BOOSTING DIVERSITY AND RELEVANCE: PRIVACY-PRESERVING DEMONSTRATION SYNTHESIS FOR IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) rely on contextual information embedded in demonstrations to perform effective in-context learning (ICL), yet using demonstrations drawn from a private data pool may expose sensitive information when interacting with LLM servers. Existing demonstration synthesis methods based on differential privacy (DP) can mitigate such privacy risks, but they struggle to balance privacy and utility. Particularly, these methods often generate demonstrations with limited diversity and task relevance, while DP noise injected into token sampling distributions introduces semantic errors that accumulate during autoregressive generation. Motivated by these limitations, we propose DPD, a novel privacy-preserving algorithm to generate high-utility demonstrations under DP guarantees. *To address diversity and relevance gaps,* we develop a task-aligned encoder that maps private demonstrations into a task-aware embedding space, where a DP coverage-aware strategy privately releases coverage-weighted centers and retrieves candidate demonstrations for promoting semantic coverage. *To mitigate error accumulation,* we avoid perturbing token autoregressive generation and instead introduce the Joint Exponential Mechanism to efficiently extract DP-based keywords from candidates, guiding the advanced LLM to synthesize demonstrations with high utility. The overall DPD framework provides demonstration-level DP. Extensive experiments on standard benchmarks show that DPD outperforms token-level DP demonstration synthesis baselines while effectively mitigating membership inference attack, providing an effective solution for privacy-preserving ICL. Our code is available in https://anonymous.4open.science/r/DPD-0E00.

1 INTRODUCTION

Through In-Context Learning (ICL) (Brown et al., 2020), LLMs can quickly adapt to unseen tasks without fine-tuning the pretrained model. Specifically, incorporating task-relevant input-output pairs (i.e., **demonstrations**) into the prompt enables the model to generate context-conditioned responses for user queries. For example, as shown in Figure 1, prompting the model with symptom-diagnosis pairs of prior patients allows it to infer the diagnosis of a new patient given their symptoms. When carefully selected, such demonstrations can further improve the compositional generalization of LLMs (Qiu et al., 2022). Currently, most existing studies retrieve demonstrations from the demonstration pool based on two key criteria: **relevance** and **diversity**. Relevance focuses on selecting demonstrations that closely match the user queries, which can be measured through semantic similarity (Rubin et al., 2022) or structural alignment (Lyu et al., 2023). Beyond relevance, prior work has also emphasized the importance of diversity in demonstration selection (Zhang et al., 2022; Wang et al., 2025; Levy et al., 2023; Xu et al., 2026b), which improves coverage of semantic and syntactic variations in user queries.

However, reliance on informative demonstrations raises serious privacy concerns. Consider the real-world medical scenario in Figure 1: if demonstrations drawn from a private dataset, such as clinical records (Huang et al., 2019; Zhang et al., 2023; Yang et al., 2026), are sent to an LLM server, an adversary may attempt to infer whether a specific demonstration associated with an individual's medical history is included in the prompt. Specifically, leveraging the memory effect of LLM (Carlini et al., 2023), the adversary conducts black-box membership inference attacks (MIAs),

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.

such as input-output pair analysis (Yang et al., 2025b) and confidence-based analysis of predicted probabilities (Duan et al., 2023), to infer private demonstrations without accessing model parameters.

To mitigate such risks, prior studies (Tang et al., 2024; Gao et al., 2025) have explored **differential privacy** (DP) (Dwork & Roth, 2014) as a rigorous privacy safeguard. DP guarantees that the model output remains statistically indistinguishable regardless of whether any individual data is present, preventing attackers from inferring sensitive information from LLM outputs. Recently, Tang et al. (2024) propose a DP few-shot generation method to synthesize demonstrations from a private demonstration pool for secure ICL. Specifically, this method partitions the private pool into disjoint subsets, uses each subset to prompt a local language model, and then privately aggregates the resulting token probability distributions to synthesize demonstrations. Building on this idea, Gao et al. (2025) introduce a data-adaptive DP synthesis (AdaDP-Syn) algorithm to avoid adding unnecessary noise to tokens with high predictability, thus improving the utility for ICL.

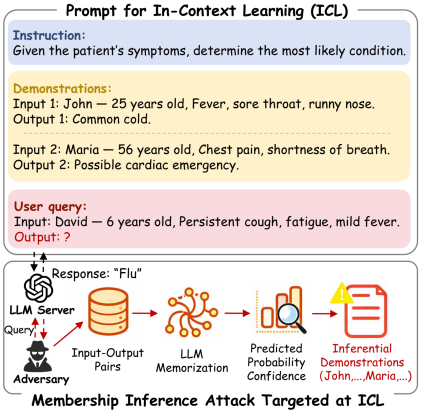


Figure 1: Illustration of ICL and the risk of MIA on private demonstration.

Despite these advances, the existing DP-based synthesis methods suffer from two clear limitations:

Limitation 1: Lack of Diversity and Task Relevance. DP noise provides formal privacy protection for generated demonstrations, but it also leads to low diversity and weak task relevance, as shown in Table 15, thereby significantly degrading ICL performance. Specifically, the low diversity arises from *majority preference bias* (Shumailov et al., 2024) introduced by DP perturbation: Since DP noise smooths the token distribution for privacy protection, common tokens with universal grammar are robustly preserved, whereas distinctive but low-probability tokens are sensitive to noise and tend to be suppressed, resulting in semantically monotonous outputs. Regarding task relevance, heavy noise induced by limited privacy budgets significantly distorts the token distribution, causing the generation process to deviate from task-relevant semantics. Consequently, the synthesized demonstrations often contain irrelevant fragments, as shown in Table 18.

Limitation 2: Semantic Error Accumulation. Existing methods generate demonstrations via a *token-to-token (autoregressive) process* (Finlayson et al., 2024), predicting each token from prior context. However, semantic errors induced by DP perturbation in early tokens will propagate and accumulate throughout subsequent generations (Zhu et al., 2026). Specifically, once initial tokens contain semantic errors due to DP noise, subsequent tokens conditioned on this corrupted context fail to recalibrate toward the proper semantics (Holtzman et al., 2020). Thus, semantic errors are progressively accumulated, not only leading to incoherent or erroneous outputs (e.g., broken strings as shown in Table 15), but also exacerbating the deviation from task-relevant topic (e.g., generating website-like text instead of number-labeled one in Table 15).

To address these limitations and improve ICL with high-quality demonstrations, we propose **Diverse Privacy-preserving Demonstration** (DPD) synthesis, a structured pipeline consisting of three key modules. To enhance diversity and task relevance, DPD employs a *Pretrained Task-aligned Encoder* to map private demonstrations into a task-aligned embedding space. Since these embeddings often overlap, indicating high semantic similarity, DPD introduces a *DP Coverage-aware Clustering Strategy* that privately releases representative cluster centers and retrieves semantically diverse demonstration candidates. To mitigate semantic error accumulation, DPD avoids directly perturbing token probability distributions. Instead, it extracts task-relevant keywords from the selected candidates and applies the *Joint Exponential Mechanism* (Gillenwater et al., 2022) to privatize them. These keywords provide semantic guidance for generating DP-protected demonstrations with high utility.

Our contributions are summarized as follows:

• *Innovative Approach.* We introduce a new DPD method to generate privacy-preserving demonstrations with semantic diversity and task relevance, thereby enabling efficient ICL. The novelty of DPD lies in our proposed *Coverage-aware Clustering Demonstration Retriever* with a *DP-Based Keyword Aggregation Mechanism*. The retriever privately releases representative cluster centers and retrieves semantically diverse demonstration candidates in the task-aligned embedding space, while

the aggregation mechanism extracts privatized keywords from these candidates as semantics anchors, guiding the generation of privacy-preserving demonstrations with high utility.

• *Theoretical Analysis.* We provide a rigorous privacy analysis showing that DPD satisfies demonstration-level DP under the addition or removal of any private demonstration record. We establish the privacy guarantees of the DP center release and conditional JEM-based keyword release, and derive the overall guarantee through the DP composition theorem. Based on the post-processing property of DP, the generated demonstrations can be queried indefinitely without additional budget.

• *Empirical Evaluation.* We empirically evaluate DPD against existing DP-based demonstration synthesis methods on standard benchmarks. DPD consistently outperforms all baselines across privacy settings and even approaches non-private baselines when $\epsilon \geq 8$. Moreover, DPD empirically improves privacy protection by reducing the average AUC of black-box MIAs from 0.91 to 0.51, approaching random guessing and substantially weakening the effectiveness of these attacks.

2 PRELIMINARIES

In-Context Learning. To predict the label $y \in \mathcal{Y}$ of a query q with ICL, we concatenate a sequence of n_{shot} demonstrations $\mathcal{D} = \{(q_1, y_1), ..., (q_{n_{\text{shot}}}, y_{n_{\text{shot}}})\}$, where (q_i, y_i) denotes a input-label pair. We consider text classification, information extraction and mathematical reasoning tasks. For the former with a limited label space $\mathcal{Y} = \{y_1, y_2, ..., y_c\}$, the probability of each label is computed by a pre-trained LLM as $\mathbb{P}_{\text{LLM}}(y_i \mid \mathcal{D} \oplus q)$, and the final prediction is obtained by selecting the label with the highest probability: $y = \arg\max_{y_i \in \mathcal{Y}} \mathbb{P}_{\text{LLM}}(y_i \mid \mathcal{D} \oplus q)$, where $\oplus$ denotes the concatenation operation. In contrast, information extraction and mathematical reasoning tasks involve an open-form label space $\mathcal{Y}$. In these tasks, the model predicts y by generating tokens sequentially until an end-of-sequence (EOS) token is produced, indicating the completion of the output.

Differential Privacy. DP ensures that an algorithm's output remains statistically indistinguishable under single-record changes. To achieve DP, several perturbation mechanisms are used: (I) the *Laplace Mechanism* (Dwork & Roth, 2014) adds calibrated noise to scalar numeric outputs; (II) the *Gaussian Mechanism* adds calibrated Gaussian noise to vector outputs; (III) the *Exponential Mechanism* (EM) (Dwork & Roth, 2014) samples discrete outputs based on utility scores; and (IV) the *Joint Exponential Mechanism* (Gillenwater et al., 2022) extends EM to jointly select a set of outputs while accounting for dependencies among them. The above concepts are detailed in Appendix B.

3 PROBLEM STATEMENT AND LIMITATIONS OF EXISTING SOLUTIONS

Problem Statement. Given a large private demonstration pool $\mathcal{D}_{\text{priv}} = \{(q_i, y_i)\}_{i=1}^{N}$ and a test query q_{test}, our goal is to privately select a compact subset of candidate demonstrations, and based on these candidates, generate a few privacy-preserving demonstrations $\mathcal{D}_{\text{syn}} = \{(q'_j, y'_j)\}_{j=1}^{n}$ $(n \ll N)$ that preserve both semantic diversity and task relevance. By feeding the concatenation $\mathcal{D}_{\text{syn}} \oplus q_{\text{test}}$ into the LLM server, the predicted output y_{test} is expected to align with the ground-truth label y_{true}.

Threat Model. As illustrated in Figure 1, we consider a black-box adversary targeting private demonstrations used in ICL. A victim submits to an LLM server a prompt composed of a private demonstration set $\mathcal{D}_{\text{priv}}$ containing sensitive personal information (e.g., patient records). The adversary has no access to internal model parameters, gradients, or hidden states. Their capability is limited to black-box API interaction with the LLM server, through which they can only observe the model's outputs. Due to the LLM memorization effect (Carlini et al., 2023; Xu et al., 2025), demonstrations used as context can induce abnormally high prediction confidence. This makes the LLM output probability $\mathbb{P}_{\text{LLM}}(y_i \mid \mathcal{D}_{\text{priv}} \oplus q)$ a side-channel, which can be exploited to perform the membership inference attack, determining whether a specific demonstration was included in the ICL prompt.

Privacy Model. To defend against the aforementioned risks, it is essential to ensure that the presence of any individual private demonstration does not significantly impact the distribution of synthesized output $\mathcal{D}_{\text{syn}}$. While satisfying this privacy constraint, the challenge is to maintain the diversity and task-relevance of the demonstrations. We formalize this privacy requirement as follows.

Definition 3.1 (Differentially Private ICL (Tang et al., 2024; Gao et al., 2025))**.** A randomized demonstration synthesis algorithm $\mathcal{M}$ satisfies (ϵ, δ)-DP if for any two adjacent demonstration pools

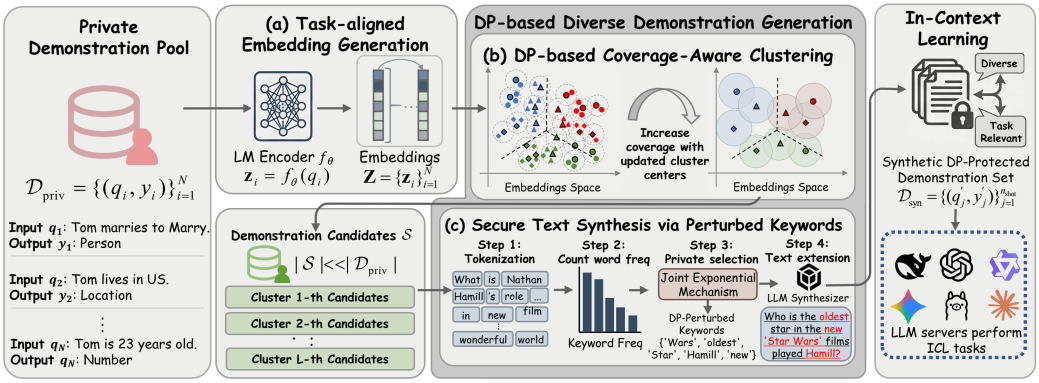


Figure 2: Overview of the DPD framework. (a) task-aligned embedding generation with a pretrained BERT model, (b) candidate retrieval via DP-based coverage-aware center release, and (c) privacy-preserving demonstration synthesis from privatized keywords using an advanced LLM API.

$\mathcal{D}, \mathcal{D}'$ that differ by adding or removing one demonstration record, and for all possible output sets $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$, it holds that $\mathbb{P}[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta$, where ϵ is the privacy budget, with smaller values providing stronger privacy. δ is the small failure probability of DP guarantees.

Existing Solutions and Limitations. To the best of our knowledge, existing DP-based ICL methods primarily follow two technical routes: **gradient perturbation** and **output distribution perturbation**. In the former category, Prompt-DPSGD (Duan et al., 2023) updates soft prompts via noisy gradients. However, its reliance on input gradients renders it incompatible with commercial black-box APIs. To relax this constraint, output distribution perturbation methods have been proposed: DP-ICL (Wu et al., 2024) performs private inference through noisy consensus over queries with private demonstrations, but consumes privacy budget per query, limiting the number of requests. To enable unlimited queries under a fixed budget, recent works (Tang et al., 2024; Gao et al., 2025) partition the private demonstration pool into disjoint subsets to prompt local LLMs, subsequently privately aggregating the resulting token probability distributions to synthesize reusable DP-based demonstrations.

Limitations. While existing methods prioritize the privacy guarantees of demonstration synthesis, they often neglect the utility of the generated demonstrations. Consequently, they suffer from two critical limitations: **lack of diversity and task relevance**, and **semantic error accumulation** (as detailed in Section 1). These deficiencies lead to suboptimal performance in downstream ICL tasks.

4 Our Proposal: DPD

To address the above limitations, we propose DPD to generate high-utility demonstrations while ensuring demonstration-level DP, offering an effective solution for privacy-preserving ICL.

Overview. Figure 2 and Algorithm 3 provide an overview of DPD framework, including: (I) *Task-aligned Embedding Generation*: A pretrained BERT model encodes private demonstrations into a *task-aligned* embedding space, after which a Johnson-Lindenstrauss (JL) projection and L_2 clipping produce bounded embeddings (Section 4.1). (II) *DP Coverage-Aware Clustering*: Based on these embeddings, this module privately releases coverage-weighted centers and retrieves candidate demonstrations, with the goal of promoting semantic coverage (Section 4.2). (III) *Privacy-preserving demonstration Synthesis*: Privatized keywords are extracted from candidate demonstrations via the Joint Exponential Mechanism and used as *task-relevant anchors* to guide privacy-preserving demonstration generation, *mitigating semantic error accumulation* (Section 4.3).

4.1 Task-aligned Embedding Generation

Motivation. A demonstration retriever is required to select representative demonstrations for ICL from a large demonstration pool (Wang et al., 2025; 2024). Liu et al. (2025) observed that an effective retriever tends to favor demonstrations that are closer to the user query in latent space, even when task-relevant and task-irrelevant demonstration candidates exhibit similar semantics and syntax.

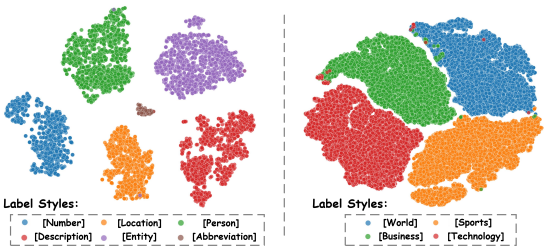
Task-aligned Embedding. Motivated by this observation, we retrain a lightweight language model (LM)-based encoder (e.g., Sentence-BERT (Reimers & Gurevych, 2019)) to encode private demonstrations in the task-aligned embedding space. Let f_θ denote the LM encoder. For text classification tasks, c_ϕ represents an additional linear classifier. Given any input text q, the encoder produces a latent embedding $\mathbf{z} = f_\theta(q)$. The classifier then estimates the probability of label y_i as

$$c_{\boldsymbol{\phi}}\left(Y = y_i \mid \mathbf{z}\right) = \exp\left(\mathbf{z} \cdot \boldsymbol{\phi}_i\right) \Big/ \sum\nolimits_j \exp\left(\mathbf{z} \cdot \boldsymbol{\phi}_j\right), \tag{1}$$

where $\boldsymbol{\phi}_i$ is the weight vector associated with label y_i. This model is trained by minimizing the cross-entropy loss over the public demonstration set $\mathcal{D}_{\text{pub}}$:

$$\mathcal{L}_c = -\sum\nolimits_{(q,y)\in\mathcal{D}_{\text{pub}}} \sum\nolimits_{i=1}^{C} \mathbf{1}\left[y = y_i\right] \log c_\phi\left(Y = y_i \mid f_\theta(q)\right), \tag{2}$$

where C is the number of classes and $\mathbf{1}[\cdot]$ is the indicator function. To prevent membership inference attacks on encoder outputs, a common approach is DP-SGD (Abadi et al., 2016), which adds Gaussian noise to gradients to ensure DP. However, applying it to large-scale models like BERT often leads to unstable optimization and poor utility (Du et al., 2023). Instead, we train a task-aligned encoder on a non-private demonstration set $\mathcal{D}_{\text{pub}}$, accessed from large public corpora (Maldonado et al., 2019) or generated by LLMs. Since LLM-generated demonstrations can replicate the general structure, topic coverage, and style of private data without including sensitive information (Yang et al., 2026), the potential domain shift between $\mathcal{D}_{\text{pub}}$ and $\mathcal{D}_{\text{priv}}$ can be largely mitigated. During training, both θ and ϕ are optimized jointly via gradient descent, so that input texts with the same labels form clusters in the embedding space, as illustrated in Figure 3. This retraining process enables the encoder to produce task-aligned embeddings of demonstration inputs, effectively aligning their semantic representations with the corresponding labels.



(a) TREC Dataset (b) AGNews Dataset

Figure 3: t-SNE visualization of task-aligned embeddings on the two datasets.

For information extraction tasks, the label space is open-form, making it infeasible to retrain the encoder with explicit label supervision. Therefore, we directly employ a pre-trained BERT model to obtain the text embeddings of private demonstrations in the semantic space, which serves as the basis for subsequent similarity computation and demonstration retrieval.

Before private-data access, a fixed JL dimension d_{JL} and random seed generate a data-independent Gaussian matrix $R \in \mathbb{R}^{d_{\text{JL}} \times d}$ with $R_{ab} \sim \mathcal{N}(0, 1/d_{\text{JL}})$, incurring no privacy cost. For each encoder output $\mathbf{z}_i = f_\theta(q_i)$, the projection $\mathbf{z}_i^{\text{JL}} = R\mathbf{z}_i$ reduces its dimension while approximately preserving pairwise distances (Johnson et al., 1984), followed by fixed L_2 clipping: $\bar{\mathbf{z}}_i = \text{Clip}_C(\mathbf{z}_i^{\text{JL}}) = \mathbf{z}_i^{\text{JL}} \min\big(1, C/\|\mathbf{z}_i^{\text{JL}}\|_2\big)$, where $C > 0$ is fixed before private-data access, ensuring $\|\bar{\mathbf{z}}_i\|_2 \le C$.

4.2 Diverse Demonstration Retrieval via DP Coverage-Aware Clustering

Motivation. As illustrated in Figure 3, it can be observed that most demonstration embeddings overlap in the mapping space, suggesting high semantic or syntactic similarity among them. Such redundancy not only slows down LLM inference when all demonstrations are provided, but also offers limited improvement for the compositional generalization of LLMs during ICL (Ye et al., 2023).

Coverage-Aware Demonstration Retrieval. We propose a *DP Coverage-Aware Clustering-Based Demonstration Retriever* that sequentially downweights regions near previously released centers. This mechanism is intended to promote semantics coverage. Specifically, let $\mathcal{E} = \{\bar{\mathbf{z}}_i\}_{i=1}^N$ denote the set of projected and clipped embeddings, where N is the number of private demonstrations. We release K centers in K successive rounds, indexed by $t \in [K] = \{1, \ldots, K\}$.

Let $\mathcal{F}_{t-1} = \{\widetilde{\mathbf{c}}_1, \ldots, \widetilde{\mathbf{c}}_{t-1}\}$ denote the DP centers released before round t, where $\widetilde{\mathbf{c}}_j$ is the center released in round j and $\mathcal{F}_0 = \emptyset$. For every embedding $\bar{\mathbf{z}}_i$, we assign a coverage-aware weight

$$w_i^{(t)} = \begin{cases} 1, & t = 1, \\ \min\left(1, \dfrac{d(\bar{\mathbf{z}}_i, \mathcal{F}_{t-1})}{r_{\text{cov}}}\right), & t > 1, \end{cases} \tag{3}$$

where $d(\bar{\mathbf{z}}_i, \mathcal{F}_{t-1}) = \min_{\widetilde{\mathbf{c}} \in \mathcal{F}_{t-1}} \|\bar{\mathbf{z}}_i - \widetilde{\mathbf{c}}\|_2$ is the distance to the nearest previously released center. The coverage radius $r_{\text{cov}} = 2C$ is fixed before accessing private data. Since both embeddings and released centers are clipped to the radius-C ball, their distance is at most $2C$. The weight $w_i^{(t)} \in [0, 1]$ means that embeddings close to existing centers receive small weights, whereas embeddings in under-covered regions retain larger weights. In the first round, all embeddings receive unit weight because no center has yet been released.

At round t, we summarize the weighted embedding set using the weighted count and vector sum

$$n_t = \sum\nolimits_{i=1}^{N} w_i^{(t)}, \qquad \mathbf{s}_t = \sum\nolimits_{i=1}^{N} w_i^{(t)} \bar{\mathbf{z}}_i. \tag{4}$$

Here, n_t is the effective weighted number of records and $\mathbf{s}_t$ is their weighted embedding sum. Their ratio points toward a region that is not sufficiently represented by $\mathcal{F}_{t-1}$. Since both n_t and $\mathbf{s}_t$ depend on the private demonstration pool, we perturb them using the Laplace and Gaussian mechanisms:

$$\widetilde{n}_t = n_t + \text{Lap}(1/\epsilon_{n,t}), \qquad \widetilde{\mathbf{s}}_t = \mathbf{s}_t + \mathcal{N}(\mathbf{0}, \sigma_t^2 C^2 I). \tag{5}$$

where $\epsilon_{n,t}$ and $\epsilon_{s,t}$ denote the privacy budgets at round t, I is the identity matrix, and the Gaussian noise scale is calibrated as $\sigma_t = \sqrt{2\ln(1.25/\delta_t)}/\epsilon_{s,t}$. Since $w_i^{(t)} \leq 1$ and $|\bar{\mathbf{z}}_i|_2 \leq C$, adding or removing one demonstration from the private pool changes n_t by at most 1 and $\mathbf{s}_t$ by at most C.

The DP center for round t is obtained by post-processing the noisy statistics:

$$\widetilde{\mathbf{c}}_t = \text{Clip}_C(\widetilde{\mathbf{s}}_t / \max(\widetilde{n}_t, n_{\min}))\,, \tag{6}$$

where $n_{\min} = 1$ is a fixed lower bound that prevents an unstable denominator when the noisy count is small or negative. The final clipping guarantees $\|\widetilde{\mathbf{c}}_t\|_2 \leq C$. We then update $\mathcal{F}_t = \mathcal{F}_{t-1} \cup \{\widetilde{\mathbf{c}}_t\}$ and repeat the procedure. For the fixed release of K centers, we allocate the vector-sum budget ϵ_1, count budget ϵ_2 and failure probability δ_1 uniformly across rounds: $\epsilon_{s,t} = \epsilon_1/K, \epsilon_{n,t} = \epsilon_2/K, \delta_t = \delta_1/K$.

By the DP composition theorem (Dwork & Roth, 2014), the center set $\mathcal{F}_K$ released over K rounds satisfies $(\epsilon_1 + \epsilon_2, \delta_1)$-DP. After obtaining the K centers, each private embedding is assigned to exactly one nearest center: Let $a(i) = \arg\min_{k \in [K]} \|\bar{\mathbf{z}}_i - \widetilde{\mathbf{c}}_k\|_2$ denote the assigned center index of the i-th demonstration, and let $\mathcal{I}_k = \{i : a(i) = k\}$ denote the demonstrations assigned to center k. We then select the nearest Top-W demonstrations within each cluster: $\mathcal{S}_k = \text{Top-}W(\{(q_i, y_i) : i \in \mathcal{I}_k\}, \widetilde{\mathbf{c}}_k)$.

Unique assignment makes the sets $\mathcal{I}_1, \ldots, \mathcal{I}_K$ disjoint, so adding or removing one demonstration affects the Top-W candidates set of at most one cluster when the DP centers are fixed. Finally, the union of all selected demonstrations forms the **candidate demonstration set** $\mathcal{S} = \bigcup_{k=1}^{K} \mathcal{S}_k$, which can significantly reduce the search space of the private demonstration pool from N to $K * W$. Table 17 provides empirical evidence on candidate diversity and demonstration-pool coverage; we do not infer semantic-mode recovery guarantees from the update rule alone.

Lemma 4.1 (The sensitivity of DP Coverage-Aware Clustering strategy)**.** *Conditioned on the preceding DP outputs, the weighted count and vector-sum queries satisfy $\Delta_{n_t} \leq 1$ and $\Delta_{\mathbf{s}_t} \leq C$ under adding or removing one demonstration from the private pool. Consequently, the complete DP Coverage-Aware Clustering strategy satisfies $(\epsilon_1 + \epsilon_2, \delta_1)$-DP. The proof is given in Appendix C.*

4.3 Privacy-preserving Demonstration Synthesis

Motivation. Given the candidate sets $\{\mathcal{S}_k\}_{k=1}^{K}$, we fix a shared vocabulary $\mathcal{V} = \{v_1, \ldots, v_b\}$ from the pretrained LLM tokenizer before accessing private data, and compute each word's document frequency within cluster k: $c_j^{(k)} = \sum_{x \in \mathcal{S}_k} \mathbf{1}[v_j \text{ occurs in } x]$, where $c_j^{(k)}$ counts the demonstrations in $\mathcal{S}_k$ containing v_j, with each demonstration contributing at most once. Compared with self-information (Li et al., 2023) and mutual information (Church & Hanks, 1990), document frequency robustly captures semantic contributions, allowing high-frequency words to summarize the text (Wu et al., 2024). Since these words may reveal sensitive information, releasing the top-T words through T applications of the Exponential Mechanism incurs repeated privacy composition. Instead, we apply the Joint Exponential Mechanism (JEM) introduced (Gillenwater et al., 2022) to select an ordered sequence of T distinct words in a single step, thereby avoiding repeated composition.

Demonstration Synthesis. Consider a fixed cluster k, and let $\mathcal{S}_T$ denote the set of all ordered sequences of T distinct words from the fixed vocabulary. Naively enumerating $\mathcal{S}_T$ is computationally

infeasible because it contains $b!/(b-T)!$ sequences. Let $\mathbf{c} = (c_1, \ldots, c_b)$ be the document-frequency vector, let $s^* = (s_1^*, \ldots, s_T^*)$ denote the true top-T sequence, and let $s = (s_1, \ldots, s_T) \in \mathcal{S}_T$ be a candidate sequence. Its utility is defined as

$$u(s^*, s) = -\max_{i \in [T]} \left(c_{s_i^*} - c_{s_i} \right), \tag{7}$$

where $c_{s_i^*}$ and c_{s_i} are the document frequencies of the i-th words in s^* and s, respectively. A larger utility indicates that the candidate sequence is closer to the true top-T sequence. Although $\mathcal{S}_T$ is combinatorially large, each utility score has the form $-(c_i - c_j)$ for $i \in [T], j \in [b]$, giving only bT possible values. Motivated by this, the **efficient sampling strategy** is carried out as follows:

Initially, all words are stably sorted by document frequency such that $c_1 \geq \cdots \geq c_b$, with ties broken by fixed vocabulary indices when words have the same frequency. We define the utility matrix with entries $\mathbf{U}_{ij} = -(c_{s_i^*} - c_j)$, and we construct its perturbed verson:

$$\widetilde{\mathbf{U}}_{ij} = \mathbf{U}_{ij} - z_{ij}, \quad z_{ij} = \big(b(T-i) + j\big)/(2bT) \in (0, \tfrac{1}{2}]. \tag{8}$$

Since $\mathbf{U}_{ij}$ is integer-valued, z_{ij} preserves the ordering between distinct utility values and only breaks ties. Importantly, $\widetilde{\mathbf{U}}$ is used only to assign each sequence to a unique layer, whereas the Exponential Mechanism weight is computed using the original utility $\mathbf{U}_{ij}$. We flatten $\widetilde{\mathbf{U}}$ and sort its entries in descending order: $\widetilde{\mathbf{U}}(1) > \widetilde{\mathbf{U}}(2) > \cdots > \widetilde{\mathbf{U}}(bT)$. Each entry $\widetilde{\mathbf{U}}(a) = \widetilde{\mathbf{U}}_{i_a, j_a}$ corresponds to the a-th sorted layer, originating from cell (i_a, j_a). For each layer a with threshold value $\widetilde{\mathbf{U}}(a)$, we compute a prefix length for every row (position) $r \in [1, ..., T]$:

$$t_r(a) = \max \left(\left\{ j' \in [b] : \widetilde{\mathbf{U}}_{rj'} > \widetilde{\mathbf{U}}(a) \right\} \cup \{0\} \right). \tag{9}$$

Because each row of $\widetilde{\mathbf{U}}$ is strictly decreasing in j, the feasible words at position r form the prefix $\{v_1, \ldots, v_{t_r(a)}\}$. We fix the word at position s_{i_a} to the j_a-th sorted word, and reserve this trigger word before selecting the remaining words. For each $r \neq i_a$, the number of available words after excluding the fixed trigger word and the words selected at preceding non-trigger positions is

$$h_r(a) = t_r(a) - (r-1) + \mathbf{1}[i_a < r] - \mathbf{1}[j_a \leq t_r(a)]. \tag{10}$$

Therefore, the exact layer multiplicity is $m(a) = \prod_{r \neq i_a} h_r(a)$ if $h_r(a) > 0 \; \forall r \neq i_a, 0$ otherwise. According to Lemma 4.2, the total Exponential Mechanism mass of layer a is $\omega(a) = m(a) \exp\left(\epsilon_3 \mathbf{U}_{i_a, j_a}/4\right)$, and the sampling probability of layer a is $p(a) = \omega(a)/\sum_{n=1}^{bT} \omega(n)$.

Assuming a layer $l \sim p(l)$ is sampled, we fix the triggered position s_{i_l} to the j_l-th sorted word, that is $s_{i_l} = v_{j_l}$. For every remaining position $r \neq i_l$ in increasing order, we sample s_r uniformly from $\{v_1, \ldots, v_{t_r(l)}\}$ after removing the trigger word and all previously selected words.

Conditional on layer l, the procedure samples each of its $m(l)$ sequences with probability $1/m(l)$. Combined with the layer probability $p(l)$, this gives each sequence a probability proportional to $\exp\left(\epsilon_3 u(s^*, s)/4\right)$, exactly matching the JEM distribution without enumerating $\mathcal{S}_T$.

Finally, the selected top-T words serve as DP keywords that guide the external LLM to generate a complete input using prompts such as *'Write a complete sentence using the keywords: <DP keywords>'*. For classification tasks, each generated input is paired with its public output label. For open-form tasks, both the input and output are generated solely from the released DP keywords. The external LLM receives only the DP keywords, public task description, and public prompt template.

Lemma 4.2 (Sensitivity of DP-Based Keyword Aggregation)**.** *Conditioned on the released DP centers, unique assignment ensures that removing or adding one demonstration affects the Top-W set of at most one cluster. Thus, the document-frequency query has $\Delta_\infty \mathbf{c} \leq 1$, and the JEM utility has $\Delta u \leq 2$. The proof is given in Appendix E.*

Theorem 4.3 (Overall Privacy Guarantee of DPD)**.** *For adjacent private demonstration pools differing by one record, adaptive composition of the $(\epsilon_1 + \epsilon_2, \delta_1)$-DP center release and the conditional ϵ_3-DP Top-W/JEM keyword release gives $(\epsilon_1 + \epsilon_2 + \epsilon_3, \delta_1)$-DP. The proof is given in Appendix F.*

5 Experiments

In this section, we conduct extensive experiments to answer: **Q1:** How does the privacy budget affect DPD and other DP-based demonstration synthesis methods on downstream ICL tasks, including text

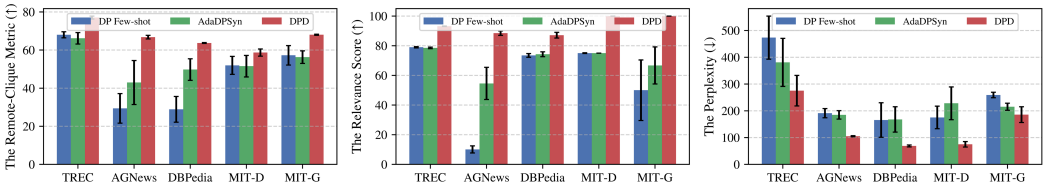
classification, information extraction and mathematical reasoning (See Section 5.1)? **Q2:** To what extent does DPD improve over baselines in the diversity, task relevance and readability of generated demonstrations (See Section 5.2 and Appendix G.8)? **Q3:** Is DPD efficient in time cost (See Section 5.3)? **Q4:** How robust is DPD to membership inference attacks (Section 5.4) and exposure-style reference attack (Section G.4)? The detailed **experimental setting** is provided in Appendix G.1. **Q5:** How do key modules and parameters affect DPD performance (See Appendix G.7)?

5.1 The Performance in ICL Tasks (Q1)

The main ICL results are reported in Table 1, showing the mean accuracy and standard deviation over five random seeds. Experiments are conducted for text classification (AGNews, DBPedia and TREC), information extraction (MIT-G and MIT-D), and mathematical reasoning (GSM8K) under varying privacy budgets (ϵ)[1]. Here, $\epsilon = 0$ denotes a fully private setting where no demonstrations are provided, while $\epsilon = \infty$ represents the comparisons without any DP perturbation, and **Non-private Baseline** refers to few-shot demonstrations randomly sampled from the private dataset.

Table 1: Few-shot ICL performance (%) and generator-controlled comparison. "Generator" denotes the demonstration generator; the downstream evaluator is Llama-2-7B for text classification and information extraction and Qwen2.5-Math-7B-Instruct for GSM8K.

Dataset	Method	Generator	$\epsilon = 0$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$	Non-private Baseline
AGNews	DP few-shot generation	Llama-2-7B	$57.42_{0.11}$	$50.67_{6.66}$	$59.94_{3.65}$	$60.23_{4.02}$	$62.03_{2.61}$	$68.57_{0.52}$	$86.71_{0.71}$
	AdaDPSyn	Llama-2-7B		$53.30_{6.99}$	$57.73_{1.19}$	$61.07_{6.10}$	$63.25_{1.20}$	$68.57_{0.52}$	
	DPD-Llama	Llama-2-7B		$81.93_{3.65}$	$82.10_{4.42}$	$81.20_{5.86}$	$83.37_{0.67}$	$83.88_{0.45}$	
	DPD-DeepSeek	DeepSeek V4		$\mathbf{82.63}_{5.99}$	$\mathbf{83.50}_{2.52}$	$\mathbf{85.07}_{3.52}$	$\mathbf{84.52}_{3.64}$	$\mathbf{85.33}_{1.66}$	
DBPedia	DP few-shot generation	Llama-2-7B	$52.21_{0.13}$	$61.41_{4.86}$	$61.71_{5.29}$	$65.55_{3.87}$	$65.97_{3.65}$	$67.19_{3.21}$	$80.10_{0.42}$
	AdaDPSyn	Llama-2-7B		$61.77_{4.22}$	$62.15_{3.19}$	$66.12_{3.11}$	$66.05_{1.07}$	$67.19_{3.21}$	
	DPD-Llama	Llama-2-7B		$74.10_{0.52}$	$75.53_{1.67}$	$76.83_{1.90}$	$74.67_{1.40}$	$75.94_{1.02}$	
	DPD-DeepSeek	DeepSeek V4		$\mathbf{78.80}_{2.95}$	$\mathbf{80.33}_{2.52}$	$\mathbf{79.83}_{1.92}$	$\mathbf{80.97}_{1.85}$	$\mathbf{81.03}_{1.66}$	
TREC	DP few-shot generation	Llama-2-7B	$70.02_{0.12}$	$63.67_{4.13}$	$67.93_{4.72}$	$70.07_{0.42}$	$71.80_{0.92}$	$70.93_{3.07}$	$77.13_{0.71}$
	AdaDPSyn	Llama-2-7B		$63.47_{3.29}$	$70.93_{4.89}$	$70.01_{1.30}$	$71.47_{1.70}$	$71.93_{3.07}$	
	DPD-Llama	Llama-2-7B		$69.21_{3.47}$	$71.60_{3.34}$	$72.53_{2.60}$	$72.33_{3.14}$	$72.57_{1.54}$	
	DPD-DeepSeek	DeepSeek V4		$\mathbf{70.13}_{1.90}$	$\mathbf{73.20}_{2.12}$	$\mathbf{74.47}_{0.81}$	$\mathbf{75.73}_{0.31}$	$\mathbf{76.88}_{3.07}$	
MIT-G	DP few-shot generation	Llama-2-7B	$32.33_{1.84}$	$51.80_{4.35}$	$52.50_{3.26}$	$56.75_{4.21}$	$55.30_{1.98}$	$61.22_{2.83}$	$76.87_{0.34}$
	AdaDPSyn	Llama-2-7B		$53.87_{6.35}$	$55.07_{3.18}$	$57.93_{3.08}$	$58.67_{2.02}$	$61.22_{2.83}$	
	DPD-Llama	Llama-2-7B		$56.37_{7.45}$	$59.70_{1.09}$	$61.50_{1.82}$	$61.52_{6.50}$	$62.03_{1.08}$	
	DPD-DeepSeek	DeepSeek V4		$\mathbf{58.95}_{2.04}$	$\mathbf{61.51}_{1.48}$	$\mathbf{64.54}_{1.50}$	$\mathbf{64.36}_{4.87}$	$\mathbf{65.09}_{0.59}$	
MIT-D	DP few-shot generation	Llama-2-7B	$50.83_{3.54}$	$80.42_{3.05}$	$83.13_{2.87}$	$83.59_{4.64}$	$84.25_{5.73}$	$86.08_{5.16}$	$92.66_{0.99}$
	AdaDPSyn	Llama-2-7B		$80.98_{4.06}$	$83.42_{2.95}$	$84.33_{2.52}$	$84.89_{5.87}$	$86.08_{5.16}$	
	DPD-Llama	Llama-2-7B		$89.32_{1.45}$	$89.80_{0.37}$	$90.04_{0.14}$	$89.56_{0.70}$	$89.95_{0.22}$	
	DPD-DeepSeek	DeepSeek V4		$\mathbf{92.13}_{1.37}$	$\mathbf{93.25}_{1.10}$	$\mathbf{93.82}_{1.09}$	$\mathbf{93.17}_{0.97}$	$\mathbf{93.90}_{0.70}$	
GSM8K	DP few-shot generation	Llama-2-7B	$44.07_{1.62}$	$20.27_{1.79}$	$25.10_{0.99}$	$24.80_{3.68}$	$26.00_{4.53}$	$31.60_{6.62}$	$88.53_{0.31}$
	AdaDPSyn	Llama-2-7B		$21.01_{1.06}$	$27.10_{3.82}$	$24.30_{2.12}$	$27.40_{0.28}$	$31.60_{6.62}$	
	DPD-Llama	Llama-2-7B		$66.27_{8.28}$	$69.73_{7.85}$	$68.20_{4.00}$	$65.07_{6.20}$	$66.68_{2.88}$	
	DPD-DeepSeek	DeepSeek V4		$\mathbf{82.13}_{5.34}$	$\mathbf{84.67}_{4.24}$	$\mathbf{87.93}_{1.86}$	$\mathbf{86.73}_{3.76}$	$\mathbf{88.40}_{1.11}$	



(a) The metric of diversity (↑) (b) The metric of task-relevance (↑) (c) The metric of perplexity (↓)

Figure 4: Evaluation of generated demonstrations for diversity, task relevance, and perplexity ($\epsilon = 4$).

As shown in Table 1, DPD consistently outperforms two baselines across most privacy settings. At $\epsilon = 1$, DPD achieves 82.63% on AGNews, compared with 50.67% for DP Few-shot Generation and 53.30% for AdaDPSyn. Similar improvements are observed on DBPedia and TREC. DPD also consistently achieves the highest accuracy across all privacy budgets on the open-form MIT-G and MIT-D tasks. On GSM8K, it substantially outperforms both baselines and approaches the non-private performance under moderate privacy budgets, demonstrating a favorable privacy-utility trade-off.

The generator-controlled rows further separate the contribution of DPD from that of the generator. When all methods use Llama-2-7B, DPD-Llama remains stronger than both baselines across most cases. For example, on GSM8K at $\epsilon = 4$, DPD-Llama achieves 68.20%, outperforming the best baseline by 43.40%. Replacing Llama-2-7B with DeepSeek V4 further improves DPD to 87.93%, indicating that the DPD framework provides substantial gains under a matched generator, while a stronger generator offers additional benefits for reasoning-intensive tasks.

[1]The ICL performance of other advanced LLM servers is discussed in Appendix G.7.7

5.2 The Quality of Generated Demonstrations (Q2)

Diversity. Diverse demonstrations can improve LLM generalization in ICL. We evaluate the semantic and syntactic diversity of generated demonstrations using the Remote-Clique (RC) metric (Huang et al., 2024) (see Appendix G.6.1). As shown in Figure 4a, DPD consistently outperforms the two baselines, with particularly clear gains on AGNews and DBPedia. Overall, DPD achieves the highest diversity across datasets, benefiting from its DP coverage-aware clustering strategy, which selects candidates from different clusters to promote semantic and syntactic diversity, as analyzed in Appendix G.2. We further analyze demonstration **readability** in Appendix G.8.

Task relevance. Task-relevant demonstrations provide examples closely aligned with user queries, promoting consistent and accurate LLM responses (Dong et al., 2024). We therefore use a Relevance Score (RS) (see Appendix G.6.2) to quantify the relevance between queries and generated demonstrations. As shown in Figure 4b, DPD achieves the highest RS across the evaluated datasets, with particularly clear gains and lower variance on AGNews and MIT-G. This result measures the relevance of the final synthetic demonstrations. Notably, it does not attribute this gain entirely to the private keyword mechanism, and we provide a direct keyword-signal diagnostic in Appendix G.3.

5.3 Time Cost Analysis (Q3)

Table 2 reports the generation time per released demonstration under $\epsilon = 4$. For DPD, the pipeline includes private pool encoding, DP center release, candidate retrieval, JEM keyword release, and API generation. DPD consistently outperforms both baselines, with substantial gains on long-text datasets such as DBPedia and AGNews. On DBPedia, it reduces generation time from about 390s to 4.14s (98.9% reduction). This improvement stems from avoiding costly token-level perturbation and autoregressive generation, enhancing scalability.

Table 2: Time cost evaluation for generating one DP-based demonstration with $\epsilon = 4$.

Dataset	DP Few-Shot	AdaDPSyn	DPD
TREC	41.50s	41.87s	**2.99s**
DBPedia	388.44s	390.01s	**4.14s**
AGNews	374.59s	388.55s	**3.22s**
MIT-D	268.74s	272.46s	**7.23s**
MIT-G	143.11s	147.06s	**7.10s**
GSM8K	143.11s	147.06s	**4.94s**

5.4 Performance Against Confidence-based Membership Inference Attacks (Q4)

To evaluate the empirical privacy of DPD, we perform confidence-based membership inference attacks (MIA) (Duan et al., 2023) targeted at few-shot ICL. The MIA setup is detailed in Appendix G.5, and attack performance is measured by the average AUC score. Table 3 and Figure 6 report AUC scores under different privacy settings. The non-private baseline ($\epsilon = \infty$), which directly includes a member record in the prompt, yields AUCs above 0.81 across all datasets, indicating severe membership exposure. In contrast, attacks against DPD-generated demonstrations yield AUCs of 0.50-0.58, close to random guessing, substantially weakening the effectiveness of the evaluated MIA. Notably, MIA offers empirical evaluation, while DPD's formal guarantee follows from the DP analysis in Appendix F.

Table 3: Average AUC results for Few-shot ICL by MIA across datasets under different privacy budgets.

Dataset	$\epsilon = \infty$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
TREC	0.9430	0.5230	0.5571	0.5685	0.5811
AGNews	0.9311	0.5217	0.5259	0.5402	0.5640
DBPedia	0.9142	0.5106	0.5041	0.5488	0.5356
MIT-D	0.8358	0.4931	0.5233	0.5621	0.5777
MIT-G	0.8148	0.5262	0.5633	0.5425	0.5457
GSM8K	0.8783	0.5473	0.5186	0.5369	0.5545

6 Conclusion

In this work, we propose DPD, a novel method for generating privacy-preserving demonstrations to protect sensitive data in ICL prompts. Formally, by applying a pretrained task-aligned encoder to the private demonstration pool, task-aligned embeddings are obtained and further bounded through JL projection and L_2 clipping. DPD then performs DP coverage-aware clustering to release cluster centers with broader semantic coverage. Representative demonstration candidates are assigned to their nearest released centers, and the cluster-wise keywords extracted from these candidates are perturbed via the Joint Exponential Mechanism. These privatized keywords are finally used to guide the LLM in generating DP-protected demonstrations that maintain task-relevance and diversity. Empirical results show that DPD effectively generates high-utility and DP-protected demonstrations and improves LLM performance on ICL tasks. The limitations of our work are discussed in Appendix I.5.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic demonstrations as an integral component of the proposed method, and to assist with the design of experiments and the analysis and interpretation of experimental results. We did not use generative AI tools to develop the proposed theoretical framework, formulate mathematical claims, or produce proofs; the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to improve the readability and linguistic clarity of the manuscript. All AI-generated demonstrations, experimental suggestions, and analyses were evaluated using the procedures reported in the paper, and all AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work proposes a privacy-preserving approach for protecting sensitive demonstration pools used in in-context learning. By providing demonstration-level differential privacy, the proposed method aims to reduce the risk of exposing information contained in individual demonstrations. Nevertheless, differential privacy does not eliminate all possible privacy or misuse risks, and the method should be deployed with appropriate safeguards, access controls, and consideration of the relevant legal and ethical requirements.

REPRODUCIBILITY STATEMENT

We provide the algorithmic details, privacy analysis, implementation settings, dataset descriptions, prompts, and additional experimental results in the main paper and appendix. The anonymized source code and materials required to reproduce our experiments are available at `https://anonymous.4open.science/r/DPD-0E00`.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, pp. 308–318, 2016.

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

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

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguistics*, 16(1):22–29, 1990.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *EMNLP*, pp. 1107–1128, 2024.

Cicero Dos Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. In *ICML*, pp. 1818–1826, 2014.

Minxin Du, Xiang Yue, Sherman S. M. Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *CCS*, pp. 2665–2679, 2023.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *NeurIPS*, volume 36, 2023.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pp. 486–503, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pp. 265–284, 2006b.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In *ICLR*, 2024.

Fengyu Gao, Ruida Zhou, Tianhao Wang, Cong Shen, and Jing Yang. Data-adaptive differentially private prompt synthesis for in-context learning. In *ICLR*, 2025.

Jennifer Gillenwater, Matthew Joseph, Andres Muñoz Medina, and Mónica Ribero Diaz. A joint exponential mechanism for differentially private top-k. In *ICML*, pp. 7570–7582, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei Xiao, Jianfeng Gao, Lichao Sun, et al. Datagen: Unified synthetic dataset generation via large language models. In *ICLR*, 2024.

Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. Pleak: Prompt leaking attacks against large language model applications. In *CCS*, pp. 3600–3614, 2024.

William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization. In *ACL*, pp. 1401–1422, 2023.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *EMNLP*, pp. 6342–6353, 2023.

Hui Liu, Wenya Wang, Hao Sun, Chris Xing Tian, Chenqi Kong, Xin Dong, and Haoliang Li. Unraveling the mechanics of learning-based demonstration selection for in-context learning. In *ACL*, pp. 2623–2641, 2025.

Jingjing Liu, Scott Cyphers, Panupong Pasupat, Ian McGraw, and James R. Glass. A conversational movie search system based on conditional random fields. In *INTERSPEECH*, pp. 2454–2457, 2012.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *USENIX Security*, pp. 1831–1847, 2024.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *ACL*, pp. 2304–2317, 2023.

Alfredo Maldonado, Filip Klubicka, and John D. Kelleher. Size matters: The impact of training size in taxonomically-enriched word embeddings. *Open Comput. Sci.*, 9(1):252–267, 2019.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Pawel Krzysztof Nowak, Tal Linzen, Fei Sha, and Kristina Toutanova. Improving compositional generalization with latent structure and data augmentation. In *NAACL*, pp. 4341–4362, 2022.

Xun Ran, Qingqing Ye, Haibo Hu, Xin Huang, Jianliang Xu, and Jie Fu. Differentially private graph neural networks for link prediction. In *ICDE*, pp. 1632–1644, 2024.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pp. 3982–3992, 2019.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *NAACL-HLT*, pp. 2655–2671, 2022.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Xinyu Tang, Richard Shin, Huseyin Inan, Andre Manoel, Niloofar Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *ICML*, pp. 33058–33077, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. In *SIGIR*, pp. 200–207, 2000.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process. In *EMNLP*, pp. 1266–1280, 2024.

Xubin Wang, Jianfei Wu, Yuan Yichen, Deyu Cai, Mingzhe Li, and Weijia Jia. Demonstration selection for in-context learning via reinforcement learning. In *ICML*, pp. 64937–64954, 2025.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *ICLR*, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Xiaoyu Xu, Minxin Du, Qingqing Ye, and Haibo Hu. OBLIVIATE: robust and practical machine unlearning for large language models. In *EMNLP*, pp. 3696–3715, 2025.

Xiaoyu Xu, Minxin Du, Zitong LI, Zi Liang, Zhibiao Guo, Shiyu Zhang, Peizhao Hu, Qingqing Ye, and Haibo Hu. From domains to instances: Dual-granularity data synthesis for LLM unlearning. In *ACL Findings*, pp. 28540–28556, 2026b.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Haocheng Yang, Xiang Cheng, Chenhao Sun, Pengfei Zhang, and Sen Su. Privsv: Differentially private steering vector for large language models. In *AAAI*, pp. 34241–34249, 2026.

Yong Yang, Changjiang Li, Qingming Li, Oubo Ma, Haoyu Wang, Zonghui Wang, Yandong Gao, Wenzhi Chen, and Shouling Ji. Prsa: prompt stealing attacks against real-world prompt services. In *USENIX Security*, 2025b.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *ICML*, pp. 39818–39833, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, pp. 649–657, 2015.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*, 2023.

Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In *EMNLP*, pp. 9134–9148, 2022.

Yifan Zhu, Huiqiang Rong, and Haoran Luo. Token-guard: Towards token-level hallucination control via self-checking decoding. In *ICLR*, 2026.

A Notation Introduction

The frequently used notations are summarized in Table 4.

Table 4: Notation Summary for the DPD Mechanism

Symbol	Description
$\mathcal{D}_{\text{priv}}, \mathcal{D}_{\text{pub}}, \mathcal{D}_{\text{syn}}$	Private raw, public, and synthesized demonstration sets
(q_i, y_i)	A raw input-label demonstration
f_θ	Task-aligned encoder trained without private-data access
R, d_{JL}	Fixed seeded JL projection and its output dimension
$\mathbf{z}_i, \bar{\mathbf{z}}_i$	Encoder output, projected and L_2-clipped embedding
C, r_{cov}	Fixed clipping radius and coverage radius, with $r_{\text{cov}} = 2C$
K, W, T	Number of released centers, candidate demonstrations, and released keywords
$\mathcal{F}_{t-1}$	DP centers released before round t
$w_i^{(t)}$	Coverage-aware weight of record i in round t
$n_t, \mathbf{s}_t$	Weighted count of records and weighted vector sum in round t
$\widetilde{n}_t, \widetilde{\mathbf{s}}_t$	Laplace-perturbed weighted count and Gaussian-perturbed weighted vector sum
$\widetilde{\mathbf{c}}_t$	The t-th released DP center
$n_{\min}$	Fixed positive lower bound for the noisy denominator
$\mathcal{S}_k$	Internal Top-W records uniquely assigned to center k
$\mathcal{V}, b$	Fixed JEM vocabulary and its size
$\mathbf{c} = (c_1, \ldots, c_b)$	Document-frequency vector over $\mathcal{V}$
s, s^*	Released and optimal ordered keyword sequences
$u_{\mathcal{D}}(s)$	Joint top-T keyword utility on dataset $\mathcal{D}$
$\epsilon_1, \epsilon_2, \epsilon_3$	The privacy budgets for vector-sum, weighted-count, and JEM
δ	Differential privacy failure probability

B Differential Privacy and Relevant Perturbation Mechanism

Differential Privacy (Dwork et al., 2006b). Differential Privacy (DP) is a rigorous privacy framework that guarantees the output of a randomized algorithm remains statistically indistinguishable when any single record in the input dataset is modified. Informally, a mechanism $\mathcal{A}$ satisfies (ϵ, δ)-DP if for any two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in one data record, and for any output set $\mathcal{S}$,

$$\mathbb{P}[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}\left(\mathcal{D}'\right) \in \mathcal{S}\right] + \delta, \tag{11}$$

where ϵ controls the privacy level (smaller means stronger privacy). δ allows a small probability of privacy failure. When $\delta = 0$, it is pure ϵ-DP. Two fundamental properties of DP are particularly useful for analyzing complex algorithms: (1) **Post-Processing**. If a mechanism $\mathcal{A}(\mathcal{D})$ satisfies (ϵ, δ)-DP, then for any function f that does not directly access the private dataset, the composition $f(\mathcal{A}(\mathcal{D}))$ also satisfies (ϵ, δ)-DP; (2) **Composition**: If $\mathcal{A}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D})$ satisfy (ϵ_1, δ_1)-DP and (ϵ_2, δ_2)-DP, then the combined mechanism $\mathcal{F}(\mathcal{D}) = (\mathcal{A}(\mathcal{D}), \mathcal{M}(\mathcal{D}))$ satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-DP.

Laplace Mechanism (Dwork & Roth, 2014). For numerical query outputs, the Laplace mechanism directly perturbs scalar- or vector-valued statistics with additive Laplace noise. Let Δ_f denote the L_1-*sensitivity* of a function f, defined as the maximum change in the function output over any two adjacent datasets: $\Delta_f = \max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_1$. The Laplace mechanism $\mathcal{A}$ perturbs the function output by adding calibrated Laplace noise:

$$\mathcal{A}(\mathcal{D}) = f(\mathcal{D}) + \text{Lap}(\frac{\Delta_f}{\epsilon}), \tag{12}$$

which satisfies ϵ-DP. The Laplace noise is scaled to Δ_f/ϵ, ensuring that higher sensitivity and smaller privacy budget require more noise.

Gaussian Mechanism (Dwork & Roth, 2014). For numerical vector-valued queries, the Gaussian mechanism instead applies Gaussian perturbation and is commonly used to achieve approximate (ϵ, δ)-DP. For a vector query g with global L_2-sensitivity Δ_g, a standard calibration for $0 < \epsilon \leq 1$ is

$$g(\mathcal{D}) + \mathcal{N}(\mathbf{0}, \sigma^2\Delta_g^2 I), \qquad \sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}}{\epsilon}, \tag{13}$$

which satisfies (ϵ, δ)-DP. Any tighter valid Gaussian calibration may be substituted without changing the subsequent composition argument.

Algorithm 1: DP Coverage-Aware Center Release

Input: Clipped embeddings $\{\bar{\mathbf{z}}_i\}_{i=1}^N$; number of centers K; fixed clipping radius C and coverage radius r_{cov}; fixed $n_{\min} > 0$; budgets $\epsilon_1, \epsilon_2, \delta_1$
Output: Released DP centers $\mathcal{F}_K$

$\mathcal{F}_0 \leftarrow \emptyset$
for $t = 1, \ldots, K$ **do**
  $\epsilon_{s,t} \leftarrow \epsilon_1/K$, $\epsilon_{n,t} \leftarrow \epsilon_2/K$, $\delta_t \leftarrow \delta_1/K$
  foreach $i \in [N]$ **do**
    $w_i^{(t)} \leftarrow 1$ if $t = 1$; otherwise $w_i^{(t)} \leftarrow \min\{1, d(\bar{\mathbf{z}}_i, \mathcal{F}_{t-1})/r_{\text{cov}}\}$
  $n_t \leftarrow \sum_i w_i^{(t)}$, $\mathbf{s}_t \leftarrow \sum_i w_i^{(t)} \bar{\mathbf{z}}_i$
  $\widetilde{n}_t \leftarrow n_t + \text{Lap}(1/\epsilon_{n,t})$
  Choose σ_t by a valid $(\epsilon_{s,t}, \delta_t)$ Gaussian calibration and release $\widetilde{\mathbf{s}}_t \leftarrow \mathbf{s}_t + \mathcal{N}(\mathbf{0}, \sigma_t^2 C^2 I)$
  $\widetilde{\mathbf{c}}_t \leftarrow \text{Clip}_C(\widetilde{\mathbf{s}}_t / \max\{\widetilde{n}_t, n_{\min}\})$
  $\mathcal{F}_t \leftarrow \mathcal{F}_{t-1} \cup \{\widetilde{\mathbf{c}}_t\}$
return $\mathcal{F}_K$

Algorithm 2: JEM for Document-Frequency Keyword Release

Input: Internal records $\mathcal{S}_k$; fixed vocabulary $\mathcal{V} = \{v_1, \ldots, v_b\}$; keyword number T; budget ϵ_3
Output: An ordered sequence $s = (s_1, \ldots, s_T)$ of distinct keyword indices

for $j = 1, \ldots, b$ **do**
  $c_j \leftarrow \sum_{x \in \mathcal{S}_k} \mathbf{1}[v_j \text{ occurs in } x]$
Stably sort $\mathbf{c}$ and let $c_{(i)}$ denote its i-th largest entry
For every ordered distinct sequence s, define $u_{\mathcal{D}}(s) \leftarrow -\max_{i \in [T]}(c_{(i)} - c_{s_i})$
Sample s from the fixed output domain with $\Pr[s] \propto \exp(\epsilon_3 u_{\mathcal{D}}(s)/4)$
return s

Exponential Mechanism (Dwork & Roth, 2014). Unlike the Laplace and Gaussian mechanisms, which perturb numerical query outputs, the exponential mechanism is designed for discrete or non-numeric outputs and randomizes the selection among candidates according to their utility scores. It employs a score function $q(\mathcal{D}, r)$ that measures the utility of selecting output r on dataset $\mathcal{D}$. Let Δq denote the sensitivity of the score function. The mechanism selects an output $r \in \mathcal{R}$ with probability

$$\mathbb{P}[\mathcal{A}(\mathcal{D}) = r] = \frac{\exp\left(\frac{\epsilon\, q(\mathcal{D}, r)}{2\Delta q}\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{\epsilon\, q(\mathcal{D}, r')}{2\Delta q}\right)}. \tag{14}$$

This mechanism ensures ϵ-DP while favoring outputs with higher utility scores.

Joint Exponential Mechanism (JEM) (Gillenwater et al., 2022). To select a set of outputs jointly, the Joint Exponential Mechanism extends the exponential mechanism by defining a utility function $q(\mathcal{D}, \mathbf{s})$ over a tuple or set of outputs $\mathbf{s} \in \mathcal{R}^k$. The mechanism samples $\mathbf{s}$ with probability

$$\mathbb{P}[\mathcal{A}(\mathcal{D}) = \mathbf{s}] = \frac{\exp\left(\frac{\epsilon\, q(\mathcal{D}, \mathbf{s})}{2\Delta q}\right)}{\sum_{\mathbf{s}' \in \mathcal{R}^k} \exp\left(\frac{\epsilon\, q(\mathcal{D}, \mathbf{s}')}{2\Delta q}\right)}. \tag{15}$$

By jointly modeling dependencies among selected outputs, JEM improves overall utility compared to independent sampling, while still satisfying ϵ-DP. However, this comes at the cost of higher computational complexity due to sampling over the combinatorial space $\mathcal{R}^k$, whereas EM operates over $\mathcal{R}$ independently.

C Sensitivity Analysis of DP Coverage-Aware Clustering

Lemma (The sensitivity of DP Coverage-Aware Clustering strategy). Conditioned on the preceding DP outputs, the weighted count and vector-sum queries satisfy $\Delta_{n_t} \leq 1$ and $\Delta_{\mathbf{s}_t} \leq C$ under adding or removing one demonstration from the private pool.

Algorithm 3: Diverse Privacy-Preserving Demonstration Synthesis

Input: Private demonstrations $\mathcal{D}_{\text{priv}}$; public data $\mathcal{D}_{\text{pub}}$; public label set $\mathcal{Y}$ for classification; fixed JL dimension and seed; fixed $C, r_{\text{cov}}, K, W, T, n_{\min}$; fixed vocabulary $\mathcal{V}$; budgets $\epsilon_1, \epsilon_2, \epsilon_3, \delta_1$

Output: DP-protected demonstrations $\mathcal{D}_{\text{syn}}$

Train f_θ on $\mathcal{D}_{\text{pub}}$ and fix its parameters
Generate the fixed JL matrix R
foreach $(q_i, y_i) \in \mathcal{D}_{\text{priv}}$ **do**
    $\mathbf{z}_i \leftarrow f_\theta(q_i)$, JL projection and clipping: $\bar{\mathbf{z}}_i \leftarrow \text{Clip}_C(R\mathbf{z}_i)$
$\mathcal{D}_{\text{syn}} \leftarrow \emptyset$
if *the task is text classification* **then**
    foreach $y \in \mathcal{Y}$ **do**
        $\mathcal{I}_y \leftarrow \{i : y_i = y\}$
        $\mathcal{F}_y \leftarrow$ DP-COVERAGE-CENTERS$(\{\bar{\mathbf{z}}_i : i \in \mathcal{I}_y\}, K, C, r_{\text{cov}}, n_{\min}, \epsilon_1, \epsilon_2, \delta_1)$
        Assign records in $\mathcal{I}_y$ to their nearest released centers
        foreach $k \in [K]$ **do**
            Let $\mathcal{S}_{y,k}$ be the nearest Top-W demonstrations assigned to center k
            Release DP keywords: $s_{y,k} \leftarrow \text{JEM}(\mathcal{S}_{y,k}, \mathcal{V}, T, \epsilon_3)$
            Generate $\mathcal{D}_{\text{syn}}^{(y,k)}$ using $s_{y,k}$, public task information, label y, and the public prompt template
            $\mathcal{D}_{\text{syn}} \leftarrow \mathcal{D}_{\text{syn}} \cup \mathcal{D}_{\text{syn}}^{(y,k)}$
else
    $\mathcal{F}_K \leftarrow$ DP-COVERAGE-CENTERS$(\{\bar{\mathbf{z}}_i\}, K, C, r_{\text{cov}}, n_{\min}, \epsilon_1, \epsilon_2, \delta_1)$
    Assign all records to their nearest released centers
    foreach $k \in [K]$ **do**
        Let $\mathcal{S}_k$ be the nearest Top-W demonstrations assigned to center k
        Release DP keywords: $s_k \leftarrow \text{JEM}(\mathcal{S}_k, \mathcal{V}, T, \epsilon_3)$
        Jointly generate the synthetic input and output in $\mathcal{D}_{\text{syn}}^{(k)}$ using only released DP keywords s_k and a fixed public task/output schema
        $\mathcal{D}_{\text{syn}} \leftarrow \mathcal{D}_{\text{syn}} \cup \mathcal{D}_{\text{syn}}^{(k)}$
return $\mathcal{D}_{\text{syn}}$

Proof. Fix a round t and condition on an arbitrary transcript of the previously released centers $\mathcal{F}_{t-1}$. Under this conditioning, $\mathcal{F}_{t-1}$ is fixed, so the map

$$\bar{\mathbf{z}} \mapsto w^{(t)}(\bar{\mathbf{z}}) = \begin{cases} 1, & t = 1, \\ \min\{1, d(\bar{\mathbf{z}}, \mathcal{F}_{t-1})/r_{\text{cov}}\}, & t > 1 \end{cases} \tag{16}$$

is a fixed record-wise function with values in $[0, 1]$.

Consider adjacent datasets in the addition case, $\mathcal{D}' = \mathcal{D} \cup \{x\}$, and let $\bar{\mathbf{z}}_x$ be the clipped embedding of the added record. All common records have identical embeddings and identical weights under the fixed transcript. Hence,

$$n_t(\mathcal{D}') - n_t(\mathcal{D}) = w_x^{(t)}, \tag{17}$$

$$\mathbf{s}_t(\mathcal{D}') - \mathbf{s}_t(\mathcal{D}) = w_x^{(t)} \bar{\mathbf{z}}_x. \tag{18}$$

Because $0 \le w_x^{(t)} \le 1$ and $\|\bar{\mathbf{z}}_x\|_2 \le C$,

$$|n_t(\mathcal{D}') - n_t(\mathcal{D})| \le 1, \qquad \|\mathbf{s}_t(\mathcal{D}') - \mathbf{s}_t(\mathcal{D})\|_2 \le C. \tag{19}$$

The removal case follows by symmetry. Therefore, conditional on every possible preceding transcript, $\Delta_{n_t} \le 1$ and $\Delta_{\mathbf{s}_t} \le C$. This completes the proof. □

D PRIVACY ANALYSIS OF ALGORITHM 1

Lemma D.1. *Algorithm 1 satisfies $(\epsilon_1 + \epsilon_2, \delta_1)$-DP with respect to add/remove one raw private demonstrations.*

Proof. Fix round t and condition on the transcript $\mathcal{F}_{t-1}$. Lemma 4.1 yields $\Delta_{n_t} \le 1$ and $\Delta_{\mathbf{s}_t} \le C$. Therefore, by the Laplace Mechanism, releasing

$$\widetilde{n}_t = n_t + \mathrm{Lap}\left(\frac{\Delta_{n_t}}{\epsilon_{n,t}}\right) = n_t + \mathrm{Lap}\left(\frac{1}{\epsilon_{n,t}}\right) \tag{20}$$

satisfies $\epsilon_{n,t}$-DP. By setting $\sigma_t \ge \sqrt{2\ln(1.25/\delta_t)}/\epsilon_{s,t}$ based on Gaussian Mechanism, releasing

$$\widetilde{\mathbf{s}}_t = \mathbf{s}_t + \mathcal{N}(\mathbf{0}, \sigma_t^2 \Delta_{\mathbf{s}_t}^2 I) = \mathbf{s}_t + \mathcal{N}(\mathbf{0}, \sigma_t^2 C^2 I) \tag{21}$$

satisfies $(\epsilon_{s,t}, \delta_t)$-DP. Finally, by sequential composition within the round, the pair $(\widetilde{n}_t, \widetilde{\mathbf{s}}_t)$ satisfies $(\epsilon_{n,t} + \epsilon_{s,t}, \delta_t)$-DP.

The center

$$\widetilde{\mathbf{c}}_t = \mathrm{Clip}_C\left(\frac{\widetilde{\mathbf{s}}_t}{\max\{\widetilde{n}_t, n_{\min}\}}\right) \tag{22}$$

is a deterministic function of the released noisy statistics and fixed public hyperparameters. It is therefore post-processing and incurs no additional privacy cost.

The preceding argument holds for every fixed transcript $\mathcal{F}_{t-1}$. Consequently, adaptive sequential composition applies across the fixed K rounds. Using

$$\epsilon_{s,t} = \epsilon_1/K, \quad \epsilon_{n,t} = \epsilon_2/K, \quad \delta_t = \delta_1/K, \tag{23}$$

we obtain

$$\sum_{t=1}^{K}(\epsilon_{s,t} + \epsilon_{n,t}) = \epsilon_1 + \epsilon_2, \quad \sum_{t=1}^{K}\delta_t = \delta_1. \tag{24}$$

Thus, the complete transcript of Algorithm 1, including all released centers, satisfies $(\epsilon_1 + \epsilon_2, \delta_1)$-DP. This completes the proof. □

E Sensitivity Analysis of DP-Based Keyword Aggregation

Lemma E.1 (Sensitivity of DP-Based Keyword Aggregation)**.** *Conditioned on the released DP centers, the document-frequency query satisfies $\Delta_\infty \mathbf{c} \le 1$, and the JEM utility satisfies $\Delta u \le 2$. Consequently, the JEM with probability proportional to $\exp(\epsilon_3 u/(2\Delta u))$ satisfies ϵ_3-DP.*

Proof. Fix any possible transcript of the released DP centers and the deterministic tie-breaking rule. For adjacent datasets $\mathcal{D}' = \mathcal{D} \cup \{x\}$, unique assignment places x in exactly one cluster k_x, while the assignments and distance order of all common records remain unchanged. Hence, only the Top-W set of cluster k_x can change, either by adding x or by replacing one boundary record.

Let $\mathbf{a}(x) \in \{0,1\}^b$ denote the fixed-vocabulary indicator of demonstration x. If x is added to the Top-W set, then $\mathbf{c}' - \mathbf{c} = \mathbf{a}(x)$. If x replaces a boundary record y, then $\mathbf{c}' - \mathbf{c} = \mathbf{a}(x) - \mathbf{a}(y)$. In either case, $\|\mathbf{c}' - \mathbf{c}\|_\infty \le 1$; removal is symmetric. Therefore, the document-frequency query has $\Delta_\infty \mathbf{c} \le 1$. For any fixed candidate sequence $s = (s_1, \ldots, s_T)$, the JEM utility is

$$u_{\mathcal{D}}(s) = -\max_{i \in [T]}\big(c_{s_i^*} - c_{s_i}\big), \tag{25}$$

where $c_{s_i^*}$ is the i-th largest document frequency and c_{s_i} is the frequency of the candidate word s_i. Since $\|\mathbf{c}' - \mathbf{c}\|_\infty \le 1$, every order statistic $c_{s_i^*}$ and every fixed-word count c_{s_i} changes by at most one. Thus, each count gap changes by at most two, and the maximum over i preserves this bound, giving $\Delta u \le 2$. The Exponential Mechanism with probability proportional to

$$\exp\left(\frac{\epsilon_3 u_{\mathcal{D}}(s)}{2\Delta u}\right) = \exp\left(\frac{\epsilon_3 u_{\mathcal{D}}(s)}{4}\right), \tag{26}$$

therefore satisfies ϵ_3-DP for the affected cluster. Since the clusters are disjoint, parallel composition makes the joint release across all K clusters ϵ_3-DP. This completes the proof. □

F Overall Privacy Analysis of DPD

Theorem (Overall Privacy Guarantee of DPD). With respect to add/remove-one adjacency on the raw private demonstration pool, DPD satisfies

$$(\epsilon_1 + \epsilon_2 + \epsilon_3, \delta_1)\text{-DP}. \tag{27}$$

Proof. The task-aligned encoder is trained only on public data. The JL projection matrix, its seed and dimension, the clipping radius, coverage radius, number of centers, JEM vocabulary, public label ontology, and prompt template are all fixed before private-data access and incur no privacy cost.

Lemma D.1 establishes that the complete adaptive center release satisfies $(\epsilon_1 + \epsilon_2, \delta_1)$-DP. Conditioned on any released center transcript, Lemma E.1 establishes that the cluster-wise Top-W/JEM keyword release satisfies ϵ_3-DP. Although Top-W accesses the private pool again, adaptive sequential composition gives $(\epsilon_1 + \epsilon_2 + \epsilon_3, \delta_1)$-DP.

For classification tasks, class labels are public task information and incur no privacy cost. For open-form tasks, the external LLM jointly generates both the synthetic input and its output from only the released DP keywords and a fixed public output schema; no raw or record-derived output is supplied. For MIT-D, the generated director name must also appear explicitly in the generated description; for MIT-G, the generated genre must be supported by the description. The external LLM never accesses raw demonstrations, private labels, Top-W records, embeddings, assignments, or exact counts. Thus, LLM synthesis and subsequent reuse are post-processing of DP outputs and incur no additional privacy cost. □

G Experimental Supplementary

G.1 Experimental Settings

Benchmark datasets. In this experiment, we study three benchmark datasets for text classification, including *TREC, DBPedia,* and *AGNews*. In addition, we conduct information extraction tasks on the *MIT-G* and *MIT-D* datasets and mathematical reasoning experiments on *GSM8K*. Their corresponding prompt formats for ICL are summarized in Table 24, and the dataset splitting follows the protocol in (Gao et al., 2025), as detailed as follows:

• **TREC**: This question classification dataset has six types: Number, Location, Person, Description, Entity, and Abbreviation (Voorhees & Tice, 2000). It consists of 5,452 training and 500 test samples.

• **DBPedia**: The DBPedia classification dataset covers 14 semantic classes, such as Company, Artist, Athlete, and Film (Zhang et al., 2015). For each class, the dataset provides 40K training samples and 5K test samples. We randomly select 49,999 training samples and 1,000 test samples for evaluation.

• **AGNews**: Such a dataset (Zhang et al., 2015) is a news classification task with four categories: World, Sports, Business, and Sci/Tech. Each category contains 30K training and 1.9K test instances. In our setting, 1,000 samples are randomly drawn from the test set.

• **MIT Movies**: This dataset (Liu et al., 2012) is designed for movie-related information extraction, including genre (**MIT-G**) and director (**MIT-D**) prediction tasks. MIT-G contains 2,953 training and 780 test samples, while MIT-D includes 1,561 training and 415 test samples.

• **GSM8K**: This dataset (Cobbe et al., 2021) is a grade-school mathematical reasoning dataset containing multi-step arithmetic word problems. We randomly evaluate 500 test samples and count a prediction as correct only when its extracted final numerical answer exactly matches the gold answer.

Baselines and LLM server. We compare the proposed DPD method against state-of-the-art baselines, namely **DP few-shot generation** (Tang et al., 2024) and **AdaDPSyn** (Gao et al., 2025). DP few-shot generation queries the LLM with randomly sampled demonstrations and aggregates the resulting next-token probability distributions in a differentially private manner to synthesize demonstrations. Based on this framework, AdaDPSyn employs a data-adaptive strategy to reduce noise injection and enhance the quality of synthetic demonstrations. They use Llama-2-7B (Touvron et al., 2023) as their demonstration generator. For the fair comparison, we adopt the optimal hyperparameters for all baselines, following the settings recommended in (Gao et al., 2025).

By default, DPD uses DeepSeek V4 (Xu et al., 2026a) to expand the released DP keywords into task-specific synthetic demonstrations. These demonstrations are subsequently used for downstream ICL. We employ Llama-2-7B (Touvron et al., 2023), Llama-3-8B (Grattafiori et al., 2024), Qwen-3.5-flash (Yang et al., 2025a), and GPT-5.5 (Singh et al., 2025) as ICL servers in the corresponding experiments. For GSM8K, we use Qwen2.5-Math-7B-Instruct (Yang et al., 2024) as the reasoning model for DPD and two baselines.

Implementation and Parameters[2]. We use the synthetic demonstrations $\mathcal{D}_{\text{syn}}$ generated by Algorithm 3 as input demonstrations for downstream ICL tasks. As shown in Table 5, for classification tasks, n_{shot} denotes the number of demonstrations per label. We set $n_{\text{shot}} = 4$ for TREC and AGNews and $n_{\text{shot}} = 2$ for DBPedia[3]. For open-form tasks, n_{shot} denotes the total number of demonstrations, and we set $n_{\text{shot}} = 4$ for MIT-G, MIT-D, and GSM8K. We set $K = n_{\text{shot}}$: K DP centers are released per public label for classification tasks, whereas K centers are released over the entire private demonstration pool for open-form tasks.

For text classification tasks, the task-aligned encoder is initialized from Sentence-BERT (Reimers & Gurevych, 2019) and trained exclusively on public data, with $|\mathcal{D}_{\text{pub}}| = 1{,}090$, $24{,}000$, and $9{,}999$ for TREC, AGNews, and DBPedia, respectively. We fine-tune the encoder for 10 epochs using AdamW, a batch size of 128, a learning rate of 5×10^{-5}, and a maximum sequence length of 128. MIT-G, MIT-D, and GSM8K require no supervised encoder training and directly use the pretrained encoder.

We set the keyword size $T = 8$ and the number of selected samples per released center $W = 50$ for all datasets. We set the embedding clipping radius $C = 1$, coverage radius $r_{\text{cov}} = 2$, and use a fixed 32-dimensional Johnson–Lindenstrauss projection ($d_{\text{JL}} = 32$) sampled independently of private data. For text classification, coverage-aware private center release is performed separately within each public label. For MIT-D and MIT-G, it is performed once over all private demonstrations because their outputs are open-ended. The JEM output domain is fixed before accessing private data. Specifically, we take the public Sentence-BERT (Reimers & Gurevych, 2019) tokenizer vocabulary, lowercase its entries, retain only tokens matching `[a-z0-9]+`, remove duplicates, and sort them deterministically. This produces a fixed vocabulary of 22,716 tokens shared by all datasets and seeds.

We use the DeepSeek V4 API to generate $\mathcal{D}_{\text{syn}}$. The generator receives only the released DP keywords, fixed public task instructions, and the public label or output schema; it never receives raw private demonstrations. The complete prompts and decoding parameters are reported in Table 23. Unless otherwise specified, we use temperature 0.3 and the API-default top-p. All main and ablation results use five matched seeds, with run configurations and output hashes recorded to ensure result traceability. Experiments are conducted on a server with eight NVIDIA A40 GPUs with 48GB memory each. Demonstration-generation time is reported in Table 2.

Privacy setting. For privacy parameters, the allocated privacy budgets for different privacy-query modules are calculated as $(\epsilon_1, \epsilon_2, \epsilon_3) = \left(\frac{1}{4}\epsilon_{\text{total}}, \frac{1}{4}\epsilon_{\text{total}}, \frac{1}{2}\epsilon_{\text{total}}\right)$. The rationale behind this allocation is that the budgets ϵ_1 and ϵ_2 are used to release the noisy vector sums and noisy coverage counts required for private center selection, respectively, while ϵ_3 is used to privately capture keywords from the selected candidates. Empirical results suggest that allocating these budgets in a 1:1:2 ratio provides a favorable balance between center quality and keyword relevance. Accordingly, for $\epsilon_{\text{total}} \in \{1, 2, 4, 8\}$, we set $(\epsilon_1, \epsilon_2, \epsilon_3)$ to $(0.25, 0.25, 0.5)$, $(0.5, 0.5, 1)$, $(1, 1, 2)$, and $(2, 2, 4)$, respectively. Moreover, we set the failure probability $\delta_1 = 10^{-5}$, which is negligible relative to the dataset size (i.e., $\delta_1 \leq 1/N$), for all benchmark datasets (Ran et al., 2024).

G.2 Diagnostic of Coverage-Aware Center Release

DP coverage-aware clustering is designed to promote semantic coverage, but does not guarantee recovery of distinct density modes. We therefore compare it with two baselines under controlled settings: (i) **Repeated Noisy Mean** repeatedly releases a noisy global mean using the same privacy budget and number of centers, but without coverage-aware weighting. (ii) **Public Prototypes** computes fixed centers solely from public embeddings, without accessing private data.

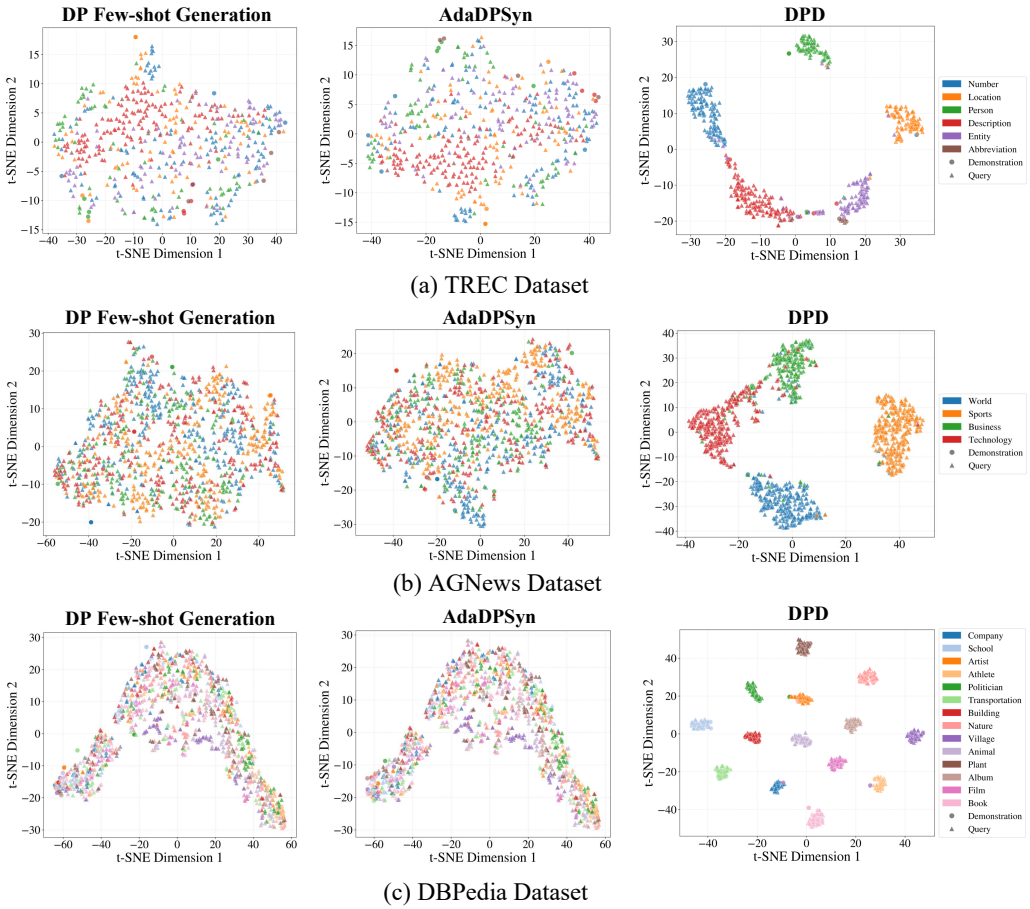
[2]All hyperparameters, including W, T, K, d_{JL}, and the privacy-budget allocation ratio, are selected on a held-out validation set independent of $\mathcal{D}_{\text{priv}}$, incurring no additional privacy cost.

[3]Llama-2-7B has a maximum context length of approximately 4,096 tokens, while DBPedia mainly contains relatively long inputs.

Table 5: Key parameter settings of DPD across different datasets.

Parameters	TREC	DBPedia	AGNews	MIT-D	MIT-G	GSM8K
Shot number ($n_{\text{shot}} = K$)*	4	2	4	4	4	4
Samples per center (W)	50	50	50	50	50	50
Keyword size (T)	8	8	8	8	8	8

*For classification tasks, n_{shot} denotes the number of demonstrations per label; for non-classification tasks, it denotes the total number of demonstrations.



(a) TREC Dataset

(b) AGNews Dataset

(c) DBPedia Dataset

Figure 5: t-SNE Visualization of query and demonstration embeddings across different generation methods.

In Table 6, **nearest distance** measures the distance from samples to their closest released center, **center distance** and **candidate distance** measure pairwise spread among released centers and candidates, respectively, and **empty-cluster rate** measures the fraction of centers without assigned samples. Lower nearest distance and empty-cluster rate indicate better coverage and balance, while a larger spread indicates greater diversity. Our DP coverage-aware release produces a larger center and candidate spread than other two baselines, supporting its semantic coverage objective.

G.3 Diagnostic of the Private Keyword Signal

We first verify the JEM keyword sampler against exact enumeration on small-scale test cases. With 100,000 samples per case, the total-variation distance between the sampled distribution and the exact JEM distribution is 0.01-0.03, validating the sampler. The released run manifest records the fixed vocabulary source and size (SHA-256 hash). Document frequency $c_j \leq W$ is binary per record.

Table 6: Center diagnostics on TREC. Nearest distance and empty-cluster rate are lower-is-better; center and candidate pairwise distances measure geometric spread.

Method	Nearest dist.↓	Center dist.↑	Empty rate↓	Candidate dist.↑
DP coverage-aware (Ours)	0.366	0.948	0.167	0.862
Repeated noisy mean	0.449	0.381	0.186	0.334
Public prototypes	0.424	0.817	0.193	0.709

Table 7: Keyword-signal diagnostic comparing JEM with uniform sampling from the same fixed vocabulary. The $W = 50, |\mathcal{V}| = 22{,}716$ rows use the main configuration.

ϵ	W	$\lvert\mathcal{V}\rvert$	Release	Candidate hit	Mean DF	Top-T recall
1	50	22,716	JEM	9.69	0.68	2.19
1	50	22,716	Uniform	8.44	0.20	1.56
4	50	22,716	JEM	11.73	1.12	5.01
4	50	22,716	Uniform	7.19	0.18	1.35
8	50	22,716	JEM	15.73	4.12	9.90
8	50	22,716	Uniform	7.19	0.18	1.35

To assess whether JEM captures candidate-specific information, we evaluate the released keywords over five TREC seeds. **Candidate hit** denotes the fraction of released keywords with nonzero document frequency in the selected Top-W records, **mean DF** is their average document frequency, and **Top-T recall** measures the overlap with the non-private document-frequency Top-T keywords. In Table 7, the main $W = 50, |\mathcal{V}| = 22{,}716$ configuration yields a 9.69-15.73% candidate-hit rate and approximately 9.90% Top-T recall. Thus, the released keywords provide meaningful semantic guidance relative to uniform sampling.

G.4 Exposure-Style Reference Attack

Attack setup. We evaluate whether synthetic demonstrations expose distinctive content from private records, such as names, identifiers, numbers, or rare phrases. Unlike confidence-based MIA, this attack requires only the released demonstrations and target records, without model log-probabilities. We lowercase each text and tokenize it into alphanumeric sequences. For a target record x and a released demonstration d, we compute

$$J(x, d) = \frac{|T(x) \cap T(d)|}{|T(x) \cup T(d)|}, \qquad S(x) = \max_{d \in \mathcal{D}_{\text{syn}}} J(x, d).$$

A threshold τ is calibrated using held-out non-member scores, and the attack predicts membership when $S(x) \geq \tau$. We use 500 members and 500 non-members whenever available and report five-seed results. For the metrics, **AUC** measures how well member scores are separated from non-member scores; values near 0.5 indicate random guessing. **TPR@5%FPR** is the fraction of detected members when 5% of non-members are incorrectly classified as members; lower is better.

Table 8: Exposure-style reference attack at $\epsilon = 4$.

Dataset	**AUC**	**TPR@5% FPR**
TREC	0.549 ± 0.068	0.045 ± 0.025
AGNews	0.460 ± 0.038	0.036 ± 0.015
DBPedia	0.497 ± 0.025	0.039 ± 0.013
MIT-G	0.517 ± 0.014	0.060 ± 0.017
MIT-D	0.517 ± 0.018	0.044 ± 0.011
GSM8K	0.494 ± 0.010	0.045 ± 0.013
Macro avg.	**0.506**	**0.045**

Table 8 shows that DPD yields a macro-average AUC of 0.506, which is close to random guessing, together with a low TPR@5% FPR of 0.045. These results indicate that the attacker has limited ability to distinguish or recover record-specific exposure from the released DPD outputs. In contrast, directly releasing the private records leads to perfect attack performance, with both AUC and TPR@5%FPR reaching 1.0. The large gap between DPD and raw release therefore provides empirical evidence that DPD substantially suppresses direct record-specific exposure beyond the confidence-based MIA.

G.5 Confidence-based Membership Inference Attack Settings

To evaluate the empirical privacy of DPD, we perform confidence-based membership inference attacks (MIA) (Duan et al., 2023) targeted at few-shot ICL. Such an attack aims to determine whether a given record was used in the private demonstration pool. Specifically, we conduct experiments by splitting each dataset into member and non-member subsets. Member samples are used to generate

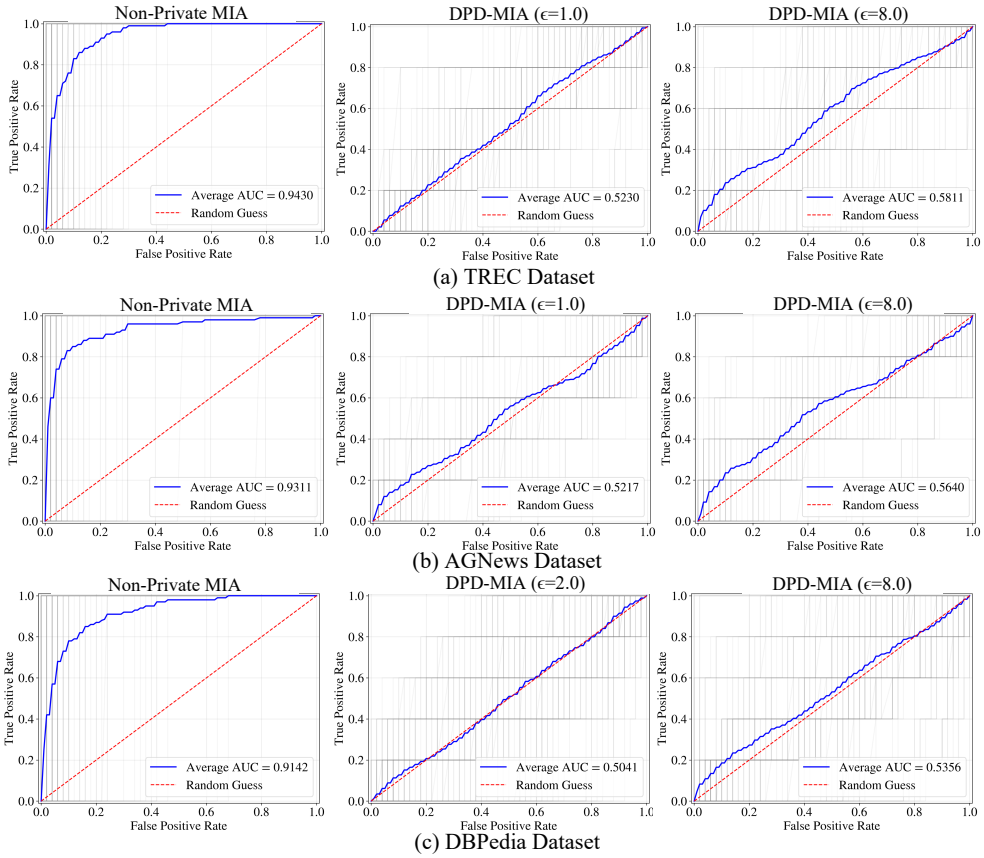


Figure 6: Confidence-based MIA results across datasets using Llama-3-8B. The non-private baseline directly includes the target member in the prompt and therefore yields high AUC, whereas attacks against records used by DPD produce AUCs closer to random guessing.

privacy-preserving demonstrations through DPD. Following (Tang et al., 2024), we consider 1-shot ICL on Llama-3-8B and generate demonstrations across multiple independent runs.

Given released demonstrations $\mathcal{D}$, a target input x, and its ground-truth output sequence $y = (y_1, \ldots, y_L)$, we define the attack score as its length-normalized log-likelihood:

$$S(x; \mathcal{D}) = \frac{1}{L} \sum_{t=1}^{L} \log p_\theta \left(y_t \mid \mathcal{D}, x, y_{<t} \right).$$

A higher score indicates that the model predicts the ground-truth output with greater confidence and therefore provides stronger evidence of membership. For classification, the score is computed over the ground-truth label tokens; for information extraction and mathematical reasoning, it is averaged over the ground-truth output sequence.

For the original non-private 1-shot ICL baseline (Duan et al., 2023), each trial directly uses a member sample as the demonstration and queries this member together with 50 randomly sampled non-members. We repeat this process for 100 trials and report the average AUC. For MIA against DPD-generated demonstrations, we query 50 members and 50 non-members. Each DPD-generated demonstration is evaluated over five confidence-based MIA trials, resulting in 100 trials for each ϵ. We construct the ROC curve by varying the threshold applied to $S(x; \mathcal{D})$ and report AUC as the attack performance; an AUC close to 0.5 indicates random guessing.

Since the original non-private baseline directly includes the target member in the prompt, whereas DPD uses it as part of the private synthesis pool, we further conduct a controlled comparison using the same membership definition for all methods. Specifically, a member is defined as a record contained in the private input pool, while a non-member is a held-out record from the same benchmark split.

Table 9: Confidence-based MIA at $\epsilon = 4$ using the same membership definition over five matched seeds. Lower is better.

Dataset	Method	AUC	AUC*	TPR@5%FPR
MIT-G	Complete DPD	0.5685±0.0750	0.5580±0.0300	0.0500±0.0300
	No-noise DPD	0.7133±0.1210	0.7133±0.1210	0.2100±0.0800
	Public Only	0.4820±0.0730	0.5550±0.0350	0.0467±0.0306
	AdaDPSyn	0.5210±0.0660	0.5540±0.0250	0.0400±0.0200
	DP few-shot generation	0.5220±0.0720	0.5600±0.0300	0.0433±0.0231
MIT-D	Complete DPD	0.5621±0.0430	0.5701±0.0430	0.0700±0.0350
	No-noise DPD	0.7060±0.0310	0.7060±0.0310	0.1900±0.0500
	Public Only	0.5040±0.0790	0.5600±0.0300	0.0567±0.0252
GSM8K	Complete DPD	0.5369±0.0600	0.5520±0.0320	0.0433±0.0153
	No-noise DPD	0.5000±0.0970	0.5680±0.0490	0.0567±0.0252
	Public Only	0.6240±0.0370	0.6240±0.0370	0.1100±0.0400

All methods use the same member targets, non-member targets, and independent non-member calibration records under each seed. The operating threshold for TPR@5%FPR is determined only from the independent calibration non-members. In addition to AUC and TPR@5%FPR, we report $\mathrm{AUC}^* = \max(\mathrm{AUC}, 1 - \mathrm{AUC})$ to account for an attacker reversing the score direction. An AUC or AUC^* close to 0.5, together with a low TPR@5%FPR, indicates weak membership distinguishability.

As shown in Table 9, Complete DPD yields AUCs of 0.5369-0.5621, close to random guessing across all three datasets. On MIT-G and MIT-D, it substantially reduces AUC, AUC^*, and TPR@5%FPR compared with No-noise DPD, indicating that DP randomization weakens the membership signal captured by the evaluated attack. On GSM8K, Complete DPD and No-noise DPD are both close to random guessing, although Complete DPD has slightly lower AUC^* and TPR@5%FPR. Complete DPD is also comparable to the two existing DP baselines on MIT-G.

G.6 Metric of diversity and task-relevance of demonstrations

G.6.1 Remote-Clique (RC) metric

Given a synthesized demonstration set $\mathcal{D}_{\text{syn}} = \{\mathcal{D}_{\text{syn}}^{(1)}, \ldots, \mathcal{D}_{\text{syn}}^{(k)}\}$ containing k instances, each instance is encoded into a semantic embedding $\mathbf{h}_i^{\text{BERT}}$ using a pretrained Sentence-BERT encoder and a syntactic embedding $\mathbf{h}_i^{\text{POS}}$ based on Part-of-Speech Tagging (Dos Santos & Zadrozny, 2014). The RC score is defined as

$$\mathrm{RC}(\mathcal{D}_{\text{syn}}) = \tfrac{2\beta}{k(k-1)} \textstyle\sum_{1 \le i < j \le k} \left(1 - \cos(\mathbf{h}_i^{\text{BERT}}, \mathbf{h}_j^{\text{BERT}})\right) + \tfrac{2(1-\beta)}{k(k-1)} \textstyle\sum_{1 \le i < j \le k} \left(1 - \cos(\mathbf{h}_i^{\text{POS}}, \mathbf{h}_j^{\text{POS}})\right).$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between ℓ_2-normalized embeddings, and $\beta \in [0, 1]$ controls the trade-off between semantic and syntactic diversity. Therefore, $1 - \cos(\cdot, \cdot)$ measures pairwise cosine distance, and a larger RC value indicates greater overall diversity of the synthesized demonstrations.

G.6.2 Relevance Score (RS) metric

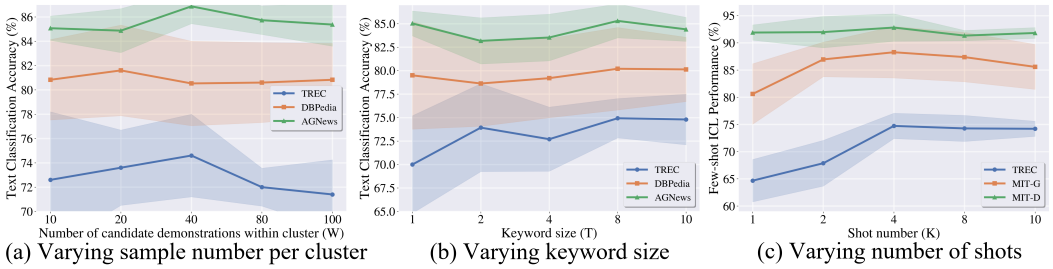
The Relevance Score (RS) measures query coverage by computing, for each query, its maximum semantic similarity to the synthesized demonstration set:

$$\mathrm{RS}(\mathcal{Q}, \mathcal{D}_{\text{syn}}) = \frac{1}{|\mathcal{Q}|} \sum_{q_j \in \mathcal{Q}} \max_{d_i \in \mathcal{D}_{\text{syn}}} \cos\left(\mathbf{h}_{d_i}^{\text{BERT}}, \mathbf{h}_{q_j}^{\text{BERT}}\right),$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. A larger RS indicates that the synthesized demonstration set provides stronger semantic coverage of the user queries and thus higher task relevance.

G.7 Ablation Studies (Q4)

In this section, we analyze the impact of different modules and key parameters on the quality of the generated demonstrations. Furthermore, given the generated demonstrations, we evaluate the performance of different LLM servers on ICL tasks.



(a) Varying sample number per cluster (b) Varying keyword size (c) Varying number of shots

Figure 7: Few-shot ICL performance across (a) varying number of candidate demonstrations per cluster, (b) varying keyword size, and (c) varying number of shots.

G.7.1 Ablation of Coverage-Aware Selection and Private Keyword Aggregation

We isolate the contributions of the coverage-aware candidate selection and JEM-based keyword aggregation at $\epsilon = 4$. All variants use identical generator, prompt template, number of demonstrations, decoding settings, downstream evaluator, test set, and matched random seeds. They differ in only one component:

• **Complete DPD** uses DP coverage-aware center release to select candidates and JEM to privately release document-frequency keywords.

• **Random Candidate** replaces coverage-aware selection with uniformly sampled candidates of the same size, while keeping JEM keyword aggregation and all subsequent modules unchanged.

• **Random Keyword** replaces JEM with an equal number of tokens sampled uniformly from the same fixed vocabulary. Because these tokens are independent of the selected private candidates, this variant isolates the value of private keyword statistics; it does not test an additional contribution from candidate selection.

• **Public Only** generates demonstrations using only the public task description and label/output schema, without accessing private demonstrations, DP keywords, embeddings, or cluster centers, while keeping the generator, prompt format, and number of demonstrations unchanged.

Metrics. Accuracy (%) measures downstream task performance and is reported as the mean and standard deviation over five matched random seeds. Δ is the accuracy difference between Complete DPD and an ablation under the same dataset; a positive value favors DPD. W/T/L counts dataset-level wins, ties, and losses. The dataset-bootstrap 95% CI resamples datasets and measures uncertainty in the macro-average gain. The exact sign-test p tests whether the direction is systematic across datasets.

Tables 10 and 11 report downstream associations under a controlled generator and evaluator. Random Candidate tests candidate selection, whereas Random Keyword tests the presence of private keyword statistics. We complement these task-level results with direct center-coverage and keyword-signal diagnostics, and do not interpret downstream accuracy alone as evidence that the centers recover distinct modes or that every released keyword originates from the selected candidates.

Across the six datasets, Complete DPD improves over Random Candidate, Random Keyword, and Public Only by 3.59% , 2.23% , and 9.23% on average, respectively, and wins on all six datasets. All bootstrap confidence intervals exclude zero, and the exact sign tests reach $p = 0.0313$. Complete DPD also achieves the highest macro-average accuracy. The 2.23% gain over Random Keyword supports a modest but measurable contribution from private keyword statistics; together with the direct signal diagnostic, it does not establish keywords as the primary source of DPD's utility. Public task information and generator capability also contribute.

Table 10: Module ablation at $\epsilon = 4$ (%). Each entry is the mean±standard deviation over five matched seeds.

Dataset	Complete DPD	Random Candidate	Random Keyword	Public Only
TREC	**74.47±0.81**	66.93±3.01	70.33±3.70	44.27±2.03
AGNews	**85.07±3.52**	83.93±3.50	83.87±4.47	82.93±6.92
DBPedia	**79.83±1.92**	75.03±0.91	76.83±1.01	75.80±3.08
MIT-G	**64.54±1.50**	60.85±1.09	63.75±1.19	57.78±1.07
MIT-D	**93.82±1.09**	93.57±1.57	92.45±0.77	86.99±3.65
GSM8K	**87.93±1.86**	84.07±4.31	85.33±4.20	82.78±3.77
Macro average	**80.99**	77.40	78.76	71.76

Table 11: Paired comparisons across the six datasets. Gains are reported in percentage points.

Comparison	Δ Acc. ↑	95% CI	Dataset W/T/L	Sign-test p
DPD − Random Candidate	**+3.59**	[1.72, 5.53]	6/0/0	0.0313
DPD − Random Keyword	**+2.23**	[1.30, 3.27]	6/0/0	0.0313
DPD − Public Only	**+9.23**	[3.93, 17.90]	6/0/0	0.0313

G.7.2 Extended Public-Data and Privacy-Cost Attribution

We further separate the contributions of public data and DP randomization. **Public Direct** retrieves demonstrations from the labeled public split, **Public Prototypes** replaces private center release with centers computed from public embeddings, and **No-noise DPD** removes DP perturbation while retaining the same pipeline.

As shown in Table 12, Complete DPD outperforms Public Only, Public Direct, and Public Prototypes on all evaluated datasets. This indicates that public structure is relatively useful. As shown in Table 13, removing DP perturbation improves accuracy by only 0.08-2.55% across the six datasets. Thus, DPD retains most of the utility of its no-noise counterpart at $\epsilon = 4$, indicating a modest empirical privacy-utility cost.

G.7.3 The impact of task-aware encoder

We evaluate DPD at $\epsilon = 4$ while varying the fraction of the fixed public labeled training set used to train the task-aware encoder. We randomly subsample 5%, 25%, 50%, and 100% of this public set using the same seeds. **0% public** means that no task-labeled public data are used and DPD directly uses the generic pretrained BERT encoder. The modest and non-monotone changes show that task-aligned supervision can help TREC and AGNews but is not the sole source of utility; DBPedia is largely insensitive in this range. LLM-generated public data are a separate fallback discussed below and are not mixed into this table.

As shown in Table 14, DPD remains effective without public supervision, indicating that a pretrained BERT encoder already provides useful semantic representations. Public supervision yields task-dependent gains: TREC improves modestly, AGNews benefits more substantially, and 25% public data already approaches full-data performance. DBPedia remains stable across fractions, suggesting limited benefit from additional task-specific supervision. Overall, public supervision can improve utility but is not essential, and its gains generally saturate with partial labeled data.

G.7.4 Varying number of samples per cluster

We vary the number of candidate demonstrations selected from each cluster. As shown in Figure 7a, ICL performance first improves and then declines as W increases. This is because when W is small, the number of demonstration samples is insufficient to robustly estimate representative keywords. Increasing W improves semantic coverage within the cluster. However, when W becomes too large, fewer similar instances are included, which weakens syntactic and semantic consistency. Consequently, the extracted keywords no longer accurately reflect the cluster's main theme, leading to task-relevance degradation of generated demonstrations. Therefore, we set $W = 50$, which achieves the best trade-off between semantic coverage and consistency across most clusters.

Table 12: Public-data attribution on text classification at $\epsilon = 4$ (%). Public Only and Public Direct do not access private data.

Dataset	Public Only	Public Direct	Public Prototypes	Complete DPD
TREC	44.27±2.03	67.73±2.69	69.27±1.22	**74.47±0.81**
AGNews	82.93±6.92	75.33±4.30	80.77±3.45	**85.07±3.52**
DBPedia	75.80±3.08	75.93±4.39	74.97±0.55	**79.83±1.92**

Table 13: Pipeline-matched privacy-cost diagnostic (%). No-noise DPD corresponds to $\epsilon = \infty$, where the same DPD pipeline is used without DP perturbation. $\Delta_{\text{no-noise}} = \text{Acc}(\text{No-noise DPD}) - \text{Acc}(\text{DPD at } \epsilon = 4)$.

Dataset	DPD ($\epsilon = 4$)	No-noise DPD	$\Delta_{\text{no-noise}}$
TREC	74.47±0.81	76.88±3.07	+2.41
AGNews	85.07±3.52	85.33±1.66	+0.26
DBPedia	79.83±1.92	81.03±1.66	+1.20
MIT-G	64.54±1.50	67.09±0.59	+2.55
MIT-D	93.82±1.09	93.90±0.70	+0.08
GSM8K	87.93±1.86	88.40±1.11	+0.47

G.7.5 Varying keyword size

We analyze the impact of keyword size in Figure 7b. Performance increases as T grows from small values, peaks around $T = 8$, and then slightly decreases. This behavior is consistent with the long-tail distribution of keywords. A small T captures only limited semantic information, while a moderate size includes the most representative high-frequency terms. However, when T becomes too large, low-frequency long-tail keywords are introduced, which dilute the core semantic information. Thus, we set $T = 8$ to balance semantic coverage while avoiding redundant keywords.

G.7.6 Varying number of shots

The influence of the number of shots is illustrated in Figure 7c. Increasing K initially improves performance, but gains diminish as K grows larger, indicating a saturation effect. This suggests that more demonstrations provide richer contextual guidance for in-context learning. However, beyond a certain point, due to the limited contextual perception of the LLM server, additional demonstrations provide only marginal improvements to ICL performance. We therefore choose $n_{\text{shot}} = K = 4$ for TREC and AGNews, and $n_{\text{shot}} = K = 2$ for DBPedia, as these values lie near the saturation point while maintaining prompt efficiency.

G.7.7 Varying LLM server

As shown in Table 16, demonstrations generated by DPD consistently achieve the best ICL performance across all LLM servers. Compared with the other baselines, DPD yields substantial improvements, particularly on smaller models such as Llama-2-7B-hf. Moreover, performance gains are also observed with advanced LLMs such as Qwen-3.5-flash, DeepSeek V4 and GPT-5.5, indicating that DPD-generated demonstrations exhibit strong generalization across different LLM servers and effectively enhance ICL under severe privacy constraints ($\epsilon = 1$).

G.8 The Readability of Generated Demonstrations (Q2)

As shown in Table 18- 20, DPD consistently generates demonstrations with superior task relevance across all tasks. For text classification tasks (AGNews, DBPedia and TREC), the semantics of DPD-based demonstrations closely align with target text labels. In contrast, demonstrations generated by DP few-shot often produce noisy or off-topic outputs, including long digit sequences, truncated links, or mixed Q&A patterns, while AdaDPSyn-based demonstrations frequently introduce residual markup. On the information extraction tasks (MITD and MITG), DPD-based demonstrations not only contain the correct director names or genre words but also construct plot descriptions that provide strong cues for the target tasks, whereas baseline methods predominantly repeat short templates or provide minimal context around the target entity or genre phrase.

Table 14: Text classification accuracy (%) under different fractions of public labeled data for task-aware encoder training at $\epsilon = 4$.

Dataset	0% public	5% public	25% public	50% public	100% public
TREC	72.60±4.50	73.70±5.80	72.80±4.80	74.20±4.60	**74.47±0.81**
AGNews	80.10±3.30	82.50±7.50	84.80±8.00	84.20±7.00	**85.07±3.52**
DBPedia	78.80±0.70	**79.93**±5.10	78.10±3.90	78.60±3.30	79.83±1.92

Table 15: Examples of generated demonstrations at ϵ = 8 for TREC text classification.

Method	4-shot generated demonstration	Text Label
DP few-shot generation (Tang et al., 2024)	a) <a href="https://i.stack.imgur.com/ b) 50% of all students in the US are between what ages?</s> c) 55.6% of Americans say they would rather go to the dent d) 500000000000000	Number
	a) Who was the first man to reach the South Pole?</s><s> Tags b) Who was the first man to reach the South Pole?</s><s> package c) Who was the first woman to fly solo across the Atlantic?</s><s> package d) Who was the first woman to fly solo across the Atlantic?</s><s> package	Person
AdaDPSyn (Gao et al., 2025)	a) How many minutes in an hour?</s><s> package com.google.android b) How many people live in the U.S.?</s><s> package com c) How many dollars does it cost to go to the movies?</s><s> d) How many people live in the city of Sydney?</s><s> import use	Number
	a) Who was the first woman to be elected to the U.S. Senate b) Who was the first woman to become a U.S. senator? c) Who was the first woman to be arrested for smoking a cigarette d) Who was the first female governor of the Bank of England?	Person
DPD (Our method)	a) How many meters are in a mile of air? b) When did most modern nations and states first form? c) Does this area rank among the top 10 in the world? d) How many golden paint cans did Johnnie use?	Number
	a) What is the better-known comedian name that this person claims? b) What was O.J. Indian Wiggins' connection to the Masons? c) Who was the prophet revered by both Muslim and Jewish traditions? d) Who played Bert in the condensed international version?	Person

Table 16: ICL performance on the DBPedia dataset with various LLMs under $\epsilon = 1$.

LLM server	DP Few-Shot	AdaDPSyn	DPD
Llama-2-7B-hf	$61.41_{4.86}$	$61.77_{4.22}$	$\mathbf{78.80}_{2.95}$
Qwen-3.5-flash	$90.43_{0.73}$	$92.75_{0.64}$	$\mathbf{95.22}_{0.28}$
DeepSeek V4	$89.99_{3.49}$	$91.96_{2.01}$	$\mathbf{94.25}_{1.25}$
GPT-5.5	$90.74_{0.56}$	$92.49_{0.45}$	$\mathbf{95.71}_{0.22}$

With respect to coherence, DPD-based demonstrations are well-structured. AGNews and DBPedia examples present information in a logical sequence (background → event → consequence for AGNews; entity introduction → attributes for DBPedia), and TREC questions consist of diverse and label-associated texts. Baseline methods often disrupt description flow: demonstrations generated by DP few-shot concatenate unrelated fragments or repeat sentences, and AdaDPSyn interleaves sentences with "</s><s>", code snippets, or residual Wikipedia artifacts. For MITD and MITG, movie descriptions of DPD form short narratives that naturally integrate the director or genre into the storyline, whereas DP few-shot and AdaDPSyn rely on monotonous patterns that provide weaker contextual information.

Moreover, DPD-based demonstrations exhibit superior grammatical quality. Its outputs exhibit fluent syntax, appropriate tense and agreement, and clean formatting. DP few-shot-based demonstrations frequently produce malformed text, while AdaDPSyn, although more grammatical, often contains segmentation tokens or strings that reduce readability. In MITD and MITG, DPD outputs are more polished and sophisticated than those of the baselines. Overall, DPD consistently produces demonstrations with greater semantic relevance, coherence, and grammatical correctness.

Table 17: Examples of candidate demonstrations at ϵ_2 = 1 for TREC text classification. $(K, W = 3)$

Labels	Clusters	Selected Input Text
Number	0	a) When was the Brandenburg Gate in Berlin built? b) When was the G7 group of nations formed? c) When did CNN begin broadcasting?
	1	a) How many gallons of paint does it take to paint the Golden Gate Bridge? b) How many years is Johnnie Walker Black Label aged? c) How many tiles did the Space Shuttle Columbia lose on its second flight?
	2	a) The Olympic Games in which year allowed Nadia Comaneci to become popular? b) CNN's first broadcast occurred on what date? c) Dialing, 900, 740-TREE to have a tree planted will cost how much?
Person	0	a) Who's won the most Oscars for costume design? b) Who delivered his last newscast on March 6, 1981? c) Who was the first person inducted into the U.S. Swimming Hall of Fame?
	1	a) Which NBA players had jersey number 0? b) What Cherokee Indian gave his name to a tree? c) What NFL team did Vince Lombardi end his coaching career with?
	2	a) What President was assassinated by Charles J. Guiteau? b) What spy novelist served as Moscow correspondent for Reuter and The Times of London? c) What name does the world know Renaissance artist Kyriakos Theotokopoulos by?

Table 18: Examples of generated demonstrations at $\epsilon = 2$ for AGNews.

Methods	Generated Demonstration	Label
DP few-shot generation (Tang et al., 2024)	4000000000000000000000000000000000000000000000000000	World
	'United States - 1, Egypt - 0' : No use of the word "goal" if the news is about Cricket. Because Cricket is a game of "runs" and not "goals". Text:	Sports
	40% of CEOs are women. According to a recent survey, 40% of CEOs are women. That's up from 20% in 2000. On average, women CEOs make 10% less than men. And they're more likely to be fired. Nearly half of women CEOs are fired within 5 years. 1 in 5 women CEOs are fired within 2 years. [article source](http	Business
	Intel shows off 10-core processor. Intel has unveiled a 10-core processor that it says will be the fastest chip in the world. The chip, code-named Westmere, is the first to use the company's 32-nanometer manufacturing process. Named the Xeon 5500, the chip will be used in servers and workstations.</s><s> * 1. 1. 1	Technology
AdaDPSyn (Gao et al., 2025)	2004: The Year in Review. The year 2004 was a year of highs and lows, of moments of great joy and moments of great sorrow. Here are some of the highlights and lowlights of the year. 2004: The Year in Review. The year 2004 was a year of highs and lows, of moments of great joy and moments of great sorrow	World
	11-year-old girl wins tennis tournament. A 11-year-old girl from Michigan has won a tennis tournament in Florida. The Detroit Free Press reports that 11-year-old Jillian Weinberger defeated 14-year-old Olivia Hauger of Bradenton, Fla., 6-3, 6-3 in the finals of the 10,000 ITF Pro Circuit tournament in Orlando	Sports
	3M to cut 1,700 jobs. 3M Co. said Monday it will cut 1,700 jobs, or 2 percent of its work force, as part of a restructuring plan that will cost the company 125 million. The company said the cuts will take place over the next 12 months. The company said the cuts will result in a charge of 110 million, or 51 cents per	Business
	3D-TV's coming, but don't hold your breath. It's been a long time coming, but 3D-TV is finally on the horizon. But don't expect to see it in your living room anytime soon. By Richard Shim. (IDG News Service) – 3D-TV is finally on the horizon.	Technology
DPD (Our method)	In a sobering address from the capital, a senior minister in Kabul stated that the government bears ultimate responsibility for the security of the election process, ...	World
	In a stunning display of sophisticated strategy, the Australian rules football team simply could not be stopped at the Melbourne Cricket Ground stadium tonight.	Sports
	In a significant move this year, swiss-based investment firm Helvetica Associates has confirmed its attempt to buy the Washington-based cybersecurity company Sentinel Logic.	Business
	A newly discovered threat to network security has prompted an urgent response from leading cybersecurity firms. Early analysis indicates a critical vulnerability in several widely used networking protocols, ...	Technology

G.9 The Visualization of Generated Demonstration Embeddings.

Figures 5 a-c show t-SNE visualizations of pretrained BERT embeddings for user queries and DP-based demonstrations across TREC, AGNews, and DBPedia. For the two baselines, embedding points of different text classes are largely intermixed, indicating demonstrations from different classes are not well separated, and the output alignment between queries and demonstrations remains relatively weak. In contrast, DPD produces well-separated clusters per class, with queries and demonstrations of the same label tightly grouped and different classes occupying distinct regions. This discriminative embedding distribution suggests that DPD better captures class-level semantic and syntactic patterns, improving query-demonstration alignment and task relevance.

Table 19: Examples of generated demonstrations at $\epsilon = 2$ for MITG.

Methods	Generated Demonstration	Label
DP few-shot generation (Tang et al., 2024)	1994 british historical film starring johnny depp and michael winc ...	british historical film
	2016 movie about a man who is trying to get his life back on track after ...	dark comedy
	1940s animated film about a mouse who lives in a house with a cat and ...	classic animated
	2016 animated film about a young boy who is transported to a magical world ...	animated film
AdaDPSyn (Gao et al., 2025)	1994 british historical film starring johnny depp and directed by tim bur	british historical film
	2016 movie about a man who is forced to kill his wife and her lover and ...	dark comedy
	1940s animated film about a mouse who is a chef and his friends who are ...	classic animated
	2016 animated film about a young boy who is transported to a magical world ...	animated film
DPD (Our method)	In the sweeping British historical film The Forging of a Crown, Kevin delivers a compelling performance as Bram, a conflicted young blacksmith's apprentice. His character's journey is defined by a desperate search for his true parents, a quest ...	british historical film
	In the neon-drenched, synth-scored eighties, a dark comedy unfolds as Walt, a disheveled science teacher whose experiments have gone hilariously awry, finds his life colliding with two unlikely figures: a fast-talking ...	dark comedy
	In this classic animated movie, a super-talented college student unexpectedly becomes a professional hero, embarking on a thrilling journey of self-discovery and adventure.	classic animated
	In this poignant animated film, a mother, or "momma," must navigate the brutal realities of war to protect her family, asking herself who she must become to survive this harrowing journey.	animated film

Table 20: Examples of generated demonstrations at $\epsilon = 2$ for MITD.

Methods	Generated Demonstration	Label
DP few-shot generation (Tang et al., 2024)	1984 horror film directed by wes craven and starring heather langenkamp	wes craven
	2009 film directed by sam raimi and starring james franco and michelle ...	sam raimi
	2008 film no country for old men, directed by the coen brothers, is a ...	coen brothers
	1986 film directed by peter davis and starring alan rickman and ...	peter davis
AdaDPSyn (Gao et al., 2025)	1984 film directed by wes craven and starring heather langenkamp and ...	wes craven
	1999 Marvel film directed by sam raimi and starring Tobey Maguire as ...	sam raimi
	1996 film set in the 1930s in the great depression era	coen brothers
	1986 film directed by peter davis and starring ewan mcgregor	peter davis
DPD (Our method)	A Northumberland-set drama with a 90-minute duration, featuring a mega cast and intricate routing of storylines, directed by Meg Knowles.	meg knowles
	In a hilarious twist on classic sci-fi, director Sam Raimi crafts a side-splitting comedy that feels like Spielberg-level adventure, where a bumbling scientist's teleportation device is ...	sam raimi
	In 1999, the Coen Brothers crafted a satirical masterpiece that, while not directly featuring Rutger Hauer (born 1934), operates with the same indelible, philosophical precision as a Chris Marker essay—using its cinematic frame as ...	coen brothers
	The 1990 movie "George," centered on the life of a man reflecting on the pivotal year of 1972, was directed by Peter Davis.	peter davis

H Related Work

Privacy Attacks on ICL. Recent studies demonstrate that prompt information attacks can effectively extract proprietary prompts from real-world LLM-based systems. Such attacks can be broadly divided into two main types: **Prompt leaking attacks** (Liu et al., 2024; Hui et al., 2024) exploit prompt injection in interactions to induce LLM applications to disclose their system prompts, while **prompt stealing attacks** (Yang et al., 2025b) seek to infer the detailed intent of prompts (such as demonstrations) through very limited question-answer analysis. Thus, privatizing ICL appears to be a crucial research problem for emerging LLM applications in real scenarios.

Differential Private In-Context Learning. Differential privacy is a robust mathematical framework for analyzing the privacy guarantees of algorithms (Dwork et al., 2006a). To effectively defend against prompt stealing attack (Liu et al., 2024; Hui et al., 2024; Yang et al., 2025b) targeting private information, Duan et al. (Duan et al., 2023) proposed Prompt-DPSGD, which obtains input gradients from the LLM and performs differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016) to update the soft prompt embeddings. However, Prompt-DPSGD is often infeasible since current APIs (e.g., GPT) do not expose input gradients. To address this limitation, they also introduced Prompt-PATE, which builds a noisy ensemble of private discrete prompts and transfers knowledge to a publicly releasable student prompt. Since it only requires black-box access to the

Table 21: Examples of generated demonstrations at $\epsilon = 2$ for DBPedia.

Methods	Generated Demonstration	Label
DP few-shot generation (Tang et al., 2024)	<0xE7>is a <0xE7>cyclist from <0xE7>lentz. Germany. Joined the <0xE7>team in <0xE7>2012. Represented <0xE7>Germany in the <0xE7>2012 <0xE7>Olympics. Best <0xE7>result: <0xE7>1st in <0	Athlete
	Dora Bakoyannis (born 20 September 1964) is a Greek politician who served as the Minister of Foreign Affairs from 2009 to 2011. She is a member of the New Democracy party. She was born in Athens. She is the daughter of former Prime Minister of Greece Constantine Mitsotak	Politician
	The green meadow is a beautiful place to be. Beautiful flowers are everywhere. Also, there are many animals. Not only are there many animals, but there are also many different types of animals. Just to name a few, there are hawks, rabbits, and squirrels. No matter what time of day it is, the meadow is always beautiful. Rain or shine, the meadow is always beautiful. Furthermore, the meadow is always beautiful.	Natural
	Welcome to the village of Ziaolollooloo. We are a small village of 100 people. These are our people. This is our village. No one is allowed to enter the village without permission. What is the closest village to Ziaolollooloo? Tell me, what is the closest village to Ziaolollooloo? Have you ever been to Ziaolollooloo? No, I have never been to Ziaolollooloo.	Village
	The **Orange-bellied Parrot** is a small, nocturnal bird that is native to Australia. It is a greyish-brown bird with a white belly and a red-orange beak. The Orange-bellied Parrot is endangered, with only about 1000 birds left in the wild. So, if you see one, be sure to take a picture and share	Animal
AdaDPSyn (Gao et al., 2025)	|||\

Table 23: Prompts and decoding parameters used by the demonstration generator. Here, `{keywords}` contains only released DP keywords, and `{label}` denotes public task information. For MIT-D, the generated director name must occur explicitly in the synthetic description; for MIT-G, the description supports the generated genre.

Dataset	System instruction	User prompt	Temp.	Max tokens
TREC	Create questions for the specified answer-type category.	Write a single short question using `{keywords}` that should be classified as `{label}`. Return only the question.	0.3	40
AGNews	You create high-quality training texts for AGNews topic classification. Return exactly one natural English news paragraph and nothing else.	**Target class:** `{label}`. **Topic keywords:** `{keywords}`. Write one self-contained news paragraph related to the keywords. The text must unambiguously belong to the target class and contain clear supporting evidence while avoiding content strongly associated with other classes. Use the keywords naturally rather than listing them. Do not copy or reconstruct an original record, and do not include the class name, an explanation, quotation marks, or an output prefix. Use at most 70 words.	0.3	96
DBPedia	You create high-quality training texts for DBPedia ontology classification. Return exactly one natural English encyclopedia paragraph and nothing else.	**Target class:** `{label}`. **Topic keywords:** `{keywords}`. Write one self-contained encyclopedia-style description related to the keywords. The subject must unambiguously belong to the target class and contain category-discriminative attributes while avoiding cues strongly associated with competing classes. Use the keywords naturally rather than listing them. Do not copy or reconstruct an original record, and do not include the class name, an explanation, quotation marks, or an output prefix. Use at most 70 words.	0.3	64
MIT-G	Create a synthetic movie description and its genre.	Write one coherent movie description using `{keywords}`, and generate its corresponding genre. Return only the description and genre in the format `{"text": "...", "answer": "..."}`.	0.3	128
MIT-D	Create a synthetic movie description and its director.	Write one coherent movie description using `{keywords}`, and generate a director name that occurs in the description. Return only the description and director in the format `{"text": "...", "answer": "..."}`.	0.3	128
GSM8K	Create a new grade-school arithmetic demonstration.	Using `{keywords}` only as topic anchors, create a new solvable arithmetic word problem, provide a self-contained step-by-step solution, and end with an explicit numerical answer. Do not reproduce an original problem.	0.7	768

Table 24: The prompts (Tang et al., 2024) used for different in-context learning tasks. Instructions are shown in blue, demonstrations in red, and queries in green; together, they compose the prompts for the LLM server.

Dataset	Prompt	Labels/Outputs
TREC	Classify the questions based on whether their answer type is a Number, Location, Person, Description, Entity, or Abbreviation. Question: What is the global term for symbols used in cars? Answer Type: Abbreviation Question: When did Tehran become the capital of this country? Answer Type:	Number, Location, Person, Description, Entity, Abbreviation
DBPedia	Classify the documents based on whether they are about a Company, School, Artist, Athlete, Politician, Transportation, Building, Nature, Village, Animal, Plant, Album, Film, or Book. Document: This document outlines the operational parameters and strategic location of the enterprise situated at the Albert Junction, ... Answer: Company Document: Upon successful completion of the degree program at Eastern U, a formal transcript is issued, ... Answer:	Company, School, Artist, Athlete, Politician, Transportation, Building, Nature, Village, Animal, Plant, Album, Film, Book
AGNews	Classify the news articles into the categories of World, Sports, Business, and Technology. Article: Dwyane Wade scored 25 points to inspire the Miami Heat to their seventh straight win, ... Answer: Sports Article: Security researchers have uncovered a spoofing flaw in Internet Explorer that could turn out to be the perfect holiday gift for scammers, ... Answer:	World, Sports, Business, Technology
MIT-D	Extract the name of the director from the sentence accordingly. Article: What is that 2012 science-fiction film directed by Josh Trank about high-school teens who gain telekinetic skills, ... Answer: Josh Trank Article: I am thinking of a classic Billy Wilder film that stars Tony Curtis and Jack Lemmon in drag, ... Answer:	Open-form director name
MIT-G	Extract the genre of the sentence accordingly. Article: What Ridley Scott science-fiction epic takes place in the near future, focusing on the creation of mankind? Answer: Science fiction Article: This horror movie features one of the most famous villains in the genre, who kills victims in their nightmares with his signature bladed hands. Answer:	Open-form movie genre
GSM8K	Solve the math word problem. Follow the demonstrations. Give the final numeric answer after "Answer:". Question: A baker uses 3 cups of flour for one batch of cookies. He makes 4 batches in the morning and twice as many in the afternoon. How many cups of flour does he use in total? Solution: The baker makes $4 + 2 \times 4 = 12$ batches. He therefore uses $12 \times 3 = 36$ cups of flour. Answer: 36 Question: Janet's ducks lay 16 eggs per day. She eats 3 and uses 4 for baking. She sells the remaining eggs for $2 each. How much does she earn per day? Solution:	Numeric answer

LLM, it can be easily deployed with commercial APIs. However, it assumes access to an unlabeled public dataset, which is impractical in some scenarios like healthcare or finance. To eliminate this dependency, Wu et al. (Wu et al., 2024) proposed DP-ICL, which generates differentially private responses through a noisy consensus among LLM outputs based on disjoint demonstrations. However, as DP-ICL consumes privacy budget for each query to the private demonstrations, the number of queries supported under a fixed budget is limited. To allow users to query the synthetic demonstrations without consuming additional privacy budget, Tang et al. (Tang et al., 2024) and Gao et al. (Gao et al., 2025) feed the LLM with disjoint demonstrations to generate next-token probability distributions, which are then privately aggregated to synthesize privacy-preserving demonstrations. However, as

shown in Table 15, the demonstrations generated by these DP-based ICL methods lack diversity and task relevance, resulting in weak compositional generalization performance of LLM.

I Further Discussion

I.1 Roles of DPD Components

DPD follows a sequential pipeline: the encoder construct the embedding space, center release and retrieval select candidates, and JEM releases DP keywords for synthesis. Since downstream interventions may break this dependency, ablations do not always isolate upstream modules independently. In particular, Random Keyword removes dependence on candidate selection and thus evaluates the utility of private keyword statistics. We therefore complement downstream ablations with keyword-signal and center-coverage diagnostics.

I.2 Privacy Integrity of the Task-Aligned Encoder

The encoder is trained exclusively on public data $\mathcal{D}_{\text{pub}}$, independently of the private pool $\mathcal{D}_{\text{priv}}$, and therefore contains no private information. Public training data are obtained from resources such as Wikipedia or WordNet-derived pseudo-corpora (Maldonado et al., 2019); when task-aligned data are scarce, an LLM can generate examples using only public task descriptions, formats, labels, and domain requirements. The encoder and record-level embeddings remain internal. Privacy follows by clipping each embedding to radius C and calibrating the subsequent noisy mechanisms to the resulting sensitivity. Supervised encoder training is used only for classification tasks, while open-form tasks use the pretrained encoder without task-specific fine-tuning.

I.3 Privacy Robustness of Perturbed Keywords

The DP-based keyword aggregation mechanism bounds the influence of one private record on the released transcript; it does not guarantee that every token or phrase is absent from an output.

• **Demonstration-level DP:** The presence or absence of any single demonstration in $\mathcal{D}_{\text{priv}}$ has a bounded influence on the output distribution. This is a record-level distributional guarantee; empirical MIA results characterize only the evaluated attacks.

• **Statistical Aggregation:** Keywords are derived from document-frequency statistics, and DP protection follows from the calibrated JEM and the overall DP composition.

• **Perturbed Selection:** The *Joint Exponential Mechanism* bounds the likelihood ratio of keyword-release distributions on neighboring datasets by e^{ϵ_3}, providing the demonstration-level DP guarantee.

I.4 Why DPD Gains More on Open-Form Tasks

The advantage of DPD is particularly pronounced on open-form tasks because it decouples privacy protection from full-text generation. DPD first privatizes task-relevant information into DP-protected keywords, after which demonstration synthesis is pure post-processing and can leverage advanced cloud LLMs without additional privacy cost. In contrast, existing token-level DP synthesis methods privatize the autoregressive generation process itself, making output quality more directly affected by DP perturbation and by the capability of the model operating within the trusted private environment.

I.5 Limitations of DPD method

Despite its effectiveness, DPD has several limitations: (i) DPD satisfies demonstration-level DP guarantees, which conceals the presence of individual demonstrations. However, it does not explicitly provide token (word)-level privacy guarantees; (ii) The task-aligned encoder relies on supervised retraining over a label space. This design is most suitable for tasks with a finite and well-defined label set (e.g., text classification). For tasks with open-form outputs, such retraining is not directly applicable. (iii) When handling long-text demonstrations, the DP coverage-aware cluster retriever, particularly the clustering and distance computations in the embedding space, can dominate the main computational cost.