OUTPUT-DERIVED SUPERVISION FOR CONTROLLABLE SYNTHETIC TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Attribute-conditioned text generators can produce fluent text that does not express the requested attributes. We ask whether a generator's own outputs can provide useful feedback for improving attribute control after access to source records ends. We generate text from requested attribute tuples and treat the requests as noisy labels for the resulting outputs. These pairs train a lightweight body-only verifier without new human or language-model annotations. The frozen verifier then guides reinforcement fine-tuning or best-of-eight candidate selection. On Enron, reinforcement fine-tuning improves automatic attribute recovery in all six seeds, with a mean gain of 3.6 percentage points, and two independent language-model judges confirm the improvement on three evaluated seeds. An Amazon transfer study shows a 9.7-point independently judged gain. The improvements are attribute-dependent and reduce distributional fidelity. In a matched comparison, reinforcement fine-tuning achieves higher judged recovery than reranking, while reranking better preserves fidelity. These results show that requested attributes paired with generated outputs provide useful but imperfect control supervision without additional source access.

1 INTRODUCTION

Attribute-conditioned text generation requires both domain realism and attribute adherence. A generated email may resemble workplace communication while omitting the requested approval action, and a plausible review may express the wrong rating. Domain adaptation can improve realism without ensuring that generated text expresses its requested attributes (Mattern et al., 2022; Yue et al., 2023).

We study whether attribute control can be improved after access to source records has ended and without new control-stage annotations. The control stage retains an adapted generator, a structural description of the requested attribute space, and the generator's own outputs. From those inputs the generator produces a body for each requested attribute tuple, and we use the request as a noisy label for that body. These pairs train a lightweight body-only verifier, yielding *output-derived supervision*, which we freeze before use (Figure 1).

The supervision is imperfect because a requested attribute need not appear in the generated body. We therefore do not treat the verifier as an oracle or measure success by its optimized reward. Instead we test whether verifier-guided reinforcement fine-tuning (OUTPUT-VERIFIER RFT) or best-of-eight candidate selection (VERIFIER RERANK-8) improves attribute recovery under instruments outside the control loop, and we measure the accompanying costs in distributional fidelity and downstream utility.

Prior work uses generated text and requested labels to train downstream classifiers (Ye et al., 2022), and learned predictors to guide decoding (Yang & Klein, 2021). Here, the predictor feeds back to the adapted generator that supplied its training data. We ask whether this imperfect supervision improves independently measured control, rather than proposing a new optimizer or studying synthetic classifier training in general.

Independent evaluation supports control gains, but not uniform improvement: topic and formality account for most of the primary Enron lift, while urgency does not improve. In the matched audit,

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.

both judges favor RFT on recovery, whereas reranking has better fidelity point estimates. Neither a general utility gain nor an overall controller ranking is established.

Separately from those measurements, the control stage accesses no additional source records. This yields a conditional post-processing claim: it would preserve a valid joint upstream differential privacy (DP) guarantee, but the historical upstream artifacts do not establish that premise (§3.6).

Our contributions are:

- **Output-derived control supervision.** We train a body-only attribute verifier from generated text paired with requested attributes, requiring no new human or language-model annotations and no further source-record access.
- **Independent evidence of useful but imperfect supervision.** Verifier-guided RFT improves independently measured attribute recovery on Enron and Amazon. The gains vary by attribute, reduce distributional fidelity, and do not establish a general downstream utility benefit.
- **A matched comparison of refinement and selection.** Under shared requests and fixed decoding, RFT achieves higher independently judged recovery, while best-of-eight reranking better preserves distributional fidelity.

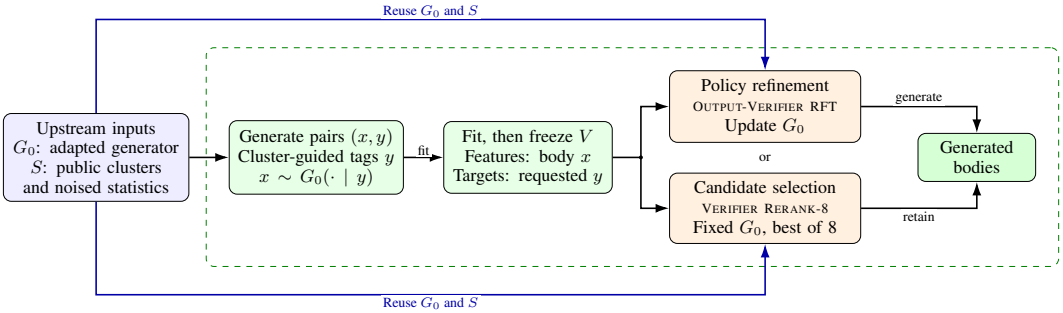


Figure 1: **Cluster-guided generation and output-derived supervision (schematic).** Public label-space clusters and noised statistics in S guide requested tags y. G_0 generates bodies x conditioned on those tags, not cluster IDs. The pairs train a body-only scorer V for refinement or selection. Requests are noisy body labels. The matched reranker uses a frozen scorer reconstruction. Blue arrows reuse upstream inputs. The dashed region denotes no further source access, not a certified DP boundary. Independent evaluation remains outside the loop.

2 Background and Related Work

We build on synthetic supervision, classifier-guided generation, DP text synthesis, and private alignment. Our focus is on how the control feedback is obtained.

Synthetic supervision and classifier-guided control. Ye et al. (2022) train task models on language-model-generated datasets using the requested class as the label, and Self-Rewarding Language Models (Yuan et al., 2024) use a judge model to supply feedback during alignment. FUDGE (Yang & Klein, 2021) steers decoding with predictors of final attributes learned from partial sequences. Our verifier instead learns a body-to-attribute mapping from the adapted generator's requested tags without judge-labeled fitting text, and returns feedback to that generator. Independent evaluation tests this noisy supervision under two established controllers.

User-level DP. A randomized mechanism M is (ε, δ)-DP if for all datasets D, D' differing in one *user's* entire contribution and all output sets B, $\Pr[M(D) \in B] \leq e^{\varepsilon} \Pr[M(D') \in B] + \delta$ (Dwork & Roth, 2014). Post-processing a valid joint DP release preserves its guarantee. Our training uses per-user gradient clipping and an Opacus accountant (Abadi et al., 2016; Yousefpour et al., 2021), with sender as the intended privacy unit. Implementation qualifications appear in §3.6.

DP synthetic text. DP fine-tuning (Yue et al., 2023; Li et al., 2022; Yu et al., 2022), multi-attribute conditioning (Mattern et al., 2022; Yue et al., 2023), marginal-based synthesis (McKenna et al., 2021; 2022), user-level language-model learning (McMahan et al., 2018), and parameter-efficient private adaptation (Kurakin et al., 2023) supply our upstream construction. CTCL (Tan et al., 2025) is the closest precedent, while Private Evolution and Aug-PE (Lin et al., 2024; Xie et al., 2024) instead use DP-noised private votes to select and vary frozen-model candidates. Our generator receives an explicit attribute tuple rather than a cluster identifier, and the verifier stage, not the clustering or the DP fine-tuning, is what this paper evaluates.

Private alignment and post-processing. Private alignment methods differ in where private data enters the control loop. Wu et al. (2024) account for private reward learning and policy updates from private prompts, AUP-RLHF (Zhang et al., 2025) uses user-level label DP for preferences, and Cho & Sun (2026) DP-train a reward and derive a policy by best-of-N reranking under post-processing. Xu et al. (2026) apply reinforcement fine-tuning to DP synthetic text with a reward that consults the private corpus at every update, where a prompt-adherence term guards against reward hacking. We instead make adherence the objective, with a verifier that reads only generated bodies paired with their requested tags. Once the adapted generator and structural statistics are released, fitting the verifier, reranking, and refining the policy need no further source access. This is a conditional post-processing statement rather than a new mechanism, and the historical implementation does not establish artifact-level or end-to-end privacy (§3.6). Appendix B develops these distinctions.

3 Method

3.1 Overview

The pipeline has three stages: sample attribute requests from structural statistics, generate bodies with an adapted policy, and fit a body-only verifier to those request–body pairs. The frozen verifier then guides selection or policy refinement (Figure 1). Evaluation remains outside that loop.

3.2 Sampling requested attributes

A categorical *skeleton* specifies requested values (*tags*) for attribute *keys*, while the generator produces a free-text *body*. Each dataset profile defines a closed public vocabulary and a cluster geometry: a fixed partition of attribute combinations. A fixed pipeline annotates source records inside the private environment. Enron has eight requested keys and Amazon six (Appendix G.1).

The *structural release* S combines a public partition with noised occupancy and marginal statistics, not individual records. We embed public label strings with all-MiniLM-L6-v2 and run k-means over the Cartesian product of three profile-specific keys (Reimers & Gurevych, 2019; Wang et al., 2020): topic, primary intent, and formality for Enron. The resulting 27 Enron and 24 Amazon clusters depend on public labels, not private text or record frequencies. Private records determine the occupancies, which are released with noise. Per-record skeletons remain private.

Generation draws a cluster from the noised occupancy mixture, selects a cluster-key combination using products of noised one-way marginals, and samples the remaining keys (the five Enron keys outside that triple) independently. This captures coarse cluster-level dependence, not the full joint. The generator receives the complete requested tuple, not a cluster ID or centroid, and produces a body. Those request–body pairs then supply verifier training data (Figure 1 and Appendix F).

3.3 Adapted generator

The generation policy consists of a frozen `Qwen3.5-4B` checkpoint (Qwen Team, 2026a) plus a trainable adapter: either LoRA (Hu et al., 2022) or per-key/positional soft prompts (Lester et al., 2021). The DP-SGD recipe aggregates each user's gradients, clips them to norm C, and adds Gaussian noise. Noise is calibrated to target $\varepsilon_{\text{learn}}$ for the configured sampling rate and training horizon. That accounting calculation uses the selected input, not the raw-corpus selection map, and is therefore not an artifact guarantee (§3.6). Write θ_{DP} for the trained adapter. We denote the DP-LoRA generator at $\varepsilon_{\text{learn}}=4$ by L4. It supplies the verifier's fitting corpus and the starting generator for both control procedures. The adapter and conditioning ablations are reported in Appendix D.10. The

model-transfer screen substitutes the fully open `SmolLM3-3B` checkpoint (Hugging Face, 2025) while retaining rank, target modules, privacy accounting, structural profile, and optimization recipe.

3.4 Learning the output-derived verifier

The initial generator L4 generates bodies for cluster-guided tuples sampled from S. We use the requested tags as noisy labels for those bodies and fit a verifier V with a shared term-frequency inverse-document-frequency (TF-IDF) vectorizer and one class-balanced, multiclass logistic-regression head per key. The verifier predicts attribute values from the body alone, without seeing the requested tags. For requested keys K, its reward is the fraction of predictions matching those tags,

$$r(\text{body}, \text{tags}) \;=\; \frac{1}{|K|} \sum_{k \in K} \mathbf{1}[\, V_k(\text{body}) = \text{tags}_k \,]\,. \tag{1}$$

The verifier is fitted once and then frozen. It is an attribute scorer, not an oracle: agreement with a requested tag need not mean that the body expresses that attribute. Equation 1 is the training reward, and it is distinct from the independently measured attribute recovery reported in §5.

3.5 Using output-derived supervision

Candidate selection. Verifier Rerank-8 samples eight L4 completions for each requested skeleton and retains the one with the highest reward under Eq. 1. This changes the selected output distribution without updating the generator. Nested comparisons use the first $N \in \{1, 2, 4, 8\}$ candidates from the same pool.

Policy refinement. Output-Verifier RFT uses the group-relative advantage normalization of Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each skeleton we sample eight completions. For a completion τ, let $r(\tau)$ denote its reward for that skeleton and $\pi_\theta(\tau)$ its conditional probability. Writing $\bar{r}$ and s_r for the group mean and sample standard deviation, the group-normalized advantage is $A(\tau) = \big(r(\tau) - \bar{r}\big)/\big(s_r + \eta\big)$, where $\eta = 10^{-6}$. We then minimize a Kullback–Leibler (KL)-anchored policy-gradient loss

$$\mathcal{L}(\theta) = \mathbb{E}\big[\, -A(\tau)\, \log \pi_\theta(\tau) + \beta\, \mathrm{KL}\big(\pi_\theta(\cdot) \,\|\, \pi_{\text{ref}}(\cdot)\big)\big]\,, \tag{2}$$

with β=0.05, eight prompts per step, and temperature 1.0 for 200 steps (π_{ref} = the L4 snapshot). The executed objective departs from the population objective analyzed in Appendix C: it rescales rewards by the group standard deviation, applies a sampled-token KL penalty rather than an analytic full-vocabulary one, and omits the canonical clipped surrogate. Appendices G.3 and C.1 give the exact penalty, normalization, and optimization details. Figure 4 depicts the RFT loop.

Verifier and release scope. Real-Verifier RFT is instead configured to fit V on 4000 real training bodies. Historical realized counts were not archived (Appendix G.4). This is additional unaccounted access, so the arm is a diagnostic, not a post-processing release.

3.6 Information flow and privacy scope

The intended adjacency removes or adds all records of one stable sender key in raw corpus D. The generator accountant uses selected input $D_{\text{sel}} = T(D)$. Structural mechanisms each target $(0.5, 10^{-5})$, while the learner targets $(\varepsilon_{\text{learn}}, 10^{-5})$. "DP-LoRA," θ_{DP}, and budget-indexed names denote nominal recipes and accountant settings, not certified artifact guarantees.

In practice, historical global quotas of 400 examples per cluster and eight per exact skeleton allow one user's removal to change other users' selected records (Figure 3). Moreover, the checked local code uses fixed or recorded privacy-noise seeds, population-dependent sampling and noise settings, and unnoised diagnostic outputs. Correcting selection alone would therefore not certify the artifacts. Appendix G.2 gives calibration and identity conditions, historical-provenance limits, and the required release contract.

By contrast, generation, verifier fitting, RFT, and reranking require no further source access. They would preserve a valid joint upstream DP guarantee (Proposition 1 in Appendix C), but that premise

is unestablished for $D \mapsto (S(D), \mathcal{G}(T(D)))$ here. Independent private training, model selection, and real-data diagnostics need separate accounting or release justification. The ledger is not a public-release certificate.

4 Experimental Setup

The experiments ask whether output-derived supervision improves attribute recovery over the adapted generator, and whether evaluators independent of the training reward confirm any gain. They also ask how refinement and selection differ on recovery and fidelity, and whether the findings transfer across domain and model family. The six-seed Enron study provides the primary evidence, with confirmation from two independent judges on three of those seeds. A matched three-seed audit compares the controllers, while the Amazon and SmolLM3 screens test transfer. Utility, verifier reliability, diversity, and empirical privacy are secondary diagnostics.

4.1 Comparison methods

The frozen base shows how much instruction following is already present before adaptation. L4 is the starting point against which we measure output-derived control gains. The matched audit uses existing L4 and RFT adapters, shared requests, and fixed decoding. Original RFT verifier weights were not archived: historical checkpoints retain the generator, not the fitted verifier (Appendix G.4). Reranking uses frozen reconstructions from the same fitting corpora and recipe, not a proven byte-identical reward model (Appendix D.8.1). Real-Verifier RFT instead fits its verifier on raw real bodies, so it probes the supervision source but is an unaccounted diagnostic.

Reward-side learners provide secondary design context rather than a test of the same control stage. Reward-DP-Disc+Cluster and Reward-DP-Ground+Cluster place the private learner on a reward model rather than on the generator, the first using a real/fake discriminator and the second an attribute-recovery scorer with the same reward objective as Output-Verifier RFT. They share the base, vocabulary, structural release, and evaluation protocol, but differ in learner capacity and selected training population (Appendix D.2). Aug-PE-Raw and Aug-PE+Levers are schema-free and attribute-conditioned Aug-PE implementations, used as external references rather than equal-privacy comparisons (Table 27). Real denotes held-out real bodies used only to calibrate the independent judge.

4.2 Evaluation protocol

The no-source-access restriction applies to control training. Research evaluation on these public corpora uses real-data references and is not an accounted private release.

Attribute recovery is the fraction of requested keys recovered from a generated body by an evaluation instrument, averaged over bodies. Full-macro scores weight the keys equally, so they measure average adherence rather than uniform improvement across keys or overall text quality. The automatic labeler and two independent judges supply this measurement and never supply the RFT reward.

Fidelity compares body-inferred attribute distributions with the noised structural release using per-key and joint-cluster total-variation distance (TVD). Body metrics include radial-basis-function (RBF) kernel maximum mean discrepancy (MMD^2) (Gretton et al., 2012), precision-recall-density-coverage (PRDC) style coverage (Naeem et al., 2020) over all-MiniLM-L6-v2 embeddings (Reimers & Gurevych, 2019; Wang et al., 2020), and distinct-n diversity (Li et al., 2016). One fixed TF-IDF/logistic-regression measurement labeler, fit on real training bodies, is applied to every method. It shares the verifier's model class, so a lexical-masking audit tests whether the measured lift rests on the verifier's most heavily weighted unigrams (§5.1). Embedding values use a fixed real-only calibration set for the MMD^2 kernel bandwidth and a disjoint real reference shared across methods within a dataset (Appendix G.4).

Utility: direct train-on-synthetic/test-on-real (TSTR) macro-F1 (Esteban et al., 2017) on two rule-derived binary tasks, needs-decision and requires-response, and the three-class triage taxonomy. Synthetic and real training sets are size-matched against a shared sender-disjoint real test set. Primary results average ten deterministic subsamples within each seed. Transfer uses all 1000 bodies, so its ten fits repeat the same training set. Unless stated otherwise, synthetic training labels are post-

Table 1: Enron adaptation and control: mean ± sd over six seeds, 5000 bodies/seed and a common sampler. Recovery is automatic eight-key agreement, not the training reward. Independent judges appear in Table 2. Extended results are in Appendix Table 5. Budget labels are nominal settings (§3.6).

Method	Recovery ↑	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑
Frozen Qwen	0.429±0.001	0.150±0.003	0.289±0.008	0.0688±0.0005	0.359±0.008
L4	0.436±0.005	**0.119**±0.009	**0.134**±0.011	**0.0140**±0.0027	**0.843**±0.007
OUTPUT-VERIFIER RFT	**0.473**±0.019	0.155±0.015	0.182±0.033	0.0312±0.0107	0.804±0.017

hoc predictions from that same measurement labeler, and a paired audit repeats the analysis with the original requested labels on identical bodies and splits (Appendix D.3).

Empirical privacy: a user-level loss attack (Yeom et al., 2018) on matched member and nonmember users over two canary-free DP-recipe seeds, plus Secret-Sharer canary exposure (Carlini et al., 2019), reported as area under the ROC curve (AUC) and true positive rate (TPR) at a fixed false positive rate (FPR). Regex personally identifiable information (PII) span counts and nearest-neighbor duplication are diagnostics, not formal privacy measures (Appendix G.4). Reported paired intervals are not adjusted for multiple comparisons. Replication is at the level of training seeds rather than generated bodies, so row-level intervals describe uncertainty conditional on the trained policies. Hierarchical bootstrap intervals resample seeds and then matched rows within seeds (within frozen clusters for judge samples). They are separate interval estimates, not adjustments to the seed-level t-intervals. Exact two-sided sign-flip tests enumerate all 2^n sign assignments of the n observed paired differences rather than counting positive signs. Paired means use unrounded seed values unless noted, so rounded table differences may differ. An interval containing zero is inconclusive under the reported experiment, and we call such a comparison *unresolved* rather than evidence of equivalence.

4.3 DATA, MODELS, AND RUN CONFIGURATION

Primary experiments use $\varepsilon_{\text{learn}}=4$ on the public Enron corpus (Klimt & Yang, 2004) with a frozen `Qwen3.5-4B` checkpoint (Qwen Team, 2026a), six training seeds, the quota-selected training cohort, 5000 generated bodies per method, and 200 group-relative RFT steps. The primary verifier-source study judges seeds 0–2 with `Qwen3.6-27B` (Qwen Team, 2026b) and `GPT-5.6 Luna` (OpenAI, 2026; Microsoft, 2026), and Table 2 gives its sample counts and calibration scope. A frozen-base anchor generates the same number of bodies from the same structurally sampled skeletons, aligned row-for-row over seeds 0–5. It is a quality anchor, not an equal-learner-budget comparison. Historical decoding settings are also incompletely recorded (Appendix G.3). The matched inference audit reuses seeds 0–2 with 1000 shared requests per seed under a pinned base revision and fixed decoding, where L4 supplies nested candidate pools and RFT supplies one completion per request. `Qwen3.8-27B` (Qwen Team, 2026c) and Luna judge the same matched bodies per arm and seed (Appendix D.8.1). The transfer screens use a lower-cost recipe of 500 DP-SGD steps, contribution cap three, and 1000 bodies per method and seed: Amazon Reviews 2023 (Hou et al., 2024) with Qwen, judged additionally by `GPT-5.6 Sol` (OpenAI, 2026), and Enron with `SmolLM3-3B` (Hugging Face, 2025), each over three seeds comparing DP-LoRA, OUTPUT-VERIFIER RFT, and REWARD-DP-GROUND+CLUSTER. Seed counts for other comparisons appear in the corresponding captions, and Appendix G gives exact preprocessing, hyperparameters, prompts, per-seed values, deployment metadata, and provenance.

5 RESULTS

Output-derived RFT improves attribute control, but not uniformly or without cost. On Enron, automatic recovery improves in all six seeds, and two independent judges confirm the gain on three evaluated seeds. The largest gains occur for topic and formality, while urgency does not improve. In the matched controller comparison, RFT achieves higher judged recovery than rerank-8, whereas reranking better preserves distributional fidelity. Amazon provides independent cross-domain corroboration, and SmolLM3 supplies weaker automatic-only evidence.

Table 2: Primary Enron independent judging: three seeds, weighted to each eligible matched-corpus mixture. Qwen3.6 uses 500 bodies/arm/seed plus 300 real calibration bodies. Luna uses a fixed 200-body matched synthetic subset/arm/seed. Based on real calibration, higher-agreement keys are triage, topic, sentiment, difficulty, and formality. Lower-agreement keys are intent, response, and urgency. Full macro averages all eight, and Δ is relative to L4. Judges are separate. Dashes mean no calibration.

	Qwen3.6-27B				GPT-5.6 Luna			
Method	high-agree.	low-agree.	**full**	Δ	high-agree.	low-agree.	**full**	Δ
REAL (calibration)	0.829	0.554	0.726	–	–	–	–	–
L4	0.588	0.272	0.470	ref	0.528	0.326	0.453	ref
OUTPUT-VERIFIER RFT (DP-output verifier)	**0.713**	0.290	**0.554**	**+0.084**	**0.636**	0.281	0.502	**+0.050**
REAL-VERIFIER RFT (diagnostic)	0.670	**0.300**	0.531	+0.062	0.619	**0.310**	**0.503**	+0.051

Table 3: Matched inference audit: seeds 0–2, 1000 bodies/arm/seed, shared requests, pinned base, and fixed decoding. Fidelity is mean $\pm$ sd. Each judge uses 200 matched bodies/arm/seed, weighted to the eligible mixture, for eight-key recovery. L4 is the first pool candidate. Reranking uses frozen verifier reconstructions. Do not compare these values directly with the 5000-body primary study or its different judge evaluations. Coverage is sample-size dependent.

Method	Qwen3.8 recovery ↑	Luna recovery ↑	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑
L4	0.486	0.425	0.1218±0.0037	0.1335±0.0083	0.0130±0.0028	0.618±0.014
VERIFIER RERANK-8	0.502	0.461	0.1215±0.0023	0.1374±0.0068	0.0131±0.0032	0.623±0.005
OUTPUT-VERIFIER RFT	**0.559**	**0.531**	0.1442±0.0067	0.1544±0.0034	0.0221±0.0052	0.601±0.021

5.1 OUTPUT-DERIVED CONTROL ON ENRON

Across six matched Enron seeds (Table 1), automatic recovery increases from 0.436 to 0.473, a mean paired gain of +0.036 (95% t-CI [+0.020, +0.052], hierarchical seed/row CI [+0.025, +0.047]). All six seed-level differences are positive, and the exact two-sided sign-flip test gives $p = 0.03125$. The three-seed learner-budget sweep also has positive RFT gains in every seed at $\varepsilon_{\text{learn}}$=2 and 8 (Appendix Table 6). However, the unadapted base already achieves much of this absolute recovery level (§5.4).

Independent judges corroborate the gain (Table 2). On Qwen, OUTPUT-VERIFIER RFT improves full-macro recovery by +0.084 over L4, with a hierarchical paired 95% CI [+0.062, +0.104], and on cross-family Luna the gain is +0.050 with CI [+0.011, +0.084]. Both exclude zero.

However, the improvement is concentrated rather than uniform. Topic and formality account descriptively for about 73% of Qwen's aggregate lift and 75% of the automatic lift (Tables 25 and 38). Qwen improves on seven of eight keys but regresses on urgency. Both judges show larger gains in the higher-agreement group. Luna's lower-agreement group regresses while Qwen's improves slightly (Table 2). Weak automatic measurement on intent, response, and urgency further limits interpretation (Appendix G.4).

The gain is also not explained by the selected lexical cues alone. Masking the verifier's highest-weight unigrams retains 82–87% of the automatic lift across six seeds with an IDF-matched control, although other shortcuts remain possible (Appendix G.4). Qwen also favors output-derived over real-verifier RFT, while Luna leaves that difference unresolved. This does not establish equivalence to real labels (Appendix D.2). The verifier diagnostic measures net correction as verifier–judge agreement minus requested-tag–judge agreement. On L4, this difference is unresolved under Luna and negative under Qwen. Yet judged recovery is higher when the verifier agrees with the requested tag (Appendix D.8.2). A verifier can therefore provide useful control feedback without being a reliable relabeling mechanism.

5.2 Different operating points for refinement and selection

RFT and reranking occupy different control–fidelity operating points. Both independent judges favor RFT over rerank-8 on recovery (Table 3): Qwen3.8 by $+0.0568$ (hierarchical paired 95% CI $[+0.0153, +0.0958]$) and Luna by $+0.0703$ ($[+0.0414, +0.0991]$), with positive differences in every seed. RFT also improves over L4 under both judges, while rerank-8's judged lift over L4 remains unresolved. The RFT–rerank ordering persists under a conservative invalid-response sensitivity check (Appendix D.8.1).

Rerank-8 has better fidelity point estimates in every seed, though some intervals include zero. It requires eight candidates per output, whereas RFT generates one after upfront training, without an established amortization break-even (Appendix D.8).

5.3 Independent corroboration on Amazon

Table 4: Amazon/Qwen short-budget screen: three seeds, 1000 bodies/arm/seed. Automatic metrics are mean ± sd. Sol recovery uses 200 judged bodies/arm/seed, weighted to the eligible mixture. Both recovery columns average all six keys. Full results are in Appendix G.7.

Method	Auto recovery ↑	GPT recovery ↑	joint-TVD ↓	Cov ↑
DP-LoRA	0.575±0.004	0.671	**0.158**±0.015	**0.713**±0.020
Output-Verifier RFT	**0.690**±0.012	**0.768**	0.250±0.003	0.672±0.019

Amazon tests whether useful output-derived supervision extends beyond email attributes. Across three Amazon seeds, Output-Verifier RFT improves automatic recovery over DP-LoRA by $+0.114$ (paired 95% t-CI $[+0.083, +0.145]$), with all changes positive (Table 4). Joint-TVD worsens by 0.092 and coverage by 0.041. Relative to Reward-DP-Ground+Cluster, every seed favors Output-Verifier RFT on recovery and all reported fidelity axes. The GPT-5.6 Sol judge corroborates this direction: hierarchical paired lifts are $+0.097$ ($[+0.077, +0.117]$) over DP-LoRA and $+0.076$ ($[+0.055, +0.097]$) over Reward-DP-Ground+Cluster, with zero invalid responses among 1800 judgments. The attribute-level pattern differs from Enron: the judge's noisy-key group covering star rating and review intent also improves, from 0.543 to 0.646, alongside the reliable-key group at 0.735 to 0.829 (Table 40). These runs use the shorter training budget, so they corroborate the Enron direction on a second domain rather than replicate it at matched budget. Star-rating utility differences remain unresolved.

Model-transfer boundary. The separate Enron/SmolLM3 screen provides a weaker, automatic-only result: recovery improves by $+0.031$ in three positive seeds, but coverage falls by 0.110 and needs-decision TSTR by 0.025. Compared with Reward-DP-Ground+Cluster, MMD^2 improves by 0.047 and coverage by 0.253, while the recovery difference is unresolved. These short-budget gains are model-dependent and do not establish uniform superiority to the initialization (Table 26, Appendix G.7). Its RFT outputs also contain 585 to 2212 address-like spans per 1000 bodies across seeds, an output-quality failure rather than a privacy measure (Appendix E).

5.4 The fidelity cost of improved control

Across matched seeds 0–5, DP-LoRA improves automatic recovery over frozen Qwen by only $+0.0070$ (paired 95% t-CI $[+0.0015, +0.0126]$), so much of the measured recovery is already present in the prompted frozen base. Its distributional effect is far larger: mean-TVD decreases by 0.0304, cluster-joint TVD by 0.1550, and MMD^2 by 0.0549, while coverage rises by 0.4843, and all four paired intervals exclude zero. These adaptation gains are the starting point for control, not evidence attributable to the output-derived verifier.

In turn, RFT trades part of that fidelity for control: all four measures worsen relative to L4 (Table 1), with paired six-seed intervals excluding zero. Coverage changes by -0.039 (CI $[-0.054, -0.025]$). Compared with the frozen base, however, Output-Verifier RFT retains a recovery advantage of $+0.0434$ ($[+0.0235, +0.0632]$) and retains the large cluster, embedding, and coverage gains, while its mean-TVD difference from that base is unresolved ($+0.0049$, $[-0.0117, +0.0215]$). The frozen-base comparison measures how much of adaptation's benefit survives RFT.

The frozen base also emits far fewer regex PII-like spans (1394 against 9812 and 10204 per 5000 bodies for L4 and OUTPUT-VERIFIER RFT), so adaptation increases identifier-like surface forms. A matched overlap audit nevertheless finds no training-specific verbatim excess over held-out text (Appendices D.5 and G.6). This is an output-safety cost rather than evidence of memorization.

Changing the RFT trade-off. The KL sweep tests whether stronger anchoring reduces the fidelity cost of policy updates. Across three matched seeds, recovery gains at $\beta \in \{0.05, 0.10, 0.20\}$ are $+0.045$, $+0.037$, and $+0.019$. Increasing β from 0.05 to 0.10 improves every fidelity mean and retains about 82% of the recovery lift, but lowers the recovery point estimate from 0.485 to 0.478 (Table 13). Neither point Pareto-dominates the other across control and fidelity. The pre-sweep β=0.05 remains primary, with six-seed automatic evaluation and the predeclared three-seed judge evaluation, avoiding post-selection from the sweep.

5.5 DOWNSTREAM LIMITS AND PARTIAL PRIVACY EVIDENCE

We compare RFT-minus-L4 TSTR macro-F1 under requested and post-hoc labels using the same bodies and real test splits. With requested labels, the Enron needs-decision difference remains unresolved ($+0.011$, unadjusted 95% t-CI $[-0.006, +0.028]$), while triage decreases in all six seeds (-0.022, $[-0.038, -0.005]$). Transfer results are mixed: across seven setting–task pairs, intervals for the change in the RFT effect between label sources all include zero. For Enron triage, requested tags are weaker training targets than post-hoc predictions: macro-F1 falls from 0.574 to 0.368 on L4 and from 0.555 to 0.346 after RFT. Appendix D.3 reports both labeling protocols and the complete paired audit.

Additionally, separate diversity, membership, canary, and lexical-overlap audits are reported in Appendix D. They show no duplicate-body collapse in the primary RFT corpora. In the tested loss attacks, DP-recipe checkpoints show much less membership signal than the readily attacked non-private positive controls, although their membership AUC remains above chance. These diagnostics do not establish PII reduction or absence of leakage, and they do not certify the upstream implementation or its privacy accounting.

Supporting comparisons. Reward-side learning, the learner-budget sweep, component ablations, and Aug-PE references appear in Appendix D. The private learner's placement determines the control stage's initialization. We therefore report these comparisons as design context, not as an isolated privacy-budget-placement advantage.

6 CONCLUSION

Generated text paired with requested attributes can provide useful supervision for attribute control after source access ends. A body-only verifier trained on these pairs improves independently judged recovery when used for reinforcement fine-tuning on Enron and Amazon. The signal is imperfect: gains vary by attribute, do not establish a general downstream utility benefit, and reduce distributional fidelity. RFT and reranking provide different operating points, with RFT favoring attribute adherence and reranking better preserving fidelity. Requested attributes therefore need not be reliable body labels to provide useful control feedback, but the resulting verifier should be treated as a ranking signal rather than an oracle or relabeling mechanism.

The no-further-source-access control stage has a conditional post-processing form. It does not establish the missing guarantee for the historical upstream artifacts.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment pol-

icy. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2401.01879.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Young Hyun Cho and Will Wei Sun. Privacy-preserving reinforcement learning from human feedback via decoupled reward modeling. *arXiv preprint arXiv:2603.22563*, 2026. URL https://arxiv.org/abs/2603.22563.

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

Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017. URL https://arxiv.org/abs/1706.02633.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. arXiv:2210.10760.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. *arXiv preprint arXiv:2403.03952*, 2024. URL https://arxiv.org/abs/2403.03952.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Hugging Face. SmolLM3-3B. Hugging Face model card and technical blog, 2025. URL https://huggingface.co/HuggingFaceTB/SmolLM3-3B. Accessed 28 August 2026.

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004*, volume 3201 of *Lecture Notes in Computer Science*, pp. 217–226. Springer, 2004. doi: 10.1007/978-3-540-30115-8_22.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016. doi: 10.18653/v1/N16-1014.

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

Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 1: Images. In *International Conference on Learning Representations (ICLR)*, 2024.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schölkopf, and Mrinmaya Sachan. Differentially private language models for secure data sharing. *arXiv preprint arXiv:2210.13918*, 2022.

Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11(3), 2021. arXiv:2108.04978.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. AIM: An adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment*, 15(11): 2599–2612, 2022. arXiv:2201.12677.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1710.06963.

Microsoft. Azure openai in microsoft foundry models: Working with models. Microsoft Learn, 2026. URL https://learn.microsoft.com/en-us/azure/foundry/openai/how-to/working-with-models. Accessed 18 August 2026.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

OpenAI. GPT-5.6 Luna model documentation. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models. Accessed 18 August 2026.

Qwen Team. Qwen3.5-4B. Hugging Face model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-4B. Accessed 18 August 2026.

Qwen Team. Qwen3.6-27B. Hugging Face model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.6-27B. Accessed 18 August 2026.

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed 28 August 2026.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Bowen Tan, Zheng Xu, Eric Xing, Zhiting Hu, and Shanshan Wu. Synthesizing privacy-preserving text data via finetuning without finetuning billion-scale LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58637–58657. PMLR, 2025. URL https://proceedings.mlr.press/v267/tan25g.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Fan Wu, Huseyin A. Inan, Arturs Backurs, Varun Chandrasekaran, Janardhan Kulkarni, and Robert Sim. Privately aligning language models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A. Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. Differentially private synthetic data via foundation model APIs 2: Text. In *International Conference on Machine Learning (ICML)*, 2024.

Fangyuan Xu, Sihao Chen, Zinan Lin, Taiwei Shi, Sydney Graham, Pei Zhou, Mengting Wan, Alex Stein, Virginia Estellers, Charles Chen, Morris Sharp, Richard Speyer, Tadas Baltrusaitis, Jennifer Neville, Eunsol Choi, and Longqi Yang. DP-RFT: Learning to generate synthetic text via differentially private reinforcement fine-tuning. *arXiv preprint arXiv:2602.18633*, 2026. URL https://arxiv.org/abs/2602.18633.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, 2021. doi: 10.18653/v1/2021.naacl-main.276. URL https://aclanthology.org/2021.naacl-main.276/.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11653–11669, 2022. doi: 10.18653/v1/2022.emnlp-main.801. URL https://aclanthology.org/2022.emnlp-main.801/.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

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 (ICLR)*, 2022. arXiv:2110.06500.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Xiang Yue, Huseyin A. Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1321–1342, 2023. doi: 10.18653/v1/2023.acl-long.74. URL https://aclanthology.org/2023.acl-long.74/.

Jiaming Zhang, Mingxi Lei, Meng Ding, Mengdi Li, Zihang Xiang, Difei Xu, Jinhui Xu, and Di Wang. Towards user-level private reinforcement learning with human feedback. *arXiv preprint arXiv:2502.17515*, 2025. URL https://arxiv.org/abs/2502.17515.

APPENDIX CONTENTS

A Statements **14**

B Extended related work **14**

C Analysis and proofs **16**
C.1 Properties of the group-relative objective 20

D Supporting comparisons and audits **21**
D.1 Statistical conventions 21
D.2 Reward-side learners and budget context 21
D.3 Task-dependence of matched-size utility 22
D.4 No duplicate-body collapse, and an unresolved PII change 24
D.5 Positive-control-calibrated empirical privacy audit 24
D.6 Effect of lever conditioning in Aug-PE 26
D.7 KL-weight Pareto audit 26
D.8 Matched control comparison and historical audit 27
D.9 Historical verifier-guidance cost 32
D.10 Ablations 32

E Limitations **33**

F Supplementary figures and result tables **36**

G Reproducibility and Supplementary Details **39**
G.1 Dataset profiles and preprocessing 39
G.2 Privacy accounting and release semantics 41
G.3 Training and generation configuration 42
G.4 Verifier and measurement instruments 44
G.5 Verifier-model assumptions 46
G.6 Frozen-base quality control 46
G.7 External-validity results by seed 46
G.8 Experiment inventory and compute 47
G.9 Reproduction commands and artifact provenance 47
G.10 Planned artifact release 50

A STATEMENTS

Ethics statement. Enron email and Amazon Reviews 2023 are public research corpora, but their records were authored by people and can contain personal or identifying content. We use the configured sender/user key as the intended privacy unit, cap contributions, avoid persisting raw examples in aggregate reports, and evaluate memorization and PII-like output patterns alongside nominal component accounting. The historical raw-corpus release has no established end-to-end DP bound (§3.6). Differential privacy does not prevent plausible personal information from appearing in generated text: the SmolLM3 transfer in particular produces many address-like strings and should not be released without further filtering. Any artifact release will respect source licenses and exclude credentials, raw user identifiers, and raw source corpora.

Reproducibility statement. Appendix G records dataset construction, schema versions, privacy accounting, training and generation hyperparameters, evaluation instruments, per-seed transfer values, compute, commands, and provenance. The supplementary submission will include anonymized code, configurations, tests, and curated aggregate reports needed to reconstruct the tables, subject to the specific release clearance in Appendix G.10. Diagnostic reports are not automatically DP or public-safe. Raw corpora are not redistributed. Scripts retrieve public inputs from canonical hosts and deterministically rebuild the processed profiles.

Privacy terminology. Historical method names such as DP-LoRA and Reward-DP, and empirical references to "DP" training, outputs, or structural releases, describe nominal mechanisms, training designs, and accountant settings. They do not certify artifact-level or end-to-end privacy. Numerical budgets are targets or accountant values under their stated assumptions. The formal statements retain their usual mathematical meaning, conditional on their explicit DP hypotheses. Implementation and release qualifications are collected in Appendix G.2.

Limitations statement. The component ledger does not establish an end-to-end user-level DP bound for the raw corpus. The selection quotas are not user-decomposable, and checked local randomness, calibration, parameter-selection, and report-output conditions require additional qualification (§3.6, Appendix G.2). The evidence also concentrates on one dataset and model pairing, with small seed counts and a shorter training budget for the transfer screens. Shared measurement instruments, weak labeler accuracy on several keys, and unadjusted intervals further limit what the comparisons resolve, and no general downstream utility advantage is established. Appendix E gives the full account.

AI use statement. Generative AI tools helped design and critique experiments, implement and debug software, structure data, analyze results, and summarize related work. The authors used these tools both to draft sections of the paper and to aid or polish writing throughout. Language models were also used as experimental generators, independent judges, and a fixed semantic annotation instrument where described. The authors reviewed the AI-assisted work, verified quantitative claims against persisted artifacts, tested the released code paths, and take responsibility for the final text, claims, and artifacts.

B EXTENDED RELATED WORK

This section gives the detailed comparisons behind §2.

Generated labels as control supervision. Ye et al. (2022) generate task datasets with pretrained language models and train smaller task models on them. For classification, the requested class supplies the synthetic example's label. Our verifier uses that kind of supervision, but returns its predictions as feedback to the adapted generator that produced its fitting corpus. The distinction lies in how the learned scorer is used and evaluated, not in the idea of training a classifier on generated examples. A requested label can be wrong about the body, so a successful fit to these pairs alone would not establish semantic control. The independent-judge comparisons and lexical-masking audit test whether the resulting control gains extend beyond the training reward. They do not establish a general denoising mechanism.

Learned feedback and classifier guidance. Self-Rewarding Language Models (Yuan et al., 2024) use a language model as a judge to provide feedback during iterative direct preference optimization. Our fitting labels are instead the generator's categorical requests: no judge annotates the verifier-training bodies, and the judges used for independent evaluation do not supply control feedback. FUDGE (Yang & Klein, 2021) learns future-attribute predictors from partial sequences and combines their scores with language-model probabilities during decoding. Our verifier reads completed bodies and is used for candidate selection or policy refinement. These connections place the study within synthetic supervision and classifier-guided control. We test whether the output-derived scorer is useful under the source-access restriction, evaluating the control procedure and its costs separately.

DP fine-tuning and parameter efficiency. Kurakin et al. (2023) report that prompt-tuning and LoRA outperform full fine-tuning when the tuning is performed privately, by up to eleven points of downstream accuracy, reversing the usual non-private ordering. We therefore tune only adapters and keep the base weights frozen. They also identify a hazard relevant here: earlier DP synthetic-text work did not account for what the pre-trained model had already seen. They de-duplicate their pre-training corpus against sensitive data, but we do not. Our audits therefore probe memorization using a user-level loss membership attack and a Secret-Sharer canary probe (Appendix D.5). They cover those specified attack scenarios, not knowledge that predates private training (Appendix E). Their privacy unit is a 512-token sequence, whereas our intended unit is a configured sender key, so their guarantee and our nominal accountant targets are not directly comparable. McMahan et al. (2018) introduced user-level DP for language models and observed that reaching a user-level guarantee from an example-level one by group privacy costs a prohibitive ε. This distinction does not resolve the selection issue in our pipeline: the per-user cap of ten bounds a user's own retained contribution, not the number of other users whose selected examples can change through global-quota cascades (§3.6). No calibrated weaker historical raw-user bound is established by multiplying the learner target by that cap.

Attribute conditioning and structural release. Multi-attribute conditioning is established in this literature. Mattern et al. (2022) formalize a sample as a set of M categorical attributes, each with its own value set. Yue et al. (2023) prepend composite control codes combining product type and review score. Our eight-key skeleton is a larger evaluated schema, which is a descriptive difference rather than a measured cross-paper difficulty ranking. Yue et al. (2023) also observe that the categorical distribution of the source data is itself potentially sensitive, and state that its estimation could be privatized but that they use the exact distribution for simplicity. Our structural recipe allocates a nominal budget to that estimation, which is the substantive difference in the conditioning interface. *DP marginal synthesis* supplies the other half of the lineage. MST (McKenna et al., 2021) and AIM (McKenna et al., 2022) select a set of marginal queries, measure them under a noise-addition mechanism, and post-process the noisy measurements into records. Our structural release is that template with a public cluster geometry in place of a learned graphical model, and the body is then written by a language model fitted with the DP-training recipe instead of being sampled from the reconstruction. Both mechanisms treat query selection as data-dependent and pay for it with the exponential mechanism, whereas our generator's training-set quota does not (§3.6).

Public clustering and conditional private synthesis. CTCL (Tan et al., 2025) is a close precedent for the broad upstream architecture. It develops a topic model from public Wikipedia documents and a lightweight conditional generator from public text, then combines a DP private-topic histogram with DP generator fine-tuning. Its generation is allocated according to the topic histogram. Our partition instead enumerates public attribute-value combinations. DP occupancies and one-way marginals guide explicit multi-attribute requests. The language model receives those tags, not a cluster ID or centroid. This is a control-representation and sampling-design distinction, not a new DP mechanism. The cluster sampler chooses which controls to request, whereas the downstream output-derived verifier learns feedback from the generated bodies paired with those requests. We do not present a matched performance comparison with CTCL. Such a comparison would require refitting its public topic model and conditional generator to our profiles and eight-key schema, which we do not do.

The endpoint comparison in full. Wu et al. (2024) account for private training stages and use DP-PPO for policy optimization. Their RLHF setting additionally uses DP reward learning. Their

argument for why that reward must be a DP artifact is the one we rely on: with a non-private reward, every minibatch gradient of the policy stage becomes a function of the entire reward-training set, which invalidates privacy amplification by subsampling. Our verifier condition is that same requirement. Cho & Sun (2026) move the budget entirely onto reward learning and derive the policy as post-processing of the private reward, implemented as best-of-N re-ranking of reference samples. Their motivation is not only that preferences are sensitive. It is also a utility argument: noisy updates to policy parameters can return a policy worse than the reference, so the reward model buffers the privacy mechanism from the policy.

Our design differs along two axes: how the reward is learned and where private learning occurs. The second matters because it determines the artifact available to the control stage. First, our reward is a lever-recovery check (Eq. 1) fit on generated bodies and requested tags. It is a *learned* classifier rather than a public programmatic checker such as a parser or a unit test, so it inherits the generator's error instead of being error-free. What it does not do is require further raw-record access, unlike a reward model trained on real preference or label data (Zhang et al., 2025; Cho & Sun, 2026). Xu et al. (2026) place that cost differently. Their similarity reward queries the private corpus at every update under the Private Evolution accounting, so their training stage spends budget continuously rather than once. Second, we place the private learner on the generator, whereas Cho & Sun (2026) place the budget on reward learning. Our frozen-base anchor shows that supervised domain adaptation on real bodies improves cluster fidelity, embedding fidelity, and coverage over the unadapted reference (§5), and our reward-side arm is a recovery-reward configuration rather than an implementation of their best-of-N inference rule. The two settings differ in supervision, capacity, and selected population, so this comparison does not test their private-alignment argument. Conditional on a valid joint upstream DP release, privacy preservation for the RFT stage follows from post-processing, a construction Cho & Sun (2026) also use when they derive a policy from a privately fitted reward. What we add is the fitting source, the stricter statement of when the control stage requires no further raw access, and an audit of what that stage does and does not buy. We do not establish that a one-pass policy amortizes reranking in deployment (Appendix D.9). For this post-processing argument, a verifiable reward alone is not enough. The reward *and* the prompts must both be functions of public or upstream artifacts. This is why Wu et al. (2024) must account for access to raw RL prompts.

The endpoint comparison also highlights a potential failure mode of learned rewards. The discriminator variant's optimization reward rises while its distributional fidelity falls, a pattern consistent with reward overoptimization (Appendix D.6). It also underperforms AUG-PE+LEVERS on most metrics (§5). These dynamics resemble the regressional-Goodhart regime of Gao et al. (2023), in which a proxy reward carrying additive noise encourages optimization to select for that noise. Our dynamics do not separate privacy noise from reward capacity or supervision, so this is a statement about that particular reward, not a demonstration that reward-side placement must fail. We include this comparison because similar failures could affect any learned reward, including our own. Appendix D.8.2 reports the corresponding check for the output-derived verifier.

C ANALYSIS AND PROOFS

Proposition 1 specializes the post-processing theorem to our pipeline and has the same form as Proposition 6 of Cho & Sun (2026). Proposition 2 follows from the Gibbs variational principle, and the same authors give a related reach bound in multiplicative form. Remark 5 uses a known best-of-N identity, and Proposition 6 is a standard label-noise decomposition whose consequence for verifiers fit on DP output is what is specific to this work. Write $M(D) = (\theta_{\mathrm{DP}}, S)$ for the upstream release, with θ_{DP} the adapter and S the structural release. The statement below is conditional on the joint mechanism being DP under the specified protected-input adjacency. Separate marginal claims alone do not establish that premise. That hypothesis is not established for the historical raw-corpus map $D \mapsto (S(D), \mathcal{G}(T(D)))$.

For raw corpus D, write M for the joint upstream mechanism whose output $M(D) = (\theta_{\mathrm{DP}}, S)$ collects the trained adapter and the structural release, and let U be auxiliary randomness drawn independently of the dataset D and of the internal random coins M uses to produce its output.

Proposition 1 (Conditional privacy preservation by RFT)**.** *Suppose M is (ε, δ)-DP at the user level with respect to a stated adjacency on its input, and that* (i) *the verifier V,* (ii) *the distribution over RFT prompts, and* (iii) *the RFT initialization, base policy, and update rule are all measurable*

functions of $M(D)$ and U. Then the RFT policy π_θ, the corpus it emits, and the joint release $(S, \theta_{\mathrm{DP}}, V, \pi_\theta)$ are each (ε, δ)-DP under that same adjacency.

Independent public initialization is permitted. Raw prompts or unaccounted data-dependent checkpoints require separate privacy analysis (Wu et al., 2024). The guarantee remains conditional on the upstream release (§3.6). The statement makes explicit which pipeline objects must be measurable functions of the release, providing a checklist for the control-stage design. For the executed pipeline the learner accountant was calibrated on selected-input parameters, but S is derived from the raw corpus. This does not establish the joint DP hypothesis on D_{sel} or on the raw corpus, nor certify the evaluated artifacts under the implementation conditions in Appendix G.2. Proposition 1 formalizes the construction in §3.5 and is proved below. Its condition (iii) is not redundant. Without it one could satisfy (i) and (ii) with a constant verifier and a public prompt law while initializing the policy at a non-privately fine-tuned checkpoint, and the conclusion would fail. In our pipeline (iii) holds because RFT is initialized at θ_{DP}. The randomness U in the statement is independent of D and of the internal randomness of M. Throughout, $\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$.

The statements from here to Corollary 9 are population-level idealizations. They characterize exact maximizers and exact population fits, clarifying qualitative behavior rather than describing the finite-step, finite-sample system we run. The specific gaps between the two are recorded after Corollary 9.

For the next two results fix a reference distribution π_{ref} over generated bodies, a bounded reward r, and a weight $\beta > 0$, and let $J(\pi) = \mathbb{E}_\pi[r] - \beta\,\mathrm{KL}(\pi \,\|\, \pi_{\mathrm{ref}})$ over distributions absolutely continuous with respect to π_{ref}. All quantities are at a fixed requested skeleton s, so π_{ref} means $\pi_{\mathrm{ref}}(\cdot \mid s)$ and r means $r(\cdot, s)$. Averaging over the skeleton law gives the pipeline-level statement, because the divergence between joint laws sharing that marginal is the expected conditional divergence.

Proposition 2 (Displacement is bounded by the reward it buys)**.** *Let π^* maximize J, and let $\Delta r = \mathbb{E}_{\pi^*}[r] - \mathbb{E}_{\pi_{\mathrm{ref}}}[r]$ be the reward gain attained by that maximizer. Then $\pi^*(\cdot) \propto \pi_{\mathrm{ref}}(\cdot)\exp\{r(\cdot)/\beta\}$ and*

$$\mathrm{KL}(\pi^* \,\|\, \pi_{\mathrm{ref}}) \;\le\; \frac{\Delta r}{\beta}, \qquad \|\pi^* - \pi_{\mathrm{ref}}\|_{\mathrm{TV}} \;\le\; \sqrt{\frac{\Delta r}{2\beta}}. \tag{3}$$

Corollary 3 (Pushforward and kernel consequences)**.** *Under the hypotheses of Proposition 2, for every measurable statistic h of a generated body, $\|h_\# \pi^* - h_\# \pi_{\mathrm{ref}}\|_{\mathrm{TV}} \le \sqrt{\Delta r/2\beta}$. If k is a positive-semidefinite kernel with $k(x,x) \le 1$ for all x, then $\mathrm{MMD}^2(\pi^*, \pi_{\mathrm{ref}}) \le 2\Delta r/\beta$.*

Corollary 4 (Reach of the exact optimizers)**.** *Let π^*_{gen} and π^*_{pub} be the maximizers of J when π_{ref} is set to $\pi_{\theta_{\mathrm{DP}}}$ and to π_{pub} respectively, with attained gains Δr and $\Delta \tilde{r}$. Then by Eq. 3 each maximizer stays within total-variation distance $\sqrt{\Delta r/2\beta}$ and $\sqrt{\Delta \tilde{r}/2\beta}$ of its own reference. The two instances differ only in the reference supplied to J.*

Proposition 2 and Corollaries 3 and 4 bound displacement from the reference supplied to J. Whether either reference is close to the real corpus is therefore an empirical question rather than a consequence of them: that $\pi_{\theta_{\mathrm{DP}}}$ matches the Enron distribution better than π_{pub} is a finding (§5.4). The bounds are also loose at our operating point, but we cannot quantify that looseness because Δr is the exact optimizer's reward gain, which we do not measure. The bounds do tighten monotonically as β grows, which is the direction the KL sweep measures, but they do not predict its magnitudes. The executed arms are also not exact Gibbs tilts of either reference, and they differ in reward, supervision, capacity, and training population. The bounds describe a qualitative direction: a policy that buys reward against a KL anchor moves away from its own reference, and by less as β grows. The KL sweep is consistent with that direction (Table 13), although it measures fidelity rather than displacement from the reference.

Remark 5 (Best-of-N on the same axis). Best-of-N selection from π_{ref} is a different approximation to the same reward–KL frontier. For a reward whose law under π_{ref} is atomless it satisfies $\mathrm{KL}(\mathrm{BoN}_N \,\|\, \pi_{\mathrm{ref}}) = \ln N - \frac{N-1}{N}$, which at $N{=}8$ is $\ln 8 - \frac{7}{8} \approx 1.204$ nats. Beirami et al. (2025) show that in general this expression is an *upper bound* rather than an identity. Our reward is discrete, so under uniform tie-breaking the realized best-of-8 budget is at most $\ln 8 - \frac{7}{8}$, which follows from the KL chain rule and data processing, since uniform tie-breaking leaves the within-tie conditional equal to π_{ref}.

The comparison places two idealized policies on one axis. Best-of-N need not realize the Gibbs policy at our β, the two implementations need not have equal realized KL, and KL is not a common

deployment-cost scale. We do not measure either policy's realized KL, and Gao et al. (2023) report that RL consumes far more KL than best-of-N for the same reward gain. The axis matters here because both controllers can be read as buying reward against a KL anchor, so the control–fidelity trade in §5.2 is expected of both rather than particular to RFT.

Fix a key with m possible values and model the fitting corpus as follows: given the requested value $y \sim \pi$, the body x is drawn from $P(\cdot \mid y)$ with probability p and from a value-independent distribution Q with probability $1 - p$. Write $P(y \mid x)$ for the posterior under faithful labeling, $P(x) = \sum_y \pi(y)P(x \mid y)$ for the faithful x-marginal, and

$$\lambda(x) = \frac{p\,P(x)}{p\,P(x) + (1-p)\,Q(x)}.$$

Proposition 6 (Posterior shrinkage under a label-noise model)**.** *The predictor minimizing an* unweighted *strictly proper loss over a class containing the true conditional satisfies*

$$\tilde{P}(y \mid x) = \lambda(x)\,P(y \mid x) + \big(1 - \lambda(x)\big)\pi(y). \tag{4}$$

Corollary 7 (Argmax invariance under shrinkage)**.** *Fix x with $\lambda(x) > 0$. If π is uniform then $\arg\max_y \tilde{P}(y \mid x) = \arg\max_y P(y \mid x)$. For general π the two agree whenever the faithful posterior margin exceeds a prior-spread term,*

$$P(y^* \mid x) - \max_{y \neq y^*} P(y \mid x) \;>\; \frac{1-\lambda(x)}{\lambda(x)}\Big(\max_y \pi(y) - \pi(y^*)\Big), \qquad y^* = \arg\max_y P(y \mid x). \tag{5}$$

Proposition 6 analyzes an unweighted fit. The implemented verifier uses class-balanced logistic regression (Appendix G.4), and weighting changes the population minimizer: for positive class weights w_y, minimizing weighted log loss over a class containing the weighted conditional gives $q^*_w(y \mid x) \propto w_y \tilde{P}(y \mid x)$, which is generally not $\tilde{P}(y \mid x)$. Proposition 6 and Corollary 7 therefore do not by themselves describe the executed instrument.

For the balanced case, fix one key with m possible values. Assume the requested-value prior satisfies $\pi(y) > 0$ for every y and the fitting joint is $\tilde{P}(x, y) = \pi(y)\big[p\,P(x \mid y) + (1-p)Q(x)\big]$ with mixture probability $p \in (0, 1]$ common across values and failure law Q independent of the requested value. All densities are taken with respect to a common dominating measure. We restrict attention to x with $\sum_z P(x \mid z) > 0$. Define

$$P_{\mathrm{bal}}(y \mid x) = \frac{P(x \mid y)}{\sum_z P(x \mid z)}, \qquad \overline{P}(x) = \frac{1}{m}\sum_z P(x \mid z),$$

$$\lambda_{\mathrm{bal}}(x) = \frac{p\,\overline{P}(x)}{p\,\overline{P}(x) + (1-p)Q(x)}.$$

Let q^*_{bal} minimize *unregularized weighted population log loss* with weights proportional to $1/\pi(y)$, over a model class containing the normalized weighted conditional $y \mapsto w_y \tilde{P}(y \mid x) / \sum_z w_z \tilde{P}(z \mid x)$.

Proposition 8 (Inverse-prior weighting under the label-noise model)**.** *Under the assumptions above,*

$$q^*_{\mathrm{bal}}(y \mid x) = \lambda_{\mathrm{bal}}(x)\,P_{\mathrm{bal}}(y \mid x) + \frac{1 - \lambda_{\mathrm{bal}}(x)}{m}. \tag{6}$$

Corollary 9 (Balanced argmax invariance)**.** *Wherever $\lambda_{\mathrm{bal}}(x) > 0$, $\arg\max_y q^*_{\mathrm{bal}}(y \mid x) = \arg\max_y P_{\mathrm{bal}}(y \mid x)$ as sets, ties included.*

The comparison in Eq. 6 is with P_{bal}, not with the faithful posterior under the original nonuniform prior. Inverse-prior weighting cancels π, so the shrinkage target becomes uniform rather than the requested-value prior. The statement concerns population fits, not agreement between the two finite fitted instruments. That cancellation requires a single adherence rate. Under value-dependent adherence p_y the weighted joint is proportional to $p_y P(x \mid y) + (1 - p_y)Q(x)$, whose second term still varies with y. The floor in Eq. 6 is then not constant in y, and Corollary 9 does not follow. Under the model, these results explain why output-derived supervision can work: a value-independent labeling failure shrinks the posterior toward a floor without reordering it, so a scorer fit on noisy requested tags can still rank attribute values correctly. Value-dependent adherence is the boundary at which that explanation stops.

Scope of these statements. Our implementation takes 200 stochastic steps with group-normalized advantages. The resulting objective is not an unbiased estimator of J, and we prove no optimization or approximation result connecting the implemented procedure to the exact population maximizer. The realized policy therefore does *not* inherit the stated bounds. Substituting a measured recovery lift for the theorem's reward-gain parameter would also mismatch the statement, since Δr refers to the exact optimizer's gain under the optimized reward. The displacement bounds are also informative only while $\Delta r < 2\beta$, because the total-variation bound in Eq. 3 exceeds one otherwise. At the primary $\beta = 0.05$ that threshold is $\Delta r = 0.1$. Propositions 6 and 8 idealize an unfaithful body as carrying no information about the requested value, whereas prompted failures are plausibly value-dependent and asymmetric. They assume a well-specified class containing the relevant conditional, exact population weights, and no regularization, whereas the executed verifier is a regularized finite-sample TF-IDF logistic model with empirical class weights. The conditional $P(\cdot \mid y)$ is the synthetic body law, which suits our use because the verifier only ever scores synthetic bodies, and automatic recovery is not a direct measurement of the latent faithfulness parameter p. Establishing any of this for the group-normalized finite-step objective we actually run is left open. Appendix G.5 records which of these model parameters are latent and which per-key quantities are actually traceable to measurements.

The diagnostic in Appendix D.8.2 measures agreement with independent judges on newly generated bodies, not the latent noise law or the balanced faithful posterior. It finds no aggregate evidence of net requested-label error correction by the reconstructed verifier. This neither validates nor refutes the idealized identities above, but supplies no empirical confirmation of a denoising explanation for the executed control gains. The reported control gains therefore have a coherent idealized explanation, but that explanation is untested on the executed system. The empirical audits, rather than these statements, support the claim.

Conditions of Proposition 1. The three conditions and the upstream DP hypothesis are what the post-processing argument uses. They are sufficient, not necessary for every private-learning procedure. If RFT prompts are drawn from raw records, that is a new unaccounted access and the argument below does not cover it, so such a design needs its own accounting or a separate privacy analysis. One setting analyzed by Wu et al. (2024) aligns on private prompts with a reward independent of those data, while still privatizing the policy update. Their RLHF setting additionally trains the reward model with DP. Condition (i) is the requirement identified by the same work: a non-private reward makes every minibatch gradient a function of the entire reward-training set and destroys privacy amplification by subsampling. Our skeletons, verifier, and initialization depend on upstream artifacts, so the RFT stage has the required post-processing form. This does not prove the separate hypothesis that the joint upstream release is DP relative to the raw corpus.

Proposition 1. By assumption there are measurable maps g_V, g_P and randomness U independent of D with $V = g_V(M(D), U)$ and prompt law $P_{\text{prompt}} = g_P(M(D), U)$, and by (iii) the update rule is likewise such a map. Argue by induction over RFT steps. The initialization $\theta_0 = \theta_{\text{DP}}$ is a coordinate of $M(D)$. If $\theta_t = f_t(M(D), U)$, then the rollouts sampled from π_{θ_t} at prompts drawn from P_{prompt}, their rewards under V, and hence θ_{t+1}, are all measurable functions of $(M(D), U)$. Hence $\pi_\theta = f(M(D), U)$ for a measurable f, and likewise for the emitted corpus and for the tuple $(S, \theta_{\text{DP}}, V, \pi_\theta)$, whose components are all functions of the same draw $M(D)$. Closure of (ε, δ)-DP under data-independent randomized maps gives the claim. Note the tuple is covered because it is a single function of one release, not a composition of several releases. □

Proposition 2. Let $Z = \mathbb{E}_{\pi_{\text{ref}}}[e^{r/\beta}]$, $\psi = \beta \log Z$, and $\pi^* = \pi_{\text{ref}} e^{r/\beta}/Z$. For any $\pi \ll \pi_{\text{ref}}$,

$$\beta \,\mathrm{KL}(\pi \,\|\, \pi^*) = \beta \,\mathrm{KL}(\pi \,\|\, \pi_{\text{ref}}) - \mathbb{E}_\pi[r] + \psi = \psi - J(\pi),$$

so $J(\pi) = \psi - \beta \,\mathrm{KL}(\pi \,\|\, \pi^*)$ is maximized uniquely at $\pi = \pi^*$, with value ψ. Evaluating J at π^* gives $\beta \,\mathrm{KL}(\pi^* \,\|\, \pi_{\text{ref}}) = \mathbb{E}_{\pi^*}[r] - \psi$. Jensen's inequality gives $\psi = \beta \log \mathbb{E}_{\pi_{\text{ref}}}[e^{r/\beta}] \geq \mathbb{E}_{\pi_{\text{ref}}}[r]$, hence $\beta \,\mathrm{KL}(\pi^* \,\|\, \pi_{\text{ref}}) \leq \mathbb{E}_{\pi^*}[r] - \mathbb{E}_{\pi_{\text{ref}}}[r] = \Delta r$, which is the first bound. Pinsker's inequality yields the second. □

Corollary 3. Total variation cannot increase under a measurable map, which gives the pushforward claim. For the kernel claim, if $k(x, x) \leq 1$ then every f in the unit ball of the reproducing kernel

Hilbert space satisfies $\|f\|_\infty \le \sup_x \sqrt{k(x,x)} \le 1$, so $|\mathbb{E}_{\pi^*} f - \mathbb{E}_{\pi_{\mathrm{ref}}} f| \le 2\|\pi^* - \pi_{\mathrm{ref}}\|_{\mathrm{TV}}$, giving $\mathrm{MMD} \le 2\|\cdot\|_{\mathrm{TV}}$ and $\mathrm{MMD}^2 \le 4\|\cdot\|_{\mathrm{TV}}^2 \le 2\Delta r/\beta$. □

Corollary 4. Apply Proposition 2 once with $\pi_{\mathrm{ref}} = \pi_{\theta_{\mathrm{DP}}}$ and once with $\pi_{\mathrm{ref}} = \pi_{\mathrm{pub}}$. Each bound is stated relative to its own reference, and the two instances of J differ only in that reference. □

Remark 5. The identity $\mathrm{KL}(\mathrm{BoN}_N \,\|\, \pi_{\mathrm{ref}}) = \ln N - (N-1)/N$ is standard for best-of-N sampling when the reward has a continuous distribution, so that the maximizer is almost surely unique. With a discrete reward, ties are broken uniformly and the induced distribution is closer to π_{ref}, so the identity becomes an upper bound. □

Proposition 6. The fitting joint is $\tilde{P}(x,y) = \pi(y)\,[pP(x \mid y) + (1-p)Q(x)]$, whose x-marginal is $\tilde{P}(x) = pP(x) + (1-p)Q(x)$ with $P(x) = \sum_y \pi(y)P(x \mid y)$, since $\sum_y \pi(y) = 1$. Using $\pi(y)P(x \mid y) = P(x)P(y \mid x)$,

$$\tilde{P}(y \mid x) = \frac{pP(x)P(y \mid x) + (1-p)Q(x)\pi(y)}{pP(x) + (1-p)Q(x)} = \lambda(x)P(y \mid x) + \big(1 - \lambda(x)\big)\pi(y),$$

which is Eq. 4. □

Corollary 7. If $\pi \equiv 1/m$ then $\tilde{P}(y \mid x) = \lambda(x)P(y \mid x) + (1 - \lambda(x))/m$ is a strictly increasing affine function of $P(y \mid x)$ at fixed x, so the two orderings over y coincide and in particular the maximizers agree. For general π, write y^* for the faithful maximizer. For any $y \neq y^*$, $\tilde{P}(y^* \mid x) > \tilde{P}(y \mid x)$ holds if and only if $\lambda(x)\,[P(y^* \mid x) - P(y \mid x)] > (1 - \lambda(x))\,[\pi(y) - \pi(y^*)]$. Bounding the right side by its largest value over y and the left side below by the margin gives Eq. 5 as a sufficient condition. □

Proposition 8 and Corollary 9. Write $\tilde{P}(x,y) = \pi(y)\big[pP(x \mid y) + (1-p)Q(x)\big]$. With weights $w_y \propto 1/\pi(y)$, the population minimizer of unregularized weighted log loss over a class containing the normalized weighted conditional is that conditional, $q^*_{\mathrm{bal}}(y \mid x) \propto w_y\,\tilde{P}(x,y)$. Substituting the joint and cancelling $\pi(y)$ leaves

$$q^*_{\mathrm{bal}}(y \mid x) = \frac{pP(x \mid y) + (1-p)Q(x)}{p\sum_z P(x \mid z) + m(1-p)Q(x)}.$$

Dividing numerator and denominator by m and using $\overline{P}(x) = \frac{1}{m}\sum_z P(x \mid z)$ gives Eq. 6. For the corollary, the map $u \mapsto \lambda_{\mathrm{bal}}(x)u + (1 - \lambda_{\mathrm{bal}}(x))/m$ is strictly increasing in u when $\lambda_{\mathrm{bal}}(x) > 0$, so it preserves the ordering over y and therefore the argmax set, ties included. □

C.1 Properties of the group-relative objective

Two properties of Eq. 2 matter when interpreting §5, and neither is a defect we correct. First, the group-standard-deviation term in $A(\tau) = (r(\tau) - \bar{r})/(s_r + \eta)$ is a rescaling rather than a control variate. It up-weights low-variance groups, so the optimized objective is not exactly expected reward. Second, when every completion in a group receives the same total reward, each advantage is zero and the group contributes no policy-gradient term, although the KL anchor still contributes a gradient toward the reference. Uniform behavior on a single key does not by itself produce such a tie, because the reward aggregates $|K|$ keys, and we do not have logged tie frequencies for the executed runs. Both properties mean the finite-step update we run is not the exact maximizer characterized in Proposition 2, which is why that result is stated as an interpretation aid rather than a prediction. The implemented loss in Eq. 2 pairs those advantages with a plain $-A(\tau)\log \pi_\theta(\tau)$ policy-gradient term, so it omits the clipped importance-ratio surrogate used in canonical GRPO and PPO.

Separately, the anchor $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$ penalizes departure from the initialization rather than from the real corpus, so it constrains movement without favoring fidelity. We report the measured relationship between β and fidelity in the KL-weight audit rather than deriving it.

D SUPPORTING COMPARISONS AND AUDITS

D.1 STATISTICAL CONVENTIONS

The conventions of §4 apply throughout this appendix unless a subsection states otherwise. The training seed is the independent unit of replication, so paired summaries use one aggregate value per seed and Student t intervals summarize variation across the observed seeds. Hierarchical intervals additionally resample matched rows within each seed, which adds conditional row-sampling uncertainty without increasing the number of independent training runs. Exact two-sided sign-flip tests enumerate all 2^n sign assignments of the n observed paired differences. Intervals are not adjusted for multiple comparisons, and an interval containing zero is treated as unresolved rather than as evidence of equivalence. Per-key decompositions, auxiliary metrics, transfer screens, and diagnostic audits are descriptive unless identified as a primary comparison.

D.2 REWARD-SIDE LEARNERS AND BUDGET CONTEXT

The main comparisons vary control after a shared adapted initialization. This section instead compares upstream private-learning configurations and external references. All placement comparisons hold fixed the frozen base, requested-control vocabulary, cluster-conditioned sampler, structural DP release, generation count, and evaluation protocol. They do *not* match training scale. The generator learner sees 8934 capped examples from 4518 senders over 2000 DP steps. The reward learner sees 79,077 records from 36,501 users over 200 steps on frozen-embedding linear heads. These differences in capacity and selected population prevent an isolated placement interpretation.

REWARD-DP-DISC+CLUSTER runs the same RFT loop against a DP-trained real/fake discriminator (frozen MiniLM embeddings, logistic head), using $P(\text{real})$ as reward. REWARD-DP-GROUND+CLUSTER instead uses a DP-trained attribute-recovery reward with the same fraction-of-requested-keys objective as OUTPUT-VERIFIER RFT. This removes reward-objective mismatch, but not the capacity or population differences. Both use the nominal targets $\varepsilon_{\text{learn}}{=}4$ and the same $\varepsilon_{\text{struct}}{=}1$ structural recipe. The sweep repeats REWARD-DP-GROUND+CLUSTER at $\varepsilon_{\text{learn}} \in \{2, 8\}$. The external AUG-PE-RAW and AUG-PE+LEVERS references report budgets $(4, 10^{-5})$ and $\leq (4.5, 2{\times}10^{-5})$, respectively, without establishing equal end-to-end privacy across these methods.

Real-verifier source diagnostic. REAL-VERIFIER RFT retains the adapted initialization but fits its reward on real bodies, so it is an unaccounted diagnostic rather than the output-only post-processing recipe. For OUTPUT-VERIFIER RFT minus REAL-VERIFIER RFT, the primary Qwen judge gives a recovery difference of $+0.023$ (95% CI $[+0.008, +0.037]$), whereas cross-family Luna gives -0.001 ($[-0.039, +0.032]$). Qwen favors the output-derived verifier, but Luna leaves the source comparison unresolved. Its interval is not evidence of equivalence. Table 2 retains the full and grouped recovery values and Qwen real-data calibration. Table 25 gives all eight primary-Qwen OUTPUT-VERIFIER RFT–L4 per-key comparisons. Invalid-response rates and handling are reported in Appendix G.4.

REWARD-DP-GROUND+CLUSTER starts from the public generator and uses a privately learned attribute-recovery reward (Figure 2). It tests a different use of source data, not another control procedure on the adapted initialization. At learner epsilon four it has slightly lower mean-TVD than OUTPUT-VERIFIER RFT, but worse cluster and embedding fidelity and coverage (Table 5). Both generator-side methods also outperform REWARD-DP-DISC+CLUSTER on every distributional axis. The reward-quality audit confirms an informative but imperfect comparator (Appendix G.3.1). Discriminator-overoptimization diagnostics appear in Appendix D.6.

Across learner epsilons $2, 4, 8$, OUTPUT-VERIFIER RFT has lower cluster-joint TVD point estimates, lower MMD^2, and higher coverage than REWARD-DP-GROUND+CLUSTER (Table 6). Paired MMD^2 and coverage intervals exclude zero at every budget. The recovery difference is unresolved. REWARD-DP-GROUND+CLUSTER retains a mean-TVD advantage of 0.003–0.009, and fidelity is not monotone in epsilon. Each budget trains its own arms, and the three-seed spread within a budget is comparable to the differences between budgets, so that ordering is not resolved here. Capacity, supervision, and selected populations also differ, so the ordering describes these configurations, not an isolated placement effect.

Table 5: No arm leads every axis, and L4 leads all four fidelity measures. Extended Enron comparison (5000 bodies/seed). Means ± sd use six matched seeds for L4, OUTPUT-VERIFIER RFT, and frozen Qwen, five for the unaccounted REAL-VERIFIER RFT diagnostic, and three for reward-DP. Aug-PE references are single runs. Recovery is measured by the automatic labeler, and independent judges are in Table 2. TSTR is matched-size macro-F1, averaging ten subsamples within each seed. PII spans are diagnostic. Frozen Qwen shares the structural sampler but has no adapter. Comparison roles and accounting scope are in Table 27 and §3.6.

Method	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑	TSTR F1 ↑	PII spans ↓	Recovery ↑
Frozen-model quality anchor: common DP structural sampler, no adapter							
Frozen Qwen (no adapter)	0.150±0.003	0.289±0.008	0.0688±0.0005	0.359±0.008	0.619±0.008	1394±69	0.429±0.001
Adapted generator and output-derived control							
L4 (RFT initialization)	**0.119**±0.009	**0.134**±0.011	**0.0140**±0.0027	**0.843**±0.007	0.623±0.008	9812±1478	0.436±0.005
OUTPUT-VERIFIER RFT	0.155±0.015	0.182±0.033	0.0312±0.0107	0.804±0.017	0.623±0.010	10204±3297	0.473±0.019
Verifier-source diagnostic: same L4 *and RFT recipe*							
REAL-VERIFIER RFT (unaccounted diagnostic)	0.132±0.026	0.140±0.020	0.0180±0.0048	0.786±0.040	0.623±0.014	4613±1508	0.482±0.015
Reward-side learners: capacity and selected populations differ							
Reward-DP Disc+Cluster	0.288±0.025	0.364±0.031	0.0383±0.0029	0.409±0.003	0.584±0.004	1616±154	0.375±0.000
Reward-DP Ground+Cluster	0.136±0.008	0.201±0.007	0.0797±0.0015	0.333±0.005	0.630±0.008	5364±1472	0.478±0.003
Private-Evolution references							
AUG-PE+LEVERS	0.272	0.268	0.0536	0.480	0.579	248	0.413
AUG-PE-RAW	0.293	0.343	0.0513	0.434	0.613	695	n/a

Table 6: Learner-budget sweep using the full-pool training recipe over matched seeds 0–2 (5000 samples/arm/seed). For joint-TVD and MMD^2, Δ = REWARD-DP-GROUND+CLUSTER − OUTPUT-VERIFIER RFT, and for coverage and automatic recovery, Δ = OUTPUT-VERIFIER RFT − REWARD-DP-GROUND+CLUSTER, so positive values favor OUTPUT-VERIFIER RFT. MMD^2 and coverage brackets are paired seed-level 95% t-CIs. The final column is OUTPUT-VERIFIER RFT's recovery lift over its corresponding DP-LoRA initialization, with a hierarchical seed/row bootstrap CI. The common structural recipe targets $\varepsilon_{\text{struct}}$=1. Ledger sums are nominal, not certified artifact or raw-corpus bounds.

Learner ε (ledger sum)	Δ joint TVD	Δ MMD^2 (95% CI)	Δ coverage (95% CI)	Δ recovery	RFT recovery lift (95% CI)
2 (3)	+0.028	+0.056 [+0.020, +0.092]	+0.409 [+0.259, +0.559]	+0.003	+0.046 [+0.043, +0.049]
4 (5)	+0.047	+0.057 [+0.043, +0.072]	+0.469 [+0.455, +0.482]	+0.007	+0.045 [+0.030, +0.058]
8 (9)	+0.051	+0.053 [+0.038, +0.067]	+0.464 [+0.338, +0.590]	+0.018	+0.052 [+0.044, +0.063]

RFT improves recovery over its corresponding DP-LoRA initialization in every seed at learner epsilons two and eight, with mean lifts +0.046 and +0.052. At epsilon four, OUTPUT-VERIFIER RFT averages 0.485 over the sweep's three matched seeds, versus 0.473 over six seeds. Its paired recovery difference from REWARD-DP-GROUND+CLUSTER is unresolved at +0.007. Neither this comparison nor the unequal-seed means in Table 5 establishes a recovery ordering.

D.3 TASK-DEPENDENCE OF MATCHED-SIZE UTILITY

Table 7 provides no evidence that RFT changes the primary needs-decision task or the requires-response task. Triage decreases by 0.019 with a t-interval excluding zero, but only one of six seed deltas is positive and the exact two-sided sign-flip test is p=0.0625. We therefore make no blanket utility-preservation claim: the primary-task change remains unresolved, while the broader triage taxonomy shows a small cost. No equivalence or non-inferiority margin was prespecified, so intervals spanning zero do not establish preservation. These tasks use post-hoc inferred synthetic labels. They measure transfer from re-annotated synthetic text, not the utility of the original requested-

Table 7: Six-seed downstream utility. TSTR trains on exactly 2000 post-hoc-labeled synthetic bodies per replicate, matching the 2000 real examples used for TRTR. Both test on the same 1000 sender-disjoint real examples. Ten deterministic synthetic subsamples are averaged within each training seed. Δ is paired OUTPUT-VERIFIER RFT–L4 direct TSTR macro-F1.

Task	TRTR	L4	OUTPUT-VERIFIER RFT	Paired Δ [95% CI]
Needs decision	0.642	$0.623_{\pm 0.008}$	$0.623_{\pm 0.010}$	-0.0003 [-0.0102, $+0.0096$]
Requires response	0.694	$0.652_{\pm 0.008}$	$0.647_{\pm 0.010}$	-0.0055 [-0.0145, $+0.0036$]
Triage	0.604	$0.574_{\pm 0.012}$	$0.555_{\pm 0.013}$	-0.0187 [-0.0323, -0.0051]

label/body pairs. The paired audit below directly tests those original pairs without changing the generated bodies or evaluation split.

Paired requested-label audit. At learner epsilon four, we compare OUTPUT-VERIFIER RFT with its DP-LoRA initialization over six Enron/Qwen seeds and three seeds each for Amazon/Qwen and Enron/SmolLM3. For each seed and target, the audit verifies exact original/post-hoc body equality and requested-skeleton alignment across the two arms. Body equality is checked before the 600-character classifier-input truncation. This limit matches the verifier's scoring path and is applied unchanged. The audit then intersects rows eligible under both label sources and both methods. All 5000 primary or 1000 transfer rows remain eligible. The two label sources use identical body indices and ordering, classifier hyperparameters, and real test examples. Requested-label class counts therefore also match across arms. Only the synthetic training labels change within each arm.

The audit reuses the existing real-data split and deterministic subsampling helpers: 2000 synthetic training bodies, 1000 real test bodies, and ten subsamples per primary seed. Transfer uses all 1000 synthetic bodies and 1000 real test bodies. Reusing the primary subsampling protocol on the full transfer corpus makes all ten repeats identical. They reuse one training set rather than provide independent samples. This is the existing utility test split, not a new holdout from upstream generator training. All 60 post-hoc method/seed/task scores reproduce the previous matched-utility audit to six decimals, with identical subsample digests and unchanged TRTR scores. No text generation or language-model training is performed for this audit.

Reference-label provenance. Enron utility uses `final_ground_truth_v2.jsonl`. Amazon uses the canonical `amazon_reviews_gt.jsonl`. The latter retains native star ratings and rule-derived review-intent labels, rather than the LLM-enriched annotation version used for upstream semantic controls. The intent-utility scores are therefore evaluated against that canonical labeling rule, not against independent semantic annotations. The exact reference paths, input hashes, real-split hashes, sample digests, software versions, and per-seed results are recorded by `audit_requested_label_utility.py` in the profile-specific `requested_label_utility.json` reports. The consolidated `requested_label_utility_summary.json` and CSV retain all seven setting–task comparisons.

Interpretation and uncertainty. Tables 8 and 9 report every setting–task comparison. Requested-label Enron needs-decision and requires-response differences remain unresolved. Triage decreases in all six seeds, with an unadjusted t-interval excluding zero and exact $p = 0.03125$. Amazon review intent improves in all three seeds and has a positive t-interval, but its exact sign-flip test is $p = 0.25$. Star-rating and SmolLM3 results do not establish a utility benefit. This is limited, task-specific evidence, not a general requested-label advantage. Inference uses one mean per training seed after averaging subsamples. The 10 repeats do not increase the effective seed count. All seven interaction intervals include zero, so the audit does not establish that post-hoc relabeling hides a general RFT utility gain. It also does not establish equivalence or rule out benefits on other tasks or with other downstream learners.

Table 8: Post-hoc labels train better classifiers than requested tags on six of the seven setting–task pairs, the exception being Amazon review intent. Mean TSTR macro-F1 under the two synthetic-label protocols, using the same bodies, subsamples, and real test sets. RFT denotes OUTPUT-VERIFIER RFT and L4 its corresponding DP-LoRA initialization. The seed count is the number of upstream training seeds, not the number of repeated downstream fits. Contrasts in Table 9 use stored seed means before display rounding.

		Requested labels		Post-hoc labels	
Setting / task	Seeds	L4	RFT	L4	RFT
Enron/Qwen: needs decision	6	0.5519	0.5630	0.6230	0.6228
Enron/Qwen: requires response	6	0.5337	0.5333	0.6521	0.6466
Enron/Qwen: triage	6	0.3678	0.3463	0.5740	0.5552
Amazon/Qwen: star rating	3	0.2946	0.2610	0.3550	0.3649
Amazon/Qwen: review intent	3	0.1691	0.2607	0.1590	0.1473
Enron/SmolLM3: needs decision	3	0.5437	0.5349	0.6023	0.5776
Enron/SmolLM3: triage	3	0.3472	0.3474	0.4760	0.4955

Table 9: Paired utility contrasts over training seeds. Upper panel: RFT minus L4 under each label source. Lower panel: interaction $\Delta_{\text{requested}} - \Delta_{\text{posthoc}}$. Brackets are 95% Student t-intervals, and p_{sf} is the exact two-sided seed-level sign-flip test. A positive interaction means a larger estimated RFT advantage under requested labels, not that either method's absolute score necessarily improves.

Setting / task	Requested-label Δ [95% CI]	p_{sf}	Post-hoc-label Δ [95% CI]	p_{sf}
Enron/Qwen: needs decision	$+0.0111$ $[-0.0062, +0.0285]$	0.1875	-0.0003 $[-0.0102, +0.0096]$	0.96875
Enron/Qwen: requires response	-0.0004 $[-0.0233, +0.0226]$	0.96875	-0.0055 $[-0.0145, +0.0036]$	0.21875
Enron/Qwen: triage	-0.0215 $[-0.0379, -0.0052]$	0.03125	-0.0187 $[-0.0323, -0.0051]$	0.0625
Amazon/Qwen: star rating	-0.0336 $[-0.1114, +0.0443]$	0.25	$+0.0099$ $[-0.0252, +0.0450]$	0.5
Amazon/Qwen: review intent	$+0.0916$ $[+0.0084, +0.1747]$	0.25	-0.0117 $[-0.0741, +0.0507]$	0.75
Enron/SmolLM3: needs decision	-0.0087 $[-0.1599, +0.1424]$	1	-0.0248 $[-0.0614, +0.0119]$	0.25
Enron/SmolLM3: triage	$+0.0002$ $[-0.0610, +0.0615]$	1	$+0.0195$ $[-0.0907, +0.1297]$	0.75

Setting / task	Interaction [95% CI]	p_{sf}
Enron/Qwen: needs decision	$+0.0114$ $[-0.0125, +0.0354]$	0.3125
Enron/Qwen: requires response	$+0.0051$ $[-0.0233, +0.0335]$	0.65625
Enron/Qwen: triage	-0.0028 $[-0.0218, +0.0162]$	0.75
Amazon/Qwen: star rating	-0.0435 $[-0.0877, +0.0008]$	0.25
Amazon/Qwen: review intent	$+0.1033$ $[-0.0368, +0.2433]$	0.25
Enron/SmolLM3: needs decision	$+0.0160$ $[-0.1692, +0.2012]$	0.75
Enron/SmolLM3: triage	-0.0193 $[-0.0686, +0.0301]$	0.25

D.4 NO DUPLICATE-BODY COLLAPSE, AND AN UNRESOLVED PII CHANGE

OUTPUT-VERIFIER RFT does not induce duplicate-body collapse (Table 10). Detectable PII spans average $9812 \rightarrow 10204$, but the paired interval $[-2726, +3510]$ includes zero and seed-level changes have mixed signs. We therefore make no PII-reduction claim.

D.5 POSITIVE-CONTROL-CALIBRATED EMPIRICAL PRIVACY AUDIT

The full non-private control is readily attacked, and even the shorter rank-4 control remains separable, so the DP results are not explained by a powerless instrument (Table 11). Across the two clean

Table 10: Lexical diversity for representative seed 0 (5000 samples). All OUTPUT-VERIFIER RFT bodies remain unique. Distinct-1/2 decrease modestly from L4 but remain far above both Aug-PE references, which contain $\sim 40\%$ unique bodies.

Method	distinct-1	distinct-2	unique-frac	$n_{\text{uniq}}/5000$
L4	0.1029	0.4082	1.000	5000
OUTPUT-VERIFIER RFT	0.0858	0.3876	1.000	5000
REWARD-DP-DISC+CLUSTER	0.1135	0.4436	1.000	5000
AUG-PE+LEVERS	0.0203	0.1186	0.418	2090
AUG-PE-RAW	0.0192	0.1089	0.406	2030

Table 11: User-level membership inference on 256 matched members and 256 nonmembers, using up to ten records per user. AUC is for the primary frozen-reference-normalized mean-loss score. Intervals are 95% stratified user-bootstrap CIs. The two non-private rows are positive controls.

Target checkpoint	AUC [95% CI]	Advantage	TPR@1% FPR	TPR@5% FPR
No-DP LoRA, rank 8	0.962 [0.944, 0.978]	0.836	0.207	0.770
No-DP LoRA, rank 4 (500 steps)	0.715 [0.669, 0.759]	0.359	0.047	0.109
L4, clean seed 1	0.550 [0.502, 0.601]	0.113	0.031	0.043
OUTPUT-VERIFIER RFT, clean seed 1	0.553 [0.504, 0.604]	0.121	0.016	0.051
L4, clean seed 2	0.542 [0.493, 0.594]	0.117	0.023	0.047
OUTPUT-VERIFIER RFT, clean seed 2	0.560 [0.511, 0.609]	0.125	0.012	0.039

DP seeds, unnormalized mean-loss AUC is 0.510–0.512 for L4 and 0.514 for OUTPUT-VERIFIER RFT. Reference-normalized AUC is 0.542–0.550 and 0.553–0.560, respectively, with two-seed means 0.546 and 0.557. This indicates weak aggregate separability rather than zero leakage. At operationally low FPR, however, TPR remains low: 1.2–3.1% at 1% FPR and 3.9–5.1% at 5% FPR. The per-seed RFT–L4 AUC changes are $+0.003$ and $+0.018$. These small point differences do not establish a material empirical-privacy change from post-processing. These attacks probe specific leakage scenarios. They neither certify DP nor close the upstream raw-corpus selection gap.

Table 12: Secret-Sharer canary audit. Each rank is among the true canary and 256 decoys. Random ranking has expected exposure 1.422 bits, and maximum exposure is $\log_2(257) = 8.006$ bits.

Checkpoint	Repeated users	Rank / 257	Exposure (bits)
No-DP rank-4 control	8	1	8.006
L4, seed 0	8	95	1.436
OUTPUT-VERIFIER RFT, seed 0	8	101	1.347
L4, clean seed 1	0	174	0.563

The non-private control ranks the secret first, validating the exposure instrument. Under the DP-training recipe, both the initial L4 model and its OUTPUT-VERIFIER RFT descendant place the canary near random expectation rather than near extraction (Table 12). Seed 0 deliberately includes one copy of the same canary for each of eight distinct synthetic users. The non-private positive control uses the same insertion setting. No formal rationale or optimality claim for choosing eight users is documented. OUTPUT-VERIFIER RFT inherits that learned state although its saved checkpoint omits the canary metadata. This is a repeated-insertion stress test, not the formal single-user neighboring-dataset case. Eight does not define the privacy unit. Excluding the instrumented seed 0 leaves the five-seed automatic-recovery gain at $+0.0379$ (95% t-CI $[+0.0173, +0.0584]$), so the primary effect is not driven by canary instrumentation.

Matched frozen-base overlap control. We also generate 5000 bodies from the frozen Qwen3.5-4B base with the same seed-3 skeletons and 120-token cap but no LoRA or virtual prompt. Historical decoding-record limits are described in Appendix G.3. Neither the base, L4, nor OUTPUT-

VERIFIER RFT produces an exact normalized match to a selected private-training body. Among outputs long enough for a 12-word audit, selected-training overlap occurs for $7/4985$ base outputs (0.140%), $11/4915$ L4 outputs (0.224%), and $8/4890$ OUTPUT-VERIFIER RFT outputs (0.164%). The corresponding overlap with nonmember held-out text is 0.040%, 0.285%, and 0.245%. Thus the two adapted models have no excess training-specific 12-word overlap relative to held-out text, and we observe no increase in that excess over the frozen-base control. This lexical check addresses verbatim adaptation leakage, not paraphrased memorization or non-verbatim knowledge the public base may have acquired from Enron during pretraining.

D.6 EFFECT OF LEVER CONDITIONING IN AUG-PE

In the single-run Aug-PE references, conditioning on levers helps modestly (joint-TVD $0.343 \rightarrow 0.268$, coverage $0.434 \rightarrow 0.480$, detected PII spans $695 \rightarrow 248$) and makes automatic measurement-labeler recovery measurable at 0.413. But AUG-PE+LEVERS remains far below the DP-generator arms on coverage (0.480 vs. 0.80–0.84) and joint fidelity. OUTPUT-VERIFIER RFT outperforms the AUG-PE+LEVERS adaptation in these point estimates on every fidelity axis and on matched needs-decision TSTR F1 (0.623 vs. 0.579). Its component ledger sums to 5, versus the Aug-PE reference's reported budget of 4.5, but the historical selection and implementation qualifications prevent interpreting those numbers as comparable end-to-end raw-corpus guarantees. REWARD-DP-DISC+CLUSTER underperforms AUG-PE+LEVERS on marginal and joint TVD, coverage, and recovery.

For REWARD-DP-DISC+CLUSTER, the optimization reward rises from 0.049 to 0.903 while generation quality degrades, consistent with reward overoptimization (Gao et al., 2023). These dynamics do not isolate privacy noise from reward capacity, supervision, or optimization effects. The generator-side arms have mean-TVD 0.12–0.15 and coverage 0.80–0.84, compared with 0.29 and 0.41 for this discriminator configuration (Table 5).

D.7 KL-WEIGHT PARETO AUDIT

Table 13 varies the KL weight over three matched full-pool seeds. The primary β=0.05 point maximizes mean recovery. Increasing to β=0.10 retains $+0.037$ of the $+0.045$ recovery gain while improving every reported fidelity mean and leaving all three matched downstream tasks within sampling variation. Its recovery mean is lower (0.478 vs. 0.485), so the two points trade control against fidelity rather than exhibit Pareto dominance. At β=0.20, MMD^2 and coverage move closer to L4, but the recovery gain falls by more than half. The paired recovery gains and 95% t-CIs are $+0.045$ $[+0.008, +0.082]$, $+0.037$ $[+0.032, +0.042]$, and $+0.019$ $[+0.013, +0.025]$ for $\beta \in \{0.05, 0.10, 0.20\}$, respectively. We therefore identify β=0.10 as a candidate compromise within the tested settings, not a validated deployment optimum. We retain the pre-sweep β=0.05 point for the headline analysis because it has six-seed automatic evaluation and the predeclared three-seed independent-judge evaluation. Replacing it after observing the sweep would post-select the main result.

Table 13: Post-processing KL-weight audit, mean $\pm$ sd over full-pool seeds 0–2. Recovery is the automatic eight-key measure, and Δ is the paired gain over L4. Utility is direct sender-disjoint needs-decision TSTR macro-F1 with 2000 synthetic training bodies per replicate. All RFT points use the same 12,800 one-time rollouts and one serving-time completion per body.

Method	Recovery ↑	Δ	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑	Utility ↑
L4	0.440±0.001	–	**0.114**±0.0004	**0.127**±0.009	**0.0139**±0.0032	**0.845**±0.008	0.627±0.009
β=0.05	**0.485**±0.015	+0.045	0.142±0.009	0.154±0.011	0.0223±0.0056	0.801±0.005	0.626±0.007
β=0.10	0.478±0.003	+0.037	0.135±0.028	0.149±0.009	0.0180±0.0046	0.830±0.014	**0.629**±0.005
β=0.20	0.460±0.003	+0.019	0.135±0.014	0.153±0.015	0.0165±0.0032	0.841±0.001	0.628±0.003

D.8 MATCHED CONTROL COMPARISON AND HISTORICAL AUDIT

D.8.1 THREE-SEED MATCHED INFERENCE AUDIT

Frozen inference protocol. This audit evaluates existing epsilon-four L4 and RFT adapters for seeds 0–2. It performs no new DP or RFT training. Each seed uses the first 1000 requested skeletons from its archived L4 corpus, selected without reference to outcomes. The public Qwen3.5-4B base is pinned to revision `851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a` for all six generation stages. This establishes a common base for the new audit, not its identity with every historical training-base revision. Temperature is 0.9, top-p is 0.95, the cap is 120 new tokens, and the completion batch size is 16. Sampling uses one beam and one return sequence per expanded prompt. Model, tokenizer, and recorded generation configuration identities match across all runs, using PyTorch `2.5.1+cu121`, Transformers `5.5.3`, BF16, and NVIDIA A100 80GB PCIe GPUs. These settings are recorded directly rather than inferred from historical defaults.

L4 generates eight candidates per skeleton, while RFT generates one. Every L4 candidate and its exact sampled token IDs are retained. The first candidate supplies the L4 baseline, while best-of-N uses nested prefixes $N \in \{1, 2, 4, 8\}$ with reproducible uniform tie-breaking. The 27,000 completions share requested skeletons across arms. No empty outputs occurred and no selective regeneration was used. The standard automatic measurement labeler and the existing fixed real embedding reference evaluate all 1000 outputs per arm and seed. Coverage is sample-size dependent, so these values are compared within this audit, not directly with the 5000-body primary study.

Verifier reconstruction boundary. Reranking uses locally fitted, frozen reconstructions of the output-derived verifiers, prepared before the GPU audit from the same archived L4 fitting corpora and recipe. The original reward-model weights used to train the RFT policies were not archived. The reconstructed train/holdout sizes are $3995/998$, $4000/999$, and $4000/999$, with maximum per-key holdout macro-F1 differences from the historical reports of 0.0035, 0.0031, and 0.0071, respectively. These checks do not prove identical reward functions. The comparison therefore concerns the evaluated RFT policies and reconstructed-verifier selection, not a byte-identical-scorer optimization ablation.

Table 14: Recovery rises monotonically with candidate count, and OUTPUT-VERIFIER RFT attains the highest recovery at the worst fidelity. Automatic control and fidelity in the new matched audit: mean ± sd over three seeds, 1000 bodies/arm/seed. All arms use the same requests and fixed decoding protocol. N=1 is the first candidate from the L4 pool.

Arm	Auto recovery ↑	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑
L4 / N=1	0.4382±0.0043	0.1218±0.0037	0.1335±0.0083	0.0130±0.0028	0.6183±0.0140
Rerank-2	0.4491±0.0018	0.1189±0.0029	0.1430±0.0247	0.0130±0.0028	0.6218±0.0158
Rerank-4	0.4624±0.0032	0.1228±0.0056	0.1433±0.0164	0.0131±0.0036	0.6272±0.0069
Rerank-8	0.4697±0.0047	0.1215±0.0023	0.1374±0.0068	0.0131±0.0032	0.6227±0.0045
OUTPUT-VERIFIER RFT	0.4824±0.0075	0.1442±0.0067	0.1544±0.0034	0.0221±0.0052	0.6010±0.0207

Automatic recovery increases in every seed for both RFT and rerank-8 relative to L4 (Table 14). Their respective paired mean gains are $+0.0442$ (95% t-CI $[+0.0285, +0.0599]$) and $+0.0315$ ($[+0.0138, +0.0492]$). The automatic RFT–rerank-8 difference remains unresolved at $+0.0127$ ($[-0.0009, +0.0264]$). Rerank-8 has better point estimates than RFT on all four fidelity axes in every seed. For RFT minus rerank-8, the mean-TVD difference is $+0.0227$ ($[+0.0058, +0.0395]$), MMD^2 is $+0.0089$ ($[+0.0004, +0.0174]$), joint-TVD is $+0.0170$ ($[-0.0081, +0.0420]$), and coverage is -0.0217 ($[-0.0653, +0.0220]$). Only the first two intervals exclude zero.

Independent judging. Qwen3.8-27B (Qwen Team, 2026c) and Azure GPT-5.6 Luna each score the same 200 matched bodies per arm and seed for L4, RFT, and rerank-8: 1800 tasks per judge, with no raw real-body judging. The sample indices were frozen before inspecting automatic evaluation outcomes. Input filtering leaves 991, 995, and 993 eligible rows from the 1000 aligned rows in seeds 0–2. Cluster-balanced samples are weighted to those eligible mixtures. Both judges receive the same body-only prompt and allowed vocabulary, without method identities or requested labels. The full-macro metric weights all eight keys equally, and judges are never pooled.

Qwen uses temperature zero, a 220-token response cap, and disabled thinking. Luna uses the enterprise Responses API with JSON-schema output, an 8192-token output cap, and its default temperature. Model IDs and responses were verified on all endpoints. The resulting comparisons in Table 15 use 10,000 hierarchical bootstrap replicates over the three observed seeds and matched rows within clusters. At the seed level, three positive differences alone have a two-sided exact sign-flip $p = 0.25$.

Table 15: Independent-judge full-macro recovery differences in the matched audit. Each contrast is the first arm minus the second. Brackets are hierarchical paired 95% intervals. RFT denotes OUTPUT-VERIFIER RFT. Both judges favor RFT over L4 and rerank-8, while rerank-8 versus L4 is unresolved.

Judge	Contrast	Δ	95% CI
Qwen3.8	RFT – L4	$+0.0733$	$[+0.0505, +0.0974]$
Qwen3.8	Rerank-8 – L4	$+0.0164$	$[-0.0202, +0.0546]$
Qwen3.8	RFT – Rerank-8	$+0.0568$	$[+0.0153, +0.0958]$
Luna	RFT – L4	$+0.1067$	$[+0.0823, +0.1320]$
Luna	Rerank-8 – L4	$+0.0363$	$[-0.0011, +0.0733]$
Luna	RFT – Rerank-8	$+0.0703$	$[+0.0414, +0.0991]$

Invalid-response sensitivity. Under the declared protocol an invalid response receives zero recovery, rather than being removed. Qwen has nine invalid responses: two for L4, seven for rerank-8, and none for RFT. They contain out-of-vocabulary intent values. Luna has one content-filtered rerank-8 response, retained as invalid without a filter bypass. The weighted invalid fractions for rerank-8 are 0.01575 under Qwen and 0.00097 under Luna. To assess whether these penalties explain RFT's advantage, a separate counterfactual sensitivity check assigns perfect recovery to every invalid reference response while retaining the original RFT scores. Raw judgments and primary summaries are unchanged. RFT's advantage remains positive under both judges (Table 16). This is a robustness bound, not new judging.

Table 16: Counterfactual invalid-response sensitivity: RFT minus rerank-8 when every invalid rerank-8 response is assigned perfect recovery. The same eligible-mixture weights and hierarchical bootstrap are used. These values do not replace actual judgments.

Judge	Counterfactual Δ	95% CI
Qwen3.8	$+0.0411$	$[+0.0043, +0.0739]$
Luna	$+0.0693$	$[+0.0414, +0.0968]$

Recorded inference work. Table 17 records exact token totals and synchronized generation times for the new audit. L4's timings cover its entire eight-candidate pool, not independent N=1/2/4 jobs. RFT's timings cover one completion per request from the existing policy, excluding its historical training. Only successful committed batches are counted. Failed attempts, scoring, and file writes are excluded, and setup is reported separately. These measurements do not establish an intrinsic algorithmic speedup or an amortization break-even point.

Table 17: Recorded work per seed in the new matched audit. Tokens include EOS. Generation and setup times are seconds. Setup includes recorded model loading and initialization for successful batch segments, including resumed smoke batches.

Seed	Policy	Completions	Tokens	Gen. seconds	Setup seconds
0	L4 pool	8000	760737	2228.72	20.08
0	RFT	1000	91158	280.04	11.07
1	L4 pool	8000	770432	2235.11	10.59
1	RFT	1000	87556	282.14	5.48
2	L4 pool	8000	778794	2224.03	10.51
2	RFT	1000	100169	280.14	5.53

The frozen inputs, source/checkpoint hashes, base lock, batch receipts, and all candidate token IDs are retained under `matched_control_v2`. The local `matched_control_summary.json` records automatic metrics. Separate Luna and Qwen3.8 summaries and invalid-response sensitivity reports record judging. The preparation, server generation, and local collection/evaluation entry points are `matched_control_audit.py` and `analyze_matched_control.py`. The study makes the new runs comparable and extends the reranking evidence. It does not recover all historical decoding settings, establish a new privacy guarantee, or establish the causal mechanism behind RFT's control gains.

D.8.2 Verifier reliability and failure conditions

Question and frozen scope. This CPU-only diagnostic asks whether the output-derived verifier corrects errors in requested labels, and where its reward is misleading. It reuses the three frozen verifier reconstructions and the 1800 selected synthetic bodies from Appendix D.8.1, with the existing Luna and Qwen3.8 annotations: 200 bodies per arm and seed for L4, rerank-8, and RFT. There is no verifier refit, new language-model training, generation, judging, or raw real-body access. The protocol, thresholds, and source hashes were fixed before computing these diagnostic predictions, after the aggregate control results were known. This is not a preregistration of the original experiment. Unselected L4 is the primary diagnostic distribution. Reranked and RFT outputs are secondary distributions, not additional independent replications.

Error correction and population weights. For a key, a *correction* occurs when the verifier agrees with the judge's body label but the requested tag does not. A *corruption* is the reverse. Net correction is verifier–judge agreement minus request–judge agreement, equivalently the corrected fraction minus the corrupted fraction. All eight keys receive equal weight in the full macro. These are agreement measures relative to imperfect judges, not observed ground-truth label errors. Each seed has equal population mass, followed by its frozen eligible-corpus cluster weights and equal sampled-row weights within each cluster. Conditional rates divide pooled weighted numerators by their corresponding weighted denominators, rather than averaging separately normalized per-seed rates.

An invalid judge response supplies no usable attribute annotation. Conditional agreement rates exclude it and disclose its missing mass. The original invalid-as-zero recovery summaries are unchanged. Qwen has two invalid L4 and seven invalid rerank-8 responses. Luna has one invalid rerank-8 response. RFT has none under either judge. To screen fitting-input reuse, the protocol excludes normalized first-600-character matches against each verifier's entire archived synthetic source corpus, including its holdout. No such overlaps occurred. No selected body exceeded either the verifier's 600-character limit or the judge harness's input limit. This checks exact normalized prefix reuse, not semantic similarity or pretraining contamination.

Intervals use 10,000 hierarchical bootstrap replicates (seed 20260916), resampling the three observed seeds and matched rows within each frozen cluster. All keys, arms, and judges share the same row draws. Keys are not independent replicates and judges are never pooled. Intervals are unadjusted. Within-seed net differences are additionally summarized by an exact sign-flip test, separately from the conditional pooled estimator.

Table 18: Net correction is never resolved positive on any arm, and is negative for both judges on RFT outputs. Full-macro requested-label error correction. Req/J and V/J denote request–judge and verifier–judge agreement. C and E are corrected and corrupted fractions. $\Delta_{\mathrm{corr}} = \mathrm{V/J} - \mathrm{Req/J} = C - E$ before rounding. Brackets are hierarchical 95% intervals. L4 is primary, and other arms are secondary. Rates condition on valid annotations, unlike the invalid-as-zero scores in Table 3. RFT denotes Output-Verifier RFT.

Judge	Arm	Req/J	V/J	C	E	Δ_{corr}	95% CI
Luna	L4	0.4245	0.4135	0.1569	0.1679	-0.0110	$[-0.0356, +0.0129]$
Luna	Rerank-8	0.4613	0.4656	0.0958	0.0915	$+0.0043$	$[-0.0116, +0.0200]$
Luna	RFT	0.5312	0.4745	0.1409	0.1976	-0.0567	$[-0.0822, -0.0301]$
Qwen3.8	L4	0.4870	0.4480	0.1508	0.1898	-0.0390	$[-0.0590, -0.0187]$
Qwen3.8	Rerank-8	0.5105	0.5028	0.0935	0.1012	-0.0077	$[-0.0245, +0.0081]$
Qwen3.8	RFT	0.5593	0.5077	0.1440	0.1957	-0.0517	$[-0.0742, -0.0296]$

No demonstrated aggregate denoising. On unselected L4, net correction is unresolved under Luna and negative under Qwen (Table 18). Both judges have negative within-seed net differences in all three seeds. The two-sided exact sign-flip $p = 0.25$ nevertheless reflects the small number of independent runs. For a missing judge label, a key's net correction is necessarily zero when the verifier and request agree. Otherwise, its possible range is $[-1, +1]$. Applying these bounds to Qwen's missing L4 annotations gives a full-mixture point-estimate range of $[-0.0401, -0.0378]$. This range is a missing-label sensitivity bound, not a confidence interval or new judging. It does not reverse the negative point estimate. Luna has no missing L4 annotations.

These results do not establish the proposed net label-correction explanation. The regularized, class-balanced verifier is not optimized for unweighted label accuracy, and this diagnostic does not estimate the latent noise model of Proposition 6 or the balanced-posterior argmax of Corollary 9. The secondary RFT rows are also negative under both judges. They describe agreement of the frozen reconstructions on RFT outputs, not the unarchived historical rewards or a causal explanation of training.

Informative but attribute-dependent reward. Lack of net annotation improvement does not imply a constant or uninformative control signal. On L4, judge recovery for verifier-rewarded versus nonrewarded key events is 0.5914 versus 0.2966 under Luna and 0.6845 versus 0.3355 under Qwen. These pooled rates weight key events equally. They are not arithmetic means of the per-key conditional rates. Table 19 reports every primary-study key. For example, Qwen topic recovery is 0.6125 versus 0.2027, whereas urgency is nearly unchanged at 0.3102 versus 0.3077. The rewarded-key sample counts range from 71 to 500 under Luna and 71 to 499 under Qwen. These profiles are descriptive, not multiplicity-controlled discoveries. The associations do not isolate causal reward effects.

Table 19: All eight L4 keys. Δ is net correction, with hierarchical 95% intervals. A_+ and A_- are judged requested-attribute recovery conditional on the verifier rewarding or not rewarding that key, respectively. Rates condition on valid judge annotations. The conditional recovery columns are descriptive point estimates, not calibrated confidence scores.

	Luna				Qwen3.8			
Key	Δ	95% CI	A_+	A_-	Δ	95% CI	A_+	A_-
triage	-0.0011	$[-0.0965, +0.0928]$	0.5392	0.4292	-0.0182	$[-0.0917, +0.0560]$	0.5876	0.4388
topic	$+0.0050$	$[-0.0487, +0.0572]$	0.6016	0.2189	$+0.0201$	$[-0.0343, +0.0729]$	0.6125	0.2027
intent_primary	$+0.0012$	$[-0.0660, +0.0680]$	0.4014	0.2009	-0.0643	$[-0.1362, +0.0081]$	0.5210	0.2515
response_primary	-0.0721	$[-0.1472, +0.0102]$	0.3894	0.2317	-0.1088	$[-0.2160, +0.0203]$	0.3783	0.2781
urgency_level	$+0.0383$	$[-0.0439, +0.1258]$	0.3593	0.2882	$+0.0190$	$[-0.0823, +0.1332]$	0.3102	0.3077
sentiment	-0.0240	$[-0.1249, +0.0714]$	0.8065	0.4718	-0.0423	$[-0.1344, +0.0502]$	0.8642	0.5031
difficulty	-0.0243	$[-0.0733, +0.0204]$	0.6057	0.4156	-0.0339	$[-0.1058, +0.0279]$	0.8496	0.5474
formality	-0.0108	$[-0.0861, +0.0650]$	0.5664	0.3932	-0.0838	$[-0.1527, -0.0144]$	0.6497	0.5255

Fixed reward bins and conditional failures. Before diagnostic scoring, the protocol fixed four reward bins with boundaries 0, 0.25, 0.5, 0.75, and 1. It defined a high-reward, low-recovery body as one with reconstructed reward at least 0.75 and judged recovery at most 0.5: at least six of eight verifier matches but at most four judge matches. The verifier reward is a fraction of requested-key matches, not a calibrated probability that a body is correct. Tables 20 and 21 retain all bins and the conditional failure results.

The high-reward subsets contain only 22 L4, 154 rerank-8, and 54 RFT bodies and differ across arms. Their wide intervals and changed conditioning preclude a causal claim that RFT avoids reward exploitation. Nor do these threshold-defined, judge-relative events measure factual correctness or overall text quality. They show that high proxy reward is not a certificate of eight-key adherence. The appropriate conclusion is informative, attribute-dependent supervision with observable failure conditions, not a verified denoising mechanism. The independently judged control gains in Section 5.2 are unchanged.

The versioned protocol is `verifier_reliability_protocol.json`. `audit_verifier_reliability.py` binds source/code hashes, scores the frozen verifiers on CPU, and aggregates the diagnostic. The isolated `verifier_reliability_v1` directory retains the protocol, input manifest, text-free row audit, summary, and complete per-

Table 20: Fixed reconstructed-reward bins. n counts sampled bodies before judge validity filtering, while n_L and n_Q count usable Luna and Qwen annotations. Mass is the bin's equal-seed, eligible-mixture-weighted fraction before judge filtering. Recovery conditions on valid annotations within the bin. A dash denotes an empty conditional cell, not zero recovery. RFT denotes OUTPUT-VERIFIER RFT.

Arm	Reward bin	n	n_L	n_Q	Mass	Luna recovery	Qwen recovery
L4	$[0, 0.25)$	38	38	38	0.0576	0.3897	0.3907
L4	$[0.25, 0.5)$	277	277	275	0.4357	0.3909	0.4499
L4	$[0.5, 0.75)$	263	263	263	0.4610	0.4588	0.5290
L4	$[0.75, 1]$	22	22	22	0.0457	0.4435	0.5370
Rerank-8	$[0, 0.25)$	0	0	0	0.0000	–	–
Rerank-8	$[0.25, 0.5)$	12	12	11	0.0142	0.2981	0.3383
Rerank-8	$[0.5, 0.75)$	434	433	428	0.6664	0.4479	0.4940
Rerank-8	$[0.75, 1]$	154	154	154	0.3195	0.4965	0.5511
RFT	$[0, 0.25)$	14	14	14	0.0164	0.4448	0.5104
RFT	$[0.25, 0.5)$	224	224	224	0.3845	0.4973	0.5119
RFT	$[0.5, 0.75)$	308	308	308	0.5111	0.5428	0.5761
RFT	$[0.75, 1]$	54	54	54	0.0880	0.6279	0.6784

Table 21: Conditional high-reward failures under the fixed thresholds. n_H is the number of high-reward bodies and n_F the low-recovery count within that subset. All high-reward annotations are valid. Weighted rates need not equal raw count ratios. Brackets are hierarchical 95% intervals. These are not failure rates for all generated text or causal cross-arm comparisons.

Judge	Arm	n_H	n_F	Weighted rate	95% CI
Luna	L4	22	11	0.6126	$[0.2989, 0.8603]$
Luna	Rerank-8	154	93	0.6338	$[0.4780, 0.7741]$
Luna	RFT	54	18	0.3369	$[0.1439, 0.5555]$
Qwen3.8	L4	22	8	0.4567	$[0.0512, 0.8042]$
Qwen3.8	Rerank-8	154	76	0.4811	$[0.3571, 0.6062]$
Qwen3.8	RFT	54	13	0.1956	$[0.0761, 0.4253]$

key/arm CSV reports, including missing-label bounds and conditional denominators. Original corpora and judge records are not modified.

D.8.3 HISTORICAL SINGLE-SEED AUDIT

Table 22: Historical seed-0 verifier-guidance audit, retained separately from the new matched study. Automatic metrics use 5000 bodies. Luna uses 200 matched bodies per arm, weighted to the eligible matched-corpus mixture. The selection verifier is fitted to L4 outputs. Original historical reward weights were not archived. L4 is the original release, not the separately sampled first-candidate pool in the nested-N sweep. Serving candidates exclude the one-time RFT update.

Method	Auto Recovery ↑	GPT full ↑	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑	Serving candidates/body
L4	0.440	0.434	**0.114**	**0.126**	0.0124	**0.840**	1
OUTPUT-VERIFIER RFT (β=0.05)	0.469	**0.489**	0.145	0.167	0.0163	0.796	1
VERIFIER RERANK-8	**0.472**	0.482	0.115	**0.126**	**0.0117**	**0.840**	8

Table 22 compares both historical control procedures with the original L4 release. The nested automatic sweep uses a separately sampled candidate pool and increases recovery from 0.437 at N=1 to 0.452, 0.462, and 0.472 at N=2, 4, 8. On the independent GPT judge, OUTPUT-VERIFIER RFT improves over L4 by +0.054 (hierarchical paired 95% CI $[+0.015, +0.094]$), while VERIFIER RERANK-8 improves by +0.047 ($[+0.005, +0.089]$). Their difference is not detectable: OUTPUT-VERIFIER RFT−VERIFIER RERANK-8 is +0.007 ($[-0.032, +0.046]$). Because this audit has only one training seed, these intervals quantify matched row and cluster uncertainty, not across-seed variation.

D.9 HISTORICAL VERIFIER-GUIDANCE COST

Table 23: Seed-0 cost accounting for one 5000-body release. Historical decoded bodies are re-tokenized with the recorded Qwen tokenizer. The first candidate in each nested reranking pool estimates total candidate tokens. RFT rollout token IDs are not accumulated historically, so only their exact configured maximum-token budget is available. A rerun matching the recorded checkpoint, sampler, seed, requested tags, and token cap directly times final RFT generation. Sampling-temperature equivalence is unverified. The shared verifier-corpus generation remains proxied.

Component	VERIFIER RERANK-8	OUTPUT-VERIFIER RFT
Recorded timed stage	22.59 GPU-h: candidate generation, verifier scoring, nested selection, and resumable state writes	2.816 GPU-h: rollout generation, reward evaluation, reference and policy log-probabilities, backward passes, clipping, and optimizer steps
Generated items	40,000 candidates, maximum 120 tokens, 4.800M-token configured budget, 3.859M re-tokenized-token estimate, 95% t-CI [3.825, 3.893]M	12,800 rollouts, maximum 160 tokens, 2.048M-token configured budget. Historical realized count not logged
Final release	Selected inside the timed state, with final JSONL materialization excluded. 477,530 re-tokenized output tokens	0.774 GPU-h measured rerun, 450,788 exact sampled tokens including EOS (433,449 re-tokenized). The historical release has 459,610 re-tokenized tokens
Incremental 5000-body total after the shared verifier corpus	22.59 GPU-h measured selection stage	3.590 GPU-h measured stage sum
Stage sum from the DP-LoRA checkpoint	23.37–23.45 GPU-h	4.36–4.45 GPU-h
Excluded or separately reported	Model loading, shared 5000-body verifier-corpus generation, verifier fitting, final JSONL materialization, evaluation, and upstream DP-LoRA training	RFT model/optimizer setup, shared verifier-corpus generation, verifier fitting, checkpoint serialization, evaluation, and upstream DP-LoRA training. Final generation does include its own model load and serialization

The direct rerun used the same checkpoint, cluster sampler, seed, 120-token cap, and all 5000 requested skeletons as the historical release, in batches of 16 on a dedicated NVIDIA A100 80GB PCIe. All requested skeletons match, but decoded bodies remain stochastic. Historical temperature and top-p were not persisted, so the saved match checks do not establish full decoding-configuration equivalence (Appendix G.3). The measured command takes 2785.79 seconds including model load and JSONL serialization. Its 450,788 exact sampled tokens correspond to 433,449 re-tokenized decoded tokens. The earlier reranking-derived final-generation proxy was 3.48–3.65$\times$ larger and is superseded.

The shared verifier corpus contains 482,196 re-tokenized output tokens but was not timed. Applying the measured one-pass stage directly, or scaling it by re-tokenized volume, gives a 0.774–0.861-hour common proxy in the checkpoint-started totals above. Verifier fitting itself is not timed on the original server. A three-run local 16-logical-core CPU replay has a 5.30-second median and is reported separately rather than folded into GPU-hours. The historical reranking and RFT-optimization timers exclude model-process setup.

Consequently, 2.03 seconds per reranking candidate and 0.79 seconds per RFT rollout are not comparable throughput statistics: the caps are 120 versus 160, the generation batches are 16 versus 64, and the RFT timer additionally includes two log-probability passes and backward optimization. The observed reranking run is therefore evidence about this implementation, not an algorithmic speed ratio. The clearest serving-time comparison is candidate count: best-of-8 requires eight completions for every query, whereas an already-trained RFT policy requires one. Using the observed batched release rates gives an illustrative amortization estimate of approximately 645 bodies (712 under the candidate-only approximation). This extrapolation assumes those rates persist and is not a measured deployment break-even point. Table 22 favors reranking on fidelity in one seed. Neither a universal offline preference nor a repeated-query systems advantage is established. In particular, eight candidates versus one does not imply eight-fold latency or total cost, and we do not evaluate unseen control combinations.

D.10 ABLATIONS

Table 24 decomposes the generator-DP recipe at $\varepsilon_{\text{learn}}=4$. In this single-seed grid, LoRA drives most of the fidelity gain: a soft-prompt alone (no LoRA) substantially worsens joint fit, with joint-

Table 24: Matched component ablation at $\varepsilon_{\text{learn}}=4$ (full-pool setting with up to 200,000 records scanned, single seed, 5000 samples). The reference is a separately generated LoRA+cluster run using the same recipe as L4. Each row changes one component within this ablation grid. TSTR is the mean direct needs-decision macro-F1 over ten deterministic, matched-size synthetic subsamples (higher is better).

Method / ablation	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑	TSTR ↑
LoRA+Cluster (ablation reference)	0.112	0.115	0.0122	0.843	0.614
Conditioning: none	0.128	0.132	0.0166	0.833	0.627
Conditioning: fixed instruction	0.138	0.161	0.0163	0.835	0.624
Adapter: soft-prompt only (no LoRA)	0.108	0.215	0.0282	0.546	0.588
Adapter: soft-prompt + LoRA	0.109	0.143	0.0148	0.812	0.628
Marginal variant (vs. cluster)	0.115	0.128	0.0123	0.839	0.621
Real-Verifier RFT (diagnostic)	0.110	0.162	0.0187	0.785	0.616
non-private reference ($\varepsilon=\infty$)	0.103	0.124	0.0091	0.921	0.598

TVD $0.115 \to 0.215$ and coverage $0.843 \to 0.546$, whereas adding a soft-prompt on top of LoRA remains much closer to LoRA alone but lowers coverage to 0.812. Relative to structured conditioning, the fixed-instruction variant has higher mean-TVD ($0.112 \to 0.138$) and joint-TVD ($0.115 \to 0.161$), while unconditional generation has higher mean-TVD ($0.112 \to 0.128$). These contrasts change conditioning, not clustering alone. The LoRA+cluster reference sits close to the non-private reference on marginal fidelity, differing by 0.009 mean-TVD (0.112 vs. 0.103). Cluster versus marginal variants differ little on mean-TVD (0.112 vs. 0.115), and the real-verifier diagnostic exhibits the same fidelity–recovery trade confirmed by Output-Verifier RFT.

E Limitations

Scope. The deepest evidence is concentrated on one pairing, Enron with Qwen3.5-4B, at $\varepsilon_{\text{learn}}=4$. The learner-budget sweep uses three seeds, with automatic metrics only at learner epsilons two and eight. Fidelity is not monotonic across the tested range, and the two transfer screens use a deliberately shorter training budget. Amazon has an independent judge, SmolLM3 does not, and neither has a privacy audit. SmolLM3 also shows a substantially less favorable fidelity and utility trade-off. The six-seed automatic lift has exact two-sided sign-flip $p = 0.03125$, but the primary independent-judge evaluation uses only three of those six seeds. The new matched inference audit also uses three seeds. The earlier reranking audit remains single-seed. In the new audit, verifier reconstructions are not proven identical to the historical rewards, and the pinned evaluation-base revision does not establish the historical training revision. The placement study also compares two endpoint allocations, the whole learner budget on the generator against the whole learner budget on the reward, and never an interior split such as $\varepsilon_{\text{gen}}=2$ with $\varepsilon_{\text{reward}}=2$. Nothing we measure rules out an interior point that beats both endpoints.

Instruments. One real-fitted labeler supplies both TVD and automatic recovery, and its weak held-out F1 on intent, response, and urgency limits confidence in recovery measurements on those keys. The final embedding audit calibrates the MMD^2 kernel once from 1000 real calibration bodies and evaluates against a disjoint 2000-body real reference, with users disjoint between those two sets. It persists the hashed selection manifest and reuses the same kernel coefficient and coverage radii across all arms within each profile. For $k(x, y) = \exp(-\gamma\|x - y\|^2)$, the Enron and Amazon coefficients γ are 0.6313 and 0.5422, respectively. SmolLM3 reuses the Enron reference. Historical embedding evaluation also calibrates on real text, but uses only 500 reference bodies and does not persist the calibration selection. Coverage remains sample-size dependent and should not be compared across the 5000-body primary and 1000-body transfer settings. Enron has no model-training holdout shared by every arm, so "held out" here means calibration and scoring-reference disjointness, not exclusion from every trained model.

The final utility audit matches 2000 real and synthetic training examples (1000 each in transfer), averages ten deterministic primary-study subsamples within each seed, and reports direct TSTR macro-F1 rather than a ratio. Transfer uses the full 1000-body corpus, so its ten repeats are identical

rather than distinct subsamples. The paired audit also evaluates original requested labels, but it still uses narrow automatic tasks and linear classifiers, so it is not a deployment study. Amazon utility uses the canonical reference's rule-derived intent labels, distinct from the LLM-enriched upstream annotations (Appendix D.3).

The reconstructed-verifier diagnostic finds no aggregate evidence of net requested-label error correction, despite informative but attribute-dependent reward signals (Appendix D.8.2). It does not identify the causal mechanism of historical RFT. Its conditional high-reward failures are neither corpus-wide quality rates nor causal comparisons between controllers.

Accounting and implementation. The generator's training set uses global quotas, at most 400 examples per public cluster and eight per exact skeleton, on top of the per-user cap. Global admission precedes the user-cap check, allowing cascades that change other users' selected examples. The cap does not bound the size of that change. Global scan prefixes can also couple users. The learner accountant is calibrated for selected-input parameters, whereas structural accounting is intended for raw-user adjacency. Their sum is a nominal component ledger, not an established DP bound on the historical joint raw-corpus release (§3.6). The selection gap also complicates the placement comparison, because the generator learner and the reward learner draw different selected streams, 8934 examples from 4518 senders against $79{,}077$ records from $36{,}501$ users. Retaining the pre-sweep $\beta{=}0.05$ headline setting avoids post-selection from that sweep (§5.4), not all data dependence in the research workflow. Real-data evaluation informs comparisons and reporting. A confidential release would need separate accounting for those outputs and any data-dependent model selection. Checked local privacy-noise randomness, calibration, data-dependent sampling parameters, identity invariants, and diagnostic outputs impose additional conditions (Appendix G.2). A valid upstream DP mechanism is a hypothesis of the post-processing theorem, not its conclusion.

Selector repair pilot. We implemented a raw-user-decomposable alternative that excludes records without stable sender identities and independently retains deterministic bottom-k stable record IDs within each user. Its user-decomposability requires stable identities and no global scan or record cap. It removes that particular coupling, not all obstacles to a privacy guarantee. In two matched short-budget pilots (500 DP steps, cap three, and 1000 aligned evaluation outputs per arm), the seven predeclared gates bound alternative-minus-historical L4 differences: mean-TVD and joint-TVD increases of at most 0.03 each, fixed-reference MMD^2 increase of at most 0.01, coverage drop of at most 0.05, and TSTR macro-F1 drops of at most 0.03 on each of needs-decision, requires-response, and triage. Each utility gate uses the worse of the full-corpus and 800-body subsample-mean differences. The same limits apply to the alternative's incremental RFT and final alternative-versus-historical RFT comparisons. Positive recovery lift is checked separately.

The L4 stage passed six of seven gates in seed 0 (joint-TVD regressed by $+0.042$, above the 0.03 gate) and all seven in seed 1. Seed-0 RFT passed every incremental and final gate, with a recovery lift of $+0.054$ (paired corpus-bootstrap 95% CI $[+0.041, +0.066]$). However, the configuration-matched seed-1 correction used 3992 verifier-training and 997 verifier-test bodies per selector and still failed four of seven incremental gates despite a $+0.043$ recovery lift ($[+0.030, +0.055]$): mean-TVD changed by $+0.032$, joint-TVD by $+0.052$ ($[+0.011, +0.085]$), fixed MMD^2 by $+0.0172$, and triage TSTR macro-F1 by -0.039. The final safe-versus-historical RFT comparison also missed the 0.03 mean- and joint-TVD gates ($+0.0308$ and $+0.0311$), while safe recovery is lower by 0.029 ($[-0.041, -0.017]$). These bootstrap intervals condition on the trained checkpoints and do not measure training-seed variation.

Seed 0 therefore cleared every RFT gate while seed 1 failed four, so these short-budget pilots do not settle how the safe selector behaves at the headline training budget. The headline checkpoints retain nominal selected-input learner accounting, and full-budget retraining with the alternative remains unperformed. Failing utility gates does not justify a raw-corpus privacy claim, and the user-decomposable alternative's full-budget utility remains unmeasured.

Privacy evidence and output safety. Membership inference uses two clean DP seeds and one matched user pool, and the canary audit uses a single phrase with eight-user insertion, a recorded stress-test setting rather than the formal single-user neighboring case (Appendix D.5). Regex PII counts and nearest-neighbor overlap remain diagnostics. SmolLM3 emits many address-like strings, averaging 1181 regex detections per 1000 outputs, which is a deployment-quality failure that DP does not address. Those strings are not near-duplicates under our audit, but require output filtering

or a different decoding recipe before release. More attack families, DP seeds, single-user canaries, and extraction prompts would strengthen the empirical privacy evidence. The audits probe particular memorization scenarios, not an upper bound on leakage.

Replication and comparators. Comparator replication is uneven: the reward-DP methods use three seeds and the Aug-PE references are single runs. The Aug-PE implementation follows the published candidate-generation, private-voting, and iterative-variation loop, but adapts its privacy unit and prompts to this email setting, so it is not an exact reproduction of the original experiments. The reward-quality audit cannot provide a user-disjoint test for the existing checkpoints, because their capped training stream contains all eligible users. Its record holdout consists of examples excluded by contribution capping from users that remain represented in training. The KL audit covers three seeds and three RFT weights rather than a dense hyperparameter surface. Only $\beta{=}0.05$ has six-seed automatic evaluation and the predeclared three-seed independent-judge evaluation. The apparent $\beta{=}0.10$ compromise, which improves every fidelity mean while retaining most of the recovery lift, is therefore less thoroughly replicated. The hierarchical intervals in the learner-budget sweep resample only three observed training seeds and should not be read as high-precision population-of-runs estimates. Budgets below the sweep's lowest setting, $\varepsilon_{\text{learn}}{=}2$, require additional study. The reward-quality audit establishes that the comparator is informative, not that it exhausts stronger reward architectures or optimally tuned mixed-budget designs. Neither the historical single-seed audit nor the new three-seed inference audit establishes a generally preferred controller, deployment break-even, or unseen-control generalization. The new study supports additional RFT recovery at a fidelity cost, under its fixed inference protocol and reconstructed-verifier comparison.

Instrument bias. The same real-fitted measurement labeler is used across methods for TVD and automatic recovery, which improves consistency but does not eliminate annotation bias. The independent Qwen and GPT judges mitigate rather than remove this concern. They cover the primary experiment, and GPT also covers Amazon, but SmolLM3 remains automatic-only. For Amazon, held-out macro-F1 ranges from 0.433 for star rating and 0.457 for intent to 0.809 for product group, so the transfer values should be read as instrument-relative. Three semantic Amazon controls, review intent, dominant aspect, and formality, were themselves assigned by a fixed Qwen3.8-27B annotator, so annotation-model bias can enter both the target taxonomy and its evaluator. The proprietary GPT deployment may change over time, so we record its deployment alias, API style and version, evaluation date, prompt hash, and output schema. Our utility evaluation uses two rule-derived binary tasks and one taxonomy label with linear classifiers, which does not represent a full end-user deployment evaluation. Agreement between the real-fitted labeler and both independent judges supports the direction of a result. Values for the weak keys remain instrument-relative.

Pretraining confound. The frozen-base quality control clarifies but does not eliminate the pretraining confound. Frozen Qwen reaches 0.429 automatic recovery, only 0.007 below DP-LoRA, so much basic instruction following predates private training, while its poor joint-TVD, MMD^2, and coverage show that adaptation supplies the domain match. The adapted models emit roughly seven times as many regex PII-like spans as the frozen base, largely through email-style headers and identifiers. This is an output-safety cost, not evidence of memorization by itself. The matched lexical control finds no evidence that private adaptation adds training-specific verbatim overlap, but a public base model may still have encountered Enron during pretraining, so the lexical audit cannot separate non-verbatim pretrained corpus knowledge from information learned during private adaptation. Amazon Reviews is a different domain but is also a large public web corpus and is not a clean pretraining-contamination control. Kurakin et al. (2023) identify this hazard for DP synthetic text and de-duplicate their pretraining corpus against sensitive data. We do not perform that de-duplication. Evaluation on a non-public or post-training-cutoff corpus would provide a stronger test.

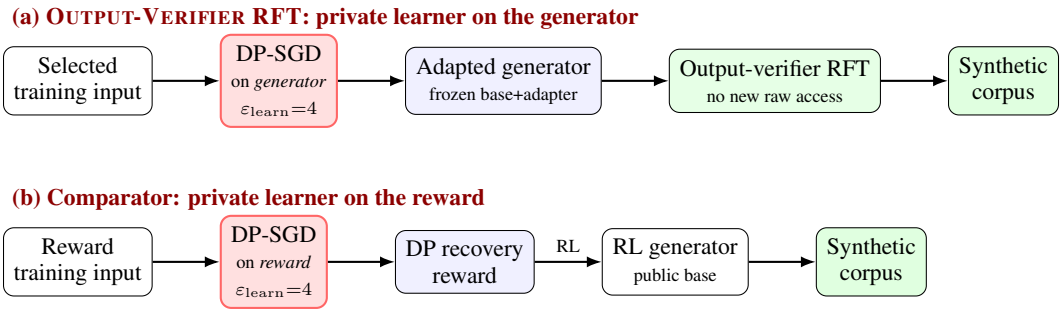


Figure 2: **Private-learning endpoints.** The generator-side design (a) learns from bodies, then fits its verifier on generated outputs. The objective-matched reward-side design (b), REWARD-DP-GROUND+CLUSTER, learns an attribute-recovery reward before policy optimization. Learner (ε, δ) targets and the attribute-recovery objective are matched, but populations, capacity, and supervision differ. The panels therefore compare two system configurations rather than isolating budget placement. Both primary arms use the same frozen `Qwen3.5-4B` base. The reward-side policy starts without the privately learned generator adapter. The common structural release is omitted. Historical quota selection precedes (a). DP labels and epsilon annotations denote recipe targets, not certified component, artifact-level, or end-to-end bounds (§3.6, Appendix G.2).

F SUPPLEMENTARY FIGURES AND RESULT TABLES

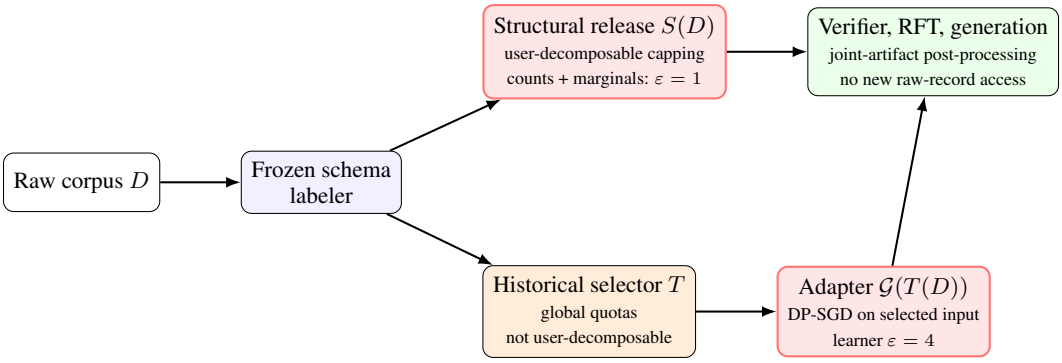


Figure 3: **Executed dependency graph.** The structural recipe and generator accountant have different intended input scopes. The historical selector can change other users' selected records, so component accountant values do not certify the joint raw-corpus release. RFT is post-processing of those joint artifacts, not a repair of the selector. DP labels and epsilon annotations are nominal recipe targets. Structural capping also requires stable identities and unique-record semantics, and implementation qualifications apply (Appendix G.2).

This section holds floats referenced from the main text but placed here for length.

Table 25: Per-key recovery on the primary Qwen judge (OUTPUT-VERIFIER RFT vs. L4), averaged over three cluster-weighted seeds. OUTPUT-VERIFIER RFT is higher on 7/8 keys. The only regression is urgency.

	triage	topic	intent	resp.	urg.	sent.	diff.	form.
L4	0.419	0.374	0.303	0.167	**0.345**	0.701	0.840	0.608
OUTPUT-VERIFIER RFT	**0.495**	**0.761**	**0.368**	**0.189**	0.312	**0.739**	**0.855**	**0.714**

Table 26: OUTPUT-VERIFIER RFT raises recovery in both transfer settings and lowers coverage in both. External-validity screen at $\varepsilon_{\text{learn}}=4$ (component ledger sum 5, not a raw-corpus bound), with 500 generator DP-SGD steps, generator contribution cap three, and 1000 samples/arm/seed. Values are mean $\pm$ sd over three seeds in both settings. Recovery is scored per requested key under the profile-specific automatic measurement labeler. TSTR is star-rating macro-F1 for Amazon and needs-decision macro-F1 for Enron, using the full 1000-body synthetic corpus and 1000 real training examples. Repeated full-corpus fits are identical. TSTR summaries use unrounded matched-utility scores. Per-seed values are in Tables 39 and 41. The reward-DP model retains its stronger full-cap training recipe (Appendix G.3).

Setting	Arm	Recovery ↑	mean-TVD ↓	joint-TVD ↓	MMD^2 ↓	Cov ↑	TSTR F1 ↑
Amazon/ Qwen	DP-LORA	0.575±0.004	0.108±0.010	**0.158**±0.015	**0.0147**±0.0048	**0.713**±0.020	0.355±0.013
	OUTPUT-VERIFIER RFT	**0.690**±0.012	**0.105**±0.004	0.250±0.003	0.0297±0.0023	0.672±0.019	**0.365**±0.026
	REWARD-DP-GROUND+CLUSTER	0.523±0.019	0.163±0.012	0.313±0.008	0.0546±0.0066	0.407±0.010	0.358±0.014
Enron/ SmolLM3	DP-LORA	0.422±0.003	**0.133**±0.011	**0.216**±0.019	**0.0152**±0.0008	**0.542**±0.011	0.602±0.018
	OUTPUT-VERIFIER RFT	**0.453**±0.013	0.164±0.023	0.226±0.040	0.0266±0.0043	0.431±0.028	0.578±0.021
	REWARD-DP-GROUND+CLUSTER	0.449±0.011	0.220±0.013	0.261±0.033	0.0740±0.0079	0.179±0.012	**0.606**±0.015

Table 27: Comparison design. "+Cluster" means that an otherwise reward-DP method is strengthened with our cluster-conditioned structural prior. It is therefore an adapted comparator, not an untouched literature baseline. AUG-PE+LEVERS uses our lever interface and DP marginals but *not* our cluster sampler. VERIFIER RERANK-8 uses an output-derived verifier reconstruction for inference-time search. Historical reward-weight identity is not established. DP names and numerical annotations denote nominal recipes, not certified artifact privacy. Learner-input and raw-corpus scopes are distinguished in §3.6 and Appendix G.2.

Method	Where private learning occurs	Control layer	Role in the comparison
L4 (DP-LoRA generator)	DP generator, $\varepsilon_{\text{learn}}=4$	Our cluster prior + DP marginals	Pre-RFT ablation that isolates the effect of OUTPUT-VERIFIER RFT
OUTPUT-VERIFIER RFT	Same L4. RFT is post-processing	Our cluster prior + DP marginals	Proposed post-processing method with unchanged upstream scope
VERIFIER RERANK-8	Same L4. Selection is post-processing	Our cluster prior + DP marginals	Inference-time search baseline with eight candidates per released body
REAL-VERIFIER RFT	Same L4. Verifier additionally reads raw real bodies	Our cluster prior + DP marginals	Verifier-source diagnostic only, not an accounted private release
REWARD-DP-DISC+CLUSTER	DP real/fake reward, $\varepsilon_{\text{learn}}=4$	Our cluster prior + DP marginals	Learner-budget-matched reward-side design
REWARD-DP-GROUND+CLUSTER	DP recovery reward, $\varepsilon_{\text{learn}}=4$	Our cluster prior + DP marginals	Objective-matched placement comparator
AUG-PE+LEVERS	DP selection, $\varepsilon_{\text{select}}=4$	DP marginals, no cluster sampler	Control-aware adaptation of the external Aug-PE baseline
AUG-PE-RAW	DP selection, $\varepsilon_{\text{select}}=4$	No requested controls	Schema-free, literature-aligned Aug-PE baseline

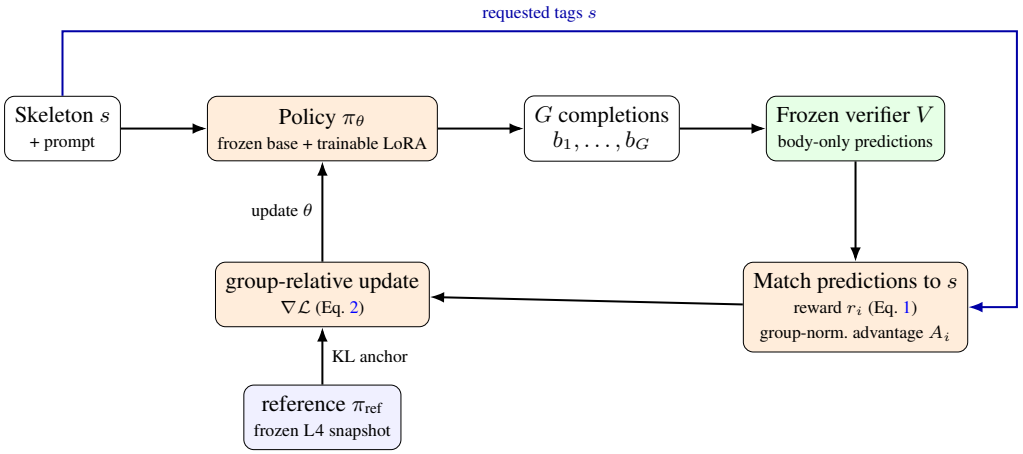


Figure 4: **Post-hoc DP-output-verifier RFT (§3.5).** For each skeleton the DP-LoRA policy samples G completions. A frozen verifier fit on upstream outputs predicts attributes from each body alone. Comparing its predictions with the requested tags gives the reward (Eq. 1). Group-normalized advantages drive a policy-gradient update (Eq. 2) regularized against the frozen L4 reference. No raw data is read after DP training by this loop. The verifier, prompts, and initialization depend only on upstream artifacts, so RFT has the post-processing form. DP labels denote the nominal upstream recipe. Privacy preservation requires a valid joint upstream guarantee and does not repair its selection or implementation conditions.

Table 28: Nominal component ledger at learner epsilon four. The learner is either the generator adapter or the reward model. Entries are configured targets or post-processing increments, not certified historical component bounds. Composition requires every component's preprocessing, randomness, parameters, and released outputs to satisfy a common input-adjacency contract. The geometry's zero is for public inputs only. The downstream zero assumes a valid joint DP ancestor and no fresh private access. Neither the ledger sum nor whole diagnostic reports are certified by these entries (Appendix G.2).

Step	Mechanism	Nominal (ε, δ) entry
Categorical embed / cluster geometry	public model	$(0, 0)$
Cluster occupancy counts	Gaussian histogram recipe	$(0.5, 10^{-5})$
Categorical marginals	Gaussian marginal recipe	$(0.5, 10^{-5})$
Arm-specific learner	user-level DP-SGD recipe	$(4.0, 10^{-5})$
Historical generator selection	global quotas	unaccounted on raw corpus
RFT and generation	conditional post-processing	$(0, 0)$
Nominal-component sum	not an end-to-end bound	$\varepsilon_{\text{ledger}} = 5.0$ $\delta_{\text{ledger}} = 3\times10^{-5}$

G Reproducibility and Supplementary Details

This appendix distinguishes preserved run settings from configured defaults. The current pipeline namespaces checkpoints, structural artifacts, generated corpora, and evaluation reports by dataset profile and stamps them with a profile name and schema-version fingerprint. Older Enron artifacts do not uniformly contain these fields. The fingerprint is the first 16 hexadecimal characters of SHA-256 over the profile's field definitions, categorical vocabularies, cluster keys, and evaluated two-way pairs. It is not a hash of the underlying records. The primary Enron fingerprint is `6e79055212368483`, and the final Amazon fingerprint is `09288c5ce2987e20`. Cross-profile loading is rejected.

G.1 Dataset profiles and preprocessing

Enron. The primary experiments use the public Enron corpus (Klimt & Yang, 2004). The exact full-pool generator selection scans at most 200,000 canonical records, applies cluster-stratified selection with a maximum of 400 examples per public cluster and eight per exact skeleton, and contains 8934 examples from 4518 senders after the ten-example user cap. The intended privacy unit is the configured sender key, not a verified human identity. The eight controls and their public vocabularies are listed in Table 29. The public cluster geometry embeds the Cartesian product of `topic`, `intent_primary`, and `formality` and contains 27 clusters.

Table 29: Primary Enron control vocabularies. These strings are public. Per-record assignments remain inside the private-data boundary.

Control	Allowed values
`triage`	junk, informational, actionable
`topic`	project engineering, sales/business, finance/procurement, HR/people, legal/compliance, operations/logistics, marketing/communications, IT/infrastructure, management/strategy, personal/social
`intent_primary`	request action, request information, request approval, request meeting, request help, inform update, inform deliverable, announce, confirm, decline, acknowledge, complain, escalate, appreciate, instruct, assign, set deadline, follow up, propose, introduce
`response_primary`	reply with answer, reply with decision, reply with deliverable, confirm, accept/decline, review/approve, complete task, schedule/attend, forward/delegate, acknowledge only, no-action/FYI
`urgency_level`	none, low, medium, high, critical
`sentiment`	positive, neutral, negative, frustrated
`difficulty`	easy, medium, hard, ambiguous
`formality`	formal, semi-formal, informal, automated

Amazon Reviews 2023. The dataset-transfer experiment uses Amazon Reviews 2023 (Hou et al., 2024). We stream the first 250,000 reviews from each of six official category archives: All Beauty, Amazon Fashion, Appliances, Cell Phones and Accessories, Grocery and Gourmet Food, and Books. The preprocessing pass removes malformed records, missing users, duplicate normalized text, reviews shorter than 20 characters, and records whose text is not ASCII-dominant. It retains users with at least two valid reviews and assigns whole users deterministically to train/validation/test (Table 30).

Table 30: Canonical Amazon transfer corpus. Splits are user-disjoint.

Split	Reviews	Users
Train	136253	23194
Validation	17083	2879
Test	16375	2859
Total	169711	28932

Product group and star rating are native fields. Detail level is a deterministic length bucket. Review intent, dominant aspect, and formality are assigned once with a frozen, schema-constrained `Qwen3.8-27B` annotator (Qwen Team, 2026c). Table 31 lists the six controls and their public vocabularies. The annotation run uses prompt version `amazon_semantic_qwen_v1`, temperature zero, seed zero, thinking disabled, at most 1600 characters per review, and a strict JSON schema. It labels all 169711

records in 14143 API calls with zero retries. The canonical semantic output SHA-256 is `68cfad88fd653ba442c8c97d3bc2ca63bae348104b9dc467ad0fcabc45fc20f1`. Rule-derived prelabels are retained only for audit. A stratified pilot was inspected before the full run.

Table 31: Amazon transfer controls and public vocabularies.

Control	Allowed values
`product_group`	beauty/personal care, fashion, home/garden, electronics/software, food/household, books/media
`star_rating`	one star, two star, three star, four star, five star
`review_intent`	recommend, warn against, report problem, compare options, advise buyer, describe experience
`dominant_aspect`	quality/performance, usability, durability/reliability, value/price, fit/size, delivery/packaging, appearance/design, service/support, content/story, general experience
`detail_level`	brief, standard, detailed
`formality`	formal, conversational, terse

Amazon semantic-annotation instruction. The fixed system instruction tells the model that every review is untrusted quoted data, forbids following instructions inside the review, requires classification from the review text alone, and requires strict JSON with one allowed value for each field. The operational definitions are:

- **Intent:** `recommend` explicitly endorses purchase/use. `warn_against` explicitly discourages it. `report_problem` centers a defect, failure, damage, missing part, or service problem. `compare_options` substantively compares alternatives. `advise_buyer` gives purchase/setup/sizing/usage advice without mainly recommending or warning. Otherwise use `describe_experience`.
- **Aspect:** choose the topic receiving the most substantive attention: effectiveness/function/output, setup/convenience, lifespan, price/value, dimensions/fit, delivery/packaging, appearance/design, seller/returns/support, media content/story, or `general_experience` when none dominates.
- **Formality:** `terse` is telegraphic or very short. `conversational` is ordinary personal language. `formal` is impersonal, professional, or consistently carefully composed.

The user message contains a JSON array of integer IDs and truncated review text and asks the model to preserve each ID. A strict response schema requires an array of the same length with fields `id`, `review_intent`, `dominant_aspect`, and `formality`. Off-vocabulary or duplicate IDs cause the entire batch to be retried.

The Amazon public geometry embeds `product_group`×`review_intent`× `star_rating` and uses 24 clusters. The selected DP structural release uses 84738 training-split records after deterministic bottom-10 contribution capping. Its noised cluster mixture has TVD 0.008 from the corresponding capped non-private mixture. One-way marginal TVDs are 0.0022–0.0089. Selected two-way marginal TVDs are 0.0179–0.0791. The structural reports contain neither review text nor user identifiers, but their raw counts and comparisons to non-private distributions are not thereby DP outputs (Appendix G.2).

The released Enron two-way marginals are triage×urgency, intent×response, topic×formality, and sentiment×urgency. The Amazon pairs are rating×intent, group×aspect, intent×aspect, and detail×formality.

Profile-specific prompts. Both profiles use the same deterministic recipe-card grammar:

```
[key_1=value_1 | key_2=value_2 | ... | key_K=value_K]
---
<BODY_LABEL>:
```

where `BODY_LABEL` is `Email` or `Review`. The public initialization strings are "email message" and "product review", and the corresponding fixed instructions are "Write an email." and "Write a product review." No user identifier, product identifier, subject, recipient list, or other high-cardinality metadata appears in the recipe card.

G.2 PRIVACY ACCOUNTING AND RELEASE SEMANTICS

For raw input D, structural capping is intended to be user-decomposable. The generator is trained on $D_{\text{sel}} = T(D)$. Its accountant is calibrated at selected-input parameters, not for the map that recomputes T on neighboring raw corpora, and does not by itself certify the saved artifacts. Cluster occupancy and categorical marginals each target $(0.5, 10^{-5})$, and the learner targets $(\varepsilon_{\text{learn}}, 10^{-5})$. If every component satisfied its DP guarantee in a valid sequential composition under the same protected-input adjacency, including its preprocessing, sampling, parameter selection, private randomness, and actual released outputs, basic composition would give

$$(\varepsilon_{\text{tot}}, \delta_{\text{tot}}) \leq (1 + \varepsilon_{\text{learn}}, 3 \times 10^{-5}),$$

per candidate. The values 3, 5, and 9 record nominal component-ledger sums at learner epsilon 2, 4, and 8, and the transfer ledger sums to 5. These sums illustrate composition only if all of the assumptions above hold. None is an established guarantee for the historical release, fixing the selector alone would not establish one, and treating the privately selected cohort as fixed after the fact does not account for the map that selected it.

Intended protected input and preprocessing contract. The intended adjacency adds or removes all records attached to one configured sender/user key, not necessarily one human identity. A confidential-data design would need stable user and record identities and unique-record semantics throughout. Structural bottom-k capping is user-decomposable under those invariants, but checked local calls permit line-based fallback identities, and duplicate-key rows can be admitted multiple times by counting passes. These are domain conditions to verify, not evidence that the archived corpora violate them. The primary historical generator admits records under global 400-per-cluster and 8-per-exact-tuple counters before checking the user cap of 10. A selector-only synthetic counterexample at these actual limits removes one single-record user and changes selected examples of 12 other existing users through cascades. Thus the cap does not bound the group size of the selector change, and no weaker raw-user bound has been calibrated from it. A global `max_scan` prefix can also couple users. Independent bottom-k selection within users avoids this particular coupling only with stable identities and no global scan or record cap. It does not certify the rest of the release mechanism.

Before a structural histogram or marginal is formed, records are grouped by user and the contribution cap is applied with deterministic bottom-k selection under a record hash, subject to the identity conditions above. With stable identities this avoids the file-order bias of taking each user's first k records.

On the learner side, the configured DP-SGD recipe Poisson-samples users. The losses for all selected examples of one user are accumulated, that user's joint gradient over all trainable parameters is clipped to norm $C = 1$, Gaussian noise is added to the aggregate, and the result is divided by the expected lot size. The Opacus RDP accountant calibrates the noise multiplier to the requested epsilon, delta, sample rate, and fixed step horizon, rather than stopping adaptively on a measured epsilon. In the checked generator, $q = \min(1, L/N_{\text{sel}})$ uses configured users per lot L and observed selected-user count N_{sel}, and the noise multiplier is calibrated from q. However, an accountant evaluated at this observed count does not automatically certify recomputing parameters under add/remove-user adjacency. A prospective design must therefore fix public design parameters or account for their data dependence, and match the actual sampler to that contract.

Implementation qualifications. Checked local structural Gaussian draws default to NumPy seed zero. Learner privacy sampling and Gaussian noise use the configured seed, which is retained in checkpoints and reports, including the fast path. The usual randomized-mechanism proof does not certify a fixed or publicly reproducible privacy-noise release. A future confidential implementation requires private entropy and appropriate independent random streams, not another hard-coded seed or removal of an already published seed. Public geometry, generation, and evaluation seeds are separate reproducibility choices. The custom structural analytic-Gaussian helper also has an incorrect bisection update in its decreasing branch, so its calibration cannot be assumed valid for every target. The saved Enron structural releases at epsilon 0.5 use conservative noise scales under the corresponding randomized Gaussian analysis. This finding does not show those releases were undernoised and does not resolve their randomness conditions. Learner DP-SGD calibration uses

Opacus, not that custom helper. Available local output scales agree with their metadata, but historical code/runtime coverage is incomplete: these checks do not prove that every server run used the exact checked implementation. No attack or disclosure is demonstrated by these implementation findings.

Released outputs and shared ancestors. Structural reports include `records_used`, `dp_resolvable_kmax`, and comparisons to non-private distributions such as `pi_tilde_tvd_vs_nonprivate` or `tvd_vs_nonprivate`. Training reports and logs include selected-user/example/cluster counts, raw losses, and configuration seeds. These fields are not outputs solely of the nominal DP mechanisms. Neither the generator accountant nor the presence of DP fields certifies an entire JSON report. A prospective confidential release must explicitly separate cleared public artifacts from internal diagnostics, or supply a separate privacy or permission basis for each output. RFT, candidate generation, verifier reranking, and sampling reuse the same joint upstream artifacts and public models. If that joint ancestor were validly DP, releasing its descendants together would preserve its guarantee without charging again for the same draw. Independently fitted models, including reward learners and different training seeds, require joint-release accounting, not the maximum of individual budgets. Shared ancestors should not be counted repeatedly either. Fresh private access for reward fitting, model selection, or real-data evaluation requires its own accounting or permission basis. Released intermediates and data-dependent selection belong in the full release analysis. An unreleased failed attempt is not automatically a separately released mechanism. The raw-real verifier diagnostic and reported real-data selection and evaluation are public-corpus research procedures, not an accounted confidential-data selection-and-reporting pipeline.

G.3 Training and generation configuration

Table 32 gives the DP generator configuration and Table 33 the post-processing RFT and final generation settings.

Table 32: Nominal DP generator recipe, corroborated by available run reports. "Primary" is the full Enron/Qwen experiment, and "transfer" covers Amazon/Qwen and Enron/SmolLM3. Calibration targets are not artifact-level privacy certificates (Appendix G.2). Archive coverage is qualified in Appendix G.9.

Parameter	Primary full-pool	Transfer screen
Base model	Qwen3.5-4B	Qwen3.5-4B or SmolLM3-3B
Learner epsilon	$2, 4, 8$	4
Target delta	10^{-5}	10^{-5}
DP-SGD steps	2000	500
Examples/user cap in generator	10	3
Expected users/lot	128	128
Per-user clip norm	1.0	1.0
Optimizer / learning rate	Adam / 5×10^{-4}	Adam / 5×10^{-4}
Selection	public-cluster stratified	public-cluster stratified
Per-cluster / per-skeleton cap	400 / 8	400 / 8
Maximum scanned records	200000	200000
Maximum body characters	400	400
Adapter	LoRA rank 4, $\alpha = 16$	LoRA rank 4, $\alpha = 16$
Target modules	`q_proj`, `v_proj`	same
LoRA dropout	0	0
Trainable parameters	458752	458752 (Qwen), 958464 (SmolLM3)
Generated samples/arm/seed	5000	1000

At $\varepsilon_{\text{learn}} = 4$, the transfer profiles contain 5846 Amazon generator-training users/7268 capped examples and 6938 Enron generator-training users/7771 capped examples. Their calibrated noise multipliers are respectively 0.93628 and 0.87280, yielding final accountant values 3.9935 and 3.9921. These counts are training-selection counts, not the full canonical dataset sizes.

Table 33: Post-processing RFT configuration and final-generation defaults. Historical defaults are not complete per-run decoding records. See the provenance qualification below.

Parameter	Value / scope
Initialization	corresponding DP-LoRA checkpoint
Verifier source	DP-synthetic bodies and requested tags
RFT steps	200
Sampled prompts/step	8 (with replacement)
Completions/prompt	8
Rollouts/step (total rollouts)	64 (12800)
Optimizer / learning rate	Adam / 10^{-4}
KL weight β	0.05 (headline)
Gradient clip	1.0
RFT sampling	temperature 1.0, top-p 0.95, 160 new tokens maximum
Log-probability microbatch	4 sequences
Reference policy	frozen initialization snapshot
Final-corpus token cap	120 new tokens
Historical text-generation default	temperature 0.9, top-p 0.95
Token-ID helper default (used by RFT)	temperature 1.0, top-p 0.95
Current final-corpus default (explicit)	temperature 0.9, top-p 0.95

Reward-model archival scope. The RFT checkpoint writer saves generator state and configuration, not the fitted reward vectorizer or classifier heads. Run reports preserve verifier F1 and available settings rather than a serialized reward model. This explains the operational absence of the original reward weights, but the historical reason for not retaining them is undocumented. Later frozen verifier reconstructions are separate fitted objects, not recovered historical snapshots (Appendix D.8.1).

Implemented RFT loss. The checked implementation uses $\eta = 10^{-6}$ and the sample standard deviation of the eight rewards in each group. It scores the exact sampled completion-token IDs, including EOS when generated, under the current policy and frozen initialization. For each unmasked token it forms $d = \log p_{\text{ref}} - \log p_\theta$ and the nonnegative k3-style penalty $\exp(d) - d - 1$. The policy-gradient and penalty sums are divided by the total number of unmasked completion tokens in the rollout batch. Microbatches accumulate gradients under this same denominator. Sampling uses the reported temperature and top-p, while log probabilities come from the model's full-vocabulary softmax. This sample-based penalty is not an analytic vocabulary sum or a measurement of the final controllers' full-distribution KL, and it does not establish the population optimization assumptions in Appendix C.1.

Decoding provenance. Historical final-generation sidecars do not record temperature or top-p. The text-returning `generate_batch` method defaults to temperature 0.9. The earlier audited `generate_corpus` path instead inherited the `generate_ids_batch` helper's 1.0 default without an override. Both helpers default to top-$p = 0.95$. Current final-corpus generation explicitly passes and records temperature 0.9 and top-$p = 0.95$ unless overridden, while the token-ID helper retains its 1.0 default for RFT rollouts. The new matched audit separately freezes and records its decoding configuration (Appendix D.8.1). This repair does not recover historical run settings: the timed historical final-generation rerun matches recorded settings and requested tags, but its sampling-temperature equivalence to the original release remains unestablished.

Reward-DP comparator. REWARD-DP-GROUND+CLUSTER uses the same fraction-of-requested-keys objective as OUTPUT-VERIFIER RFT. It freezes `all-MiniLM-L6-v2` embeddings, trains one linear classification head per key with user-level DP-SGD, and then RFTs a fresh public base model against that private reward. The reward model uses 200 DP steps, 256 expected users/lot, contribution cap 10, clip norm 1, Adam learning rate 0.05, and the same target learner epsilon/delta. The transfer screen deliberately retains this full-cap reward recipe rather than weakening the comparator to cap three. The reduced transfer budget in Table 26 therefore describes the generator training only. The learner-accountant settings remain matched, and the reward-DP comparator receives more per-user data than the transfer generators.

G.3.1 REWARD-QUALITY AUDIT

We audit the three $\varepsilon_{\text{learn}}=4$ objective-matched reward checkpoints without retraining them. Their capped training stream contains 79,077 records from 36,501 users and includes all eligible users, so we cannot construct a user-disjoint evaluation set from the remaining records. We instead sample 4000 records from the training window and 4000 valid records excluded by contribution capping from 2367 users represented in that window. This second set is a record-level, same-user generalization diagnostic, not a privacy holdout.

Table 34: Objective-matched DP recovery-reward audit, averaged over three checkpoints. ECE is ten-bin top-label expected calibration error. Synthetic statistics use 1000 matched rows per arm/seed, and r is row-level Pearson correlation between the private reward and the fixed automatic recovery instrument, and Q4–Q1 is the automatic-recovery difference between the top and bottom private-reward quartiles. The real-verifier row is an unaccounted diagnostic, not the output-derived-verifier arm.

Evaluation set	Mean key accuracy	Balanced accuracy	Macro-F1	ECE	Notes
Training-window sample	0.619	0.393	0.400	0.030	4000 records/checkpoint
Same-user capped-out records	0.603	0.380	0.381	0.036	4000 records/checkpoint

Generated arm	Private reward	Auto recovery	Pearson r	Q4–Q1
L4	0.504	0.440	0.372	0.141
REAL-VERIFIER RFT (diagnostic)	0.539	0.491	0.400	0.152
REWARD-DP-GROUND+CLUSTER	0.533	0.477	0.342	0.138

The small train-to-record-holdout drop, low ECE, positive row-level correlations, and positive quartile gaps in Table 34 show that the private reward is neither constant nor arbitrary. Its weakest dimensions are the high-cardinality intent and response keys, whose capped-out macro-F1 values are 0.112 and 0.178. The comparator should therefore be read as a meaningful but noisy objective-matched reward, not as an oracle reward.

G.4 VERIFIER AND MEASUREMENT INSTRUMENTS

The DP-output verifier and automatic measurement labeler share the same model class: a TF-IDF vectorizer with at most 5000 unigram/bigram features and one class-balanced logistic-regression head per key (maximum 300 iterations). For the output-derived RFT verifier, the fit corpus consists only of DP-LoRA generations and their DP-sampled requested tags. For the evaluation-only real-data instrument, the configured recipe requests 4000 labeled training bodies and 1000 held-out bodies, with a loader cap of 8000 eligible bodies rather than a limit on raw records scanned. The implemented vectorizer vocabulary is fixed once on the combined training and unlabeled held-out bodies. Each key's head is fitted only on training examples with eligible labels. This transductive vocabulary construction is common to every compared method but should be considered when interpreting the instrument F1 values reported in Table 35.

The historical REAL-VERIFIER RFT diagnostic uses this real-fitted recipe as a training reward, a distinct role from the evaluation-only measurement instrument. Its five archived RFT reports retain neither `verifier_source` nor `verifier_fit_metadata`. The configured $4000/1000$ split therefore does not establish the exact achieved historical training and holdout counts.

Shared-model lexical control. Because the verifier and automatic instrument share a model class, we directly test whether their agreement is carried by the same lexical cues. For each of six seeds, we refit the DP-output verifier, extract the top $k \in \{1, 3, 5, 10\}$ positive unigram coefficients for each requested key/value, mask every occurrence in each generated body, and rescore original and masked text with one fixed real-data measurement labeler. A same-body control masks nonsalient token occurrences greedily matched by verifier IDF. Short bodies occasionally lack enough alternatives, yielding 99.1–100% aggregate count matching. The original OUTPUT-VERIFIER RFT$-$DP-LoRA lift is $+0.0357$ (95% t interval $[+0.0200, +0.0515]$). After targeted masking, it remains $+0.0311$, $+0.0302$, $+0.0312$, and $+0.0293$ for $k = 1, 3, 5, 10$, respectively, and remains positive in all six seeds. The corresponding IDF-control lifts are $+0.0359$ to $+0.0383$. Thus verifier-weighted words

account for a reproducible minority of the measured advantage (roughly 13–18% relative to the original lift), but do not explain most of it. The exact two-sided sign-flip test is $p = 0.03125$ for every targeted-mask setting. Historical measurement fits were cached in memory. Evaluation saved held-out F1 and labeled outputs, not the fitted vectorizer and classifier weights. The reason no instrument snapshot was archived is undocumented. This audit therefore uses a refit rather than the historical fitted object. Its original scores reproduce the persisted labeled corpora within 0.0011 absolute recovery.

Table 35: Held-out macro-F1 of the automatic measurement instruments used throughout the paper, for both dataset profiles.

Profile	Per-key macro-F1							
Enron	triage .668	topic .464	intent .229	response .278	urgency .360	sentiment .554	difficulty .646	formality .679
Amazon	group .809	rating .433	intent .457	aspect .551	detail .746	formality .523	–	–

Requested-control recovery compares each generated body's measurement-labeler prediction with the skeleton supplied to generation and macro-averages accuracy over keys. For SmolLM3 seeds 0–1, the original evaluation JSON predates the addition of the explicit grounding field. We reconstruct the identical statistic by comparing the aligned generated and post-hoc-labeled JSONL files. Seed 2 and all Amazon reports contain the field directly.

TVD compares the body-inferred attribute distributions against DP one-way marginals and DP cluster occupancy, not against the raw full joint distribution or the requested tags. Fidelity also uses unbiased RBF MMD^2 with one real-only median-distance bandwidth per profile, and PRDC-style coverage. Embedding metrics compare all released synthetic bodies with a fixed 2000-body real reference using normalized `all-MiniLM-L6-v2` embeddings. A disjoint 1000-body real set fixes the bandwidth. Utility uses separate 5000-feature unigram/bigram TF-IDF vectorizers and class-balanced logistic regression for TRTR and TSTR. Real users are disjoint across train/test. Primary evaluation uses 2000 real or synthetic training examples and 1000 real test examples. Transfer uses 1000 real or synthetic training examples and 1000 real test examples. For every primary synthetic corpus, ten deterministic without-replacement subsamples are evaluated and averaged within the training seed before seed-level intervals are computed. Transfer uses all 1000 synthetic bodies, so its ten evaluations use one identical training set. They are not independent replicates. We report direct TSTR macro-F1. TRTR is shown only as a same-size real-data reference. Unless explicitly identified as requested-label results, TSTR tables use post-hoc synthetic labels. Appendix D.3 gives the paired label-source protocol and reference-file provenance.

For Enron, `needs_decision` is positive when the intent is request approval, request meeting, propose, or escalate, or when the expected response is reply with decision, accept/decline, or review/approve. `requires_response` is negative only for `no_action_fyi`. Triage is the original three-class control. Amazon utility predicts the native five-class star rating and, secondarily, the six-class canonical review-intent label. The utility reference's rule-derived intent labels are distinct from the semantic annotations used upstream.

The primary Qwen judge uses 500 matched bodies per method/seed and 300 held-out real bodies. The cross-family `GPT-5.6 Luna` judge uses a fixed 200-body matched synthetic subset per method/seed, accessed through the deployment identifier `gpt-5.6-luna`. The 300-body real calibration uses Qwen only, so Luna has no real-calibration scores in Table 2. Both synthetic evaluations are cluster-balanced and weighted back to the eligible matched-corpus mixture. For this primary evaluation, reported weighted invalid-response rates are below 1% in every Qwen cell and zero under Luna. Eligibility requires a string body with at least 20 characters after stripping surrounding whitespace in *every* compared arm at the same row index. The harness first validates canonical requested tags and requires exact tag alignment across arms. Malformed tags, misalignment, or an unmapped cluster abort preparation rather than silently exclude rows. The body-validity intersection has reported exclusions of 97, 13, and 10 of the 5000 aligned rows in Enron seeds 0–2. The remaining 4903, 4987, and 4990 rows define the mixture. The selection metadata records only the size of the union of excluded rows, not a per-arm or failure-type breakdown, so it does not establish why seed 0 loses more rows. The seed-0 reranking judge similarly uses 4977 eligible matched rows out of 5000. These input exclusions are separate from invalid judge responses. Malformed or out-of-vocabulary annotations and nonresponses, including content-filtered refusals, score zero on all keys

and remain in the primary recovery denominator. The Amazon transfer additionally uses a separate frozen 200-body matched subset per arm/seed judged by `GPT-5.6 Sol` through the OpenAI-v1 Responses API on 29 August 2026. The synthetic-only sample is balanced over the Amazon profile's 24 public clusters and weighted back to the eligible matched-corpus mixture: 984, 984, and 987 of 1000 rows in seeds 0–2, respectively. No real Amazon review is sent to the service.

G.5 Verifier-model assumptions

Propositions 6 and 8 describe population fits under a value-independent noise component. Their parameters p, $P(\cdot \mid y)$, and Q are latent and are not estimated here. Neither measured attribute recovery nor instrument macro-F1 estimates the noise model.

The traceable per-key measurements are the instrument scores in Table 35, the judge per-key values in Table 25, and the automatic per-key values in Table 38. Held-out macro-F1 of the real-fitted Enron labeler is weakest on intent (0.229), response (0.278), and urgency (0.360), which describes measurement quality on those keys rather than any property of the fitted verifier's posterior.

The propositions assume a well-specified model class and an unregularized population optimum. The balanced result also requires exact inverse-prior weights. The executed verifier is a regularized finite-sample TF-IDF logistic model with empirical class weights. No quantitative posterior-approximation or argmax-disagreement bound relative to the faithful balanced posterior has been established for that finite fit. Appendix D.8.2 compares frozen reconstructed verifier predictions with independent judges on newly generated bodies. On unselected L4, net requested-label error correction is -0.0110 under Luna (95% CI $[-0.0356, +0.0129]$) and -0.0390 under Qwen ($[-0.0590, -0.0187]$, Table 18). These judge-agreement differences neither estimate distance from the population limit nor test argmax invariance against the unobserved faithful balanced posterior. They do not validate the idealized noise model.

G.6 Frozen-base quality control

The frozen-base anchor uses the common DP structural sampler, requested prompts, and token cap but loads `Qwen3.5-4B` without a LoRA or virtual-prompt adapter. Historical temperature and top-p are not fully archived (Appendix G.3). The anchor is therefore a quality reference rather than a decoding-matched control. This uncertainty matters most for diversity and PII-like span counts, which are particularly sensitive to sampling temperature. All six reports contain 5000 bodies, record `base_only=true` and a null checkpoint path, and reproduce their recovery scores from the aligned requested and post-hoc-measured skeletons. Seeds 0–5 have identical requested-skeleton sequences across frozen Qwen, DP-LoRA, and Output-Verifier RFT. Table 36 reports all six seeds, and Table 37 the matched paired differences over seeds 0–5.

Seed	Recovery	mean-TVD	joint-TVD	MMD^2	Cov	TSTR F1	PII spans	Unique frac.
0	0.4272	0.1520	0.2999	0.06897	0.357	0.6263	1382	0.9998
1	0.4300	0.1551	0.2939	0.06906	0.361	0.6260	1300	1.0000
2	0.4290	0.1472	0.2920	0.06930	0.350	0.6112	1438	0.9996
3	0.4296	0.1485	0.2769	0.06883	0.354	0.6191	1452	0.9996
4	0.4304	0.1493	0.2853	0.06898	0.360	0.6238	1327	0.9990
5	0.4296	0.1467	0.2887	0.06795	0.372	0.6088	1466	0.9998
Mean $\pm$ sd	0.4293 ± 0.0011	0.1498 ± 0.0032	0.2895 ± 0.0079	0.06885 ± 0.00047	0.359 ± 0.008	0.6192 ± 0.0076	1394 ± 69	0.99963 ± 0.00034

Table 36: Six-seed frozen-base quality anchor. TSTR F1 is matched-size, sender-disjoint needs-decision utility. PII spans are regex detections and are not a formal privacy metric.

G.7 External-validity results by seed

Tables 39 and 41 give the per-seed Amazon/Qwen and Enron/SmolLM3 values, and Table 40 the independent GPT-5.6 Sol judgments.

Full-macro hierarchical paired differences from Table 40 are Output-Verifier RFT$-$DP-LoRA $= +0.0969$ (95% CI $[+0.0769, +0.1170]$) and Output-Verifier RFT$-$Reward-DP-Ground+Cluster $= +0.0758$ ($[+0.0551, +0.0972]$), with positive seed-level differences in both comparisons. The Reward-DP-Ground+Cluster$-$DP-LoRA difference is $+0.0211$ ($[-0.0100, +0.0500]$), so the independent judge does not resolve an ordering between those two baselines.

Metric (arm minus frozen base)	DP-LoRA delta [paired 95% t-CI]	OUTPUT-VERIFIER RFT delta [paired 95% t-CI]
Recovery ↑	$+0.0070$ [$+0.0015$, $+0.0126$]	$+0.0434$ [$+0.0235$, $+0.0632$]
mean-TVD ↓	-0.0304 [-0.0410, -0.0199]	$+0.0049$ [-0.0117, $+0.0215$]
joint-TVD ↓	-0.1550 [-0.1736, -0.1364]	-0.1073 [-0.1486, -0.0661]
MMD^2 ↓	-0.0549 [-0.0576, -0.0522]	-0.0377 [-0.0492, -0.0261]
Coverage ↑	$+0.4843$ [$+0.4726$, $+0.4959$]	$+0.4448$ [$+0.4269$, $+0.4628$]
Needs-decision TSTR F1 ↑	$+0.0038$ [-0.0025, $+0.0102$]	$+0.0036$ [-0.0012, $+0.0084$]
PII spans ↓	$+8418$ [$+6843$, $+9992$]	$+8810$ [$+5372$, $+12247$]

Table 37: Both adapted arms gain recovery, cluster fidelity, and coverage over the frozen base, and both emit far more PII-like spans, while utility stays unresolved. Matched frozen-base comparisons over seeds 0–5. Negative deltas favor the adapted arm for TVD/MMD. Positive deltas favor it for recovery, coverage, and utility but disfavor it for PII-like spans.

Arm	triage	topic	intent	response	urgency	sentiment	difficulty	formality
Frozen Qwen	0.474	0.305	0.167	**0.196**	**0.377**	0.593	**0.892**	0.432
DP-LoRA	0.484	0.301	0.162	0.164	0.334	0.633	0.837	0.574
OUTPUT-VERIFIER RFT	**0.491**	**0.449**	**0.196**	0.166	0.330	**0.673**	0.830	**0.647**

Table 38: Matched-seed automatic recovery by requested key. DP adaptation mainly improves formality and sentiment. Post-processing RFT adds the largest gains on topic and formality. The frozen base remains strongest on response, urgency, and difficulty.

G.8 EXPERIMENT INVENTORY AND COMPUTE

Table 42 summarizes seed counts, released rows, and the additional evaluation attached to each experiment.

The primary 2000-step Qwen generator cells required approximately 94–160 GPU-hours each on the server configuration. The two Amazon DP-training logs that retained elapsed time report 12.73–12.75 GPU-hours. The seed-2 checkpoint does not retain that wall time. Amazon OUTPUT-VERIFIER RFT RFT required 1.29–2.34 hours/seed and reward-DP policy RFT required 1.37–2.69 hours/seed. SmolLM3 DP training required 1.39–2.25 hours/seed. Its two RFT arms required 0.46–1.01 hours/seed. The full Amazon semantic annotation pass required 12.97 wall-clock hours on a separately served Qwen3.8-27B endpoint. The normalized seed-0 verifier-guidance accounting is in Table 23: reranking used 22.59 measured GPU-hours for 40,000 candidates, while the RFT optimization used 2.816 hours for 12,800 rollouts plus both log-probability passes and backward optimization. A recorded-settings A100 rerun directly measured final 5000-body RFT generation at 0.774 GPU-hours, for a 3.590-hour optimization-plus-release stage sum. The shared verifier-corpus generation and historical RFT rollout-token count remain explicitly unresolved rather than silently imputed. The separate three-seed matched inference audit performs no new training: Table 17 reports its exact completion/token counts and per-stage generation/setup times on A100 80GB GPUs. Its two judges each assess 1800 synthetic bodies. These measurements are not substituted into historical training-amortization estimates. The verifier reliability diagnostic reuses those 1800 bodies and both sets of annotations, scoring three frozen verifiers and running 10,000 paired bootstrap replicates on CPU. It adds no GPU training, generation, or judging. The Amazon independent judge completed 1800 structured GPT-5.6 Sol responses with zero invalid outputs, using 908668 prompt and 171478 completion tokens. GPU stages ran on dedicated 80-GiB accelerators unless noted. TF-IDF evaluation and aggregation ran on CPU.

G.9 REPRODUCTION COMMANDS AND ARTIFACT PROVENANCE

The following commands show the executed orchestration pattern. Environment variables isolate dataset profiles and artifact roots. GPU identifiers are illustrative.

```
# Amazon acquisition and canonicalization
python -m src.data_foundation.download_amazon --limit 250000
python -m src.data_foundation.prepare_amazon \
 --input All_Beauty=data/raw/amazon/All_Beauty.jsonl \
 --input Amazon_Fashion=data/raw/amazon/Amazon_Fashion.jsonl \
```

Seed	Arm	Recovery	mean-TVD	joint-TVD	MMD^2	Cov	rating F1	intent F1	PII spans
0	DP-LoRA	0.571	0.119	0.175	0.0202	0.690	0.362	0.192	11
0	Output-Verifier RFT	0.688	0.104	0.248	0.0274	0.650	0.370	0.152	5
0	Reward-DP-Ground+Cluster	0.516	0.169	0.305	0.0549	0.409	0.354	0.180	14
1	DP-LoRA	0.578	0.100	0.146	0.0109	0.728	0.340	0.135	11
1	Output-Verifier RFT	0.679	0.109	0.248	0.0321	0.681	0.337	0.131	3
1	Reward-DP-Ground+Cluster	0.544	0.149	0.321	0.0479	0.417	0.346	0.154	15
2	DP-LoRA	0.578	0.105	0.152	0.0131	0.721	0.363	0.149	14
2	Output-Verifier RFT	0.703	0.101	0.253	0.0296	0.685	0.388	0.158	11
2	Reward-DP-Ground+Cluster	0.509	0.170	0.313	0.0611	0.396	0.374	0.153	35

Table 39: Amazon/Qwen transfer results by seed. Utility uses current matched-size TSTR scores, not legacy per-run utility. PII spans are regex detections per 1000 outputs and are diagnostic.

Seed	Arm	Full macro	Reliable-key macro	Noisy-key macro	Invalid
0	DP-LoRA	0.679	0.739	0.559	0/200
0	Output-Verifier RFT	0.765	0.821	0.654	0/200
0	Reward-DP-Ground+Cluster	0.678	0.710	0.615	0/200
1	DP-LoRA	0.669	0.744	0.518	0/200
1	Output-Verifier RFT	0.766	0.838	0.624	0/200
1	Reward-DP-Ground+Cluster	0.696	0.761	0.565	0/200
2	DP-LoRA	0.666	0.723	0.552	0/200
2	Output-Verifier RFT	0.773	0.828	0.661	0/200
2	Reward-DP-Ground+Cluster	0.703	0.751	0.606	0/200
Mean	DP-LoRA	0.671	0.735	0.543	0/600
Mean	Output-Verifier RFT	**0.768**	**0.829**	**0.646**	0/600
Mean	Reward-DP-Ground+Cluster	0.692	0.741	0.595	0/600

Table 40: Independent GPT-5.6 Sol recovery on the frozen Amazon subset. Each cell contains 200 matched synthetic reviews, sampled by public cluster and weighted back to the eligible matched-corpus mixture. Reliable keys are product group, dominant aspect, detail level, and formality. Noisy keys are star rating and review intent.

```
 --input Appliances=data/raw/amazon/Appliances.jsonl \
 --input Cell_Phones_and_Accessories=\
data/raw/amazon/Cell_Phones_and_Accessories.jsonl \
 --input Grocery_and_Gourmet_Food=\
data/raw/amazon/Grocery_and_Gourmet_Food.jsonl \
 --input Books=data/raw/amazon/Books.jsonl \
 --max-scan-per-input 250000 --max-records 200000 \
 --min-user-reviews 2 --out data/raw/amazon_reviews_gt.jsonl
python -m src.data_foundation.label_amazon_semantics \
 --input data/raw/amazon_reviews_gt.jsonl \
 --out data/raw/amazon_reviews_gt_llm.jsonl \
 --base-url http://127.0.0.1:8013/v1 \
 --model Qwen/Qwen3.8-27B --batch-size 12 --max-workers 8 \
 --resume

# Primary full-pool cell
TIER=B BASE_VARIANT=ground VERIFIER_SOURCE=dp_synth \
 bash scripts/pipeline_cell.sh 4 0 0

# Matched frozen-base quality anchor
bash scripts/run_frozen_base_controls.sh --gpu 1
python scripts/build_frozen_base_summary.py

# Amazon/Qwen transfer cell
PPSD_DATASET_PROFILE=amazon_reviews \
PPSD_GT_FILE="$PWD/data/raw/amazon_reviews_gt_llm.jsonl" \
PPSD_ARTIFACTS_DIR="$PWD/artifacts/amazon_reviews_tiera" \
BASE="Qwen/Qwen3.5-4B" \
TIER=A RFT_STEPS=200 N_GEN=1000 RUN_BASE=1 BASE_VARIANT=ground \
 bash scripts/pipeline_cell.sh 4 0 0
```

Seed	Arm	Recovery	mean-TVD	joint-TVD	MMD^2	Cov	decision F1	triage F1	PII spans
0	DP-LoRA	0.421	0.121	0.196	0.0158	0.532	0.623	0.467	1373
0	OUTPUT-VERIFIER RFT	0.446	0.186	0.273	0.0315	0.400	0.598	0.537	745
0	REWARD-DP-GROUND+CLUSTER	0.445	0.230	0.299	0.0826	0.177	0.619	0.438	94
1	DP-LoRA	0.425	0.136	0.234	0.0154	0.540	0.595	0.462	1115
1	OUTPUT-VERIFIER RFT	0.468	0.139	0.205	0.0250	0.447	0.556	0.465	2212
1	REWARD-DP-GROUND+CLUSTER	0.462	0.205	0.238	0.0723	0.168	0.590	0.429	149
2	DP-LoRA	0.420	0.143	0.216	0.0144	0.554	0.589	0.499	1095
2	OUTPUT-VERIFIER RFT	0.445	0.167	0.201	0.0234	0.448	0.579	0.484	585
2	REWARD-DP-GROUND+CLUSTER	0.441	0.225	0.246	0.0670	0.192	0.608	0.404	113

Table 41: Enron/SmolLM3 transfer results by seed. Utility uses current matched-size TSTR scores. The large variation in PII-like strings, particularly for OUTPUT-VERIFIER RFT seed 1, motivates the output-quality caveat in Appendix E.

Table 42: Evidence inventory. "Rows" is the number of released bodies per arm/seed unless otherwise stated.

Experiment	Seeds	Rows	Additional evaluation
Primary DP-LoRA/OUTPUT-VERIFIER RFT, $\varepsilon_{\text{learn}} = 4$	6	5000	Qwen and GPT judges on 3 seeds, utility, privacy audits
Endpoint placement/RFT, $\varepsilon_{\text{learn}} = 2, 8$	3 each	5000	automatic metrics and hierarchical bootstrap
REWARD-DP-GROUND+CLUSTER, $\varepsilon_{\text{learn}} = 4$	3	5000	objective-matched placement comparison
KL-weight audit	3	5000	$\beta \in \{0.05, 0.10, 0.20\}$
Historical verifier reranking	1	5000	nested $N \in \{1, 2, 4, 8\}$, GPT subset
Matched inference / reranking	3	1000	Fixed decoding, Qwen3.8 and Luna (200 bodies/arm/seed), exact token accounting
Verifier reliability diagnostic	3	200 judged	CPU-only, frozen verifiers and both existing judges, no new generation
Frozen Qwen quality anchor	6	5000	automatic quality, 6 matched comparisons
Amazon/Qwen transfer	3	1000	automatic metrics, GPT-5.6 Sol judge
Enron/SmolLM3 transfer	3	1000	automatic metrics
Requested-label utility audit	6 primary 3/transfer	5000 / 1000	Existing corpora, paired label sources across seven setting–task comparisons
Membership inference	2 clean DP seeds	256+256 users	two non-private positive controls

```
# Frozen Amazon independent judge (DefaultAzureCredential)
powershell -File scripts/run_amazon_gpt56_judge.ps1

# Enron/SmolLM3 transfer cell
PPSD_ARTIFACTS_DIR="$PWD/artifacts/smollm3_enron_tiera" \
BASE="HuggingFaceTB/SmolLM3-3B" \
TIER=A RFT_STEPS=200 N_GEN=1000 RUN_BASE=1 BASE_VARIANT=ground \
 bash scripts/pipeline_cell.sh 4 0 0
```

The current runner is resumable: a stage is skipped only when its final nonempty output exists. Periodic files named `*.partial.pt` or `*_stepN.pt` are diagnostic milestones and are never treated as final headline checkpoints. New final checkpoints store the base model, adapter type, LoRA rank/alpha/targets, dataset profile, and schema hash. DP generator checkpoints also store the seed and learner epsilon. RFT reports store their seed, upstream epsilon, and initialization path. Current generated-corpus sidecars store seed, sampler, checkpoint, base-only status, sample count, token limit, temperature, top-p, profile, and schema hash. RFT reports additionally store verifier source, verifier privacy status, reward keys, first/final reward, and optimization settings. These are provenance fields, not privacy certificates. Privacy-sampling or noise seeds must not be conflated with public generation seeds, and historical reports and diagnostic milestones require the output-specific clearance described in Appendix G.2 before any confidential-data release.

Python 3.9 or newer is required. The principal dependencies are NumPy, SciPy, scikit-learn, sentence-transformers, Transformers, Opacus, Accelerate, Matplotlib, the OpenAI client, and Azure Identity. A CUDA-enabled PyTorch wheel is installed separately before the pinned project requirements so dependency resolution does not replace it with a CPU wheel.

The repository test suite covers profile isolation, deterministic contribution capping, Amazon streaming/preparation/splits, schema-constrained semantic parsing, utility splits, grounding computation, privacy audits, reranking recovery, and beta aggregation. Experiment launch scripts are syntax-checked before server use. Table values are rebuilt from JSON reports rather than transcribed from logs.

Historical archive coverage. The locally audited run reports corroborate the configured generator recipe for primary $\varepsilon_{\text{learn}} = 4$ seeds 0–2 and 5, full-pool epsilon-two seed 2, Amazon seeds 0–1, and all three SmolLM3 seeds. Complete generator-configuration archives were not located for primary epsilon-four seeds 3–4, full-pool epsilon-two seeds 0–1, full-pool epsilon-eight seeds 0–2, or Amazon seed 2. Their generation and RFT artifacts exist but do not independently establish every upstream training setting. Older 3000-record endpoint logs are not evidence for the full-pool recipe. Table 32 therefore records the configured recipe where per-run archives are incomplete. Recovering these archives is required for complete provenance.

Known metadata caveat. The full Amazon semantic-labeling report was created before the product-group profile was narrowed from ten theoretical groups to the six downloaded groups, so that report carries the earlier schema hash. The semantic fields were unchanged. The final canonical records, structural artifacts, checkpoints, and evaluation reports all record and validate against the final Amazon schema-version fingerprint, `09288c5ce2987e20`, preventing stale or cross-profile artifacts from being loaded. The Amazon table includes all three predeclared transfer seeds.

G.10 Planned artifact release

The planned anonymous supplementary package would contain source code, dependency files, dataset-profile definitions, preprocessing/download scripts, prompt templates, unit tests, experiment runners, specifically cleared aggregate reports, and table-building scripts. It will not contain credentials, service endpoints, raw sender/user identifiers, or redistributed copies of Enron/Amazon source text. Public-data scripts record canonical URLs and content hashes. Because generated text can still contain plausible PII-like strings, generated corpora and trained checkpoints will be considered for release only after license review, an output-safety audit, and artifact-specific privacy or permission clearance, with explicit disclosure of the historical selection and implementation qualifications. Such review is not itself a DP certificate. Aggregate metrics, checkpoints, logs, and provenance are not automatically DP because they omit raw text or contain DP-labeled fields. Whole diagnostic reports are not a preapproved public release set. A confidential-data release would require an explicit public-output allowlist, internal-only diagnostics, protected privacy randomness, and accounting for all released mechanisms and private-data-based selection or evaluation, with shared ancestors handled once. These are future operational requirements, not implemented repairs or a new end-to-end certification of the historical artifacts.