Differentially Private Learning Needs Better Model Initialization and Self-Distillation

Ivoline C. Ngong[1] **Joseph P. Near**[1] **Niloofar Mireshghallah**[2]

[1]University of Vermont [2]University of Washington
kngongiv@uvm.edu,jnear@uvm.edu,niloofar@cs.washington.edu

Abstract

Differentially private SGD (DPSGD) enables privacy-preserving training of language models, but often reduces utility, diversity, and linguistic quality. We introduce DPRefine, a three-phase method that initializes a model using data synthesis from a small pre-trained LM with rigorous filtering, applies DP finetuning on private data, and performs self-distillation to refine outputs. This approach significantly outperforms vanilla DPSGD, with AlpacaEval preferring DPRefine's generations in 78.4% of cases across all datasets. Our analysis reveals that DPRefine reduces linguistic errors in generated text by 84.0%, mitigating grammar and spelling errors, commonly associated with DPSGD. It also reduces inconsistencies of non-private models, such as hallucinated details and misattributed quotes. We find that small models like GPT-2 can be effective for initialization and distillation, highlighting their potential in enabling scalable and efficient deployment of privacy-preserving language.

1 Introduction

Training machine learning models on private data offers significant potential for enhancing performance on domain-specific tasks, particularly in natural language processing (Li et al., 2021; Yu et al., 2021; Liu et al., 2023; Cummings et al., 2023). However, the use of sensitive information raises critical privacy concerns. Differentially Private Stochastic Gradient Descent (DPSGD) has emerged as a prominent technique to bound information leakage during model training by clipping gradients and adding calibrated noise during optimization (Abadi et al., 2016). While DPSGD provides strong privacy guarantees, the introduced noise and gradient modifications lead to significant challenges: decreased model utility (Yu et al., 2021; Ponomareva et al., 2022), less diverse text generation due to distribution smoothing (Bagdasaryan et al., 2019; Mireshghallah et al., 2022b), and notably, as our analysis reveals, increased linguistic errors in generated text (Table 1).

A common scenario in practice involves domain-specific tasks with limited private labeled data, where organizations aim to leverage existing pre-trained language models while preserving privacy. Simply applying DPSGD to fine-tune these models on private data often yields poor results, particularly when the private dataset is small (Tramèr et al., 2022; Mireshghallah et al., 2021). Recent work has shown that leveraging better hand-crafted features (Tramer and Boneh, 2020) or features from large pre-trained language models (Li et al., 2022, 2021) can improve the privacy-utility trade-off in differentially private learning. However, these approaches have limitations: smaller pre-trained models offer limited benefits, and fine-tuning larger models on private data may be infeasible due to proprietary concerns or infrastructure limitations. This raises a critical question: *Can we develop small, domain-specific language models that achieve high performance without requiring large private datasets or large, pre-trained models?*

In this paper we introduce DPRefine, building on the intuition that model performance under differential privacy heavily depends on model initialization and feature representation (Tramer and Boneh, 2020; Li et al., 2022). Additionally, recent advances in NLP have demonstrated the effectiveness of data synthesis (Liu et al., 2024; Flemings and Annavaram, 2024), filtering and self-distillation in improving the performance of smaller models(Jung et al., 2023; Bansal et al., 2024). Our method has three phases, depicted in Figure 1: In the first phase we create a strong initialization by generating high-quality synthetic data using a small pre-trained language model (e.g., GPT-2) with rigorous filtering - importantly, this phase operates entirely independently of any private data. In the second phase we apply DPSGD to fine-tune this initialized model on the private labeled data, representing the only stage where private data is accessed.

Finally, we apply self-distillation where the DP model generates new training data for further refinement, again without accessing the original private data. We evaluate DPRefine on three datasets across two domain-specific tasks: summarization, XSum (Narayan et al., 2018) and PubMed (National Library of Medicine, 2024), and paraphrasing, MRPC (Dolan and Brockett, 2005). Our experiments demonstrate that DPRefine significantly outperforms vanilla DPSGD, with AlpacaEval (Li et al., 2023) preferring DPRefine's generations in 78.38% of cases across all datasets and metrics.

Beyond standard metrics, we conduct a fine-grained manual error analysis of the generated text, constructing a taxonomy of both linguistic errors (e.g., grammar

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9009–9027
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

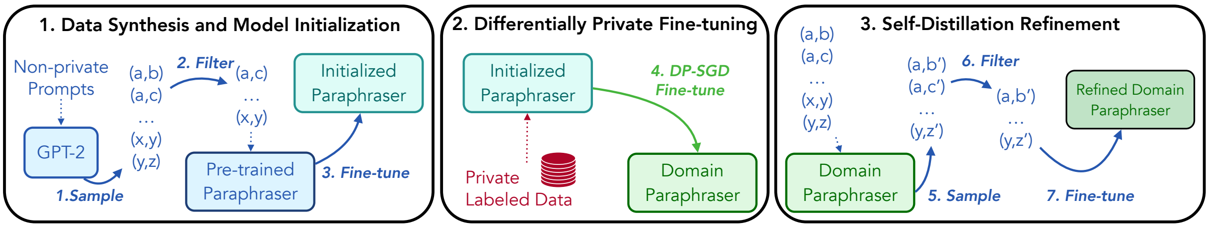


Figure 1: Overview of DPRefine's three-phase approach: Phase 1: Data Synthesis and Model Initialization generates synthetic training pairs using GPT-2, applies quality filtering, and performs initial fine-tuning on a pre-trained T5 (encoder-decoder model) to create an initialized paraphraser/summarize, all without accessing private data. Phase 2: Differentially Private Fine-tuning applies DP-SGD on private labeled data to create a privacy-preserving domain paraphraser/summarizer. Phase 3: Self-Distillation Refinement uses the DP model to generate new training pairs, applies filtering, and performs final fine-tuning to produce a refined domain paraphraser/summarizer.

and spelling mistakes) and non-linguistic errors, which we refer to throughout the paper as inconsistencies and hallucinations (e.g., misattributed quotes or ungrounded contextual details), as shown in Table 1. Our analysis reveals that DPRefine reduces linguistic errors by 84.0% compared to vanilla DPSGD while also mitigating inconsistencies present in non-private models. Our implementation of DPRefine is available as open source.[1]

2 Preliminaries

Differential Privacy: Differential privacy (DP) (Dwork et al., 2006, 2014) is a formal privacy definition that provides strong guarantees by limiting the influence any single data point can have on the output of an algorithm. Formally, a randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if for any 2 neighboring datasets $D, D' \in \mathcal{D}$ that differ in exactly one data sample, and for all sets of outcomes S, $\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta$. To train DP LLMs, *Differentially Private Stochastic Gradient Descent (DPSGD)* (Abadi et al., 2016) is typically used. DPSGD computes gradients for individual data points, clips each individual gradient, and adds Gaussian noise to the average of clipped gradients to ensure (ϵ, δ)-DP.

Knowledge Distillation: In Knowledge Distillation (KD), a small student model is trained to replicate the behavior of a larger teacher model. The student model learns by imitating the outputs of the teacher model. Our approach utilizes a variant known as Impossible Distillation (Jung et al., 2023) which leverages smaller, pre-trained models to generate high-quality training data.

3 Proposed Method

We present DPRefine, depicted in Figure 1, a three-phase method designed to enhance the linguistic quality of differentially private large language models for domain-specific tasks like summarization and paraphrasing. Our method integrates data synthesis, differentially private fine-tuning, and self-distillation to produce high-quality outputs while maintaining strong privacy guarantees. Our method integrates three key phases: data synthesis, differentially private fine-tuning, and self-distillation. In the first phase, data synthesis provides better initialization and richer feature representations, allowing the model to learn key patterns without privacy concerns. In the second phase, differentially private fine-tuning preserves these learned features while ensuring privacy, adding noise selectively to maintain model robustness. Finally, in the third phase, self-distillation combined with careful filtering further refines the model's outputs, correcting privacy-induced errors and boosting linguistic quality, all while maintaining strong privacy guarantees.

3.1 Phase 1: Data Synthesis and Model Initialization

As depicted in Figure 1 (left section), DPRefine begins by generating high-quality input-output pairs $\{(a, b), (a, c), \ldots, (x, y)\}$ using a small pre-trained language model (e.g GPT-2 for general tasks or BioGPT for biomedical tasks). We choose a smaller model because it offers a balance between efficiency and quality, allowing for fast generation of synthetic data without the computational overhead of larger models. This synthetic data serves as the foundation for fine-tuning the base T5-large model, M_{base}.

We begin by generating a **context** c based on a domain-specific **prefix**. The prefix ensures alignment with the target domain and can either be generated by a language model or sourced from a human-written corpus to ensure meaningful data. For example, using the prefix *"NYC (Reuters) –"* for the news domain, the model generates a context; *c = "The mayor announced a new climate initiative."* Multiple sentence completions $\{a, b, c, \ldots\}$ are then generated based on the context c. (a) -*"The initiative focuses on creating new parks, reducing emissions, and implementing stricter environmental regulations to combat climate change."* , (b)- *"The plan includes new parks and emission controls."* , (c)-*"New parks and stricter emission laws are planned."*...

These generated completions $\{a, b, c, \ldots\}$ are then paired to form input-output pairs (x, y), where x is an

---

[1] `https://github.com/uvm-plaid/private_llm_generation`

Category	Type	Definition	Example	Explanation
Inconsistencies	Extrinsic information	Summary contains new information not grounded in the source document	*settling companies..., he added, noting that the companies had not yet settled.*	Source does not mention whether companies had settled
	Mis-referencing (quote or attribution)	Property or event found in source, but incorrectly quoted or associated with wrong entity	*A tropical storm is expected to make landfall in the Gulf of Mexico on Monday night, the National Hurricane Center (NHC) has said.*	Source does not attribute prediction to the NHC
	Mis-referencing (name or entity)	Person, property, or event found in source, but associated with wrong name or entity	*British Cycling has apologised for "failings" in the organisation's World Class programme, according to former world champion Laura Houvenaghel.*	Houvenaghel's first name is Wendy, not Laura
	Contradiction	Summary contradicts the source material	*The Dow Jones Industrial Average .DJI > closed at a record high on Tuesday, down 0.2 percent.*	Source indicates the market went up, not down
Language Errors	Duplicated input	Model's output is identical to the input	*According to the federal Centers for Disease Control and Prevention (news-web sites)...*	Generated sentence is identical to input
	Grammar error	Grammatically incorrect output	*man, 26, arrested and charged after search properties in the city*	Missing word "of" makes sentence non-grammatical
	Incomplete thought	Output begins describing a thought or concept but does not complete it	*Queen's will cut the number of students and the staff it. The cuts will come...*	Sentence ends without completion
	Missing punctuation	Output missing clearly required punctuation	*More than 100 people have been killed in floods in the state Gujarat*	Sentence ends without period
	Spelling mistake	Output with at least one spelling mistake	*East Sussex NHS NHS has apo to patients who were sent leaflets in the mistake*	Incorrect spelling of "apologized"

Table 1: Taxonomy and examples of generation errors in language models, categorized into two main types: (1) Inconsistencies (hallucinations) - factual errors including extrinsic information, misattributions, and contradictions, and (2) Language Errors - structural issues such as grammar mistakes, incomplete sentences, and spelling errors. Each error type is illustrated with a representative example (errors highlighted in red) and detailed explanation. Inconsistency categories are adapted from Tang et al. (2024).

input and y is the corresponding output. For instance, potential pairs could be

$$(x, y) = (a, b), (a, c), (b, c)$$

providing a diverse set of synthetic input-output pairs for training.

To ensure high-quality and meaningful pairs, we apply the following filters:

1. **Entailment Filtering:** Using a pre-trained NLI model (Liu et al., 2022a), we ensure that the generated pair (x, y) holds logical entailment in both directions, meaning $x \to y$ and $y \to x$.
2. **Length Filtering:** Ensure the length of the response y is appropriate for the input x. For summarization, y should be shorter than x , $|y| < |x|$; for paraphrasing, x and y should have similar length, $|x| \approx |y|$.
3. **Diversity Filtering:** Remove pairs that are too similar to each other to ensure a diverse dataset. Pairs (x_1, y_1) and (x_2, y_2) are duplicates if one pair entails the other: $x_1 \to x_2$ and $y_1 \to y_2$.
4. **Grammar Filtering**: Apply the `language-tool-python` library (lan, 2024) to check for grammatical correctness in both x and y. Any pairs with significant grammatical errors are removed.
5. **Numerical Consistency Filtering**: For pairs containing numerical data, ensure that numbers appearing in x are consistent with those in y, ensuring no significant deviations between the input and output.
6. **Redundancy Filtering:** Remove pairs where more than 30% of the tokens in y are repeated from x, ensuring minimal redundancy in the generated text.

These customized filters inspired by the principles in Impossible Distillation (Jung et al., 2023), designed to ensure that the generated samples are highly suited for tasks like summarization and paraphrasing. The filters also counteract common errors exacerbated by DPSGD, such as language errors, as identified in our manual error analysis (see Table 1). By addressing these task-specific and error-related challenges, DPRefine produces cleaner, more accurate outputs for downstream tasks.

After filtering, we compile the curated dataset D_{base} and fine-tune M_{base}, effectively distilling the knowledge from the pre-trained model to prepare it for specialized tasks like paraphrasing or summarization.

3.2 Phase 2: DP Task-Specific Fine-Tuning

In this phase (see middle section in Figure 1), we fine-tune M_{base} using DPSGD (Abadi et al., 2016) on a private dataset while ensuring differential privacy. We assume access to this private dataset, which contains sensitive, domain-specific data relevant to the task at hand such as medical text for summarizing biomedical data. This private dataset allows the model to specialize in the target domain while ensuring differential privacy.

This fine-tuning process results in a differentially private model $M_{private}$, which preserves privacy under (ϵ, δ)-DP. The post-processing property of DP ensures that any further use of $M_{private}$ maintains the same privacy guarantee. The private dataset used in this phase consists of domain-specific data, allowing $M_{private}$ to specialize while ensuring privacy guarantees.

3.3 Phase 3: Self-Distillation Refinement

As shown in Figure 1 (right section), the final phase of DPRefine involves using self-distillation to further refine the model. Here, $M_{private}$ generates new outputs based on input contexts and self-corrects using its own predictions. For each context c, $M_{private}$ generates multiple output completions $\{a', b', c', \dots\}$, which are then paired to form new input-output pairs: $(x', y') = ((a', b'), (b', c'), (a', c'))$. The same filtering criteria from Phase 1 are applied to ensure high-quality pairs. After filtering, the data set $D_{refined}$ is used to fine-tune $M_{private}$, resulting in the final model $M_{refined}$, which balances privacy and output quality.

4 Experimental Setup

4.1 Dataset, Models and Data

Datasets and models. We evaluated DPRefine across three datasets for two domain-specific tasks: summarization and paraphrasing. For summarization, we used the XSum dataset (Narayan et al., 2018), containing 204,045 training samples and 11,334 test samples. XSum consists of BBC articles paired with single-sentence summaries, making it particularly challenging due to the requirement for highly concise yet informative summaries. We also used the PubMed dataset (National Library of Medicine, 2024), which contains 119,924 training samples and 6,658 test samples of biomedical research articles with structured abstracts. This dataset demands the summarization of technical and domain-specific content. For paraphrasing, we utilized the MRPC dataset from the GLUE benchmark (Wang et al., 2018), which contains 3,668 training samples and 1,725 test samples of sentence pairs automatically extracted from news articles, with human annotations for semantic equivalence. The smaller size of MRPC makes it particularly well-suited for tasks requiring careful paraphrasing and semantic retention.

While the privacy guarantees of DP-SGD mathematically hold regardless of pretraining data, we analyzed potential data overlap concerns (Magar and Schwartz, 2022) by examining 100 randomly sampled examples from each dataset using "What's In My Big Data?" (WIMBD)(Elazar et al., 2023) to check for overlap with C4, T5's pretraining corpus. Our investigation found no full document overlap in any of the analyzed samples (0/100 for each dataset). We chose these benchmark datasets as they are well-studied and enable reproducible research, and prior work has successfully applied differential privacy to them - for instance, Lyu et al. (Lyu et al., 2020) demonstrated privacy-preserving text representations using MRPC.

Algorithm 1 DPRefine

```
Input: Pre-trained language model $M_{pre}$, dataset $D$, privacy parameters $(\epsilon, \delta)$, learning rate $\eta$, gradient clipping norm $C$
Output: Final refined model $M_{refined}$
Phase 1: Data Synthesis & Base Model Training
Generate a set of contexts $\{c_i\}_{i=1}^{N}$, where each $c_i$ is generated by a language model (LM) using a domain-specific prefix.
For each $c_i$, generate completions $\{a, b, c, \dots\}$, where each letter represents a distinct completion for that context.
Form input-output pairs $\mathcal{P} = \{(a, b), (b, c), \dots\}$ for each context.
Define a filtering function $\mathcal{F}(\mathcal{P}) \to \mathcal{P}_{filtered}$, applying filters: entailment, length, diversity, grammar, numerical consistency, and redundancy.
Fine-tune the pre-trained model $M_{pre}$ on the filtered dataset $\mathcal{P}_{filtered}$, resulting in $M_{base}$.

Phase 2: DP Fine-Tuning with DPSGD
Compute noise multiplier $\sigma$ based on $(\epsilon, \delta)$.
for each minibatch $B_t \subseteq D$ from the private dataset do
    Compute gradients $g_t = \nabla \mathcal{L}(M_{base}; B_t)$.
    Clip the gradient: $\hat{g}_t = \frac{g_t}{\max(1, \frac{\|g_t\|_2}{C})}$.
    Add noise: $\tilde{g}_t = \hat{g}_t + \mathcal{N}(0, \sigma^2 C^2)$.
    Update model: $M_{private} \leftarrow M_{private} - \eta \tilde{g}_t$.
end for

Phase 3: Self-Distillation Refinement
Generate new completions $\{a', b', c', \dots\}$ from $M_{private}$ for each input context.
Form new input-output pairs $\mathcal{P}' = \{(a', b'), (b', c'), \dots\}$.
Apply the same filtering $\mathcal{F}(\mathcal{P}')$ to form $\mathcal{P}'_{filtered}$.
Fine-tune $M_{private}$ on $\mathcal{P}'_{filtered}$, resulting in the final model $M_{refined}$.

return $M_{refined}$
```

To generate synthetic data, we used GPT-2 (Radford et al., 2019) for general tasks like XSum and MRPC, while BioGPT (Luo et al., 2022) was employed for domain-specific biomedical text in PubMed. These smaller models were chosen due to their efficiency in producing high-quality data quickly, without the computational overhead of larger models. The synthetic data generated by these models was then used to fine-tune the base paraphrasing/summarization model, T5-large, referred to as M_{base}. While the training data for GPT-2 and T5-large is not publicly available, we use XSum, PubMed, and MRPC as proxies for sensitive data, ensuring transparency and reproducibility.

Baselines. We compared DPRefine against several baselines. The non-private baselines included Copy-Input, GPT-4, and T5-large. Copy-Input provides a simplistic baseline for paraphrasing tasks by replicating the input directly. GPT-4 represents an upper bound for performance without privacy constraints, and T5-large was fine-tuned directly on the datasets without privacy mechanisms, serving as a middle-ground non-private baseline. For private baselines, we applied DPSGD to T5-large (DPSGD_{T5}) to evaluate how much our approach improves over traditional differentially private training.

4.2 Implementation Details

Synthetic Data Generation: In Phase 1, we used GPT-2 and BioGPT to generate synthetic data with nucleus sampling (top-p=0.9, temperature=0.1) and a token limit of 150. Contextual prefixes (e.g., "New York (Reuters) –") were used for XSum and MRPC to enhance diversity, while BioGPT generated PubMed data without prefixes.

Filtering: We applied several filters to improve the generated data. First, length filtering ensured the output text was concise ($\leq 75\%$ of the input length). We then applied semantic equivalence filtering using RoBERTa-large-WANLI, retaining only pairs with a bidirectional entailment score above 0.95. Reverse-NLI filtering ensured logical consistency between input and output, with a threshold of 0.7. Additional filters removed redundant tokens ($\leq 30\%$ repetition), ensured numerical consistency, and checked for grammatical errors using the language-tool-python library. Finally, a graph-based approach identified and removed duplicate paraphrases based on entailment scores.

Model Training: Both the base model, M_{base} and the final model $M_{refined}$ were trained using the same configuration. We used T5-large with an AdamW optimizer, a learning rate of $5e^{-5}$, gradient clipping set at 1.0, and Perplexity (PPL) as the main evaluation metric. Training was conducted over 1 epoch with a batch size of 8 for training and 16 for validation. Beam search and top-p sampling (top-p = 0.9) ensured output diversity.

DP Fine-Tuning: In Phase 2, using the private-transformers codebase (Li) we fine-tuned the T5-large model (M_{base} using DPSGD. For each dataset, we set a privacy budget ϵ to 8, $\delta = \frac{1}{2N}$ and a clipping parameter of 1.0. After basic hyperparameter tuning, training was conducted over four epochs with a batch size of 4 with a learning rate of $5e^{-5}$, and batch size of 4 resulting in $M_{private}$.

Self Distillation: In Phase 3, $M_{private}$ generated new outputs using inputs from Phase 1 data. These outputs were filtered and used to further fine-tune $M_{private}$, resulting in $M_{refined}$, following the same training setup as in earlier phases.

5 Experimental Results

In this section we evaluate the efficacy of DPRefine. We conduct extensive quantitative and qualitative analysis: (1) overall evaluation using LLM-as-a-judge, (2) reference-based, targeted assessments and (4) diversity evaluations. Then, we provide a fine-grained, qualitative error-analysis which categorizes the types of errors different models make, and exemplifies the improvements that DPRefine provides. Finally, we perform ablation studies.

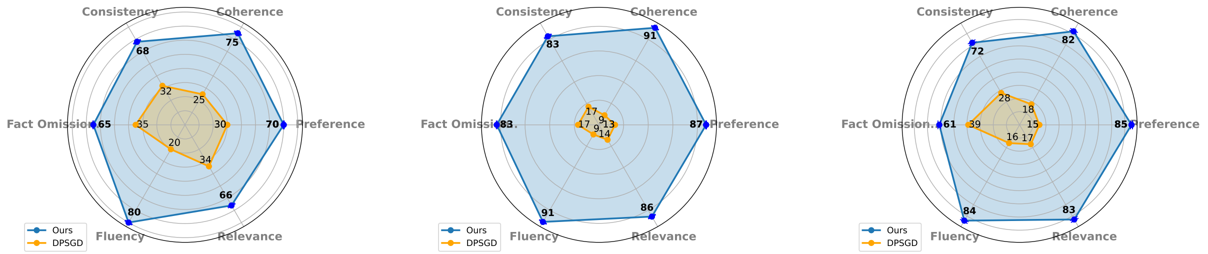
5.1 LLM-as-a-judge Evaluation

Setup: We evaluated DPRefine and DPSGD using AlpacaEval (Li et al., 2023), performing pairwise comparisons on the test sets of all datasets. *GPT-4-1106-preview* was used to assess the models across six metrics: preference, coherence, consistency, fact omission, fluency, and relevance. The evaluation was based on modified prompts(see Appendix C) tailored to measure the quality of generated outputs on these dimensions. To ensure a fair comparison, we include both non-private and private baselines. The non-private baselines (e.g., GPT-4, T5-large) demonstrate the upper bounds of performance without privacy constraints, while private baselines (e.g., DPSGD) allow us to compare how our approach improves over traditional differentially private models.

Results: As shown in Figure 2, DPRefine consistently outperforms DPSGD across all evaluation metrics. On average, AlpacaEval preferred DPRefine's generations in 78.38% of cases across all datasets and metrics. The model shows higher scores in relevance and consistency, indicating better alignment with the input and fewer contradictions. While fluency was slightly lower, DPRefine's superior performance in coherence and fact omission suggests more logically structured and accurate outputs. These results demonstrate that DPRefine generates outputs with stronger contextual and factual alignment compared to DPSGD. These findings are further supported by our manual analysis, which shows that DPRefine significantly reduces inconsistencies and language errors compared to DPSGD, validating the LLM-as-judge results.

5.2 Targeted Reference-based Evaluation

Setup: For reference-based metrics, we evaluate DPRefine using ROUGE-L (Lin, 2004) and BERT-F1 (Zhang et al., 2019) to assess token overlap and semantic similarity with reference outputs. Additionally, we use iBLEU (Sun and Zhou, 2012) and BERT-iBLEU (B-iB), which offer a more comprehensive assessment of output quality, with BERT-iBLEU being particularly useful for capturing semantic preservation, as it correlates better with human judgments than token-based metrics (Niu



(a) XSum Dataset (b) PubMed Dataset (c) MRPC Dataset

Figure 2: Comparison of DPRefine and DPSGD across multiple metrics (Preference, Coherence, Consistency, Fact Omission, Fluency, and Relevance) for the XSum, PubMed, and MRPC datasets. Error bars represent standard deviations. DPRefine demonstrates consistently stronger performance showing its ability to generate more contextually aligned and factually accurate outputs compared to DPSGD.

et al., 2020). iBLEU is a variant of the BLEU metric that balances adequacy and diversity by rewarding similarity to the reference output while penalizing excessive overlap with the input, making it well-suited for paraphrasing tasks. Given that our tasks include summarization and paraphrasing, we use these automated metrics to measure performance against ground truth outputs. We compare both non-private and private baselines to evaluate DPRefine's performance under privacy constraints.

Results: Table 2 shows that Copy-Input, a simple baseline that copies the input directly, achieves high scores on ROUGE-L and BERT-F1, indicating that token overlap metrics often favor models that produce outputs closely resembling the input especially for paraphrasing tasks. However, BERT-iBLEU highlights the limitations of Copy-Input, as it performs poorly on semantic similarity with the reference paraphrase. DPRefine outperforms DPSGD across iBLEU and BERT-iBLEU, particularly in MRPC (80.55 in BERT-iBLEU) and PubMed (75.55 in BERT-iBLEU), reflecting its ability to generate semantically accurate outputs. While privacy-preserving models naturally experience some performance trade-offs compared to non-private models, DPRefine maintains high semantic accuracy and fluency even in privacy-constrained settings. Although DPRefine's scores on ROUGE-L and BERT-F1 are slightly lower than DPSGD in some tasks, its superior performance on BERT-iBLEU suggests that it captures deeper semantic meaning.

These results indicate that reference-based metrics like ROUGE may not fully reflect the improvements in output quality brought by DPRefine, particularly in tasks requiring greater semantic accuracy. This aligns with prior findings (Cohan and Goharian, 2016; Liu and Liu, 2008; Reiter and Belz, 2009; Stent et al., 2005) that traditional metrics often fail to evaluate the true quality of summaries, especially in datasets like XSum where reference summaries are of questionable quality (Zhang et al., 2024).

5.3 Diversity Evaluation

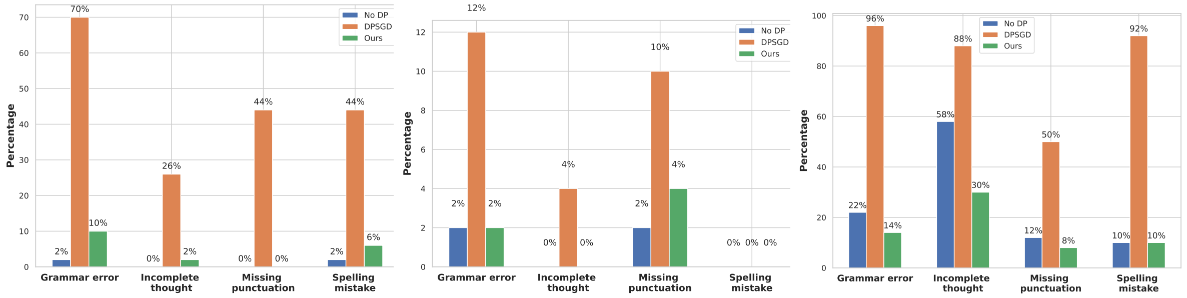
Setup: For lexical diversity, we follow Liu et al. to compute Lexical Deviation (LD) and Word Pair Deviation (WPD), which assess the degree of variation between input and output (Liu et al., 2022b). We also measure vocabulary richness using the mean-segmented token type ratio (MSTTR) (Torruella and Capsada, 2013) and the token-level Jaccard similarity between source and predicted output.

Results: DPRefine performs consistently well in MSTTR and Jaccard Similarity, slightly exceeding DPSGD across most datasets, with only a marginal difference in PubMed. This indicates that DPRefine improves vocabulary richness and reduces word overlap more effectively than DPSGD, though the improvements are modest. These results suggest that DPRefine introduces incremental enhancements in lexical diversity compared to DPSGD, especially in general tasks.

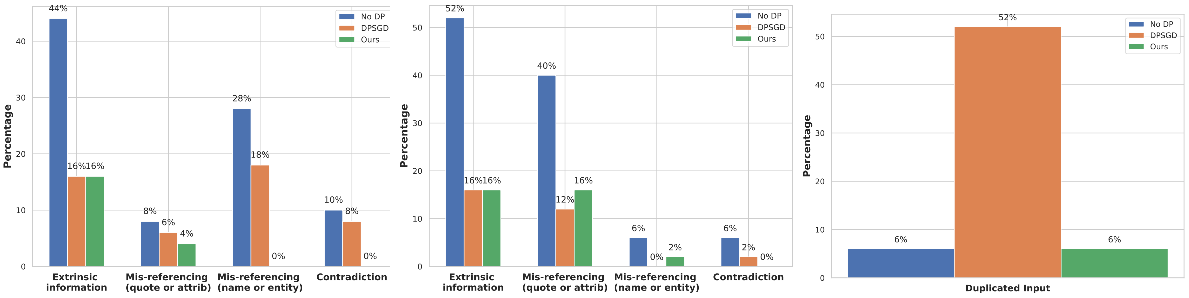
DPRefine's most significant gains are observed in WPD and LD, particularly in the MRPC paraphrasing task. In MRPC, DPRefine achieves significantly higher scores in both LD and WPD, demonstrating its ability to introduce greater structural variation and diversity between input and output. This is especially important for paraphrasing tasks, where generating diverse sentence structures is essential for varied and meaningful outputs.

5.4 Qualitative Error Analysis

To investigate the impact of DPSGD and distillation on inconsistencies and linguistic quality, we manually annotated 50 outputs each from the XSum, MRPC, and Pubmed test sets. We annotated only language errors in the Pubmed results, since the authors lack the necessary domain knowledge to evaluate inconsistencies for the associated task. To categorize inconsistencies and language errors, we developed the taxonomy shown in Table 1. For inconsistencies and hallucinations, we adopt the types and definitions proposed by Tang et al. (Tang et al., 2024). For language errors, we developed our taxonomy based on the errors observed in the model outputs. The results of our qualitative analysis



(a) XSum - Language Errors (b) MRPC - Language Errors (c) PubMed - Language Errors



(d) XSum - Inconsistency Errors (e) MRPC - Inconsistency Errors (f) MRPC - Duplicated Input Errors

Figure 3: Manual analysis of error types for three models: No DP, DPSGD, and DPRefine across XSum, MRPC, and PubMed datasets. The first row shows language errors, and the second row shows inconsistency errors. DPRefine consistently reduces both language errors and inconsistencies compared to DPSGD, leading to more accurate and fluent outputs.

appear in Figure 3.

Inconsistencies. Our results (Figures 3d and 3e) show that non-private models introduce all kinds of inconsistencies listed in Table 1, and both DPSGD and DPRefine *reduce these inconsistencies* significantly.

In both the summarization and paraphrasing tasks, the non-private model introduces *extrinsic information* in nearly 50% of all outputs; both DP methods reduce these inconsistencies to 16% for both tasks. In the paraphrasing task (MRPC), the non-private model *mis-references quotes* about 40% of the time, and *mis-references entities* about 6% of the time; both DP methods reduce these inconsistencies significantly. The non-private model introduces *contradictions* in 10% (XSum) and 6% (MRPC) of cases; both DP methods reduce them.

Language errors. Our results (Figures 3a, 3b, and 3c) show that DPSGD introduces language errors of all types at significant rates, and that DPRefine *reduces language errors to nearly the level of the non-private model*. On the PubMed dataset—the most challenging task in our evaluation—DPRefine consistently introduced fewer language errors than even the non-private model. In the paraphrasing task (MRPC), DPSGD often duplicated the input; DPRefine eliminated this behavior. See Appendix A for more discussion, and Appendix D for example outputs from all models.

5.5 Ablation Studies

5.5.1 Disentangling The Impact of Each Phase

We systematically evaluated the impact of each phase in DPRefine by comparing the performance of three models: M_{base}, $M_{private}$, and $M_{refined}$. These correspond to models trained on synthetic data (Phase 1), fine-tuned with differential privacy (Phase 2), and refined via self-distillation (Phase 3), respectively.

Results: As shown in Table 5, M_{base} (trained on synthetic data) demonstrates strong diversity but lacks consistency and fluency due to the limitations of training on synthetic data alone. Fine-tuning with DPSGD ($M_{private}$) improves fluency and semantic accuracy, though at the cost of some diversity due to the privacy noise. Self-distillation ($M_{refined}$) recovers much of the lost diversity while maintaining or improving semantic accuracy across datasets, particularly in MRPC and XSum.The results confirm that each phase plays a critical role in balancing privacy and output quality. Removing any phase leads to a reduction in performance, underscoring the necessity of the full DPRefine pipeline. See Appendix E for qualitative examples.

5.5.2 Privacy-utility Trade-off

To explore the impact of differential privacy on output quality, we conducted experiments using two privacy budgets: $\epsilon = 3$ and $\epsilon = 8$. For each ϵ, we measured the sampling efficiency—defined as the ratio of outputs that passed our quality filters—as well as the semantic

		XSUM				PUBMED				MRPC			
		iBLEU ↑	*B-iB* ↑	*R-L* ↑	*BERT-F1* ↑	*iBLEU* ↑	*B-iB* ↑	*R-L* ↑	*BERT-F1* ↑	*iBLEU* ↑	*B-iB* ↑	*R-L* ↑	*BERT-F1* ↑
Non-Private	Copy-Input	-	-	-	-	-	-	-	-	24.36	78.75	62.08	93.80
	GPT-4	28.93	73.69	13.91	86.36	32.07	74.97	19.44	83.54	30.91	82.57	42.61	91.35
	T5-large	35.0	79.96	31.60	90.59	31.83	73.67	18.10	83.18	30.25	78.92	54.89	91.85
Private	DPSGD$_{\text{T5}}$	**30.49**	74.69	**19.35**	**87.02**	30.80	71.89	15.13	81.37	28.72	78.30	**52.43**	**90.76**
	DPRefine$_{\text{T5}}$	30.10	**75.55**	17.73	86.51	**31.06**	**72.93**	**16.08**	**82.10**	**30.65**	**80.55**	42.49	90.56

Table 2: Reference-based evaluation results across XSum, PubMed, and MRPC. The first group lists non-private baselines, including Copy-Input, GPT-4, and T5. The second group presents private models (DPSGD and DPRefine). We highlight the top-performing scores for each model in bold, and underline the best non-private model results.

		Membership Inference Attack Metrics			
	Model	*AUC*	*TPR@0.01%*	*TPR@0.1%*	*TPR@1%*
XSum	Non-Private	0.56	0.00	0.00	3.00
	DPSGD	0.51	0.00	0.00	0.40
	Ours	**0.51**	0.00	0.00	**0.20**
MRPC	Non-Private	0.58	0.00	0.00	1.40
	DPSGD	0.52	0.00	0.00	1.00
	Ours	**0.51**	0.00	0.00	**1.00**
PubMed	Non-Private	0.54	0.00	0.00	1.20
	DPSGD	0.52	0.00	0.00	1.20
	Ours	0.53	0.00	0.00	**0.80**

Table 3: Membership inference attack results across datasets. Lower TPR values indicate stronger privacy protection. AUC near 0.5 indicates attack performance close to random guessing. Best results for privacy-preserving models are in **bold**, while non-private baseline results are underlined.

accuracy.

The results in Table 6 show that, at $\epsilon = 3$, the model's performance declines in both sampling efficiency and output quality compared to $\epsilon = 8$. For instance, in the MRPC dataset, DPRefine shows lower BERT-iBLEU scores at $\epsilon = 3$. However, with $\epsilon = 8$, both DPRefine and DPSGD produce higher-quality outputs, with DPRefine outperforming DPSGD.

The sampling efficiency results further emphasize the trade-off: a lower ϵ value introduces more noise, reducing the number of high-quality outputs. At $\epsilon = 8$, DPRefine not only improves utility but also generates a higher proportion of usable outputs (e.g., 51.7% in MRPC compared to 46.93% at $\epsilon = 3$), demonstrating that less noise allows the model to retain more valuable content, balancing privacy and output quality.

5.5.3 Empirical Privacy Evaluation

To complement the theoretical privacy guarantees of DP-SGD, we conducted loss-based membership inference attacks (MIA) (Yeom et al., 2018) across all three datasets. These attacks attempt to determine whether a particular example was used during training, with the attack's success rate indicating potential privacy leakage.

Table 3 presents the performance of these attacks across different false positive rate (FPR) thresholds. Both DP methods show strong protection against membership inference compared to non-private training, particularly at lower FPR thresholds where privacy protection is most critical. For instance, on the XSum dataset at 1% FPR, the true positive rate (TPR) drops from 3.00% for non-private training to 0.40% and 0.20% for DPSGD and DPRefine respectively. Similar patterns emerge across MRPC and PubMed, with both DP methods showing comparable protection levels (TPR $\leq$ 1.20% at 1% FPR).

Both DP methods achieve AUC scores near 0.5 across all datasets, indicating near-random attack performance while maintaining the utility improvements demonstrated in previous sections.

		Lexical Diversity		Lexical Deviation	
	Model	*MSTTR* ↑	*Jaccard Similarity* ↑	*LD* ↑	*WPD* ↑
XSum	Non-Private	92.37	93.73	89.39	32.20
	DPSGD	82.03	83.90	**87.08**	29.93
	Ours	**85.24**	**87.18**	84.83	**30.47**
PubMed	Non-Private	78.74	82.49	91.73	33.70
	DPSGD$_{\text{T5}}$	81.30	81.18	**92.63**	31.76
	Ours$_{\text{T5}}$	**81.31**	**81.36**	91.49	**33.11**
MRPC	Non-Private	91.92	92.53	22.50	11.33
	DPSGD$_{\text{T5}}$	**89.93**	91.26	23.63	13.47
	Ours$_{\text{T5}}$	89.39	**91.91**	**45.59**	**18.08**

Table 4: Results across multiple datasets for different models, comparing Lexical Diversity (MSTTR and Jaccard Similarity) and Lexical Deviation (LD and WPD).

6 Related Work

Our work builds on existing efforts to balance privacy and utility in NLP, particularly through the use of DPSGD for fine-tuning models on both private and public data (Kerrigan et al., 2020; Li et al., 2021; Yu et al., 2021; Mireshghallah et al., 2022b; Ganesh et al., 2023). DPSGD has been widely adopted for maintaining privacy guarantees during training, but it often comes at the cost of reduced utility (Yu et al., 2021; Ponomareva et al., 2022; Bagdasaryan et al., 2019; Mireshghallah et al., 2022b).

Knowledge Distillation (KD) has also been used to improve model performance and reduce the size of LLMs(Jiao et al., 2019; Sun et al., 2019). However, combining KD with DP remains under explored, and only a few studies address the utility losses that arise from using both techniques together(Mireshghallah et al., 2022a; Xie et al., 2024). Our work seeks to fill this gap by enhancing the linguistic quality of DP-trained LLMs through a combination of KD and self-distillation.

		iBLEU ↑	BERT-iBLEU ↑	MSTTR ↑	Jaccard Similarity ↑
XSum	M_{base} (Pretrained on Synthetic Data)	29.92	74.0	**87.48**	**89.68**
	$M_{private}$ (DP Fine-tuned)	**30.42**	75.04	81.79	83.96
	$M_{refined}$ (Self-Distilled)	30.13	**75.55**	85.20	87.18
PubMed	M_{base} (Pretrained on Synthetic Data)	**32.06**	**74.97**	71.87	76.18
	$M_{private}$ (DP Fine-tuned)	30.86	72.33	79.09	79.36
	$M_{refined}$ (Self-Distilled)	31.05	72.93	**81.05**	**81.36**
MRPC	M_{base} (Pretrained on Synthetic Data)	30.04	80.36	87.56	90.24
	$M_{private}$ (DP Fine-tuned)	30.55	80.53	86.87	89.79
	$M_{refined}$ (Self-Distilled)	**30.65**	**80.55**	**89.38**	**91.90**

Table 5: Performance comparison across multiple datasets for M_{base} (Pretrained on Synthetic Data), $M_{private}$ (DP Fine-tuned), and $M_{refined}$ (Self-Distilled) models.

		BERT-iBLEU		Sampling Efficiency
Dataset	ϵ	**DPSGD**	**DPRefine**	**DPRefine**
XSum	3	74.78	75.01	28.23% (847/3000)
	8	74.68	**75.55**	30.4% (912/3000)
MRPC	3	78.25	80.54	46.93% (1408/3000)
	8	78.30	**80.55**	51.7% (1551/3000)

Table 6: Comparison of BERT-iBLEU and sampling efficiency for DPSGD and DPRefine models across different datasets at privacy budgets of $\epsilon = 3$ and $\epsilon = 8$.

Recent work has also focused on the use of synthetic data to enhance private model training. Flemings et al. introduced DistilDP, which leverages DP synthetic data for knowledge distillation to minimize utility loss in compressed models under privacy constraints (Flemings and Annavaram, 2024), while Yu et al. explored DP synthetic data generation for training lightweight models (Yu et al., 2023). Xie et al. proposed AUG-PE, generating DP synthetic text via API access to large models without fine-tuning (Xie et al., 2024). Unlike these works, DPRefine emphasizes not only model compression and privacy but emphasizes on enhancing linguistic quality and model scalability without relying on large foundational models or relying on API access.

Hallucinations in LLMs are another critical concern, where models generate incorrect or unsubstantiated information. Some recent methods for addressing hallucinations include entropy-based detection (Farquhar et al., 2024), finetuning unfamiliar examples (Kang et al., 2024), and self-reflection techniques to reduce errors in domain-specific tasks (Ji et al., 2023). Kang et al. further observed that neural networks tend to extrapolate predictably toward a constant value when faced with out-of-distribution (OOD) inputs, contributing to predictable hallucinations in certain contexts (Kang et al., 2023). DPRefine incorporates these insights to minimize hallucinations while maintaining privacy and output quality.

7 Conclusion

We introduced DPRefine, a three-phase method to improve the utility and linguistic quality of differentially private language models. DPRefine addresses DPSGD's limitations by combining data synthesis, privacy-preserving fine-tuning, and self-distillation, resulting in more accurate, coherent, and diverse outputs. Our experiments show DPRefine outperforms DPSGD, reducing language errors and inconsistencies while preserving privacy. By using synthetic data for initialization and self-distillation for refinement, DPRefine mitigates privacy noise and balances quality with privacy guarantees. Fine-grained analysis confirms that DPRefine significantly reduces inconsistencies and errors compared to DPSGD, making it a promising approach for developing high-performing, privacy-preserving models for sensitive tasks. Future work could explore its application to additional NLP tasks and larger models.

8 Social Impacts Statement

In this work, we utilized pre-trained LLMs and well-known language modeling datasets accessed from the Hugging Face API, which are publicly available and free to use. Specifically, we employed GPT-2 and T5 models licensed under the Apache License, Version 2.0. The datasets used in our experiments, including XSum, PubMed, and MRPC, are widely recognized in the academic community and are used under their respective licenses.

Our intended use of these artifacts is strictly for academic research, aligning with the intended use specified by the creators. The datasets were chosen for their relevance to the tasks of summarization and paraphrasing, ensuring that they do not contain personally identifiable information (PII). The PubMed dataset, while containing potentially sensitive information, was handled with care to focus solely on text content without disclosing any personal or identifiable data.

By utilizing publicly available models and datasets and implementing differential privacy techniques, we minimize any unintended privacy leakage that could result from experimenting.

9 Limitations

Our approach, while showing promising improvements in the linguistic quality of DP LLMs, comes with certain limitations. First, the computational cost of the multi-phase distillation process is significant, particularly in Phase 1 where large amounts of synthetic data are generated and filtered. This might be impractical for settings with limited computational resources. Additionally, the effectiveness of our approach heavily relies on the quality and diversity of the initial synthetic dataset. If the generated data is not representative or diverse enough, the benefits of the subsequent distillation phases may be diminished.

Furthermore, our method has been evaluated on specific datasets and tasks, such as summarization and paraphrasing. The generalizability of our findings to other NLP tasks or different types of datasets remains to be explored. Lastly, while our approach reduces inconsistencies, it does not entirely eliminate them, and some errors still persist, which could affect the reliability of the generated outputs.

Acknowledgments

We would like to thank Sean Welleck for helpful discussion at the conception of the idea, and Jaehun Jung for their help with the experimental setup and evaluation choices. We would also like to thank the members of the UVM PLAID lab and UW NLP for helpful discussion and feedback.

This material is based upon work supported by the National Science Foundation under Grant No. 2238442. Computations were performed on the Vermont Advanced Computing Center supported in part by NSF award No. OAC-1827314. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300), the National Science Foundation (NSF DMS-2134012), the Intelligence Advanced Research Projects Activity (IARPA) HIATUS program (Agreement No. 2022-22072200003), and the Office of Naval Research (ONR N00014-24-1-2207). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense, the U.S. Government, or any of its agencies.

References

2024. language-tool-python: A python wrapper for languagetool. `https://pypi.org/project/language-tool-python/`. Version 2.8.1.

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

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q Tran, and Mehran Kazemi. 2024. Smaller, weaker, yet better: Training llm reasoners via compute-optimal sampling. *arXiv preprint arXiv:2408.16737*.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. *arXiv preprint arXiv:1604.00400*.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. 2023. Advancing differential privacy: Where we are now and future directions for real-world deployment. *arXiv preprint arXiv:2304.06929*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

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

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

James Flemings and Murali Annavaram. 2024. Differentially private knowledge distillation via synthetic text generation. *arXiv preprint arXiv:2403.00932*.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. 2023. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pages 10611–10627. PMLR.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2023. Impossible distillation: from low-quality model to high-quality dataset & model for summarization and paraphrasing. *arXiv preprint arXiv:2305.16635*.

Katie Kang, Amrith Setlur, Claire Tomlin, and Sergey Levine. 2023. Deep neural networks tend to extrapolate predictably. *arXiv preprint arXiv:2310.00873*.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. *arXiv preprint arXiv:2009.05886*.

Xuechen Li. private-transformers. `https://github.com/lxuechen/private-transformers`. Accessed: 2024-09-12.

Xuechen Li, Daogao Liu, Tatsunori B Hashimoto, Huseyin A Inan, Janardhan Kulkarni, Yin-Tat Lee, and Abhradeep Guha Thakurta. 2022. When does differentially private learning not suffer in high dimensions? *Advances in Neural Information Processing Systems*, 35:28616–28630.

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

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022a. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Feifan Liu and Yang Liu. 2008. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, short papers*, pages 201–204.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Timothy Liu et al. 2022b. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601.

Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. 2023. Differentially private low-rank adaptation of large language model using federated learning. *ACM Transactions on Management Information Systems*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Lingjuan Lyu, Yitong Li, Xuanli He, and Tong Xiao. 2020. Towards differentially private text representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1813–1816.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Fatemehsadat Mireshghallah, Arturs Backurs, Huseyin A Inan, Lukas Wutschitz, and Janardhan Kulkarni. 2022a. Differentially private model compression. *Advances in Neural Information Processing Systems*, 35:29468–29483.

Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. 2021. Privacy regularization: Joint privacy-utility optimization in languagemodels. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3799–3807.

Fatemehsadat Mireshghallah, Yu Su, Tatsunori Hashimoto, Jason Eisner, and Richard Shin. 2022b. Privacy-preserving domain adaptation of semantic parsers. *arXiv preprint arXiv:2212.10520*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

National Library of Medicine. 2024. Pubmed. `https://pubmed.ncbi.nlm.nih.gov/`. Accessed: 2024-06-12.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2020. Unsupervised paraphrasing with pretrained language models. *arXiv preprint arXiv:2010.12885*.

Natalia Ponomareva, Jasmijn Bastings, and Sergei Vassilvitskii. 2022. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2182–2193.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International conference on intelligent text processing and computational linguistics*, pages 341–351. Springer.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual smt system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*.

Joan Torruella and Ramón Capsada. 2013. Lexical statistics and tipological structures: a measure of lexical richness. *Procedia-Social and Behavioral Sciences*, 95:447–454.

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

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. 2022. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, et al. 2024. Differentially private synthetic data via foundation model apis 2: Text. *arXiv preprint arXiv:2403.01749*.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Da Yu, Arturs Backurs, Sivakanth Gopi, Huseyin Inan, Janardhan Kulkarni, Zinan Lin, Chulin Xie, Huishuai Zhang, and Wanrong Zhang. 2023. Training private and efficient language models with synthetic data from llms. In *Socially Responsible Language Modelling Research*.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Discussion & Examples of DPSGD Grammatical Errors

DPSGD repeats input instead of paraphrasing. In the paraphrasing task, we observed that models trained with DPSGD often produce output that is identical (or nearly identical) to the input. Non-private models and models trained with DPRefine occasionally reproduce portions of the input unchanged, but far less often than the DPSGD models—suggesting that DPSGD's noise interferes with the ability of the finetuning process to train models that effectively perform the desired task.

DPSGD misspells uncommon words. Both DPSGD and DPRefine models misspelled technical words (especially in the PubMed dataset) and names of people or places (especially in the XSum dataset). Misspellings seemed most common for uncommon words or names; for example, both models misspelled medical terms in the PubMed dataset (e.g. "Parkinson's disease" as "Parkson's disease"), and DPSGD misspelled "Hannah Ennis-Hill" as "Hillisis-hill" in the XSum dataset, but both models correctly spelled common city names like "Edinburgh." The DPRefine models produced many fewer misspellings than the DPSGD models, perhaps because the filtering step removed training examples with misspellings.

DPSGD makes grammatical errors and produces incomplete sentences. The DPSGD models made significant grammatical errors in all datasets. The DPSGD outputs often left out articles and mis-conjugated verbs; for example, in an example about flooding in Gujrat, the DPSGD model produced the sentence, "death in floods in Gujarat have more than double in two days." DPRefine models had many fewer grammatical errors, and were comparable to non-private models, perhaps because the filtering process eliminates training examples with grammatical errors.

The DPSGD models also produced some sentences that were incomplete. For example, when summarizing an article about an election, the DPSGD model produced the sentence, "out of the 28 candidates, 70 are women. the election will take place on 2 March. the D are the party with the most candidates with 38 and the." DPRefine models produced complete sentences, perhaps because the entailment filtering step fails to infer the meaning of incomplete sentences.

B Phase-wise Dataset Generation and Models

Here we provide details on the models used across different phases of dataset generation for the MRPC, XSUM, and PubMed datasets (see Table 7. The substantial increase in the number of examples generated in Phase 3 highlights the effectiveness of self-distillation in enhancing the dataset and preparing it for task-specific fine-tuning.

C Example AlpacalEval Prompt

We show an example of the prompts used in AlpacaEval for evaluating language model performance based on fluency. This prompt is a modified version of the default AlpacaEval prompt with minor changes focused on the specific metric (fluency) for evaluation. All other prompts for the remaining six dimensions (e.g., coherence, consistency, relevance, etc.) were designed in a similar way to assess specific aspects of output quality.

```
<|im_start|>system
You are a highly efficient assistant, who evaluates and selects
the best large language model (LLMs) based on the quality of their
responses to a given instruction. This process will be used to
create a leaderboard reflecting the most accurate and
human-preferred answers.
<|im_end|>
<|im_start|>user
I require a leaderboard for various large language models.
I'll provide you with prompts given to these models and their
corresponding outputs.

Your task is to assess these responses, specifically for fluency,
and select the model that demonstrates the highest fluency in
its output.

## Instruction

{
    "instruction": """{instruction}""",
}

## Model Outputs

Here are the unordered outputs from the models. Each output is
associated with a specific model, identified by a unique
model identifier.

{
    {
        "model_identifier": "m",
        "output": """{output_1}"""
    },
    {
        "model_identifier": "M",
        "output": """{output_2}"""
    }
}

## Task

Evaluate the models based on how fluently they deliver their
responses, and select the model that best demonstrates high
fluency. Answer by providing the model identifier of the
best model. We will use your output as the name of the
best model, so make sure your output only contains one of
the following model identifiers and nothing else
(no quotes, no spaces, no new lines, ...): m or M.

## Best Model Identifier
<|im_end|>
```

D Example Outputs

D.1 XSum Dataset

In this section, we present example outputs from the XSum dataset, showing the reference summaries and the summaries generated by non-private, DPSGD, and $M_{refined}$ models. These examples highlight the differences in summarization quality across different models.

D.2 MRPC Dataset

In this section, we present example paraphrases from the MRPC dataset, showing the reference paraphrases and the paraphrases generated by non-private, DPSGD, and $M_{refined}$ models. These examples highlight the differences in paraphrasing quality across different models.

Dataset	Phase 1 generation model	Phase 1 generated examples	Phase 2 fine-tuned model	Phase 3 generated examples	Phase 3 fine-tuned model
MRPC	GPT2	73706	T5-large	70143	T5-large
XSUM	GPT2	151779	T5-large	202707	T5-large
Pubmed	BioGPT	45484	T5-large	234048	T5-large

Table 7: Phase-wise dataset generation and models used.

E Example Outputs for Different Phases

In this section, we present example outputs from the XSum and MRPC dataset, showing the reference summaries/paraphrases and the summaries/paraphrases generated by M_{base}, $M_{private}$, and $M_{refined}$ models. These examples highlight the differences in summarization quality across the different stages of DPRefine.

Reference Summary	There is a ""chronic"" need for more housing for prison leavers in Wales, according to a charity.
Non-private	A charity has said there is a ""desperate need"" for housing for ex-offenders after their release from prison in Wales.
DPSGD	prison leavers need help to find a home, charity says. the charity said there was ""a desperate need for accommodation"" in Wales
$M_{refined}$	Police found three firearms, ammunition and a five-figure sum of money in Edinburgh's property on Wednesday. he was arrested and charged at Edinburgh Sheriff Court on Thursday."
Reference Summary	Four people accused of kidnapping and torturing a mentally disabled man in a ""racially motivated"" attack streamed on Facebook have been denied bail.
Non-private	Four teenagers accused of kidnapping and abducting a mentally ill man in Chicago have been denied bail, prosecutors say.
DPSGD	suspects charged with hate crimes and kidna of a man with disabilities in attack man. sasts are accused of beating the victim, who suffers from schizophrenia and mental disorder
$M_{refined}$	The four were charged with hate crimes and aggravated kidnapping and battery in the case of a white man who was found dead in his home on Friday."
Reference Summary	West Brom have appointed Nicky Hammond as technical director, ending his 20-year association with Reading.
Non-private	West Bromwich Albion have announced the retirement of former England goalkeeper Alan Shearer, who has been director of football at the club since 2003.
DPSGD	former Arsenal goalkeeper is West Brom's director of football. has been for the club since 2003. the former goal for for four years at the Royals.
$M_{refined}$	West Brom said the former goalkeeper was a key player in the club's promotion to the Premier League in 2006 and 2012."
Reference Summary	The pancreas can be triggered to regenerate itself through a type of fasting diet, say US researchers.
Non-private	A ""fasting-mimicking diet"" can help regenerate the pancreas, a study has shown, by starving mice for five days..
DPSGD	scientists have found diet can reverse diabetes in mice by reprogramming cells in the pan. the diet is low in calories, high in fat and high fats
$M_{refined}$	Researchers have found that the diet can be used to reverse the effects of diabetes in mice by reprogramming the pancreas to respond to the sugar in the blood."
Reference Summary	Since their impending merger was announced in January, there has been remarkably little comment about the huge proposed deal to combine Essilor and Luxottica.
Non-private	There is not a lot of interest in the proposed merger of Essilor and Luxottica, the two world's biggest eyewear companies.
DPSGD	silor and Luxottia are merging to create the world's largest eyewear company Esilor. the deal is not in the public interest, experts say ,
$M_{refined}$	Essilor is the world's largest manufacturer of lenses and contact lenses, and Luxottica, the top of the frame industry, says the merger would be ""in the public interest"" but the industry is concerned that the deal could be a ""stanglehold"" on the supply of high-end glasses."
Reference Summary	A ""medal at any cost"" approach created a ""culture of fear"" at British Cycling, says former rider Wendy Houvenaghel.
Non-private	British Cycling has apologised for ""failings"" in the organisation's World Class programme, according to former world champion Laura Houvenaghel..
DPSGD	Olympic gold Honaghel British Cycling has admitted it has ""fas"" in the World Class programme, she says. she said the British had ""zero regard"" for the athletes' welfare
$M_{refined}$	Hounaghel said she was ""vindicated"" by the report into the treatment of athletes by British Cycling, which she said had ""sanitised"" the organisation's ""res"" after the Olympics."
Reference Summary	Have you heard the one about the computer programmer who bought a failing comedy club in Texas and turned it into a million dollar a year business?
Non-private	""We bought a sinking ship and we brought buckets to bail it out,"" says the owner of one of the US's best-known comedy clubs.
DPSGD	comedy clubs in the US are growing in popularity, says founder eem Badr. he owns the Hiout, a comedy club in Austin, Texas. the club was founded in 2009 after two friends bought it for $20,000.
$M_{refined}$	""The comedy industry is in its golden age,"" says Kareem Badr, who is the founder of the Hideout in Austin, Texas."
Reference Summary	The reaction from BT's investors told us much about media regulator Ofcom's ruling on the fate of Openreach, the BT subsidiary that provides much of the UK's broadband infrastructure.
Non-private	Shareholders in BT have reacted with glee to the news that the government's competition watchdog, Ofcom, has said it will not take action against the company.
DPSGD	s of the telecoms is to be split by Ofcom. its pension scheme, BT, has a deficit of £10b -
$M_{refined}$	BT's pension scheme, which is owned by the government, has been the subject of debate, and the Ofcom chief has said that the company is not going to be broken up."

Table 8: Example summaries from non-private, DPSGD, and $M_{refined}$ models on the XSum dataset.

Reference Paraphrase	PCCW 's chief operating officer , Mike Butcher , and Alex Arena , the chief financial officer , will report directly to Mr So .
Non-private	PCCW's board of directors will be led by chief executive John So, according to a statement from the organisation he heads.
DPSGD	PCCW's chief operating officer, Mike Butcher and chief financial officer Alex Arena will report directly to Mr So. Mr Arena is a member of the board of directors of pccw.
$M_{refined}$	PCCW's chief operating officer and chief financial officer will report to Mr So, who will be in charge of the company' s operations."
Reference Paraphrase	The world 's two largest automakers said their U.S. sales declined more than predicted last month as a late summer sales frenzy caused more of an industry backlash than expected .
Non-private	Two of the world's largest automakers have said their U.S. sales fell more than expected last month as a late summer sales frenzy caused more of an industry backlash.
DPSGD	the world's two largest automakers said their U.S. sales declined more than predicted last month as a sales frenzy caused more of an industry backlash than expected.
$M_{refined}$	The world's two largest automakers said their U.S. sales fell more than expected last month as a late summer sales surge slowed sales."
Reference Paraphrase	A tropical storm rapidly developed in the Gulf of Mexico Sunday and was expected to hit somewhere along the Texas or Louisiana coasts by Monday night .
Non-private	A tropical storm is expected to make landfall in the Gulf of Mexico on Monday night, the National Hurricane Center (NHC) has said.
DPSGD	A tropical storm rapidly developed in the Gulf of Mexico Sunday and was expected to hit somewhere along the Texas or Louisiana coasts by Monday night.
$M_{refined}$	A tropical storm was developing in the Gulf of Mexico Sunday and was expected to make landfall in Texas or Louisiana by Monday night, according to the National Weather Service."
Reference Paraphrase	The company didn 't detail the costs of the replacement and repairs .
Non-private	The company has said it will replace the car's engine, but hasn't said how much it would cost to replace it with a new one.
DPSGD	The company didn 't detail the costs of the replacement and repairs. the company did not provide a breakdown of costs.
$M_{refined}$	The company did not provide a cost estimate for the replacements or repairs. - paraphrased , pg."
Reference Paraphrase	The settling companies would also assign their possible claims against the underwriters to the investor plaintiffs , he added .
Non-private	The settling companies would also assign their possible claims against the underwriters to the investor plaintiffs, he said ; paraphrase.
DPSGD	The settling companies would also assign their possible claims against the underwriters, the investor plaintiffs added. '
$M_{refined}$	The settling companies would also assign the investor plaintiffs' claims against the underwriters, he added, noting that the companies had not yet settled."
Reference Paraphrase	Air Commodore Quaife said the Hornets remained on three-minute alert throughout the operation .
Non-private	The Royal Navy's Hornets remained on high alert throughout the operation, the Royal Air Force has said, paraphrasing.
DPSGD	Air Commodore Quaife said the Hornets remained on three-minute alert throughout the operation.
$M_{refined}$	The Hornets were on three-minute alert during the operation, Air Commodore Quaife said. - paraphrased"
Reference Paraphrase	A Washington County man may have the countys first human case of West Nile virus , the health department said Friday .
Non-private	A Washington County man may have the first human case of West Nile virus, the health department has said. Paraphrased
DPSGD	A Washington County man may have the countys first human case of West Nile virus, the health department said Friday.
$M_{refined}$	The state health department said Friday that a man in Washington County may have the countys first West Nile virus case, the first in the state."
Reference Paraphrase	Moseley and a senior aide delivered their summary assessments to about 300 American and allied military officers on Thursday .
Non-private	US Secretary of Defense Robert Moseley and a senior aide have delivered their summary assessments of the Iraq war to senior military officers in Washington.
DPSGD	Moseley and a senior aide delivered their summary assessments to about 300 allied military officers on Thursday. he paraphrase:
$M_{refined}$	Moseley and a senior adviser delivered the summary assessments to about 300 U. S. and NATO military officers on Thursday. - paraphrased.".

Table 9: Example paraphrases from non-private, DPSGD, and $M_{refined}$ models on the MRPC dataset.

Reference Summary	—
Non-private	background : congenital adrenal hyperplasia (cah) refers to a group of autosomal recessive disorders caused by an enzyme deficiency which leads to defects in biosynthesis of steroid precursors. during childhood, the management is largely focused on achieving normal growth and attaining appropriate final adult height.aim:the aim of this study was to evaluate the clinical and biochemical features of patients diagnosed to have
DPSGD	the relationship between anxiety and cogn in parkson's disease (pd) has not been explored. this study investigated whether anxiety differentially affects cogn in patients with left-sided dominant d (ldd) compared p patients with anxiety. the results showed that apathy, but not anxiety, was associated with executive dysfunction, and pathy. in the lp patients,pathy was not associated cognitive
$M_{refined}$	anxiety affects quality of life in those living with parkinson's disease (pd) more than overall cognitive status, motor deficits, apathy, and depression, but the relationship between anxiety and cognition in d is still meager and lags far behind that of depression. however, recent studies have shown that anxiety is more prominent and prevalent than depression and is associated with cognitive performance."
Reference Summary	—
Non-private	mirnas are a class of multifunctional singled - stranded small RNAs which regulate gene expression and play important roles in the onset and progression of tumorigenesis. the current treatment of radiation therapy, chemotherapy and hormone therapy slow mammary gland tumor growth and prolong survival but metastatic breast cancer still remains incurable due to our limited understanding of the molecular mechanisms that drive tumor formation and metastasis in eukary
DPSGD	nns are non-coding rs tran into mna but remain untranslated in e otic cells. they includern, mira,pirn andsnolarrra. mirn are class of multi d-stranded small aa 20nt in length and regulate range of processes including development, cell cycle progression, stem cell renewal, differentiation, proliferation, metsis and
$M_{refined}$	rnas are transcribed into mra but remain untranslated in eukaryotic cells. they include sir (small interfering), mir (micror) and snolar (small nucleolar). mira is class of single - stranded small RNAs that regulate expression and expression of genes and are involved in the initiation and progression of tumorigenesis."
Reference Summary	—
Non-private	ovarian hyperstimulation syndrome (ohss) is a serious complication of in vitro fertilization, occurring in 1 - 10% of patients. vascular endothelial growth factor has emerged as one of the factors most likely involved in the pathophysiology of this iatrogenic condition and human chorionic gonadotrophin has been implicated in its etiopathogenesis
DPSGD	hs is complication of oulation induction, occurring in 1- 10% of ino fertilization patients. it has spectrum of clinical and laboratory manifestationsranging from mild to life-threatening conditions.vascular endhelial growth factor has emerged as one factor most likely involved in the pathophysio ofos. endoheli growth is an ogenic kin that is stimulator of thevascular thelium with its sequelae of ascite and
$M_{refined}$	hs is a serious complication of ovulation induction, occurring in 1 - 10% of in vitro fertilization patients. vascular endothelial growth factor is an angiogenic factor that causes capillary permeability with its sequelae of ascites and pleural effusion. the early form is caused by exogenous chorionic gonadotrophin (hcg) and occurs before the 10"
Reference Summary	—
Non-private	background : congenital adrenal hyperplasia (cah) refers to a group of autosomal recessive disorders caused by an enzyme deficiency which leads to defects in biosynthesis of steroid precursors. during childhood, the management is largely focused on achieving normal growth and attaining appropriate final adult height.aim:the aim of this study was to evaluate the clinical and biochemical features of patients diagnosed to have
DPSGD	congenital hyperpa refers to group of autosom reces disorders caused by an enzyme de which leads to defects in biosynthesis ofsteroid precursors. the non classical cah variant is more common with prevalence of 1 in 1000 the classical h is more common with a prevalence rate of 1000. management is focused on normal growth and attaining appropriate adult height and height sds andbi data were collected
$M_{refined}$	congenital adrenal hyperplasia (cah) is group of autosomal recessive disorders caused by an enzyme deficiency which leads to defects in synthesis of steroid precursors. johns medical college hospital, bangalore by the department of endocrinology on patients diagnosed to have ca and seen in outpatient clinic between january 2012 and october 2012. all patients underwent bio"

Table 10: Example summaries from non-private, DPSGD, and $M_{refined}$ models on the PubMed dataset.

Reference Summary	—-
Non-private	type 1 diabetes (t1d) results from the destruction of insulin producing pancreatic cells by a cell specific autoimmune process. in order to obtain insight into the pathogenic mechanisms in humans and to test novel therapeutic approaches for its treatment, different preclinical models of the disease such as spontaneous and accelerated diabetes in the non - obese diabetic (nod) mice or diabetes induced in susceptible rodent strains by multiple low doses of str
DPSGD	pentoxlline (px) is a xanthin-derived general phosphosterase ((de) inhibitor that has immunomodulatory and anti-inflammatory activity. tx is an p - inhibitor of dieas (de), which has been available for many years to treat vascular disorder. the st and cc phosphate (cgp) andcic
$M_{refined}$	pt is a methyl xanthine-derived general phosphodiesterase (pde) inhibitor that has been used for many years to treat vascular disorders, and it is thought to have anti- inflammatory activity, which may be useful in the treatment of type 1 diabetes. t1d is caused by the destruction of insulin producing pancreatic cells by specific autoimmune process. the pathogenesis of 1d was mediated by"
Reference Summary	—
Non-private	objective : to determine the presence of specific gastrointestinal antimicrobials and their sensitivity, as well as whether these are associated with severity of asthma in adults.methods:a cross - sectional study including adult asthmatics in ambulatory care at a tertiary care hospital in the city of rio de janeiro. patients were randomly assigned to two groups of severity according to the global initiative for asthma criterion
DPSGD	os srics de antipos ige antitas estoccicas, as asa em adulto in hospital universitrio no rii (rj) study included adultms in acomament ambulatorial hospital universityir io dejano (j. determined nves specifics of corposige andoxinsaafil
$M_{refined}$	estudo transversal incluindo asmticos adulto ambulatorial in um hospital universitrio no janeiro (rj). o pacientes foram alocados consecutivamente in two groups de gravidade da asa segundo criteria of global initiative for asthma: as leve (al) and asmoderada ou grave (amg).
Reference Summary	—
Non-private	background : the family is the cornerstone of human social support network and its presence is essential in everyone s life. changes inevitably occur in families with illness and hospitalization of a family member.objectives:the purpose of this study was to investigate the satisfaction of the families of patients in the intensive care unit (icu) in tehran.methods:this cross - sectional survey was conducted on the patients, their families and their
DPSGD	the satisfaction of the family of hospitalized patients in intensive care units was investigated. the families of intensive units are facing life-threatening illness experiences and complex treatments, different technologies and different equipment are used which could lead to the dissatis of family members. this study aimed to investigate satisfaction satisfaction of families in the intensive unit. results showed that satisfaction family satisfaction was higher in intensive units than in other , the satisfaction and satisfaction were higher than the
$M_{refined}$	the family is the cornerstone of human social network and its presence is essential in everyone s life, and changes inevitably occur in families with illness and hospitalization, especially intensive care unit (icu). the needs of the families of hospitalized patients in the i are summarized in five sections, which include: assurance, communication, information, support, comfort and satisfaction. fox et al. found that the satisfaction of family members is the most important"
Reference Summary	—
Non-private	Non-private: background. coronary artery disease is one of the most widespread chronic diseases in most countries and it is associated with high mortality, limitations in life and disability in a large part of productive forces of iran.objectives.the aim of this study was to investigate the relationship between anxiety and depression in patients undergoing heart surgery.patients and methods.a cross - sectional survey was conducted among patients who had undergone cardiac surgery from january 2011 to
DPSGD	coronaryartery disease is the most cardiovascular disorder in developing and developed countries. it is not only chronic disease with high mortality, but it causes limitations in life and disability in a large part of productive forces of the country. it also associated with increased production and medical costs. coronar artery bypass surgery is one of surgical procedures for treatment of cardiovascular patients. on the other, complications are anxiety and depression. anxiety is common psychological response of patients which is associated quality of
$M_{refined}$	the development of human societies and industrialization as well as changes in stress sources has changed the disease pattern in civilized societies. coronary artery disease is the most widespread chronic disease in most countries. according to forecasts, the mortality rate in eastern mediterranean countries including iran will be 30 - 35%. anxiety and depression are associated with the quality of life of patients with cardiovascular diseases. in fact, anxiety is a common psychological response of cardiovascular patients."
Reference Summary	—
Non-private	objective : influenza a virus (h3-f7) was isolated from patients in nepal, thailand and the united states in 2009. rna was extracted from 48-hour shell vial cultures and analyzed by reverse transcription - polymerase chain reaction and ferret antisera using the superscript pcr system (invitrogen ; carlsbad / california
DPSGD	ra was extracted from 48-hour shell cultures by using the magnapur lx (roche mocular) andrii kiti (roch mo) according to the manufacturer's protocols. t-pr amplification was performed by the one- step reverse - polymerase reaction system (rogen, carbad, ca, usa). gtg- ct atc-gc
$M_{refined}$	rna was extracted from 48-hour shell cultures (10) and was amplification by reverse transcription- polymerase chain reaction (rt-pcr). the swab samples were collected within the first 72 hours of onset of symptoms, and were analyzed by immunoassay and monoclonal antibody staining (11). for antigenic analysis of select isolates, mra was added to"

Table 11: Example summaries from non-private, DPSGD, and $M_{refined}$ models on the PubMed dataset.

Original Text	PCCW's chief operating officer, Mike Butcher, and Alex Arena, the chief financial officer, will report directly to Mr So.
M_{base}	The company's chief operating officer, Mike Butcher, and chief financial officer Alex Arena will report directly to Mr So and will not be subject to the oversight of the PCCW board of directors.
$M_{private}$	The company's chief operating officer, Mike Butcher, and chief financial officer Alex Arena will report directly to Mr So, who will be in charge of the company's operations.
$M_{refined}$	PCCW's chief operating officer and chief financial officer will report to Mr So, who will be in charge of the company's operations.
Original Text	The world's two largest automakers said their U.S. sales declined more than predicted last month as a late summer sales frenzy caused more of an industry backlash than expected.
M_{base}	The world's two largest automakers, General Motors and Chrysler, said their U.S. sales declined more than they had expected last month, as a weak economy and weak consumer demand weighed on their bottom lines.
$M_{private}$	The world's two largest automakers, Ford and General Motors, said their U.S. sales declined more than expected last month as a slump in sales slowed sales of both cars and trucks.
$M_{refined}$	The world's two largest automakers said their U.S. sales fell more than expected last month as a late summer sales surge slowed sales.
Original Text	According to the federal Centers for Disease Control and Prevention (CDC), there were 19 reported cases of measles in the United States in 2002.
M_{base}	According to the Centers for Disease Control and Prevention (CDC), there were 19 cases of measles in the United States in 2002.
$M_{private}$	According to the Centers for Disease Control and Prevention (CDC), there were 19 cases of measles in the United States in 2002.
$M_{refined}$	In 2002, the U.S. Centers for Disease Control and Prevention reported 19 cases of measles in the United States.

Table 12: Example paraphrases from M_{base}, $M_{private}$, and $M_{refined}$ models on the MRPC dataset.

Reference Summary	There is a "chronic" need for more housing for prison leavers in Wales, according to a charity.
M_{base}	Prison Link Cymru said it had received 1,099 referrals in 2015-16, with some ex-offenders living on the streets for up to a year before being able to find accommodation.
$M_{private}$	Prisons in Wales have been more than ever, but the charity says they are struggling to find accommodation for offenders in the country.
$M_{refined}$	Wales' prisons have said that the number of people living on the streets is increasing, but the charity says the need for more one-bedroom flats is more than ever.
Reference Summary	A man has appeared in court after firearms, ammunition and cash were seized by police in Edinburgh.
M_{base}	Officers searched properties in the city's Waterfront Park and Colonsay View on Wednesday. A 26-year-old man was arrested and charged with robbery.
$M_{private}$	Police found three firearms and a five-figure sum of money in Edinburgh on Wednesday. A man was arrested and charged in the attack in the city.
$M_{refined}$	Police found three firearms, ammunition, and a five-figure sum of money in Edinburgh's property on Wednesday. He was arrested and charged at Edinburgh Sheriff Court on Thursday.
Reference Summary	Four people accused of kidnapping and torturing a mentally disabled man in a "racially motivated" attack streamed on Facebook have been denied bail.
M_{base}	Prosecutors said the suspects stuffed a sock into the victim's mouth, taped his mouth shut, and beat the 18-year-old white victim to death, according to the Chicago Tribune.
$M_{private}$	Suspects are accused of beating the 18-year-old victim in a van after he was taken to the police station, police said on Friday.
$M_{refined}$	The four were charged with hate crimes and aggravated kidnapping and battery in the case of a white man who was found dead in his home on Friday.

Table 13: Example summaries from M_{base}, $M_{private}$, and $M_{refined}$ models on the XSum dataset.