DualTune-GhostDP: A Unified Framework for Synergistic Differentially Private Fine-Tuning of Prompt-Based Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

With growing concerns about data privacy and confidentiality, there has been increased attention on privacy preserving integration in many applications, particularly data driven ones like Large Language Models (LLMs). LLMs are powerful in-context learners and are widely adopted in real world products. However, their dependence on sensitive private data in training and prompts exposes them to potential data leakage and privacy breaches. Differential Privacy (DP) delivers a rigorous, mathematically provable safeguard against these vulnerabilities; however, this assurance often comes with considerable reductions in model performance and increased computational cost. While prior work has highlighted the inherent trade-off between privacy and utility, our proposed method, DualTune-GhostDP, shows that strong privacy guarantees can be maintained under a controlled budget without sacrificing high model performance. Our method adopts a two-phase fine-tuning pipeline that integrates Ghost Clipping with an EdgeWorth (EW) Advanced Privacy Accountant, replacing conventional DP accounting mechanisms. Experimental results show that the principled integration of these components in DualTune-GhostDP consistently outperforms the individual benefits of each and both the single-phase Differentially Private Stochastic Gradient Descent (DP-SGD) baseline and a two-phase fine-tuning variant using standard clipping. Specifically, it achieves higher accuracy, faster convergence, and improved computational efficiency while maintaining differential privacy guarantees. In addition, we assess robustness to Membership Inference Attacks (MIA), which aim to determine whether a particular sample was used during training. Our findings demonstrate that DualTune-GhostDP substantially mitigates membership leakage across all training stages, strengthening both the privacy assurances and the overall stability of the approach against such attack relative to existing baselines.

1 Introduction

With advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP), large language models (LLMs), including encoder-only transformer-based models such as the BERT family, decoder-only models such as the GPT family, and encoder–decoder models such as T5 and BART, have demonstrated extraordinary capabilities across multiple domains, including language generation, sentiment analysis, question answering, and text classification Das et al. (2025). Their ability to generate and understand human-like contextual language has fostered significant advancements across diverse sectors and transformed multiple industries Trummer (2023). However, alongside these advancements, LLMs introduce significant privacy concerns Gupta et al. (2024).

LLMs rely on vast amounts of data that often contain personal or sensitive information, increasing the risk of inadvertent memorization and data leakage. Recent studies show that private information can leak from LLMs, leading to growing interest in privacy-preserving techniques Duan et al. (2024); Shi et al. (2021); Li et al. (2021a); Shi et al. (2022); Hong et al. (2023); Behnia et al. (2022). These risks are further exemplified by Membership Inference Attacks (MIA), where adversaries exploit model outputs to determine whether a

specific record was included in the training data. While MIA does not directly reconstruct sensitive records, it reveals weaknesses in how models handle training data and highlights that LLMs may overfit or retain information about individual samples, thereby undermining data confidentiality and user privacy Carlini et al. (2022b).

To mitigate such threats, various approaches have been proposed, with Differential Privacy (DP) emerging as a leading solution. However, as LLMs and attack strategies continue to evolve, ensuring strong privacy guarantees remains an ongoing challenge. Existing DP-based methods for LLMs, reviewed in Section 3, often suffer from a privacy–utility trade-off that leaves room for further optimization.

To address this challenge, we propose *DualTune-GhostDP*, a unified hybrid framework that leverages the strengths of multiple complementary techniques. Specifically, our approach adopts a two-phase fine-tuning strategy inspired by Shi et al. (2022), integrates the soft-prompting method of PromptDP-SGD Duan et al. (2023), enhances efficiency using ghost clipping as suggested in Li et al. (2021a), and employs the EW-Tune privacy accountant Behnia et al. (2022) in place of traditional RDP to maintain tighter privacy bounds. This combination enables improved privacy accounting while maximizing model performance under DP-SGD with ghost clipping.

Unlike prior work that applies dual-phase fine-tuning, ghost clipping, or advanced privacy accounting in isolation, DualTune-GhostDP is the first framework to jointly analyze their interaction. We show that this integration yields non-additive, mechanistically coupled benefits: sanitized pre-fine-tuning reshapes the per-example gradient distribution in the private phase, reducing gradient sensitivity and heavy-tailed behavior; Ghost Clipping then operates with less unnecessary gradient suppression; and the Edgeworth (EW) accountant exploits this stabilized training regime to provide tighter finite-sample privacy estimates. Together, these interactions enable higher accuracy, faster convergence, and lower memory consumption under the same privacy budget gains that cannot be explained by the sum of individual component improvements, but instead arise from their synergy, as validated by our ablations and gradient sensitivity analysis.

In this work, we use the term *differentially private prompting* to refer to privacy-preserving fine-tuning of language models that are subsequently deployed in prompt-based inference settings. Importantly, we do not apply differential privacy to user prompts at inference time. Instead, privacy guarantees are enforced during the fine-tuning stage, ensuring that the learned model parameters—and thus the model's responses to downstream prompts—do not leak information about individual training records. This formulation aligns with prior white-box prompt-learning and fine-tuning approaches, where prompt behavior is determined by representations acquired during training rather than by inference-time mechanisms. Consequently, mitigating privacy risks in such settings requires privacy-preserving fine-tuning, rather than relying solely on defenses applied at inference time Li et al. (2021b); Yu et al. (2021).

The remainder of this paper is organized as follows. Section 2 presents essential background concepts, while Section 3 reviews related work. Section 4 describes the proposed approach, including its motivation, building blocks, and threat model. Section 5 presents experimental results and discussion, followed by an evaluation of robustness to MIA in Section 6. Section 7 concludes the paper and outlines future directions. The code for our experiments is publicly available at `https://github.com/AnonymousSatMLll/Differentially-Private-Prompting-in-Large-Language-Models.git`.

2 Background and Preliminaries

This section presents the foundational background on key topics relevant to our study.

2.1 The Hype of Large Language Models

Artificial Intelligence (AI) allows machines to perform tasks that traditionally require human intelligence. Within this scope, Natural Language Processing (NLP) focuses on understanding and generating human language, leading to the development of Large Language Models (LLMs). These models leverage transformer-based deep learning architectures to process and generate text with remarkable fluency.

LLMs have gained attention in both academia and industry for their ability to solve diverse general-purpose tasks, rather than being limited to specific applications which made them highly versatile Chang et al. (2024). They are trained on vast datasets across various domains and require significant computational resources Myers et al. (2024). The different types of LLMs excel in tasks ranging from simple sentence classification, question-answering, and sentiment analysis to more complex applications such as sentence completion, conversational AI, and advanced text generation Anisuzzaman et al. (2024).
A key aspect of using LLMs is fine-tuning, which involves taking a pre-trained model and retraining it on domain-specific data to improve efficiency and performance for specialized tasks. Different modern fine-tuning techniques have been developed to enable effective adaptation to specific applications Parthasarathy et al. (2024).

2.2 Differential Privacy (DP)

As data usage continues to grow, the protection of sensitive and private information requires a rigorous approach. DP, introduced in 2006 Dwork (2006), provides a mathematical guarantee that ensures that an attacker, regardless of their computational power or access to data, cannot extract meaningful information about any individual.
DP guarantees that an individual's participation in a study or model has minimal impact on the outcome. In other words, any information inferred about a person comes solely from the model's output, not from their direct involvement. Consequently, DP makes it difficult for adversaries to infer sensitive information.
The core idea behind DP is the controlled addition of noise to model outputs or query responses. Given two neighboring datasets, D and D', which differ by only a single data point (through addition, removal, or substitution), DP guarantees that the statistical behavior of a mechanism responding to queries (i.e., the curator's method of answering) remains nearly identical while still providing useful responses. This in turn ensures that the answer does not reveal whether the mechanism was applied on D or D'.
DP can be applied in two primary ways: Global DP (GDP) and Local DP (LDP), but our focus will be on GDP. **Global DP**: The noise is added centrally by a trusted server. **Local DP**: Users perturb their data locally before sending it to an untrusted server, ensuring privacy even when the server is not trusted.

Definition 1 ((ϵ, δ)-DP)**.** *Given $\epsilon \geq 0$ and $\delta \geq 0$, a randomized algorithm M is (ϵ, δ)-differentially private if, for all adjacent datasets D, D' (differing by a single data sample) and all measurable subsets S in the output space, the following condition holds:*

$$\Pr(M(D) \in S) \leq e^{\epsilon} \Pr(M(D') \in S) + \delta.$$

ϵ is termed the privacy budget. For both ϵ and δ, smaller values correspond to stronger privacy guarantees.

2.3 Privacy Accountants and their Role in DP

DP mechanisms provide mathematical privacy guarantees, but these guarantees must be carefully tracked across repeated queries or training steps, as each application consumes part of the overall privacy budget. Without proper monitoring, cumulative privacy loss may exceed the allocated budget and breach the promised guarantees. To address this, *privacy accountants* were introduced to measure and track cumulative privacy loss, ensuring it remains within a predefined bound while enabling fine-grained control over different operations.

Several types of privacy accountants have been proposed for different mechanisms and noise distributions, including the Moments Accountant Abadi et al. (2016), Rényi divergence-based accounting (RDP), which offers tighter bounds and flexible choice of divergence order α Wang et al. (2019), Privacy Random Variable (PRV) accounting Gopi et al. (2021), and specialized accountants for Gaussian Koskela et al. (2022) and Laplace mechanisms Dwork et al. (2006). Among these, RDP has become widely adopted in practice due to its strong privacy bounds and flexibility. More recently, the **Edgeworth Privacy Accountant (EW)** was proposed in Wang et al. (2022), further improving over RDP by providing tighter finite-sample privacy estimates and better adaptability to practical DP training settings, as summarized in Table 5 in the Appendix A.1. The choice of privacy accountant ultimately depends on system requirements for accuracy, scalability, and implementation complexity.

3 Related Work

The success of LLMs in generating and processing information has impacted nearly every sector of our life with remarkable capabilities Duan et al. (2023); Devlin et al. (2019); Achiam et al. (2023); Zhang et al. (2022). However, this advancement has also raised critical privacy concerns, posing risks to personally identifiable information and the potential leakage of sensitive data Singh et al. (2024). In LLMs, the Prompting paradigm can be broadly studied under two paradigms: white-box and black-box settings. Our work operates in the white-box paradigm, where access to model gradients is available and fine tuning is duable, so this section reviews works primarily in this direction.

Several studies have focused on prompt fine-tuning in the white-box setting, where users have access to the model gradients. Li et al. Li et al. (2021a) addressed the challenge of integrating DP-SGD into LLMs while maintaining both performance and computational efficiency. They investigated the use of large pretrained language models, such as RoBERTa, and emphasized the critical role of selecting appropriate hyperparameters for optimal performance. Contrary to non-private setups where smaller batch sizes and learning rates are preferable, their findings demonstrated the advantages of using larger values in the private fine-tuning setting. Additionally, to improve memory efficiency, they introduced ghost clipping, an approach that optimizes memory usage when fine-tuning large transformers under DP-SGD. This technique extends the method in Lee & Kifer (2020) while avoiding the instantiation of per-example gradients, even for individual linear layers.

In conventional DP-SGD, gradients are computed for each sample, individually clipped to a predefined bound C, and then averaged with added Gaussian noise. Ghost clipping replaces explicit per-sample gradient computation with an analytical approximation of each sample's gradient norm, referred to as the *ghost norm*. The ghost norm g_i for a sample i is estimated using the activations and back-propagated gradients as $\|g_i\|_2^2 = \sum_l \|J_l(x_i) \cdot \delta_l\|_2^2$. where $J_l(x_i)$ denotes the Jacobian of the l-th layer activations with respect to the input, and δ_l represents the back-propagated error. This estimate enables efficient clipping of gradients without instantiating them individually in memory, thereby substantially reducing computation time and GPU usage. After estimating the ghost norms, gradients are rescaled by $\min(1, \frac{C}{\|g_i\|_2})$ and aggregated across the batch before Gaussian noise is added, consistent with the standard DP-SGD update rule. Experimental results on sentence classification datasets in this work, confirmed the effectiveness of this approach reducing memory consumption by at least a factor of 22. The study demonstrated that leveraging large pretrained models, well-chosen hyperparameters, and direct application of DP optimization during fine-tuning leads to high-performing DP language models, even under a moderate privacy budget, achieving a strong balance between privacy and utility.

Another line of work introduced Just Fine-Tune Twice "JFT" Shi et al. (2022), a framework that involves fine-tuning the model twice to safeguard against privacy leakage. The first fine-tuning phase is performed using redacted in-domain data, where sensitive information is concealed and no DP optimizer is applied. The second fine-tuning phase is conducted with the original data under a private mechanism. The proposed framework achieves Selective Differential Privacy (SDP), an extension of DP formalized by Shi et al. Shi et al. (2021), based on the understanding that sensitive information is typically sparse. SDP defines neighboring datasets to differ only in the sensitive part of a training example and as a result, SDP selectively hides the difference in the sensitive part only where they stated that it's particularly suitable for NLP tasks. In the first phase, redacted data is extracted from the original data using a secret detector. Based on the detector's performance through recall metric evaluation between the redacted data and the original, three methods are then employed to fine-tune the redacted data. If the detector successfully masks all sensitive information so recall score of 100, the model is fine-tuned directly with a public unnoised optimizer. If the detector is imperfect, an affordable subset is manually screened and fine-tuned using the public optimizer. In cases where the detector is imperfect and manual screening misses some sensitive data, light noise is added, and a private optimizer is used, where the missed sensitive information receives smaller epsilon values compared to the rest of the redacted data. After fine-tuning on the redacted data, the model is further fine-tuned on the original data using DP-SGD with a private optimizer. For their experiments, the authors used natural language understanding (NLU, on GLUE) and language generation datasets, demonstrating that JFT outperforms DP-SGD and Redacted-only models.

Building on DP fine-tuning, EW-Tune Accountant Behnia et al. (2022) improved privacy accounting by leveraging Edgeworth approximations for finite-step training. This method tightens privacy guarantees and optimizes the noise multiplier, thereby reducing the utility loss typically incurred in DP-SGD. Experimental results on GLUE benchmarks demonstrated up to a 1.1% accuracy gain with reduced noise levels, showing its effectiveness for practical DP fine-tuning.

Some hybrid works also consider overlapping paradigms where gradient-based and gradient-free methods overlap Duan et al. (2023). In this work, they proposed PromptDP-SGD, a gradient-based method that privately tunes soft prompt embeddings, achieving comparable performance to private fine-tuning but with reduced storage and training costs. In cases where gradients are unavailable, they extended the approach with PromptPATE, a black-box compatible method, though the white-box variant remains directly relevant to our setting.

While black-box approaches such as DP-OPT Hong et al. (2023) and DP-GTR Li et al. (2025) offer privacy-preserving solutions when gradients are inaccessible, they fall outside the scope of our contribution. Our focus remains on the white-box paradigm, where gradient access enables more direct integration of privacy-preserving optimizations.

Safeguarding LLMs is still an ongoing challenge, due to its devastating impacts. Existing white-box solutions address memory efficiency (ghost clipping), selective privacy (JFT), and tighter privacy accounting (EW-Tune). However, these contributions remain fragmented, with no unified approach combining their strengths. To bridge this gap, we propose DualTune-GhostDP, a two-phase fine-tuning strategy that integrates ghost clipping for efficient gradient handling and EW-Tune for advanced privacy accounting, thereby mitigating privacy leakage while maintaining high utility.

4 Proposed Approach: DualTune-GhostDP

We next outline the motivation for our model and its underlying core building blocks.

4.1 Motivation

Current methods for defending against privacy leakage demonstrate fair and consistent accuracy values, yet they still lack a solid guarantee for privacy in LLMs Duan et al. (2023); Li et al. (2021a); Shi et al. (2022); Hong et al. (2023); Behnia et al. (2022). While each approach excels in suggesting a specific module for enhancing either accuracy, memory or convergence, achieving superior results, they still fall short in achieving an optimal privacy-utility trade-off that leaves a room for enhancement. Moving forward, our contribution lies in the principled integration and empirical validation of these components as a unified end-to-end system, which has not been previously studied, being the first framework to jointly analyze this interaction. We will explore advanced approaches that extend beyond the black-box setting, where only model outputs are observable, to the white-box gradient setting, in which fine-tuning is performed with full access to gradients. A promising direction is a two-layer framework: an initial privacy-free relaxation phase, followed by a Gaussian DP-SGD phase that leverages Gaussian Differential Privacy (GDP). In this setup, noise is injected during training by a trusted entity operating under a rigorous privacy accountant.

We emphasize that the privacy guarantee in this framework applies to the second fine-tuning phase only. The first phase operates on sanitized data and is treated as public preprocessing under the differential privacy paradigm Parthasarathy et al. (2024). Consequently, the overall framework provides formal (ϵ, δ)-differential privacy with respect to the original unredacted dataset through the second phase, while the first phase does not consume any privacy budget.

4.2 Model building blocks

The proposed solution **DualTune-GhostDP** integrates DualTune, a two-phase fine-tuning method, with GhostDP, which utilizes Ghost Clipping instead of the canonical clipping technique used in traditional DP. In addition, it utilizes the EW privacy accountant instead of the RDP that prior work used. The selection of this clipping technique and the privacy accountant is based on their demonstrated superiority over the

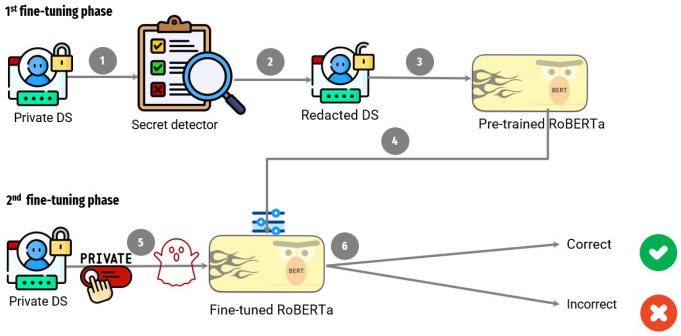

Figure 1: DualTune-GhostDP workflow.

other adopted methods, as shown in tables 5 and 6 in the Appendix, which compare the selected technique to the most commonly used methods in the literature.

The DualTune-GhostDP approach consists of the following steps as shown in Figure 1:

1. Fine-Tuning on Redacted Data: The large pre-trained model is first fine-tuned on redacted (sanitized) data, generated by applying a high-contextual secret detector to the original dataset. This allows the model to learn useful task representations without the addition of privacy noise, ensuring that sensitive details are not memorized during training.The **Secret Detector** is a deterministic preprocessing function that removes predefined sensitive attributes, including named entities, pronouns, proper nouns, verbs, and syntactic roles (subjects/objects). The detector is implemented using SpaCy-based NER, POS tagging, and dependency parsing, and its output is treated as non-sensitive.

2. Fine-Tune using DP-SGD with Ghost Clipping: The original (unredacted) private data is used to further fine-tune the model obtained from Stage One. This is done under a DP setting, incorporating Ghost Clipping. The model is then re-evaluated to assess improvements.

3. Privacy controlling using EW Privacy Accountant: Optimize noise levels to add less noise than RDP while still ensuring DP guarantees. This improves model utility by accounting for finite training steps, rather than assuming unlimited iterations. It also enhances compatibility with Ghost Clipping, as it does not assume per-step clipping but instead dynamically optimizes privacy bounds.

4.3 Formal Algorithm Description

Let M_0 denote a pre-trained language model and D the original training dataset. DualTune-GhostDP proceeds as follows. **Sanitization:** a sanitized dataset $D_{\text{san}} = f(D)$ is computed using a secret detector $f(\cdot)$ that removes predefined sensitive attributes. **Phase 1 (Public Fine-Tuning):** the model is fine-tuned on D_{san} using a standard optimizer, yielding $M_1 = \text{FineTune}(M_0, D_{\text{san}})$. **Phase 2 (Private Fine-Tuning):** starting from M_1, DP-SGD with Ghost Clipping is applied on D, where per-sample gradient norms are approximated, clipped to ℓ_2 bound C, aggregated, and perturbed with Gaussian noise calibrated to multiplier σ. **Privacy Accounting:** cumulative privacy loss from Phase 2 is tracked using the Edgeworth (EW) accountant, yielding a final budget (ϵ, δ).

This process produces the final model M^*, which satisfies differential privacy guarantees with respect to D, as summarized in Pseudocode 1.

4.4 Privacy Analysis

Differential Privacy Guarantee. Phase 2 of DualTune-GhostDP is trained using DP-SGD with Ghost Clipping, clipping norm C, noise multiplier σ, and sampling rate q. When privacy loss is tracked using the

Pseudocode 1 Privacy-Preserving Dual-Phase Fine-Tuning (DualTune-GhostDP)

Require: Pre-trained RoBERTa model M, Dataset D
Ensure: Final fine-tuned model M^*
Initialize $M \leftarrow$ Load RoBERTa
$D \leftarrow$ Load and preprocess(D)
$D_{balanced} \leftarrow$ Balance(D)
$D_{sanitized} \leftarrow$ ApplySecretDetector($D_{balanced}$)
$M \leftarrow$ FineTune($M, D_{sanitized}$) ▷ First Fine-tuning Phase
$M^* \leftarrow$ FineTuneDP($M, D_{balanced}$, ghost clipping) ▷ Second Fine-tuning Phase
Monitor $\epsilon \leftarrow$ EW(M^*)
Measure performance metrics: accuracy, execution time, memory usage
Compare model accuracy under varying ϵ values

Edgeworth (EW) accountant, the resulting model M^* satisfies (ϵ, δ)-differential privacy with respect to the original dataset D.

Ghost Clipping preserves the sensitivity of standard per-sample clipping by providing exact or conservative estimates of per-sample gradient norms Lee & Kifer (2020), ensuring that each update corresponds to a Gaussian mechanism with bounded sensitivity. The EW accountant composes these mechanisms over a finite number of training steps using Edgeworth expansions, yielding tighter privacy estimates than RDP-based accounting. Since Phase 1 operates solely on sanitized data, it incurs no privacy cost; thus, the overall guarantee is fully determined by Phase 2.

4.5 Threat Model

We adopt the *global differential privacy* setting, where training is managed by a trusted controller enforcing privacy guarantees through Gaussian DP-SGD. The adversary is modeled as an external analyst with black-box access to model predictions, attempting to infer whether a specific data record was used during training or to reconstruct sensitive inputs, without access to raw data or internal gradients.

The trusted controller is responsible for clipping gradients and injecting calibrated Gaussian noise before aggregation. The *secret detector* is assumed to be a semi-trusted local component that redacts sensitive entities prior to training, ensuring that no raw identifiers enter the pipeline.

5 Experiments and Discussion

In this section, we present four experiments designed to evaluate the effectiveness of each building block of our proposed method and to demonstrate improvements over prior work. All experiments are documented in the GitHub link provided at the end. Experiment 1 investigates the optimal epoch count for each dataset and analyzes how accuracy varies with this parameter for a single-phase DP-SGD model. Experiment 2 reproduces the approach of Shi et al. (2022), but substitutes the RDP accountant with the EW accountant to assess its impact. Experiment 3 evaluates our proposed method, DualTune-GhostDP, which builds on Experiment 2 by replacing standard clipping with Ghost Clipping. This modification emphasizes improvements in runtime, memory efficiency, convergence, and accuracy. Finally, Experiment 4 demonstrates the consistency of the privacy–utility trade-off on range of epsilon values achieved by our proposed method. It shows the consistent performance of DualTune-GhostDP under different epsilon values. The implementation and evaluation of **DualTune-GhostDP** are performed using an HPC cluster to ensure efficient code execution and result visualization.

This section will start with overview on model, dataset and mechanism selection with justification, then defines threat model and the trusted entities, after that the experimenys and their results are demonstrated in details. A final subsection shed the lights on the comparison of our model: DualTune-GhostDP over the existing methods to assess the relative effectiveness of the proposed approach.

5.1 Model, Datasets and Mechanism Selection

The **RoBERTa** model is chosen for its strong, well-established performance and its prevalent use in related literature, ensuring a fair basis for comparison. Additionally, it is well-suited for deployment in resource-constrained environments. Three binary classification datasets from the GLUE benchmark were used for evaluating our suggested model: SST-2 (Stanford Sentiment Treebank 2), QNLI (Question Natural Language Inference) and QQP (Quora Question Pairs). **SST-2** is a sentiment classification dataset that contains movie reviews from Rotten Tomatoes, labeled as either positive (1) or negative (0), and is widely regarded for its effectiveness in evaluating sentiment analysis models. **QNLI** is a dataset containing questions/answers task classifying pairs as *Entailment* where the answer sentence logically answers the question or *Not Entailment* where it does not. **QQP** is a paraphrase detection task that detects *Duplicate* versus *Not Duplicate* question pairs. These datasets were chosen to assess the generalizability of our method across different natural language understanding tasks. Table 7 in the Appendix demonstrates some statistics on the used datasets. All experiments in this work employ the *Gaussian differential privacy mechanism* implemented through *DP-SGD*, where Gaussian noise is added to the aggregated, clipped gradients after each mini-batch update. The privacy guarantees follow $(\epsilon = 3, \delta = 10^{-5})$-differential privacy, with a fixed noise multiplier of $\sigma = 1$ for Experiments 1–3 to enable consistent and fair comparison with prior work. Experiment 4 further explores different values of ϵ while maintaining the same δ and σ to analyze the privacy–utility trade-off.

5.2 Experiments

To demonstrate the effectiveness of each building block, four experimental setups are developed.

To generate the redacted dataset used in experiments 2 (Two-phase fine-tuning with DP-SGD and standard clipping) and 3 (Two-phase fine-tuning with DP-SGD and Ghost clipping), SpaCy's off-the-shelf NLP tools were utilized, including Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and dependency parsing. Since sentiment classification tasks often involve sensitive information that is semantically or contextually implied rather than explicitly stated, a High Contextual Secret Detector was adapted to redact a broader set of features, including: All 18 NER entity types, Pronouns (POS tag: PRON), Proper nouns (PROPN), All verbs (VERB), Subjects and objects (dependency tags: nsubj, dobj, pobj).

Experiment 1: Baseline: Single-phase private fine-tuning using DP-SGD
This experiment evaluates single-phase private fine-tuning using DP-SGD with standard clipping on SST-2, QNLI, and QQP. The number of training epochs is varied to identify optimal convergence under a fixed privacy budget $(\epsilon = 3, \delta = 10^{-5})$. Figure 4 in Appendix A.4.1 reports accuracy as a function of epoch count. The optimal epoch counts were four for SST-2 and QNLI, and two for QQP. These results establish the baseline performance used for comparison in subsequent experiments.

Experiment 2: Two-phase fine-tuning with DP-SGD and standard (normal) clipping and EW Privacy Accountant

In this experiment, the model undergoes two sequential fine-tuning phases. The first phase uses a public optimizer on redacted data where sensitive information is masked using a high-contextual detector without any privacy protection. The second phase applies DP-SGD with standard clipping on the original, unredacted data along with EW privacy accountant.

Each task (SST-2, QNLI, and QQP) is fine-tuned using varying combinations of epoch counts across the two phases, aiming to identify the optimal configuration that achieves the highest accuracy under a fixed privacy budget of $\epsilon = 3$. Epoch combinations start from (1,1), representing 1 epoch for each phase and increase in multiple of 2 until the maximum epoch count used in Experiment 1 (e.g., up to (4,4) for SST-2).

Discussion:
The optimal utility is achieved with fewer total epochs used in experiment 1 particularly at the (2,2) for SST-2, (4,4) for QNLI and at (2,1) for QQP configuration as shown in Tables 8, 9 and 10 which also corresponds to shorter training time. Beyond this point, additional epochs result in diminishing returns, where more computation time yields lower accuracy, making the setup both inefficient and suboptimal.
Comparing with experiment 1, the results indicate that the two-phase fine-tuning approach significantly

outperforms the traditional single-phase fine-tuning with standard DP-SGD clipping under identical settings. Notably, this improvement in accuracy is achieved with fewer total epochs. For SST-2, accuracy increased by 3%, from 87.76.90% to 91.38%; for QNLI, by 5%, from 79.82% to 84.53%; and for QQP, by 6%, from 77.44% to 82.69% as shown from the tables.

A sanity check confirms that any configuration from Exp2 (not necessarily the optimal) achieves higher accuracy than all configurations in Exp1 across all datasets. These results highlight the effectiveness of introducing an initial redaction phase. This phase not only preserves privacy by removing sensitive information but also appears to enhance model performance and sufficiently initializes useful representations which advocates the idea of Tramer & Boneh (2020) that access to features learned on public data from a same domain can enhance privacy-preserving learning performance. The findings support the claim that sensitive data is not essential for effective training and that its careful removal can, in fact, be beneficial.
Gradient Sensitivity Analysis: Beyond accuracy improvements, we analyze gradient sensitivity during the private fine-tuning phase of Exp1 and Exp2 to better understand the effect of sanitized initialization on DP-SGD optimization. We summarize gradient statistics using the mean, median, and 95th percentile ($p95$) of per-example ℓ_2 gradient norms, together with the clipping rate. These metrics jointly capture the central tendency, tail behavior, and the extent of clipping-induced distortion introduced by DP-SGD.

As shown in Appendix A.4.2 (Table 12), per-example gradient sensitivity statistics are reported for all evaluated datasets. To discuss SST-2 as a representative case, Exp2 substantially reduces per-example gradient sensitivity relative to Exp1: the mean gradient norm decreases by nearly an order of magnitude, indicating a more stable optimization regime, while the median norm approaches zero, reflecting tightly concentrated per-example updates. In addition, the $p95$ norm is significantly reduced, evidencing a marked attenuation of heavy-tailed gradient behavior known to exacerbate DP-SGD clipping and noise amplification. This effect is further reflected in the clipping rate: whereas more than 41% of gradients are clipped in Exp1, only 15.82% are clipped in Exp2. Together, these results indicate that Exp2 preserves a larger fraction of the original gradient signal, thereby reducing clipping-induced bias and improving optimization efficiency under the same DP configuration.

Experiment 3: Two-phase fine-tuning with DP-SGD and Ghost Clipping

In this experiment, the setup and hyperparameters are identical to those used in Experiment 2, with the only difference being the clipping technique. Ghost Clipping is employed here instead of standard clipping. As before, each task (SST-2, QNLI, and QQP) is fine-tuned using various combinations of epoch counts across the two training phases, with the objective of identifying the optimal configuration that maximizes accuracy under a fixed privacy budget of $\epsilon = 3$.
The tables below summarize the best-performing epoch combinations for each dataset. Highest accuracy was achieved using (2,1) epochs for SST-2 with 92.29%, (2,4) epochs for QNLI with 86.03% and (2,2) for QQP with 83.72%.

Discussion:
Compared to Experiment 2, Ghost Clipping provided significant memory efficiency gains, reducing usage by approximately three times from 24,880.69 MB for SST-2 and 23,850.41 MB for QNLI and QQP (under normal clipping) to a uniform 8,694.16 MB across all datasets as shown in Table 1. This memory reduction stems from the reduced need for large accumulated gradient buffers when using ghost norm estimation, which also accelerated model convergence. The identical memory usage observed under Ghost Clipping highlights its independence from dataset structure and sequence length, leading to equalized memory usage. In contrast, the higher memory consumption of SST-2 under normal clipping can be explained by its longer average input sequences (full movie review sentences), which inflate per-sample gradient storage requirements despite the dataset's smaller size. Meanwhile, QNLI and QQP, both sentence-pair tasks with similar average tokenized lengths, naturally exhibit comparable memory usage.

Consequently, the model achieved faster convergence in terms of runtime. Ghost Clipping reduced the training time by 25.6% for SST-2, 17.9% for QNLI, and 5.4% for QQP, as shown in Table 2. While QQP required the same number of epochs, both SST-2 and QNLI benefited from fewer total epochs count, further contributing to computational efficiency.

Table 1: Peak Memory Usage (MB): Normal vs. Ghost Clipping

Dataset	Normal Clipping	Ghost Clipping	Reduction
SST-2	24,880.69	8,694.16	∼3×
QQP & QNLI	23,850.41	8,694.16	∼2.7×

Among the datasets, SST-2 exhibited the most pronounced reduction. This can be attributed to the interaction between dataset size and clipping strategy. While SST-2 is smaller in scale compared to QNLI and QQP, the per-sample gradient computation in traditional DP-SGD introduces significant overhead. Ghost Clipping eliminates this overhead, leading to greater efficiency gains. In contrast, for larger datasets such as QQP, runtime is dominated by dataset scale that is the high number of training instances rather than per-sample operations, resulting in smaller relative improvements.

Table 2: Runtime Comparison: Normal vs Ghost Clipping

Dataset	Normal Clipping	Ghost Clipping	Reduction
SST-2	00:57:30	00:42:46	↓ 25.6%
QNLI	03:28:20	02:51:06	↓ 17.9%
QQP	05:19:29	05:02:05	↓ 5.4%

Interestingly, this reduction in memory was accompanied by improved optimal accuracy values across the three datasets as shown in Figure 3. In this experiment, Ghost Clipping was shown to enhance utility likely because it removes less private information and avoids unnecessary gradient suppression during updates, requiring less noise calibration.

Experiment 4: Privacy-utility trade off
The objective of this experiment is to evaluate the performance of our proposed method in balancing privacy and utility across different privacy budgets, ranging from 1 to 8. Since lower values of ϵ correspond to stronger privacy guarantees, often at the cost of reduced model accuracy, this study examines the robustness of our approach across this spectrum. We conduct the evaluation on two benchmark datasets—SST-2 and QQP—chosen to represent the highest (SST-2) and lowest (QQP) model performance, respectively.

Discussion:
Figure 2 demonstrates that accuracy generally improves as ϵ increases. For SST-2, our model sustains high accuracy even under strict privacy ($\epsilon = 1$), achieving 91.53% with minimal variation across seeds, which reflects strong robustness to privacy noise. As ϵ relaxes, accuracy improves steadily and peaks at 92.23% ($\epsilon = 8$), showing that the model benefits from larger budgets without overfitting.

On QQP, accuracy improves as ϵ increases: it is already strong at low values, achieving 82.88% when $\epsilon = 1$, and then peaks at 84% when $\epsilon = 8$. This trend highlights the privacy–utility trade-off, where tighter privacy guarantees (smaller ϵ) slightly reduce performance, while relaxing privacy constraints (larger ϵ) provides higher utility, though with diminished privacy protection.

Overall, the proposed method consistently surpasses prior work across all privacy settings, while maintaining robustness at low ϵ. These findings underscore the effectiveness of combining the EW privacy accountant with Ghost clipping in dual tune fine-tuning phases model for stable privacy-preserving fine-tuning.

5.3 Comparison with Prior Work

We compare against three representative DP fine-tuning baselines covering single-phase, dual-phase, and memory-efficient optimization These findings strongly support the effectiveness of our suggested model that uses Ghost Clipping over standard clipping in the DualTune fine-tuning process and EW privacy accountant over RDP, offering advantages in memory usage, time, and utility. Table 3 shows a comparison between our suggested model and the three works in literature under same privacy budget and using the same pre-trained model RobERTa. **PromptDP-SGD:** (single-phase DP fine-tuning with RDP accountant) Duan et al. (2023), **JFT** (Just fine tune twice,two-phase fine-tuning with RDP and normal Clipping) Hong et al. (2023)

Table 3: Performance comparison with prior work at $\epsilon = 3$. Best results are in **bold**.

Model (Prior Work ↓, Dataset →)	**SST-2**	**QNLI**	**QQP**
PromptDP-SGD	90.48	83.62	80.29
JFT	89.22	84.02	84.77
RoBERTa with DP-Adam	86.12	84.62	**85.41**
Ours (EW Accountant + Ghost)	**92.08±0.38**	**85.95±0.18**	84.40±0.14

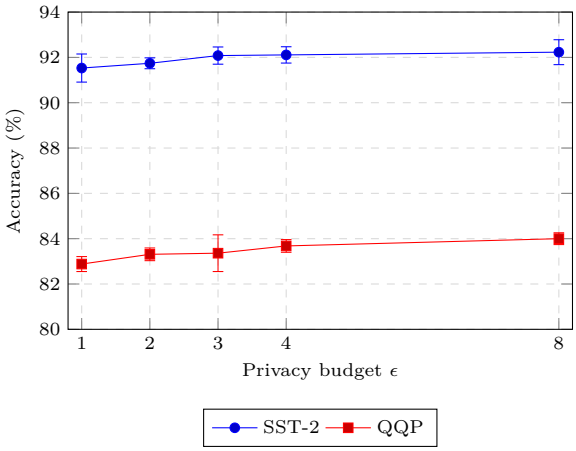

Figure 2: Privacy–utility trade-off for our model on SST-2 and QQP datasets across different privacy budgets ϵ. Error bars indicate standard deviation over multiple seeds.

Table 4: Accuracy Evolution Across All Experiments (mean ± std).

Dataset	**E1:Single (DP-SGD)**	**E2:Dual+Normal+EW**	**E3:Dual + Ghost**
SST-2	87.76 ± 1.43	91.38 ± 0.43	92.08 ± 0.38
QNLI	79.82 ± 0.34	84.80 ± 0.33	85.95 ± 0.18
QQP	77.44 ± 0.87	82.69 ± 0.21	84.40 ± 0.14

and **RoBERTa with DP-Adam** (single phase fine-tuning using RDP and ghost clipping) Li et al. (2021a). Our model, DualTune-GhostDP, clearly outperforms all the other works emphasizing the architecture benefits for the two datasets SST-2 and QNLI. However, for QQP, it outperforms PromptDP-SGD by 4%, while remaining comparable to the other two techniques, which could potentially surpass it when multiple runs and additional resources are considered.

Our results confirm the effectiveness of the proposed unified DualTune-GhostDP approach. The two-phase fine-tuning strategy combined with the EW Accountant, along with the integration of Ghost Clipping, yields a high-performing private model. This model outperforms both the two-phase variant with normal clipping under the RDP Privacy Accountant and the baseline single-phase private fine-tuning. The accuracy gains and their standard deviations over five runs, contributed by each component of our framework, are presented in Figure 3 and Table 4. Specifically, Experiment 2 extends the work of Shi et al. (2022) by replacing the privacy accountant, while Experiment 3 represents our complete proposed method, incorporating both the new privacy accountant and Ghost Clipping. Together, these experiments validate our DualTune-GhostDP framework as a high-performing, fast, efficient, and strong privacy-preserving fine-tuning unified strategy. It achieves strong utility without compromising privacy, efficiency, or scalability.

While RoBERTa is used for consistency with prior work, DualTune-GhostDP is model-agnostic and applies to any transformer fine-tuned with DP-SGD. The reliance on RoBERTa reflects experimental control rather than architectural dependence. All components: sanitization, ghost clipping, and EW accounting operate independently of model architecture.

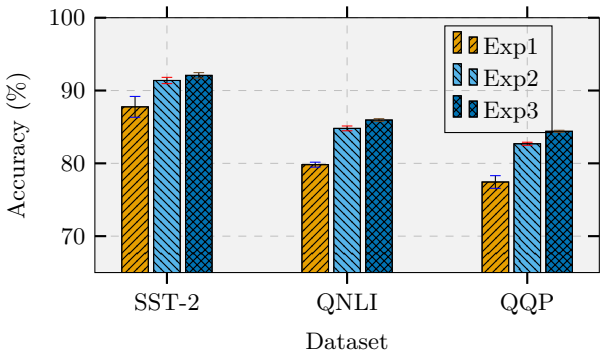

Figure 3: Accuracy evolution across all experiments (mean ± std).

5.4 Generalization to Other Transformer-based Backbones

To assess whether the observed benefits of DualTune-GhostDP are specific to RoBERTa or extend to other transformer architectures, we repeat our experiments using BERT-base (uncased) under the same privacy budget and comparable training settings. Table 11 in the Appendix isolates the incremental contributions of each training stage. Transitioning from single-phase DP-SGD (E1) to dual-phase training (E2) yields a substantial accuracy improvement of 5.02%, while introducing ghost clipping (E3) preserves accuracy and reduces peak GPU memory usage over E2 by approximately 4 times. These results confirm that each stage contributes a distinct and complementary benefit.

6 Membership Inference Attack Results

We evaluate the privacy guarantees of DualTune-GhostDP under a strong Membership Inference Attack (MIA) setting, where an adversary attempts to determine whether a specific sample was used during training based on model outputs. Following standard practice for DP-SGD evaluation, we consider a direct MIA on the SST-2 dataset, in which the adversary has access to model posteriors and ground-truth membership labels, representing a worst-case threat model Shokri et al. (2017); Carlini et al. (2022a). Results show that the post-sanitization model and the full DualTune-GhostDP pipeline both achieve attack performance indistinguishable from random guessing, while a single-phase DP-SGD baseline exhibits noticeably higher and less stable attack success. These findings confirm that combining data sanitization with DP-SGD and Ghost Clipping effectively suppresses membership leakage, yielding stronger privacy protection without sacrificing model utility. A detailed experimental setup is described in the Appendix A.5 and Table 13.

7 Conclusion

In this paper, we proposed DualTune-GhostDP, an approach designed to enable differentially private prompting in large language models (LLMs). The approach consists of two fine-tuning phases: the first phase employs a high-context secret detector to sanitize the training data, ensuring that sensitive information is removed before further training. The second phase fine-tunes the model on the original dataset under DP constraints, using Ghost Clipping to reduce computational overhead compared to traditional clipping techniques with increased accuracy. To ensure tight and accurate privacy guarantees, the framework incorporates a privacy accountant mechanism, EW Privacy Accountant that proved to be better for LLM fine-tuning tasks. Experimental results demonstrate that DualTune-GhostDP maintains strong model performance while saving memory, reducing runtime, speeding up convergence and incorporating privacy compared to prior works.

Limitations and Future Work. In our experiments, we fine-tuned the encoder-only models as the pre-trained backbone to maintain consistency with prior work, but exploring other modern decoder-only LLMs is our future direction which raises additional architectural considerations related to parameter sharing and Ghost Clipping efficiency, which we discuss in Appendix.

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report (2023). *arXiv preprint arXiv:2303.08774*, 2023.

DM Anisuzzaman, Jeffrey G Malins, Paul A Friedman, and Zachi I Attia. Fine-tuning llms for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 2024.

Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 560–566, 2022.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022a.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022b.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *Journal of the ACM*, 15(3):1–45, 2024.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *Communications of the ACM*, 57(6):1–39, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *Adv. Neural Inf. Process. Syst.*, 36: 76852–76871, 2023.

Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the privacy risk of in-context learning. *arXiv preprint arXiv:2411.10512*, 2024.

Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pp. 1–12, 2006.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51, 2006. doi: 10.1145/1130000.1130002. URL `https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf`.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

Brij B Gupta, Akshat Gaurav, Varsha Arya, Wadee Alhalabi, Dheyaaldin Alsalman, and P Vijayakumar. Enhancing user prompt confidentiality in large language models through advanced differential encryption. 116:109215, 2024.

Junyuan Hong, Jiachen T Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. Dp-opt: Make large language model your privacy-preserving prompt engineer. *arXiv preprint arXiv:2312.03724*, 2023.

Antti Koskela, Marlon Tobaben, and Antti Honkela. Individual privacy accounting with gaussian differential privacy. *arXiv preprint*, 2022. URL `https://arxiv.org/abs/2209.15596`. arXiv:2209.15596.

Jaewoo Lee and Daniel Kifer. Scaling up differentially private deep learning with fast per-example gradient clipping. *arXiv preprint arXiv:2009.03106*, 2020.

Mingchen Li, Heng Fan, Song Fu, Junhua Ding, and Yunhe Feng. Dp-gtr: Differentially private prompt protection via group text rewriting. *arXiv preprint arXiv:2503.04990*, 2025.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021a.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021b.

Devon Myers, Rami Mohawesh, Venkata Ishwarya Chellaboina, Anantha Lakshmi Sathvik, Praveen Venkatesh, Yi-Hui Ho, Hanna Henshaw, Muna Alhawawreh, David Berdik, and Yaser Jararweh. Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. 27(1):1–26, 2024.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2408.13296`. arXiv:2408.13296.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. Selective differential privacy for language modeling. *arXiv preprint arXiv:2108.12944*, 2021.

Weiyan Shi, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. Just fine-tune twice: Selective differential privacy for large language models. *arXiv preprint arXiv:2204.07667*, 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

Tanmay Singh, Harshvardhan Aditya, Vijay K Madisetti, and Arshdeep Bahga. Whispered tuning: Data privacy preservation in fine-tuning llms through differential privacy. *Journal of Software Engineering and Applications*, 17(1):1–22, 2024.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.

Immanuel Trummer. From bert to gpt-3 codex: harnessing the potential of very large language models for data management. *arXiv preprint arXiv:2306.09339*, 2023.

Hua Wang, Sheng Gao, Huanyu Zhang, Milan Shen, and Weijie J Su. Analytical composition of differential privacy via the edgeworth accountant. *arXiv preprint arXiv:2206.04236*, 2022.

Yu-Xiang Wang, Borja Balle, and Shiva P. Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pp. 1226–1235, 2019. URL `https://proceedings.mlr.press/v89/wang19b.html`. Proceedings of Machine Learning Research, PMLR.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Appendix: Supplementary Material

This appendix provides additional theoretical justification, dataset statistics, extended experimental results, and detailed privacy and attack evaluations omitted from the main paper due to space constraints. All results are reported to ensure reproducibility and completeness.

A.1 Comparison of Privacy Accounting and Gradient Clipping Mechanisms

This section provides a comparative analysis of the core building blocks adopted in DualTune-GhostDP against traditional alternatives commonly used in differentially private training.

Table 5: Comparison between Rényi Privacy Accounting (RDP) and Edgeworth Privacy Accounting (EW).

Aspect	**Rényi DP (RDP)**	**Edgeworth (EW)**
Privacy regime	Asymptotic (many iterations)	Finite-sample training
Composability	Moment-based composition	Edgeworth expansion
Noise calibration	Conservative	Tighter for same ϵ
Accuracy–privacy trade-off	Lower utility	Improved utility
Best suited for	Large-scale DP training	DP fine-tuning of LLMs

Table 6: Comparison between Standard DP-SGD Clipping and Ghost Clipping.

Aspect	**Standard DP-SGD**	**Ghost Clipping**
Gradient computation	Explicit per-sample gradients	Implicit norm estimation
Computational cost	High	Low
Memory usage	High	Low
Accuracy impact	Standard DP trade-off	Comparable with efficiency gains

A.2 Dataset Statistics

Table 7 reports detailed statistics for all datasets used in this work, including task type, dataset size, and balanced training splits.

Table 7: GLUE Dataset Statistics with Task Types and Balanced Train Sizes

Dataset	**Task Type**	**Split**	**Size (MB)**	**#Samples**
SST-2	Sentiment Analysis	Train	11.82	67,349 (59,560 balanced)
		Validation	0.15	872
		Test	0.32	1,821
QQP	Paraphrase Detection	Train	63.85	363,846 (268,756 balanced)
		Validation	7.09	40,430
		Test	68.61	390,965
QNLI	QA/NLI	Train	18.38	104,743 (104,732 balanced)
		Validation	0.96	5,463
		Test	0.96	5,463

A.3 Extended Accuracy Results and Epoch Sensitivity

A.3.1 Accuracy Variation with Epoch Combinations

Figures 8, 9 and 10 detail the full accuracy vs epoch sensitivity.

Table 8: Accuracy Variation with Epoch Combinations for SST2

# Epochs (Phase 2 ↓, Phase 1 →)	1	2	4
1	90.42%	89.72%	91.00%
2	91.36%	**91.38**%	90.77%
4	89.72%	91.12%	91.36%

Table 9: Accuracy Variation with Epoch Combinations for QNLI

# Epochs (Phase 2 ↓, Phase 1 →)	1	2	4
1	83.49%	81.94%	83.64%
2	83.31%	83.59%	83.88%
4	84.14%	84.2%	**84.53**%

Table 10: Accuracy Variation with Epoch Combinations for QQP

# Epochs (Phase 2 ↓, Phase 1 →)	1	2
1	82.53%	73.39%
2	82.16%	**82.69**%

A.3.2 Generalization to BERT-base (Uncased)

To assess the generality of the proposed framework beyond the primary backbone, we evaluate DualTune-GhostDP on **BERT-base (uncased)** under the same privacy budget with the results documented in table 11.

Table 11: Targeted improvements on **BERT-base (uncased)** for SST-2 at $\epsilon = 3$: accuracy gain from E1→E2 and memory reduction from E2→E3.

Improvement Target	**Transition (Before → After)**	**Net Change**
Accuracy (%)	85.51 → 90.53	↑ **+5.02**
Peak GPU Reserved (MB)	27748.00 → 7038.00	↓ **−20710.00**

A.4 Additional Experimental Results

A.4.1 Experiment 1: Single-Phase DP-SGD Baseline

This experiment establishes the baseline against which all subsequent dual-phase results are compared.

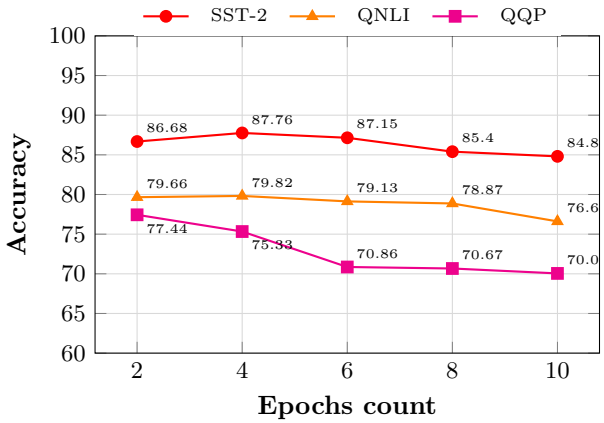

Figure 4: Accuracy variation vs epochs count for each DS

In this experiment, the model undergoes a single phase of fine-tuning using DP-SGD with standard clipping using (ϵ, δ) values as declared above on the original training datasets for the three tasks (SST-2, QNLI, and QQP). During this stage, the per-sample gradients are clipped to an ℓ_2-norm bound of 0.1 to control sensitivity, after which Gaussian noise is added to ensure differential privacy guarantees. The clipping is applied at the gradient level rather than on the input space, following the conventional DP-SGD formulation. The number of training epochs is varied to determine the optimal epoch count that yields the highest accuracy for each dataset. This setup serves as the baseline for privacy-preserving fine-tuning approaches against which subsequent experiments are compared. The results, illustrated in Figure 4, show the relationship between accuracy and the number of training epochs for the SST-2, QNLI, and QQP datasets. The optimal epoch counts observed were four for SST-2 with an accuracy of **87.76%**, four for QNLI with an accuracy of **79.82%**, and two for QQP with an accuracy of **77.44%**.

This pattern can be partly explained by the impact of dataset size on training under DP-SGD: larger datasets often lead to improved model utility within a fixed privacy budget, as each training step benefits from more diverse examples, potentially requiring fewer epochs to converge. However, excessive training can harm the learning process, leading to overfitting and a subsequent decrease in accuracy.

A.4.2 Per-Example Gradient Sensitivity Statistics

To analyze how the proposed two-phase training strategy reshapes optimization under differential privacy, we report per-example gradient sensitivity statistics for all evaluated datasets (SST-2, QQP, and QNLI). These statistics are collected during the DP training phase and summarize the distribution of per-example gradient norms together with the fraction of gradients affected by clipping.

Table 12 reports four DP-relevant measures: the mean, median, and 95th percentile ($p95$) of per-example gradient norms, as well as the clipping rate. Collectively, these metrics capture (i) the overall scale of gradient sensitivity, (ii) the typical per-example contribution, (iii) the heaviness of the gradient tail, and (iv) the degree of optimization distortion introduced by DP clipping.

Across SST-2 and QQP, Exp2 substantially reduces the mean and $p95$ gradient norms while also lowering the clipping rate. This indicates a shift toward a lower-sensitivity, lighter-tailed gradient regime, in which fewer examples are aggressively rescaled by the clipping operation. Such behavior directly limits clipping-induced bias and is consistent with the observed gains in validation accuracy.

QNLI exhibits a distinct but still informative behavior under Exp2. While absolute gradient magnitudes are drastically reduced (mean 3.92, $p95$ 9.05), all gradients exceed the fixed clipping threshold ($C = 0.1$), resulting in a 100% clipping rate. This outcome reflects a scale mismatch between the dataset-specific gradient distribution and the globally fixed clipping bound, rather than instability or heavy-tailed behavior. Notably, the narrow gap between the median and $p95$ norms indicates a highly concentrated and well-controlled

Table 12: Per-example gradient sensitivity statistics during the DP phase across datasets. Exp2 consistently reduces gradient magnitudes and tail heaviness relative to Exp1, with dataset-dependent effects on clipping behavior under a fixed clipping bound.

Dataset	**Exp**	**Mean**	**Median**	$p95$	**Clip (%)**
SST-2	Exp1	302.9	1.72	1321.1	41.14
	Exp2	43.7	0.0036	191.3	15.82
QQP	Exp1	310.2	2.06	1796.1	49.96
	Exp2	98.3	0.31	518.8	30.05
QNLI	Exp1	298.8	3.75	1747.9	52.14
	Exp2	3.92	3.46	9.05	100.00

gradient distribution, implying that gradients are uniformly rescaled rather than selectively distorted by clipping.

Overall, these results show that Exp2 consistently reshapes the per-example gradient landscape toward lower sensitivity and reduced tail risk across datasets. At the same time, they highlight how dataset-dependent gradient scales interact with fixed DP hyperparameters, emphasizing the importance of distributional diagnostics beyond clipping rate alone.

A.5 Membership Inference Attack Evaluation

To provide a complete assessment of the privacy guarantees of DualTune-GhostDP, this section presents the full Membership Inference Attack (MIA) evaluation across different stages of the training pipeline. MIA is a privacy attack in which an adversary attempts to determine whether a given data sample was included in the model's training set based solely on the model's outputs. Models that memorize training data tend to exhibit higher MIA success rates, whereas privacy-preserving mechanisms, such as dataset sanitization and differential privacy, aim to reduce attack performance toward random guessing ($\approx 50\%$). In all experiments, we adopt a strong adversarial setting consistent with standard practice in DP-SGD privacy analysis. Specifically, we simulate a *direct membership inference attack* on the SST-2 dataset, which is widely considered one of the strongest threat models for evaluating privacy leakage in classification settings Shokri et al. (2017); Carlini et al. (2022a). In this setting, the adversary (i) has access to the model's output posteriors and (ii) knows the true membership status of samples used during training. This configuration provides an upper bound on the adversary's advantage and represents a worst-case estimate of potential privacy leakage.

Table 13: Comparison of MIA Evaluation Metrics Across Model Pipelines on SST-2.

Metric	**Post-Sanitization**	**DualTune-GhostDP**	**Baseline Model**
Accuracy	0.512	0.508	0.501
AUC	0.5150	0.5033	0.5909
Precision	0.5071	0.5045	0.5082
Recall	0.8741	0.9189	0.1206
F1-Score	0.6418	0.6514	0.1950

A.5.1 MIA on Stage 1 Model: Post-Sanitization Model

The first evaluation is conducted on the model trained solely on the sanitized dataset, prior to applying any differentially private optimization. Under this setting, the attack achieves an accuracy of 0.512 and an AUC

of 0.5150, both of which are effectively equivalent to random guessing. Although the recall appears high, this behavior is attributed to the attacker over-predicting the 'member" class, which results in a large number of false positives as reflected in the confusion matrix. Importantly, the precision remains close to 0.50, indicating that the adversary cannot reliably distinguish between member and non-member samples. These results suggest that dataset sanitization alone already provides a strong degree of protection by reducing memorization and limiting privacy leakage. Consequently, the post-sanitization model serves as a robust initialization point for subsequent private fine-tuning.

A.5.2 MIA on Proposed Approach: DualTune-GhostDP Model

We next evaluate the full DualTune-GhostDP pipeline, which combines dataset sanitization with DP-SGD using Ghost Clipping and the EW privacy accountant. Under the same direct-MIA setting, the resulting model achieves an attack accuracy of 0.508 and an AUC of 0.5033. These values correspond to pure random guessing, demonstrating that the adversary gains no meaningful advantage even under this strong threat model. The near-random attack performance confirms that integrating differential privacy on top of sanitized training further suppresses membership leakage. Compared to the post-sanitization model, the slight reduction in attack success indicates that DP-SGD with Ghost Clipping provides an additional layer of protection, reinforcing the privacy guaranties of the proposed framework.

A.5.3 MIA on Baseline: Single-Phase DP-SGD Model

For comparison, we evaluate a baseline model trained using standard single-phase DP-SGD without dataset sanitization or Ghost Clipping. Despite operating under the same privacy budget, this baseline exhibits noticeably weaker and less stable privacy behavior. In one representative run, the attack AUC reaches 0.5909, indicating that the adversary can more reliably distinguish training samples compared to both the post-sanitization model and DualTune-GhostDP. Even in configurations where the baseline AUC approaches 0.50, the attack metrics remain highly unstable, with extremely low recall and F1 scores. This instability highlights the baseline model's reduced robustness to membership inference and suggests that standard DP-SGD alone may be insufficient to consistently suppress membership leakage in practice.

A.5.4 Discussion

Across all evaluated settings, DualTune-GhostDP consistently provides the strongest protection against membership inference. Dataset sanitization substantially reduces memorization in the initial training stage, while the subsequent application of DP-SGD with Ghost Clipping further suppresses residual leakage. In contrast, the single-phase DP-SGD baseline demonstrates higher attack success and instability, despite operating under the same privacy budget. Overall, these results demonstrate that the combined use of sanitization and differentially private fine-tuning yields privacy guarantees that are indistinguishable from random guessing under a worst-case direct-MIA setting, while simultaneously maintaining higher model utility. This highlights the advantage of the proposed DualTune-GhostDP framework in achieving a more favorable privacy–utility trade-off than existing baselines.

A.6 Future Work: Architectural Considerations for Decoder-Only LLMs

While our experiments focus on encoder-only backbones for consistency with prior DP fine-tuning work, extending DualTune-GhostDP to modern decoder-only LLMs (e.g., GPT-style architectures) introduces additional architectural considerations. In particular, decoder-only models rely heavily on parameter sharing mechanisms such as tied input and output embeddings and repeated reuse of projection matrices across layers which complicate the assumptions underlying Ghost Clipping's efficiency gains. Although Ghost Clipping remains theoretically applicable under DP-SGD, the presence of extensive parameter sharing can mask its practical memory and runtime benefits by collapsing multiple gradient paths onto shared parameters. We are actively investigating decoder-only variants that explicitly account for these interactions, including controlled untying and architectural adaptations, and leave a full empirical study of decoder-only LLMs to future work.