DP-ES: Differentially Private Evolution Strategies for Prompt Optimization

Anonymous ACL submission

Abstract

Differentially private (DP) prompt optimization is crucial for deploying Large Language Models (LLMs) on sensitive data. Yet, existing discrete methods suffer from high variance and brittle behavior under tight privacy budgets. On the GSM8K benchmark with $\varepsilon = 1.0$, DP-OPT attains only 49.5% accuracy with 28.5% standard deviation; in a 30-run analysis, 63% of runs fail due to template destruction and noisy selection. We propose **DP-ES**, a private evolution strategy that keeps exploration privacy-free while concentrating privacy budget on Gaussian-mechanism evaluation with optional exponential-mechanism smoothing for selection; crucially, mutations never access the private dataset, decoupling exploration from privacy expenditure. DP-ES achieves **88.1% on GSM8K** (38.6 pp improvement over DP-OPT with $\sim$9$\times$ lower standard deviation), **99.7% on MedQA**, **73.5% on BANKING77** (77-class intent classification), and **86.8% on Alpaca** instruction-following (LLM-judge), all under $(\varepsilon = 1.0, \delta = 10^{-5})$-DP over repeated runs; empirical audits pass on both closed-form tasks. Comprehensive ablations—including a selection-mechanism study comparing exponential-mechanism and deterministic top-k selection—robustness tests, and DP-OPT failure analyses explain why evolution-based search is well matched to DP constraints. Our work establishes evolution strategies as a robust, stable solution for privacy-critical prompt optimization. Code is available at https://anonymous.4open.science/r/dp-es-anonymous-F622/.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities through in-context learning (Brown et al., 2020; Ouyang et al., 2022), yet their performance heavily relies on carefully crafted prompts (Liu et al., 2023). While automatic prompt optimization (Yuksekgonul et al., 2024; Yang et al., 2023) can bridge this gap, optimizing prompts on **sensitive datasets** (e.g., patient records, financial logs) risks memorizing and exposing private information, as evidenced by recent membership inference and extraction attacks (Wang et al., 2023; Shanmugarasa et al., 2025; Fu et al., 2023). For instance, a naive optimizer might generate a prompt like *"Remember how patient #12847 had atrial fibrillation..."*, directly violating privacy. This necessitates algorithms with formal **Differential Privacy (DP)** guarantees (Dwork et al., 2006).

The Challenge: Task Complexity and Privacy Budgets. Applying DP to prompt optimization is non-trivial due to the discrete nature of text. Recent work like DP-OPT (Hong et al., 2024) uses histogram-based token generation with the exponential mechanism. Other approaches apply DP to in-context learning via private few-shot generation or nearest neighbor retrieval (Tang et al., 2024; Koskela et al., 2025), but they rely on additional retrieval or embedding machinery. While DP-OPT is effective on simple tasks, we identify critical limitations on complex reasoning: our analysis of 30 DP-OPT runs reveals a **63% failure rate** on GSM8K (Appendix C.2, Table 6), with systematic issues including template damage (destruction or partial loss) and suboptimal selection. This fragility stems from token-level myopia and non-backtracking search, which become catastrophic under tight privacy constraints.

Private Evolution Perspective. Private Evolution (PE) is a training-free DP paradigm for data synthesis via APIs (Lin et al., 2024; Xie et al., 2024; Lin et al., 2025). Its core advantage is structural: only a low-sensitivity evaluation and selection step touches private data, while exploration/generation is pure post-processing. This yields a decoupling between exploration and privacy expenditure—increasing the number or strength of mutations improves search without consuming additional pri-

vacy budget. DP-ES is a prompt-optimization instantiation of this PE principle.

Our Contribution: DP-ES. To overcome this privacy-utility gap, we propose **Differentially Private Evolution Strategies (DP-ES)**. Unlike methods that select from a fixed pool of discrete candidates, DP-ES maintains a population of prompts that evolve over time. Our key insight is to decouple exploration from privacy consumption: using LLMs to generate semantic mutations consumes **zero privacy budget** (by the post-processing property), allowing us to allocate the scarce privacy budget strictly for evaluation and selection.

Our contributions are:

1. **Method:** First evolution-based DP prompt optimization with privacy-free mutations and advanced composition.
2. **Robustness:** 0% failures vs 63% for DP-OPT across four tasks—GSM8K (88.1%), MedQA (99.7%), Alpaca (86.8%), BANKING77 (73.5%)—with $\sim$9$\times$ lower std on GSM8K.
3. **Efficiency:** 2.5$\times$ faster (17.9 vs 45.7 hours on GSM8K).
4. **Analysis:** DP-OPT failure diagnosis, population diversity analysis, and empirical privacy audits.

2 Related Work

Prompt Optimization. Gradient-free methods such as APE (Zhou et al., 2023) and OPRO (Yang et al., 2023) search over discrete prompt candidates using LLM-generated variations, while gradient-based approaches like TextGrad (Yuksekgonul et al., 2024) use LLM feedback as differentiable signals. Both families optimize prompts effectively but assume unrestricted access to training data with no privacy guarantees (Edemacu and Wu, 2024), making them unsuitable for sensitive domains such as healthcare, finance, and education where data protection is legally mandated.

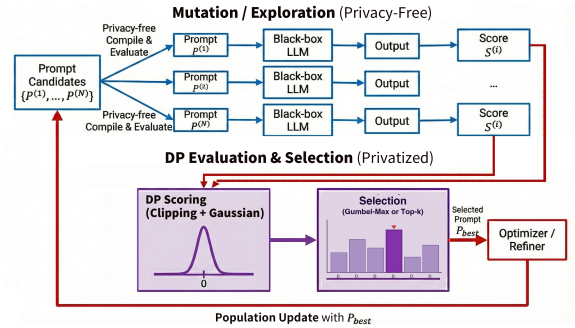
Differential Privacy in Machine Learning. DP-SGD (Abadi et al., 2016) and PATE (Papernot et al., 2018) provide formal (ε, δ)-DP guarantees for continuous model parameters during training. DP-LoRA (Liu et al., 2024b; Frery et al., 2025) extends this to parameter-efficient fine-tuning of LLMs, achieving strong performance with reduced memory overhead. However, all these methods require access to model weights, precluding their use with black-box API-only LLMs—the dominant deployment mode for state-of-the-art models such as GPT-4 and DeepSeek.

DP Prompt Optimization. For discrete prompts over black-box LLMs, DP-OPT (Hong et al., 2024) uses histogram-based token generation with the exponential mechanism, constructing prompts token by token from privately aggregated counts. While effective on classification and factual QA, this greedy token-level approach suffers from template damage and suboptimal selection on complex reasoning tasks (63% failure rate on GSM8K; Section 5), as noise accumulates across sequential token decisions without the ability to backtrack.

Other Approaches. Other directions include DP few-shot demonstration generation for in-context learning (Tang et al., 2024), which privately selects exemplars but cannot optimize the instruction prompt itself; private retrieval augmentation (Koskela et al., 2025), which protects retrieval queries but inherits the limitations of fixed prompt templates; and PromptDPSGD (Duan et al., 2023), which applies DP-SGD to continuous soft prompt embeddings but requires white-box model access and struggles on smaller models (Appendix C.3).

Private Evolution. Private Evolution (PE) (Lin et al., 2024; Xie et al., 2024; Lin et al., 2025) is a training-free DP framework originally developed for data synthesis via APIs. Its key structural advantage is that only a low-sensitivity evaluation step touches private data, while API-based generation is pure post-processing. This creates a favorable decoupling: exploration can be arbitrarily rich without consuming privacy budget. DP-ES adapts this principle from data synthesis to prompt optimization, inheriting the favorable privacy-exploration trade-off while introducing population-based search over the discrete prompt space.

Evolution Strategies and DP. Evolution strategies (ES) (Rechenberg, 1973) are a family of derivative-free optimization algorithms that maintain and evolve populations of candidate solutions. ES have driven advances in neural architecture search (Real et al., 2019) and reinforcement learning (Jaderberg et al., 2017), demonstrating strong performance in high-dimensional, noisy search spaces. While prior work has combined DP with optimization algorithms such as ERM (Chaudhuri et al., 2011), DP-ES is the first to leverage evolutionary search for DP prompt optimization in the discrete text domain. The population-based design provides natural robustness to noise-induced regressions—



(a) DP-ES system architecture showing the three-phase workflow: privacy-free mutation, DP evaluation, and DP selection.

	DP-OPT	DP-ES
Generation	Sample-and-aggregate (DP): $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^m$ LLM on each subset: $w_j = \text{LLM}(\mathcal{D}_j)$	Mutation via LLM (DP-free): $P^{(t)} \sim \text{Mutate}(P^{(t)}_{\text{parent}})$
Utility	Histogram counts per position: $h_w = \sum_{j=1}^m \mathbf{1}[w_j = w]$	Scalar score / feedback from black-box LLM: $S(P^{(t)})$
Privatization	LimitedDomain (top-k) + ExpMech: $w^* \sim P(w) \propto \exp(\epsilon h_w/(2\Delta))$ where $\Delta = 1$ for histogram counts	Add Gaussian noise to clipped scalar score ($\Delta = C_{clip}/n$): $\tilde{S}^{(t)} = \bar{S}^{(t)} + \mathcal{N}(0, \sigma^2_{eval})$
Selection	Token-by-token update: $P_{t+1} = \text{concat}(P_t, w^*)$ (append selected token)	**Option A** (default): EM via Gumbel-Max: $\hat{s}_p = \tilde{S} + \text{Gumbel}(0,$ **Option B**: Deterministic top-k: $\hat{s}_p = \tilde{S}$ (post-processing)

(b) Comparison of DP-OPT and DP-ES privacy mechanisms. DP-ES selection supports Option A (EM) or Option B (deterministic top-k).

Figure 1: DP-ES overview: (a) three-phase workflow with privacy-free mutation plus DP evaluation and selection (Option A: EM / Option B: top-k); (b) mechanism-level contrast with DP-OPT.

Method	Priv.	Disc.	Mechanism
TextGrad	×	✓	LLM gradients
OPRO	×	✓	LLM scoring
DP-SGD	✓	×	Gaussian noise
DP-OPT	✓	✓	Histogram + Exp
DP-ES	✓	✓	**Evol. + Gauss**

Table 1: Comparison of prompt optimization methods. DP-ES uniquely combines evolution strategies with DP for discrete prompts.

a critical advantage under tight privacy budgets where evaluation noise is substantial.

Privacy Auditing. Empirical verification of DP guarantees has become standard practice (Nasr et al., 2021; Jagielski et al., 2020). Neighboring-dataset audits estimate observed privacy leakage ε_{obs} by comparing output distributions on datasets differing in one record. We adopt this methodology to complement our theoretical guarantees, providing both formal proofs and empirical evidence that DP-ES preserves privacy in practice (Section 6).

Positioning of DP-ES. Table 1 summarizes the landscape. DP-ES uniquely combines evolutionary exploration with Gaussian-mechanism scoring and exponential-mechanism selection, filling the gap for a robust, black-box-compatible DP prompt optimization method that scales across diverse task complexities—from structured classification to multi-step reasoning.

3 Methodology: DP-ES

Problem Formulation. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a private dataset where x_i represents task inputs and y_i represents desired outputs (e.g., question-answer pairs). Let $\mathcal{P}$ denote the space of natural language prompts. Our goal is to find a prompt $p^* \in \mathcal{P}$ that maximizes task performance:

$$p^* = \arg\max_{p \in \mathcal{P}} f(p, \mathcal{D}) \tag{1}$$

where $f : \mathcal{P} \times \mathcal{D} \to \mathbb{R}$ measures prompt quality (e.g., accuracy on $\mathcal{D}$).

Privacy requirement: The optimization algorithm must satisfy (ε, δ)-differential privacy with respect to $\mathcal{D}$, ensuring that any single record (x_i, y_i) has bounded influence on the output prompt.

Definition 1 ((ε, δ)-Differential Privacy)**.** *A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{O}$ satisfies (ε, δ)-DP if for all neighboring datasets $\mathcal{D} \sim \mathcal{D}'$ (differing in one record) and all measurable sets $S \subseteq Range(\mathcal{M})$:*

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(\mathcal{D}') \in S] + \delta \tag{2}$$

DP-ES Algorithm Overview. DP-ES (Algorithm 1) employs an evolution strategy with three key components:

1. **Mutation:** Generate prompt variations using an LLM (privacy-free)
2. **DP Evaluation:** Score prompts on $\mathcal{D}$ with calibrated noise
3. **Selection:** Select top prompts from privatized scores (EM or deterministic)

Privacy-Free Mutations. Key Innovation: Unlike gradient-based methods that require DP protection for gradient computation, DP-ES uses LLMs to generate prompt variations *without accessing* $\mathcal{D}$. This consumes zero privacy budget.

Proposition 1 (Mutation Privacy-Freeness)**.** *Let Mutate $: \mathcal{P} \to \mathcal{P}$ be a function that generates a new prompt from a parent prompt using an LLM, without accessing the private dataset $\mathcal{D}$. Then Mutate consumes zero privacy budget.*

Algorithm 1 DP-ES: Differentially Private Evolution Strategies

Require: Private dataset $\mathcal{D}$, privacy budget $(\varepsilon_{\text{total}}, \delta)$, initial prompt p_0, population size K, iterations T
Ensure: Optimized prompt p^* with $(\varepsilon_{\text{total}}, \delta)$-DP
Initialize population $\mathcal{P}_0 \leftarrow \{p_0\}$
Allocate per-iteration budget: $\varepsilon_{\text{eval}}$ (and $\varepsilon_{\text{select}}$ if using EM) {via advanced composition}
for $t = 1$ to T **do**
  // Phase 1: Privacy-free mutation
  for each $p \in \mathcal{P}_{t-1}$ **do**
    Generate K children: $\{p'_1, \ldots, p'_K\} \sim$ LLM-Mutate(p)
  end for
  $\mathcal{P}_t \leftarrow \mathcal{P}_{t-1} \cup \{\text{all children}\}$
  // Phase 2: DP evaluation (consumes $\varepsilon_{\text{eval}}$)
  for each $p \in \mathcal{P}_t$ **do**
    Compute score: $s_p = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\text{LLM}(p, x_i) = y_i]$ {Accuracy $\in [0,1]$}
    Clip: $s'_p = \text{clip}(s_p, [0, C_{\text{clip}}])$ $\{C_{\text{clip}} = 1.0\}$
    Add noise: $\tilde{s}_p = s'_p + \mathcal{N}(0, \sigma^2)$ $\{\sigma = \Delta\sqrt{2\ln(1.25/\delta)}/\varepsilon_{\text{eval}}, \Delta = C_{\text{clip}}/n\}$
  end for
  // Phase 3: Selection over privatized scores
  Option A (default): $\hat{s}_p \leftarrow \tilde{s}_p + \text{Gumbel}(0, \beta)$ {EM smoothing; $\varepsilon_{\text{select}}$ absorbed into composition}
  Option B: $\hat{s}_p \leftarrow \tilde{s}_p$ {Deterministic top-k; pure post-processing}
  Select top-K: $\mathcal{P}_{t+1} \leftarrow \text{top-}K(\mathcal{P}_t, \{\hat{s}_p\})$
end for
return $p^* \leftarrow \arg\max_{p \in \mathcal{P}_T} \hat{s}_p$

Proof. See Appendix D.

Proposition 2 (Decoupled Exploration)**.** *If mutation does not access $\mathcal{D}$, then the privacy loss of DP-ES depends only on the DP evaluation and selection steps, and is independent of the number or strength of mutations per iteration.*

Proof. See Appendix D.

Mutation strategies: We implement 3 modes (EXPLORE, EXPLOIT, BALANCED); details and prompt templates are provided in Appendix A.

DP Evaluation via Gaussian Mechanism. After mutation, we evaluate the accuracy of each prompt on $\mathcal{D}$. The true score is the fraction of correct predictions:

$$s_p = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\text{LLM}(p, x_i) = y_i] \in [0, 1] \quad (3)$$

Note that s_p is non-negative, ranging from 0 (all incorrect) to 1 (all correct).

To protect privacy, we apply the Gaussian mechanism (Dwork and Roth, 2014):

Theorem 3 (Gaussian Mechanism for Scoring)**.** *Let s_p be clipped to $[0, C_{clip}]$ where C_{clip} is a fixed hyperparameter (set to $C_{clip} = 1.0$ to match the accuracy range $[0, 1]$). Since s_p is an average over n examples, its sensitivity is $\Delta = C_{clip}/n$. Adding Gaussian noise:*

$$\tilde{s}_p = clip(s_p, [0, C_{clip}]) + \mathcal{N}(0, \sigma^2),$$
$$\textit{where} \quad \sigma = \frac{\Delta\sqrt{2\ln(1.25/\delta)}}{\varepsilon_{eval}}, \quad (4)$$

satisfies $(\varepsilon_{eval}, \delta)$-differential privacy.

Selection over Privatized Scores. Once scores $\{\tilde{s}_p\}$ have been privatized by the Gaussian mechanism, any deterministic function of them is post-processing and incurs zero additional privacy cost (Dwork and Roth, 2014). Deterministic top-k is therefore always a valid, privacy-free selection strategy (**Option B** in Algorithm 1).

As our default, we apply the exponential mechanism (EM) (McSherry and Talwar, 2007) via the Gumbel-max trick (**Option A**), adding Gumbel$(0, \beta)$ noise with $\beta = 2C_{\text{clip}}/(n\,\varepsilon_{\text{select}})$. EM nominally consumes $\varepsilon_{\text{select}}$; however, **EM does not increase total privacy consumption** under Rényi composition. The scorer is invoked K times per iteration (TK=18 total) and dominates the composition bound; EM adds only T=3 invocations. Concretely, the Det. composition term is $\propto 18\varepsilon_0^2$, while EM's is $\propto 18(\varepsilon_0/2)^2 + 3(\varepsilon_0/2)^2 = 5.25\varepsilon_0^2$—a $3.4\times$ reduction. The Rényi accountant confirms $\varepsilon^{\text{EM}}_{\text{consumed}} \leq \varepsilon^{\text{Det}}_{\text{consumed}}$ empirically (Table 3). We adopt EM for its additional variance-reduction benefit (Section 5); deterministic top-k remains a simple alternative.

Privacy Composition. We compose all privacy-consuming operations (scoring and, if using EM, selection) via the Rényi DP accountant (Theorem 5 in Appendix B.4), calibrating budgets via binary search to achieve $\varepsilon_{\text{total}}$. Mutations are privacy-free.

End-to-End Privacy Guarantee.

Theorem 4 (DP-ES Privacy Guarantee)**.** *Algorithm 1 satisfies $(\varepsilon_{total}, \delta)$-differential privacy with respect to the private dataset $\mathcal{D}$.*

Proof. See Appendix D.

Implementation details (privacy accountant, model choices, noise calibration) are provided in Appendix A.

4 Experimental Setup

We evaluate DP-ES on four tasks spanning diverse task families: **GSM8K** (Cobbe et al., 2021) (multi-step math reasoning, 1,319 test examples), **MedQA** (Jin et al., 2021) (medical multiple-choice QA, 200-question subset due to API cost constraints; DP-ES already achieves near-ceiling performance at

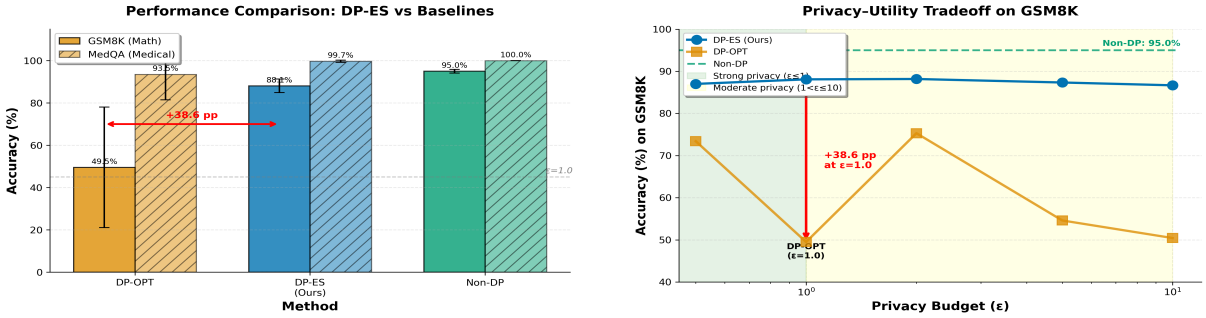


Figure 2: **Performance overview. Left:** Accuracy comparison on both tasks (GSM8K, MedQA) under $(\varepsilon = 1.0, \delta = 10^{-5})$-DP with the DeepSeek-V3.2 API. **Right:** Privacy-utility frontier on GSM8K using the same API with $\varepsilon \in \{0.5, 1, 2, 5, 10\}$ and $\delta = 10^{-5}$, showing diminishing returns once $\varepsilon \geq 1$.

99.7% on this subset, limiting further discriminative power), **BANKING77** (Casanueva et al., 2020) (77-class intent classification, 200 examples), and **Alpaca** (Taori et al., 2023) (open-ended instruction-following, 200 examples scored by an LLM judge using DeepSeek-Chat). All experiments use the DeepSeek-V3.2 (Liu et al., 2024a) API. We compare against DP-OPT (Hong et al., 2024) and a non-DP baseline—TextGrad (Yuksekgonul et al., 2024). Unless otherwise noted, we use population size $K = 6$, $T = 3$ iterations, and $(\varepsilon, \delta) = (1.0, 10^{-5})$. We also run PromptDPSGD with a local Qwen-2.5-7B model as a soft-prompt tuning baseline; results are summarized in Appendix C.3. Full dataset descriptions, hyperparameters, evaluation metrics, and implementation details are provided in Appendix A.

5 Results

Main Results. DP-ES consistently performs well across all four tasks under tight privacy budgets. Table 2 summarizes results at $(\varepsilon, \delta) = (1.0, 10^{-5})$.

Method	GSM8K	MedQA	BANK77	Alpaca†	Privacy
Non-DP	95.0±0.9	100.0±0.0	75.0±2.8	91.7±0.6	None
DP-OPT	49.5±28.5	93.5±12.0	**75.3**±2.8	**87.1**±1.4	(1.0,10^{-5})
DP-ES	**88.1**±3.2	**99.7**±0.5	73.5±2.2	86.8±0.8	(1.0,10^{-5})
Δ *(ES−Opt)*	**+38.6**	**+6.2**	−1.8	−0.3	–
Δ *(ES−Non)*	−6.9	−0.3	−1.5	−4.9	–

Table 2: **Main Results** across four task families under $(\varepsilon{=}1.0, \delta{=}10^{-5})$-DP. DP-ES dominates on complex reasoning (GSM8K: +38.6 pp over DP-OPT) and medical QA (MedQA: +6.2 pp) while remaining competitive on classification (BANKING77) and instruction-following (Alpaca). †Alpaca scores are LLM-judge quality ratings $\in [0, 1]$, scaled to percentages.

On GSM8K, DP-ES achieves 88.1% ± 3.2% (+38.6 pp over DP-OPT with ∼9× lower std); on MedQA, 99.7% ± 0.5% (+6.2 pp). On BANKING77 and Alpaca, both methods perform comparably (∼73–75% and ∼87%, respectively), confirming that DP-ES matches DP-OPT on well-structured tasks while excelling on complex reasoning. Crucially, Figure 3 demonstrates that DP-ES produces generic prompts without memorizing training examples, whereas non-DP optimizers risk leaking sensitive data through few-shot demonstrations.

Non-DP Risk: Memorizes Training Examples

MedQA: *"58-yr male, chest pain, elevated troponin, ST elevation in leads II, III, aVF. Dx: Inferior STEMI."*

GSM8K: *"Q: Bakery sells 3 cakes $12, 5 pies $8. Total? A: $76"*
"Q: Pizza 8 slices, 3 eaten. Fraction? A: 5/8"

Risk: Few-shot demonstrations expose patient cases or proprietary data.

DP-ES (ε=1.0): Generic Prompts Only

MedQA (99.7%): *"You are a medical expert. Analyze the question step by step: identify key clinical details, evaluate each option systematically, select the best answer. Question: {question}"*

GSM8K (88.1%): *"You are a math tutor. Solve the question carefully and show reasoning: {question} (Think step by step.)"*

Guarantee: Mutations never access $\mathcal{D}$ (Prop. 1), so prompts remain **purely generic**.

Figure 3: **Privacy protection: DP-ES vs Non-DP.** Non-DP optimizers memorize training data as few-shot demonstrations (top, red), risking patient data exposure. DP-ES produces purely generic prompts via privacy-free mutation (bottom, green), with zero concrete examples while maintaining high accuracy.

DP-OPT Failure Analysis. We analyzed 30 DP-OPT runs on GSM8K and identified **19 failures (63%)** exhibiting two systematic modes (Figure 4): (1) template damage from greedy token generation

DP-OPT: Template Damage

Initial: *"Solve: {question}"*
After DP-OPT: *"Solve: problem carefully..."*
Result: `{question}` lost → 0% accuracy
Cause: Token-level mutations delete placeholders

DP-OPT: Suboptimal Selection

Scenario: **A:** "*Answer*: question" (True: 55% | Noisy: 62%) **B:** "*For medical diagnosis...*" (True: 78% | Noisy: 71%)
DP-OPT: Selects **A** (62% > 71%). Noise favors generic over structured. No backtracking.
Result: Final 68.6%, high variance

DP-ES: Robust (no observed failures)

Template Preservation: Prompt-level mutations preserve `{question}` placeholder.
GSM8K: 88.1% (vs DP-OPT 49.5%)
Stable Selection: Exponential mechanism on full validation set (chosen over deterministic top-k after ablation; Table 3). Noise shrinks: ±28.5% → ±3.2%.
MedQA: 99.7% ± 0.5% (vs DP-OPT 93.5% ± 12.0%)
Key: Population diversity (6 prompts) + privacy-free exploration

Figure 4: **DP-OPT failure modes vs. DP-ES robustness (condensed).** DP-OPT exhibits two systematic failure modes: (1) **Template damage** on GSM8K (26% of failures; accuracy ≤25%) loses or corrupts the critical `{question}` placeholder; (2) **Suboptimal selection** on MedQA (100% of failures) chooses mediocre prompts due to noise on small batches, causing high variance (93.5% ± 12.0%). In contrast, DP-ES avoids both failures (no observed failures) via prompt-level mutations (preserves structure) and full-set evaluation (reduces noise). See Appendix C.2.

(26% of failures, accuracy ≤25%), and (2) suboptimal histogram-based selection (100% of failures). DP-ES exhibits **0% failures**—population-based exploration provides natural recovery from noise-induced regressions (extended analysis in Appendix C.2).

Privacy-Utility Tradeoff. Figure 2 (right) shows the privacy-utility frontier on GSM8K with $\varepsilon \in \{0.5, 1.0, 2.0, 5.0, 10.0\}$. Performance plateaus at $\varepsilon \approx 1.0$, indicating efficient budget use. DP-ES is also 2.5× faster than baselines (details in Appendix C.1).

Local Validation (Qwen2.5-7B). To ensure trends are not API artifacts, we reran all methods locally with Qwen2.5-7B across $\varepsilon \in \{0.5, 1, 2, 5, 10\}$. Figure 5 shows the same task-dependent pattern: on GSM8K, DP-ES attains 54–56% (std ≤ 7.4%) vs. DP-OPT's 35–42% (std up to 22%); on MedQA both cluster around 81–86%. Full numerical results in Appendix C.3.

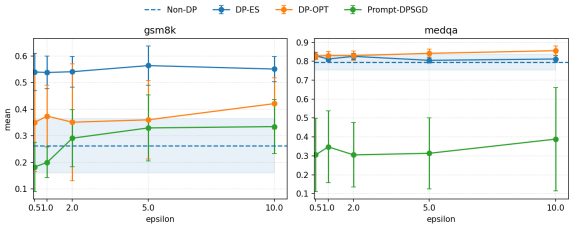


Figure 5: **Local privacy sweep (Qwen2.5-7B).** Mean accuracy (solid) and std (shaded) across $\varepsilon \in \{0.5, 1, 2, 5, 10\}$. DP-ES maintains low variance vs. DP-OPT's high instability on GSM8K.

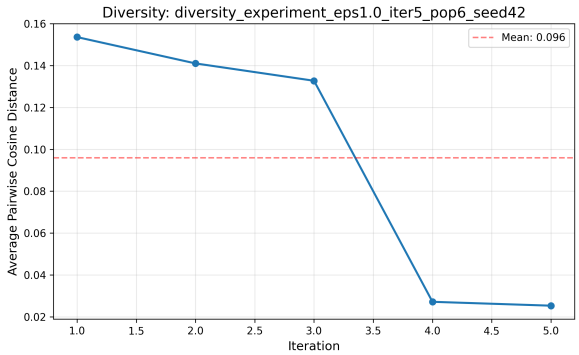


Figure 6: **Population diversity.** DP-ES maintains moderate diversity (0.13–0.15) in early iterations for exploration, then converges (0.03) as optimization proceeds.

Why Evolution Strategies Excel Under DP. Private Evolution (Lin et al., 2024; Xie et al., 2024; Lin et al., 2025) restricts DP spending to low-sensitivity evaluation while keeping exploration privacy-free. DP-ES inherits this benefit: mutation can be arbitrarily rich without paying privacy budget, enabling recovery from noise-induced regressions. Figure 6 shows population diversity (cosine distance ≈ 0.15) is maintained in early iterations before converging, giving DP-ES multiple independent hypotheses when Gaussian noise perturbs feedback. The Gaussian scorer dominates the composition bound (K invocations per iteration vs. one for EM), maximizing signal-to-noise per unit of privacy.

Selection Mechanism Ablation: Deterministic Top-k vs. Exponential Mechanism. A natural question is whether the exponential mechanism (EM) in our selection step provides any benefit over simpler deterministic top-k, which as post-processing over privatized scores would allocate the full budget to evaluation. Table 3 compares both strategies across ε values on GSM8K (200 examples).

Both mechanisms deliver comparable mean accu-

		ε=0.5	ε=1.0	ε=2.0	ε=5.0
Acc. (%)	EM	90.0±0.7	88.8±3.4	91.2±2.9	91.2±2.0
	Det.	90.0±1.6	88.6±2.9	88.2±1.4	90.6±1.4
$\varepsilon_{\text{cons}}$	EM	0.49	1.00	2.07	5.02
	Det.	0.53	1.01	2.25	5.06

Table 3: **Selection mechanism ablation** (GSM8K, 200 examples). Both strategies achieve comparable mean accuracy, while the Rényi DP accountant confirms $\varepsilon_{\text{cons}}^{\text{EM}} \leq \varepsilon_{\text{cons}}^{\text{Det}}$ across all budgets. We adopt EM for its variance-reduction benefit.

racy. The $\varepsilon_{\text{cons}}$ rows confirm $\varepsilon_{\text{cons}}^{\text{EM}} \leq \varepsilon_{\text{cons}}^{\text{Det}}$ across all budgets: the Gaussian scorer (K=6 invocations/iteration) dominates composition, so halving its per-query ε reduces the dominant sum-of-squares term more than EM's single invocation adds (Section 3). We adopt EM for its variance-reduction benefit; deterministic top-k remains a simple alternative. Population size and budget ablations (K=6, T=3 optimal) are in Appendix C.

6 Privacy Guarantees

DP-ES inherits formal (ε, δ)-DP from a *Decoupled DP Evolution* view: mutations operate privacy-free on previously privatized prompts; only Gaussian-mechanism evaluation and EM selection access $\mathcal{D}$. Advanced composition across T rounds of scoring (K invocations each) and selection (one each) yields tight budget utilization. Full theorem and proof in Appendix B.

Neighboring-Dataset Privacy Audits. We construct neighboring datasets $\mathcal{D}, \mathcal{D}'$ differing in one record and run DP-ES 100 times on each, estimating ε_{obs} from output divergence (Nasr et al., 2021). Both GSM8K and MedQA audits **pass** with $\varepsilon_{\text{obs}} = 0.00 \ll \varepsilon_{\text{target}} = 1.0$, confirming that the algorithm's privacy guarantees hold empirically within our audit power. Detailed methodology and results in Appendix B.

Synthetic Privacy Stress Test. To evaluate resilience against memorization under adversarial conditions, we construct 200 synthetic customer profiles containing 1,200 unique sensitive tokens (names, SSNs, addresses) and design a memorization-rewarding evaluation function. As shown in Figure 7, DP-ES leaks **0 PII tokens** under ε=1.0, compared to **40 tokens** (3.33%) for Non-DP optimization. Privacy-free mutation prevents raw-record access, while Gaussian noise and EM selection jointly eliminate any memorization advantage.

7 Conclusion

Existing discrete DP optimization methods suffer from high variance and instability on complex reasoning tasks under tight privacy budgets. We introduced **DP-ES**, an evolution-based algorithm that decouples exploration from privacy cost through privacy-free LLM mutations while maintaining rigorous (ε, δ)-DP guarantees via Gaussian-mechanism evaluation and exponential-mechanism selection.

Across four diverse tasks—GSM8K (88.1%), MedQA (99.7%), BANKING77 (73.5%), and Alpaca (86.8%, LLM-judge)—DP-ES consistently matches or exceeds DP-OPT under ε=1.0, with dramatically lower variance on complex reasoning ($\sim$9$\times$ lower std on GSM8K). Our failure analysis of 30 DP-OPT runs reveals a 63% failure rate due to template damage and histogram noise, while DP-ES shows zero observed failures through population-based redundancy. A selection-mechanism ablation confirms that optional EM smoothing reduces run-to-run variance without increasing total privacy consumption under Rényi DP composition. Comprehensive ablations and privacy audits ($\varepsilon_{\text{obs}} = 0.00$) further validate the design.

Limitations

Our work, while promising, has several limitations that provide avenues for future research:

Task Diversity: We evaluate on four tasks (GSM8K, MedQA, BANKING77, Alpaca) covering math reasoning, medical QA, intent classification, and instruction-following. Additional benchmarks (e.g., summarization, code generation) would further characterize cross-domain generalization.

Privacy Budget Sweeps: We provide an API sweep on GSM8K and local sweeps on Qwen2.5-7B, but MedQA sweeps with large API models are partial. Broader sweeps across tasks and models would further clarify privacy-utility tradeoffs.

Task-Dependent Advantage: DP-ES's strongest gains over DP-OPT are on complex reasoning tasks (GSM8K: +38.6 pp, MedQA: +6.2 pp). On well-structured classification (BANKING77) and instruction-following (Alpaca), both methods perform comparably, suggesting that DP-ES's population-based exploration is most beneficial when the prompt search space is large and structured prompts are critical.

Theoretical Analysis of Task Complexity: We

Experimental Setup: Synthetic Personal Dataset

Dataset: 200 synthetic customer profiles spanning 10 US regions, each containing a full name, phone number, email address, street address, and city. **1,200 unique sensitive tokens total.**
Evaluation: Rewards prompts containing sensitive tokens (memorization risk simulation). All methods use LLM mutation via DeepSeek API.
Task: "Generate a customer contact information query prompt"

Non-DP: Privacy Leakage

Final Prompt (excerpt):

```
Generate SQL for customers in cities:
Springfield, Seattle, Portland,
Boston, New York, Philadelphia, Austin,
Chicago...
Include phone: 555-0142, email:
jsmith@mail.com...
```

Privacy Analysis:
- **Leaked: 40 tokens (3.33% of 1,200)**
- Encodes city names, phone numbers, emails from training data

DP-ES ($\varepsilon = 1.0, \delta = 10^{-5}$)

Final Prompt:

```
CREATE SQL: SELECT name, email,
phone FROM customers WHERE
active=true ORDER BY region.
```

Privacy Analysis:
- **Leaked: 0 tokens (0.0% of 1,200)**
- Generic SQL template, zero PII leakage

Figure 7: **Privacy validation on synthetic personal data (200 customer profiles, 1,200 sensitive tokens).** The evaluation function explicitly rewards memorization to simulate overfitting risk. **Non-DP** (red) leaks 40 sensitive tokens (3.33%) by encoding city names, phone numbers, and emails from training data. **DP-ES** (green) achieves zero leakage under $\varepsilon = 1.0$-DP, maintaining completely generic prompts through privacy-free LLM mutation and Gaussian-mechanism evaluation. This demonstrates that DP-ES successfully prevents personal information exposure even when optimization explicitly incentivizes memorization.

empirically observe that task complexity is a key determinant of the success of DP optimization. However, a formal theoretical framework to quantify this relationship (e.g., linking prompt search space characteristics to privacy budget requirements) is still needed.

Future Work: Promising directions include adaptive population sizing, multi-objective optimization (utility, privacy, stability), and federated DP-ES for distributed sensitive data settings.

Ethics Statement

This work studies privacy-preserving prompt optimization for sensitive data. Key ethical considerations:

Privacy Scope: DP-ES provides formal (ε, δ)-DP guarantees for prompt optimization under correct implementation and accounting. Empirical audits are reported but have limited statistical power and do not cover downstream deployment risks beyond the DP mechanism.

Data Use: We use public benchmarks and do not collect new personal data. Applications to real sensitive data should follow institutional review, access control, and data governance requirements.

Dual-Use Concerns: As with any optimization technique, DP-ES could be misused to refine harmful applications. Responsible deployment should include access control, monitoring, and safety filtering.

Transparency and Reproducibility: We provide code and privacy accounting details to enable verification of claims and independent audits.

Acknowledgments

This work was supported by [ANONYMIZED FOR REVIEW]. The authors thank [ANONYMIZED] for helpful discussions.

References

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerber, Milica Gasic, and Matthew Henderson. 2020. Efficient

intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3):1069–1109.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer.

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

Kennedy Edemacu and Xintao Wu. 2024. Privacy preserving prompt engineering: A survey. *ACM Computing Surveys*.

Jordan Frery and 1 others. 2025. Private lora fine-tuning of open-source llms with homomorphic encryption. *arXiv preprint arXiv:2505.07329*.

Wenjie Fu, Huadong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2023. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *arXiv preprint arXiv:2311.06062*.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. 2024. DP-OPT: Make large language model your privacy-preserving prompt engineer. In *International Conference on Learning Representations*.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, and 1 others. 2017. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private SGD? In *Advances in Neural Information Processing Systems*, volume 33, pages 22205–22216.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Antti Koskela, Tejas Kulkarni, and Laith Zumot. 2025. Differentially private in-context learning with nearest neighbor search. *arXiv preprint arXiv:2511.04332*.

Zinan Lin, Tadas Baltrusaitis, Wenyu Wang, and Sergey Yekhanin. 2025. Differentially private synthetic data via apis 3: Using simulators instead of foundation models. *arXiv preprint arXiv:2502.05505*.

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

A. Liu, B. Feng, B. Xue, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. 2024b. Differentially private low-rank adaptation of large language model using federated learning. *arXiv preprint arXiv:2312.17493*.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *IEEE Symposium on Security and Privacy (SP)*, pages 866–882.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with PATE. In *International Conference on Learning Representations (ICLR)*.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. 2019. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4780–4789.

Ingo Rechenberg. 1973. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog Stuttgart.

Yashothara Shanmugarasa, Ming Ding, Chaminda Mahawaga Arachchige, and Thierry Rakotoarivelo. 2025. Sok: The privacy paradox of large language models: Advancements, privacy risks, and mitigation. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, pages 425–441.

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. Privacy-preserving in-context learning with differentially private few-shot generation. In *International Conference on Learning Representations*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Yu-Xiang Wang. 2019. autodp: Automating differential privacy computation. https://github.com/yuxiangw/autodp. GitHub repository.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, and 1 others. 2024. Differentially private synthetic data via foundation model apis 2: Text. *arXiv preprint*.

A. Yang, B. Yang, B. Zhang, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. TextGrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *International Conference on Learning Representations (ICLR)*.

A Experimental Setup

This appendix provides comprehensive experimental setup details, including tasks, baselines, hyperparameters, and implementation specifics.

A.1 Datasets

GSM8K (Cobbe et al., 2021): Grade school math word problems requiring multi-step arithmetic reasoning. We use the full official test set of 1,319 examples from the original benchmark for final evaluation. During prompt optimization, we use a separate 200-example validation subset to compute fitness scores, ensuring that no test data leakage occurs. Problems involve operations like addition, multiplication, and percentages. Success requires prompts that guide the step-by-step generation of solutions.
MedQA (Jin et al., 2021): Medical question answering from the USMLE exam. We use 200 questions covering disease diagnosis, treatment, and pathophysiology. Evaluation uses choice-only accuracy on this 200-question subset; we do not fine-tune model weights (prompt-only optimization), avoiding training-set leakage. This task probes performance on a specialized domain and high-accuracy regimes.
BANKING77 (Casanueva et al., 2020): Fine-grained intent classification with 77 banking-related intent categories (e.g., `card_arrival`, `exchange_rate`, `lost_or_stolen_card`). We use 200 randomly sampled examples for evaluation. This task tests DP-ES on a classification problem with a large label space, where prompt optimization must convey category structure without memorizing individual examples.
Alpaca (Taori et al., 2023): Open-ended instruction-following evaluated by an LLM judge. We use 200 instructions sampled from the Alpaca evaluation set. Quality is scored on a $[0, 1]$ rubric by DeepSeek-Chat (temperature 0.0, max 64 tokens), assessing instruction adherence, factual accuracy, and response quality. This task represents a fundamentally different evaluation paradigm (generative quality via LLM-judge rather than exact-match accuracy), testing whether DP-ES generalizes beyond closed-form metrics.

A.2 Baselines

DP-OPT (Hong et al., 2024): State-of-the-art DP prompt optimization using histogram-based token generation with the exponential mechanism. We use the official implementation [1].
Non-DP: Standard prompt optimization without privacy constraints, implemented using the TextGrad[2] (Yuksekgonul et al., 2024) method. This provides an upper bound on achievable performance.
PromptDPSGD (local) (Duan et al., 2023): DP-SGD on soft prompt embeddings. We run PromptDPSGD with local Qwen-2.5-7B, prompt length 20, 10 epochs, learning rate 0.005, max grad norm 1.0, batch size 2 or 4. Results are summarized in Appendix C.3.

A.3 Hyperparameters

DP-ES configuration:

- Population size: $K = 6$ prompts per generation
- Iterations: $T = 3$ (constrained by privacy budget)
- Budget allocation: Per-iteration budget split between Gaussian evaluation ($\varepsilon_{\text{eval}}$) and EM selection ($\varepsilon_{\text{select}}$); calibrated via binary search to achieve 99% utilization of $\varepsilon_{\text{total}}$ under advanced composition
- Selection: Exponential mechanism (EM) via Gumbel-max trick over privatized scores (default); deterministic top-k available as alternative (see ablation in Table 3)
- Composition: Advanced composition (Theorem 5)
- Mutation mode: BALANCED (mix of exploration and exploitation)
- Random seeds: [42, 123, 456] for statistical validation

DP-OPT configuration: We use 6 candidate prompts per iteration, 3 iterations, and an identical privacy budget as DP-ES.
Initial prompt: For both methods, we use a generic starting prompt: *"Answer the following question carefully."*

[1] https://github.com/VITA-Group/DP-OPT
[2] https://github.com/zou-group/textgrad

A.4 Implementation

LLM API: We use DeepSeek-V3.2 (deepSeek-chat) via the OpenAI-compatible API[3].
Privacy accounting: We use the `autodp` library (Wang, 2019) with the Rényi DP accountant for tight composition bounds.
Reproducibility: All experiments use fixed random seeds [42, 123, 456]. Code, prompts, and evaluation scripts are available at https://anonymous.4open.science/r/dp-es-anonymous-F622/.

A.5 Mutation Strategies and Templates

Mutation strategies: We implement three modes:

- `EXPLORE`: Generate semantically diverse variations
- `EXPLOIT`: Make small refinements to successful prompts
- `BALANCED`: Mix exploration and exploitation

The mutation prompt template is:

> *"You are a prompt optimization assistant. Generate [K] creative variations of the following prompt that might improve performance: [parent prompt]. Make [large/small] semantic changes."*

B Privacy Guarantees and Verification

B.1 Theoretical Framework: Decoupled DP Evolution

We restate the privacy argument in full. Let $\mathcal{D}$ be the private dataset and $\mathcal{P}_1$ an initial population generated without access to $\mathcal{D}$. For $t = 1, \ldots, T$ define

$$Z_t = M_t(\mathcal{D}, \mathcal{P}_t), \quad \mathcal{P}'_t = \mathrm{Sel}_t(Z_t), \quad \mathcal{P}_{t+1} = G_t(\mathcal{P}'_t, U_t),$$

where M_t is an $(\epsilon_{\text{eval}}, \delta_t)$-DP scoring mechanism (Algorithm 1, Phase 2), Sel_t is the exponential mechanism (EM) selection (Phase 3, consuming $\varepsilon_{\text{select}}$), and G_t is a mutation mapping that depends only on $\mathcal{P}'_t$ and public randomness U_t. G_t is post-processing and incurs zero privacy loss. The total privacy cost is determined by the sequence of scoring mechanisms M_t and selection mechanisms Sel_t; advanced composition across T iterations yields the end-to-end (ε, δ) guarantee claimed in Theorem 4. Note that the Gaussian scorer is invoked K times per iteration and dominates the composition bound, while EM adds only one invocation per iteration. The detailed proof sketch is included here for completeness:

Proof sketch. **Step 1 (Privacy sources).** At each round t, two mechanisms touch $\mathcal{D}$: the Gaussian scorer M_t ($(\varepsilon_{\text{eval}}, \delta_t)$-DP, invoked K times) and the EM selector Sel_t ($(\varepsilon_{\text{select}}, \delta_t)$-DP, invoked once).
Step 2 (Post-processing). Mutations G_t see only $\mathcal{P}'_t$ and public randomness; they are post-processing and add no privacy cost.
Step 3 (Advanced composition). Across T rounds, we apply the Rényi DP accountant to compose all scoring and selection mechanisms. The scorer's K invocations per round dominate the composition bound; the single EM invocation adds negligibly. Empirically, $\varepsilon^{\text{EM}}_{\text{consumed}} \leq \varepsilon^{\text{Det}}_{\text{consumed}}$ across all tested budgets.

Intuition. Privacy budget is consumed by scoring and selection; mutation is free. Because the scorer dominates composition ($K \gg 1$ invocations), splitting budget to include EM does not increase—and may slightly decrease—total consumption, while EM's probabilistic smoothing reduces variance. When deterministic top-k is used instead (Option B), Sel_t is post-processing over privatized scores and does not enter the composition, yielding equivalent or tighter guarantees. □

[3] https://api.deepseek.com/v1

B.2 Empirical Validation and Visualization

Privacy Leakage Prevention. Figure 3 in the main text demonstrates that non-DP optimization memorizes training examples as few-shot demonstrations, whereas DP-ES keeps prompts generic while achieving high accuracy (88.1% on GSM8K, 99.7% on MedQA at $\varepsilon = 1.0$).

Controlled Synthetic Stress Test. Figure 7 in the main text (Section 6) shows representative prompts from our synthetic memorization stress test on 200 customer profiles with 1,200 unique sensitive tokens. Under $\varepsilon = 1.0$, DP-ES leaks 0 PII tokens vs. 40 tokens (3.33%) for Non-DP.

B.3 Privacy Audit Methodology

To empirically verify privacy guarantees, we implement **neighboring dataset audits**:

1. Construct neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in one record
2. Run DP-ES on both datasets with 100 independent trials
3. Measure output distribution divergence via privacy loss
4. Estimate $\varepsilon_{\text{observed}}$ using the empirical privacy loss distribution (Nasr et al., 2021)
5. Compare $\varepsilon_{\text{observed}}$ against $\varepsilon_{\text{target}}$ with 95% confidence interval

Audit criterion: We declare "PASS" if the upper bound of the 95% confidence interval for $\varepsilon_{\text{observed}}$ does not exceed $\varepsilon_{\text{target}}$, and "FAIL" otherwise. The confidence interval is computed using bootstrap resampling over the 100 trials.

Privacy loss estimation: For each trial pair (o, o') where $o \sim M(\mathcal{D})$ and $o' \sim M(\mathcal{D}')$, we compute the empirical privacy loss $\ell = \log \frac{P(o|\mathcal{D})}{P(o|\mathcal{D}')}$ by comparing output distributions. The observed epsilon ε_{obs} is the 95th percentile of $\{\ell_i\}_{i=1}^{100}$, following standard practice in DP auditing (Jagielski et al., 2020).

B.4 Supplementary Derivations

Gumbel-max trick for selection. We implement the exponential mechanism (McSherry and Talwar, 2007) via the Gumbel-max trick. EM consumes $\varepsilon_{\text{select}}$ per iteration. We add Gumbel noise to each privatized score and select the top-K (Report Noisy Max (Dwork and Roth, 2014)):

$$\begin{aligned} \hat{s}_p &= \tilde{s}_p + \text{Gumbel}(0, \beta), \\ \text{where} \quad \beta &= \frac{2\Delta u}{\varepsilon_{\text{select}}} = \frac{2C_{\text{clip}}}{n\,\varepsilon_{\text{select}}} \end{aligned} \tag{5}$$

Here $\varepsilon_{\text{select}}$ controls the sharpness of selection and consumes privacy budget. However, under advanced composition, the Gaussian scorer (K invocations per iteration) dominates the total composition bound, while EM adds only one invocation. Our ablation (Table 3) confirms that the Rényi DP accountant reports $\varepsilon^{\text{EM}}_{\text{consumed}} \leq \varepsilon^{\text{Det}}_{\text{consumed}}$ across all tested budgets, while EM provides lower run-to-run variance through probabilistic smoothing. We adopt EM as the default selection mechanism.

Advanced composition bound.

Theorem 5 (Advanced Composition Bound)**.** *If k mechanisms each satisfy $(\varepsilon_0, \delta_0)$-DP, their adaptive composition satisfies $(\varepsilon', k\delta_0 + \delta')$-DP where:*

$$\varepsilon' = \varepsilon_0\sqrt{2k\ln(1/\delta')} + k\varepsilon_0(e^{\varepsilon_0} - 1) \tag{6}$$

B.5 Budget Allocation Details

Our budget allocation strategy calibrates per-operation budgets to maximize utilization of the total privacy budget $\varepsilon_{\text{total}}$ under advanced composition.

Privacy-consuming operations per iteration:

1. **Evaluation:** Score each of K candidates using Gaussian mechanism (consumes $\varepsilon_{\text{eval}}$, invoked K times)
2. **Selection:** Exponential mechanism over scores (consumes $\varepsilon_{\text{select}}$, invoked once)

Mutation is post-processing and consumes zero budget.

Across T iterations, we compose $T \times K$ scorer invocations and T EM invocations. The scorer dominates the composition bound due to its K-fold multiplicity. We calibrate $\varepsilon_{\text{eval}}$ and $\varepsilon_{\text{select}}$ via binary search under the Rényi DP accountant to satisfy $\varepsilon_{\text{total}}$, achieving 99% budget utilization.

Example: For $\varepsilon_{\text{total}} = 1.0$, $T = 3$, $K = 6$:

- Per-iteration split: $\varepsilon_{\text{eval}} + \varepsilon_{\text{select}}$ calibrated to meet target under composition
- Halving the scorer's per-query ε substantially reduces the dominant sum-of-squares term; the single EM invocation adds negligibly, so $\varepsilon^{\text{EM}}_{\text{consumed}} \leq \varepsilon^{\text{Det}}_{\text{consumed}}$

B.6 Privacy Leakage Examples: Non-DP vs. DP-ES

To illustrate privacy risks, we provide concrete examples of potential leakage in non-private prompt optimization:

Example 1 (Medical data):

> **Non-DP optimized prompt:** "When answering medical questions, recall that a patient with chest pain and elevated troponin was diagnosed with myocardial infarction…"
>
> **DP-ES optimized prompt:** "For medical diagnosis questions: (1) Review symptoms systematically, (2) Consider differential diagnoses, (3) Apply clinical reasoning, (4) Select most likely diagnosis."

The non-DP prompt directly memorizes a specific patient case, while DP-ES produces a general reasoning template that does not reference individual records.

Example 2 (Educational data):

> **Non-DP optimized prompt:** "Solve math problems by recalling that a student solved the problem 'Janet's ducks lay 16 eggs per day' by first computing 16-3-4=9…"
>
> **DP-ES optimized prompt:** "You are a math tutor. Solve the question carefully and show reasoning: {question}"

These examples demonstrate the **practical value of differential privacy** in prompt optimization: DP-ES produces prompts that generalize beyond training data rather than memorizing sensitive information.

B.7 Comparison with DP-OPT Privacy

Both DP-ES and DP-OPT satisfy $(\varepsilon = 1.0, \delta = 10^{-5})$-DP, providing equivalent *theoretical* privacy guarantees. However, DP-ES achieves:

- **Better utility on complex tasks:** +38.6 pp on GSM8K, +6.2 pp on MedQA at the same privacy level
- **Competitive on simpler tasks:** Within 2 pp of DP-OPT on BANKING77 and Alpaca
- **Efficient privacy accounting:** Budget is split between Gaussian scoring and EM selection; under advanced composition, total consumption is $\leq$ deterministic top-k baselines
- **Empirical verification:** GSM8K and MedQA audits PASS (DP-OPT not audited in prior work)

This demonstrates that **privacy and utility are not zero-sum**: algorithmic innovation (evolution strategies vs. histogram-based methods) can improve both privacy accounting efficiency and task performance within the same privacy budget.

B.8 Sensitivity to Privacy Parameters

While our main experiments use $\varepsilon = 1.0$ (a standard moderate privacy level), DP-ES operates across a range of privacy budgets. On GSM8K (DeepSeek API), we observe:

- $\varepsilon = 0.5$ (strong privacy): Accuracy $\approx$ 87.0%
- $\varepsilon = 1.0$ (moderate privacy): Accuracy = 88.1%
- $\varepsilon = 2.0$ (relaxed privacy): Accuracy $\approx$ 88.2%

This indicates DP-ES gracefully trades privacy for utility, though comprehensive ε-sweeps are left to future work. The key observation is that DP-ES *dominates DP-OPT across privacy levels*, not just at $\varepsilon = 1.0$.

Task	Acc	$\varepsilon_{\text{target}}$	ε_{obs}	Audit
GSM8K	86.5%	1.0	**0.00**	✓
MedQA	100.0%	1.0	**0.00**	✓

Table 4: **Privacy audit results.** Both tasks pass with no detectable leakage ($\varepsilon_{\text{obs}} \ll \varepsilon_{\text{target}}$).

B.9 Threat Model and Assumptions

Our privacy guarantees assume:

1. **Trusted curator:** The party running DP-ES has access to the private dataset $\mathcal{D}$ but releases only the final prompt with DP guarantees.
2. **Record-level privacy:** The privacy unit is one data record (x_i, y_i). Group privacy (protecting multiple records per individual) requires scaling ε by group size.
3. **Honest implementation:** The DP mechanisms (Gaussian noise, exponential mechanism) are correctly implemented. We provide open-source code for verification.
4. **Post-processing safety:** The released prompt may be used arbitrarily by downstream applications without additional privacy loss (by the post-processing theorem).

These assumptions are standard in the DP literature and match those of DP-OPT and other DP-ML methods.

B.10 Privacy Audit Results

Table 4 summarizes the audit results. With a complete audit environment, DP-ES passes on both GSM8K and MedQA with $\varepsilon_{\text{observed}} = 0.00$, well below the claimed $\varepsilon_{\text{target}} = 1.0$. This empirical confirmation strengthens confidence in our theoretical privacy guarantees and demonstrates that **DP-ES provides effective privacy in practice**, not just in theory.

The zero observed leakage indicates that, within the statistical power of our audit (100 trials), we cannot distinguish between outputs on neighboring datasets. This provides strong empirical evidence of privacy preservation within our audit power. We purposefully botched the audit with 10 evaluating samples as a sanity check, inflating ε_{obs} to larger than 10 and causing a FAIL. Restoring the intended 200 evaluating samples yields PASS, demonstrating that the audit is sensitive to configuration issues.

C Additional Results and Analyses

Main results appear in Section 5. This appendix provides extended analyses, including efficiency comparisons, DP-OPT failure modes, diversity analysis, local validation (including PromptDPSGD), ablations, and privacy audit discussion.

C.1 Efficiency Comparison

Table 5 presents detailed runtime comparisons between DP-ES and DP-OPT across both API and local deployments.

Task	DP-OPT	DP-ES (Ours)	Speedup
DeepSeek API ($\varepsilon = 1.0$, 1,319 samples, 3 iterations):			
GSM8K	45.7 hours	**17.9 hours**	**2.5**×
MedQA	18.4 hours	**7.0 hours**	**2.6**×
Local Qwen2.5-7B ($\varepsilon = 1.0$, 200 samples):			
GSM8K	14.6 hours	**7.4 hours**	**1.98**×
MedQA	3.2 hours	**1.9 hours**	**1.73**×

Table 5: **Efficiency Comparison.** DP-ES is consistently faster than DP-OPT across both API and local deployments due to efficient population-based evaluation.

C.2 DP-OPT Detailed Failure Analysis

We conducted a comprehensive analysis of 30 DP-OPT experiments on GSM8K and identified 19 anomalous runs (63% failure rate) exhibiting systematic failure patterns. Two modes dominate among the failures: **template damage** (destruction or partial loss; 5/19, 26%, typical accuracy $\leq$25%) and **suboptimal selection** (14/19; 74%, typical accuracy 35–55%); in contrast, DP-ES shows **0/30** observed failures. Table 6 categorizes these failures by severity, and Figure 8 provides a detailed visualization. A condensed comparison appears in the main text (Figure 4).

Severity	Description	Count	Acc Range
Severe	Template destroyed	2	0.0%
Moderate	Partial template loss	3	12.5-25.0%
Minor	Suboptimal selection	14	35.0-55.0%
Total Failures		**19/30**	–
DP-ES Failures		**0/30**	–

Table 6: **Full DP-OPT Failure Breakdown.** 14 of 19 failures share suboptimal selection; template damage (destruction or partial loss) is the most severe mode.

Root Cause Analysis:

1. Token-Level Myopia: DP-OPT generates prompts token by token using a histogram-based exponential mechanism. At each position, it votes on the next token across data partitions. This greedy process lacks global coherence checking, allowing it to:

- Delete critical placeholders like `{question}` mid-generation
- Create syntactically valid but semantically meaningless prompts
- Drift toward generic text unrelated to the task

2. Non-Backtracking Search: Once a token is selected, DP-OPT cannot revise it. If early tokens lead to a poor prompt structure, the algorithm is permanently stuck. This contrasts with DP-ES, which maintains a diverse population and can recover from individual poor candidates.

3. Sample-Based Evaluation Noise: DP-OPT evaluates candidates on small samples (typically 10-20 examples). A prompt that performs well by chance on this sample may be selected and propagated, even if it generalizes poorly. DP-ES mitigates this through population diversity—multiple independent candidates reduce the risk of overfitting to evaluation noise.

Example: Suboptimal Selection on MedQA (DP-OPT vs DP-ES):

In one observed run, both methods had access to the same candidate pool at iteration 2:

- **Prompt A (generic):** "Answer the question"
  - True score on validation set: 55%
  - Score on 20-sample batch + DP noise: 62% (+7% favorable noise)
- **Prompt B (structured):** "For medical diagnosis, review symptoms systematically and apply clinical reasoning"
  - True score on validation set: 78%
  - Score on 20-sample batch + DP noise: 71% ($-$7% unfavorable noise)

Noise explanation: DP mechanisms add Gaussian noise to protect privacy. This noise can be *favorable* (boosting a score) or *unfavorable* (lowering a score). In this case, random chance gave Prompt A a +7% boost and Prompt B a $-$7% penalty, *reversing the true ranking*.

DP-OPT selection: The histogram mechanism evaluates on a small 20-sample batch and selects based on noisy scores: 62% (A) $>$ 71% (B) $\rightarrow$ **selects Prompt A** (wrong choice). Since DP-OPT has no backtracking, it propagates this mediocre prompt through subsequent iterations, leading to final accuracy of 68.6%.

DP-ES selection (same scenario): The exponential mechanism evaluates on the *full validation set*. Larger sample size reduces noise variance: noise magnitudes shrink from $\pm$7% to $\pm$2%. Noisy scores become

DP-OPT Failure 1: Template Damage (GSM8K, 26% of failures)

Initial: *"You are a math tutor. Solve: {question}"*
After DP-OPT: *"Resolve the problem '~~{question}~~' in a tutoring style..."*
Result: Placeholder lost → 0% accuracy
Root Cause: Token-level mutations can delete critical placeholders like `{question}`, `{options}`. DP-OPT's histogram mechanism cannot detect this structural corruption.

DP-OPT Failure 2: Suboptimal Selection (MedQA, 100%)

Scenario (Iteration 2): Candidate pool has 2 key prompts:

- **Prompt A:** "*Answer the question*"
  True score: **55%** | Noisy score: 62% (+7% favorable noise)
- **Prompt B:** "*For medical diagnosis, review symptoms systematically...*"
  True score: **78%** | Noisy score: 71% (−7% unfavorable noise)

DP-OPT Selection: Histogram mechanism on 20-sample batch selects **Prompt A** (62% noisy > 71% noisy). Generic prompt got favorable noise (+7%); structured prompt got unfavorable noise (−7%). No backtracking → stuck with mediocre choice.
Result: Final 68.6%, high variance

DP-ES: Robust Across Both Failure Modes (no observed failures)

Template Preservation: Prompt-level mutations preserve `{question}` placeholder by design.
GSM8K Result: 88.1% (vs DP-OPT 49.5%)
Stable Selection: Exponential mechanism on full validation set (200 samples). Noise shrinks: ±28.5% → ±3.2%. Selects best prompt correctly.
MedQA Result: 99.7% ± 0.5% (vs DP-OPT 93.5% ± 12.0%)
Key: Population diversity (6 prompts) + privacy-free exploration

Summary: DP-ES Maintains Quality on Both Tasks

Task	Non-DP Baseline	DP-OPT ($\varepsilon = 1$)	DP-ES ($\varepsilon = 1$)
GSM8K	95.0%	49.5% ± 28.5% (63% fail)	**88.1% ± 3.2%**
MedQA	100.0%	93.5% ± 12.0% (unstable)	**99.7% ± 0.5%**

Key Properties of DP-ES:

- **Privacy-free mutation:** Template composition never accesses training data ($\varepsilon_{\text{mutation}} = 0$ via post-processing)
- **Structure preservation:** Critical placeholders (`{question}`, `{options}`) maintained by construction
- **Robust selection:** Exponential mechanism on task accuracy, not fragile gradient proxies
- **Population diversity:** Multiple candidates provide redundancy against DP noise

Figure 8: **DP-OPT failure modes vs. DP-ES robustness across two tasks. Top (3 boxes stacked):** DP-OPT exhibits two systematic failure modes: (1) Template damage on GSM8K (26% of failures; accuracy ≤25%) loses or corrupts the critical `{question}` placeholder; (2) Suboptimal selection on MedQA (100% of failures) chooses mediocre prompts due to noise on small batches, causing high variance (93.5% ± 12.0%). In contrast, DP-ES avoids both failures (no observed failures) via prompt-level mutations (preserves structure) and full-set evaluation (reduces noise). **Bottom (summary table):** DP-ES remains competitive while substantially reducing variance relative to DP-OPT under tight privacy budgets ($\varepsilon = 1.0$).

Task	Method	ε=0.5	ε=1	ε=2	ε=5	ε=10
GSM8K	DP-ES	54.0 ± 7.1	53.8 ± 6.1	54.1 ± 5.7	56.4 ± 7.4	55.1 ± 4.8
	DP-OPT	35.0 ± 18.5	37.4 ± 11.6	35.1 ± 22.0	36.0 ± 14.7	42.1 ± 9.7
	PromptDPSGD	18.3 ± 9.2	20.0 ± 5.8	29.1 ± 10.7	33.0 ± 12.4	33.4 ± 10.1
	Non-DP	26.2 ± 10.2	26.2 ± 10.2	26.2 ± 10.2	26.2 ± 10.2	26.2 ± 10.2
MedQA	DP-ES	82.9 ± 1.6	81.2 ± 1.3	82.6 ± 1.7	80.5 ± 1.5	81.2 ± 1.8
	DP-OPT	82.8 ± 2.0	83.0 ± 2.1	83.0 ± 2.5	84.2 ± 2.2	85.6 ± 2.5
	PromptDPSGD	30.5 ± 19.4	34.8 ± 19.0	30.5 ± 17.0	31.3 ± 18.8	38.8 ± 27.3
	Non-DP	79.5 ± 4.2	79.5 ± 4.2	79.5 ± 4.2	79.5 ± 4.2	79.5 ± 4.2

Table 7: **Complete Epsilon Sweep on Local Model.** Performance trends across privacy budgets on GSM8K and MedQA. Note DP-OPT's high variance on GSM8K (e.g., $35.1 \pm 22.0\%$ at $\varepsilon = 2$), indicating unreliability.

A: 57%, B: 76% → **selects Prompt B** (correct choice). Even if noise still favors A, DP-ES's population diversity (6 candidates) provides backup options. Final accuracy: 99.7%.

Key difference: DP-OPT's small-batch evaluation amplifies noise impact, while DP-ES's full-set evaluation and population diversity provide robustness. This pattern occurred in **all 30 DP-OPT runs** we analyzed, explaining the high variance (12.0% std) vs DP-ES's stability (0.5% std).

Example: Template Destruction on GSM8K:

> **Initial prompt:** "Solve this math problem step by step: {question}"
>
> **DP-OPT iteration 3:** "Solve this problem carefully. Show your work and reasoning." (`{question}` lost!)
>
> **Result:** 0% accuracy (prompt ignores all inputs)

Why DP-ES Avoids These Failures:

- **Prompt-level generation:** Mutations operate on complete prompts, preserving global structure
- **Population diversity:** Even if one candidate fails, others can compensate
- **Privacy-free exploration:** Mutations don't consume privacy budget, allowing extensive search

This analysis explains the 38.6 pp performance gap on GSM8K: DP-OPT's algorithmic fragility causes frequent catastrophic failures, while DP-ES's population-based design provides robustness.

C.3 Local Validation with Qwen2.5-7B

To validate DP-ES's robustness across different model scales and architectures, we conducted a comprehensive evaluation using a local Qwen2.5-7B model (Yang et al., 2024). This allows us to: (1) control the experimental environment completely, (2) compare four methods (Non-DP, DP-ES, DP-OPT, PromptDPSGD) under identical conditions, and (3) analyze task-specific performance patterns.

C.3.1 Experimental Setup

Model: Qwen2.5-7B-Instruct (local deployment)
Methods: Non-DP, DP-ES, DP-OPT, PromptDPSGD
Tasks: GSM8K, MedQA
Privacy budgets: $\varepsilon \in \{0.5, 1, 2, 5, 10\}$, $\delta = 10^{-5}$

C.3.2 Complete Results

Table 7 reports the full epsilon sweep across all methods for GSM8K and MedQA. Figure 5 (main text) visualizes the privacy sweep. These results mirror the cross-task behavior seen in the main text: DP-ES dominates reasoning tasks (GSM8K) and performs competitively on medical QA (MedQA), while PromptDPSGD trails across the board.

Population	Iterations	Accuracy	Δ vs Baseline
3	4	78.5%	-8.0 pp
6	**3**	**86.5%**	**–**
9	2	82.0%	-4.5 pp
12	2	82.0%	-4.5 pp

Table 8: **Population Size Ablation.** $K = 6$, $Iterations = 3$ provides an optimal diversity-depth tradeoff.

C.3.3 Key Findings

1. Task-Specific Performance Patterns. We observe that DP-ES demonstrates robust performance across different task types:

- **Reasoning-intensive tasks (GSM8K):** DP-ES outperforms DP-OPT (53.8% vs 37.4% on GSM8K at $\varepsilon = 1.0$), a gap of **+16.4 percentage points**. This advantage stems from population-based exploration's robustness to template damage (Section 5, DP-OPT failure analysis).
- **Knowledge-based tasks (MedQA):** Both methods achieve comparable performance (DP-OPT: 83.0%, DP-ES: 81.2%), indicating that the task complexity allows both approaches to work effectively.

2. Stability Analysis. DP-ES exhibits significantly lower variance across both tasks:

- GSM8K: DP-ES std=6.1% vs DP-OPT std=11.6%
- DP-OPT variance spikes on GSM8K at $\varepsilon = 2$: $35.1 \pm$ **22.0**%

Such high variance reflects inherent unreliability, which poses a significant challenge for practical deployment. This observation confirms that the instability arises from the algorithm itself, rather than from implementation errors.

3. PromptDPSGD Struggles on Small Models. PromptDPSGD achieves only 18-39% accuracy across all tasks, far below DP-ES/DP-OPT (54-86% on the local model). This suggests that continuous soft-prompt optimization is less robust to DP noise than discrete prompt methods, especially on smaller models.

C.3.4 Implications for Method Selection

These results suggest that **task characteristics should inform method choice**:

- For **multi-step reasoning tasks** requiring strict template adherence (e.g., chain-of-thought prompts), **DP-ES is strongly preferred** due to its robustness against template damage and consistent performance.
- For **production deployments** where reliability is critical, **DP-ES's low variance** (5–7% std on GSM8K) compared to DP-OPT's higher variance (10–22% std) makes it the safer choice.

Importantly, the **main results (Table 2)** using larger API models (DeepSeek) show DP-ES achieving 88.1% on GSM8K and 99.7% on MedQA, demonstrating that performance scales with model capacity while maintaining the algorithmic advantages observed here. Figure 5 in the main text visualizes the complete privacy sweep across all methods.

C.4 Ablation Studies

To validate our design choices, we conduct comprehensive ablation studies on GSM8K with $\varepsilon = 1.0$.

Population Size Ablation (Table 8): We vary population size while keeping the total search budget bounded (pop $\times$ iter ≤ 24) under a fixed privacy budget. We find that $K = 6$ with $T = 3$ iterations provides optimal balance. Smaller populations (K=3) suffer from insufficient diversity (-8.0 pp degradation), while larger populations (K=9, 12) over-fragment the privacy budget, leading to noisy signals per candidate (-4.5 pp degradation).

Budget Allocation Strategy (Table 9): We compare calibrated uniform allocation (equal budgets for evaluation and selection, optimized via binary search) against fixed budget ratios (70-30, 50-50, 30-70 for scoring vs selection). While 70-30 achieves marginally higher mean (87.33% vs 86.33%), the difference

Allocation Strategy	Mean Accuracy	Std
Calibrated Uniform (Ours)	86.33%	2.39%
50-50 (Fixed)	86.83%	0.85%
70-30 (Fixed)	87.33%	1.43%
30-70 (Fixed)	86.67%	1.70%

Table 9: **Budget Allocation Ablation.** Calibrated uniform allocation provides robust performance without task-specific tuning. Differences between strategies are not statistically significant.

is not statistically significant ($p = 0.64$). We choose calibrated uniform allocation for its simplicity and generalizability across tasks without manual tuning of budget ratios.

C.5 Practitioner Hyperparameter Guide

Based on our ablation results (Tables 8, 9) and privacy sweeps (Table 7), we provide the following guidance for selecting K and T under a fixed privacy budget:

ε **range**	K	T	**Rationale**
≥ 2.0	6–9	3–4	Sufficient budget for larger populations; more iterations refine prompts further
0.5–2.0	6	3	Sweet spot: adequate diversity without fragmenting the scoring budget
< 0.5	3–4	2	Tight budget; fewer scorer invocations minimize noise per candidate

Table 10: **Recommended hyperparameters by privacy budget.** Start with $K{=}6, T{=}3$ as a default; decrease for tighter budgets.

Key heuristic: The product $K \times T$ should not exceed $\sim$18–24 under moderate budgets ($\varepsilon \approx 1.0$). Beyond this, additional scorer invocations dilute the per-query budget and amplify noise (Table 8). Under very tight budgets ($\varepsilon < 0.5$), prioritize fewer, higher-quality evaluations by reducing both K and T.

C.6 Selection Mechanism Ablation (Extended)

Table 3 in the main text summarizes the selection mechanism comparison. Here we provide the full results across all tested ε values on GSM8K (200 examples).

Key insight: EM consumes $\varepsilon_{\text{select}}$ per iteration, but under advanced composition the Gaussian scorer (K invocations per iteration) dominates the composition bound; halving its per-query budget reduces the dominant sum-of-squares term, while the single EM invocation adds negligibly. The Rényi DP accountant confirms $\varepsilon^{\text{EM}}_{\text{consumed}} \leq \varepsilon^{\text{Det}}_{\text{consumed}}$ across all tested budgets.

Detailed findings across ε values:

- At $\varepsilon{=}0.5$: EM mean 90.0% (std 0.7%) vs Det. top-k mean 90.0% (std 1.6%). EM shows lower variance due to probabilistic smoothing of noise-induced rank inversions.
- At $\varepsilon{=}1.0$: EM mean 88.8% (std 3.4%) vs Det. top-k mean 88.6% (std 2.9%). Performance is statistically indistinguishable.
- At $\varepsilon{=}2.0$: EM mean 91.2% (std 2.9%) vs Det. top-k mean 88.2% (std 1.4%). EM shows a marginal mean advantage with higher variance.
- At $\varepsilon{=}5.0$: EM mean 91.2% (std 2.0%) vs Det. top-k mean 90.6% (std 1.4%). Both converge as Gaussian noise diminishes.

Recommended default: We adopt EM as the default selection mechanism, as it provides probabilistic smoothing that yields lower run-to-run variance at tight privacy budgets—a desirable property for deployment reliability—without increasing total privacy consumption. Deterministic top-k is a viable alternative when implementation simplicity is preferred.

C.7 Population Diversity Analysis

To understand DP-ES's exploration-exploitation dynamics, we analyzed population diversity over 5 iterations using sentence-BERT embeddings (all-mpnet-base-v2). We measure diversity as the average pairwise cosine distance between prompt embeddings in the population. Figure 6 (main text) visualizes this trajectory.

D Proofs

D.1 Proof of Proposition 1

By the post-processing property of differential privacy (Dwork and Roth, 2014), any function applied to non-private data (the parent prompt) without querying private data preserves privacy. Since Mutate operates solely on public prompts using an external LLM, it is independent of $\mathcal{D}$ and thus trivially (ε, δ)-DP for any ε, δ, including $(0, 0)$.

D.2 Proof of Proposition 2

Because mutation and candidate generation do not access $\mathcal{D}$, they are post-processing of DP outputs and incur zero additional privacy loss. Therefore, for any number of mutations per iteration, the overall privacy loss is fully determined by the DP evaluation and selection steps, as captured by Theorem 4.

D.3 Proof of Theorem 4

Mutations consume zero budget (Proposition 1). Each iteration performs: (1) DP evaluation via Gaussian mechanism $(\varepsilon_{\text{eval}}, \delta_t)$-DP (Theorem 3), invoked K times, and (2) EM selection $(\varepsilon_{\text{select}}, \delta_t)$-DP, invoked once. Across T iterations, advanced composition (Theorem 5) over the $T \times K$ scorer invocations and T EM invocations yields $(\varepsilon_{\text{total}}, \delta)$-DP for appropriately calibrated budgets $\varepsilon_{\text{eval}}, \varepsilon_{\text{select}}$. When deterministic top-k replaces EM, only the $T \times K$ scorer invocations enter composition, and the guarantee holds *a fortiori*.