PRIVACY NOISE MASKS THEN REVEALS ORDER AND PARTITION EFFECTS IN SAMPLE-AND-AGGREGATE ICL

Anonymous authors
Paper under double-blind review

ABSTRACT

In sample-and-aggregate private in-context learning (ICL) for classification, privacy noise changes not only the amount but also the *source* of released-label instability. Using a full-domain Gumbel/exponential count channel, we instantiate a law-of-total-variance identity, valid for any count-based categorical release kernel, that separates released Gini exactly into release randomness, within-partition order variation, and partition variation. Under tight Gumbel privacy, a near-uniform release *masks* order and partition structure. In three sampled designs, grid-interpolated crossovers lie in $\varepsilon \in [3.2, 4.2]$; small measured order/partition components near the commonly evaluated $\varepsilon = 1$ need not imply weak upstream structure. Gaussian report-noisy-max on the same counts shows the same qualitative masking, without implying cross-mechanism ε equivalence. Nested and matched designs attribute most aggregate upstream variation to ordering and reveal task-dependent partition concentration. Signed multiclass margins prospectively detect synchronized deterministic plurality-order activity on Text REtrieval Conference (TREC) question classification, including a weaker 14B replication, but do not universally rank vote strength or severity. A prospectively frozen sizing forecast predicts an unseen $S = 14$ Banking77 configuration within .007 on accuracy and .002 on fidelity. Holding 72 private records fixed, 24 three-example voters preserve similar pre-release plurality accuracy yet achieve substantially higher released utility than eight nine-example voters, exposing a condition-specific transmission window. Low measured order components under tight privacy can therefore be channel artifacts. Internal, privacy-accounted audits of pre-noise votes—or analyses on public development data—should accompany released-label stability.

1 INTRODUCTION

In-context learning (ICL) predictions can change when labeled demonstrations are reordered (Lu et al., 2022; Zhao et al., 2021). In private ICL, randomized release can hide this upstream order sensitivity.

Treating privacy randomization as an independent increment on fixed upstream instability is a tempting shortcut. In sample-and-aggregate classification ICL it fails: tight privacy raises released-label randomness while a near-uniform channel *masks* the contributions of order and partitioning. As privacy relaxes, upstream structure *re-emerges* past a configuration-dependent crossover. Privacy changes both the amount and the *source* of variation.

We study the sample-and-aggregate private-ICL pipeline of Wu et al. (2024): disjoint subsets of a private exemplar pool vote through a categorical count-release channel. We use full-domain Gumbel report-noisy-max (RNM), an exponential-mechanism variant with a closed-form distribution. Figure 1 shows the three randomization levels; Figure 3 shows masking and crossover.

Reordering fixed subsets is not repartitioning the pool. Our nested design varies within-subset order at fixed membership, then resamples partitions.

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

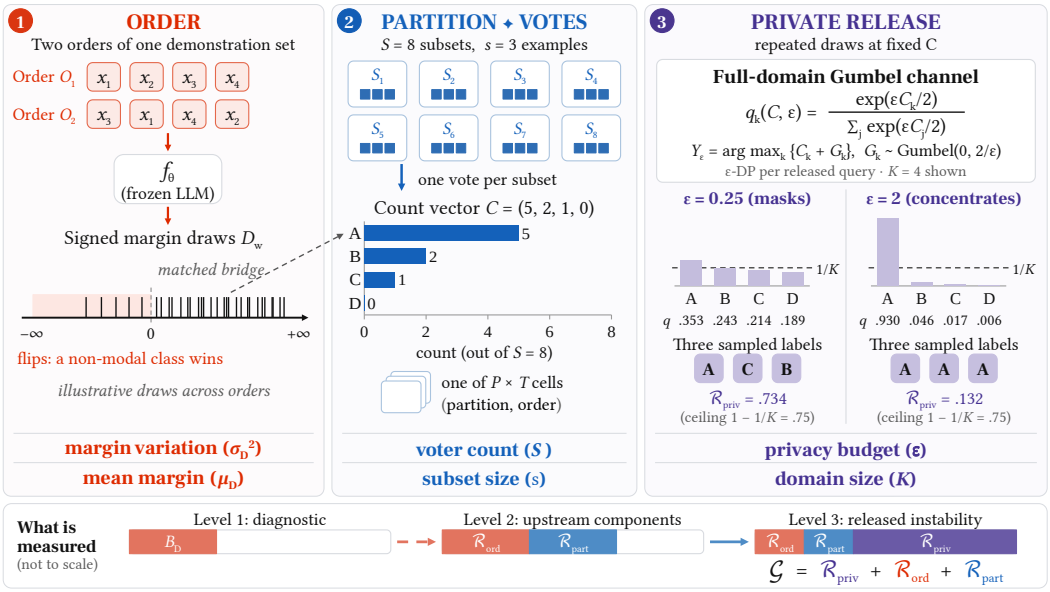

Figure 1: **Three randomization levels in sample-and-aggregate ICL.** Full-context orders yield a signed-margin diagnostic; partitioned subset predictions form $C = (5, 2, 1, 0)$ in the illustrated $K = 4$ example; and the full-domain Gumbel channel maps the same C to an ε-dependent released distribution. The dashed bridge is empirical, not an identity: exact $\mathcal{R}_{\text{ord}}/\mathcal{R}_{\text{part}}$ attribution starts from Level-2 vote distributions. Sampled labels and the bottom rail are illustrative; displayed probabilities are computed from C and rounded independently.

Contributions (by role).

1. **Primary finding.** Across three audited fixed-domain conditions, tight Gumbel privacy masks upstream order and partition effects, which re-emerge after a configuration-dependent crossover. Gaussian report-noisy-max on the same counts shows qualitative masking only; we make no cross-mechanism ε equivalence claim.
2. **Explanatory tool.** An exact law-of-total-variance identity attributes released instability to privacy, order, and partition for any categorical kernel whose private-data dependence is through counts. This is attribution, not a new privacy guarantee.
3. **Structural refinement.** Post-hoc exact 6^8 joint-order reconstruction retains aggregate order-share intervals above one half in all five matched conditions; finite-design estimates exceed one half in all three nested conditions. Partition effects are more right-tail concentrated on AGNews; complementary shares are not independent evidence.
4. **Upstream diagnostic.** Across twelve collision-free clean conditions, signed multiclass margins track order sensitivity and, under frozen TREC protocols, detect synchronized deterministic order activity at 7B with a weaker 14B replication. Vote-strength and active-only severity rankings are not universal.
5. **Scoped design validation.** A frozen Qwen/Banking77 $S = 8, 14, 24$ forecast passes all eight directional checks and both $S = 24$ point targets; $S = 14$ accuracy/fidelity errors are $.007/.002$. A separate fixed-72-record PATE analogue improves released utility with a near-zero pre-release plurality gap. A further two-model, two-task $n = 24$ panel passes its frozen TREC rule at $\varepsilon = 1$; post-hoc replay reverses accuracy gains by $\varepsilon = 8$ (Section 6). These are scoped interventions, not a scaling or optimality law.

Scope of the clean–private bridge. On matched queries and pools, signed margins screen synchronized deterministic plurality activity. This frozen bridge does not establish finite-budget or independent-order detection, nor severity ranking.

2 Related Work

ICL mechanisms and order. ICL has Bayesian, circuit, and forward-pass optimization accounts (Brown et al., 2020; Dong et al., 2024; Xie et al., 2022; Wang et al., 2023; Olsson et al., 2022; Dai et al., 2023; Von Oswald et al., 2023), with scale changing what demonstrations teach (Wei et al., 2023). Performance depends on selection, mappings, formatting, and label semantics (Min et al., 2022; Yoo et al., 2022; Liu et al., 2022; Perez et al., 2021); order induces model-specific shifts (Lu et al., 2022; Sclar et al., 2024) and positional asymmetries (Xiang et al., 2024; Guo et al., 2024; Cobbina & Zhou, 2025). Unlike training-time resilience (Chen et al., 2025) or general prompt bounds (Liu & Chu, 2026), we isolate the distribution of order-induced multiclass boundary crossings.

Private in-context learning. Sample-and-aggregate originates with Nissim et al. (2007); DP-ICL instantiates it with private-subset responses and Gaussian report-noisy-max (Wu et al., 2024). Alternatives generate private demonstrations, privatize representations, or optimize prompts (Tang et al., 2024; Duan et al., 2023; Hong et al., 2024), with extensions using adaptive noise, token amplification, or public examples (Gao et al., 2025; Yamasaki et al., 2025; Joo et al., 2025). We analyze the inference-time count channel and separate release, order, and partition randomness.

Private ensembles, release, and decomposition. Disjoint subsets connect to PATE and private prediction (Papernot et al., 2017; 2018; Dwork & Feldman, 2018), but our "teachers" are one frozen model under different contexts. The exponential mechanism (McSherry & Talwar, 2007) and report-noisy-max representations (Dwork et al., 2006; Ding et al., 2021) give our categorical channel; permute-and-flip and smooth sensitivity provide adjacent release tools (McKenna & Sheldon, 2020; Nissim et al., 2007). Unlike private training (Abadi et al., 2016; McMahan et al., 2018; Li et al., 2022; Yu et al., 2022), each released query composes under pool reuse (Dwork & Roth, 2014). Prior decompositions address synthetic-data error or aleatoric/epistemic uncertainty (Räisä & Honkela, 2025; Depeweg et al., 2018); ours attributes the pipeline stage injecting randomness. This query-level view matters because DP effects can be heterogeneous (Bagdasaryan et al., 2019).

3 A Hierarchical Theory of Prediction Instability

Preview. For a fixed query, the pipeline randomizes demonstration order, pool partition, and private release (Figure 1). Any categorical count-release kernel admits the exact attribution $\mathcal{G} = \mathcal{R}_{\text{priv}} + \mathcal{R}_{\text{ord}} + \mathcal{R}_{\text{part}}$ (Proposition 2); a count-independent tight limit masks the upstream terms. We proceed order $\rightarrow$ partition/votes $\rightarrow$ release; upstream terms are counterfactual sensitivities when deployment fixes those choices.

Categorical instability. For a fixed query x and execution model, let $Y \in [K]$ be one output and $p_k(x) = \mathbb{P}(Y = k \mid x)$. Independent draws resample O at Level 1, or (Π, O) and release noise in the private pipeline. Categorical Gini is

$$\mathcal{G}(x) = 1 - \sum_{k=1}^{K} p_k(x)^2 = \mathbb{P}(Y \neq Y' \mid x), \tag{1}$$

where Y' is an independent execution. If $M(x) = 1 - \max_k p_k(x)$ is modal disagreement, then $M(x) \leq \mathcal{G}(x) \leq 2M(x)$ (proofs are in Section A).

3.1 Order: a multiclass signed-margin gate

Let O order a fixed demonstration set and let $z_k(O, x)$ be the declared verbalizer-sequence log-probability score. Fix w as the modal full-context prediction over estimation orders and define

$$D_w(O, x) = z_w(O, x) - \max_{j \neq w} z_j(O, x). \tag{2}$$

$D_w < 0$ implies a flip; ties follow the tie rule. Let $\mu_D = \mathbb{E}[D_w]$, $\sigma_D^2 = \text{Var}(D_w)$, $B_D = \sigma_D^2/(\sigma_D^2 + \mu_D^2)$ (the Cantelli score), and let $\mathcal{G}_{\text{L1}}$ be Equation (1) under random O.

Proposition 1 (Multiclass margin bound)**.** *If $\mu_D > 0$, then*

$$\mathbb{P}(\widehat{Y}(O, x) \neq w) \leq B_D, \qquad \mathcal{G}_{\text{L1}}(x) \leq 2B_D. \tag{3}$$

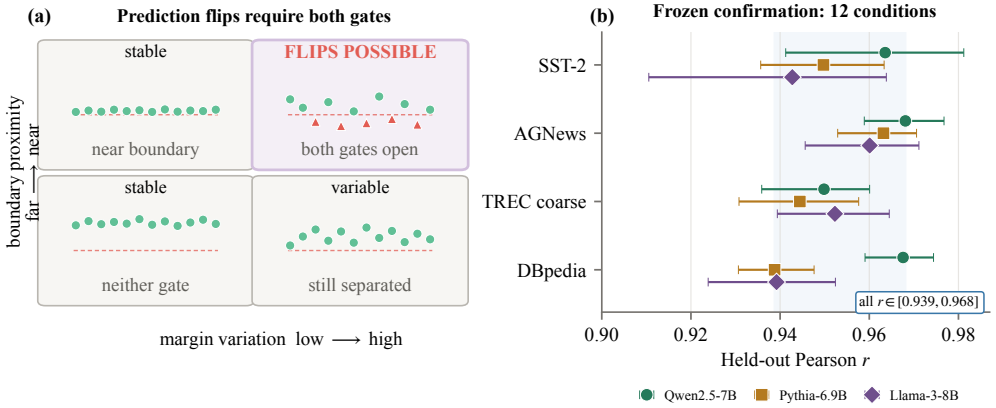

Figure 2: **Clean margins formalize order activity and track Gini order sensitivity. (a)** Flips require both margin variation and boundary proximity. **(b)** Frozen correlations across 12 model–task conditions; full values are in Table 5. Figure 7 instead shows the separate retrospective four-model AGNews comparison.

Thus ordering must vary the signed margin and bring it close enough to zero: B_D combines the variation and boundary gates. Correct-label loss variance is only a proxy because competing class scores also vary. For $K > 2$, accuracy alone does not determine the categorical prediction distribution, while the signed margin still measures distance to the strongest competitor (Figure 2). The proof is in Section A.

3.2 Partition: Subset Votes and the Margin Bridge

Level 1 permutes a full n-example context ($n = 24$ in matched bridges) and yields D_w. Level 2 draws a data-independent partition Π into S disjoint subsets of size $s = n/S$, orders each, and sums their predictions into $C(\Pi, O, x) \in \mathbb{N}^K$, $\sum_k C_k = S$, with top-two gap Δ. Reordering varies C at fixed membership; resampling Π varies membership. Data-dependent partitioning requires a new privacy analysis.

Subset composition precludes a deterministic map $D_w \mapsto \Delta$. Matched endpoints test strength Δ/S, detection of deterministic plurality activity $\mathcal{R}_{\text{ord}}^{\text{det,sync}} > 0$ by B_D, and active-only severity. "Sync" denotes six shared-index schedules, not independent-order or finite-ε activity. AGNews split detection/severity post hoc; TREC froze it.

3.3 Private Release: Exact Count Channel

The Gumbel/exponential mechanism releases the Section 3.2 count vector over the complete public class domain with exact distribution

$$q_k(C, \varepsilon) = \mathbb{P}(Y_\varepsilon = k \mid C) = \frac{\exp(\varepsilon C_k/2)}{\sum_{j=1}^{K} \exp(\varepsilon C_j/2)}. \tag{4}$$

Equivalently, it reports $\arg\max_k\{C_k + G_k\}$ for independent $G_k \sim \text{Gumbel}(0, 2/\varepsilon)$. Under replacement adjacency this is ε-DP per released query (Section A); repeated queries compose. Count vectors and margins are private intermediate statistics, not free deployment outputs. For tie-broken plurality label $k^\star(C)$, *plurality fidelity* is $q_{k^\star(C)}(C, \varepsilon)$. Condition-level *exposure* is

$$\text{Exposure}(\varepsilon) = \frac{\bar{\mathcal{R}}_{\text{ord}}(\varepsilon) + \bar{\mathcal{R}}_{\text{part}}(\varepsilon)}{\bar{\mathcal{R}}_{\text{ord}}^{\text{det}} + \bar{\mathcal{R}}_{\text{part}}^{\text{det}}}.$$

Bars average cells and queries; the denominator must be positive. Ties make this deterministic-reference ratio potentially exceed one.

3.4 EXACT PRIVACY–ORDER–PARTITION DECOMPOSITION

Let $Q_\theta : \mathbb{N}^K \to \Delta^{K-1}$ be any categorical release kernel with channel parameter θ ($\theta = \varepsilon$ for Gumbel), $q_{\Pi,O} = Q_\theta(C(\Pi, O, x))$, $\bar{q}_\Pi = \mathbb{E}_O[q_{\Pi,O} \mid \Pi]$, and $p = \mathbb{E}_\Pi[\bar{q}_\Pi]$.

Proposition 2 (Hierarchical instability decomposition)**.** *For every query and categorical release kernel,*

$$\mathcal{G}(x, \theta) = \mathcal{R}_{\text{priv}}(x, \theta) + \mathcal{R}_{\text{ord}}(x, \theta) + \mathcal{R}_{\text{part}}(x, \theta). \tag{5}$$

Here, exactly, $\mathcal{R}_{\text{priv}} = \mathbb{E}_{\Pi,O}[1 - \|q_{\Pi,O}\|_2^2]$, $\mathcal{R}_{\text{ord}} = \mathbb{E}_\Pi \mathbb{E}_{O|\Pi}[\|q_{\Pi,O} - \bar{q}_\Pi\|_2^2]$, *and* $\mathcal{R}_{\text{part}} = \mathbb{E}_\Pi[\|\bar{q}_\Pi - p\|_2^2]$.

The identity needs Q_θ to depend on private data through C and all randomized variables to be independently redrawn across executions. Fixed Π gives a Π-specific conditional Gini; its average is $\mathcal{R}_{\text{priv}} + \mathcal{R}_{\text{ord}}$, so fixing Π removes $\mathcal{R}_{\text{part}}$, and fixing schedules leaves $\mathcal{R}_{\text{priv}}$. Vector total variance attributes disagreement to release, order, or partition; the terms are channel-coupled, not additive on a clean baseline (Equation (6) and Figure 3). The identity covers Gumbel and Gaussian channels and can recurse, but experiments validate only these two upstream factors.

For the Gumbel channel, as $\varepsilon \to 0$,

$$\mathcal{R}_{\text{priv}} \to 1 - 1/K, \qquad \mathcal{R}_{\text{ord}} \to 0, \qquad \mathcal{R}_{\text{part}} \to 0, \tag{6}$$

so tight privacy masks upstream variation. Writing $\bar{\mathcal{R}}_{\text{priv}}(\varepsilon) = \mathbb{E}_x[\mathcal{R}_{\text{priv}}(x, \varepsilon)]$ for the query mean over a condition, and likewise for $\bar{\mathcal{R}}_{\text{ord}}, \bar{\mathcal{R}}_{\text{part}}$, we report the condition-level first crossover

$$\varepsilon^* = \inf\{\varepsilon > 0 : \bar{\mathcal{R}}_{\text{priv}}(\varepsilon) \le \bar{\mathcal{R}}_{\text{ord}}(\varepsilon) + \bar{\mathcal{R}}_{\text{part}}(\varepsilon)\}, \tag{7}$$

when the set is nonempty, without assuming global monotonicity or uniqueness. Empirically ε^* is the first sign change of $\bar{\mathcal{R}}_{\text{priv}} - \bar{\mathcal{R}}_{\text{ord}} - \bar{\mathcal{R}}_{\text{part}}$, located by linear interpolation on the logged ε grid and right-censored at the search limit $\varepsilon = 16$ when no crossing is observed. Existence, continuity, and the tie-limited finite-budget endpoint are detailed in Section A.4.

Any kernel with a C-independent tight limit has $\mathcal{R}_{\text{ord}}, \mathcal{R}_{\text{part}} \to 0$; Gaussian report-noisy-max does so as its noise scale diverges.

Let $m_{\max}(C)$ count top-count ties. As finite $\varepsilon \to \infty$, the channel selects uniformly among them: $\bar{\mathcal{R}}_{\text{priv}} \to \mathbb{E}_{x,\Pi,O}[1 - 1/m_{\max}(C)]$. The upstream limit uses the same randomized tie distribution, not fixed tie-breaking (Section A.4).

Proposition 3 (Existence of a crossover)**.** *If the finite-budget tie floor* $\mathbb{E}_{x,\Pi,O}[1 - 1/m_{\max}(C)]$ *is strictly below* $\lim_{\varepsilon\to\infty}(\bar{\mathcal{R}}_{\text{ord}} + \bar{\mathcal{R}}_{\text{part}})$*, then* ε^* *in Equation* (7) *exists.*

Continuity proves existence, not uniqueness or global monotonicity. The separately plotted deterministic endpoint resolves ties by a fixed rule; it is not the finite-budget limit. The same $P = T = 5$ cells as the curves satisfy the sufficient condition:

Condition	Tie floor	Tie-aware upstream limit	Check
Pythia/AGNews	.0469	.1187	$<$
Qwen/AGNews	.0100	.0173	$<$
Qwen/Banking77	.0124	.0382	$<$

3.5 COUNT-GAP CONTROL AND FINITE DESIGNS

For a unique plurality winner $k^\star = k^\star(C)$, let $d_j = C_{k^\star} - C_j$ and $\Delta = \min_{j \neq k^\star} d_j$. Its exact release error, one minus plurality fidelity, is

$$r(C, \varepsilon) = \frac{\sum_{j \neq k^\star} e^{-\varepsilon d_j/2}}{1 + \sum_{j \neq k^\star} e^{-\varepsilon d_j/2}}. \tag{8}$$

Factoring out the nearest-competitor gap gives an exact effective competitor mass,

$$N_{\text{eff}}(C, \varepsilon) = \sum_{j \neq k^\star} e^{-\varepsilon(d_j - \Delta)/2} \in [1, K-1], \qquad \log \frac{q_{k^\star}}{1 - q_{k^\star}} = \frac{\varepsilon \Delta}{2} - \log N_{\text{eff}}. \tag{9}$$

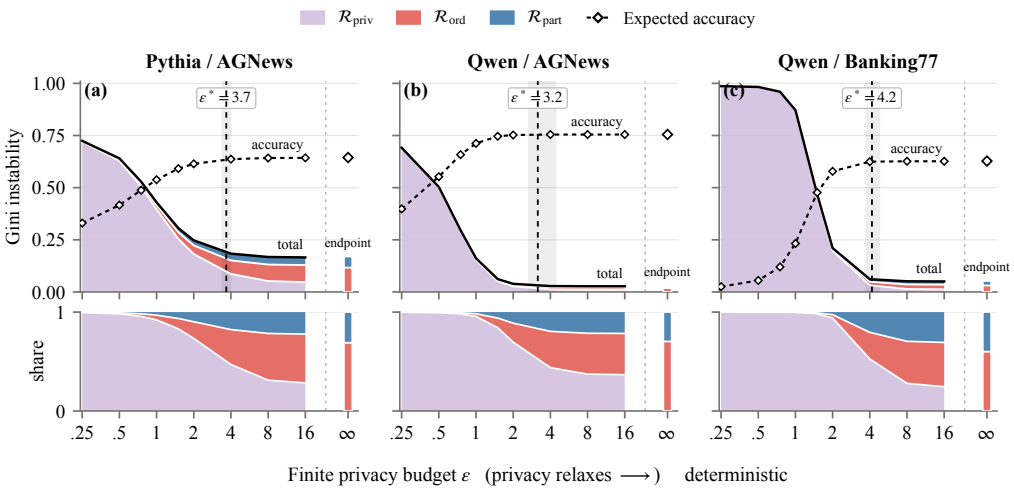

Figure 3: **Tight privacy masks upstream variation; relaxing ε re-exposes it.** Exact Gini components (lavender: privacy; coral: order; blue: partition), sampled-grid-conditional crossover intervals, component shares, and expected accuracy; finite curves end at $\varepsilon = 16$. Detached deterministic endpoints use fixed tie-breaking and are not the finite-ε limit when top counts tie.

Thus release fidelity is controlled by both the top-two gap and how much near-top competitor mass remains; replacing N_{eff} by $K - 1$ gives a worst-case bound. A per-vector sufficient condition and the exact minimum ensemble size for which any count vector can attain error at most α are

$$\varepsilon\Delta \geq 2 \log \frac{(K-1)(1-\alpha)}{\alpha}, \qquad S_{\min} = \left\lceil \frac{2}{\varepsilon} \log \frac{(K-1)(1-\alpha)}{\alpha} \right\rceil. \tag{10}$$

Unanimity attains the maximum $e^{\varepsilon S/2}/(e^{\varepsilon S/2} + K - 1)$, making this a feasibility boundary rather than a typical-profile guarantee. Section 5.2 tests that distinction; Section A.5 gives the full bounds and proof. Mean gap alone cannot determine ε^*, which also depends on N_{eff}, ties, K, and the upstream level.

Our nested estimates use P partitions and T seeded joint within-subset schedules per partition. The decomposition is exact for that empirical design; population random-effects components require finite-design correction, reported separately and derived in Section A.6.

4 Experimental Design

Categorical Gini—pairwise execution disagreement—is primary because its quadratic form admits exact attribution; accuracy is reported separately. Vote counts, margins, and adaptive diagnostics are private internal statistics; associations are descriptive, not causal mediation. Table 4 records each layer's randomization unit, evidence status, and permitted claim.

Private channel and nested decomposition. Figure 3 uses 100 queries, fixed $n = 24$ pools, $S = 8$ subsets of $s = 3$, and $P = T = 5$ partitions/schedules (25 cells/query) for AGNews/Banking77. Gumbel curves are reconstructed from pre-noise counts, not repeated releases; intervals condition on the sampled grid. A separate 20-realization non-nested audit assesses channel and Gaussian sensitivity, without query-wise identity across designs. Grids, ties, and bootstrap details are in Section B.

Prospective sizing, fixed-record, and taxonomy interventions. Before Qwen2.5-7B/Banking77 inference, we froze predictions, criteria, code, and hashes for paired $S = 8, 14, 24$ prefixes ($K = 77$, $\varepsilon = 1$); $S = 24$ targets were .55 accuracy and .90 fidelity. A second frozen protocol holds 72-record pools fixed for $S = 8, s = 9$ versus $S = 24, s = 3$, with accuracy and fidelity co-primary. Six seeded $s = 9$ orders preclude cross-arm component attribution; its offline ε curve is post hoc. An

exploratory CPU intervention contrasts $K = 50$ release-then-map with aggregation before $K = 6$ release. Protocols and paired-bootstrap details are in Section B.5.

Clean margins and matched bridges. Clean confirmation covers three models and four tasks ($K = 2, 4, 6, 14$): twelve collision-free conditions, 200 queries, and 100 orders, with disjoint splits for estimation and 5/10/20/50-order ablations. Matched bridges use 100 shared queries and six within-subset-exhaustive schedules for Qwen-7B/Pythia on AGNews and TREC coarse. AGNews detection/severity is post hoc; TREC froze AUC detection and active-only severity. The 14B protocol was frozen after 7B outcomes but before 14B inference, so it is a prospective scale replication, not an independent task preregistration. Full model, seed, and audit details are in Section B.

5 RESULTS

5.1 PRIVACY MASKING AND CROSSOVER

In all three audited nested conditions, corrected full-domain results are inconsistent with fixed-baseline additive accounting: the private channel changes which source of variation is visible in the released label. At small ε, release randomness dominates and upstream components approach zero; at the deterministic endpoint, all observed variation is upstream (Figure 3). Masking is not zero utility, as the audited $\varepsilon = 1$ slice in Table 1 illustrates. Stability and utility must therefore be reported jointly.

Table 1: **Audited $\varepsilon = 1$ slice of Figure 3.** Upstream is $\mathcal{R}_{\mathrm{ord}} + \mathcal{R}_{\mathrm{part}}$. The Qwen/Banking77 accuracy .232 is the $S = 8$ forecast basis audited in Table 6; Table 2 uses a separate 72-record sizing design.

Condition	Expected accuracy	$\mathcal{R}_{\mathrm{priv}}$	Upstream
Pythia/AGNews	.538	.396	.034
Qwen/AGNews	.712	.154	.007
Qwen/Banking77	.232	.869	.002

Grid-interpolated Gumbel crossovers on the logged finite grids (95% query-bootstrap percentile ranges), rounded to one decimal, are 3.7 [3.4,3.9], 3.2 [2.7,4.4], and 4.2 [3.8,4.7] for Pythia/AGNews, Qwen/AGNews, and Qwen/Banking77. On the same counts, Gaussian RNM crosses at $\sigma = .659, .793, .655$; analytic accounting maps these to $\varepsilon = 9.709, 7.687, 9.780$ at $\delta = 10^{-4}$. This is qualitative mechanism sensitivity, not equal-ε replication. The ranges resample queries on logged $P = T = 5$ grids, not the population crossover; finite-design correction shifts point roots downward by .16–.20 (Table 14). Tie floors and coverage audits are in Sections C.8 and C.10.

5.2 PROSPECTIVE ENSEMBLE SIZING

The channel analysis yields a forward design question: can increasing S move a previously unusable high-K configuration toward faithful release? For $K = 77$, $\varepsilon = 1$, and $\alpha = .1$, Equation (10) gives $S_{\min} = 14$. This is only the first ensemble size at which *some* count vector can reach .90 plurality fidelity. We therefore froze directional predictions across $S = 8, 14, 24$ and a practical target for $S = 24$ before running the new 72-record master ensemble.

Table 2: **A prospectively frozen sizing intervention behaves as predicted.** Qwen2.5-7B/Banking77, $\varepsilon = 1$, 100 paired-prefix queries; $n = 3S$. Brackets are query-bootstrap 95% intervals.

S	n	Expected accuracy	Plurality fidelity	$\mathcal{G}$	$\mathcal{R}_{\mathrm{priv}}$	Upstream
8	24	.237 [.202,.273]	.335 [.312,.356]	.870	.867	.003
14	42	.550 [.466,.632]	.808 [.769,.846]	.309	.299	.010
24	72	.614 [.521,.708]	.965 [.949,.979]	.066	.054	.012

Table 2 passes the conjunctive set of all eight directional interval criteria. All twelve tolerance checks also pass but are calibration diagnostics, not inferential criteria for this decision. The frozen $S = 14$ accuracy/fidelity predictions, .543/ .807, become .550/ .808 (errors .007/ .002). Its fidelity shows that feasibility is not a guarantee. At $S = 24$, fidelity .965 $[.949, .979]$ exceeds .90; accuracy .614

$[.521, .708]$ exceeds .55 as a point estimate although its interval includes .55. These two $S = 24$ targets were frozen as point criteria, weaker than the directional interval criteria.

Across $S = 8 \rightarrow 24$, mean N_{eff} falls 59.625→ 48.415 and exposure rises .047→.396. Privacy still exceeds upstream, so we claim neither a crossover nor a universal εS law. Exact release-error envelopes for the logged $K = 4, 50, 77$ profiles are in Figure 4.

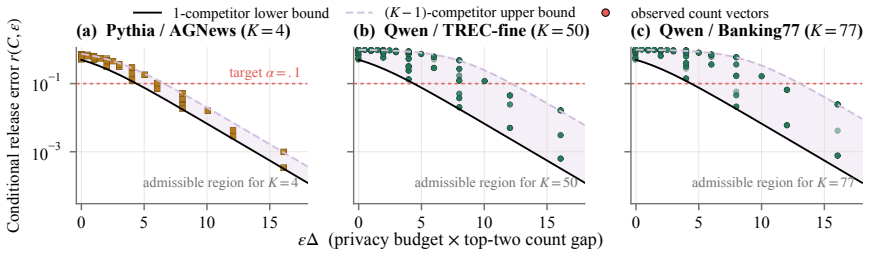

Figure 4: **Gap and competitor mass jointly control release error.** Exact one-competitor lower and $(K - 1)$-competitor upper bounds with logged profiles; the dotted line is the $\alpha = .1$ sizing target.

5.3 Fixed-record transmission window

A separately frozen fixed-record ICL analogue of PATE's teacher-count trade-off (Papernot et al., 2018) holds the 100 queries and 72-record pools fixed; our voters are contextual evaluations of one frozen model, not trained teachers. Pre-release plurality accuracy differs by only .008 $[-.032, .050]$. At $\varepsilon = 1$, $S = 24, s = 3$ instead of $S = 8, s = 9$ raises accuracy by .379 $[.318, .441]$ and fidelity by .636 $[.619, .653]$. The near-zero plurality gap limits a label-information explanation, while the count vectors still differ in the geometry the channel transmits. The unanimous $S = 8$ fidelity ceiling is only .418, whereas $S = 24$ can reach 1.000, consistent with the $S_{\min} = 14$ feasibility boundary.

The post-hoc offline curve descriptively shows double-ended convergence: the gain is negligible near uniform and deterministic release and peaks at .484 $[.408, .559]$ near $\varepsilon = .707$ (Figure 5). Random $s = 9$ orders restrict cross-arm claims to accuracy and fidelity; the window is a single $K = 77$ condition, not a universal curve.

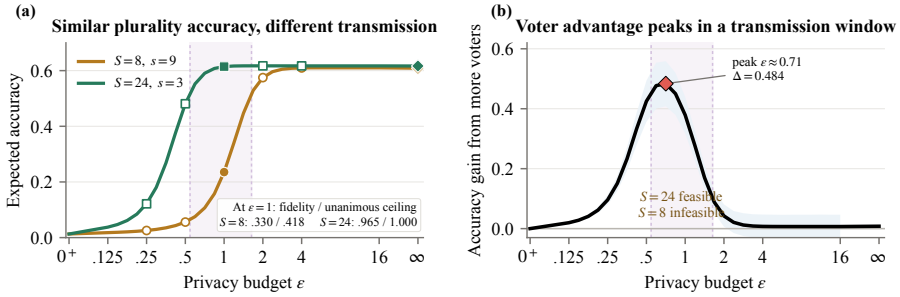

Figure 5: **At fixed record count, voter count matters inside a condition-specific transmission window. (a)** Exact offline curves and frozen $\varepsilon = 1$ marks. **(b)** Paired accuracy gap with 95% bootstrap band; vertical lines delimit where only $S = 24$ can attain .90 fidelity under unanimity. Curves are post hoc; $\varepsilon = 1$ endpoints are frozen.

5.4 Upstream structure and its clean-margin signature

Corrected nested-design order shares span .682–.819. In the matched design, synchronizing one permutation index across subsets gives shares .694–.848, whereas the post-hoc exact 6^8 independent-schedule reconstruction gives .606–.755. Independent subset changes can cancel, explaining the systematic reduction; every exact 95% interval nevertheless remains above one half (Table 15). Complementary partition shares are arithmetic, not independent evidence. Exact query-level partition estimates are nonnegative; trimming shows stronger right-tail concentration on AGNews than TREC.

Across twelve collision-free clean conditions, held-out Pearson correlations between signed-margin Cantelli score and Gini span .939–.968. Frozen TREC detection AUCs are .745/.748 at 7B and a weaker .630 [.520,.734] at 14B. Active-only severity is unresolved in all three and vote strength is task-dependent (Figure 6 and Table 15). The frozen detection target is synchronized deterministic plurality activity, not finite-budget detection; post-hoc independent-order sensitivity is in Table 16.

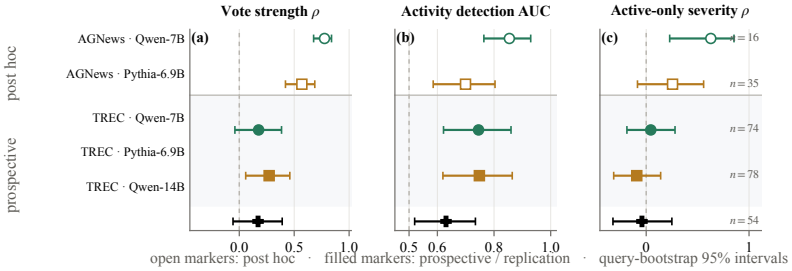

Figure 6: **Margins detect synchronized deterministic plurality activity, not universal severity.** TREC detection replicates prospectively; severity does not. Dashed lines mark zero correlation and chance AUC (.5).

6 Design Implications

Audit what the release channel hides. Equation (10) and Figure 4 give sufficient fidelity checks, not correctness guarantees. Internal count/gap audits require privacy accounting; Q pure-DP releases on a reused pool cost at most $Q\varepsilon$ under basic composition. Public-development calibration also requires public exemplar pools to avoid this cost.

Use the mechanism levers jointly. At fixed $n = 24$, more, weaker voters (Table 9) improve released accuracy at $\varepsilon = 1$ in all four conditions but lose accuracy at $\varepsilon = 8$ (Table 3); TREC-fine fidelity improves at both. Gaussian replay is similar (Section C.6). Fidelity-oriented allocation need not maximize utility: capacity screening still requires accuracy validation. Public taxonomy coarsening changes the task (Section C.7); private label restriction needs new accounting.

Table 3: Fixed-n allocation: $C-A$ accuracy changes for $(S, s) = (24, 1)$ versus $(8, 3)$, with original $\varepsilon = 1$ and post-hoc $\varepsilon = 8$ paired query-bootstrap 95% intervals. The frozen $\varepsilon = 1$ panel treats TREC as primary and DBpedia as descriptive.

Condition	$\varepsilon = 1$	$\varepsilon = 8$
Mistral / DBpedia	$+0.056$ $[+0.028, +0.085]$	-0.096 $[-0.144, -0.048]$
Mistral / TREC-fine	$+0.036$ $[+0.019, +0.055]$	-0.134 $[-0.186, -0.085]$
Qwen-1.5B / DBpedia	$+0.139$ $[+0.102, +0.175]$	-0.079 $[-0.126, -0.035]$
Qwen-1.5B / TREC-fine	$+0.061$ $[+0.042, +0.081]$	-0.055 $[-0.103, -0.006]$

7 Limitations

Gumbel crossovers do not imply equal-ε Gaussian replication. The bridge supports synchronized deterministic activity detection, not universal strength/severity ranking; both levels admit stable errors (Section C.9). Sizing/fixed-record validation covers only Qwen/Banking77. Allocation covers two models and two tasks at $n = 24$, $\varepsilon = 1$; replay is post hoc. TREC-fine accuracy remains low (.086/.111), with mixed or refuted secondary predictions (Section C.4). Attribution uses $s = 3$; larger-s persistence is untested; $s = 9$ is nonexhaustive. Taxonomy coarsening is exploratory and changes the target; nested intervals condition on sampled grids. Internal counts, gaps and margins are private; no deployable private audit is provided. Plug-in selectors do not beat maximum S (Section C.5). Other pipelines require fresh audits (Table 4).

Reproducibility Statement

The anonymous artifact includes configurations, frozen protocol files and their SHA-256 records, audited score/count vectors, stable query/exemplar identifiers, seeds, public model identifiers, package versions, analysis and figure scripts, and source hashes. Logged counts reconstruct every private curve without model inference; acceptance checks and compute details are in Section B. Exact revision hashes are unavailable for legacy model checkpoints, as disclosed in Section B.6.

Ethical Considerations

The mechanisms studied here protect records in an inference-time exemplar pool for each released query under the stated adjacency relation. They do not make the frozen language model, public benchmarks, prompts, or unreleased diagnostic artifacts differentially private. Reusing a private pool across released queries requires composition accounting, and inspecting vote counts or margins to adapt the deployment consumes privacy unless it is covered by an additional private mechanism. Our experiments use established text-classification benchmarks and do not collect new human-subject data. We report poor-utility regimes explicitly because low measured instability under a nearly uniform private channel should not be interpreted as a safe or useful deployment by itself.

AI Use Statement

During research and manuscript preparation, the authors used OpenAI Codex to assist with grammar checking, linguistic polishing, standardizing reference formatting, as well as code debugging and implementation suggestions. All research design, experiments and analysis were conducted by the authors. All AI-generated outputs were manually reviewed and revised. The authors take full responsibility for all content of this paper.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016. doi: 10.1145/2976749.2978318.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5.

Liang Chen, Li Shen, Yang Deng, Xiaoyan Zhao, Bin Liang, and Kam-Fai Wong. PEARL: Towards permutation-resilient LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=txoJvjfI9w.

Kwesi Adu Cobbina and Tianyi Zhou. Where to show demos in your prompt: A positional bias of in-context learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29560–29593, 2025. doi: 10.18653/v1/2025.emnlp-main.1503.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019, 2023. doi: 10.18653/v1/2023.findings-acl.247.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1184–1193. PMLR, 2018.

Zeyu Ding, Daniel Kifer, Sayed M. Saghaian Nejad Esfahani, Thomas Steinke, Yuxin Wang, Yingtai Xiao, and Danfeng Zhang. The permute-and-flip mechanism is identical to report-noisy-max with exponential noise. arXiv:2105.07260, 2021.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.64.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76852–76871, 2023. doi: 10.52202/075280-3358.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1693–1702. PMLR, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284, 2006. doi: 10.1007/11681878_14.

Fengyu Gao, Ruida Zhou, Tianhao Wang, Cong Shen, and Jing Yang. Data-adaptive differentially private prompt synthesis for in-context learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/96d00450ed65531ffe2996daed487536-Abstract-Conference.html.

Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

Qi Guo, Leiyu Wang, Yidong Wang, Wei Ye, and Shikun Zhang. What makes a good order of examples in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14892–14904, 2024. doi: 10.18653/v1/2024.findings-acl.884.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. DP-OPT: Make large language model your privacy-preserving prompt engineer. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ifz3IgsEPX.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. arXiv:2310.06825, 2023.

Seongho Joo, Hyukhun Koh, and Kyomin Jung. Public data assisted differentially private in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16129–16152. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.875.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia: A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6 (2):167–195, 2015. doi: 10.3233/SW-140134.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150/.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=bVuP3ltATMz.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022. doi: 10.18653/v1/2022.deelio-1.10.

Yang Liu and Chenhui Chu. Understanding the prompt sensitivity. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44362–44388. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2053.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556.

Ryan McKenna and Daniel R. Sheldon. Permute-and-flip: A new mechanism for differentially private selection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 193–203, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/01e00f2f4bfcbb7505cb641066f2859b-Abstract.html.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ0hF1Z0b.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pp. 94–103, 2007. doi: 10.1109/FOCS.2007.66.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pp. 75–84. ACM, 2007. doi: 10.1145/1250790.1250803.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning

and induction heads, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. Transformer Circuits Thread.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkwoSDPgg.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZB1XbRZ.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11054–11070, 2021.

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

Ossi Räisä and Antti Honkela. A bias–variance decomposition for ensembles over multiple synthetic datasets. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 307–315. PMLR, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZtt0pRnOl.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15614–15638, 2023. doi: 10.52202/075280-0687.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023. doi: 10.48550/arXiv.2303.03846.

Tong Wu, Ashwinee Panda, Jiachen T Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=x4OPJ7lHVU.

Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. Addressing order sensitivity of in-context demonstration examples in causal language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6467–6481, 2024. doi: 10.18653/v1/2024.findings-acl.386.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Yusuke Yamasaki, Kenta Niwa, Daiki Chijiwa, Takumi Fukami, and Takayuki Miura. Plausible token amplification for improving accuracy of differentially private in-context learning based on implicit Bayesian inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70205–70248. PMLR, 2025.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437, 2022. doi: 10.18653/v1/2022.emnlp-main.155.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021.

A Proofs and Technical Details

A.1 Categorical Gini facts

Lemma 1 (Modal and pairwise disagreement)**.** *For* $M(x) = 1 - \max_k p_k(x)$, $M(x) \le \mathcal{G}(x) \le 2M(x)$.

Proof. Let $p_* = \max_k p_k$. Since $\sum_k p_k^2 \le p_* \sum_k p_k = p_*$, $\mathcal{G} \ge 1 - p_* = M$. Also $\mathcal{G} \le 1 - p_*^2 = (1 - p_*)(1 + p_*) \le 2M$. □

Lemma 2 (Deterministic label coarsening lowers measured Gini)**.** *Let* $f : [K] \to [J]$ *merge labels and* $q_g = \sum_{k:f(k)=g} p_k$*. Then*

$$\mathcal{G}(p) - \mathcal{G}(q) = 2 \sum_{g=1}^{J} \sum_{\substack{i<j \\ f(i)=f(j)=g}} p_i p_j \ge 0. \tag{11}$$

Proof. Expand $q_g^2 = (\sum_{k:f(k)=g} p_k)^2$ group by group and subtract $1 - \sum_g q_g^2$ from $1 - \sum_k p_k^2$. □

A.2 Proof of the clean margin bound

Proof of Proposition 1. The error event $\{\widehat{Y} \ne w\}$ is contained in $\{D_w - \mu_D \le -\mu_D\}$, including when ties are broken against w. Cantelli's one-sided inequality gives the first bound. Applying the modal-disagreement bound with $1 - p_w = \mathbb{P}(\widehat{Y} \ne w)$ gives the second; the argument remains valid if w is not modal because $\mathcal{G}_{\mathrm{L1}} \le 1 - p_w^2 \le 2(1 - p_w)$. □

A.3 Privacy and decomposition proofs

Per-query privacy. For Equation (4), a replacement affects one disjoint subset vote, so each count utility changes by at most one. Sampling proportionally to $\exp(\varepsilon C_k/2)$ is therefore the exponential mechanism with sensitivity one. The guarantee holds conditional on data-independent partitioning and ordering and hence for their mixture. □

Replacement is the natural adjacency for a fixed-size exemplar pool: it moves one vote between two classes, so the utility is non-monotone and the $\varepsilon/2$ exponent is necessary. For example, moving one vote from the second class to the first gives $K = 2$, $C = (0, S)$ versus $(1, S-1)$. The corresponding privacy-loss ratio is $(1 + a^S)/(1 + a^{S-2})$ for $a = e^{\varepsilon/2}$; its log approaches ε as $S \to \infty$. The guarantee is therefore asymptotically tight rather than attained at finite S. The monotone-utility variant $\exp(\varepsilon C_k)$ of McSherry & Talwar (2007) applies only under an add/remove adjacency in which a single vote is inserted or deleted with all others fixed — a different neighboring relation than the one protected here.

Proof of Proposition 2. Let Z be the one-hot released label. Conditional on (Π, O), $\mathbb{E}[Z \mid \Pi, O] = q_{\Pi,O}$ and $\mathbb{E}[\|Z - q_{\Pi,O}\|_2^2 \mid \Pi, O] = 1 - \|q_{\Pi,O}\|_2^2$. Apply vector total variance over release randomness, then O conditional on Π, and finally Π. Since Z is one-hot, $\mathbb{E}\|Z - p\|_2^2 = 1 - \|p\|_2^2$. □

A.4 Finite-budget limits and ties

Each component is a finite average of functionals of $q(C, \varepsilon)$ over the finite count support, and $q(C, \varepsilon)$ is smooth in ε on $(0, \infty)$; hence $\bar{\mathcal{R}}_{\mathrm{priv}}, \bar{\mathcal{R}}_{\mathrm{ord}}, \bar{\mathcal{R}}_{\mathrm{part}}$ are continuous and the limits below exist. As $\varepsilon \to 0$, $q(C, \varepsilon)$ becomes uniform, proving Equation (6), so $g(\varepsilon) := \bar{\mathcal{R}}_{\mathrm{priv}}(\varepsilon) - \bar{\mathcal{R}}_{\mathrm{ord}}(\varepsilon) - \bar{\mathcal{R}}_{\mathrm{part}}(\varepsilon) > 0$ for small ε. Let $m_{\max}(C)$ be the number of classes tied at the maximum count. Along finite budgets approaching infinity, the channel converges to $t_k(C) = \mathbf{1}\{k \in M(C)\}/m_{\max}(C)$, not to the fixed tie-breaking plurality label. For each query,

$$\mathcal{R}_{\mathrm{priv}} \to \mathbb{E}_{\Pi,O}[1 - 1/m_{\max}(C)], \tag{12}$$

while $\mathcal{R}_{\mathrm{ord}}+\mathcal{R}_{\mathrm{part}} \to \mathbb{E}_{\Pi,O}\|t(C)\|_2^2-\|\mathbb{E}_{\Pi,O}t(C)\|_2^2$; the condition-level limit averages this quantity over queries. A fixed-rule deterministic endpoint can differ. These limits supply the continuity and endpoint facts used by Proposition 3. Its reported limits are recomputed from the same $P = T = 5$ cells as the crossover curves, not imported from the separate fixed-domain audit in Section C.10.

A.5 Top-two count-gap bounds

Proposition 4 (Top-two count-margin bounds)**.** *For finite $\varepsilon > 0$,*

$$\frac{e^{-\varepsilon\Delta/2}}{1+e^{-\varepsilon\Delta/2}} \le r(C,\varepsilon) \le \frac{(K-1)e^{-\varepsilon\Delta/2}}{1+(K-1)e^{-\varepsilon\Delta/2}}. \tag{13}$$

Conditional privacy Gini also satisfies $2r(1-r) \le 1-\|q(C,\varepsilon)\|_2^2 \le 2r$.

Proof. At least one competitor has gap Δ, and all $K-1$ competitors have gap at least Δ, which bounds the sum in Equation (8). Competitor probabilities are nonnegative and sum to r; their squared sum lies between zero and r^2, yielding the Gini bounds. □

Solving the upper bound for $r(C,\varepsilon) \le \alpha$ gives Equation (10). This is sufficient, not necessary. For the ensemble-level statement in Equation (10), a label's softmax probability is maximized by assigning all S votes to that label and zero to every competitor. The resulting minimum possible release error is $(K-1)/(e^{\varepsilon S/2}+K-1)$. Solving this exact optimum for error at most α yields $S_{\min}$. If this optimum is at most α, the unanimous count vector witnesses feasibility; if it exceeds α, every other count vector has at least as much release error, so no feasible vector exists. This proves both directions of the if-and-only-if claim.

A.6 Finite nested-design correction

For P independently sampled partitions and T conditionally independent orders per partition, Equation (5) holds exactly over the empirical discrete distribution. Under the balanced random-effects model, population components use

$$\widehat{R}_{\mathrm{ord}} = \frac{T}{T-1}R_{\mathrm{ord}}^{\mathrm{emp}}, \qquad \widehat{R}_{\mathrm{part}} = \frac{P}{P-1}R_{\mathrm{part}}^{\mathrm{emp}} - \frac{\widehat{R}_{\mathrm{ord}}}{T}. \tag{14}$$

The exact finite-design identity and corrected population estimates answer different questions and are reported separately. Because the partition estimator subtracts an estimated order contribution, it is not constrained to be nonnegative query by query; such estimates must be retained rather than clipped. The post-hoc exact-order reconstruction removes this subtraction for the order component and yields only nonnegative partition estimates, including structural zeros, in the five audited conditions (Section C.9).

A.7 A finite-budget attenuation bound

The tight-privacy limit also admits a conservative finite-budget bound. We give this elementary supplement to distinguish a guaranteed upper bound from the empirical forecasts; it is not a new privacy theorem.

Lemma 3 (Softmax attenuation of count variation)**.** *For fixed $K \ge 2$, $\varepsilon \ge 0$ and a count vector C with finite second moment, let $q(C) = \mathrm{softmax}(\varepsilon C/2)$ and $\mathrm{Var}_2(X) = \mathbb{E}\|X-\mathbb{E}X\|_2^2$. Then*

$$\mathcal{R}_{\mathrm{ord}}+\mathcal{R}_{\mathrm{part}} = \mathrm{Var}_2(q(C)) \le \min\left\{\frac{\varepsilon^2}{16}\mathrm{Var}_2(C),\, 1-\frac{1}{K}\right\}. \tag{15}$$

Also $\mathcal{R}_{\mathrm{ord}} \le (\varepsilon^2/16)\mathbb{E}_\Pi[\mathrm{Var}_2(C \mid \Pi)]$.

Proof. The softmax Jacobian is $J = \mathrm{diag}(q) - qq^\top$. Its absolute row sum is $2q_i(1-q_i) \le 1/2$, and symmetry gives $\|J\|_2 \le 1/2$. Thus $C \mapsto q(C)$ is $\varepsilon/4$-Lipschitz. For independent copies C, C', $\mathrm{Var}_2(q(C)) = \frac{1}{2}\mathbb{E}\|q(C)-q(C')\|_2^2 \le (\varepsilon^2/16)\,\mathrm{Var}_2(C)$. The simplex bound follows from $\mathbb{E}\|q\|_2^2 - \|\mathbb{E}q\|_2^2 \le 1-1/K$. Applying the same independent-copy argument conditional on Π proves the order bound. □

There is no corresponding substitution of $\mathrm{Var}_{\Pi}(\mathbb{E}[C \mid \Pi])$ for the partition term without additional assumptions: nonlinear release and conditional averaging do not commute. In the deterministic-consensus panel of Section C.6, all order terms are zero, so the bound applies to partition variation directly. At $\varepsilon = 1$ it is loose: the Mistral/TREC arm-A mean upper bound is .135 against .00117 measured partition variation. Arm C has constant counts across partitions and both sides are zero. This bound quantifies a guaranteed attenuation rate near zero privacy budget, but is not used to locate a crossover or select an arm.

B Extended Experimental Details

B.1 Clean ICL evidence

The historical clean study records categorical predictions, full class scores where available, correct-label loss variance, and per-order accuracy. We use collision-free AGNews artifacts for retrospective mechanism comparison but do not reuse their scalar loss threshold as a theorem. For each signed-margin analysis, the modal prediction across estimation orders fixes w and D_w is computed from log probability ratios. Fifty random 50/50 splits report mean correlation and the 2.5–97.5 percentile range; this range measures order sampling sensitivity, not confidence over queries. The extended nested order-budget battery uses 100 splits with seed 20260722 and never excludes uninformative queries post hoc. Its secondary rank and active-only analyses use the same frozen splits.

An artifact audit found first-token collisions in the first-token high-cardinality classifier: Banking77 has 45 unique first-token columns for 77 labels and TREC-fine has 47 for 50. Repeated logits were duplicated before normalization, changing ties and probabilities. Those conditions are excluded from primary clean evidence. First-token verbalizers are injective for SST-2, AGNews, TREC coarse and DBpedia; complete label sequences are unique on all six datasets. Pythia and Qwen have identical uniqueness counts. The collision-free TREC-fine `sequence_mean` run has 200 queries and 100 orders; Banking77 compares mean and summed sequence log likelihood on the same 50 queries and 30 orders.

B.2 Private channel details

Every subset prediction is constrained to one canonical class, including zero-vote labels in the public domain. All primary corrected conditions pass zero-invalid-vote and fixed-domain audits. Under replacement adjacency, one record can move one vote between two classes, giving count-vector ℓ_2 sensitivity $\sqrt{2}$. For the Gaussian comparison, noise scale σ is the primary channel parameter. The classical expression used by Wu et al. (2024), $\sigma = 2\sqrt{\log(1.25/\delta)}/\varepsilon$ with $\delta = 10^{-4}$, defines the prior-work-aligned grid only. Reported (ε, δ) labels use analytic Gaussian accounting (Balle & Wang, 2018), because the classical formula is not a valid calibration for $\varepsilon \geq 1$. Equal ε is not a like-for-like comparison with pure-DP Gumbel RNM. For each count vector, Gaussian RNM uses $q_k = \int \phi(z) \prod_{j \neq k} \Phi(z + (C_k - C_j)/\sigma)\, dz$, with standard-normal density ϕ and CDF Φ. The original channel audit integrates over $[-10, 10]$ with absolute/relative tolerances $2 \times 10^{-10}/2 \times 10^{-8}$ and subdivision limit 120. Allocation replay instead uses $[-10, 10]$, absolute/ relative tolerances $10^{-10}/10^{-9}$, and subdivision limit 200. Every replay profile is independently checked on $[-12, 12]$ at $10^{-12}/10^{-11}$, also with limit 200; the reported probabilities normalize the replay integral, not the stricter check. These are deterministic quadrature results, not Monte Carlo estimates.

The nested design reconstructs the complete channel curve offline. We report exact empirical components, query-bootstrap intervals with 5,000 resamples, and the population correction in Equation (14). The two AGNews runs use identical query sources, labels, extraction, and randomization. Crossover intervals resample queries while holding each query's sampled partition/order grid fixed. They therefore quantify query uncertainty conditional on that finite upstream design, not superpopulation uncertainty from redrawing partitions and orders.

B.3 Synthetic estimator-audit protocol

Before simulation, we froze a CPU-only estimator audit over finite, exactly enumerable count-vector hierarchies. The primary grid crosses $K \in \{4, 6, 50, 77\}$ with three plurality geometries and six

structural regimes (null, order-only, partition-only, and three mixed dispersion patterns), always using $S = 8$, 100 queries, and $P = T = 5$. Each hierarchy assigns known probabilities to partition and conditional order states, and every state maps to a valid integer count vector. We compute oracle channel components by exact weighted summation, then run 1,000 sampled designs with 1,000 query-bootstrap resamples each, at ten budgets $\varepsilon \in \{.25, .5, .75, 1, 1.5, 2, 4, 8, 16, \infty\}$. The 324 secondary cells are separate sweeps, not a Cartesian product: $S \in \{4, 16\}$ across all four K values gives $2 \cdot 4 \cdot 3 \cdot 6 = 144$; $Q \in \{50, 200\}$ at $K \in \{4, 77\}$ gives $2 \cdot 2 \cdot 3 \cdot 6 = 72$; and $P = T \in \{3, 5, 10\}$ at those two K values gives $3 \cdot 2 \cdot 3 \cdot 6 = 108$. Unvaried parameters retain primary values; each secondary cell uses 300 trials and 300 bootstraps. The separate crossover audit uses 200 trials and 500 bootstraps per population on 400 geometrically spaced budgets from .05 to 16. All phases derive seeds from base seed 20260810 and condition/phase/trial IDs.

Disclosed v1.1 numerical amendment. The original frozen run exposed floating-point cancellation at mathematically zero order components (maximum magnitude 1.49×10^{-15}). Version 1.1 canonicalizes nonnegative variance components at or below 10^{-12} to zero; it changes no count population, seed, grid, bootstrap, or interpretation threshold, and retains the original outputs. Corrected partition estimates are never clipped. Full manifests, both result versions, trial arrays, and the amendment are included in the artifact.

B.4 Matched bridge endpoints and audits

The Level-1 strength endpoint is Spearman correlation between mean signed margin and mean normalized Level-2 top-two vote margin. The pooled endpoint is Spearman correlation between Level-1 Cantelli score and deterministic $\mathcal{R}_{\mathrm{ord}}$ under six synchronized schedules. The decomposed endpoints are tie-aware AUC for detecting $\mathcal{R}_{\mathrm{ord}}^{\mathrm{det,sync}} > 0$ and Spearman severity correlation conditional on this activity. These frozen endpoints do not identify finite-ε or independently permuted order activity. Population order shares are ratios of cross-query mean components and retain all queries, including structural zeros. Query-level shares are diagnostic only and are defined only when total upstream variation is positive; no such conditional mean is a primary endpoint. Bootstrap resampling clusters by query and never treats orders, subsets, or count vectors as independent observations. The Qwen2.5-14B TREC protocol was frozen after the 7B TREC outcomes were known but before 14B inference. It is therefore a prospective model-scale replication, not an independent task-level preregistration. Its paired comparison uses the same query, exemplar, partition, and order identifiers as Qwen-7B.

For structural attribution only, we additionally perform a post-hoc exact CPU reconstruction. Each subset was evaluated under every one of its six permutations, so repeated sparse convolution gives the categorical vote-count distribution under all $6^8 = 1{,}679{,}616$ independent joint schedules without new inference. This removes the finite-T correction from the order component; the partition component retains the $P/(P-1)$ population correction. It does not replace the synchronized schedules in the prospectively frozen bridge detection and severity endpoints.

B.5 Prospective sizing and taxonomy protocols

The prospective Qwen2.5-7B/Banking77 (Casanueva et al., 2020) master design uses 100 stratified queries, one 72-example pool per query, five partitions, and six synchronized schedules: all permutations within each three-example subset, not crossed across subsets. Arms $S = 8, 14, 24$ are paired prefixes, not independent runs. The analyzer requires complete query/membership/partition/order/vote grids and applies the exact 77-way Gumbel channel offline at $\varepsilon = 1$.

Predictions, protocol, analyzer and hashes were committed before smoke/full inference. Both adjacent comparisons require accuracy/fidelity increases and Gini/privacy decreases: all eight paired 95% intervals must lie wholly in their frozen directions. This is one intersection–union decision, not eight opportunities for discovery; union-of-rejections family-wise-error arithmetic does not apply. Separately, $S = 24$ point estimates must reach accuracy .55 and fidelity .90. The twelve deterministic calibration checks allow .08 accuracy, .12 Gini/privacy and .10 fidelity error; they are not significance tests and cannot override failure. Intervals use 10,000 paired query resamples, seed 20260813.

The separately frozen fixed-record test compares $S = 8, s = 9$ with the existing $S = 24, s = 3$ arm on the same 100 queries and 72-example pools. Both co-primary paired accuracy/fidelity intervals at $\varepsilon = 1$ must exclude zero in favor of $S = 24$; neither endpoint alone establishes success. Six

seeded $s = 9$ orders are not exhaustive 9!, precluding exact cross-arm component comparisons. Three complete, non-overlapping query shards (33/33/34) match a subsequent single-GPU run record-for-record (SHA-256). The later offline ε curve is descriptive.

The earlier frozen C6 panel crosses Mistral/Qwen-1.5B with DBpedia/TREC-fine. Each subset emits its modal label across all six orders, compared with the average of six fixed-order policies. Success requires privacy/Gini decreases, a gap increase and accuracy non-inferiority ($-.02$) in all four conditions. Both TREC-fine conditions fail the mechanism rule. A post-outcome CPU replay of five soft aggregators qualifies none for fresh inference; it cannot alter that verdict or validate a replacement (Section C.3).

After this outcome, the separately frozen allocation study uses the same four conditions with 200 queries each, holding each $n = 24$ pool fixed across A $(S, s) = (8, 3)$, B $(12, 2)$ and C $(24, 1)$. A reuses the earlier artifacts; B/C are fresh, matched on models/tasks/queries/pools/seeds/scoring. Both TREC-fine models must have $C-A$ fidelity lower bounds above zero and accuracy lower bounds at least $-.02$. Fewer than 180 audited queries in either condition means incomplete, not passed. Intervals use 10,000 paired query resamples, seed 20260827. Fresh $C-B$ contrasts check provenance without replacing the primary test.

All three frozen secondary predictions, including the initially omitted third, are reported in Section C.4. Post-outcome utility/cost/selection and channel audits (Sections C.5 and C.6) leave the primary verdict unchanged; neither is new prospective evidence or a privacy-free deployment procedure.

The exploratory taxonomy intervention reuses the logged TREC-fine count vectors and pairs every observation across two mechanisms: 50-way release followed by the public fine-to-coarse map, and public count aggregation followed by a declared six-way release. Accuracy, plurality fidelity, release error, and N_{eff} are primary descriptive quantities; Gini is secondary because deterministic coarsening lowers it mechanically. This comparison was not prospectively frozen and changes the prediction target.

Table 4 records the randomization unit, evidence status, and permitted claim for every experimental layer.

The artifact package records query/exemplar hashes, partition memberships, ordered subset IDs, full subset scores, vote vectors, tokenizer audits, seeds, and source hashes. Automated checks reject incomplete grids or label domains, invalid votes, duplicated first-token verbalizers, repeated permutations, membership changes, and decomposition-identity failures.

B.6 Implementation and compute

The evaluated model families are Qwen2.5, Pythia, Llama 3, and Mistral (Qwen Team, 2024; Biderman et al., 2023; Grattafiori et al., 2024; Jiang et al., 2023). Benchmarks are SST-2, AGNews, TREC, DBpedia, and Banking77 (Socher et al., 2013; Zhang et al., 2015; Li & Roth, 2002; Lehmann et al., 2015; Casanueva et al., 2020). Local Hugging Face checkpoints are evaluated in float16 on a server equipped with NVIDIA A6000 GPUs (48 GB each). The original 100-query matched-bridge runs use one visible GPU and label batch size one; the later allocation panel uses label batches of eight (Section C.5). Completed-run metadata record 28–89 minutes per full 100-query bridge condition; the paired Qwen2.5-14B TREC condition took 50 minutes. The ten logged bridge smoke/full runs sum to 4.463 GPU-hours, a strict lower bound because legacy runs lack complete timing logs. Exact private-channel curves, bootstrap intervals, and the 7B–14B paired comparison are reconstructed from logged scores and counts on CPU. The release package records public model identifiers and recoverable software versions. Exact checkpoint revision hashes were not embedded in the legacy inference metadata, which is a reproducibility limitation of the completed runs.

C Extended Results

C.1 Clean-margin figures and high-cardinality scoring

Consistent with Figure 8, at 50 orders all-query Spearman correlations range from .392 to .911 because inactive-query ties differ substantially by condition. Conditioning on positive held-out Gini, Pearson remains .887–.954 and Spearman remains .882–.932. These are distinct estimands: Table 5

Table 4: Evidence inventory. Each row has a distinct randomization unit and supports only the claim in the final column. Section pointers form an appendix roadmap.

Evidence layer	Scope and design	Permitted claim
Retrospective clean (§C.1)	4 models on AGNews; 200 queries $\times$ 100 orders; first-token scoring	Mechanism comparison
Clean confirmation (§C.1)	3 models $\times$ 4 tasks; 200 queries $\times$ 100 orders; held-out, collision-audited splits	Cross-model/task margin relation
High-K sensitivity (§C.1)	Qwen on TREC-fine and Banking77; 200×100 and 50×30; sequence mean/sum	Scoring sensitivity, not pooled magnitude
Fixed-domain private (§B.2)	3 model–task conditions; 100 queries, 20 upstream replicates; closed-form Gumbel and quadrature-evaluated Gaussian channels	Independent channel audit and Gaussian sensitivity
Nested decomposition (§B.2)	3 model–task conditions; $P = 5$, $T = 5$, fixed 24-example pool; full-domain votes	Figure 3 masking/crossover and order/partition shares
Exact order reconstruction (§C.9)	5 matched conditions; all 6^8 independent joint schedules per partition, reconstructed on CPU	Post-hoc aggregate attribution and tail-concentration audit
Synthetic estimator audit (§C.8)	72 primary and 324 secondary cells; enumerable count hierarchies; frozen bias/coverage criteria	Finite-design reliability and crossover-interval scope
Matched bridge (§C.9)	2 models on AGNews and TREC; 100 queries/task; 100 Level-1 orders; $P = 5$, six synchronized schedules, marginally exhaustive	Prospective detection; task-specific strength/severity
Scale bridge (§C.9)	Qwen 7B/14B on paired TREC; same queries, pools, partitions, orders, scores, and votes	Prospective scale detection; descriptive paired magnitude
Bridge robustness (§C.9)	5 matched conditions; logged Level-1/2 artifacts; 10,000 query bootstraps	Post-hoc baselines, reliability, stable-wrong rates
Subset thinning (§C.10)	5 conditions at $S = 4/6/8$; exhaustive retention from logged $S = 8$ votes; exact Gumbel channel	Mechanism sensitivity, not an end-to-end S effect
Prospective sizing (§C.2)	Qwen/Banking77 at $S = 8/14/24$; 100 paired queries; new 72-record pools; frozen predictions and criteria	One-setting forward design validation, not a scaling law
Fixed-record count (§B.5)	Qwen/Banking77, $S = 8$ vs. 24; same 100 queries and 72-record pools; frozen paired $\varepsilon = 1$ endpoints	Voter-count effect on accuracy/fidelity only; one high-K condition
Within-subset consensus (§C.3)	Mistral/Qwen-1.5B on DBpedia/TREC-fine; exhaustive six-order consensus; frozen joint rule plus post-outcome CPU replay	Negative operating boundary and diagnosis only; no universal intervention
Fixed-record allocation (§C.4)	Mistral/Qwen-1.5B on DBpedia/TREC-fine; $n = 24$ across $(S, s) = (8, 3), (12, 2), (24, 1)$; frozen TREC joint rule	Voter-count effect at fixed n in this panel; no (S, s) optimality
Allocation audit (§C.5)	Same 12 arms; post-hoc utility identity, individual-vote accuracy, compute counts and development/test split simulation	Accounting and selection limits only; no new primary verdict or adaptive DP policy
Allocation channel replay (§C.6)	Same 12 arms; Gumbel ε and Gaussian σ grids; post-hoc, paired query intervals and tie-aware limits	Budget-dependent utility/fidelity trade-offs; not independent replication, equal-ε mechanism comparison or an optimal allocation rule
Taxonomy intervention (§C.7)	Qwen/TREC-fine, $K = 50 \to 6$; paired logged counts; release–map vs. public aggregation–release	Exploratory mechanism guidance, not prospective evidence
Gap attainability (§C.10)	2 TREC models plus Qwen/Banking77; 3,000 vectors/model and 2,000 Banking77 vectors; pre-noise counts	Feasibility of sufficient sizing thresholds

Table 5: Held-out Pearson correlation between the 50-order signed-margin score and Level-1 Gini order sensitivity on disjoint orders (200 queries and 100 orders per condition). Main-text summary: $r \in [.939, .968]$ (Figure 2).

Model	SST-2	AGNews	TREC coarse	DBpedia
Qwen2.5-7B-Instruct	.964	.968	.950	.968
Pythia-6.9B	.950	.963	.944	.939
Llama-3-8B	.943	.960	.952	.939

reports all-query held-out Pearson, the active-only range above conditions on positive held-out Gini, and the structural reference below uses the training-defined informative subset $\mu_D > 0$. These secondary checks show that the high Pearson relation is not generated solely by a mass at zero. Figure 8, rather than the structural reference below, supports the practical scope: Cantelli's advantage is largest at 5–10 orders, while direct training Gini catches up as its budget grows.[1]

[1] A post-hoc Gaussian-margin reference, recomputed on the same 100 disjoint 50/50 splits and informative-query subsets, gives Pearson .979–.993, versus empirical .934–.966 across 12 clean model–task conditions. The reference is higher in all condition-level means, consistent with shared margin geometry explaining much of the raw association. It is neither an upper bound nor a competing estimator: it replaces empirical outcomes with Gaussian outcomes constructed from the same training moments. It is distribution-dependent and motivates rather than establishes the order-budget scoping.

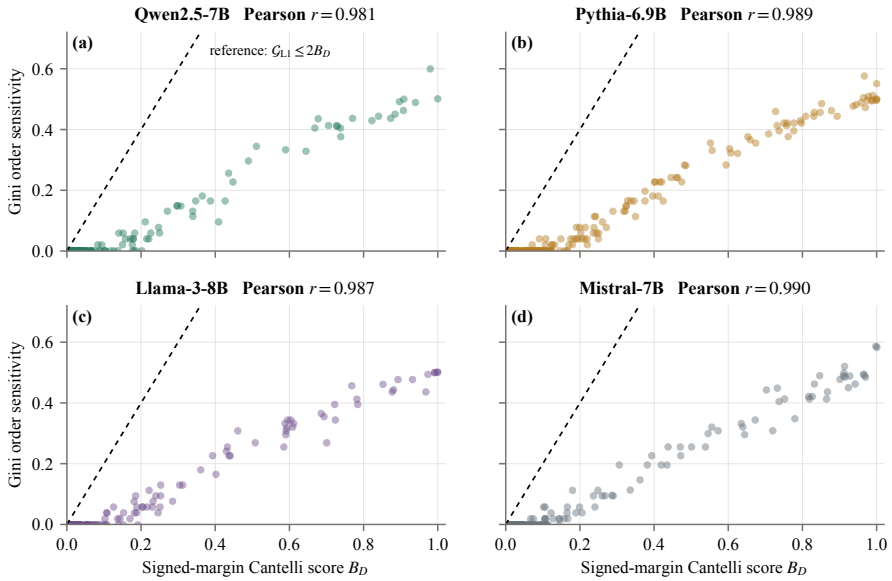

Figure 7: Signed-margin Cantelli score versus clean AGNews order sensitivity (200 queries, 100 orders per model), with Pearson r per panel. Dashed lines are the population reference $\mathcal{G}_{\mathrm{L1}} \leq 2B_D$; correlations use complete empirical order distributions.

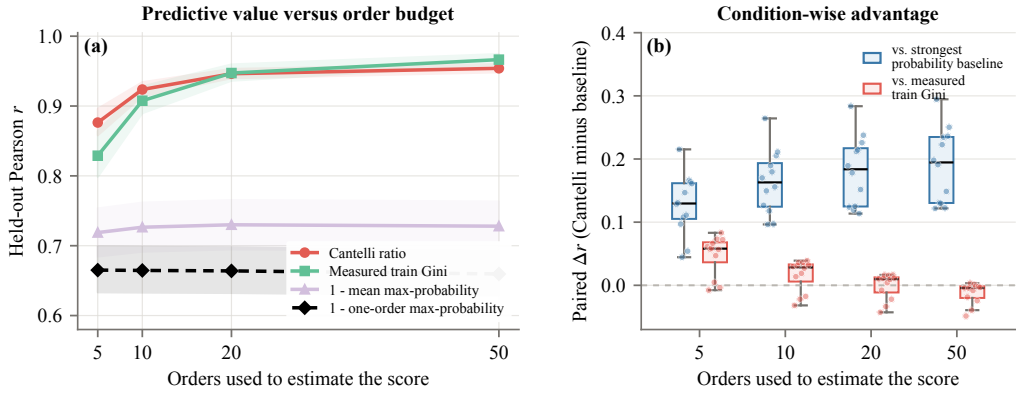

Figure 8: Held-out order-budget comparison across twelve collision-free conditions. The signed-margin Cantelli score is compared with direct prediction-histogram estimators and cheap probability-only diagnostics. It uniformly exceeds the probability-only baselines under the frozen Pearson endpoint; direct training Gini catches up as its order budget grows. Rank-correlation results are less uniform and are not used to claim universal estimator dominance.

The mean margin is positive for every query except one Mistral (Jiang et al., 2023) query. Loss variance omits changes in competing scores; the loss-variance $\times$ accuracy composite improves on it but remains below the direct margin quantity.

The audit in Section B.1 shows that the first-token evaluator duplicated shared logits before normalization, changing the class distribution rather than merely making decoding ambiguous. Deterministic coarsening alone can only reduce Gini (Lemma 2), but duplication adds a separate distortion. Complete label-to-token audits should therefore fail before inference unless the chosen representation is injective.

The collision-free Qwen/TREC-fine `sequence_mean` run has 200 queries, 100 orders, accuracy .526, and mean Gini .172. Held-out margin correlations at 5, 10, 20, and 50 estimation orders are .872, .930, .956, and .964; the 50-order split range is [.955,.973]. Loss variance and the loss-

variance × accuracy composite reach only .271 and .336. For Banking77, mean and summed sequence likelihood on the same 50 queries and 30 orders give correlations .883/.890 but change 35.7% of predictions. The relation is robust; absolute accuracy and order sensitivity are not scoring-rule invariant.

C.2 Prospective sizing forecast audit

The frozen, parameter-free forecast scales each inspected $S = 8$ count vector from the earlier $n = 24$ audit: $p = C^{(8)}/8$, $\widehat{C}^{(S)} = Sp$, and $q_k \propto \exp(\varepsilon \widehat{C}_k^{(S)}/2)$. A post-hoc executable transcription reproduces the nine frozen accuracy, plurality-fidelity, and privacy values to 10^{-12} from the committed source counts. The three Gini forecasts in Table 6 remain the values in the frozen file but are outside this claim because finite-design decomposition depends on the sampling grid. The transcription does not retroactively become part of the frozen protocol. The separate 72-record master run keeps $s = 3$ but resamples the pool; its $S = 8$ calibration directly audits the frozen transfer approximation without establishing transportability beyond this setting.

Table 6: Complete forecast audit. Signed error is observation minus prediction; the $S = 8$ row is a cross-design calibration check, whereas $S = 14, 24$ were unrun configurations when frozen. Predictions, observations, and signed errors are rounded independently; errors are computed from full-precision values.

S	Endpoint	Prediction → observation (error)
8	Accuracy	.232 → .237 (.005)
	Gini	.871 → .870 (-.001)
	Privacy	.869 → .867 (-.002)
	Fidelity	.331 → .335 (.004)
14	Accuracy	.543 → .550 (.007)
	Gini	.308 → .309 (.001)
	Privacy	.297 → .299 (.002)
	Fidelity	.807 → .808 (.002)
24	Accuracy	.619 → .614 (-.004)
	Gini	.085 → .066 (-.019)
	Privacy	.062 → .054 (-.008)
	Fidelity	.958 → .965 (.008)

Across all eight $S = 14/24$ forecasts (four endpoints per size, including the two frozen Gini entries excluded only from transcription parity), mean absolute error is .006. A declared post-hoc no-growth baseline that reuses each observed $S = 8$ endpoint for both larger sizes has mean absolute error .567. This descriptive comparison is not a second prospective test.

C.3 A failed within-subset lever and its diagnosis

The frozen C6 intersection–union test failed. Both DBpedia conditions pass all mechanism criteria, but neither TREC-fine condition establishes a gap increase, and Mistral/TREC-fine does not establish a privacy-component decrease. Accuracy non-inferiority holds in all four conditions but cannot override the failed joint rule. This test does not support a general improvement from within-subset consensus; unresolved directional tests do not establish ineffectiveness.

The post-outcome replay also refutes its motivating explanation. Mean within-subset modal fractions are already .850–.908, with strict majorities in 87.3–93.4% of subsets; soft aggregation changes few TREC votes and no tested rule meets the declared screen of at least two mechanism improvements in both TREC conditions together with the accuracy guard. A descriptive separator is cross-subset concentration: N_{eff} is only 7.0–8.4 on DBpedia but 23.9–24.5 on TREC-fine. Individual subsets are internally stable yet disagree with one another. The decomposition locates execution-level randomization, whereas N_{eff} exposes dispersion inside a fixed count vector. This post-hoc diagnosis motivates the separately frozen allocation study below; it is not itself a validated targeting rule.

C.4 FIXED-RECORD VOTER-ALLOCATION DETAILS

Table 7 reports the frozen TREC-fine decision at $\varepsilon = 1$. Both models pass both requirements: fidelity lower bounds exceed zero and accuracy lower bounds exceed the $-.02$ non-inferiority guard. Ensemble pre-release plurality accuracy nevertheless declines. The post-hoc accounting in Section C.5 explains how reduced release loss offsets that decline, and measures individual emitted-vote accuracy separately; the two pre-release metrics are not interchangeable.

Table 7: Frozen fixed-n allocation result on TREC-fine. Entries are paired $C-A$ differences with 95% query-bootstrap intervals; A is $(S, s) = (8, 3)$ and C is $(24, 1)$. Pre-release plurality accuracy was a mechanism diagnostic, not part of the joint decision.

Model	Released accuracy	Plurality fidelity	Pre-release accuracy
Mistral-7B-v0.3	+.036 [+.019,+.055]	+.251 [+.221,+.283]	-.131 [-.189,-.074]
Qwen2.5-1.5B	+.061 [+.042,+.081]	+.202 [+.177,+.229]	-.069 [-.122,-.014]

Table 8: Arm-level means at $\varepsilon = 1$, generated from the frozen analyzer output. "Pre" is pre-release plurality accuracy; "Tie" is the top-count tie rate; "Suff." is the per-partition fraction meeting the sufficient condition in Equation (10). Means are rounded independently; reported contrasts use full precision.

Condition	Arm	S	s	Acc	Pre	Fid	$\mathcal{G}$	$\mathcal{R}_{\text{priv}}$	Gap	N_{eff}	Tie	Suff.
Mistral / DBpedia	A	8	3	.309	.670	.344	.816	.806	2.333	7.0	.201	.000
Mistral / DBpedia	B	12	2	.329	.623	.380	.778	.766	2.422	6.1	.168	.002
Mistral / DBpedia	C	24	1	.365	.550	.431	.698	.698	2.770	5.4	.210	.025
Mistral / TREC-fine	A	8	3	.050	.306	.078	.974	.973	1.111	24.5	.348	.000
Mistral / TREC-fine	B	12	2	.057	.232	.115	.964	.961	1.399	19.2	.283	.000
Mistral / TREC-fine	C	24	1	.086	.175	.329	.821	.821	2.920	10.4	.145	.010
Qwen-1.5B / DBpedia	A	8	3	.525	.795	.565	.617	.609	4.690	8.4	.072	.000
Qwen-1.5B / DBpedia	B	12	2	.676	.771	.753	.364	.356	7.048	7.4	.043	.365
Qwen-1.5B / DBpedia	C	24	1	.664	.715	.837	.235	.235	9.970	4.4	.025	.495
Qwen-1.5B / TREC-fine	A	8	3	.051	.304	.078	.974	.973	1.091	23.9	.363	.000
Qwen-1.5B / TREC-fine	B	12	2	.068	.272	.125	.959	.956	1.534	19.0	.248	.000
Qwen-1.5B / TREC-fine	C	24	1	.111	.235	.280	.860	.860	2.300	10.4	.175	.000

The arm-level means in Table 8 show that this is a scoped channel-side validation, not a practical resolution of TREC-fine. Arm C reaches only .086/.111 released accuracy, and the fraction meeting the sufficient gap threshold $\Delta \geq 13$ is .010/.000. Mean gaps rise to 2.920/2.300 and N_{eff} falls to 10.4/10.4, but 24 one-example voters still scatter. Thus the experiment demonstrates consolidation without typical threshold attainment at this record budget.

Capacity staging differs by task and model. DBpedia first becomes feasible in principle at arm B ($S = 12$ against integer gap 10), whereas TREC-fine does so only at arm C ($S = 24$ against integer gap 13). Qwen/DBpedia follows this staging, with its largest fidelity step in $B-A$ and sufficient-condition fractions $.365/.495$ in arms B/C. Mistral/DBpedia improves only gradually and reaches a fraction of $.025$ at arm C, showing that existence of a feasible count vector does not imply that typical empirical profiles attain it.

Frozen secondary predictions are reported without changing the joint verdict. The prediction that the TREC-fine $B \to C$ fidelity step exceeds $A \to B$ holds for both models (Mistral: +.214 versus +.037; Qwen: +.155 versus +.047). The prediction that $C-A$ gains are larger on TREC-fine than DBpedia holds for Mistral (+.251 versus +.087) but reverses for Qwen (+.202 versus +.272); this mixed result is not reconciled post hoc. The third frozen prediction—fidelity gains accompanied by preserved pre-release plurality accuracy or an accuracy-guard failure—is contradicted in both TREC-fine models: plurality accuracy decreases with intervals entirely below zero while the released-accuracy guards pass. This secondary prediction was omitted from the initial outcome report and is restored here as a reporting correction, not reclassified as a successful prediction. The primary joint verdict is unchanged. All eight new arms contain 200 audited queries. The analyzer for the frozen primary verdict was executed once after panel closure and verified cross-arm identity and the decomposition identity to a maximum absolute residual of 6.20×10^{-16} (frozen tolerance 10^{-10}).

Table 9: Post-hoc paired $C-A$ changes at $\varepsilon = 1$ with descriptive 95% query-bootstrap intervals. Negative L and positive H changes increase utility. Values are rounded independently; differences use full precision. These intervals do not enter the frozen joint verdict.

Condition	Individual-vote accuracy V	Lost-correct mass L	Rescue mass H
Mistral / DBpedia	-.223 [-.245,-.202]	-.171 [-.211,-.131]	+.005 [-.002,+.013]
Mistral / TREC-fine	-.077 [-.094,-.060]	-.158 [-.208,-.109]	+.009 [+.004,+.013]
Qwen-1.5B / DBpedia	-.192 [-.220,-.165]	-.223 [-.250,-.195]	-.005 [-.012,+.003]
Qwen-1.5B / TREC-fine	-.059 [-.075,-.043]	-.117 [-.164,-.072]	+.012 [+.007,+.018]

C.5 Post-hoc utility accounting and decision audit

Utility accounting. Let $w(C)$ be the plurality label under the public tie rule, $A_{\text{pre}} = \mathbb{E}[\mathbf{1}\{w(C) = y\}]$, and $U = \mathbb{E}[q_y(C)]$. Splitting on $w(C) = y$ gives, for any categorical release kernel,

$$U = A_{\text{pre}} - \underbrace{\mathbb{E}[\mathbf{1}\{w(C) = y\}(1 - q_{w(C)}(C))]}_{L\text{:lost-correct mass}} + \underbrace{\mathbb{E}[\mathbf{1}\{w(C) \neq y\}q_y(C)]}_{H\text{:wrong-plurality rescue mass}}. \tag{16}$$

Expectations are over the declared count-generating design and queries. Utility is not generally A_{pre} times mean fidelity: correctness and fidelity can covary, and wrong-plurality releases contribute H. Rescue is an accounting term, not a noise recommendation or causal-mediation claim. Individual emitted-vote accuracy, $V = \mathbb{E}[C_y/S]$, is a different quantity.

Table 9 shows that V decreases in all four conditions. On TREC-fine, Mistral's plurality change $-.131$, minus its L change $-.158$, plus its H change $+.009$, yields $+.036$ released accuracy. Thus reduced release loss offsets worse upstream accuracy in these data. Table 8 reports absolute arm outcomes; the artifact also retains the complete $V, A_{\text{pre}}, L, H, U$ table. Means and contrasts are rounded independently; the identity uses full precision.

Compute cost. The $S(s!)$ Level-2 scoring calls per partition are 48, 24 and 24 for arms A, B and C, respectively; over five partitions they total 240, 120 and 120. Their observed unique-input ranges are 222–240, 96–118, and 24, identical across conditions. Full-domain sequence-mean scoring batches eight labels per call, requiring $\lceil K/8 \rceil$ forward batches; two additional Level-1 calls per query are audit diagnostics, not deployment requirements. Tokens and energy were not logged. Per-run elapsed metadata remain in the artifact but measure shared-server process time, potentially reset on resume, not controlled latency.

At $s = 1$, partitions merely rearrange the same singleton voters: all five count vectors coincide and $\mathcal{R}_{\text{ord}} = \mathcal{R}_{\text{part}} = 0$. The 120 logged calls comprise 24 unique inputs, permitting memoization; no cache speedup or proportional latency reduction is claimed because prompt lengths and label scoring differ.

Configuration selection: a negative result. We retrospectively simulate 100 fixed splits per condition, each with 100 development and 100 evaluation queries. Two plug-in policies scale development arm-A profiles to $C^{(S)} = (S/8)C^{(8)}$ and select A/B/C by predicted mean accuracy or fidelity; accuracy also uses development labels. Baselines are maximum S, minimum capacity-feasible S, and direct development-accuracy tuning over all three arms. Ties prefer smaller S; test outcomes never select configurations.

Both plug-ins select C in all 400 condition/splits and offer no gain over maximum S (Table 10). They miss B's accuracy advantage on Qwen/DBpedia, where direct tuning selects B in 73/100 splits. Minimum-feasible-S instead loses accuracy on Mistral/DBpedia. Capacity is not an optimality rule: A-profile scaling holds voter quality fixed, an assumption violated when s changes.

Plug-ins cost 240 Level-2 calls per development query versus 480 for all-arm tuning and zero for the fixed rules. This saves calibration calls but not utility. All original outcomes were known before split specification, and exemplar pools can overlap across queries. Deployment would require public development queries *and* exemplar pools (or additional private selection accounting), then fresh frozen evaluation; no adaptive DP selector is claimed.

Table 10: Retrospective mean held-out expected accuracy over 100 splits. Min-feasible selects B on DBpedia and C on TREC-fine; Plug-in-U/F optimize predicted accuracy/fidelity from arm A. Split averages are descriptive, not independent replications or evidence of public/private transfer.

Condition	Max-S	Min-feasible	Direct-dev	Plug-in-U	Plug-in-F
Mistral / DBpedia	.364	.330	.364	.364	.364
Mistral / TREC-fine	.087	.087	.087	.087	.087
Qwen-1.5B / DBpedia	.663	.675	.666	.663	.663
Qwen-1.5B / TREC-fine	.112	.112	.112	.112	.112

Shared post-hoc audit protocol. This section and Section C.6 reuse the frozen counts; neither is independent replication. The extractor verifies 12 raw/meta hash pairs, identities, exhaustive permutations and normalized scores, reconstructs consensus counts, and reproduces 72 frozen means within 10^{-12}. Equation (16) closes to the same tolerance. New paired intervals use 10,000 query resamples, seed 20260905 plus condition index, keeping all cells of a query together; split simulations use seed 20260905. Intervals are pointwise, not simultaneous tests, and overlapping exemplar pools limit generalization. Frozen primary intervals are unchanged. Tables are generated from audited JSON; complete arm-level accounting and timing records remain in the artifact.

C.6 Allocation depends on the release channel and budget

Replay protocol. We hold all predictions fixed across the 12 allocation arms (four conditions, 200 queries, five partitions). Deterministic consensus gives $\mathcal{R}_{\text{ord}} = 0$; singleton arm C also has $\mathcal{R}_{\text{part}} = 0$, without implying accuracy. The following complete grids were specified after the original outcomes but before replay, so this is post hoc, not an independent replication:

$$\text{Gumbel (privacy relaxes} \rightarrow\text{):} \quad \varepsilon \in \{0, .125, .25, .5, 1, 2, 4, 8, 16, \infty\},$$
$$\text{Gaussian (noise increases} \rightarrow\text{):} \quad \sigma \in \{0, .25, .5, 1, 2, 4, 8, 16, \infty\}.$$

Intervals use the shared post-hoc protocol in Section C.5, not the original seed for Table 7. Complete grids, all three pairwise contrasts and per-query outputs remain in the artifact; ordering the displayed set does not reorder stored grid indices.

Accuracy and fidelity can favor different allocations. C's accuracy gain at $\varepsilon = 1$ reverses at $\varepsilon = 8$ in all four conditions, although TREC-fine fidelity still favors C (Table 3 and Figure 9). Adjacent-grid mean sign changes fall between $\varepsilon = 1, 2$ on DBpedia and $2, 4$ on TREC-fine; these are not population crossover estimates. Mistral/TREC at $\varepsilon = 2$ and Qwen/TREC at $\varepsilon = 4$ remain unresolved.

The full three-arm audit also rules out fidelity as an accuracy selector: Qwen/DBpedia's best mean accuracy is B at $\varepsilon = 1, 2$, but its best fidelity is C. Mistral/DBpedia's fidelity ranking itself changes at low noise with tie structure. These are observed rankings, not deployment policies; the unsuccessful plug-in selectors (Table 10) miss the voter-quality change caused by varying s.

Gaussian sensitivity, separately accounted. We apply Gaussian RNM to the same counts using the quadrature formula in Section B.2. C's accuracy advantage is positive at $\sigma = 4, 8$ and negative at $\sigma = .5$ in all four conditions, with pointwise intervals excluding zero (Table 11). Mistral/DBpedia and Qwen/DBpedia at $\sigma = 2$, and Qwen/TREC at $\sigma = 1$, remain unresolved. The full Gaussian curves are retained in the artifact; this same-trace sensitivity check does not establish transfer to new data or private-ICL pipelines.

At $\sigma = .5/8$, analytic vector-Gaussian accounting (Balle & Wang, 2018) gives sufficient $\varepsilon = 13.884/0.524$ for $\delta = 10^{-4}$ and replacement sensitivity $\Delta_2 = \sqrt{2}$. The full mapping is in the artifact. Argmax inherits this guarantee by postprocessing; it is not tight RNM accounting, nor is equal numerical ε equivalent to pure-DP Gumbel. Releasing multiple replay answers or choosing mechanisms from protected diagnostics would require privacy accounting.

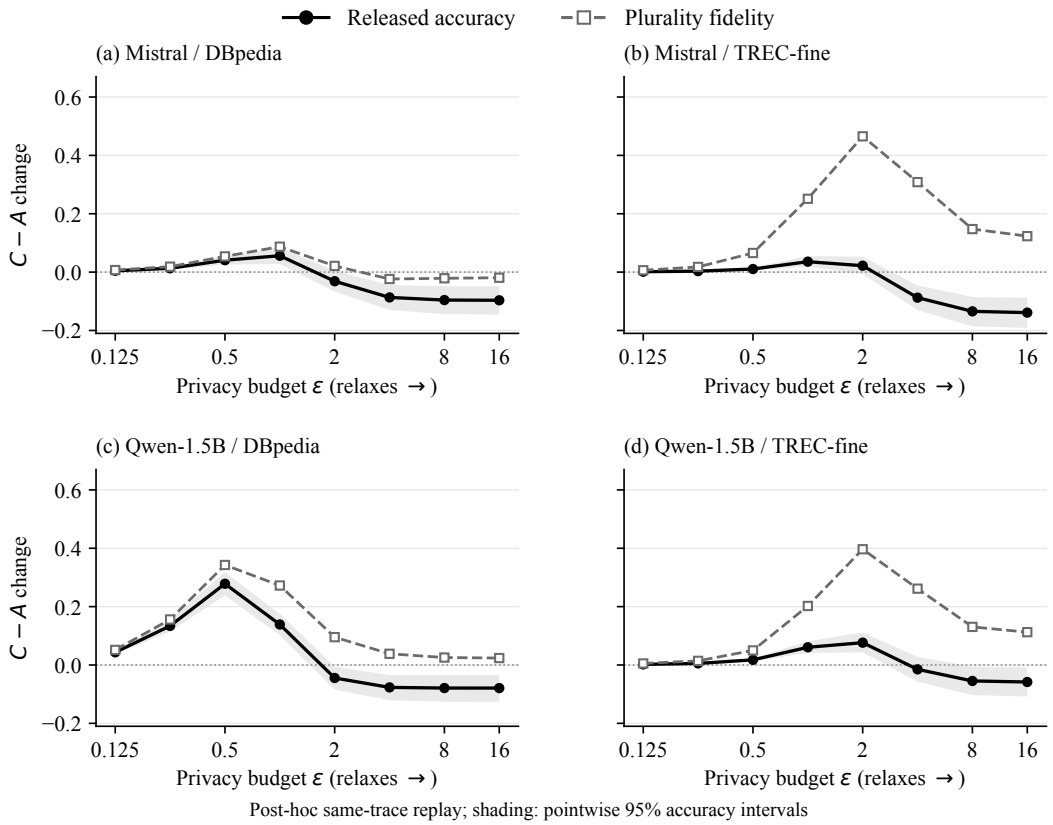

Figure 9: Post-hoc allocation replay: C's finite-budget accuracy benefit reverses as privacy relaxes, while TREC-fine fidelity remains higher. Curves show $C-A$ differences at fixed $n = 24$ without smoothing; shading is pointwise 95% accuracy intervals. Low-noise limits are reported separately in Table 12.

Table 11: Post-hoc Gaussian $C-A$ accuracy contrasts with pointwise paired 95% intervals. Noise SD decreases from left to right; these are not Gumbel-equivalent budgets. Complete curves and intermediate points remain in the artifact.

Condition	$\sigma = 8$	$\sigma = .5$
Mistral / DBpedia	$+0.026$ $[+0.016, +0.037]$	-0.094 $[-0.140, -0.049]$
Mistral / TREC-fine	$+0.010$ $[+0.006, +0.015]$	-0.128 $[-0.176, -0.081]$
Qwen-1.5B / DBpedia	$+0.223$ $[+0.191, +0.255]$	-0.079 $[-0.124, -0.036]$
Qwen-1.5B / TREC-fine	$+0.017$ $[+0.012, +0.022]$	-0.050 $[-0.096, -0.003]$

Tie-aware limits. At $\varepsilon = 0$ or $\sigma \to \infty$, accuracy and fidelity equal $1/K$ for every arm and contrasts vanish. At $\varepsilon \to \infty$ or $\sigma \to 0$, both channels select uniformly among maximal-count ties, unlike the fixed public plurality rule: accuracy can differ and fidelity tends to $1/m$ for tie size m (Table 12).

C's limiting accuracy is lower in all four conditions. Unlike the near-matched upstream accuracies of the 72-record Banking77 study (Figure 5), its accuracy contrast need not return to zero. Continuity and opposite-signed finite-budget/low-noise contrasts imply at least one sign change, not uniqueness or a universal transmission window.

Numerical audit. Frozen $\varepsilon = 1$ means and both utility/Gini identities agree within 10^{-12}. All 4,613 unique Gaussian profile/scale pairs passed the two-integral check specified in Section B.2; raw normalization residuals and between-integral differences are below 2×10^{-10}. Binary closed forms,

Table 12: Fixed-tie pre-release accuracy versus low-noise limits for arms A/C. $U_{\lim}$ averages over maximal-count ties; $F_{\lim}$ is fidelity to the fixed plurality label. Infinite limits are not plotted at finite coordinates.

	A_{pre}		$U_{\lim}$		$F_{\lim}$	
Condition	A	C	A	C	A	C
Mistral / DBpedia	0.670	0.550	0.653	0.556	0.888	0.868
Mistral / TREC-fine	0.306	0.175	0.305	0.166	0.801	0.924
Qwen-1.5B / DBpedia	0.795	0.715	0.798	0.719	0.963	0.986
Qwen-1.5B / TREC-fine	0.304	0.235	0.296	0.238	0.792	0.904

symmetry, label permutations and tie limits have separate tests. Dependencies, hashes and complete numerical diagnostics are recorded in the artifact. These checks establish numerical consistency, not new statistical or privacy guarantees.

C.7 Public taxonomy as a different deployment trade-off

The exploratory Qwen/TREC intervention uses 100 queries and the public TREC fine-to-coarse mapping. It compares 50-way release followed by the public map with aggregating the same counts into six public categories before release. Both arms are evaluated against the coarse target; the distinction is where aggregation occurs, not a comparison of fine-label accuracy against coarse-label accuracy. At fixed S and $\varepsilon = 1$, expected accuracy rises from .509 to .800 (paired gain .291 [.259, .322]), and plurality fidelity rises from .547 to .861 (gain .314 [.290, .339]). Native-channel N_{eff} changes from 34.162 to 4.471. Intervals use 10,000 paired query resamples, seed 20260813.

This changes the deployment objective relative to fine-grained classification: neither coarse output can recover the original 50-way answer. Deterministic mapping alone lowers Gini (Lemma 2), so that reduction is not evidence of improved model behavior. The result supports an exploratory choice of public output granularity, not a free utility improvement, prospective validation, or data-dependent restriction of private labels.

C.8 Synthetic finite-design estimator audit

Table 13 reports the audit defined in Section B.3. It sharply separates finite-design bias from bootstrap behavior: across 72 primary conditions, correction lowers absolute component bias in 90.0% of frozen cells; remaining ties are structural-zero cases rather than material reversals.

Table 13: Exactly enumerable synthetic audit. Bias is measured against oracle population components across 1,000 sampled $Q = 100$, $P = T = 5$ designs per condition. All frozen primary conditions are retained.

Estimator	Median absolute bias	95th percentile
Empirical order	.00762	.12457
Corrected order	.000009	.000235
Empirical partition	.01291	.10678
Corrected partition	.000027	.000327
Corrected order share	.000175	.000664

Partition share needs no separate audit row: whenever total upstream variation is positive, it is exactly one minus order share. The corrected-order-share row therefore certifies the complementary partition share under the same design.

Across the ten budgets, eligible nonzero endpoints total 480 order $+472$ partition $+368$ order-share $= 1{,}320$. Share eligibility requires oracle upstream variation $\geq .01$; positivity uses the disclosed numerical-zero convention. All 1,320 nonzero component/share endpoints attain empirical 95% query-bootstrap coverage in the frozen [.90,.98] range (mean .944). Exact-zero order and share endpoints have coverage 1.00, so the deliberately strict all-cell frozen verdict remains "bias-reducing but imperfect" rather than "well-calibrated." Across the 324 secondary cells, nonzero coverage falls

Table 14: Post-hoc crossover sensitivity on the same $P = T = 5$ count grids. Finite-grid roots use empirical components; corrected roots replace them with the untruncated finite-design correction of Equation (14) under its independent nested-sampling assumptions. Values are point estimates, not population confidence intervals. Bisection refines the first crossing on the continuous Gumbel channel; linear-interpolation errors are below .0012.

Condition	Finite grid	Corrected	Shift
Pythia/AGNews	3.668	3.467	-.201
Qwen/AGNews	3.164	2.968	-.196
Qwen/Banking77	4.162	3.998	-.164

Table 15: Matched bridge results with 10,000 query-cluster bootstrap resamples. AGNews pooled endpoints were frozen and its detection/severity split was post hoc. TREC detection/severity were prospectively frozen primary endpoints; pooled correlations were secondary. The Qwen-14B row is a prospective scale replication frozen after the 7B TREC outcomes were known.

Bridge endpoints

Task	Model	$\rho(D, \Delta)$	Detection AUC	Active ρ	n
AGNews	Qwen	.776 [.677,.840]	.854 [.764,.929]	.628 [.228,.854]	16
	Pythia	.570 [.420,.687]	.699 [.585,.804]	.256 [-.085,.558]	35
TREC	Qwen	.176 [-.040,.385]	.745 [.622,.860]	.045 [-.188,.280]	74
	Pythia	.272 [.058,.460]	.748 [.620,.864]	-.092 [-.316,.140]	78
	Qwen-14B	.171 [-.057,.391]	.630 [.520,.734]	-.041 [-.323,.250]	54

Decomposition diagnostics

Task	Model	Pooled $\rho(B_D, \mathcal{R}_{\text{ord}})$	Sync. share	Exact share
AGNews	Qwen	.460 [.314,.581]	.848	.719 [.630,.822]
	Pythia	.346 [.153,.520]	.844	.755 [.678,.830]
TREC	Qwen	.301 [.096,.483]	.773	.706 [.655,.756]
	Pythia	.214 [.007,.403]	.774	.696 [.650,.740]
	Qwen-14B	.194 [-.005,.384]	.694	.606 [.540,.670]

in the frozen range for 98.8–100% of cells, and corrected-component 95th percentile absolute bias remains at most .00173.

Forty synthetic populations have an oracle crossover within the search range; all 200 trials per population recover its observed/censored status. Mean absolute root error is .094 (median .069). Query-only bootstrap intervals cover the superpopulation crossover in none of these conditions. The resampling procedure holds each sampled partition/order grid fixed, and the empirical upstream function itself has finite-design bias. This negative result motivates the conditional interpretation in Section B.3; it does not invalidate the exact finite-design curves.

For independent P partitions and T orders per partition, $\mathbb{E}[\hat{\mathcal{R}}_{\text{ord}}^{\text{emp}}] = (1 - 1/T)W$ and $\mathbb{E}[\hat{\mathcal{R}}_{\text{part}}^{\text{emp}}] = (1 - 1/P)(B + W/T)$; hence pooled empirical upstream has expectation $(1 - 1/PT)W + (1 - 1/P)B$. At $P = T = 5$ this is $.96W + .80B$. Corrected point roots above are therefore a sensitivity analysis, not a retrospective replacement for the observed finite-grid crossings. Bootstrap coverage for a corrected population root remains unestablished.

C.9 Matched-bridge intervals and finite-design checks

In Table 15, exact-minus-drop-one share differences are -.002 [-.043,.043], -.007 [-.041,.029], -.009 [-.026,.009], -.005 [-.020,.010], and .008 [-.017,.032]. All intervals include zero; aggregate ratios retain inactive queries. The post-hoc independent-schedule shares are lower than the frozen *Sync.* estimand in all five conditions, e.g., Qwen/AGNews changes from .848 to .719, and Qwen-14B/TREC from .694 to .606. Independent combinations permit partial cancellation instead of synchronized vote co-movement. Both estimands are valid for their declared schedules; frozen bridge endpoints remain synchronized. Every exact share interval stays above one half.

Table 16: Post-hoc bridge target sensitivity, using the same 100 queries per condition. Each entry is Cantelli-score detection AUC (active-query count). The original frozen target uses six synchronized schedules and deterministic plurality; the other columns use the exact 6^8 independent-order convolution. The finite-budget column detects $\mathcal{R}_{\mathrm{ord}}(\varepsilon = 1) > .01$, an exploratory absolute Gini-component threshold, not a preregistered success criterion.

Condition	Sync. deterministic	Independent deterministic	Independent $\varepsilon = 1$
Qwen/AGNews	.854 (16)	.870 (21)	.882 (24)
Pythia/AGNews	.699 (35)	.757 (48)	.777 (50)
Qwen/TREC	.745 (74)	.870 (82)	.858 (81)
Pythia/TREC	.748 (78)	.856 (91)	.720 (81)
Qwen-14B/TREC	.630 (54)	.702 (69)	.745 (62)

At $\varepsilon = 1$, all 100 TREC queries in each model have strictly positive finite-budget order variation, making a literal $\mathcal{R}_{\mathrm{ord}} > 0$ detection AUC undefined; the exploratory $.01$ cutoff restores a nontrivial endpoint. The choice of cutoff matters: at $.05$, none of these TREC queries is active. The table tests scope robustness, not prospective replication or a universal finite-budget diagnostic. Partition intervals are arithmetic complements, not independent evidence.

Query-level partition distribution and tail robustness. Table 17 reports the full distribution rather than only the aggregate share. Exact reconstruction produces structural zeros but no negative partition estimate in any condition. Removing the upper and lower 10% of query-level partition components yields the trimmed share; its larger change on AGNews shows that the positive aggregate there is more concentrated in a right tail than on TREC.

Table 17: Post-hoc exact-order partition audit. Five-number summaries are min/Q1/median/Q3/max of query-level corrected partition components. "Full / trim" gives aggregate partition share before and after symmetric 10% tail trimming.

Condition	Zeros	Five-number summary	Full / trim share
AGNews/Qwen	79/100	.000/.000/.000/.000/.311	.281/.079
AGNews/Pythia	52/100	.000/.000/.000/.007/.297	.245/.109
TREC/Qwen	17/100	.000/.002/.038/.111/.428	.294/.213
TREC/Pythia	9/100	.000/.009/.063/.172/.572	.304/.245
TREC/Qwen-14B	31/100	.000/.000/.008/.147/.474	.394/.301

The separate nested $P = T = 5$ designs give corrected aggregate order shares .802 (Pythia/AGNews), .819 (Qwen/AGNews), and .682 (Qwen/Banking77). They are retained as an independent finite-design check rather than pooled with the matched estimand.

As a post-hoc baseline audit, Level-1 empirical Gini and modal disagreement predict the same Level-2 activity target. Cantelli improves over Level-1 Gini on Qwen-7B/TREC by .120 $[.057, .186]$ AUC and on Pythia/TREC by .071 $[.020, .124]$; differences are unresolved in the other three matched conditions. The result supports incremental detection value in the prospective 7B task, not universal dominance. Five-partition severity-target intraclass correlation coefficients (ICCs) are .621–.708 except for the small active Qwen/AGNews subset (.313, $n = 16$). As a separate post-hoc diagnostic, empirical stable-wrong rates range from 1.0–11.0% at Level 1 and 5.0–19.0% at Level 2 across the five matched conditions. Stability is therefore not synonymous with correctness.

Continuous strength/severity diagnostics appear in Figure 10; activity distributions and receiver-operating characteristic curves are in Figure 11. At the post-hoc 25% audit budget, condition-specific Cantelli screens have recall 28.2–68.8% and precision 44.0–88.0%; complete threshold sweeps remain in the artifact and are not deployment rules.

C.10 Channel diagnostics and model-gap attenuation

C.10.1 Gap attainability, ties, and subset thinning

The exact envelopes in Figure 4 show why both gap and competitor mass matter. For $K = 77$ and $\alpha = .1$, Equation (10) gives $S_{\min} = 27/53$ at $\varepsilon = .5/.25$, or 81/159 records with three-example

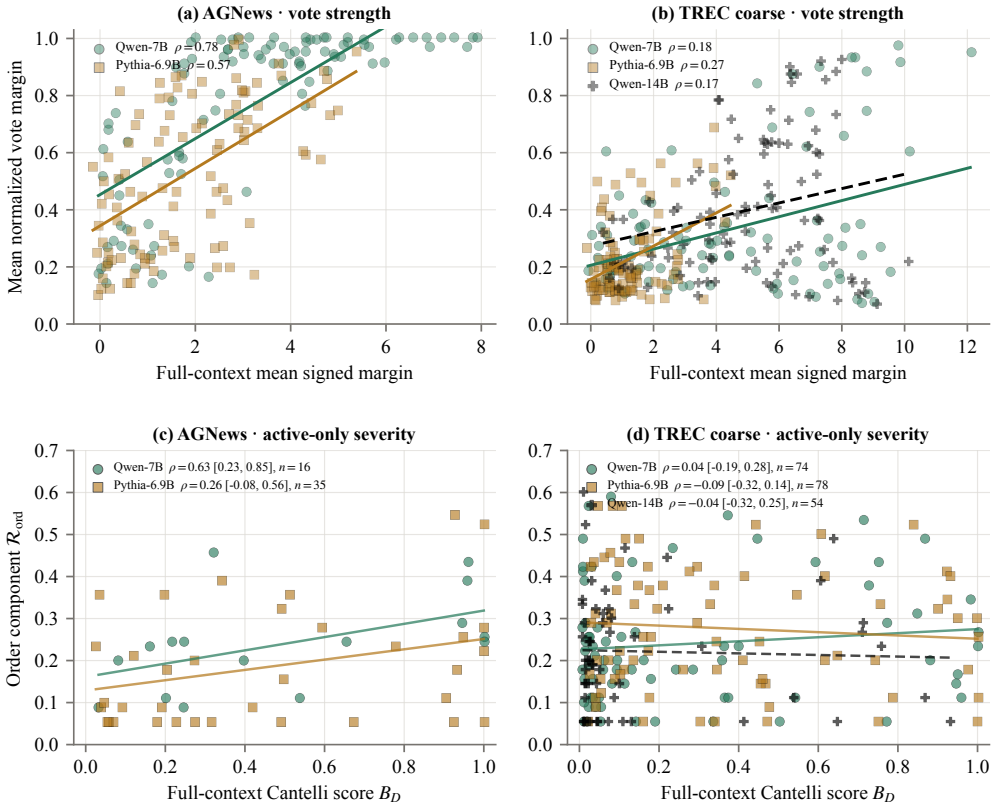

Figure 10: Per-query diagnostics for the two continuous matched-bridge endpoints. Top: Level-1 mean signed margin versus Level-2 mean normalized vote margin over all queries. Bottom: full-context Cantelli score versus deterministic $\mathcal{R}_{\text{ord}}$ among active queries only. Lines are descriptive linear fits. Vote strength is task-dependent, while active-only severity does not prospectively replicate on TREC coarse.

subsets. Even unanimous votes cap winner probability at 8.9%/3.5%. Indeed, 1,256 of the 2,000 logged $S = 8$ Banking77 vectors are unanimous ($\Delta = 8$), yet none can meet the required $\Delta = 53$. For TREC coarse ($K = 6, \varepsilon = 2$), the integer threshold is four and is met by 28.0% of Qwen and 15.0% of Pythia vectors; this post-hoc audit measures attainability, not necessity.

Top-count tie rates are 9.60% for Pythia/AGNews, 1.60% for Qwen/AGNews, 6.25% for Qwen/TREC-fine, and 3.45% for Qwen/Banking77, producing privacy floors .0483, .0080, .0325, and .018, respectively. The deterministic baseline resolves these ties and differs from the limit through finite budgets. For the matched nested crossover designs, the tie floors are checked against the tie-aware finite-budget upstream limits in Proposition 3. The respective fixed-rule deterministic upstream values are .1695, .0264, and .0522; they are not used in the sufficient-condition check.

The exhaustive subset-thinning diagnostic enumerates every retained set of four or six votes from each logged eight-vote cell. At $\varepsilon = 2$, reducing $S = 8$ to $S = 4$ increases release randomness by .148–.238 and changes expected accuracy by -.144– -.045. First crossovers span 5.231–5.405 for $S = 4$ and 4.124–5.428 for $S = 8$. Because the retained-subset choice expands the partition variable and reduces the effective pool, this isolates channel sensitivity to logged vote removal, not a causal subset-size effect.

For Qwen/AGNews, 48 of 5,000 crossover bootstrap samples are right-censored at $\varepsilon = 16$; the other conditions have none. This censoring concerns sampled bootstrap grids, not the empirical sufficient check in Proposition 3.

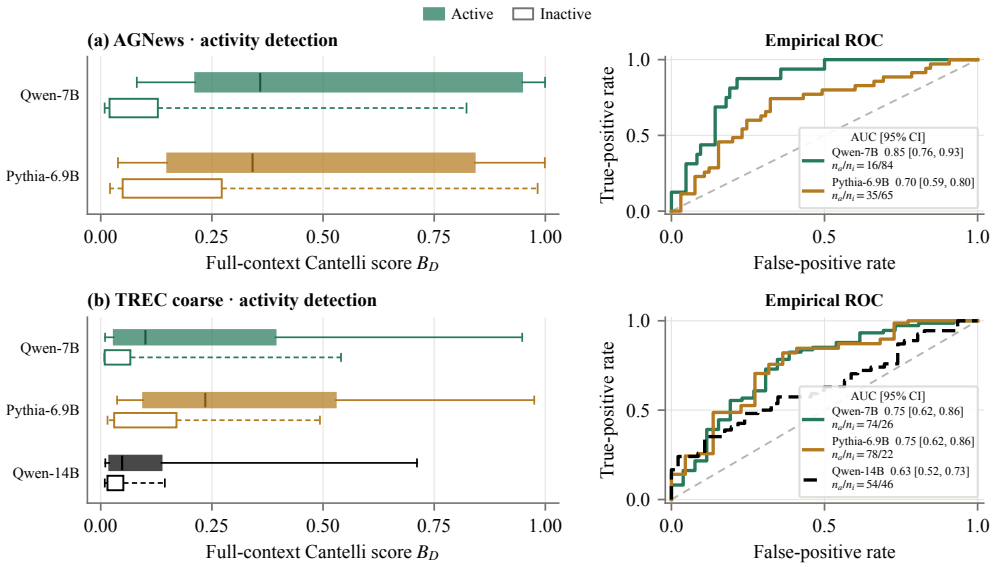

Figure 11: Activity-detection diagnostics for the zero-inflated order component under the original synchronized-order, deterministic-plurality target. Left: full-context Cantelli-score distributions for active ($\mathcal{R}_{\text{ord}}^{\text{det,sync}} > 0$, filled) and inactive ($\mathcal{R}_{\text{ord}}^{\text{det,sync}} = 0$, open) queries, summarized by interquartile boxes and 5th–95th percentile whiskers. Right: empirical receiver-operating-characteristic (ROC) curves with AUC, 95% bootstrap intervals, and active/inactive counts.

C.10.2 Descriptive model-gap attenuation

On the same 100 AGNews queries, Pythia exceeds Qwen in deterministic Gini by .1432 [.1030,.1855], split into .0986 order and .0446 partition variation. Adding standardized mean vote margin and polynomial terms attenuates the model coefficient by 85.6–99.5% across budgets. Residuals remain and reverse at high budgets, so this is descriptive attenuation rather than causal mediation.