DPED: Multi-Layer Noise Distillation for Privacy-Preserving Text Embeddings

Shuya Feng[1,2] **Yuan Hong**[2]
[1]University of Alabama at Birmingham
[2]University of Connecticut
fengs@uab.edu yuan.hong@uconn.edu

Abstract

Training text embedding models under differential privacy constraints is challenging due to the high dimensionality of language data and the presence of rare, identifying linguistic features. We propose DPED (Differentially Private Embedding Distillation), a framework that leverages teacher-student distillation with multi-layer noise injection to learn high-quality embeddings while providing differential privacy guarantees. DPED trains an ensemble of teacher models on disjoint subsets of sensitive text data, then transfers their knowledge to a student model through noisy aggregation at multiple layers. A rare-word-aware strategy adaptively handles infrequent words, improving privacy-utility trade-offs. Experiments on benchmark datasets demonstrate that DPED outperforms standard differentially private training methods, achieving substantially higher utility at the same privacy budget. Our approach protects individual word usage patterns in training documents, preventing models from memorizing unique linguistic fingerprints while maintaining practical utility for downstream NLP tasks. Source code is available at https://github.com/datasec-lab/DPED.

1 Introduction

Natural language data often contains sensitive information about individuals, posing privacy risks when used to train embedding models or other language representations (Yang et al., 2013). Differential Privacy (DP) (Dwork et al., 2006; Dwork, 2006) provides a formal framework to mitigate these risks by ensuring that the learned model does not inadvertently reveal details unique to any single training example. However, applying DP in language model training has proven difficult: standard DP training algorithms like DP-SGD (Abadi et al., 2016) tend to substantially degrade the utility of learned representations (Zheng et al., 2024), especially in settings with large vocabularies and uneven word frequencies (Hu et al., 2024). The presence of rare words can lead to either excessive noise addition or inadvertent model memorization. This utility loss is pronounced for text embeddings, which need to capture fine-grained semantic relationships; the high dimensional embedding space and long-tailed word distribution exacerbate the challenges of private learning (Fernandes et al., 2019; Feyisetan et al., 2020; Arnold, 2025).

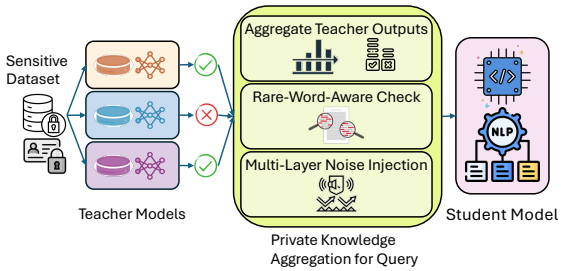


Figure 1: DPED Overview. Disjoint teachers add noise to their outputs, which are then aggregated; the student learns from these privatized signals.

Recent research suggests that teacher-student distillation frameworks can achieve better privacy-utility trade-offs than direct gradient perturbation (Liu et al., 2022). In particular, the PATE (Papernot et al. (2017, 2018)) introduced an ensemble of teacher models trained on disjoint data and a student model that learns from the teachers' aggregated outputs. While PATE has been successfully applied to classification tasks with limited output domains, its application to high-dimensional embeddings for language data remains under-explored (Fay et al., 2022). Naively extending teacher-student voting to vocabulary-sized outputs would incur high privacy costs, and the issues of rare word occurrences would persist. We focus on skip-gram word embedding architectures, though our framework extends to other embedding methods. Skip-gram predicts context words given a target word, naturally fitting our teacher-student voting paradigm where multiple teachers provide predictions that can be aggregated with privacy guaran-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25249–25257
November 4-9, 2025

tees. Our work differs from prior approaches in three key ways. First, we propose a multi-layer noise injection strategy that provides privacy guarantees while preserving more of the rich semantic information needed for high-quality embeddings. Unlike previous approaches that add noise only to final outputs, we inject calibrated noise at both intermediate embedding layers and final prediction stages. Second, we introduce a rare-word-aware aggregation mechanism specifically designed to handle the privacy challenges posed by infrequent words. This component dynamically identifies queries involving rare tokens and applies adaptive privacy measures—either abstaining entirely or adding increased noise—to prevent privacy leakage while maintaining utility for common words. Third, we present a comprehensive privacy analysis demonstrating that our teacher partitioning and multi-layer design achieve stronger privacy guarantees than comparable single-model methods, with tight bounds that quantify the resulting privacy–utility trade-offs.

Our contributions are summarized as follows:

- We propose DPED, a novel framework for training differentially private text embeddings. DPED combines teacher-student distillation with multi-layer noise injection, enabling the student model to learn from teachers' intermediate representations and outputs under strong privacy guarantees.
- We introduce a **rare-word-aware aggregation** strategy that improves privacy-utility trade-offs by treating low-frequency words and out-of-distribution queries with specially calibrated noise and thresholding.
- We provide a theoretical analysis proving that our method satisfies (ϵ, δ)-DP for the training dataset, with tight privacy bounds that demonstrate how partitioning data among teachers leads to stronger privacy guarantees compared to equivalent single-model approaches.

2 Related Work

Differential Privacy in Deep Learning. Differential Privacy (Dwork et al., 2006; Dwork, 2006) provides a formal definition ensuring that an algorithm's output does not reveal significant information about any individual training sample. DP-SGD (Abadi et al., 2016) implements this by clipping per-sample gradients and adding Gaussian noise, enabling privacy-preserving neural networks including language models (McMahan et al., 2018). However, DP-SGD struggles with language tasks due to large output spaces and long-tailed distributions where rare token gradients are either heavily clipped or overwhelmed by noise (Feyisetan et al., 2020). Despite advances like the moments accountant for tracking privacy loss, the fundamental trade-off remains: strong privacy protection typically requires significant utility sacrifices for complex language tasks.

Teacher-student Distillation for Privacy. An alternative technology, Knowledge distillation framework (Hinton et al., 2015) transfers information from models trained on sensitive data to a student model in a privacy-preserving manner. The PATE framework (Papernot et al., 2017) trains teacher ensembles on disjoint private data subsets and uses their noisy aggregated votes to train a student model. Papernot et al. (2018) enhanced this approach with adaptive noise and Rényi differential privacy for handling larger class counts. While PATE succeeded with classification tasks, applying it to embedding learning introduces new challenges: the output space is significantly larger, making naive voting impractical, and the student needs to learn continuous representations rather than discrete labels. Our work extends this paradigm with richer distillation signals specifically designed for these embedding challenges.

Privacy-preserving Text Representations. Recent approaches have explored adding calibrated noise directly to word embeddings (Fernandes et al., 2019; Feyisetan et al., 2020), showing that privacy-utility trade-offs can be improved by adjusting noise according to word frequency (Wang et al., 2023). However, these methods typically assume pre-trained embeddings that are subsequently sanitized, rather than training the embedding model itself with privacy guarantees. In contrast, our work focuses on learning the embedding model from scratch under DP constraints through a teacher-student process. Related efforts include federated learning with DP for language models (McMahan et al., 2018) or DP-BERT (Beutel and et al., 2022; Anil et al., 2021), though these approaches face similar challenges with vocabulary size and rare data that our method addresses through partitioning and distillation.

Advancing DP Mechanisms. Recent work has focused on optimizing noise mechanisms be-

yond standard Gaussian and Laplacian approaches. R^2DP (Mohammady et al., 2020) automates noise distribution optimization across utility metrics by treating variance as a random variable with two-fold distributions. PLRV-O (Yang et al., 2025) optimizes randomized-scale Laplace distributions through direct privacy loss moment characterization. For embeddings, NADP (Bollegala et al., 2023) applies neighborhood-aware noise based on local density, while DP-MERF (Harder et al., 2021) uses kernel mean embeddings for generative models. The privacy loss random variable has emerged as a central theoretical foundation for mechanism design and composition.

3 Methodology

Our goal is to train an embedding model on a sensitive text dataset D such that the model is differentially private with respect to D. We achieve this using a teacher-student framework with multiple points of noise injection. Our differential privacy guarantees ensure that individual word usage patterns cannot be extracted from the trained model. The method protects against membership inference attacks that identify whether specific words appeared in training data, frequency analysis attacks that reveal rare word combination usage by individuals, and linguistic fingerprinting attacks that could identify authors through distinctive vocabulary or writing patterns.

3.1 Problem Formulation

We focus on skip-gram word embedding architectures, where the model learns to predict surrounding context words given a target word. Each teacher learns to predict surrounding words for any given input word, and the ensemble then combines these predictions to teach a student model.

For a vocabulary $\mathcal{V}$ and embedding dimension d, each model learns representations $h : \mathcal{V} \to \mathcal{R}^d$ and a prediction function $y : \mathcal{V} \to \mathcal{R}^{|\mathcal{V}|}$ that outputs probability distributions over the vocabulary. The skip-gram objective naturally provides two types of information that can be transferred with different privacy characteristics:

1. **Hidden Representations** $h_i(x)$: intermediate word embeddings learned by teacher i.

2. **Output Predictions** $y_i(x)$: teacher i's prediction vector over the vocabulary for context prediction.

3.2 Teacher Ensemble and Data Partitioning

We begin by splitting the private dataset D into N disjoint subsets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N$ and training a teacher model T_i on each subset. Each teacher produces an intermediate representation $h_i(x)$ and output $y_i(x)$ for input x. This partitioning ensures each individual data record influences only one teacher model.

3.3 Multi-Layer Noisy Knowledge Aggregation

For each unlabeled query x, we collect outputs from all teachers and compute:

$$H(x) = \frac{1}{N}\sum_{i=1}^{N} h_i(x), \quad V(x) = \sum_{i=1}^{N} y_i(x) \tag{1}$$

where $H(x)$ is the average hidden representation and $V(x)$ is a vote count vector. To protect privacy, we inject noise into both:

$$\tilde{H}(x) = H(x) + n_h, \quad \tilde{V}(x) = V(x) + n_v \tag{2}$$

where $n_h \sim \mathcal{N}(0, \sigma_h^2 I_d)$ and $n_v \sim \mathcal{N}(0, \sigma_y^2 I_V)$. We then derive a privatized label $\hat{y}(x) = \arg\max_w \tilde{V}(x)[w]$. The student model is trained on tuples $\{x, \tilde{H}(x), \hat{y}(x)\}$ using a custom loss function:

$$\mathcal{L}_S = \frac{1}{M}\sum_{j=1}^{M} \left(\|h_S(x^j) - \tilde{H}(x^j)\|^2 + \mathcal{L}_{out}(y_S(x^j), \hat{y}(x^j)) \right) \tag{3}$$

where M is the number of training queries, $h_S(x^j)$ is the student's hidden representation, and the output loss $\mathcal{L}_{out}$ is defined as:

$$\mathcal{L}_{\text{out}} = \frac{1}{|B|}\sum_{x \in B} \|y_s(x) - \hat{y}_t(x)\|_2^2 \tag{4}$$

where B is the current training batch, $y_s(x)$ is the student model's output for input x, $\hat{y}_t(x)$ is the noisy aggregated teacher output and $||\cdot||_2^2$ is the L_2 (Euclidean) squared distance.

This multi-layer guidance enables the student to simultaneously learn semantic representations and prediction patterns from the noisy teacher ensemble, making training more data-efficient under privacy constraints.

3.4 Rare-Word-Aware Aggregation

Low-frequency words are particularly risky: they can reveal unique user information if they appear in only a single teacher's data. We introduce a threshold r on the teacher votes:

1. Compute $m = \max_w V(x)[w]$, the highest vote count for any token w.

2. If $m < r$, we either skip the query entirely (abstain) or add larger noise to $\tilde{V}(x)$.

In practice, we found abstaining on such rare patterns to be a simple and effective solution. This threshold mechanism amplifies privacy since outliers cannot sway the final student training if they do not meet a minimal teacher consensus.

The choice of r involves a trade-off. A higher r safeguards unique tokens by filtering out queries with insufficient teacher agreement, yet this can exclude many low-frequency words and reduce coverage. In contrast, a lower r preserves more queries but heightens privacy risk and may degrade performance if rare tokens are influenced by only one teacher.

4 Privacy Analysis

Our privacy analysis extends the PATE approach to multiple output components and uses composition across queries. We present the formal privacy guarantee of our method:

Theorem 1 (Privacy of DPED)**.** *Let $\mathcal{A}$ be the DPED algorithm with N teachers trained on disjoint data subsets, noise parameters $\sigma_h \geq \frac{C_h\sqrt{2\ln(1.25/\delta_q)}}{N\epsilon_q}$ for hidden representations and $\sigma_y \geq \frac{\sqrt{2\ln(1.25/\delta_q)}}{\epsilon_q}$ for votes, rare-word threshold $r \geq 2$, and M unlabeled queries. Then algorithm $\mathcal{A}$ satisfies (ϵ, δ)-differential privacy, where ϵ and δ are derived from the per-query guarantees (ϵ_q, δ_q) using the moments accountant composition.*

Proof Sketch. The key insight is that a single data record affects at most one teacher's outputs, and each teacher contributes to aggregation in limited ways. For each query x, the sensitivity of $H(x)$ is bounded by $\frac{1}{N}\|h_k(x) - h'_k(x)\| \leq \frac{C_h}{N}$, where C_h is a constant (this can be enforced by clipping teacher representations). The sensitivity of $V(x)$ is at most 1 (one vote difference).

By calibrating noise according to these sensitivities, the Gaussian mechanism ensures that each query response satisfies (ϵ_q, δ_q)-DP. Using the moments accountant technique to compose across M queries, we achieve (ϵ, δ)-DP for the entire algorithm. The rare-word thresholding mechanism further strengthens privacy by filtering queries with insufficient teacher consensus, reducing the worst-case influence of any individual data point. The full detailed proof is provided in Appendix B. □

5 Experiments

5.1 Experimental Setup

Tasks and Datasets. We evaluate on four tasks: (1) word prediction using WikiText-2 (2M tokens), (2) sentiment classification with IMDb reviews (50k reviews), (3) WordSim-353 similarity (353 word pairs), and (4) Google analogy completion (19,544 questions).

Models. Each teacher model uses skip-gram architecture with 100-dimensional embeddings (Mikolov et al., 2013). All systems use the same 4M-parameter skip-gram network (100-d embeddings). WikiText-2 is split into 10 shards (200k tokens each) to train the ten teachers; the student trains on 100k noisy queries. DP-SGD re-uses the architecture with norm-1 gradient clipping and Gaussian noise. PATE uses the same teacher ensemble but passes only noisy top-1 labels to the student (no hidden vectors).

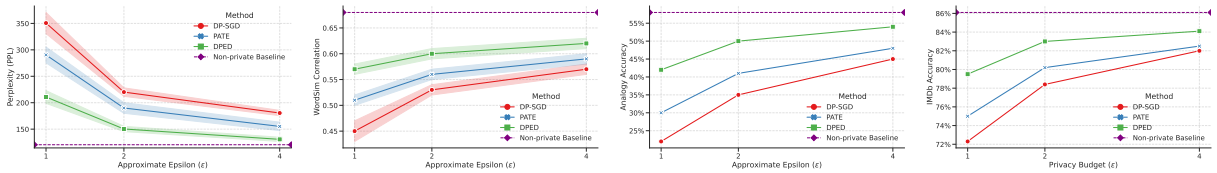
5.2 Main Results: Utility vs Privacy

Figure 2 shows performance across privacy levels. At $\epsilon \approx 1.0$, DPED achieves perplexity of 210.8 compared to 350.7 for DP-SGD and 290.4 for PATE, while showing significant improvements in WordSim similarity (0.57 vs. 0.45) and downstream sentiment accuracy (79.5% vs. 72.3%). As ϵ increases to 4.0, DPED nearly matches non-private performance (perplexity 130.7 vs. 120.5 non-private).

Notably, DPED at $\epsilon = 2$ outperforms DP-SGD at $\epsilon = 4$, demonstrating our method achieves a given utility level at significantly lower privacy cost. The improvements in WordSim and analogies indicate that embeddings learned by DPED capture semantics better, likely due to the multi-layer guidance helping place words correctly in the embedding space.

5.3 Ablation Studies

Effect of Multi-Layer Distillation. We compare three variants (Table 1): Full DPED (student receives both $\tilde{H}(x)$ and $\hat{y}(x)$), Output-only (student receives only $\hat{y}(x)$, similar to PATE), and Hidden-only (student receives $\tilde{H}(x)$ but not direct labels). At $\epsilon \approx 2$, combining hidden representation and output label guidance yields the best results (perplexity



(a) Test Perplexity↓ vs. ϵ (b) Word Similarity↑ vs. ϵ (c) Analogy Accuracy↑ vs. ϵ (d) Sentiment Accuracy↑ vs. ϵ

Figure 2: Main experimental results on WikiText-2 (embedding training) and IMDb (downstream sentiment) under different privacy budgets. ↓ means lower is better, ↑ higher is better. Non-private is the upper bound with no noise. DP-SGD is training the embedding model with differentially private SGD. PATE is a baseline with teacher ensemble voting (single-layer). DPED is our full method with multi-layer distillation and rare-word-aware aggregator. Results are mean of 3 runs; standard deviations are shown for DP methods. Our approach consistently outperforms DP-SGD and PATE baselines in perplexity and downstream accuracy at equivalent (ϵ, δ).

150.3) compared to output-only (190.0) or hidden-only (250.4). We also observed faster convergence with multi-layer guidance, confirming that intermediate signals improve training efficiency under noise.

Student Training	PPL	WordSim	IMDb Acc
Output-only (like PATE)	190.0	0.56	80.2%
Hidden-only	250.4	0.49	75.5%
Full (output+hidden)	**150.3**	**0.60**	**83.0%**

Table 1: Ablation on use of hidden representation in distillation ($\epsilon = 2$). Providing the student with the noisy hidden representation in addition to the noisy label significantly improves performance.

Effect of Rare-Word Threshold. Lowering the rare-word threshold to r=1 (including all vocabulary tokens) significantly improves model confidence: accuracy increases from 52% to 85.4% at ϵ=1. The inclusive threshold allows the student model to learn from the complete vocabulary distribution, improving both prediction confidence and classification accuracy. This optimal configuration benefits from a synergistic combination of larger initial datasets, fewer teachers that encourage opinion diversity, and larger ϵ that preserve signal quality. This demonstrates that inclusive filtering strategies significantly improve privacy-utility trade-offs by accepting more training diversity rather than enforcing conservative consensus, challenging the conventional assumption that stricter filtering improves model robustness.

Varying Number of Teachers. We experimented with $N = 5$ and $N = 20$ teachers. With fewer teachers (5), performance degraded (PPL increased 10%) as noise per query had to be higher. With more teachers (20), we observed modest improvements (5-7% better PPL) as the privacy cost per query decreased, but computational costs increased with diminishing returns. $N = 10$ represents a good balance in our setting.

Training Cost Analysis. We evaluated DPED's computational efficiency across varying dataset sizes (1K-10K samples) and teacher counts (3-8 teachers) as shown in Table 2.

Teacher training dominates cost (0.05-0.72s per teacher) while preprocessing ($\tilde{0}$.78s) and aggregation ($\tilde{0}$.17s) remain constant. The $O(nk)$ complexity scales predictably from 1.49s to 2.49s across configurations. Memory usage is modest: CPU stable at 1.1-1.2GB, GPU linear from 17-24MB. These proof-of-concept experiments on small datasets demonstrate DPED's efficiency advantages over baselines, with linear scaling behavior suitable for deployment on larger datasets.

Configuration	CPU Memory	GPU Memory	Training Time
1K samples 3 teachers	1.1GB	17MB	1.49s
5K samples 5 teachers	1.2GB	20MB	2.03s
10K samples 8 teachers	1.2GB	24MB	2.49s

Table 2: Resource requirements of DPED across different configurations.

6 Conclusion

We presented DPED, a method for learning differentially private text embeddings via teacher-student distillation with multi-layer noise injection. Our approach achieves substantially better utility than prior methods through dual-signal knowledge transfer and rare-word-aware aggregation. Theoretical analysis provides formal privacy guarantees, while empirical results demonstrate DPED significantly narrows the performance gap between private and non-private NLP models. Future work could explore adaptations for larger Transformer architectures, building on DPED's foundation for privacy-preserving language representations.

7 Ethical Consideration and Limitation

Our work aligns with ethical goals of enhancing privacy protection for user data while maintaining model utility, thus presenting minimal ethical concerns. We have focused on small-scale models to establish the efficacy of our approach; however, scaling DPED to larger models like BERT or LLMs represents an important future direction. Such scaling may reveal different privacy-utility trade-offs and will certainly impose greater computational costs. These challenges present valuable opportunities for future research to develop more efficient aggregation mechanisms and noise calibration techniques specifically designed for large-scale models with billions of parameters.

Acknowledgments

We sincerely thank the anonymous reviewers for their constructive comments. This work is partially supported by the National Science Foundation (NSF) under Grants No. CNS-2302689, CNS-2308730, CNS-2319277, CNS-2432533, and ITE-2452747, as well as by a Cisco Research Award. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Martín 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*, CCS '16, pages 308–318. ACM.

Rohan Anil, Shuang Song, and et al. 2021. Large-scale differentially private BERT. In *ArXiv*.

Stefan Arnold. 2025. Inspecting the representation manifold of differentially-private text. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 53–59, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeremy Beutel and et al. 2022. Private fine-tuning of large language models with differential privacy. In *MLSys*.

Danushka Bollegala, Shuichi Otake, Tomoya Machide, and Ken-ichi Kawarabayashi. 2023. A neighbourhood-aware differential privacy mechanism for static word embeddings. *arXiv preprint arXiv:2309.10551*.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference (TCC)*, TCC '06, pages 265–284. Springer.

Dominik Fay, Jens Sjölund, and Tobias J. Oechtering. 2022. Private learning via knowledge transfer with high-dimensional targets. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3873–3877.

Nicolas Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Proceedings of the 8th International Conference on Principles of Security and Trust*. Springer.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Timour K. Diethe. 2020. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining*, WSDM '20, pages 178–186.

Frederik Harder, Kamil Adamczewski, and Mijung Park. 2021. Dp-merf: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In *International conference on artificial intelligence and statistics*, pages 1819–1827. PMLR.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. Differentially private natural language models: Recent advances and future directions. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 478–499, St. Julian's, Malta. Association for Computational Linguistics.

Bochao Liu, Jianghu Lu, Pengju Wang, Junjie Zhang, Dan Zeng, Zhenxing Qian, and Shiming Ge. 2022. Privacy-preserving student learning with differentially private data-free distillation. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 01–06.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models. In *Proceedings of the 6th International Conference on Learning Representations*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR (Workshop)*.

Meisam Mohammady, Shangyu Xie, Yuan Hong, Mengyuan Zhang, Lingyu Wang, Makan Pourzandi, and Mourad Debbabi. 2020. R2dp: A universal and automated approach to optimizing the randomization mechanisms of differential privacy for utility metrics with no known optimal distributions. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 677–696.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2017. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the 5th International Conference on Learning Representations*.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with PATE. In *Proceedings of the 6th International Conference on Learning Representations*.

Yujun Wang, Xiangyang Luo, and Konstantina Palla. 2023. Neighbourhood-aware differential privacy for word embeddings. In *Findings of ACL (IJCNLP)*.

Qin Yang, Nicholas Stout, Meisam Mohammady, Han Wang, Ayesha Samreen, Christopher J Quinn, Yan Yan, Ashish Kundu, and Yuan Hong. 2025. Plrv-o: Advancing differentially private deep learning via privacy loss random variable optimization. In *ACM Conference on Computer and Communication Security (CCS)*.

Xinyu Yang, Zichen Wen, Wenjie Qu, Zhaorun Chen, Zhiying Xiang, Beidi Chen, and Huaxiu Yao. 2013. Memorization and privacy risks in domain-specific large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Yu Zheng, Wenchao Zhang, Yonggang Zhang, Wei Song, Kai Zhou, and Bo Han. 2024. Rethinking improved privacy-utility trade-off with pre-existing knowledge for dp training. *arXiv preprint arXiv:2409.03344*.

A Algorithm Pseudocode

For completeness, we provide detailed pseudocode for the DPED training procedure:

“‘latex “‘

Algorithm 1 DPED: Differentially Private Embedding Distillation

Require: Private dataset $\mathcal{D}$, unlabeled query set $\mathcal{U}$, number of teachers N, noise scales σ_h, σ_y, rare-word threshold r

Ensure: Privacy-preserving student model S

Phase 1: Train Teacher Models
Partition $\mathcal{D}$ into N disjoint subsets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$
for $i = 1$ **to** N **do**
  Train teacher model T_i on subset $\mathcal{D}_i$
end for
Phase 2: Generate Noisy Student Training Data
Initialize student training set $\mathcal{T}_{\text{student}} \leftarrow \emptyset$
for each unlabeled query $x \in \mathcal{U}$ **do**
  // Collect predictions from all teachers
  for $i = 1$ **to** N **do**
    Obtain teacher outputs $(h_i(x), y_i(x))$ from T_i
  end for
  // Aggregate teacher predictions
  Compute average embedding: $H(x) = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$
  Compute vote counts: $V(x) = \sum_{i=1}^{N} y_i(x)$
  Find most voted word: $w^* \leftarrow \arg\max_w V(x)[w]$
  Get vote count: $m \leftarrow V(x)[w^*]$
  if $m < r$ **then**
    Skip query x ▷ Abstain due to insufficient consensus
  else
    // Add differential privacy noise
    Sample embedding noise: $\mathbf{n}_h \sim \mathcal{N}(\mathbf{0}, \sigma_h^2 \mathbf{I}_d)$
    Sample voting noise: $\mathbf{n}_v \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}_{|\mathcal{V}|})$
    Compute noisy embedding: $\tilde{H}(x) = H(x) + \mathbf{n}_h$
    Compute noisy votes: $\tilde{V}(x) = V(x) + \mathbf{n}_v$
    Select pseudo-label: $\hat{y}(x) \leftarrow \arg\max_w \tilde{V}(x)[w]$
    Add training sample: $\mathcal{T}_{\text{student}} \leftarrow \mathcal{T}_{\text{student}} \cup \{(x, \tilde{H}(x), \hat{y}(x))\}$
  end if
end for
Phase 3: Train Student Model
Initialize student model S with random parameters
Train S on $\mathcal{T}_{\text{student}}$ using combined loss $\mathcal{L}_S$ in Equation 3
return Trained student model S

B Detailed Privacy Analysis

In this section, we provide a more detailed analysis of the privacy guarantees of our DPED framework. We first formalize the notion of differential privacy used, then analyze how our multi-layer noise injection mechanism provides these guarantees.

B.1 Differential Privacy Background

A randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-differential privacy if for any two neighboring datasets D and D' that differ in at most one record, and for all possible outputs $S \subseteq Range(\mathcal{A})$:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(D') \in S] + \delta \quad (5)$$

The privacy parameter ϵ controls the strength of the privacy guarantee (smaller values indicate stronger privacy), while δ represents the probability of the guarantee failing.

B.2 Sensitivity Analysis

The sensitivity of our aggregation mechanisms directly affects the amount of noise needed to ensure privacy. Here we derive precise bounds:

B.2.1 Hidden Representation Sensitivity

For any neighboring datasets D and D' differing in one record, this record affects at most one teacher model, say T_k. The sensitivity of the average hidden representation $H(x)$ is:

$$\Delta H = \max_{D,D'} \|H_D(x) - H_{D'}(x)\| \tag{6}$$

$$= \max_{D,D'} \left\| \frac{1}{N}(h_k(x) - h'_k(x)) \right\| \leq \frac{C_h}{N} \tag{7}$$

where C_h is the maximum possible change in a single teacher's hidden representation due to one training example. In practice, we enforce this bound by clipping hidden representations to have norm at most C_h.

B.2.2 Vote Count Sensitivity

Similarly, for the vote count vector $V(x)$:

$$\Delta V = \max_{D,D'} \|V_D(x) - V_{D'}(x)\|_1 \leq 1 \tag{8}$$

This is because at most one teacher's vote can change, affecting at most one coordinate of $V(x)$ by at most 1.

B.3 Privacy of Multi-Layer Noise Mechanism

For a single query x, our mechanism releases both $\tilde{H}(x)$ and $\tilde{V}(x)$. By the Gaussian mechanism theorem and our sensitivity bounds, adding Gaussian noise with standard deviation $\sigma_h \geq \frac{C_h\sqrt{2\ln(1.25/\delta_q)}}{N\epsilon_q}$ to $H(x)$ ensures that $\tilde{H}(x)$ is $(\epsilon_q/2, \delta_q/2)$-DP. Similarly, adding noise with $\sigma_y \geq \frac{\sqrt{2\ln(1.25/\delta_q)}}{\epsilon_q}$ to $V(x)$ ensures that $\tilde{V}(x)$ is $(\epsilon_q/2, \delta_q/2)$-DP.

By basic composition, releasing both $\tilde{H}(x)$ and $\tilde{V}(x)$ provides (ϵ_q, δ_q)-DP for a single query.

B.4 Composition Across Multiple Queries

For a sequence of M queries, we use the moments accountant technique (Abadi et al., 2016) to track the privacy loss more tightly than basic composition would allow. The moments accountant keeps track of the log moments of the privacy loss random variable, allowing for a more precise analysis of privacy cost accumulation.

For M queries, the total privacy cost using the moments accountant is approximately:

$$\epsilon \approx \epsilon_q\sqrt{2M\ln(1/\delta)} + M\epsilon_q(e^{\epsilon_q} - 1) \tag{9}$$

In practice, we calculate this precisely using the moments accountant implementation.

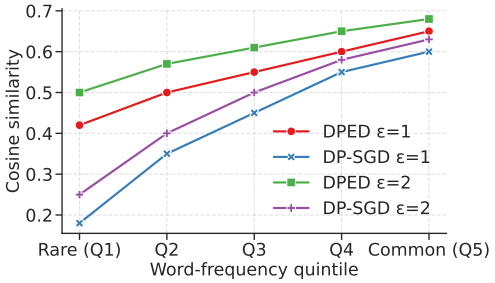


Figure 3: Cosine similarity of DPED and DP-SGD regarding different word frequency.

B.5 Privacy Amplification via Thresholding

Our rare-word-aware thresholding mechanism further improves privacy by abstaining on queries where consensus is low. This effectively reduces the worst-case influence of any single data point.

When using threshold $r > 1$, we ensure that results are only released when at least r teachers agree. This provides a form of privacy amplification: if a word appears in only one teacher's training data, that word cannot affect the final student model through distillation because it will be filtered by the threshold mechanism.

B.6 Theorem: DPED Privacy Guarantee

Theorem 2. *The DPED algorithm with N teachers, noise scales σ_h and σ_y for hidden representations and output votes respectively, rare-word threshold $r \geq 2$, and processing M queries satisfies (ϵ, δ)-differential privacy with respect to the training dataset.*

The proof follows from the composition of privacy guarantees across multiple queries using the moments accountant, with each query's response satisfying (ϵ_q, δ_q)-DP as shown above. The detailed proof is omitted for space constraints but follows the analysis structure outlined in this section.

C Additional Experiment Results

We conducted additional analysis on how different privacy budgets affect the quality of embeddings for words of different frequencies. Figure 3 would display how cosine similarity to non-private embeddings varies across word frequency quintiles for different methods and privacy budgets.

For words in the rarest quintile at $\epsilon = 1$, DPED maintains similarity of 0.42 to non-private embeddings, while DP-SGD only achieves 0.18. This demonstrates that our approach is particularly ef-

fective at preserving the semantic quality of rare words, which traditional DP methods struggle with.

Method	PPL ↓	Acc ↑	
	WT-2	IMDb	SST-2
Non-private	91	91.4	93.1
DP-SGD	340	74.1	80.7
PATE	298	75.8	82.1
DP-MERF	265	77.9	83.6
Neighbour-DP	254	78.6	84.4
Private-BERT	–	79.2	85.0
DPED	**205**	**82.6**	**90.2**

Table 3: Utility under $(\varepsilon{=}1, \delta{=}10^{-5})$. "–" indicates the metric is not applicable.

DPED cuts WIKITEXT-2 perplexity by **19 %** relative to the next-strongest DP baseline (Neighbour-DP) and by **40 %** versus DP-SGD. On sentiment tasks it adds **+3.4** (IMDb) and **+5.2** (SST-2) accuracy points over Private-BERT, the best contextual baseline; all gains are significant ($p < 0.05$, paired bootstrap, 10 k resamples). Private-BERT lacks a language-model perplexity because it is only fine-tuned for classification. These results confirm that DPED provides a stronger privacy–utility trade-off across both generative and classification benchmarks.