FROM ROW-LEVEL TO TABLE-LEVEL: ZERO-SHOT PRIVATE ATTRIBUTE INFERENCE IN TABULAR DATA WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Withholding sensitive columns is a common precaution when sharing tabular data, yet the remaining attributes may still support inference of the hidden values. We study this privacy risk under a zero-shot adversary that observes the released dataset and uses pretrained large language models (LLMs), without receiving ground-truth examples of the target attribute. The attack is evaluated under different levels of feature disclosure and with or without aggregate target proportions. Beyond direct LLM predictions for individual samples, our analysis asks whether statistical dependencies across the dataset can enable additional attribute recovery. We develop an evidence-based pseudo-context strategy that uses the LLM's scores to select pseudo-labeled samples for a tabular learner, deriving a separate context budget for each predicted class without ground-truth target labels. On demographic and health survey data, direct LLM inference is competitive with zero-shot baselines, and the table-level extension improves average accuracy and macro recall over direct prediction. These findings demonstrate how pretrained semantic knowledge can supply supervision for learning from an unlabeled dataset, exposing a source of privacy risk beyond independent sample-level inference.

1 INTRODUCTION

Tabular data are widely used across research and industry (Borisov et al., 2024), supporting applications in demographic analysis (Ding et al., 2021), finance (Khandani et al., 2010), healthcare (Shickel et al., 2018), and AI-driven scientific discovery (Hollmann et al., 2025). When these data describe individuals, they may contain sensitive attributes such as income, occupation, marital status, and place of birth (Fung et al., 2010; Staab et al., 2024). To make the data available for analysis while limiting disclosure, a publisher may remove selected sensitive columns entirely and share only the remaining attributes. However, removing a column does not necessarily remove the information needed to infer its values: visible attributes may remain related to the hidden target, and an attacker may exploit those relationships together with auxiliary knowledge (Machanavajjhala et al., 2007). The resulting privacy risk therefore depends on what the released dataset reveals and what the attacker's models already know about the hidden attribute and its relationships with visible fields.

The ability of pretrained large language models (LLMs) to interpret textual cues has already raised privacy concerns. Staab et al. (2024) study eight personal attributes and show that LLMs can infer information such as location, income, and sex from user-written text, highlighting the privacy implications of interpreting what people write.[1] Separately, TabLLM (Hegselmann et al., 2023) shows that tabular rows can be serialized into natural-language descriptions and used for zero-shot and few-shot classification with LLMs. This provides a route for applying knowledge encoded in language models to tabular samples through the meanings of their fields and values. Together, these findings motivate a privacy question: can an LLM use the visible attributes of a tabular sample to infer a private attribute that the publisher has withheld, without access to labeled target examples?

[1] In addition to their main study of text, Staab et al. (2024) report zero-shot attribute inference on an American Community Survey (ACS) dataset in Appendix D. We build on this evidence and examine how access to an entire table can support further inference from the released cohort.

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.

In this paper, we study *zero-shot private attribute inference in tabular data* when the attacker observes a released dataset but receives no ground-truth examples of the hidden target. Unlike ordinary zero-shot classification with opaque class identifiers, the target and candidate values have known meanings that the attacker can use. We first evaluate *row-level inference*, where the LLM predicts each sample independently, and then a *table-level extension* that constructs pseudo-labeled context from predictions across the table. Their difference measures the additional recovery enabled by the released cohort beyond independent semantic predictions. Figure 3 in Appendix C.1 illustrates this distinction with a held-out education sample.

The LLM uses field and value meanings to produce initial predictions, while a statistical learner models dependencies between visible features and the resulting pseudo-labels. We use E0 predictions as pseudo-labels, following the broader pseudo-labeling paradigm (Xie et al., 2020; Sohn et al., 2020), and derive an adaptive per-class context budget from the LLM scores. The released cohort supplies both the samples being attacked and the pseudo-labeled examples used to learn dataset-specific relationships, allowing the same construction to work with different supervised tabular learners.

Following Staab et al. (2024), we study eight demographic, socioeconomic, and geographic targets. The *strict*, *weak*, and *prior* settings progressively expose other private attributes or aggregate target proportions as an additional source of population-level information.

We evaluate this threat on 2024 ACS data from four states and the 2024 National Health Interview Survey (NHIS). Direct LLM inference is competitive with feasible zero-shot baselines, and table-level inference yields further but non-universal recovery. The gains depend on pseudo-context quality and coverage; aggregate proportions can additionally guide context allocation without changing sample assignments.

Our main contributions are threefold.

- We systematically study zero-shot private-attribute inference across eight targets, two datasets, and multiple LLM families. The results characterize how feature exposure and aggregate target information affect recovery of the hidden attributes.
- We separate direct semantic inference from table-level statistical transfer. By comparing the two on identical LLM scores, we identify when the unlabeled cohort adds useful evidence or instead propagates errors through the released dataset.
- We introduce an evidence-based pseudo-context strategy that derives label-specific budgets from LLM scores without target labels. It supports table-level statistical transfer and can incorporate released aggregate proportions into context allocation.

2 Related Work

2.1 Private-attribute inference from released information

Classical privacy research shows that suppressing explicit identifiers does not prevent disclosure when released attributes can be combined with background knowledge (Sweeney, 2002; Machanavajjhala et al., 2007). Attribute inference makes this risk explicit by predicting a hidden sensitive value from available information. Such predictability can arise from population-level statistical dependencies rather than information specific to a trained model (Jayaraman & Evans, 2022); even a limited collection of released aggregate statistics can support attribute inference (Mao et al., 2026).

Pretrained LLMs introduce another source of inference through semantic and world knowledge. Staab et al. (2024) infer personal attributes such as location, income, and sex from user text without task-specific training and also report experiments on structured ACS samples. SynthPAI provides a controlled benchmark based on synthetic Reddit-style comment threads and profile-level collections of free-form text (Yukhymenko et al., 2024), while inference-based evaluation of text anonymization shows that masking explicit PII can still leave subject attributes recoverable from the remaining context (Oh et al., 2026). Our attacker instead observes a structured dataset whose target column is hidden for every sample. We separate individual feature disclosure from aggregate target statistics and ask whether predictions across the unlabeled dataset reveal statistical information beyond independent LLM inference under this threat model.

2.2 Zero-shot tabular prediction and pseudo-supervision

LLMs can use feature and label meanings for tabular prediction without labeled examples from the target dataset. TabLLM serializes each sample as text and demonstrates zero- and few-shot classification from pretrained language knowledge (Hegselmann et al., 2023). Recent methods elicit that knowledge in other forms: LLM-generated decision trees provide executable predictors without training data (Knauer et al., 2025), while ProtoLLM generates class-specific feature values and constructs example-free prototypes (Wang et al., 2025). LLMTabBench evaluates zero- and few-shot LLMs on general binary classification tasks (Grushina et al., 2026). It does not study privacy disclosure or additional inference from an unlabeled cohort. Our direct row-wise attack instead evaluates semantic knowledge under this privacy threat model across multiclass and open-vocabulary targets.

We further use zero-shot predictions as pseudo-labels for learning feature–target dependencies in the released dataset. Conventional pseudo-labeling methods such as Noisy Student and FixMatch begin with a predictor trained on ground-truth examples from the same task (Xie et al., 2020; Sohn et al., 2020); our initial labels come only from zero-shot LLM inference. Label-free tabular methods such as TableDC can instead learn representations and clusters from the dataset itself (Rauf et al., 2025), but the resulting clusters do not identify which semantic value of the hidden attribute they represent. We use LLM predictions to supply this connection, select pseudo-context without target labels, and test whether a supervised tabular learner can recover additional information from the released dataset as a source of statistical evidence.

3 Problem Setting and Threat Model

3.1 Problem setting and adversary access

Let $X = (\boldsymbol{x}_i)_{i=1}^n$ denote a released tabular dataset containing n samples, where $\boldsymbol{x}_i \in \mathcal{X}$ contains the visible attributes of individual i. The corresponding values $\boldsymbol{y} = (y_i)_{i=1}^n$ of a private target attribute are withheld, with $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the space of possible attribute values. The attacker seeks an estimate $\hat{y}_i \in \mathcal{Y}$ for each sample using the dataset and pretrained models. For a *closed-set target*, the candidate values are specified in advance and shared across samples, with $|\mathcal{Y}| = C$. For an *open-vocabulary target*, $\mathcal{Y}$ is too large to enumerate as a shared candidate set, and valid predictions are expressed as attribute values rather than selected from a fixed label list. We study location (LOC) in this form, where the hidden values are place names drawn from a large geographic space and are represented by canonical geographic entities for subsequent inference and evaluation.

The attacker observes the entire released dataset while receiving no samples paired with ground-truth values of the hidden target for fitting or tuning the attack. Each LLM prompt describes the visible feature values and the target attribute, together with candidate-label meanings when a label list is supplied (see Appendix B.1 for prompt examples). Aggregate target proportions are included only under the *prior* condition described in Section 3.2. We use $\mathcal{K}$ to denote the supplied semantic descriptions and, when disclosed, aggregate target proportions.

3.2 Disclosure conditions and inference scope

Disclosure conditions. We compare *strict* and *weak* releases to investigate whether exposing some private attributes makes others easier to infer from the remaining released features.

For each target, the *strict* condition withholds the target and restricts the other private fields according to the target-specific release policy. The *weak* condition keeps the target hidden but exposes a selected subset of the other attributes that also serve as private-inference targets in our study, giving nested feature sets $F_{\text{strict}} \subseteq F_{\text{weak}}$. The *prior* condition uses the *weak* feature set and additionally supplies approximate target proportions $\boldsymbol{\rho} = (\rho_c)_{c \in \mathcal{Y}}$ in the prompt. Table 1 summarizes these conditions, and Appendix A lists the visible fields for every target and release condition used in the experiments.

Table 1: Disclosure conditions for a fixed hidden target.

Condition	Visible features	Released target proportions
Strict	F_{strict}	Unavailable
Weak	F_{weak}	Unavailable
Prior	F_{weak}	Approximate ρ

Inference scope. A row-level attack predicts $\hat{y}_i = f(\boldsymbol{x}_i; \mathcal{K})$ independently for each sample. A table-level attack instead produces $\hat{\boldsymbol{y}} = G(X; \mathcal{K})$, allowing information derived across samples to affect each estimate. Our attack begins with independent LLM predictions and then selects pseudo-labeled samples from X as context for a tabular learner; the second stage therefore extends the attack from row-level semantic inference to table-level statistical inference.

4 From Instance-Level Predictions to Table-Level Inference

The attack uses independently generated LLM predictions to label selected examples from the released dataset. These examples connect target-label meanings to the visible features and allow a tabular learner to model their statistical dependencies.

4.1 Zero-shot semantic inference

We begin by asking an LLM to infer the hidden attribute independently for each sample from its visible features. For each sample $\boldsymbol{x}_i$, we convert its visible fields and values into text and provide the LLM with the target description and candidate-label meanings. The LLM assigns a numerical score u_{ic} to each candidate $c \in \mathcal{Y}$, indicating how well that value fits the sample.

Then these scores are normalized and the highest-scoring candidate is taken as the initial prediction. For each $i \in \{1, \ldots, n\}$, we define

$$s_i(c) \equiv s_{ic} = \frac{[u_{ic}]_+}{\sum_{d \in \mathcal{Y}} [u_{id}]_+}, \qquad \hat{y}_i^{(0)} = \arg\max_{c \in \mathcal{Y}} s_{ic}, \tag{1}$$

where $[a]_+ = \max(a, 0)$. The notation s_{ic} identifies an entry of the closed-set score matrix, while $s_i(c)$ treats the same quantity as a score assigned to candidate attribute value c. We refer to this direct row-level inference as E0. The normalized scores represent the LLM's relative assessments of the candidate labels. E1 subsequently uses these same scores and initial predictions to construct pseudo-context for learning statistical dependencies across samples.

4.2 Evidence-based pseudo-context construction

The direct predictions from E0 provide initial estimates of the hidden attribute, drawing on knowledge the LLM may have acquired during pretraining. To use them as supervision for learning statistical dependencies in the released dataset, we must decide which samples enter the pseudo-context and how many to draw from each predicted class using only the LLM's output.

In the zero-shot setting, the attacker has no ground-truth target labels with which to tune the amount of pseudo-context used for each class. A fixed per-class quota requires choosing a budget in advance and assigns the same number of pseudo-labeled samples to classes with different levels of semantic support. We derive separate budgets from the observed class-wise score distributions; together, these budgets fix the size of the context supplied to the tabular learner.

Arrange the scores as an $n \times C$ matrix, with samples as rows and candidate labels as columns. A row-wise comparison of $s_{i1}, \ldots, s_{iC}$ determines which label best fits sample i and gives its E0 prediction. A column-wise comparison of $s_{1c}, \ldots, s_{nc}$ instead ranks samples by their support for label c. Among samples assigned to class c, this ranking determines which samples should enter the context first. The score distribution must additionally determine how many of them to select.

Background-adjusted evidence. A high absolute score alone may reflect a general preference for a class: a score that is typical across the dataset provides little basis for selecting one sample over another. We therefore compare each score with the mean score for the same class and normalize its positive excess by the remaining score range:

$$b_c = \frac{1}{n} \sum_{i=1}^{n} s_{ic}, \qquad \theta_{ic} = \frac{[s_{ic} - b_c]_+}{\max\{\varepsilon, 1 - b_c\}}, \tag{2}$$

where $\varepsilon = 10^{-12}$ ensures a positive denominator. Here, b_c summarizes the model's background tendency toward class c on the released dataset. The normalized excess $\theta_{ic} \in [0, 1]$ measures how

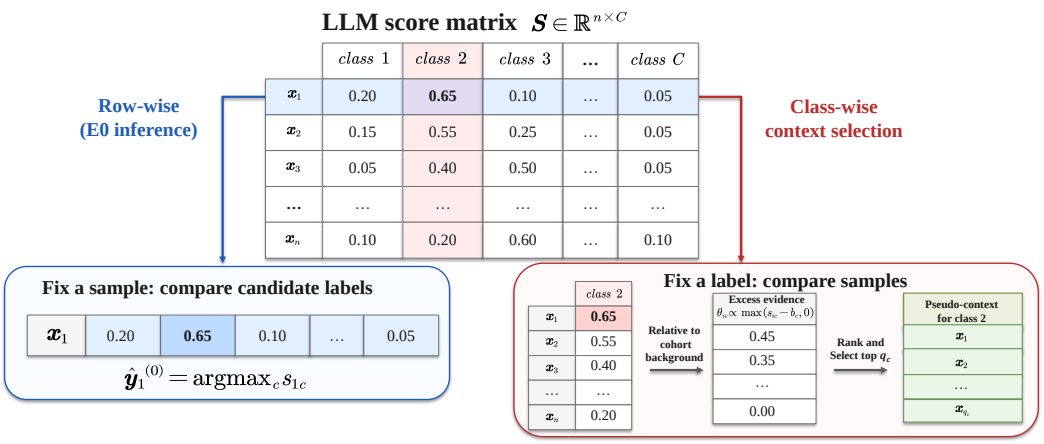


Figure 1: Two uses of the LLM score matrix. The entry s_{ic} is the normalized LLM support for value c on sample i. E0 compares these scores within a row. For context selection, b_c denotes the mean score for value c across samples, and θ_{ic} measures the positive support of sample i above this background. Samples are ranked by θ_{ic} within each predicted value, and q_c is the resulting number selected for that value.

strongly a sample supports that class above this background. For example, a score of 0.8 has little excess when the class mean is 0.79, but substantially more when the mean is 0.4.

Adaptive context allocation. We retain the original semantic prediction as the pseudo-label and partition the samples into disjoint sets $B_c = \{i \in \{1, \dots, n\} : \hat{y}_i^{(0)} = c\}$. To account for both the number of supporting samples and the strength of their excess scores, we sum the evidence within each predicted class and round it to obtain a selection budget:

$$M_c = \sum_{i \in B_c} \theta_{ic}, \qquad q_c = \left\lfloor M_c + \tfrac{1}{2} \right\rfloor. \tag{3}$$

Let S_c contain the indices of the q_c highest-ranked samples in B_c, ordered by decreasing θ_{ic}. We form $S = \bigcup_{c \in \mathcal{Y}} S_c$ and the pseudo-labeled context $\mathcal{C}_S = ((\boldsymbol{x}_i, \hat{y}_i^{(0)}))_{i \in S}$. This combines row-wise label assignment with class-wise context allocation: E0 determines each pseudo-label, while accumulated excess determines how many samples represent it. For a fixed class, θ_{ic} is monotone in s_{ic}, so background adjustment changes the budget while preserving the original score ordering.

Figure 1 summarizes the row-wise prediction and class-wise context construction defined above. Appendix F.2 evaluates the empirical quality of the class-wise ranking in Figure 5.

Prior-aware allocation. Under the *prior* condition, the disclosed target proportions ρ_c change only the number of context samples drawn from each E0 bucket. Let $q_c^{(0)}$ denote the budget in Equation 3, $Q = \sum_c q_c^{(0)}$, and $\bar{\theta}_c = |B_c|^{-1} \sum_{i \in B_c} \theta_{ic}$ for a nonempty bucket. We reallocate the fixed total Q using

$$\omega_c = \rho_c \bar{\theta}_c, \qquad q_c^{\text{prior}} \approx Q \frac{\omega_c}{\sum_d \omega_d}, \qquad 0 \le q_c^{\text{prior}} \le |B_c|, \tag{4}$$

with an integer, capacity-constrained apportionment detailed in Appendix C.1. The revised budgets select the highest-ranked samples within each unchanged E0 bucket B_c.

The accumulated evidence M_c is also the expected context size under an auxiliary Bernoulli selector with inclusion probabilities θ_{ic}. Rounding it gives the nearest deterministic integer budget; Appendix C.2 provides the derivation. This interpretation does not treat θ_{ic} as a calibrated probability that the associated pseudo-label is correct in the selected context.

4.3 TABULAR INFERENCE FROM PSEUDO-CONTEXT

The selected pseudo-context $\mathcal{C}_S$ pairs visible feature vectors with E0 predictions. A supervised tabular learner can use these pairs to model dependencies between the features and pseudo-labels. Let $\mathcal{A}$ denote its learning algorithm and $f_S = \mathcal{A}(\mathcal{C}_S)$ the resulting predictor. We instantiate $\mathcal{A}$ with TabPFN (Hollmann et al., 2025), which applies pretrained tabular inference to the selected context.

Closed-set targets. Let $\mathcal{Y}_S = \{\hat{y}_i^{(0)} : i \in S\}$ be the labels represented in $\mathcal{C}_S$ and let $m_S = |\mathcal{Y}_S|$. The predictor returns a probability vector over these labels:

$$\boldsymbol{t}_i = f_S(\boldsymbol{x}_i) = (t_{ic})_{c\in\mathcal{Y}_S} \in \mathbb{R}^{m_S}, \qquad t_{ic} \geq 0, \qquad \sum_{c\in\mathcal{Y}_S} t_{ic} = 1. \tag{5}$$

We use the tabular output as a relative compatibility correction to the semantic score. For each represented class, its tabular support is normalized by the uninformative uniform baseline $1/m_S$:

$$w_i(c) = \begin{cases} m_S t_{ic}, & c \in \mathcal{Y}_S, \\ 1, & c \notin \mathcal{Y}_S, \end{cases} \qquad a_i(c) = s_i(c) w_i(c). \tag{6}$$

Thus, $w_i(c) = 1$ leaves the E0 score unchanged, whereas values above or below one respectively strengthen or weaken a candidate according to table-level structure. A class absent from the context also receives the neutral factor, since the tabular learner provides no evidence about it. Without introducing an additional fusion coefficient, the final prediction is

$$\hat{y}_i^{(1)} = \begin{cases} \hat{y}_i^{(0)}, & i \in S, \\ \arg\max_{c\in\mathcal{Y}} a_i(c), & i \notin S. \end{cases} \tag{7}$$

Open-vocabulary targets. When the target space is too large to enumerate, the prompt requests a fixed-length top-k list and one score for each candidate instead of assigning a score to every possible label. After canonicalization, let $\mathcal{Z}_i$ be the sample-specific candidate set, $k_i = |\mathcal{Z}_i| \leq k$, and $s_i(z)$ the normalized semantic score of $z \in \mathcal{Z}_i$. Because these sets differ across samples, context selection first maps their scores to shared support units on which cohort-level background correction can be applied. The tabular stage then assigns a normalized compatibility $t_i(z)$ to each $z \in \mathcal{Z}_i$. Using the uniform baseline $1/k_i$ gives the same rule:

$$w_i(z) = k_i t_i(z), \qquad a_i(z) = s_i(z) w_i(z). \tag{8}$$

For samples outside the selected context, E1 chooses the candidate maximizing $a_i(z)$; context samples retain their selected pseudo-values. We evaluate this extension on LOC, the hidden location target introduced in Section 3.1. The prompt requests five locations per sample, while the union of candidate sets across the dataset can be larger. Place names are mapped to coordinates, and spatial support above the regional background replaces θ_{ic} when selecting S_{loc}. A TabPFN regressor supplies candidate compatibilities; Appendix C.3 specifies the spatial representation and context rule for this target.

We refer to the combined attack as E1. If no tabular predictor can be formed, E1 returns E0; fallback and tie-handling rules are detailed in Appendix C.1.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

5.1.1 DATASETS AND TARGETS

We use 2024 person-level ACS data from four states[2] and the 2024 NHIS Sample Adult file.[3] ACS covers eight targets and NHIS provides five aligned targets. Each test set contains 1,200 held-out samples; the *strict* and *weak* conditions use matched samples. Table 2 summarizes the consolidated target spaces and labeled training sets available only to supervised references. Complete construction details and feature policies appear in Appendix A.

[2]Official ACS page

[3]Official NHIS page

Qwen3-14B is the main LLM (Qwen Team, 2025). Meta-Llama-3.1-8B-Instruct provides cross-model evaluation (Llama Team, 2024), and DeepSeek-R1-Distill-Qwen-14B supports a targeted analysis of explicit reasoning traces (DeepSeek-AI, 2025). GPT-4.1[4] provides a closed-source comparison on ACS New York and NHIS. All checkpoints predate the 2024 public-use ACS and NHIS releases evaluated here, reducing direct exposure to these records during pretraining.

5.1.2 Evaluation protocol and metrics

E0 and E1 are evaluated on the same test set and share the same LLM outputs. Closed-set targets are assessed using accuracy and macro recall. For the ordered AGE and SCH targets, let $r(y)$ denote the rank of level y. The supplementary measures are mean absolute level error $n^{-1}\sum_i |r(\hat{y}_i) - r(y_i)|$ and accuracy within one level $n^{-1}\sum_i \mathbb{I}[|r(\hat{y}_i) - r(y_i)| \leq 1]$. ACS metrics are computed separately in each state and then averaged equally; NHIS metrics use its test set.

Table 2: Datasets and private targets. ACS sizes are per state; training data are available only to supervised references. C denotes consolidated classes; LOC is open-vocabulary.

Dataset	Target	Meaning	C	Train	Test
ACS	AGE	Age group	4	10,000	1,200
ACS	SCH	Educational attainment	5	10,000	1,200
ACS	SEX	Sex	2	10,000	1,200
ACS	MAR	Marital status	5	10,000	1,200
ACS	INC	Personal income above $50,000	2	10,000	1,200
ACS	OCC	Broad occupation group	12	10,000	1,200
ACS	POB	Place of birth	20	10,000	1,200
ACS	LOC	Residential location	Open	10,000	1,200
NHIS	AGE	Age group	4	30,577	1,200
NHIS	SCH	Educational attainment	5	30,482	1,200
NHIS	SEX	Sex	2	30,624	1,200
NHIS	MAR	Marital status	5	27,352	1,200
NHIS	OCC	Broad occupation group	10	17,286	1,200

Following prior geolocation work that evaluates predictions by great-circle error and distance-based accuracy (Liu et al., 2026; Haas et al., 2024), location predictions are assessed by the great-circle distance d_i between the predicted and true coordinates. We report $\mathrm{Acc}\,@100\,\mathrm{km} = n^{-1}\sum_i \mathbb{I}[d_i \leq 100]$ and median distance in kilometers. For ACS, both metrics are computed per state before equal averaging across California, Texas, Illinois, and New York in the final summary.

5.1.3 Compared approaches

Label-dependent references include Train Majority, supervised XGBoost (Chen & Guestrin, 2016), and TabPFN (Hollmann et al., 2025); they provide scale references outside the zero-shot threat model. Cluster-based attacks comprise K-means, SCAN (Van Gansbeke et al., 2020), SCARF (Bahri et al., 2022), INSTRUCTOR (Su et al., 2023), CLIP clustering (Radford et al., 2021), ClusterLLM (Zhang et al., 2023), GoalEx (Wang et al., 2023), and TableDC (Rauf et al., 2025). They map anonymous clusters to target values using released proportions and are therefore prior-aware. Semantic zero-shot comparisons comprise MiniLM (Wang et al., 2020), Gen-Z (Kumar et al., 2024), CLIP prototype (Radford et al., 2021), and LLM-induced decision trees (Knauer et al., 2025). E0 is the direct LLM attack; E1 constructs pseudo-context from the same scores. Appendix D.2 gives configurations and information access, while Appendix F.1 reports method variants.

5.2 Private attribute recovery and table-level inference

Private-attribute recovery. Table 3 reports the ACS comparison under *weak* disclosure. E0 recovers substantial information from visible features, and additional sample-specific attributes generally strengthen direct inference relative to *strict*. Accuracy can hide recovery across classes: on ACS SCH, E0 falls below Train Majority (34.02% versus 44.52%) but reaches higher macro recall (41.97% versus 20.00%) and within-one-level accuracy (90.77% versus 80.46%). Its predictions cover more than the dominant class and often remain close to the true education level.

Additional table-level recovery. E1 improves average recovery beyond E0. Across seven closed-set ACS targets, mean accuracy/macro recall rises by 3.35/2.20, 2.43/1.73, and 1.56/2.22 percentage points under *strict*, *weak*, and *prior* disclosure, respectively; the NHIS changes are 0.22/0.71, 0.85/1.32, and 0.23/1.20 points. On ACS *weak* targets, E1 improves accuracy on 6/7 and exceeds the best semantic zero-shot accuracy on 5/7. It also remains competitive with prior-assisted clustering

[4]GPT-4.1 model documentation; we use the `gpt-4.1-2025-04-14` snapshot.

Table 3: ACS *weak* results (four-state mean). A/R/W denote accuracy, macro recall, and within-one-level accuracy (%); LOC reports Acc@100 km (%) and median km. Bold and underline mark the best and second-best attack values per metric; supervised methods and Train Majority are references. Shaded E0 cells have lower accuracy than Train Majority.

Method	AGE (A/R/W ↑)	SCH (A/R/W ↑)	SEX (A/R ↑)	MAR (A/R ↑)	INC (A/R ↑)	OCC (A/R ↑)	POB (A/R ↑)	LOC (100 ↑/Med ↓)
Train Majority	31.33/25.00/81.94	44.52/20.00/80.46	51.23/50.00	44.50/20.00	62.02/50.00	27.23/8.33	30.00/5.00	45.25/149.86
XGBoost supervised	60.48/60.41/94.42	55.56/47.76/92.06	73.46/73.46	61.60/38.18	83.54/82.81	54.77/36.03	35.15/13.81	45.29/146.04
TabPFN supervised	63.17/62.86/96.06	56.77/47.31/92.25	73.35/73.37	62.90/37.87	83.90/83.18	56.75/35.30	36.19/11.51	46.29/141.27
K-means	38.06/34.80/73.84	24.68/21.97/65.46	49.70/49.49	22.74/18.54	41.96/39.10	11.07/9.27	4.96/4.94	–/–
SCAN	27.88/26.79/66.20	20.90/18.44/55.87	52.12/52.07	19.95/20.06	41.99/41.21	9.77/8.53	4.90/4.90	–/–
SCARF	23.93/23.23/62.03	19.69/19.61/55.75	48.46/48.43	20.59/20.55	43.37/42.27	8.06/7.92	4.92/4.98	–/–
INSTRUCTOR	25.63/23.81/62.03	21.81/17.10/59.04	40.11/40.02	25.45/21.29	45.94/44.34	7.15/6.10	6.37/5.03	–/–
CLIP clustering	28.76/28.00/64.51	20.06/20.10/56.19	44.49/44.46	19.91/19.96	43.28/42.18	7.45/7.86	4.97/5.05	–/–
ClusterLLM	27.48/24.43/64.25	24.91/18.30/62.01	52.68/51.99	30.12/19.00	52.51/47.10	16.61/11.08	5.76/5.01	–/–
GoalEx	26.30/24.24/75.42	33.50/21.83/76.39	51.61/50.81	37.08/20.30	50.62/43.96	18.37/9.47	**30.00**/5.00	–/–
TableDC	31.92/29.05/68.41	21.96/17.33/59.55	50.86/50.22	28.50/23.23	56.46/49.34	13.14/8.24	4.92/4.98	–/–
MiniLM	31.44/26.30/80.65	**41.02**/21.70/78.21	48.77/50.00	44.50/20.00	37.98/50.00	6.88/7.12	11.27/5.23	–/–
Gen-Z	38.88/43.08/67.79	23.10/24.87/55.58	51.92/52.83	48.17/37.28	63.73/53.35	26.81/14.97	9.08/5.59	–/–
CLIP prototype	37.17/36.75/71.69	24.77/34.22/58.98	58.48/58.65	30.04/31.51	63.50/64.27	10.35/9.95	6.58/5.44	–/–
LLM tree induction	**41.13/43.35**/74.85	15.15/24.79/68.81	50.69/51.51	55.38/27.28	72.71/68.20	35.23/15.61	8.58/5.06	–/–
E0 (direct LLM)	36.85/40.96/78.77	34.02/41.97/90.77	63.60/64.31	54.25/31.16	**78.31/78.07**	39.98/17.16	24.71/**7.12**	39.06/162.74
E1 (ours)	41.02/42.16/**85.77**	35.75/**42.78/92.04**	**65.77/66.26**	**62.52/39.11**	78.06/78.04	**40.48/17.58**	25.17/6.93	**41.08/156.24**

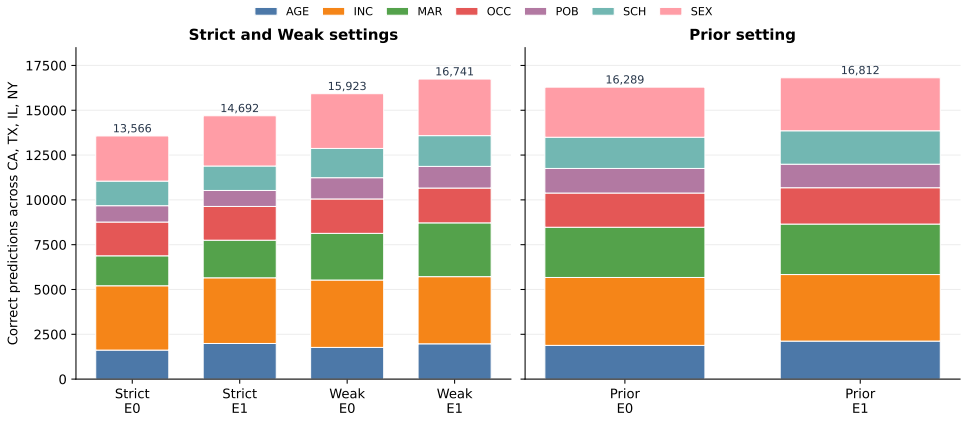


Figure 2: Qwen3-14B exact-label recovery on seven ACS targets. Each segment counts correct predictions across four states (4,800 samples per target). LOC uses geographic metrics and is excluded.

references without using released proportions for label alignment. Clustering seeks feature structure before aligning groups to labels, while semantic baselines predict samples independently. E1 combines label semantics with dependencies learned from the cohort. LOC also improves from 39.06% to 41.08% Acc@100 km, with median error falling from 162.74 to 156.24 km. Because E0 and E1 use the same LLM scores, their difference isolates recovery enabled by pseudo-context constructed from released samples.

When does table-level inference help? E1 uses class-wise score rankings to select samples whose LLM predictions provide stronger evidence for their assigned labels. Figure 5 shows that higher-ranked samples have higher pseudo-label precision for INC and SCH. The adaptive selection also outperforms fixed per-class budgets (Section 5.3). These results link the table-level gains to informative pseudo-context rather than to additional LLM queries: E1 turns selected semantic predictions into examples from which the tabular learner can exploit dataset-specific dependencies.

Effect of disclosure. Figure 2 shows higher ACS recovery with *weak* than *strict* features. Additional visible attributes can strengthen both E0 inference and the pseudo-context supplied to E1. Released target proportions instead provide aggregate $P(Y)$ information: under *prior* disclosure, they change how many context samples E1 selects per predicted class without changing E0 labels. E1 raises

correct ACS predictions from 16,289 to 16,812 under *prior*. The aggregate pattern demonstrates the table-level risk: a released cohort can reveal more than independent row-wise inference.

5.3 Further analysis

Context construction. Adaptive selection exceeds fixed-100 score and margin ranking on 14 ACS target means (46.67% versus 44.35% and 43.94% accuracy). Across all seven closed-set ACS targets, matched-Q adaptive selection yields larger average gains than frequency-proportional allocation, while permuting context labels sharply reduces recovery. Under *prior* disclosure, prior-aware allocation improves mean accuracy/macro recall by +1.01/+1.80 points across 12 targets, versus +0.23/+0.61 for ordinary E1. Appendix F.1 reports these comparisons.

Downstream learner. Replacing TabPFN with XGBoost on the same context improves ACS accuracy over E0 by 2.95 and 2.15 points under *strict* and *weak* disclosure. The gains weaken on NHIS and reverse under *prior* disclosure, showing that transfer depends on the learner's inductive bias. Full results appear in Appendix F.1 for each dataset and disclosure condition.

Class-wise ranking. The score-matrix comparison in Figure 1 motivates the selector; Figure 5 tests it against held-out labels. High-ranked samples are more precise for INC and SCH, while OCC provides weaker separation. This class-wise precision analysis tests the ranking signal separately from the context size determined by evidence mass. Appendix F.2 examines context quality and class coverage across targets.

6 Discussion and Limitations

The released cohort is an attack resource even without any disclosed target labels. E0 uses pretrained semantic knowledge to infer attributes from each sample, while E1 converts selected predictions into pseudo-supervision for learning dependencies within the cohort. The matched E0–E1 comparison therefore measures an increment that a row-wise privacy audit would miss. The gains in both accuracy and macro recall show that this increment can recover additional values across classes, rather than merely reinforcing the most common prediction in the released cohort.

Disclosure changes both sources of information. Under *weak* disclosure, additional visible attributes can strengthen an individual's semantic prediction and the context from which E1 learns. Aggregate target proportions instead guide how much context to select for each predicted class; they do not supply individual target labels. These distinctions matter when assessing a proposed release: concealing the target column alone does not determine which other disclosed information remains useful to an attacker under the stated release conditions.

The attack also depends on the available model. GPT-4.1 exceeds Qwen3-14B in direct-inference accuracy on 34/36 matched closed-set cells, illustrating why a single LLM is an incomplete measure of risk. E1 requires no additional LLM calls beyond E0, and Appendix D.4 reports the observed inference time and API-cost estimates for the studied backbones. The table-level step can therefore add recovery without a second generation pass over the released cohort.

Our evidence is limited to the sampled ACS and NHIS cohorts, the studied attributes, and the tested pretrained models. E1 relies on the quality of its initial pseudo-labels and may propagate their errors. Checkpoints predating the evaluated public-use files reduce, but do not eliminate, contamination concerns. These experiments measure attribute inference, not identification of named individuals or a formal privacy guarantee for a specific released dataset or population.

7 Conclusion

Withholding a sensitive column does not preclude zero-shot inference from the remaining tabular features. We showed that LLM predictions can supply pseudo-labels for learning dependencies across a released cohort, yielding additional recovery beyond direct predictions on average across ACS and NHIS. This table-level gain depends on pseudo-context quality and coverage. Future privacy audits should assess both row-level and table-level inference under the intended feature and aggregate-statistics disclosure policy for each sensitive target.

AI USE STATEMENT

Generative AI tools were used for brainstorming and discussion of research ideas, literature search and discovery, drafting parts of the manuscript, and editing the paper for clarity and readability. AI-generated suggestions were critically reviewed rather than treated as authoritative sources. The authors verified the technical content and references, revised all AI-assisted material, and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is intended to study and audit privacy risks in released tabular data. Our experiments use public-use ACS and NHIS survey data and evaluate attribute inference at the aggregate level; we do not attempt to identify named individuals. The same techniques could, however, be misused to infer sensitive attributes that individuals did not intend to disclose. We therefore report aggregate attack performance rather than row-level inferred attributes and frame the results as evidence of inferential privacy risk, not as identification of particular individuals or a formal privacy guarantee. Researchers applying these methods to other data should follow the applicable data-use and ethical requirements.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the experimental pipeline in the appendices. Appendix A describes the data sources, sample construction, target definitions, disclosure policies, and aggregate priors. Appendix B specifies the LLM prompts, output parsing and score normalization, and model configurations. Section 4 and Appendix C describe pseudo-context selection, tabular inference, fusion, and the LOC-specific procedure, while Appendix D specifies the evaluation protocol, baselines, and implementation details. Random seeds and the principal model configurations are reported with these specifications.

REFERENCES

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. SCARF: Self-supervised contrastive learning using random feature corruption. In *Proceedings of the 10th International Conference on Learning Representations*, pp. 1–16, Virtual Conference, 2022.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, California, USA, 2016.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, pp. 1–86, 2025.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems 34*, pp. 6478–6490, Virtual Conference, 2021.

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14:1–14:53, 2010.

Daria Grushina, Kseniia Kuvshinova, Alina Kostromina, Aziz Temirkhanov, Mile Mitrovic, and Dmitry Simakov. LLMTabBench: Evaluating LLMs on binary tabular classification from zero to few shots. *arXiv preprint arXiv:2605.24417*, pp. 1–23, 2026.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. PIGEON: Predicting image geolocations. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12893–12902, Seattle, Washington, USA, 2024.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 5549–5581, Valencia, Spain, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1569–1582, Los Angeles, California, USA, 2022.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Ricardo Knauer, Mario Koddenbrock, Raphael Wallsberger, Nicholas M. Brisson, Georg N. Duda, Deborah Falla, David W. Evans, and Erik Rodner. "Oh LLM, I'm Asking Thee, Please Give Me a Decision Tree": Zero-shot decision tree induction and embedding with large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1196–1206, Toronto, Canada, 2025.

Sachin Kumar, Chan Young Park, and Yulia Tsvetkov. Gen-Z: Generative zero-shot text classification with contextualized label descriptions. In *Proceedings of the 12th International Conference on Learning Representations*, pp. 1–24, Vienna, Austria, 2024.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Simeng Qin, and Xiaochun Cao. GeoShield: Safeguarding geolocation privacy from vision-language models via adversarial perturbations. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pp. 35653–35661, Singapore, 2026.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, pp. 1–92, 2024.

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

Yifeng Mao, Bozhidar Stevanoski, and Yves-Alexandre de Montjoye. Automated privacy risk estimation of limited fixed aggregate statistics. *Proceedings on Privacy Enhancing Technologies*, 2026(3):100–117, 2026.

Myeong Seok Oh, Dong-Yun Kim, Hanseok Oh, Chaean Kang, Joeun Kang, Xiaonan Wang, Hyunjung Park, Young Cheol Jung, and Hansaem Kim. Subject-level inference for realistic text anonymization evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17100–17135, San Diego, California, United States, 2026.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, pp. 1–35, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, Virtual Conference, 2021.

Hafiz Tayyab Rauf, André Freitas, and Norman W. Paton. TableDC: Deep clustering for tabular data. *Proceedings of the ACM on Management of Data*, 3(3):229:1–229:28, 2025.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2018.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33*, pp. 596–608, Virtual Conference, 2020.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *Proceedings of the 12th International Conference on Learning Representations*, pp. 1–47, Vienna, Austria, 2024.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121, Toronto, Canada, 2023.

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

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to classify images without labels. In *Proceedings of the 16th European Conference on Computer Vision*, pp. 268–285, Glasgow, United Kingdom, 2020.

Peng Wang, Dongsheng Wang, He Zhao, Hangting Ye, Dandan Guo, and Yi Chang. LLM empowered prototype learning for zero and few-shot tasks on tabular data. *arXiv preprint arXiv:2508.09263*, pp. 1–23, 2025.

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

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. Goal-driven explainable clustering via language descriptions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10626–10649, Singapore, 2023.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, Virtual Conference, 2020.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin T. Vechev. A synthetic dataset for personal attribute inference. In *Advances in Neural Information Processing Systems 37*, pp. 120735–120779, Vancouver, Canada, 2024.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13903–13920, Singapore, 2023.

A Dataset Construction and Disclosure Policies

A.1 Data sources and sample construction

ACS experiments use the 2024 one-year person-level public-use files for California, Texas, Illinois, and New York. A separate pool is constructed for each target and state, with disjoint training and test participants. The supervised references use 10,000 labeled training samples. A deterministic selection with seed 42 retains 1,200 test samples. Target counts are allocated by the largest-remainder rule so that integer sample counts preserve the source proportions as closely as possible. The *weak* condition reuses the sample identifiers selected under *strict* disclosure, so all attacks compared within a target use matched samples.

NHIS experiments use the 2024 national Sample Adult public-use file. After target-specific missing-value filtering, the supervised training sets contain the counts reported in Table 2 and do not overlap the test cohort. A deterministic joint selection with seed 42 chooses 1,200 participants while preserving the marginal proportions of all five targets. The *strict* and *weak* conditions use the same participants. ACS person weights and NHIS survey weights are retained in the source data but are neither prompt features nor weights in the main metrics.

A.2 Target definitions and candidate values

Table 4 summarizes the candidate values. The mappings are fixed before model evaluation and applied identically to predictions and ground truth.

Table 4: Target label spaces. OCC and POB show representative values because their complete spaces contain 10–20 categories.

Dataset	Target	C	Candidate values
ACS/NHIS	AGE	4	18–29; 30–44; 45–64; 65+
ACS/NHIS	SCH	5	No high-school diploma; high-school diploma; some college or associate's degree; bachelor's degree; graduate degree
ACS/NHIS	SEX	2	Female; male
ACS/NHIS	MAR	5	Married; widowed; divorced; separated; never married
ACS	INC	2	Income at most \$50,000; income above \$50,000
ACS	OCC	12	Broad groups; e.g., management/business/science/arts, service, sales/office, production, and transportation/material moving
NHIS	OCC	10	Broad groups; e.g., management/business/science/arts, healthcare, service, sales/office, and production
ACS	POB	20	Countries or regions; e.g., United States, Mexico, India, Philippines, China, other Europe, and other Africa
ACS	LOC	Open	Human-readable place names within the known state

LOC remains an open-vocabulary LLM target. The prompt requests five human-readable places within the known state rather than exposing the state-specific evaluation labels. For supervised references, source geographic codes are consolidated into 31 areas in California, 34 in Texas, 15 in Illinois, and 23 in New York. Their predicted CBSA labels are mapped to official TIGER/Line representative points, while each state-specific non-CBSA label uses the mean representative point of its constituent PUMAs; the resulting coordinates are evaluated with the same distance metrics as the LLM outputs. Generated place names are canonicalized to geographic entities and coordinates for evaluation and the regression-based stage in Appendix C.3. Top-k candidate coverage measures whether any generated candidate is geographically close to the true location; it is reported separately from the accuracy of the final selected candidate.

A.3 Visible features and aggregate proportions

Table 5 uses source-variable names from the official ACS PUMS data dictionary[5] and NHIS Sample Adult codebook.[6] Each target's *strict* input is the union of its dataset's common base row and its target-specific additions. The count d_{strict} includes both; the final column lists fields added only under *weak* disclosure.

[5] 2024 ACS PUMS Data Dictionary.

[6] 2024 NHIS Sample Adult Codebook.

Table 5: Visible source variables by dataset and target. Each *strict* input combines the base row and its target-specific additions. The current target is always withheld.

Dataset	Target	d_{strict}	Base or target-specific *strict* fields	*Weak*-only fields
ACS	Base	6	COW, WKHP, JWMNP, JWTRNS, HICOV, WRK	–
ACS	AGE	10	INDP, CIT, ENG, RELSHIPP	SCHL, MAR, OCCP, PINCP
ACS	SCH	10	INDP, CIT, ENG, RELSHIPP	AGEP, MAR, OCCP, PINCP
ACS	SEX	10	INDP, CIT, ENG, RELSHIPP	AGEP, OCCP, MAR, PINCP
ACS	MAR	9	INDP, CIT, ENG	AGEP, SCHL, SEX, OCCP, PINCP
ACS	INC	10	INDP, CIT, ENG, RELSHIPP	AGEP, SCHL, SEX, OCCP, MAR
ACS	OCC	9	CIT, ENG, RELSHIPP	AGEP, SCHL, SEX, MAR, PINCP
ACS	POB	8	INDP, RELSHIPP	AGEP, SCHL, SEX, OCCP, MAR, PINCP
ACS	LOC	10	INDP, CIT, ENG, RELSHIPP	AGEP, SCHL, SEX, OCCP, PINCP
NHIS	Base	12	PHSTAT_A, SLPHOURS_A, HICOV_A, USUALPL_A, LASTDR_A, EMPLASTWK_A, EMPWRKFT1_A, EMPWKHRS3_A, PCNTADLT_A, PCNTKIDS_A, REGION, URBRRL23	–
NHIS	AGE	18	BMICAT_A, SMKEV_A, DRKSTAT_A, DIFF_A, VISIONDF_A, HEARINGDF_A	EDUCP_A, MARSTAT_A
NHIS	SCH	18	SMKEV_A, DRKSTAT_A, EMPHEALINS_A, EMPSICKLV_A, HOUTENURE_A, FDSCAT4_A	AGEP_A, EMDOCCUPN2_A
NHIS	SEX	17	SMKEV_A, EMPHEALINS_A, EMPSICKLV_A, HOUTENURE_A, FDSCAT4_A	AGEP_A, EDUCP_A, MARSTAT_A, EMDOCCUPN2_A
NHIS	MAR	14	HOUTENURE_A, FDSCAT4_A	AGEP_A, SEX_A, EDUCP_A, EMDOCCUPN2_A
NHIS	OCC	19	BMICAT_A, SMKEV_A, DRKSTAT_A, EMPHEALINS_A, EMPSICKLV_A, HOUTENURE_A, FDSCAT4_A	SEX_A, EDUCP_A

Identifiers, survey-design fields, target source fields, and sampling weights are excluded from both conditions. Feature names and their natural-language renderings are fixed by the dataset metadata, so the LLM and tabular learner receive two representations of the same disclosed fields.

For a task in the *prior* condition with training counts N_c, the released vector is $\rho_c = N_c / \sum_{c'} N_{c'}$. ACS vectors are state and target specific, whereas NHIS vectors are national. Percentages are rounded before insertion into the prompt and presented as approximate population frequencies. They are never joined to individual samples. The *prior* vector ρ is external information and is distinct from the LLM score background b_c in Equation 2.

B Prompts, Parsing, and Model Configurations

B.1 Prompt examples

Each closed-set prompt contains a task instruction, one sample rendered as a list of field meanings and values, the target name and meaning, the complete candidate set, and a JSON output schema. The model is asked to return every candidate exactly once with a self-reported score between zero and one. For example, an AGE prompt has the following structure:

> Infer one hidden attribute from public profile features.
> *Public profile:* class of worker: [value]; usual weekly work hours: [value]; . . .
> *Allowed labels for AGE (age group):* 18–29; 30–44; 45–64; 65+.
> Assign a confidence to every allowed label and return only the specified JSON object.

Weak disclosure changes only the fields in the public profile. Under the *prior* condition, the approximate value frequencies are inserted between the candidate list and prediction instruction, together with an instruction to treat them as population-level information rather than an individual label. LOC follows the same profile structure but cannot enumerate its large place-name space. Its prompt instead has the following target and output instructions:

Hidden target: current location within [state].
Return five distinct plausible locations, each as a human-readable place name with its place type and a score between zero and one.
Return only the specified JSON object containing the five candidates.

The five locations are generated separately for each sample rather than selected from a shared label list. All sample-specific prompts are saved before inference.

B.2 Output parsing and candidate normalization

For closed-set targets, parsing requires one finite, nonnegative score for every member of $\mathcal{Y}$. Label aliases are mapped to the fixed candidate strings, after which scores are normalized by their sum as in Equation 1. E0 uses the largest normalized score; ties follow the emitted order and then the fixed candidate order. A model–task cell enters the reported all-score comparisons only after the output matrix contains all n samples and all C candidates. The raw response, parsed order, unnormalized scores, and normalized matrix are retained separately.

For LOC, duplicate textual candidates referring to the same canonical entity are merged before score normalization. Canonicalization uses only the generated name, stated place type, and known state. The resulting entity coordinates support Haversine evaluation and the method in Appendix C.3; no ground-truth location is used to resolve a candidate. The emitted order is retained to break equal-score ties.

B.3 Pretrained models and inference configurations

The main score matrices are generated by Qwen3-14B in bfloat16 with greedy decoding, an 8,192-token input limit, a 1,024-token output limit, and seed 42. Closed-set prompts request all labels in one response. LOC requests five candidates and uses a 700-token output limit. Meta-Llama-3.1-8B-Instruct uses the same task content in the cross-model evaluation. DeepSeek-R1-Distill-Qwen-14B is used for the targeted reasoning analysis in Appendix F.3; its generated rationales are treated as observable outputs rather than faithful causal explanations of its predictions. Each score matrix is generated once and shared by E0, E1, and their matched selection variants.

For the tabular stage, numerical fields are converted to 32-bit floating-point values, with missing or non-numeric entries represented by -1. Categorical fields are ordinal-encoded using categories observed in the released dataset; unseen values are also assigned -1. The resulting categorical column indices are supplied to TabPFN. This encoding uses only the released feature matrix and is shared by the selected context and remaining samples. Pseudo-label strings are mapped to consecutive integer indices and mapped back after prediction. The main instantiation uses the TabPFN v3 classifier for closed-set targets and its regressor for LOC, with eight estimators, random seed 82, and `fit_preprocessors` mode. The XGBoost control uses 300 trees of maximum depth four, learning rate 0.05, row and column subsampling rates of 0.8, and the same random seed. Degenerate contexts follow the E0 fallback in Appendix C.1.

C Additional Method Details

C.1 Complete inference procedure

Algorithm 1 summarizes closed-set inference from the normalized LLM score matrix. E0 score ties follow the emitted candidate order and then the fixed candidate order; context-ranking ties follow sample index. Our main instantiation of the tabular learner $\mathcal{A}$ is TabPFN.

Under *prior* disclosure, $\mathrm{Apportion}$ keeps the ordinary total Q fixed. For every nonempty bucket, Equation 4 gives the real-valued target allocation $r_c = Q\omega_c / \sum_d \omega_d$; an empty bucket has weight and capacity zero. We initialize $q_c^{\mathrm{prior}} = \min\{|B_c|, \lfloor r_c \rfloor\}$ and distribute the remaining units one at a time to the non-saturated class with the largest residual $r_c - q_c^{\mathrm{prior}}$, using remaining capacity and then label order to break ties. This yields integer budgets satisfying $\sum_c q_c^{\mathrm{prior}} = Q$ and $q_c^{\mathrm{prior}} \le |B_c|$. When $Q = 0$, all apportioned budgets are zero.

Algorithm 1 Closed-set zero-shot private-attribute inference (E1)

Require: Released dataset $X = (\boldsymbol{x}_i)_{i=1}^n$, LLM scores (s_{ic}), candidate values $\mathcal{Y}$, optional proportions $\boldsymbol{\rho}$

Ensure: Predictions $(\hat{y}_i^{(1)})_{i=1}^n$

E0: Direct LLM inference

$\hat{y}_i^{(0)} \leftarrow \arg\max_{c \in \mathcal{Y}} s_{ic}$ for every i

E1: Evidence-based pseudo-context and tabular inference

for $c \in \mathcal{Y}$ **do**

  $B_c \leftarrow \{i : \hat{y}_i^{(0)} = c\}$; $b_c \leftarrow n^{-1} \sum_i s_{ic}$

  $\theta_{ic} \leftarrow [s_{ic} - b_c]_+ / \max\{\varepsilon, 1 - b_c\}$ for every i

  $q_c \leftarrow \lfloor \sum_{i \in B_c} \theta_{ic} + 1/2 \rfloor$

end for

if ρ is disclosed **then**

  $Q \leftarrow \sum_c q_c$; $\bar{\theta}_c \leftarrow |B_c|^{-1} \sum_{i \in B_c} \theta_{ic}$ for nonempty B_c, and 0 otherwise

  $(q_c)_c \leftarrow \mathrm{Apportion}(Q, (\rho_c \bar{\theta}_c, |B_c|)_c)$

end if

$S \leftarrow \bigcup_{c \in \mathcal{Y}} \mathrm{Top}_{q_c}(B_c; \theta_{ic})$; $\mathcal{C}_S \leftarrow \{(\boldsymbol{x}_i, \hat{y}_i^{(0)}) : i \in S\}$

if $|S| < 2$ or $|\mathcal{Y}_S| < 2$ **then**

  $S \leftarrow \{i : s_{i,\hat{y}_i^{(0)}} > b_{\hat{y}_i^{(0)}}\}$; $\mathcal{C}_S \leftarrow \{(\boldsymbol{x}_i, \hat{y}_i^{(0)}) : i \in S\}$

end if

if $|S| < 2$ or $|\mathcal{Y}_S| < 2$ **then**

  return $(\hat{y}_i^{(0)})_{i=1}^n$

end if

$f_S \leftarrow \mathcal{A}(\mathcal{C}_S)$; $m_S \leftarrow |\mathcal{Y}_S|$

$\hat{y}_i^{(1)} \leftarrow \hat{y}_i^{(0)}$ for $i \in S$

$\hat{y}_i^{(1)} \leftarrow \arg\max_{c \in \mathcal{Y}} s_{ic} w_i(c)$ for $i \notin S$, where $w_i(c) = m_S f_S(\boldsymbol{x}_i)_c$ if $c \in \mathcal{Y}_S$, and 1 otherwise

return $(\hat{y}_i^{(1)})_{i=1}^n$

If the mass-budget context has fewer than two samples or represented classes, we first expand it to samples with positive evidence for their E0 label. This rule is used in the cross-model completion runs; it leaves the main Qwen results unchanged because none of their reported closed-set contexts require it. If the expanded context is still degenerate, E1 returns $\hat{y}_i^{(0)}$ for every sample. Otherwise, the learner is conditioned on $\mathcal{C}_S$, and Equations 5–7 produce predictions for the complete dataset. The same principle applies to LOC: fewer than two spatial context samples cause an E0 fallback. Neither fallback inspects ground-truth targets.

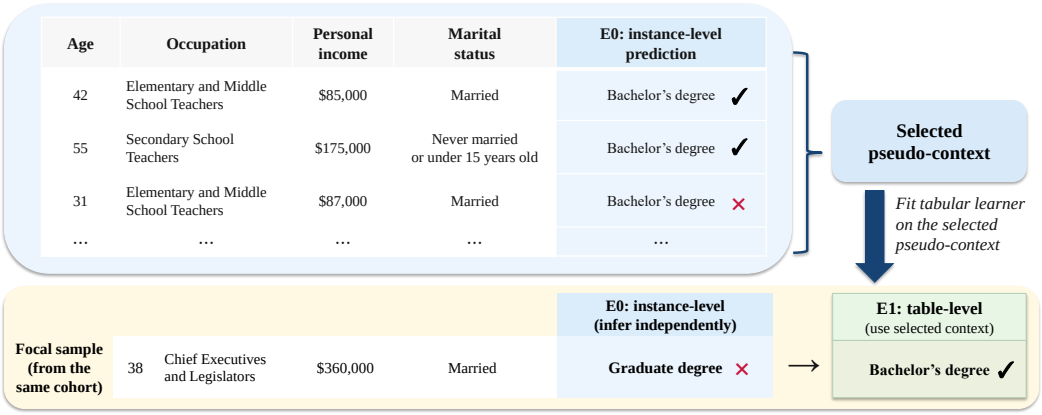


Figure 3: An education (SCH) example from a held-out cohort. The focal sample is predicted independently by E0; E1 uses selected pseudo-labeled samples from the same cohort and recovers its hidden value. Checks and crosses indicate correctness against withheld labels, which are used only to annotate this figure. The selected context can include an incorrect pseudo-label.

C.2 Expected-size interpretation of the context budget

Condition on the observed score matrix and consider an auxiliary randomized selector in which each candidate $i \in B_c$ is included independently according to

$$Z_{ic} \sim \text{Bernoulli}(\theta_{ic}), \qquad N_c = \sum_{i \in B_c} Z_{ic}. \tag{9}$$

Then $\mathbb{E}[N_c] = \sum_{i \in B_c} \theta_{ic} = M_c$. The deterministic budget in Equation 3 satisfies

$$q_c \in \operatorname*{arg\,min}_{q \in \{0,\dots,|B_c|\}} \mathbb{E}\left[(q - N_c)^2\right], \tag{10}$$

because $\mathbb{E}[(q - N_c)^2] = (q - M_c)^2 + \text{Var}(N_c)$. Thus rounding M_c gives a minimizing integer, with half-integer ties rounded upward. The executed selector remains deterministic and takes the top-q_c samples by θ_{ic}; the construction does not claim that θ_{ic} is a correctness probability or that q_c is optimal for downstream prediction.

C.3 Candidate representation and statistical prediction

Location context and compatibility. For resolved candidates $\mathcal{Z}_i$ with normalized scores $s_i(z)$, the representative location and dispersion are

$$z_i^\star = \operatorname*{arg\,min}_{z \in \mathcal{Z}_i} \sum_{z' \in \mathcal{Z}_i} s_i(z') d(\boldsymbol{r}(z), \boldsymbol{r}(z')), \qquad D_i = \sum_{z \in \mathcal{Z}_i} s_i(z) d(\boldsymbol{r}(z_i^\star), \boldsymbol{r}(z)). \tag{11}$$

Here d is great-circle distance in kilometers. Let D_{ref} be the median of strictly positive D_i values, or 1 km if none are positive. Spatial concentration is $\gamma_i = (1 + D_i / D_{\text{ref}})^{-1}$. Each candidate is assigned a region identifier $\rho(z)$ using its county when available, then its metropolitan area, and otherwise its canonical entity identifier. This region map is distinct from the released target proportions denoted by ρ_c in Section 3.2.

For region v, let $p_{iv} = \sum_{z \in \mathcal{Z}_i : \rho(z) = v} s_i(z)$ and $b_v = n^{-1} \sum_i p_{iv}$. Samples are grouped by $v_i = \rho(z_i^\star)$. Their selection evidence is $e_i = \gamma_i [p_{iv_i} - b_{v_i}]_+ / (1 - b_{v_i})$ when $b_{v_i} < 1$, and zero otherwise. For $B_v = \{i : v_i = v\}$, we set $q_v = \min\{|B_v|, \lfloor \sum_{i \in B_v} e_i + 1/2 \rfloor\}$ and select its q_v highest-evidence samples. Their union forms S_{loc}, and we set their representative pseudo-locations to $\bar{z}_i = z_i^\star$. Their coordinates supply the regression targets and are retained as the final outputs for context samples.

The sample-specific distance scale used for location compatibility is

$$\tau_i = \max\left\{10^{-9}\ \text{km}, \begin{cases} D_i, & D_i > 0, \\ D_{\text{ref}}, & D_i = 0. \end{cases}\right\}. \tag{12}$$

Let $\tilde{\boldsymbol{r}}_i$ be the coordinate predicted by the TabPFN regressor for sample i. Its normalized compatibility with each candidate is

$$t_i(z) = \frac{\exp[-d(\tilde{\boldsymbol{r}}_i, \boldsymbol{r}(z))/\tau_i]}{\sum_{z' \in \mathcal{Z}_i} \exp[-d(\tilde{\boldsymbol{r}}_i, \boldsymbol{r}(z'))/\tau_i]}, \qquad z \in \mathcal{Z}_i. \tag{13}$$

Substituting this distribution into Equation 8 gives the LOC prediction. The distance-decreasing form assigns greater structural support to candidates closer to the location inferred from the visible tabular features, while the sample-specific scale reflects the dispersion of the LLM's own candidates. Both τ_i and $t_i(z)$ are computed without ground-truth locations. Spatial concentration affects context selection; it is not an additional candidate-wise multiplier at prediction time. For a fixed sample, the normalization in Equation 13 multiplies all product scores by the same positive constant and therefore preserves their ranking relative to the corresponding unnormalized distance product.

Closed-set and open-vocabulary targets share the same principle but differ in how their candidate spaces are obtained. For a closed-set target, every value in $\mathcal{Y}$ is supplied in the prompt and receives an LLM score, so the classifier output and semantic scores are indexed by the same fixed set. For LOC, the prompt instead generates the sample-specific set $\mathcal{Z}_i$. Aliases resolving to the same geographic entity are merged before normalization, and the regressor supplies a compatibility value only for candidates in $\mathcal{Z}_i$. A place absent from this generated set is therefore unobserved rather than assigned a known score of zero. Product fusion reranks the generated candidates and cannot introduce a new place after E0. In both cases, the tabular prediction is mapped back to the canonical attribute value used by the prompt, and an unusable context invokes the deterministic E0 fallback in Appendix C.1.

C.4 Selection and prediction variants

The reported variants change one component of the attack at a time. The two fixed-budget controls retain the E0 class buckets and select at most 100 samples per predicted class, ranking them either by the winning score or by the row-wise margin; their downstream endpoint is otherwise matched. The adaptive variant replaces this fixed cap with Equation 3. A matched-total-budget control allocates context in proportion to E0 bucket frequency while retaining the evidence ranking within each bucket. A separate diagnostic permutes pseudo-labels over the adaptively selected context samples while preserving their label counts. The learner-backend control retains the adaptive context and pseudo-labels but replaces TabPFN with XGBoost. E0 omits the tabular stage entirely. These comparisons do not add labels missing from the E0 buckets or choose a variant separately for each target. Coverage-completion and alternative fusion variants are not components of the reported attack.

D Evaluation Protocol and Baselines

D.1 Metrics and aggregation

Accuracy and macro recall use the fixed target label set $\mathcal{Y}$. Their definitions are $n^{-1}\sum_i \mathbb{I}[\hat{y}_i = y_i]$ and $C^{-1}\sum_{c\in\mathcal{Y}} \mathrm{TP}_c/(\mathrm{TP}_c + \mathrm{FN}_c)$, respectively. Every predefined class is included in the macro average even when the attack never predicts it. AGE and SCH additionally use the rank function specified in Section 5.1.2: mean absolute level error averages $|r(\hat{y}_i) - r(y_i)|$, and within-one-level accuracy counts errors no greater than one rank.

For LOC, d_i is the Haversine distance in kilometers between the predicted and reference coordinates. We report $n^{-1}\sum_i \mathbb{I}[d_i \le 100]$ and the sample median of d_i. ACS metrics are first computed independently for each state and then averaged with equal state weight; hence the distance summary is the mean of four state-level medians. NHIS metrics are computed once on the national cohort. No survey weights are used. Averages across targets are reported only when the participating targets and weighting rule are stated explicitly, and LOC is never mixed into a closed-set average.

D.2 Baseline information and configurations

Table 6 separates the information used by each comparison. All methods receive the same disclosed feature columns. Candidate semantics comprise the target name, value names, and their natural-language meanings; aggregate proportions are supplied only where indicated.

Table 6: Information available to comparison methods. A check in the final column places the method outside the zero-shot threat model.

Method family	Candidate semantics	Target proportions	Training labels
Train Majority	–	–	✓
Supervised XGBoost / TabPFN	–	–	✓
K-means / SCAN / SCARF / TableDC	–	✓	–
INSTRUCTOR / CLIP clustering	✓	✓	–
ClusterLLM / GoalEx	Target goal	✓	–
MiniLM / Gen-Z / CLIP prototype	✓	–	–
LLM tree induction	✓	–	–
E0 / E1, *strict* or *weak*	✓	–	–
E0 / E1, *prior*	✓	✓	–

Train Majority, XGBoost, and supervised TabPFN are fitted to the training data listed in Table 2. Unsupervised table baselines are transductive on the disclosed test samples. K-means clusters encoded features directly, using 20 initializations and at most 300 iterations. SCAN uses 20-nearest-neighbor consistency with a linear head. The SCARF baseline first learns features from the disclosed table through contrastive corruption and then applies SCAN. INSTRUCTOR embeds natural-language renderings of samples before clustering; CLIP clustering instead uses representations from a contrastively trained table–text encoder. ClusterLLM and GoalEx guide clustering with LLM-generated semantic information. TableDC jointly learns a tabular representation and C clusters from the disclosed samples. It is evaluated on all 99 closed-set cells: 84 ACS state–condition–target cells and 15 NHIS condition–target cells. LOC is omitted because categorical clusters do not provide a geographic candidate decoder. MiniLM and CLIP prototype instead select the nearest candidate description in their respective representation spaces. The number of clusters is fixed to C.

The LLM tree induction baseline adapts the zero-shot decision-tree procedure of Knauer et al. (2025). Qwen3-14B receives feature and candidate meanings but no labeled examples, then produces a single tree for each dataset, disclosure condition, and target. Its structured JSON tree is validated and applied to every test sample; this replaces the paper's text-to-code output format without adding supervision. Trees have maximum depth two, except the saved depth-three tree for ACS *strict* OCC. The baseline covers all closed-set *strict* and *weak* cells; it has no geographic decoder for LOC. Released target proportions are not inputs to its tree construction.

PriorSize alignment ranks learned clusters by size and target values by their released proportion, with fixed-seed averaging when ranks cannot be distinguished. It does not inspect sample-level targets, but it does require aggregate priors. These rows are therefore *prior*-aware comparisons even when their input features follow the *strict* or *weak* policy. For TableDC, this alignment is applied after clustering with the feature policy shown by each table; the *prior* condition reuses the *weak*-feature clustering. Hungarian alignment uses ground-truth labels and is used solely as a cluster-recovery diagnostic; it is never included in attack rankings.

Gen-Z uses Qwen3-14B as the channel model and selects the label description yielding the lowest mean input-token negative log likelihood. ClusterLLM and GoalEx follow their published LLM-guided clustering stages with tabular samples rendered as text. E0 and E1 use the identical saved LLM score matrix. Fixed-budget variants alter only pseudo-context selection; a learner-backend comparison is reported only when TabPFN and XGBoost are trained on exactly the same selected context.

D.3 Evaluation protocol consistency

All main E0–E1 comparisons use the all-label score protocol, matched sample identifiers, and the complete evaluation cohort. Single-answer prompts and alternative fusion rules use different protocols and are excluded from the main comparisons; fixed-100 context selection appears only as an ablation. A result is included in an aggregate only when every required state or national cell has completed the same protocol; incomplete cells are marked missing rather than replaced or averaged over the available subset.

D.4 Attack computation and cost accounting

Table 7 reports inference time for 1,200-sample cohorts. Qwen timings cover 36 completed ACS cells for SCH, MAR, and OCC; DeepSeek timings cover two NHIS SCH cells. Open-weight models run locally without an API charge. GPT-4.1 costs use the standard rates of US$2 per million input tokens and US$8 per million output tokens listed in the model documentation. Its time estimate is derived from the final throughput recorded for 18 complete 1,200-sample API runs.

Table 7: E0 inference time and API cost per 1,200-sample cell. Ranges across available cells are in parentheses. GPT-4.1 time uses 18 complete-run logs; its cost uses 38 saved cells.

LLM	Coverage	Hardware or access	Time	API cost / cell (USD)
Qwen3-14B	36 ACS cells	V100	0.718 s/sample (0.573–3.146)	$0 (local model)
DeepSeek-14B	NHIS SCH, *weak*	A800	6.29 h (18.9 s/sample)	$0 (local model)
DeepSeek-14B	NHIS SCH, *prior*	A800	7.55 h (22.6 s/sample)	$0 (local model)
GPT-4.1	38 saved cells	API	25.2 min/cell (19.0–41.7)	$1.68 median (1.00–3.71)

The GPT-4.1 cost estimate covers 45,600 saved final requests across 38 cells. Approximating tokens as characters divided by four yields 24.75 million input and 2.87 million output tokens. Its time estimate divides 1,200 by the final cumulative requests-per-second rate in each of 18 complete-run logs. Local GPU use has no API charge but still requires compute; model loading and experimental setup are outside these per-cohort measurements.

Table 8 separates the shared LLM inference from E1's extra work. Here n is the number of attacked samples, C is the closed-set target cardinality, and k is the requested number of open-vocabulary candidates per sample. A prompt evaluation processes one sample; local batching may combine several evaluations into one model invocation. E1 calculates closed-set background evidence in $O(nC)$ time and selects within predicted-class buckets in at most $O(n \log n)$ time. The tabular learner is then fit once to the selected context and predicts the remaining samples. Its runtime depends on the learner and hardware; the *prior*-aware quota requires no additional LLM pass.

Table 8: Computational requirements per target and disclosure condition. The E1 row gives work additional to E0; tabular runtime depends on the learner and hardware.

Attack	LLM prompt evaluations	Saved scores	Additional tabular work
E0, closed set	n	nC	None
E0, open vocabulary	n	At most nk	None
E1, either target type	0 additional	Reuses E0	Select context; fit once; predict remaining samples

E Complete Experimental Results

E.1 Attribute recovery across disclosure conditions

Table 9: Mean closed-set recovery and E1–E0 changes. A/R denote accuracy and macro recall (%); changes are percentage points.

Dataset	Disclosure	E0 A/R	E1 A/R	ΔA/R	A wins / R wins
ACS	*Strict*	40.37/34.31	43.73/36.51	+3.35/+2.20	4/7 / 4/7
ACS	*Weak*	47.39/40.11	49.82/41.84	+2.43/+1.73	6/7 / 5/7
ACS	*Prior*	48.48/39.09	50.04/41.31	+1.56/+2.22	5/7 / 5/7
NHIS	*Strict*	37.58/30.89	37.80/31.60	+0.22/+0.71	1/5 / 2/5
NHIS	*Weak*	40.15/34.51	41.00/35.83	+0.85/+1.32	3/5 / 3/5
NHIS	*Prior*	36.40/30.59	36.63/31.79	+0.23/+1.20	2/5 / 2/5

Table 10: ACS results under *strict* feature disclosure, averaged equally over four states. Closed-set cells contain A/R, with A/R/W for AGE and SCH; LOC contains $\mathrm{Acc}\,@100\,\mathrm{km}$/median km. Bold and underline mark the highest and second-highest attack values per metric; supervised models and Train Majority are references. Shaded E0 cells have lower accuracy than Train Majority.

Method	AGE (A/R/W ↑)	SCH (A/R/W ↑)	SEX (A/R ↑)	MAR (A/R ↑)	INC (A/R ↑)	OCC (A/R ↑)	POB (A/R ↑)	LOC (100 ↑/Med ↓)
Train Majority	31.33/25.00/81.94	44.52/20.00/80.46	51.23/50.00	44.50/20.00	62.02/50.00	27.23/8.33	30.00/5.00	45.25/149.86
XGBoost supervised	56.02/55.57/91.85	51.17/40.09/87.88	68.81/68.82	49.75/26.17	80.42/79.89	50.75/30.11	31.10/10.83	44.67/150.32
TabPFN supervised	59.31/58.15/92.90	53.27/39.75/88.23	68.94/68.95	51.90/26.35	81.46/80.80	52.52/27.57	33.33/8.58	45.92/143.61
K-means	39.59/36.18/76.11	25.39/20.18/62.85	48.00/47.37	27.18/22.38	47.54/41.14	8.67/8.35	5.38/4.98	–/–
SCAN	30.18/28.91/67.62	21.77/19.10/57.44	49.38/49.35	20.96/20.46	44.35/43.30	13.56/10.01	5.58/5.00	–/–
SCARF	26.26/26.31/65.10	19.91/19.16/56.89	50.08/50.08	19.84/19.78	53.76/53.84	14.54/10.37	4.92/4.93	–/–
INSTRUCTOR	28.29/26.08/62.26	24.25/19.84/61.28	50.41/50.26	30.41/**24.97**	36.98/33.65	5.72/4.53	6.59/5.10	–/–
CLIP clustering	25.72/25.07/63.61	20.75/20.28/56.93	48.60/48.53	21.43/19.70	46.39/46.05	8.44/8.26	5.72/5.08	–/–
ClusterLLM	31.90/27.83/67.80	26.97/18.86/65.10	51.86/51.11	29.90/18.96	48.33/40.19	13.60/9.29	6.80/5.14	–/–
GoalEx	26.78/22.25/77.51	**44.46**/20.00/80.33	50.66/49.71	42.58/19.80	42.98/38.26	14.83/7.18	**29.22**/4.96	–/–
TableDC	38.85/35.46/74.89	32.72/22.46/70.35	50.35/49.74	26.03/18.28	52.40/44.77	19.01/10.56	7.22/4.94	–/–
MiniLM	31.10/25.04/79.81	44.44/19.96/80.27	48.77/50.00	**44.50**/20.00	37.98/50.00	11.08/7.88	18.58/4.85	–/–
Gen-Z	39.12/43.02/67.00	28.40/23.11/59.85	50.50/51.50	38.46/21.86	63.21/52.91	25.56/13.85	5.35/5.48	–/–
CLIP prototype	31.06/30.27/65.46	23.02/31.10/59.02	54.54/54.71	27.19/22.14	58.96/60.80	5.90/7.83	6.12/5.65	–/–
LLM tree induction	38.50/37.17/**82.69**	41.42/22.80/82.15	46.27/46.56	33.19/19.44	57.75/47.79	34.06/14.51	27.56/4.83	–/–
E0 (direct LLM)	33.73/36.92/74.96	28.56/**31.82/88.00**	52.50/53.57	34.83/21.10	74.75/73.62	**39.25/16.38**	19.00/**6.77**	**39.15**/184.42
E1 (ours)	**41.42/43.57**/81.17	28.40/31.45/87.33	**58.40/58.96**	43.73/21.19	**76.33/77.39**	39.23/16.37	18.58/6.65	38.83/**178.37**

The complete NHIS baseline matrices under *strict* and *weak* disclosure are reported in Appendix E.1. The main text retains the compact *prior* comparison below and discusses NHIS jointly with the ACS results.

Table 11: ACS results when aggregate target proportions are disclosed. Cells contain accuracy/macro recall (%), with within-one-level accuracy added for AGE and SCH. Bold and underline mark the best and second-best reported attack values per metric. Shaded E0 cells have lower accuracy than Train Majority on the matched test set. TableDC uses cluster-size alignment with the released proportions.

Method	AGE	SCH	SEX	MAR	INC	OCC	POB
TableDC	31.92/29.05/68.41	21.96/17.33/59.55	50.86/50.22	28.50/23.23	56.46/49.34	13.14/8.24	4.92/4.98
E0 (direct LLM)	39.21/38.22/**88.15**	36.27/39.16/**90.60**	58.23/57.82	58.40/37.09	**78.92/78.03**	39.75/16.82	**28.58/6.51**
E1 (ours)	**44.17/43.62**/86.00	**38.85/39.35**/87.75	**61.60/61.64**	**58.67/41.01**	77.38/77.83	**42.23/19.41**	27.35/6.35

Table 12: NHIS results when aggregate target proportions are disclosed. Cells contain accuracy/macro recall (%), with within-one-level accuracy added for AGE and SCH. Bold and underline mark the best and second-best reported attack values per metric. TableDC uses cluster-size alignment with the released proportions.

Method	AGE	SCH	SEX	MAR	OCC
TableDC	19.19/18.50/52.43	21.94/20.53/55.90	50.83/50.80	27.58/25.72	11.35/9.52
E0 (direct LLM)	**32.83**/28.79/81.25	**28.58**/24.38/**83.92**	50.58/50.42	53.08/36.76	**16.92/12.58**
E1 (ours)	27.83/26.87/**94.25**	26.92/23.22/78.33	**56.42/56.33**	**55.50/40.21**	16.50/12.31

Table 3 and Tables 10–14 give the complete comparisons across disclosure conditions. A/R/W denote accuracy, macro recall, and within-one-level accuracy; supervised methods and Train Majority are reference points and are excluded from the attack-method ranking.

Negative E1–E0 changes are retained rather than omitted from aggregates.

E.2 Transfer across language models

Tables 15 and 16 evaluate whether the table-level increment transfers from the main Qwen experiments to another LLM family. Llama covers all eight ACS targets under *strict* and *weak* disclosure,

Table 13: NHIS results under *strict* feature disclosure. Bold and underline mark the highest and second-highest values separately for accuracy, macro recall, and W among attack methods. Cluster-based rows use released target proportions for cluster-to-value alignment.

Method	AGE (A/R/W ↑)	SCH (A/R/W ↑)	SEX (A/R ↑)	MAR (A/R ↑)	OCC (A/R ↑)
Train Majority	18.17/25.00/50.00	24.08/20.00/69.75	50.17/50.00	36.92/20.00	14.58/10.00
XGBoost supervised	52.17/50.13/86.67	30.58/27.47/78.08	62.50/62.50	56.17/37.33	18.92/15.71
TabPFN supervised	54.50/52.35/88.42	32.83/28.67/82.00	63.92/63.92	58.83/38.81	21.58/17.37
K-means	25.60/23.92/62.74	20.83/20.76/58.71	45.92/45.84	29.08/25.60	9.72/9.73
SCAN	19.31/17.92/57.45	20.26/20.17/54.45	46.92/46.89	34.09/27.87	11.77/10.59
SCARF	23.19/23.33/68.50	20.35/20.28/54.34	50.03/50.03	18.74/17.53	10.14/10.13
INSTRUCTOR	24.34/25.03/65.05	21.65/19.17/60.16	49.50/49.46	25.90/20.05	10.62/8.26
CLIP clustering	28.26/28.04/69.33	19.77/19.74/53.56	49.25/49.21	24.95/22.26	10.19/10.18
ClusterLLM	27.50/27.31/65.17	19.52/18.77/53.72	49.42/49.38	20.53/17.15	11.14/10.28
GoalEx	18.33/25.13/50.00	23.40/20.11/65.00	44.58/44.51	30.71/18.63	14.28/11.11
TableDC	17.67/18.96/64.67	19.95/20.06/54.34	50.92/50.88	24.92/20.91	15.76/**12.91**
MiniLM	31.83/25.00/79.50	24.00/19.94/69.67	49.83/50.00	30.08/17.61	7.50/10.00
Gen-Z	23.25/28.83/59.92	21.75/19.92/56.67	52.67/52.69	36.92/20.00	10.75/10.95
CLIP prototype	26.58/26.48/70.67	23.00/**26.56**/56.75	50.92/50.94	29.25/23.75	12.50/11.99
LLM tree induction	29.42/27.07/76.08	26.00/22.55/64.00	56.58/56.60	31.08/20.52	8.75/9.07
E0 (direct LLM)	**36.50/30.10/80.75**	**28.67**/25.70/**74.17**	56.75/56.85	**49.42**/29.46	**16.58**/12.32
E1 (ours)	35.00/29.01/79.75	28.08/25.43/72.42	**61.83/61.84**	48.25/**29.87**	15.83/11.83

Table 14: NHIS results under *weak* feature disclosure. Notation and metric-wise ranking follow Table 13.

Method	AGE (A/R/W ↑)	SCH (A/R/W ↑)	SEX (A/R ↑)	MAR (A/R ↑)	OCC (A/R ↑)
Train Majority	18.17/25.00/50.00	24.08/20.00/69.75	50.17/50.00	36.92/20.00	14.58/10.00
XGBoost supervised	58.83/58.05/95.83	31.58/29.23/84.25	66.08/66.07	66.75/47.74	21.00/18.69
TabPFN supervised	61.92/60.37/96.75	36.08/32.60/88.25	66.42/66.40	69.92/51.27	26.67/23.06
K-means	25.43/23.73/62.45	21.53/20.79/55.93	50.00/49.97	19.10/17.47	10.08/10.02
SCAN	22.78/21.75/60.46	22.05/21.98/53.60	49.58/49.57	22.17/19.63	11.96/10.98
SCARF	24.72/25.98/67.68	20.16/20.22/54.05	50.02/50.02	19.97/19.84	10.23/10.23
INSTRUCTOR	24.13/25.21/66.89	21.46/19.09/62.12	51.92/51.89	22.50/21.31	11.77/8.83
CLIP clustering	28.26/28.04/69.33	20.17/20.32/54.18	53.75/53.74	20.19/20.14	10.27/10.29
ClusterLLM	27.50/27.31/65.17	23.58/23.72/57.41	49.42/49.32	23.08/17.31	11.09/10.44
GoalEx	15.92/15.29/64.67	23.25/20.23/63.72	50.08/50.07	30.71/22.24	13.90/10.02
TableDC	19.19/18.50/52.43	21.94/20.53/55.90	50.83/50.80	27.58/25.72	11.35/9.52
MiniLM	31.83/25.00/79.50	24.00/19.94/69.67	49.83/50.00	25.25/22.84	7.50/10.00
Gen-Z	27.75/34.76/61.58	21.83/20.08/56.75	56.42/56.51	36.92/20.00	10.67/10.96
CLIP prototype	**38.75/40.39**/77.58	29.00/**31.45**/65.58	57.58/57.63	35.83/34.84	12.00/11.91
LLM tree induction	29.42/27.07/76.08	17.42/21.10/43.00	51.42/51.44	25.17/20.12	14.17/10.53
E0 (direct LLM)	32.92/28.92/93.33	33.50/30.08/**88.25**	58.50/58.60	57.67/41.30	**18.17/13.65**
E1 (ours)	32.08/28.40/**93.67**	**34.17**/31.19/85.08	**62.92/62.93**	**57.92/43.14**	17.92/13.48

the seven closed-set ACS targets under *prior* disclosure, and all five NHIS targets under the three disclosure conditions. The *strict* and *weak* results use the final candidate-score, adaptive-context, and product-fusion protocol. The available *prior* rows retain the ordinary evidence budgets rather than the *prior*-aware reallocation in Equation 4; they therefore measure sensitivity to *prior* prompting and are not used to validate that extension across models.

Table 15: Llama results on ACS, averaged equally over four states. Closed-set cells contain accuracy/macro recall, with within-one-level accuracy added for AGE and SCH; LOC contains $\mathrm{Acc}\,@100\,\mathrm{km}$/median km. *Prior* disclosure was evaluated for the closed-set targets. Bold marks the better E0 or E1 value for each metric; ties are unmarked.

Disclosure	Stage	AGE (A/R/W ↑)	SCH (A/R/W ↑)	SEX (A/R ↑)	MAR (A/R ↑)	INC (A/R ↑)	OCC (A/R ↑)	POB (A/R ↑)	LOC (100 ↑/Med ↓)
Strict	E0	30.38/34.11/63.83	25.81/21.72/81.00	48.77/50.00	30.94/19.98	62.02/50.00	36.46/**16.48**	30.00/5.00	38.23/203.35
	E1	**32.06/34.15/70.00**	**25.94/22.03**/81.00	**49.85/51.01**	30.94/19.98	62.02/50.00	**36.69**/16.29	30.00/5.00	**39.31/202.44**
Weak	E0	33.42/**38.43**/71.69	25.62/21.07/81.69	59.04/59.67	43.23/25.04	63.04/51.43	**38.88/17.54**	30.00/5.00	**37.10**/203.68
	E1	**33.79**/38.40/**72.50**	**25.98/21.53/82.31**	**59.44/59.93**	**46.65/26.45**	**71.98/64.34**	38.81/17.48	30.00/5.00	36.96/**198.73**
Prior	E0	32.46/26.63/81.90	30.67/32.72/88.29	**58.65/57.86**	57.00/30.91	**62.23**/50.39	**38.21**/16.10	28.33/4.76	–
	E1	**35.12/30.86/81.98**	**31.06/32.97/89.19**	56.29/55.81	**59.27/34.19**	59.10/**52.19**	38.19/**16.23**	**29.17/4.90**	–

Table 16: Llama results on NHIS. Cells contain accuracy/macro recall (%), with within-one-level accuracy added for AGE and SCH. Bold marks the better E0 or E1 value for each metric; ties are unmarked.

Disclosure	Stage	AGE (A/R/W ↑)	SCH (A/R/W ↑)	SEX (A/R ↑)	MAR (A/R ↑)	OCC (A/R ↑)
Strict	E0	31.83/25.00/79.42	**25.25/22.88**/63.58	50.17/50.00	40.33/22.57	14.17/13.94
	E1	31.83/25.00/79.42	24.17/21.68/**64.08**	50.17/50.00	**40.92/24.03**	**14.50/14.65**
Weak	E0	**32.75/26.16**/80.92	**35.75/31.57**/81.17	57.50/57.36	55.58/37.71	**19.50**/15.86
	E1	32.17/25.79/**82.50**	35.25/30.92/**82.08**	**59.67/59.53**	**57.58/43.29**	18.58/**16.19**
Prior	E0	**31.83/26.48/79.25**	25.08/20.85/70.42	50.17/50.00	39.25/21.72	**14.83/10.24**
	E1	21.92/20.95/72.50	**25.50/21.20/70.75**	50.17/50.00	**41.92/23.65**	12.75/9.03

Llama yields positive transfer for several targets, but the complete results show that the effect is concentrated rather than uniform. On ACS, the *strict* and *weak* gains are concentrated in a subset of targets, and *prior* disclosure improves AGE, SCH, MAR, and POB while reducing SEX and INC accuracy. POB collapses to a dominant prediction under *strict* and *weak* disclosure, and LOC improves both geographic metrics only under *strict* disclosure. On NHIS, *weak* disclosure improves SEX and MAR, whereas *prior* disclosure sharply degrades AGE despite improving MAR. Four degenerate contexts fall back exactly to E0 under the predefined rule. These results support transfer of the mechanism across an LLM family and a second dataset without establishing model-independent or target-independent reliability. Older single-answer runs and the superseded missing-label-completion variant are excluded because their prediction and context definitions do not match the final protocol.

E.3 Closed-source model comparison

Tables 17 and 18 report the GPT-4.1 snapshot on matched ACS New York and NHIS samples. Across the 36 closed-set cells, its E0 accuracy exceeds Qwen3-14B in 34 cells and its macro recall in 27. Unweighted cell means are 47.92% versus 42.14% for accuracy and 39.47% versus 34.92% for recall. The same E1 construction does not produce uniformly larger increments from GPT-4.1: mean accuracy changes are +0.92, -0.63, and -0.67 points on ACS under *strict*, *weak*, and *prior* disclosure, and -0.13, -1.15, and +0.57 points on NHIS. GPT-4.1's *prior* rows retain ordinary evidence budgets, so they do not validate the prior-aware allocation used in the main results. These are matched-model comparisons within the covered cells, not four-state ACS averages.

A separate New York LOC diagnostic reranks GPT-4.1's original top-five place candidates using a hierarchical geographic predictor. Under *strict* disclosure, its $\mathrm{Acc}\,@100\,\mathrm{km}$ changes from 46.83% to 47.00% and median error from 106.7 to 106.1 km; under *weak* disclosure, threshold accuracy

Table 17: GPT-4.1 closed-set results on ACS New York (%). Each cell gives A/R, with W added for AGE and SCH. The *prior* rows use weak features and the same E1 context rule as the other GPT-4.1 rows.

Disclosure	Target	E0 A/R(/W)	E1 A/R(/W)
Strict	AGE	37.92/35.68/89.00	37.83/36.58/89.75
	SCH	38.58/33.08/88.08	40.83/35.28/87.92
	SEX	63.00/63.18	63.08/63.18
	MAR	47.00/24.15	45.75/23.69
	INC	73.17/66.27	79.08/76.83
	OCC	41.17/19.12	40.50/18.81
	POB	30.42/5.48	30.58/5.59
Weak	AGE	47.92/46.84/94.33	47.17/47.77/94.25
	SCH	34.83/38.27/82.17	33.67/38.18/79.67
	SEX	68.08/68.42	66.17/66.30
	MAR	57.00/44.25	54.00/45.00
	INC	80.67/77.29	81.75/80.47
	OCC	42.42/19.60	43.67/20.79
	POB	30.42/5.35	30.50/5.50
Prior	AGE	46.08/43.30/92.50	48.50/47.12/94.58
	SCH	39.00/37.71/82.00	37.17/39.38/78.17
	SEX	69.08/69.10	68.00/68.09
	MAR	58.33/44.79	52.83/45.54
	INC	79.67/77.63	80.42/80.58
	OCC	42.58/19.68	43.17/20.41
	POB	30.75/6.09	30.75/6.13

Table 18: GPT-4.1 closed-set results on NHIS (%). Notation follows Table 17.

Disclosure	Target	E0 A/R(/W)	E1 A/R(/W)
Strict	AGE	40.08/32.27/84.00	43.83/35.62/87.25
	SCH	31.83/27.00/76.92	30.00/25.37/76.00
	SEX	60.17/60.09	57.17/57.07
	MAR	52.67/32.15	53.33/32.73
	OCC	19.00/14.71	18.75/14.04
Weak	AGE	42.83/37.90/95.58	45.92/40.55/97.50
	SCH	38.92/34.62/87.08	31.67/28.70/85.67
	SEX	66.17/66.13	66.67/66.63
	MAR	64.83/50.57	64.08/50.64
	OCC	22.75/19.01	21.42/18.36
Prior	AGE	36.50/32.28/90.33	43.00/39.19/95.58
	SCH	36.42/32.38/86.92	34.17/30.33/84.92
	SEX	66.08/66.03	64.67/64.65
	MAR	65.92/51.40	63.67/50.43
	OCC	22.92/19.21	25.17/21.30

changes from 47.08% to 47.17% while median error remains 104.4 km. This location-specific reranker differs from the main LOC instantiation and is not included in the closed-set means.

E.4 Class-level and resolution-dependent recovery

Macro recall in every table averages over the predefined label set, including values never predicted by the attack. Consequently, a gain in accuracy can coexist with unchanged or lower macro recall when propagation favors already common values. This pattern is most evident for OCC and POB, whose larger label spaces contain values receiving no selected context. Class-level quantities are used diagnostically rather than to select the reported method.

AGE and SCH are also evaluated at their native ordinal resolution. Within-one-level accuracy counts a prediction as close when its rank differs from the reference by at most one, while exact accuracy remains the private-value recovery measure. For ACS AGE, E1 raises within-one-level accuracy from 74.96% to 81.17% under *strict* disclosure and from 78.77% to 85.77% under *weak* disclosure. SCH changes are smaller and can be negative, showing that coarse ordinal agreement does not imply exact recovery. The NHIS tables display the same distinction.

LOC separates candidate generation from final selection. Every Qwen LOC result used in the main tables contains 1,200 geocoded final predictions. The LLM's five generated places define the recoverable candidate set; the tabular stage only reranks that set. Figure 4 shows that expanding this list improves both closed-set OCC top-k recovery and geographic LOC coverage, leaving room for later candidate selection but not constituting a table-level gain by itself. Under *strict* disclosure, E1 changes $\mathrm{Acc}\,@100\,\mathrm{km}$ from 39.15% to 38.83% while reducing median error from 184.42 to 178.37 km. Under *weak* disclosure, it raises the threshold accuracy from 39.06% to 41.08% and reduces median error from 162.74 to 156.24 km. These distance measures capture geographic proximity and must not be interpreted as exact place-name recovery or mixed with closed-set macro averages.

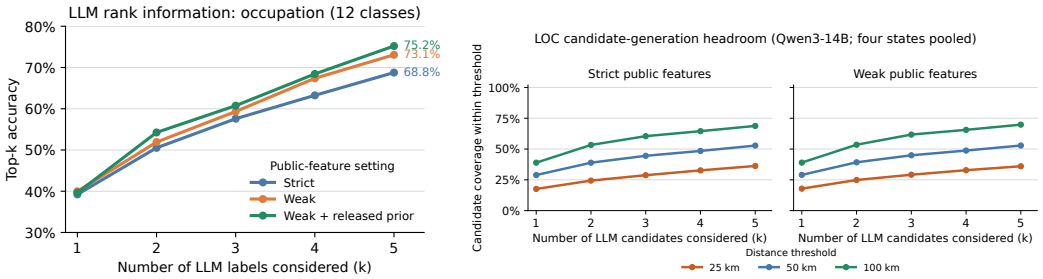


Figure 4: Candidate coverage in Qwen3-14B outputs before table-level inference. **Left:** for the 12-class ACS occupation target, top-k accuracy rises as more entries from the closed-set LLM ranking are retained. **Right:** for open-vocabulary ACS location, coverage is the fraction of samples with at least one of the first k generated places within the stated distance of the reference location, pooled over four states. These curves characterize the candidate space available to E1; they do not measure its final recovery.

F Ablations and Diagnostic Analyses

F.1 Context selection and downstream learner

Table 19 compares the adaptive selector with two fixed-budget controls on the same all-label Qwen3-14B score matrices and a common direct tabular endpoint. For each class c, both controls use the E0 bucket B_c and select at most 100 samples; if $|B_c| < 100$, all samples in that bucket are selected. Margin ranking orders B_c by the difference between the largest and second-largest normalized LLM scores for each sample. Score ranking instead orders it by s_{ic}, the normalized score of the E0-predicted class c. Thus the controls share the same candidate pools and per-class cap and differ only in their within-class ranking. The summary averages the seven ACS closed-set targets equally under *strict* and *weak* disclosure. The adaptive selector improves both accuracy and macro recall over either fixed construction, although the target-wise matrix contains losses as well as gains.

Table 19: Context-budget ablation over 14 ACS setting–target cells. Values are unweighted means (%).

Method	Accuracy ↑	Macro recall ↑
E0 (direct LLM)	43.88	37.21
Fixed-100, margin ranking	43.94	38.62
Fixed-100, score ranking	44.35	38.99
Adaptive evidence budget	**46.67**	**39.24**

Table 20 holds the total context size Q, LLM scores, and TabPFN configuration fixed within each task. Frequency-Q assigns each E0 bucket a share of Q proportional to its size and keeps the original evidence ranking within buckets. The permutation diagnostic keeps the adaptive context samples and label counts but randomly reassigns those labels among the selected samples. All changes are measured against the same E0 predictions.

Table 20: Matched-Q context diagnostics. Entries are mean accuracy/macro-recall changes from E0 in percentage points, averaged over four states and all seven closed-set targets for ACS and over five target cells for NHIS.

Dataset	Disclosure	Adaptive	Frequency-Q	Permuted labels
ACS	*Strict*	+3.35/+2.20	+1.93/+1.14	−4.86/−4.66
ACS	*Weak*	+2.43/+1.73	+1.37/+0.99	−9.47/−9.66
NHIS	*Strict*	+0.22/+0.71	−0.85/−0.42	−5.65/−4.85
NHIS	*Weak*	+0.85/+1.32	+0.87/+1.11	−5.27/−5.83

Table 21 isolates the effect of the *prior*-aware allocation. Both variants use the same Qwen score matrices, *prior* prompts, E0 buckets, total context budget, and product fusion. Ordinary E1 retains the evidence budgets from Equation 3; PRIORQUOTA reallocates that fixed total according to Equation 4. Entries are changes from E0 in percentage points.

Table 21: Prior-allocation ablation for Qwen3-14B. Both variants retain E0 pseudo-label assignments. A/R changes are relative to E0 in percentage points.

Dataset	Target	Ordinary E1 ΔA/ΔR	PRIORQUOTA ΔA/ΔR
ACS	AGE	+3.40/+3.16	+4.96/+5.40
	SCH	−0.60/−1.33	+2.58/+0.19
	SEX	+0.81/+1.15	+3.38/+3.82
	MAR	−0.08/+2.53	+0.27/+3.93
	INC	−1.92/−0.43	−1.54/−0.20
	OCC	+1.40/+0.99	+2.48/+2.59
	POB	−0.06/+0.07	−1.23/−0.16
NHIS	AGE	−0.08/+0.99	−5.00/−1.92
	SCH	+0.00/+0.10	−1.67/−1.16
	SEX	+0.50/+0.50	+5.83/+5.91
	MAR	+0.58/+0.92	+2.42/+3.45
	OCC	−1.17/−1.28	−0.42/−0.28

Table 22 reports the matched-context XGBoost control. It uses the same Qwen score matrices and adaptive context selected for E1; XGBoost is trained only on the pseudo-labels in that context, and all metrics cover the complete test set. This differs from the supervised XGBoost reference, which receives ground-truth training labels.

For LOC, replacing the regressor with XGBoost reduces $\mathrm{Acc}\,@100\,\mathrm{km}$ relative to E0 by 14.38 and 6.81 percentage points under *strict* and *weak* disclosure, while increasing median distance by 18.2 and 19.6 km. We therefore retain the TabPFN regressor for the main LOC results.

F.2 PSEUDO-LABEL QUALITY AND ERROR PROPAGATION

Table 23 audits the selected context and the prediction changes after the attack has been run. Pseudo-label precision is weighted by context size, and class coverage is the mean fraction of predefined target values represented in the selected context. Corrected errors count samples that are wrong

Table 22: XGBoost downstream-learner control for closed-set targets. Metric changes are percentage points relative to E0; wins count cells with higher accuracy than E0. ACS cells are state–target pairs.

Dataset	Disclosure	Δ accuracy	Δ macro recall	Δ macro-F1	Accuracy wins
ACS	*Strict*	+2.95	+2.07	+3.42	17/28
ACS	*Weak*	+2.15	+0.92	+2.40	19/28
ACS	*Prior*	−1.01	−0.47	−1.24	10/28
NHIS	*Strict*	+0.13	+0.22	+0.93	2/5
NHIS	*Weak*	−0.37	−0.24	+0.80	1/5
NHIS	*Prior*	−0.83	+0.03	−0.65	1/5

under E0 and correct under E1; introduced errors count the reverse transition. Ground-truth labels are used only for this retrospective analysis and never by context selection or prediction.

Table 23: Post-hoc context and error analysis for Qwen closed-set results. Context size is the mean per dataset–target cell; precision and class coverage are percentages.

Dataset	Disclosure	Context size	Precision	Class coverage	Corrected	Introduced
ACS	*Strict*	245.5	58.9	69.6	3,190	2,064
ACS	*Weak*	369.6	66.4	75.4	2,184	1,366
ACS	*Prior*	164.4	67.3	77.3	2,829	2,306
NHIS	*Strict*	159.6	49.2	69.0	350	337
NHIS	*Weak*	209.6	58.7	79.0	312	261
NHIS	*Prior*	73.6	47.6	67.0	384	370

Weak disclosure selects the largest contexts on both datasets. Under *prior* disclosure, the smaller contexts produce a positive net correction on ACS and NHIS, consistent with the accuracy changes in Table 9. Feature disclosure and aggregate target proportions thus provide different information to the attack.

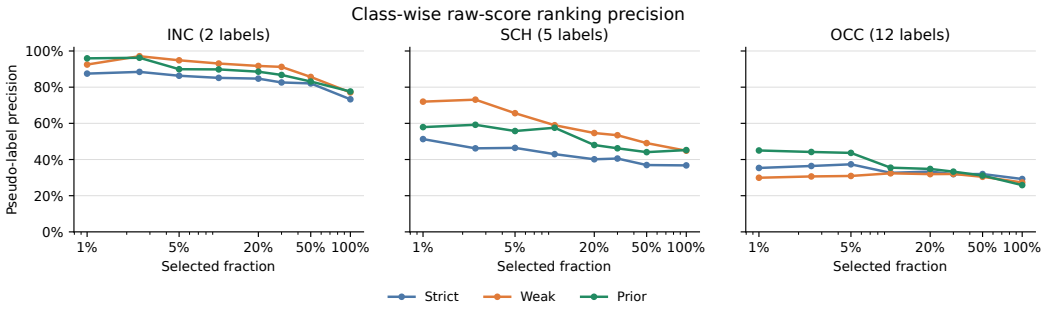


Figure 5: Class-wise raw-score ranking precision for three ACS targets. For each state, target value, and disclosure condition, samples in the E0 bucket for that value are ordered by their raw LLM score; precision is then computed for the leading fraction and macro-averaged across target values and states. High-ranked INC and SCH samples are substantially more reliable than the full bucket, whereas OCC shows weaker separation. The *prior* curve uses *weak* features plus released aggregate target proportions. Ground truth is used only for this post-hoc diagnostic.

Across 198 nonempty state–condition–target–value cells in the ACS *strict*/*weak* analysis, accumulated evidence mass correlates with the number of correctly pseudo-labeled selected samples (Pearson's $r = 0.956$). Since evidence mass also determines how many samples are selected, this count correlation does not by itself establish pseudo-label precision. Figure 5 instead evaluates ranking quality at fixed leading fractions of each predicted-class bucket. Correctness generally rises across evidence bins and then falls in the highest bin, so an extreme LLM score is not itself a correctness probability. Missing context values are concentrated among low-prevalence values with little average model support.

Visible profile (ACS New York)
Self-employed in an incorporated business; lawyer; legal-services industry; age 49; married; 60 working hours per week; personal income $150,000; health insurance; 30-minute car commute.

Released SCH prior
Some college/associate: 43.0%; high school: 26.0%; bachelor's: 12.4%; no diploma: 11.0%; graduate: 7.5%.

DeepSeek reasoning excerpt
"Becoming a lawyer typically requires a Juris Doctor (J.D.) degree, which is a graduate degree. ... Since the individual is a lawyer, it is more likely they have a graduate degree."

Returned scores
Graduate degree 0.50; some college/associate 0.25; bachelor's 0.15; high school 0.05; no diploma 0.05.

Withheld SCH label: Graduate degree.

Figure 6: A real DeepSeek-R1-Distill-Qwen-14B response for one held-out ACS New York sample under *prior* disclosure. The profile and released proportions are the actual prompt content, and the reasoning excerpt and scores are taken from the saved response; the row identifier is omitted. The model uses occupational semantics to place Graduate degree above the more prevalent population category, producing the score vector later consumed by E1.

F.3 Reasoning-model analysis

DeepSeek-R1-Distill-Qwen-14B provides explicit rationales that allow us to inspect how a reasoning model uses visible attributes. We focus on educational attainment (SCH), the target available under all three disclosure conditions on both datasets. Table 24 reports ACS four-state means and the NHIS results. This deliberately partial coverage is a diagnostic analysis rather than a benchmark-wide cross-model comparison; additional OCC, POB, and LOC runs are not combined into an incomplete cross-target average.

Table 24: DeepSeek SCH results. ACS values are four-state means; all entries are percentages. Each cell reports E0 $\rightarrow$ E1.

Dataset	Disclosure	Accuracy	Macro recall	Macro-F1
ACS	*Strict*	32.38 $\rightarrow$ 31.38	32.76 $\rightarrow$ 34.68	28.83 $\rightarrow$ 28.67
ACS	*Weak*	34.40 $\rightarrow$ 35.60	38.67 $\rightarrow$ 40.25	30.70 $\rightarrow$ 30.33
ACS	*Prior*	37.67 $\rightarrow$ 37.60	34.85 $\rightarrow$ 34.93	26.09 $\rightarrow$ 26.21
NHIS	*Strict*	25.50 $\rightarrow$ 24.92	23.60 $\rightarrow$ 23.35	21.11 $\rightarrow$ 18.34
NHIS	*Weak*	35.75 $\rightarrow$ 38.92	32.78 $\rightarrow$ 35.92	27.10 $\rightarrow$ 29.12
NHIS	*Prior*	31.58 $\rightarrow$ 33.58	28.42 $\rightarrow$ 30.79	27.00 $\rightarrow$ 28.31

The rationales often connect education to plausible correlates such as industry, employment status, working hours, occupation, income, and job benefits. They also repeatedly treat health behavior, sleep, commuting, insurance, and household stability as individualized evidence of educational attainment. These explanations are therefore coherent at the level of population associations but frequently rely on weak proxies that do not identify the hidden value of a particular sample. Reasoning length provides no corrective signal: across the audited DeepSeek tasks, correct outputs are not consistently longer or shorter than incorrect ones.

On ACS, E1 increases macro recall under all three conditions, although accuracy improves only under *weak* disclosure. On NHIS, both *weak* and *prior* disclosure improve all three metrics, whereas *strict* disclosure declines. Thus, table-level gains can also arise from a reasoning model's outputs. More broadly, explicit reasoning can connect visible attributes to a hidden value, while subsequent statistical transfer uses the reliability of these predictions across the released cohort.

F.4 Semantic inference and statistical dependencies

The experiments distinguish three sources of predictive information. E0 measures semantic inference from a single sample's visible fields and candidate meanings. The clustering and representation baselines use the released feature distribution without target labels, while supervised XGBoost and TabPFN show how predictable the target is when true training labels connect features to values. Their separation in Table 3 and Tables 10–14 shows that semantic prediction, unlabeled structure, and labeled statistical prediction are not interchangeable.

The unsupervised table baselines generally remain below E0 despite observing the full cohort, so access to other samples alone does not identify which anonymous structure corresponds to each private value. Conversely, supervised references often exceed E0, showing that useful feature–target dependencies exist but require a label correspondence. E1 occupies the intermediate setting: its only correspondence comes from selected LLM predictions. Its gains demonstrate that semantic outputs can sometimes bootstrap statistical transfer, while the per-target win counts in Table 9 show that additional rows do not guarantee usable supervision.

Earlier batch-prompt experiments asked the LLM to inspect groups of 64 samples directly, using 2,000-row cohorts and a different single-answer or top-k output protocol. They did not provide the all-label score matrix required by the present selector and are therefore treated only as historical diagnostics. Similarly, agreement between a learned representation and target labels does not establish recovery of the true conditional distribution. The matched E0–E1 comparison is the evidence used for the table-level claim because both stages start from the same per-sample score matrix and differ only by the constructed pseudo-context and downstream learner.

F.5 Comparison with prior tabular attribute-inference experiments

Appendix D of Staab et al. (2024) provides the closest prior tabular experiment. It evaluates zero-shot GPT-4 on the 2018 New York split of ACSIncome after filtering to people born outside the United States. Its 1,000-sample test set covers five targets—place of birth, race, education, income, and sex—with a different target-specific feature subset for each task. Predictions are made independently from serialized rows, and a majority classifier and supervised XGBoost models trained on the remaining labeled data provide reference points.

Our E0 follows the same broad question of whether an LLM can infer a withheld ACS attribute without labeled examples, but the protocols are not numerically interchangeable. We use 2024 data, four ACS states, matched 1,200-sample cohorts, consolidated target spaces fixed for the present study, and an all-label score protocol. We additionally evaluate NHIS, distinguish *strict*, *weak*, and *prior* disclosure, and include age, marital status, occupation, and current location among the targets. Most importantly, the prior ACS experiment evaluates independent row-wise inference; it does not construct pseudo-label context from the released cohort or measure an E1–E0 table-level increment. We therefore cite it as evidence that direct tabular LLM inference is a privacy threat, while treating the conversion of those predictions into dataset-level statistical supervision as the separate question studied here.