10^{-5}	3.30×10^{-5}
Abstract	Leak	4564	0.045	0.077	+0.033	1.00×10^{-5}	3.30×10^{-5}
Abstract	Ref. match	4526	2.947	3.018	+0.071	2.40×10^{-4}	6.70×10^{-4}
Abstract	Util	4547	2.518	2.638	+0.120	1.00×10^{-5}	3.30×10^{-5}
Review	ROUGE-L	1774	0.135	0.138	+0.003	1.00×10^{-5}	3.30×10^{-5}
Review	True-label score	1774	0.567	0.615	+0.048	1.00×10^{-5}	3.30×10^{-5}
Review	Top-1	1774	0.626	0.675	+0.049	1.00×10^{-5}	3.30×10^{-5}
Review	Leak	1774	0.050	0.098	+0.048	2.00×10^{-5}	6.10×10^{-5}
Review	Ref. match	1774	1.806	1.749	-0.056	0.023	0.047
Review	Util	1774	1.134	1.149	+0.015	0.32	0.42
Topic	ROUGE-L	2248	0.116	0.117	+0.001	0.27	0.37
Topic	True-label score	2248	0.560	0.699	+0.140	1.00×10^{-5}	3.30×10^{-5}
Topic	Top-1	2248	0.619	0.761	+0.142	1.00×10^{-5}	3.30×10^{-5}
Topic	Leak	2248	0.075	0.218	+0.143	1.00×10^{-5}	3.30×10^{-5}
Topic	Ref. match	2248	1.480	1.452	-0.028	0.17	0.25
Topic	Util	2247	1.223	1.209	-0.015	0.46	0.57

Qwen records and $4{,}487/4{,}500$ Llama records; the rest have four. No record is excluded for score coverage, and there are no failed requests or all-label abstentions.

Table 30 reports all 12 task–evidence–attacker contrasts. Pointwise percentile intervals use $10{,}000$ user-cluster bootstrap draws; user sign-flip tests use $100{,}000$ draws with Holm correction over these 12 contrasts (seed 20260915, one random-number stream). The intervals are not simultaneous confidence intervals. With document evidence, both Topic reductions survive correction. With summary evidence, the Topic differences are negative for both attackers, but only Qwen survives correction. News and Review comparisons are inconclusive.

Table 30: **All frozen-input replay contrasts.** DP/SP are document/summary candidates; the query-only control and candidate identities are matched within each item and representation. Negative Δ means lower group attribution. Pointwise user-bootstrap 95% intervals and Holm-adjusted p_H use the replay's 12-contrast family. These selected 150-item task samples do not represent all main-table cells.

Task	Evidence	Attacker	n/users	Own Leak	Group Leak	Δ [95% CI]	p_H
News	DP	Qwen	150/126	0.0567	0.0533	-0.0033 $[-0.0616, +0.0552]$	1
News	DP	Llama	150/126	-0.0156	-0.0056	$+0.0100$ $[-0.0399, +0.0616]$	1
News	SP	Qwen	150/126	0.0500	0.0200	-0.0300 $[-0.0780, +0.0162]$	1
News	SP	Llama	150/126	0.0633	0.0400	-0.0233 $[-0.0788, +0.0285]$	1
Review	DP	Qwen	150/150	-0.0133	-0.0367	-0.0233 $[-0.0833, +0.0367]$	1
Review	DP	Llama	150/150	-0.0189	-0.0089	$+0.0100$ $[-0.0433, +0.0633]$	1
Review	SP	Qwen	150/150	-0.0533	-0.0444	$+0.0089$ $[-0.0411, +0.0600]$	1
Review	SP	Llama	150/150	0.0633	0.0356	-0.0278 $[-0.0700, +0.0133]$	1
Topic	DP	Qwen	150/150	0.1611	-0.0289	-0.1900 $[-0.2633, -0.1200]$	1.20×10^{-4}
Topic	DP	Llama	150/150	0.1900	0.0167	-0.1733 $[-0.2500, -0.1000]$	2.20×10^{-4}
Topic	SP	Qwen	150/150	0.1033	-0.0100	-0.1133 $[-0.1867, -0.0433]$	0.0317
Topic	SP	Llama	150/150	0.1600	0.0633	-0.0967 $[-0.1733, -0.0233]$	0.1499