The Privacy-Hallucination Tradeoff in Differentially Private Language Models

Krithika Ramesh
Johns Hopkins University
kramesh3@jh.edu

Krishna Pillutla
Indian Institute of Technology, Madras
krishnap@dsai.iitm.ac.in

Danish Pruthi
Indian Institute of Science
danishp@iisc.ac.in

Anjalie Field
Johns Hopkins University
anjalief@jhu.edu

Abstract

Both privacy and factual accuracy are paramount in high-stakes domains like healthcare. Concerningly, we uncover and investigate a *privacy-hallucination tradeoff* in differentially private (DP) language models. First, we empirically show that models pre-trained or fine-tuned with DP tend to produce more hallucinations than non-DP counterparts, with increased severity as the privacy budget grows stricter. Second, we investigate model properties driving this tradeoff, demonstrating that DP mechanisms flatten output distributions, potentially redistributing probability mass toward factually incorrect alternatives. Third, through experiments where we control fact frequency in training data, we characterize how information frequency can reduce hallucination risks in DP models. Overall, our findings underscore the need for more nuanced privacy-preserving interventions that offer rigorous privacy guarantees without compromising factual accuracy.

1 Introduction

The development and deployment of large language models (LLMs) in high-stakes settings requires non-negotiable standards for both privacy-preservation and factual accuracy. LLMs that are exposed to sensitive information during training are susceptible to reproducing it in subsequent interactions, resulting in privacy violations (Carlini et al., 2021; Chu et al., 2024; Meeus et al., 2024; Kandpal et al., 2024), even in non-adversarial settings (Aerni et al., 2025); this is legally and ethically unacceptable in high-stakes domains such as healthcare and law, which involve sensitive data. As simple anonymization offers insufficient protection (Staab et al., 2024; Xin et al., 2024; Pang et al., 2024), differential privacy (DP) (Dwork et al., 2006) has emerged as the de-facto paradigm for provably mitigating privacy risks in language models (Carlini et al., 2019; Li et al., 2022; Xu et al., 2023; Yan et al., 2024; Hu et al., 2024).

Alongside privacy risks, LLMs are known to *hallucinate*, i.e., to generate factually incorrect outputs (Wang et al., 2024; Jiang et al., 2024a; Das et al., 2022; Asgari et al., 2025).[1] This issue poses serious risks, particularly in high-stakes tasks such as generating patient discharge summaries (Chung et al., 2025). These concerns have motivated research aiming to evaluate and improve factual correctness in LLMs (e.g. Ji et al., 2023; Li et al., 2024a). While prior work has investigated privacy and factual accuracy independently, no work has investigated the interaction between them, despite the clear need to achieve both in high-stakes settings.

In this work, we empirically investigate the tradeoff between these two critical properties, specifically focusing on the guiding question: does differentially private training increase factual hallucinations in models? Our work is motivated by the potentially conflicting conditions conducive to each property. Privacy-preserving strategies are inherently designed to counteract memorization (Miranda et al., 2025; Kassem et al., 2023; Hans et al., 2024), which may inhibit the acquisition and, thus, the output of factual information (Lu et al., 2024; Merullo et al., 2025).

In particular, because example-level DP limits the influence of individual training examples on the model, one might expect a privately trained model to struggle with reproducing *rare facts* (i.e., those that appear only a few times in the fine-tuning data), while still capturing information that is repeated more frequently and is less likely to be privacy-sensitive. However, our findings belie this expectation; rather than only reducing the generation of rare facts, private training leads to a problematic increase in hallucinations, where models do not simply omit what they failed to learn, but actively

[1] We use "hallucination" to refer to LLM generation of false information. Unless stated otherwise, we restrict ourselves to information not supported by the training (pre-training or fine-tuning) data. Appendix B gives a definition.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 44441–44489
October 24-29, 2026 ©2026 Association for Computational Linguistics

generate incorrect content instead.

In this work, we specifically investigate:

(RQ1) What impact does DP training have on hallucinations in model outputs?

(RQ2) What impact does DP training have on model properties related to hallucination?

(RQ3) Under what conditions could DP be usable without increasing hallucinations?

Findings.

- In RQ1, we find consistent evidence that DP training, in both fine-tuning and pre-training settings, leads to increased hallucinations, with a more pronounced degradation in the pre-training setting.
- In RQ2, we find that DP noise leads to flatter predictive distributions, dispersing the next-token probability mass across a larger set of candidate tokens, which can increase the risk of generating factually incorrect content.
- Finally, in RQ3 we show that a fact must recur several times before a DP-trained model acquires it at all, and we find that under stricter privacy budgets, facts are not acquired even at the substantially higher frequencies that we test for.

Our results call for increased caution in turning to DP as a solution to privacy in high-stakes settings and further investigation of when DP can be used without increasing hallucinations.[2]

2 Experimental Design

Our goal is to measure how models are affected by DP at various privacy budgets. To this end, we carefully construct experimental setups where we can fine-tune valid differentially private models, control for overlap between fine-tuning and pre-training data, and evaluate factual accuracy in open-ended model outputs. We situate our study in related research on privacy-preserving approaches and factuality in language models, reviewed in detail in §A.1.

2.1 Models and Training Setup

DP Fine-tuning. We fine-tune LLMs for controllable text generation, similar to previously proposed applications for DP in LLMs, particularly for privacy-preserving synthetic and open-ended text generation (Yue et al., 2023; Mattern et al., 2022; Ramesh et al., 2024).

[2]Our code is available at: https://github.com/kr-ramesh/privacy-hallucination-tradeoff.

All fine-tuning (DP and non-DP) is achieved using Low-Rank Adaptation (LoRA) for computational efficiency (Hu et al., 2022). We specify additional details about DP in §A.2 and about LoRA in §C.2, including all hyperparameter settings.

The base language model behind all fine-tuning experiments is GPT-J 6B (Wang and Komatsuzaki, 2021), a decoder-only transformer-based model. We make this choice for a crucial reason: its pre-training dataset, namely The Pile, is fully open with a known cutoff date, ensuring that GPT-J is not pre-trained on any post-2020 data. This knowledge of the pre-training data mixture and cut-off date allows us to select fine-tuning data that has no overlap with pre-training data; cf. §2.2.

Private pre-training. We address the effect of DP pre-training (as opposed to fine-tuning) on factuality using the DP-pre-trained VaultGemma (VaultGemma Team, 2025). This is a 1B-parameter open-weights model fully pre-trained with DP ($\epsilon = 2$).

We compare VaultGemma with the same two models as in VaultGemma Team (2025): Gemma3-1B, which can be viewed as the "non-private counterpart", and GPT-2 XL (1.5B), which has performance similar to VaultGemma's on standard benchmarks, although it has a much earlier knowledge cut-off date, as it was released in 2019. We also evaluate the non-private Gemma2-2B, which is data-matched with VaultGemma, and Gemma-2B as additional baselines.

2.2 Datasets

We use Wikipedia data for fine-tuning and evaluation because it meets two key criteria. First, metadata allows us to select articles created after 2020, ensuring they were not included in GPT-J 6B's pre-training corpora. Overlap with pre-training data would invalidate DP guarantees and artificially inflate performance (Tramèr et al., 2024; Cummings et al., 2024; Igamberdiev et al., 2022). Second, content is constructed to contain verifiable facts rather than opinions or speculation, and automated fact-checking methods have been previously validated over Wikipedia (Min et al., 2023), which ensures that the setup can answer RQ1. We use three datasets for factuality evaluation as summarized in Table 1 and described below. An exact list of topics included in these datasets is given in Appendix C.4.

Wikipedia Science. We collect 231 Wikipedia articles on science topics that were created after the

cutoff date for GPT-J 6B's pre-training data, where we use keyword searches of Wikipedia meta-data to identify science articles. We focus on science topics as they contain detailed technical language, which is also common in sensitive data settings (e.g., clinical notes). While Wikipedia articles on these topics did not exist before 2020, we do expect some concepts to exist in other pre-training data sources, which makes it feasible for a DP model to produce facts on these topics, even without memorizing individual data points.

Wikipedia AI. We collect 124 Wikipedia articles on AI topics, where we hand-curate products and models that did not exist before 2020, along with related articles we expect to mention them. Unlike the Wikipedia Science articles, GPT-J 6B cannot have any knowledge of most of these concepts without fine-tuning, as they could not have existed in pre-training data. However, by constructing our data to contain articles that mention overlapping topics, we ensure that it is feasible for a DP model to learn them. For example, if our dataset only contained *DeepSeek (chatbot)*, DP would preclude learning of information isolated to one data point. By including *DeepSeek (chatbot)*, *DeepSeek*, and *DeepSeek (disambiguation)*, a DP model can hypothetically learn information about DeepSeek, as it is mentioned in multiple data points.

Wikipedia pre-training. To investigate effects of DP fine-tuning on knowledge acquired during pre-training, we randomly sample 250 Wikipedia articles from the GPT-J pretraining data that are not included in the fine-tuning data.

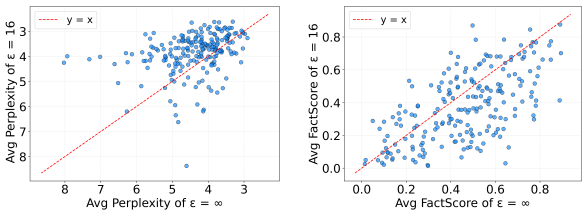
For fine-tuning, since DP requires sufficiently large datasets and batch sizes (McMahan et al., 2018; Ponomareva et al., 2023), we intersperse our curated evaluation articles with 20,000 additional randomly sampled Wikipedia articles (ensuring no overlap with the pre-training evaluation set). The data is divided into 512-token sequences as the unit of privacy protection. We fine-tune the model to produce an article when prompted on the article title (e.g., a topic).

2.3 Evaluation of Factual Accuracy in Open-Ended Text Generation

We evaluate the factual accuracy in open-ended text generation settings where we prompt the model with a title and generate a full Wikipedia article. We then measure factual accuracy of the generated article, where higher accuracy indicates a lower hallucination rate. This setup reflects a more modern paradigm for evaluating factuality in open-ended generation, as opposed to traditional factuality evaluations based on cloze-style or short-form response queries designed to probe models (Youssef et al., 2023; Petroni et al., 2019). We note that utility measures such as perplexity—a common metric used for evaluating the efficacy of DP methods in prior work (Yu et al., 2022; Thareja et al., 2026; Ma and Rajtmajer, 2026)—do not serve as a proxy for hallucinations, and therefore fail to capture factual inaccuracies (see Figure 1; further details provided in Appendix D.4). We use both automated and human assessments of factuality, as described below.

Dataset	Article Count	Topics
Wikipedia Science	231	Follicular drug delivery; Malaria therapy; Eurotrac
Wikipedia AI	124	DeepSeek; DALL-E
Wikipedia Pretraining	250	Randomly sampled articles
Fine-tuning data	20355	Wikipedia Science + Wikipedia AI + 20k randomly sampled articles

Table 1: Datasets used for evaluation and fine-tuning.



(a) Avg. perplexity (b) Avg. FactScore

Figure 1: Average perplexity (a) and average FactScore (b) per topic on Wikipedia Science, $\varepsilon = 16$ vs. $\varepsilon = \infty$. DP-finetuned models often achieve lower average perplexity (points above the $y = x$ line), yet the non-DP model attains higher average FactScores across topics—showing that lower perplexity does not imply better factuality.

Automated Evaluation via FactScore. Given a generated text d_i, FactScore (Min et al., 2023) operates in two distinct phases: (i) **atomic claim extraction**, where d_i is decomposed into a set of minimal, verifiable claims $\mathcal{AF}_{d_i}$, and (ii) **claim verification**, where each claim $\alpha_j^{(d_i)} \in \mathcal{AF}_{d_i}$ is evaluated for factuality using a verifier $\mathcal{V}$ conditioned on both intrinsic language model judgments and evidence retrieved from an external knowledge source $\mathcal{K}$, as detailed in Algorithm 1, Appendix A.3. We use the original Wikipedia article as the external knowledge source for verifying generated claims. As claim decomposition methods can produce redundant claims that artificially inflate scores, we use the CORE module (Jiang et al., 2024b) to filter

down the superfluous and repetitive claims. We use Llama-3.1-8B-Instruct for both (i) and (ii). We verify in Appendix D.9 that trends are consistent with other choices of models.

Human Evaluation. As FactScore is an automated metric relying on LLM judgments that may not be accurate, we conduct human evaluations to validate results and provide finer-grained analysis of generated information. We use the Wikipedia AI dataset, as this dataset most carefully separates the pre-training and fine-tuning data. For the annotation task, we recruited 5 computer science graduate students, whom we expect to have high AI literacy. We ensure that the human evaluators see generations with both high and low factual accuracy using stratified sampling: we sample 15 articles where both DP ($\epsilon = 16$) and non-DP models have FactScore $\geq 50\%$ and 15 where both models have FactScore $< 50\%$. Since FactScore judgements may be imprecise, this stratification allows us to evaluate whether human annotators can distinguish between DP and non-DP generations with comparable FactScores, and informs us of whether FactScore underestimates DP's effect on the factuality of the generation. Annotators rated automatically decomposed atomic claims for their overall veracity (correct/incorrect/unclear) and their support within the source text using our custom annotation interface (Appendix §D.8). Annotators were also asked to flag quality issues in decomposed claims, marking them as vague or subjective. Two annotators rated each claim's veracity and support, and we report moderate-to-strong inter-annotator agreement (Cohen's $\kappa = 0.57$ for both veracity and support at $\epsilon = \infty$, and 0.84 and 0.73 respectively at $\epsilon = 16$). [3]

3 RQ1: What impact does DP training have on hallucinations in model outputs?

3.1 DP fine-tuning increases hallucinations

Automated Evaluation. We generate multiple articles per topic and report both the response-level statistics, where every generation is scored as an independent observation, and topic-level averages, where FactScores for generations are averaged within a topic first, so every topic is weighted equally, and the CIs are bootstrapped by resampling topics. In addition to fine-tuned models, we also evaluate the "base" pre-trained model and the task-tuned model, which denotes the base model fine-tuned without DP on pre-training Wikipedia data only, excluding the unseen WIKI$_{\text{AI + SCIENCE}}$ articles. These baselines offer comparison between the DP models and models never trained on this data.

	Response-level				Topic-level
DP Setting	**Avg FS**	**Med**	**Q1**	**Q3**	**Avg FS [95% CI]**
Wiki AI					
$\epsilon = \infty$	37.7	33.3	13.3	57.1	37.5 [33.2, 42.0]
$\epsilon = 16$	33.8	27.3	10.0	54.0	33.6 [29.2, 38.2]
$\epsilon = 8$	31.6	25.0	10.0	50.0	31.4 [27.3, 35.7]
Base	32.9	27.3	**9.1**	50.0	32.9 [28.4, 37.5]
Task-tuned	**30.3**	**23.1**	**9.1**	**48.3**	**30.3** [26.3, 34.4]
Wiki Science					
$\epsilon = \infty$	56.1	58.3	33.3	80.0	56.1 [52.6, 59.5]
$\epsilon = 16$	53.6	55.6	28.6	80.0	53.6 [50.1, 57.1]
$\epsilon = 8$	53.5	54.5	28.6	80.0	53.6 [50.1, 57.1]
Base	**52.1**	**50.0**	27.3	**75.0**	**51.9** [48.9, 55.0]
Task-tuned	52.6	**50.0**	**25.0**	80.0	52.5 [49.0, 56.1]

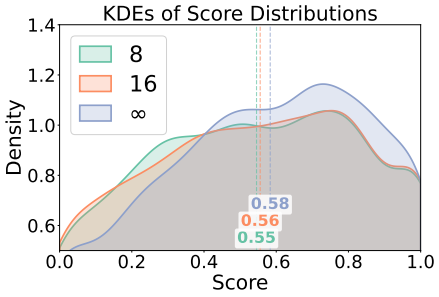
Table 2: FactScore (FS; %) of GPT-J 6B fine-tuned under different DP budgets, evaluated at temperature $\tau = 0.3$. Response-level statistics pool all generations; topic-level averages within each entity first, with bootstrap 95% CIs resampled over topics. The row with the lowest FactScore (i.e. most hallucinations) is bolded. Factual consistency decreases with stricter privacy budgets (smaller ϵ), highlighting the tradeoff between privacy and hallucinations. Significance tests for these differences against the non-private baseline ($\varepsilon = \infty$) in §D.2.

Wikipedia AI	**FactScore < 0.5**		**FactScore ≥ 0.5**	
	$\epsilon = \infty$	$\epsilon = 16$	$\epsilon = \infty$	$\epsilon = 16$
Avg. Veracity Score	0.18	0.17	0.51	0.43
Avg. Support Ratio	0.19	0.09	0.54	0.32
Quality Issues	87	79	55.5	79

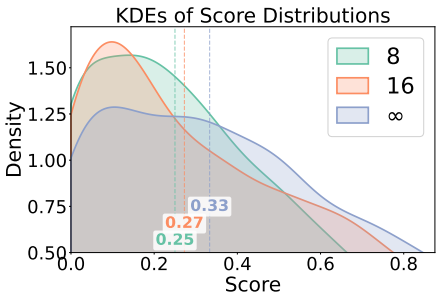
Table 3: Human annotation results on Wiki AI , stratified by response groups with FactScore < 0.5 and ≥ 0.5. In both cases, hallucinations increase with DP ($\epsilon = 16$) relative to no DP ($\epsilon = \infty$). "Quality Issues" is the number of claims annotated as having quality issues.

The automated factuality evaluation for unseen fine-tuning data in Table 2 suggests that DP fine-tuning leads to more frequent hallucinations. There is a consistent decrease in the average FactScore for the text generated from models fine-tuned with DP compared to those trained without DP: non-DP ($\epsilon = \infty$) $\succ$ DP ($\epsilon = 16$) $\succ$ DP ($\epsilon = 8$) on both datasets. Performance decline is more pronounced for the stricter privacy budget setting $\varepsilon = 8$ on the Wikipedia AI data, and is further illustrated by the skew (toward lower factuality scores) in the distributions of FactScores for DP-trained models

[3] Annotators were offered compensation of $\sim$ \$20/hr.



(a) Wikipedia Science



(b) Wikipedia AI

Figure 2: Kernel density (KDE) plots comparing FactScore distributions when finetuning GPT-J 6B under different DP budgets. As the privacy budget becomes more stringent (ϵ from ∞ to 8), the FactScore distribution shifts leftward, indicating that stronger privacy constraints result in less factually consistent outcomes.

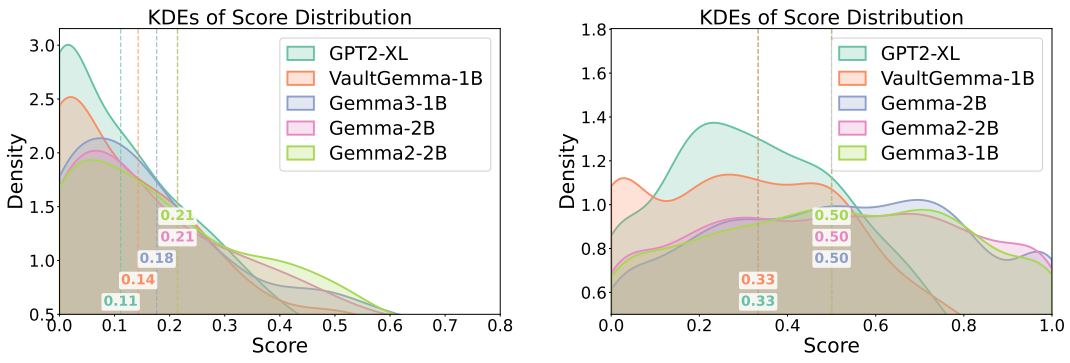
in Figure 2. The differences in the FactScores of the non-DP and DP-finetuned models are significant at the response level and at the topic-level, except for $\varepsilon = 16$ on Wikipedia AI. (cf. §D.2).

As a sanity check, we observe that the model trained without DP outputs a higher percentage of correct claims than the base model (32.9% $\rightarrow$ 37.7% for the AI articles and 52.1% $\rightarrow$ 56.1% for the Science articles), demonstrating successful fine-tuning. Notably, on the Wikipedia AI dataset, the model fine-tuned with $\varepsilon = 8$ performs worse or comparably (mean: 31.6%; median: 25.0%), to the base model (mean: 32.9%; median: 27.3%), despite the training loss decreasing progressively when fine-tuning with this privacy budget (Appendix C.5). While this marginal degradation in FactScore for Wikipedia AI articles could be statistical noise, it could also indicate the gradient signal from the AI articles being masked by the larger levels of noise introduced by DP-SGD when fine-tuning with lower privacy budgets ($\varepsilon = 8$). Under a more generous budget ($\varepsilon = 16$), FactScore does increase compared to the baseline (mean: 33.8%), but still falls short of the model trained without DP (mean: 37.7%).

FactScore is higher for the Science articles compared to AI articles; median scores are consistently $\geq 50\%$. Furthermore, the DP model with $\varepsilon = 8$ achieves a higher mean FactScore than the base model for the Science articles, suggesting that DP fine-tuning hallucinates less than the base model. The differences between the AI and Science datasets likely results from their differing overlap with pre-training data: concepts from Science articles are more likely to occur in pre-training data, even if exact articles are non-overlapping. While these experiments evaluate factuality on articles not seen during pre-training, in Appendix D.5, we also test whether DP fine-tuning perturbs knowledge already encoded in the base model from pre-training. We find insignificant differences between the DP and non-DP models, which suggests that the noise injected during DP fine-tuning does not disrupt knowledge acquired in non-private pre-training.

Human Evaluation. Table 3 reports the average ratings selected by human annotators. In lower-quality generations (FactScore ≤ 0.5), annotators rated both the non-DP and DP models with equally low veracity, while the DP model exhibited more unsupported facts. In higher-quality generations (FactScore > 0.5), annotators rated the DP model as outputting both lower veracity information and more unsupported facts. There are not conclusive differences in counts of quality issues. While agreement between human annotators was generally high, model-human agreement was not always high , and the model–annotator agreement on support was substantially lower ($\kappa = 0.23$–0.49; see Table 15), indicating that automated evaluations of factuality are an imperfect proxy for human judgment of how factual a response is. Regardless, overall trends are consistent between human and automated evaluations: both indicate greater hallucination in the DP model, even the model with a more generous privacy budget.

Additional diagnostics. We conduct further diagnostics to analyze the behavior of DP models, and report our findings in the appendix. First, we find that DP fine-tuning does not appear to disrupt knowledge already acquired during pre-training (§D.5). Second, a natural concern is that the DP fine-tuned models revert to their pre-trained priors during generation, rather than acquiring knowledge from unseen fine-tuning data; our analyses in §D.7



(a) Wikipedia Pretrained

(b) Wikipedia Science

Figure 3: Kernel density (KDE) plots comparing FactScore distributions for models pre-trained with (VaultGemma $\epsilon = 2$) and without DP (Gemma models & GPT2-XL). GPT2-XL, which has not seen the data during training, and VaultGemma (DP pre-trained) both lag behind the Gemma models in factual consistency.

demonstrates that this is not the case. Third, our experiments in §D.1 show that DP models tend to output the same repeated hallucinations, suggesting that there may be structured shifts in the output distribution that increase the likelihood of certain incorrect alternatives. Finally, in addition to standard DP-SGD, we evaluate the adaptive noise allocation algorithm proposed in (Li et al., 2024b) and find that it does not help mitigate the privacy-hallucination tradeoff; we defer the results and the discussion of its privacy analysis to §E.1.

3.2 DP pre-training and hallucinations

We investigate the effect of DP pre-training, as opposed to fine-tuning, on hallucinations in Table 4. It is worth noting that the pre-training data mixtures for the Gemma models likely include *all* the datasets used in our factual evaluations, whereas GPT-2 is not expected to have been trained on any of our factual evaluation datasets.

We find that the DP pre-trained model (VaultGemma) consistently outputs a greater percentage of inaccurate facts relative to every non-private Gemma baseline. On the Wikipedia pre-training data, Gemma3, the lowest-scoring Gemma baseline for this data, achieves an average FactScore of 26.6, compared to 22.0 for VaultGemma. Gemma-2B and Gemma2-2B score higher still (with average FactScores of 29.3 and 28.7), despite Gemma2-2B sharing VaultGemma's pre-training mixture. This gap in factual correctness is more pronounced for the domain-specific scientific and AI articles, where the difference between Gemma3-1B and VaultGemma reaches 10.9 points (50.4 vs. 39.5) and 8.7 points (51.9 vs. 43.2), respectively, with the other Gemma models ahead of VaultGemma by comparable margins. All the differences between the VaultGemma and Gemma models are statistically significant (§D.2), and the FactScore distributions reflect this as well, (Figure 3) where VaultGemma outputs have a higher density of low FactScores than any of the Gemma models.

	Response-level				Topic-level
Model	**Avg FS**	**Med**	**Q1**	**Q3**	**Avg FS [95% CI]**
Wikipedia AI					
Gemma3	51.9	55.6	25.0	79.6	51.6 [46.4, 56.7]
VaultGemma	43.2	43.3	17.0	66.7	43.1 [38.5, 47.7]
GPT-2 XL	**27.7**	**22.2**	**7.6**	**45.5**	**27.7** [24.3, 31.5]
Gemma2	49.4	50.0	21.4	77.8	49.4 [44.2, 54.5]
Gemma	49.6	50.0	23.1	77.1	49.4 [44.3, 54.5]
Wikipedia Science					
Gemma3	50.4	50.0	25.0	75.0	50.6 [47.4, 53.7]
VaultGemma	39.5	**33.3**	**14.6**	60.0	**39.1** [36.4, 41.9]
GPT-2 XL	**39.1**	**33.3**	16.9	**57.1**	39.3 [36.7, 42.0]
Gemma2	50.3	50.0	25.0	75.0	50.4 [47.4, 53.5]
Gemma	51.4	50.0	25.0	76.9	51.5 [48.3, 54.8]
Wikipedia Pretrain					
Gemma3	26.6	17.6	7.1	42.9	26.7 [23.9, 29.6]
VaultGemma	22.0	14.3	**0.0**	33.3	22.0 [19.6, 24.7]
GPT-2 XL	**17.6**	**11.1**	**0.0**	**25.0**	**18.0** [15.9, 20.2]
Gemma2	28.7	21.4	7.7	44.4	28.7 [25.9, 31.6]
Gemma	29.3	21.4	7.1	45.5	29.3 [26.2, 32.7]

Table 4: FactScore (FS; %) of pre-trained models at $\tau = 0.3$. Response-level statistics pool all generations; topic-level averages within each entity first, with bootstrap 95% CIs resampled over topics. Bolding indicates lower FactScore (increased hallucinations).*Gemma3 = Gemma3-1B-PT, VaultGemma = VaultGemma-1B (DP, $\epsilon = 2$), GPT-2 XL (1.5B), Gemma2 = Gemma-2-2B, Gemma = Gemma-2B. Significance tests for these differences against VaultGemma are reported in §D.2

Compared to GPT-2 XL, VaultGemma does have higher average FactScores for AI and pre-training articles. As the cutoff date for GPT-2 XL's pre-training data was in 2019, this model had no ex-

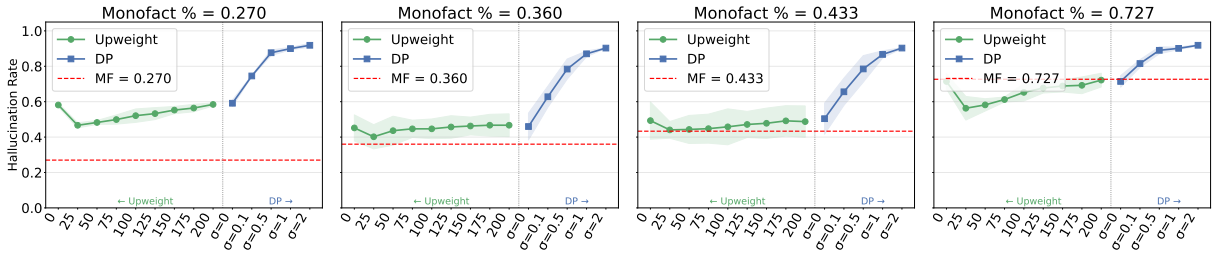


Figure 4: Hallucination rate of a bigram model under the selective upweighting strategy vs. DP Laplace noise.

posure to most concepts in the AI articles, thus constituting an extremely low bar for factual correctness in this setting. The improved FactScores of VaultGemma over the Wikipedia pre-training may be a reflection of general improvements in LLM development over the last 5 years that are not undone by DP training. More surprisingly, VaultGemma fails to output more factually correct information than GPT-2 XL in the Wikipedia Science setting, even though these articles were all created after 2020, suggesting they were included in VaultGemma training data and not GPT-2 XL data (the difference between scores is not significant at the response level ($p = .752$), and the topic-level test favors GPT-2 XL ($p = .028$)). These results suggest DP-SGD pre-training can substantially weaken a model's ability to encode and output factually correct information, with DP model outputs sometimes as hallucinated as outputs from a model never directly exposed to the targeted information.

4 RQ2: What impact does DP training have on model properties related to hallucination?

4.1 DP miscalibration increases hallucinations

Having established that DP training increases hallucinations, we investigate possible underlying mechanisms. We ground this investigation in recent theory (Kalai and Vempala, 2024) that shows that the hallucination rate of a model is governed by its calibration and the frequency of facts in its training data.

Formally, Kalai and Vempala (2024) represent this using the monofact rate $\widehat{\text{MF}}$, the fraction of facts that appear exactly once in the training corpus, which is used to approximate the probability mass of facts the model has never seen. A calibrated model, whose predicted probabilities match the observed frequencies from the training data, thus spreads the probability mass over the *unobserved* candidates from the unseen data, and (Kalai and Vempala, 2024) show that this yields a lower bound on the hallucination rate such that:

$$f_{\text{gen}} \gtrapprox \widehat{\text{MF}} - \text{Mis}(g, p) \quad (1)$$

where $\widehat{\text{MF}}$ is the monofact rate, $\text{Mis}(g, p)$ is the divergence of the model's learned distribution from the true distribution of factual frequencies. This implies that a calibrated model cannot do better than the monofact rate, but that miscalibration could decrease hallucinations.

Building on this, Miao and Kearns (2026) construct and validate an empirical analog of the theory of Kalai and Vempala (2024) and show that targeted miscalibration (via selective upweighting of training samples) that concentrates probability mass on high-confidence factual associations, can reduce hallucinations by up to 40%. We investigate the effects of DP-induced miscalibration on hallucination rates by replicating the bigram model setup from Miao and Kearns (2026), as it allows us to study the effect of the upweighting and DP-based interventions on the model's hallucination rate using a controlled synthetic setup where the true distribution of facts is known.

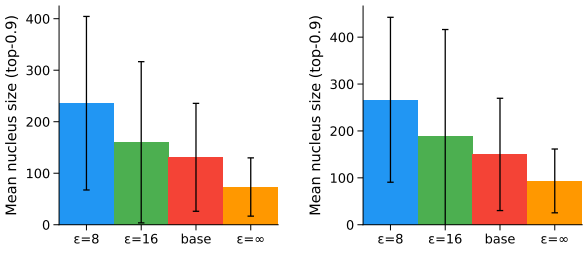
Setup. We replicate the bigram setup almost identically to the experiments from Miao and Kearns (2026), except that we use synthetic facts of the form $\texttt{Person}_i\ \texttt{Food}_j\ \texttt{Place}_k$, with the vocabulary constituting 40, 25 and 25 unique entities respectively, giving us a total of 25000 possible statements, out of which we designate $\mathcal{F} = 2000$ of these as "true" facts. The frequency of each true fact is drawn from a Pareto distribution, which also allows us to modify the $\widehat{\text{MF}}$ (a small α concentrates frequency on few facts, which lowers $\widehat{\text{MF}}$ whereas a large α spreads it across many unique facts, raising $\widehat{\text{MF}}$). From this, a training set of $|S| = 200$ samples is drawn with replacement, on which we fit a bigram model to predict the token that succeeds a given token. During the evaluation phase, the model generates 3000 tuples, over which we report

the hallucination rate (defined as the proportion of generated tuples not belonging to $\mathcal{F}$) in Figure 4.

The upweighting strategy from Miao and Kearns (2026) injects k training facts, each of which we duplicate 10 times so as to make the model overconfident on those tuples when renormalizing the modified transition count tables into updated probabilities. For our DP comparison, we add Laplacian noise to every entry of all possible transition counts from the vocabulary space, with L_1 sensitivity of 3, with all negative counts being clamped to zero. DP differs from strategic upweighting in that upweighting concentrates probability mass on observed facts, whereas DP is a non-targeted intervention that can disperse mass onto unseen transitions as well. We report the results in Figure 4, where upweighting is found to produce a modest, non-monotonic decrease in hallucination, in contrast to DP noise that progressively inflates hallucination rates as the scale of noise added increases.

4.2 Hallucinations in DP fine-tuned LLMs

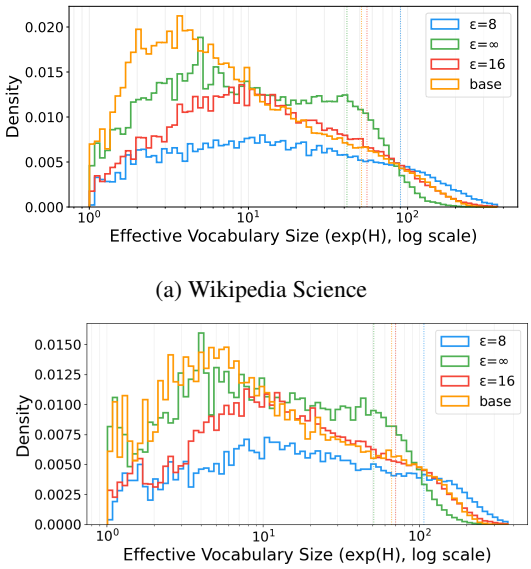
The results over the bigram models raise the question: do we observe the same shifts in probability mass in DP fine-tuned LLMs? To study this, we quantify the model's uncertainty at each position in the sequences from the unseen evaluation data in two ways: i) using the nucleus size (the smallest set of tokens accounting for 90% of the probability mass) (Figure 5) and ii) the effective vocabulary size, which is the exponentiated entropy of the next-token distribution. It represents how many tokens the model's predictive uncertainty is spread across (Figure 6). Both metrics reveal the same distributional dispersion as in the bigram models, which grow more pronounced under stricter privacy budgets.



(a) Wikipedia Science (b) Wikipedia AI

Figure 5: Mean nucleus size (top-$p = 0.9$) for models fine-tuned under different DP budgets. DP models exhibit larger nucleus sizes, indicating greater dispersion of probability mass across tokens and increased next-token uncertainty.

While higher entropy alone does not entail hallucination, it can foster conditions under which it becomes more likely. DP mechanisms produce effects on the next-token probability distribution that are analogous to increasing the sampling temperature, which has been shown to increase hallucination rates (Chang et al., 2025). However, altering the temperature rescales the distribution in a manner that preserves the rank order of tokens. DP mechanisms do not flatten the distribution evenly and the results from the bigram experiments suggest that they can spread mass from plausible sequences toward incorrect alternatives, increasing the likelihood of hallucination. Our results in Appendix D.6 support this possibility; on pre-training data, the token rankings are preserved and hallucination rates are largely unaffected, whereas on the unseen data, these rankings shift and we observe an increase in hallucination rates.



(a) Wikipedia Science

(b) Wikipedia AI

Figure 6: Distribution of effective vocabulary size ($\exp(H)$), where higher values indicate probability mass spread across more tokens. DP models show higher effective vocabulary sizes, reflecting increased distributional uncertainty.

5 RQ3: Under what conditions could DP be usable without increasing hallucinations?

Having shown that DP training increases hallucinations and investigated output distribution flattening as a possible underlying mechanism, we now ask a more fundamental question: under what conditions can a model trained with DP learn correct factual associations? The practical feasibility of

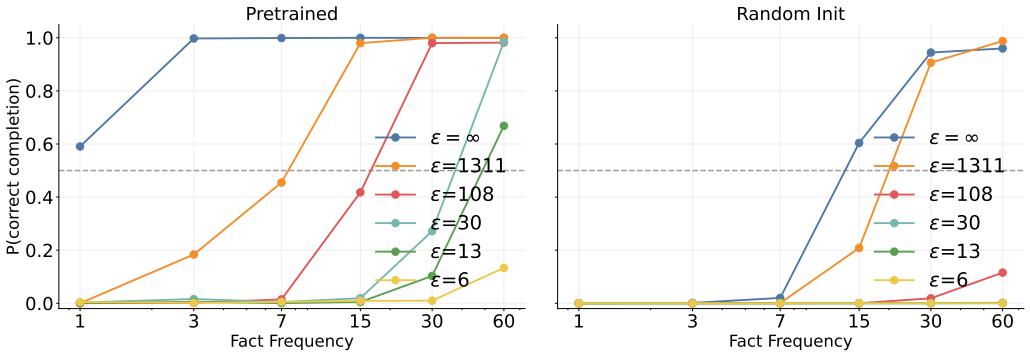


Figure 7: Factual recall vs. fact frequency under DP-SGD for pre-trained (left) and randomly initialized (right) GPT-2. Stricter privacy budgets (ϵ) require increasingly frequent repetitions for learning.

DP training hinges on this question.

To study this, we construct a controlled memorization experiment using synthetic factual associations that take the form of entity triples (e.g., the capital of [subject] is [object]) sampled from a set of fictional subjects and objects to ensure that none of these facts can be answered from prior knowledge. Every fact appears f times per epoch ($f \in \{1, 3, 7, 15, 30, 60\}$), i.e. $15f$ times over the full training run, with five distinct facts per frequency tier. To approximate knowledge acquisition in fine-tuning and pre-training, we use the pre-trained version of GPT-2 and a randomly initialized version of GPT-2, both of which we train for 15 epochs on the synthesized facts with different privacy budgets. We fix the clipping norm at $C = 1.0$ and vary the amount of noise added during training, $\sigma \in \{0, 0.1, 0.3, 0.5, 0.7, 1.0\}$, where a higher σ corresponds to a stronger privacy guarantee. Each setting corresponds to a privacy budget $\varepsilon \in \{\infty, 1311, 108, 30, 13, 6\}$. We report a detailed version of this setup in Appendix E.2. In this setup, we treat factual recall as the probability that the model assigns to the ground-truth answer given the fact prefix.

As ε decreases (and the corresponding noise multiplier σ increases), the minimum frequency required for the pre-trained model to learn an association rises sharply as well (Figure 7). Although at $\varepsilon = \infty$, the pre-trained model recalls facts that appear only once per epoch, at $\varepsilon = 1311$, this increases to 30 repetitions, and to 60 repetitions at $\varepsilon = 30$. For stricter budgets (where $\varepsilon \leq 13$), no fact is learned even at 60 repetitions per epoch, indicating that DP noise does not simply slow down learning, but can prevent it entirely within the frequency range that we test. Furthermore, there is a stark gap between the pre-trained model and the randomly initialized setup. While the pre-trained model requires 7 repetitions per epoch at $\varepsilon = 1311$, the randomly initialized model requires 30 repetitions for the same ε. Beyond this, no fact is learned at any strict ε within our frequency range.

This has implications for DP fine-tuning and DP pre-training. In practice, factual associations are unlikely to appear so frequently in a training corpus and will likely not be acquired by models fine-tuned over sensitive data with meaningful DP guarantees. For DP pre-training, which already entails high compute and memory costs, these frequency thresholds are considerably worse. Overall, our findings highlight the urgent need for methods that can balance meaningful privacy guarantees while still preserving the reliability of model outputs.

6 Discussion and Conclusions

We reveal and systematically analyze a privacy–hallucination tradeoff in differentially private language models. Our results show that differential privacy significantly hinders the acquisition of new factual associations during pre-training and fine-tuning, leading to hallucinations. Our investigation into possible underlying mechanisms reveals that DP shifts model outputs toward factually incorrect content, and that moderate levels of noise can preclude the acquisition of a factual association.

Overall, our findings highlight how existing DP training methods can inadvertently compromise factual reliability, which is of particular concern in high-stakes applications. This underscores the need for refined privacy-preserving approaches that balance privacy guarantees with factual accuracy. Addressing this tradeoff is essential for deploying trustworthy AI systems in sensitive domains such as healthcare and law, where both privacy and factual consistency are non-negotiable.

Limitations

Our experiments are limited to open-source models with known pre-training data cutoffs, to ensure the availability of sufficient unseen evaluation and training data for our experimental setup to adhere to differentially private constraints.

We do not investigate methods to decrease hallucinations, such as specialized post-training objectives to steer models toward expressing uncertainty or refusal over generating incorrect content. Given that fully mitigating hallucinations remains an unsolved problem even in non-DP settings, we do not expect such approaches to eliminate the risk of hallucination, although they could still reduce the tradeoff characterized in this work. Additionally, while this work focuses on DP training, we leave the investigation of this tradeoff in other DP approaches (such as DP retrieval-augmented generation (Koga et al., 2024; Grislain, 2025)) to future work, as maintaining privacy guarantees under multiple queries remains challenging, and more research is needed to assess if hallucinations persist in these settings.

Acknowledgements

The authors would like to thank the reviewers for their helpful feedback. This work was supported in part by the AI2050 Fellowship program by Schmidt Sciences.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Michael Aerni, Javier Rando, Edoardo Debenedetti, Nicholas Carlini, Daphne Ippolito, and Florian Tramèr. 2025. Measuring non-adversarial reproduction of training data in large language models. In *The Thirteenth International Conference on Learning Representations*.

Michael Aerni, Jie Zhang, and Florian Tramèr. 2024. Evaluations of machine learning privacy defenses are misleading. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1271–1284.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. *Differential privacy has disparate impact on model accuracy*. Curran Associates Inc., Red Hook, NY, USA.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, pages 267–284. USENIX Association.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, and Tagyoung Chung. 2025. REAL sampling: Boosting factuality and diversity of open-ended generation by extrapolating the entropy of an infinitely large LM. *Transactions of the Association for Computational Linguistics*, 13:760–783.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junjie Chu, Zeyang Sha, Michael Backes, and Yang Zhang. 2024. Reconstruct your previous conversations! comprehensively investigating privacy leakage risks in conversations with GPT models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6584–6600, Miami, Florida, USA. Association for Computational Linguistics.

Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes, Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra, and 10 others. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, and 1 others. 2024. Advancing differential privacy: Where we are now and future directions for real-world deployment. *Harvard Data Science Review*.

Souvik Das, Sougata Saha, and Rohini Srihari. 2022. Diving deep into modes of fact hallucinations in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 684–699, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anderson Santana de Oliveira, Caelin Kaplan, Khawla Mallat, and Tanmay Chakraborty. 2024. An empirical analysis of fairness notions under differential privacy. *Preprint*, arXiv:2302.02910.

Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. 2022. Connect the dots: Tighter discrete approximations of privacy loss distributions. *Proceedings on Privacy Enhancing Technologies*, 4:552–570.

Jiacheng Du, Zhibo Wang, and Kui Ren. 2024. Textual unlearning gives a false sense of unlearning. *CoRR*, abs/2406.13348.

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

Cynthia Dwork, Aaron Roth, and 1 others. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. *Preprint*, arXiv:2009.06389.

Ferdinando Fioretto, Pascal Van Hentenryck, and 1 others. 2025. Differential Privacy in Artificial Intelligence: From Theory to Practice.

Nicolas Grislain. 2025. Rag with differential privacy. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 847–852. IEEE.

Abhimanyu Hans, John Kirchenbauer, Yuxin Wen, Neel Jain, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. 2024. Be like a goldfish, don't memorize! mitigating memorization in generative LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Victor Hansen, Atula Neerkaje, Ramit Sawhney, Lucie Flek, and Anders Søgaard. 2024. The impact of differential privacy on group disparity mitigation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3952–3965, Mexico City, Mexico. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. DP-rewrite: Towards reproducibility and transparency in differentially private text rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2927–2933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shotaro Ishihara. 2023. Training data extraction from pre-trained language models: A survey. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 260–275, Toronto, Canada. Association for Computational Linguistics.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024a. On large language models' hallucination with regard to known facts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053, Mexico City, Mexico. Association for Computational Linguistics.

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. 2024b. Core: Robust factual precision with informative sub-claim identification. *Preprint*, arXiv:2407.03572.

Seongho Joo, Hyukhun Koh, and Kyomin Jung. 2026. PEARL: Differentially private and entropy-aware regulated language generation. In *Forty-third International Conference on Machine Learning*.

Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, page 160–171, New York, NY, USA. Association for Computing Machinery.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, and Zheng Xu. 2024. User inference attacks on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18238–18265. Association for Computational Linguistics.

Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

Daniel Kifer and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204.

Tatsuki Koga, Ruihan Wu, and Kamalika Chaudhuri. 2024. Privacy-preserving retrieval-augmented generation with differential privacy. *arXiv preprint arXiv:2412.04697*.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2024. Harnessing large-language models to generate private synthetic text. *Preprint*, arXiv:2306.01684.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024a. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Xianzhi Li, Ran Zmigrod, Zhiqiang Ma, Xiaomo Liu, and Xiaodan Zhu. 2024b. Fine-tuning language models with differential privacy through adaptive noise allocation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8368–8375, Miami, Florida, USA. Association for Computational Linguistics.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11263–11282, Miami, Florida, USA. Association for Computational Linguistics.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365, Online. Association for Computational Linguistics.

Qian Ma and Sarah Rajtmajer. 2026. Private seeds, public LLMs: Realistic and privacy-preserving synthetic data generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 189–210, San Diego, California, United States. Association for Computational Linguistics.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cleo Matzken, Steffen Eger, and Ivan Habernal. 2023. Trade-offs between fairness and privacy in language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6948–6969, Toronto, Canada. Association for Computational Linguistics.

Ryan McKenna, Yangsibo Huang, Amer Sinha, Borja Balle, Zachary Charles, Christopher A Choquette-Choo, Badih Ghazi, George Kaissis, Ravi Kumar, Ruibo Liu, and 1 others. 2025. Scaling laws for differentially private language models. *arXiv preprint arXiv:2501.18914*.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2024. Did the neurons read

your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. 2025. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations*.

Miranda Muqing Miao and Michael Kearns. 2026. Hallucination, monofacts, and miscalibration: An empirical investigation. *Proceedings of the National Academy of Sciences*, 123(8):e2533582123.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, Fabio Massimo Zanzotto, Sébastien Bratières, and Emanuele Rodolà. 2025. Preserving privacy in large language models: A survey on current threats and solutions. *Preprint*, arXiv:2408.05212.

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, Online. Association for Computational Linguistics.

Ivoline C. Ngong, Joseph Near, and Niloofar Mireshghallah. 2025. Differentially private learning needs better model initialization and self-distillation. In *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, Albuquerque, New Mexico. Association for Computational Linguistics.

Eshaan Nichani, Jason D. Lee, and Alberto Bietti. 2024. Understanding factual recall in transformers via associative memories. *Preprint*, arXiv:2412.06538.

Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, Minhui Xue, and Bo Li. 2024. Reconstruction of differentially private text sanitization via large language models. *Preprint*, arXiv:2410.12443.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Krishna Pillutla, Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. 2024. Federated learning with superquantile aggregation for heterogeneous data. *Machine Learning*, 113(5):2955–3022.

Krishna Pillutla, Jalaj Upadhyay, Christopher A Choquette-Choo, Krishnamurthy Dvijotham, Arun Ganesh, Monika Henzinger, Jonathan Katz, Ryan McKenna, H Brendan McMahan, Keith Rush, and 1 others. 2025. Correlated Noise Mechanisms for Differentially Private Learning. *arXiv preprint arXiv:2506.08201*.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ML: A practical guide to machine learning with differential privacy. *J. Artif. Intell. Res.*, 77:1113–1201.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15254–15269, Miami, Florida, USA. Association for Computational Linguistics.

Igor Shilov, Matthieu Meeus, and Yves-Alexandre de Montjoye. 2026. The mosaic memory of large language models. *Nature Communications*, 17(1):2142.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Bowen Tan, Zheng Xu, Eric P. Xing, Zhiting Hu, and Shanshan Wu. 2025. Synthesizing privacy-preserving text data via finetuning *without* finetuning billion-scale LLMs. In *Forty-second International Conference on Machine Learning*.

Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, and Nils Lukas. 2026. DP-fusion: Token-level differentially private inference for large language models. In *The Fourteenth International Conference on Learning Representations*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *Preprint*, arXiv:2311.08401.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. 2024. Position: Considerations for differentially private learning with large-scale public pretraining. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. 2021. Decision making with differential privacy under a fairness lens. In *IJCAI*, pages 560–566.

VaultGemma Team. 2025. VaultGemma: A differentially private Gemma model. Google Technical Report.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

Xinwei Wu, Weilong Dong, Shaoyang Xu, and Deyi Xiong. 2024. Mitigating privacy seesaw in large language models: Augmented privacy neuron editing via activation patching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5319–5332, Bangkok, Thailand. Association for Computational Linguistics.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.

Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2024. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. In *Neurips Safe Generative AI Workshop 2024*.

Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan Mcmahan, Jesse Rosenstock, and Yuanbo Zhang. 2023. Federated learning of gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 629–639, Toronto, Canada. Association for Computational Linguistics.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On Protecting the Data Privacy of Large Language Models (LLMs): A Survey . In *2024 International Conference on Meta Computing (ICMC)*, pages 1–12, Los Alamitos, CA, USA. IEEE Computer Society.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. 2022. Opacus: User-friendly differential privacy library in pytorch. *Preprint*, arXiv:2109.12298.

Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give me the facts! a survey on factual knowledge probing in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15588–15605, Singapore. Association for Computational Linguistics.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *Preprint*, arXiv:2404.05868.

Xinyu Zhou and Raef Bassily. 2024. Differentially private worst-group risk minimization. *arXiv preprint arXiv:2402.19437*.

Erion Çano and Ivan Habernal. 2025. Differentially-private text generation degrades output language quality. *Preprint*, arXiv:2509.11176.

Appendix Contents

A Background and Related Work **15**
- A.1 Related Work 15
- A.2 Background: Differential Privacy . 16
- A.3 FactScore and Recurring Claims Algorithms 17

B Definition of Hallucination **17**

C Experimental Setup **18**
- C.1 Datasets 18
- C.2 Experimental Setup: Methods and Hyperparameters 19
- C.3 Dataset : Examples of Input . . . 21
- C.4 List of topics in the datasets 24
- C.5 Training Loss Curves for Wikipedia Fine-tuning 25

D Detailed Experimental Results **27**
- D.1 Analysis of Recurring Hallucinations 27
- D.2 Statistically Significant Differences in FactScore 29
- D.3 Generation Length Statistics . . . 29
- D.4 Lower Perplexity Does Not Correlate With Factual Reliability . . . 30
- D.5 Hallucination of pre-training facts in DP fine-tuned models 32
- D.6 Distributional flattening for pre-training data 32
- D.7 Is the model reverting to its pre-training priors? 32
- D.8 Manual Annotations: Claim Annotation Interface 36
- D.9 Expanded FactScore Evaluations . 39
- D.10 Variability in FactScore 39

E Supplementary Experiments **45**
- E.1 ANADP: Adaptive Noise Allocation for Differential Privacy 45
- E.2 Controlled Fact Acquisition Experiment 46
- E.3 Layer-wise Signal-to-Noise Ratio 47
- E.4 Additional Examples of Claim Clusters 47

F LLM Usage **47**

A Background and Related Work

A.1 Related Work

Privacy in LLMs. This risk of privacy leakage by language models has inspired work on provable privacy-preserving strategies such as differential privacy (Li et al., 2022; Miranda et al., 2025), as well as heuristics such as knowledge unlearning to reduce the influence of sensitive data points on the model parameters (Jang et al., 2023; Zhang et al., 2024) and knowledge editing to locate and modify neurons containing private information (Wu et al., 2023, 2024).

Privacy-preserving methods typically (either directly or indirectly) involve changes to the model's parameters. For instance, DP training is typically accomplished through a modified version of SGD, which involves clipping and noising the gradient update (Abadi et al., 2016). Knowledge unlearning methods trace and remove an approximate estimate of the influence of a training point on the model's parameters. All of these approaches contribute to some degradation in model utility: DP limits the information the model learns, while knowledge unlearning and knowledge editing affects useful non-private information the model has encoded. We focus on the rigorous and future-proof guarantees of DP since heuristic privacy defenses can often be broken (Aerni et al., 2024; Du et al., 2024).

Tradeoffs from DP. The classical no-free-lunch theorem of DP states that DP necessarily incurs a penalty on utility (Kifer and Machanavajjhala, 2011). In practice, this results in a privacy-utility-compute tradeoff (McMahan et al., 2018; Ponomareva et al., 2023). Research has since established that privacy also comes at a cost to fairness in statistical estimation tasks (Tran et al., 2021), discriminative models (Bagdasaryan et al., 2019; Farrand et al., 2020; de Oliveira et al., 2024), and LLMs (Lyu et al., 2020; Matzken et al., 2023; Ramesh et al., 2024; Hansen et al., 2024). Ngong et al. (2025) explore the adverse effect of DP on elements such as grammatical correctness, fluency and the coherence of model-generated text. We note that it is possible to optimize for more balanced privacy-utility tradeoffs (Mireshghallah et al., 2021) or privacy-fairness tradeoffs (Pillutla et al., 2024; Zhou and Bassily, 2024).

Specifically in the text domain, prior work has empirically examined the privacy-utility tradeoffs with task-specific measures of utility, including

classification accuracy, linguistic aspects such as fluency, grammatical correctness and lexical diversity. In contrast, we focus on tradeoffs between DP and hallucination. Minimizing hallucinations is highly desirable, and is often distinct from other task-specific measures of utility surveyed above. Concurrent with our work, Joo et al. (2026) find that DP-decoding in RAG settings increases hallucinations in models, attributing this to a similar DP-induced flattening effect which amplifies knowledge conflict, which is measured as the gap between the retrieval-conditioned predictions and the model's parametric prior.

Factuality in Language Models. No single factuality metric generalizes across settings, so it is common to use task- or domain-specific methods. Early research focused on measures such as the factual precision of cloze-style or short-form responses (Youssef et al., 2023; Petroni et al., 2019), and NLI-based methods to determine whether generated summaries are consistent with their source document (Chen et al., 2021; Fabbri et al., 2022; Laban et al., 2022). The rise of LLM applications with open-domain, free-form model-generated text, where there can be multiple plausible responses has led to the development of factuality metrics in these settings (Min et al., 2023; Wei et al., 2024; Song et al., 2024). This is significantly more challenging, as the decoding strategy also influences model outputs (Wang et al., 2024), and models can generate correct answers across multiple attempts (Tian et al., 2023). Further, they may generate factually accurate content that contradicts or is unsupported by the training data (Cao et al., 2022).[4] We leverage existing state-of-the-art methods for fact-checking in open-ended generation along with human evaluations to analyze privacy-hallucination tradeoffs in LLMs (Min et al., 2023).

A.2 Background: Differential Privacy

Differential privacy offers a formal privacy guarantee that ensures that any individual's data cannot be inferred from a query applied to a dataset (Dwork et al., 2006). In other words, the result of such a query is nearly indistinguishable from the result of the same query applied to a dataset that either includes a modified version of the individual's data or excludes the record entirely, thereby preserving the individual's privacy. In this case, the notion of adjacency specifies exactly how changing single record in the original dataset D yields the modified dataset D'.

We review the definition of DP here; we refer to the textbooks (Dwork et al., 2014; Fioretto et al., 2025) for specific details and the guide (Ponomareva et al., 2023) for details on DP-SGD.

Formally, differential privacy is defined as follows:

Definition: Two datasets D and D' are said to be neighboring (in the add-or-remove sense) if $D = D' \cup \{x\}$ or $D' = D \cup \{x\}$ for some training example x.

A randomized algorithm A is (ϵ, δ)-private for some $\epsilon > 0$ and $\delta \in [0, 1]$if for *any* two neighboring datasets D, D', the following holds true for all sets Y in the range of A:

$$\Pr[A(D) \in Y] \leq e^{\epsilon} \Pr[A(D') \in Y] + \delta.$$

The value of ϵ denotes the privacy budget, while δ specifies the likelihood that the privacy guarantee may fail. If δ is set to 0, this implies a purely differentially private setting with no probability of the guarantee being broken. The value of ϵ constrains how similar the outputs of both distributions are; a higher ϵ value indicates a greater privacy budget, meaning the algorithm is less private. DP guarantees that even if an adversary has access to any side-knowledge, the privacy leakage of (ϵ, δ)-DP algorithms will not increase. Additionally, another property of DP is that it ensures that any post-processing on the outputs of (ϵ, δ)-differentially private algorithms will remain (ϵ, δ)-differentially private.

We use DP-SGD (Abadi et al., 2016), a modification to the stochastic gradient descent (SGD) algorithm, which is typically used to train neural networks. DP-SGD clips the gradients to limit the contribution of individual samples from the training data and subsequently adds noise from the Gaussian distribution to the sum of the clipped gradients across all samples.[5] DP-SGD thus provides a differentially private guarantee to obfuscate the gradient update, thereby ensuring that the contribution of any given sample in the training data is indistinguishable due to the aforementioned post-processing property. This process ensures (ϵ, δ)-differential privacy for each model update. Given a

[4] We regard this as hallucinations for the purpose of this work as the training dataset is considered as the sole *source of truth.*

[5] The noise added is independent coordinate-wise and across time, although DP-SGD with temporal correlations has recently been of growing interest (Pillutla et al., 2025).

privacy budget, number of epochs, and other training parameters, we can estimate the privacy parameters using standard privacy accounting algorithms, which implemented in common software.

A.3 FactScore and Recurring Claims Algorithms

We recall the pseudo-code of FactScore in Algorithm 1 and describe our algorithm to cluster repeated claims in Algorithm 2 (cf. §D.1).

A.3.1 FactScore Algorithm

Algorithm 1 FActScore: Atomic Fact Extraction and Verification

Input: Generated texts $\mathcal{D} = d_1, d_2 ... d_n$, atomic fact extractor module $\mathcal{E}$, claim verification model $\mathcal{V}$, knowledge source $\mathcal{K}$

Output: FactScore for each document in the generated corpus : $\mathcal{S}(\mathcal{E}, \mathcal{D})$

1: **for** each document $d_i \in \mathcal{D}$ **do**

2: Extract a candidate set of atomic claims:

$$\mathcal{AF}_{d_i} = \mathcal{E}(d_i)$$

3: **for** each atomic claim $\alpha_j^{(d_i)} \in \mathcal{AF}_{d_i}$ **do**

4: Verify factuality if $\alpha_j^{(d_i)}$ is supported by knowledge source $\mathcal{K}$:

$$\hat{y}_j^{(d_i)} = \mathcal{V}(\alpha_j^{(d_i)}, \mathcal{K}) \quad \text{where } \hat{y}_j^{(d_i)} \in \{0, 1\}$$

5: **end for**

6: Compute per-document precision:

$$\mathcal{S}(\mathcal{E}, g) = \frac{1}{|\mathcal{AF}_{d_i}|} \sum_{j=1} \mathbb{I}(\hat{y}_j^{(d_i)} = 1)$$

7: **end for**

Algorithm 1 describes how the FactScore is computed for a set of generated documents. For each document d_i, the atomic fact extractor $\mathcal{E}$ which is an instruction-tuned LLM, is prompted with in-context examples to decompose the text into a set of atomic claims, denoted $\mathcal{AF}_{d_i}$.

Each atomic claim $\alpha_j^{(d_i)} \in \mathcal{AF}_{d_i}$ is then independently verified using an external knowledge source $\mathcal{K}$. The knowledge source is the reference data against which claims are verified, and in our setup, this consists of the relevant Wikipedia articles for the evaluation domain. For each atomic claim, evidence passages are retrieved (e.g. via BM25) from these articles which are then provided to the claim verification model. The claim verification model ($\mathcal{V}$) then judges whether the claim is supported by this knowledge, producing a binary label $\hat{y}_j^{(d_i)} \in \{0, 1\}$ that indicates whether or not the claim is factually supported.

The FactScore for the document is computed as follows:

$$\mathcal{S}(\mathcal{E}, d_i) = \frac{1}{|\mathcal{AF}_{d_i}|} \sum_j \mathbb{I}(\hat{y}_j^{(d_i)} = 1),$$

which yields a scalar score corresponding to the factual correctness of the information in the generated document with respect to the knowledge source. A higher score translates to a greater proportion of verified claims, and conversely, a lower score is indicative of fewer supported claims.

We use Llama-3.1-8B-Instruct, to perform both the atomic fact extraction as well as the atomic claim verification. To demonstrate that our results remain consistent across different choices in the claim decomposition and claim verification modules used, we also report results over other instruction-tuned LLMs such as DeepSeek-R1-Distill-Qwen-7B and Llama 3.2-3B Instruct in Table 16 and Table 17.

B Definition of Hallucination

The term "hallucination" typically refers to the generation of factually incorrect information by generative models. For this work, we focus on a class of facts supported by the model's data.

Definition: We say a model *hallucinates* if it generates a claim that is not supported by the model's training data. Hallucinations w.r.t. finetuning are claims that are not supported by the finetuning data and hallucinations w.r.t. pretraining are claims not supported by the pretraining data.

Here are a few consequences of our definitions:

- If the model generates a correct fact that is not grounded in its finetuning (resp. pretraining) data, we consider it a hallucination.
- If the model generates a correct fact that is grounded in the pretraining data but not the finetuning data, we consider it a finetuning hallucination.
- Conversely, a factually correct claim is not considered a hallucination provided it appears somewhere in the training data, even if it is absent from the specific source article being evaluated.

Our evaluation is tailored to this definition: we specifically target the model's ability to acquire and reproduce facts from previously unseen data. This definition allows us to directly assess how DP influences model learning. Failure to learn a fact is a different problem from hallucination: in the absence of knowledge acquisition, a model could output no content or generic content, which we might expect to be the case, as prior work has shown DP outputs tend to be shorter than non-DP outputs (Ngong et al., 2025; Çano and Habernal, 2025). Instead our work highlights that DP increases hallucinations, which is not inherently implied by the definition of DP.

C Experimental Setup

C.1 Datasets

Choosing a dataset for this study requires careful consideration of two factors. First, DP guarantees hinge on the assumption that the private fine-tuning data should not have appeared in the pre-training corpora of the LLM (Tramèr et al., 2024; Cummings et al., 2024). The importance of not violating this condition can be attributed to i) the potential for pre-training data to be adversarially extracted (Ishihara, 2023), and ii) evidence that pre-training and fine-tuning on the same data artificially inflates performance estimates (Igamberdiev et al., 2022). Second, the inclusion of factually verifiable information and statements in the fine-tuning data is essential to evaluate changes in the factual correctness of the model's outputs; domains and datasets (e.g., social media posts) without clear factual content cannot be assessed for factuality.

In view of these two factors, we focus on Wikipedia data for fine-tuning and evaluation, where content is constructed to contain verifiable facts rather than opinions or speculation, and automated fact-checking methods have been previously validated (Min et al., 2023). Additionally, metadata allows us to select articles created after 2020, ensuring they were not included in GPT-J 6B's pre-training corpora. We expect our fine-tuning data to have some overlap with text in the pre-training data in terms of linguistic patterns, broad concepts, and topics (in some settings). This overlap is not inherently problematic, as it reflects natural language settings, where syntactic and semantic structures are rarely novel and learning dynamics are influenced by previously learned distributions.

C.1.1 Evaluation Datasets

We use three datasets for factuality evaluation as summarized in Table 1 and described below. An exact list of topics included in these datasets is given in Appendix C.4.

Wikipedia Science. We collect 231 Wikipedia articles on science topics that were created after the cutoff date for GPT-J 6B's pre-training data, where we use keyword searching of Wikipedia meta-data to identify science articles. We focus on science topics as they contain detailed technical language, which is also common in sensitive data settings (e.g., clinical notes). While Wikipedia articles on these topics did not exist before 2020, we do expect some concepts to exist in other pre-training data sources, which makes it feasible for a DP model to produce facts on these topics, even without memorizing individual data points.

Wikipedia AI. We collect 124 Wikipedia articles on AI topics, where we hand-curate products and models that did not exist before 2020, along with related articles we expect to mention them. Unlike the Wikipedia Science articles, GPT-J 6B cannot have any knowledge of most of these concepts without fine-tuning, as they could not have existed in pre-training data. However, by constructing our data to contain articles that mention overlapping topics, we ensure that it is feasible for a DP model to learn them. For example, if our dataset only contained *DeepSeek (chatbot)*, DP would preclude learning of information isolated to one data point. By including *DeepSeek (chatbot)*, *DeepSeek*, and *DeepSeek (disambiguation)*, a DP model can hypothetically learn information about DeepSeek, as it is mentioned in multiple data points.

Wikipedia pre-training. To investigate effects of DP fine-tuning on knowledge acquired during pre-training, we randomly sample 250 Wikipedia articles from the GPT-J pretraining data that are not included in the fine-tuning data.

C.1.2 Fine-tuning Dataset

Fine-tuning with differential privacy generally requires large enough datasets and large batch sizes (e.g. $\Omega(10^3)$ or more) (McMahan et al., 2018; Ponomareva et al., 2023). However, since our curated evaluation sets are insufficient to meet these batch size specifications, we intersperse our collected articles with an additional 20,000 randomly sampled Wikipedia articles that likely occurred in the pre-training data. We ensure that these sam-

ples do not overlap with the Wikipedia pre-training dataset used for factuality evaluation.

The data is divided into sequences of 512 tokens, which is the unit of privacy protection. We fine-tune the model to produce an article when prompted on the article title (e.g., a topic). For evaluations, we similarly prompt the model with article titles and evaluate the factual accuracy of the generated text.

C.2 Experimental Setup: Methods and Hyperparameters

DP fine-tuning is achieved using a stochastic gradient optimization approach known as DP-SGD (Abadi et al., 2016). This algorithm bounds the information learned from each sample by clipping the per-sample gradients to a fixed ℓ_2 norm bound, and perturbs them (for DP) with white Gaussian noise. The scale of the Gaussian noise is calibrated to the desired (ϵ, δ)-DP guarantee. The DP guarantees are provided with respect to the add-or-remove adjacency at the sequence-level, i.e., the model outputs should be nearly indistinguishable if a new sequence of k tokens is added to or removed from the training dataset. We set privacy budgets of $\epsilon \in \{8, 16\}$, and $\delta = n^{-1.1}$, where n is the dataset size (in terms of number of sequences). We use the DP-SGD implementation from Opacus (Yousefpour et al., 2022) and measure the privacy budget consumed using the PLD accountant (with amplification by sampling) (Doroshenko et al., 2022).

All experiments were conducted on NVIDIA A100 GPUs. DP fine-tuning was performed using 1×A100 GPUs, with each DP fine-tuned model requiring approximately 24 GPU hours.

The PLD accounting algorithm proposed in Doroshenko et al. (2022) provides us with tighter estimates of the privacy loss as compared to alternate accounting techniques. This in turn allows us to more accurately determine the noise multiplier required to satisfy the specified privacy budget for our fine-tuning setups.

While VaultGemma's DP pre-training uses a unit of privacy of 1024 tokens (which includes separating and merging documents based on this fixed sequence length), our Wikipedia-based fine-tuning experiments use a single document as the privacy unit, which is reflective of settings where each document corresponds to a distinct individual.

Why VaultGemma is not data-matched. We evaluate against the two baselines used in the VaultGemma technical report (VaultGemma Team, 2025), Gemma3 (1B) and GPT-2 (1.5B), together with Gemma2 (2B) and Gemma (2B). No exactly matched non-private counterpart exists, but VaultGemma does share the same Gemma2 pretraining mixture as the Gemma2 (2B) model, and VaultGemma is approximately the same size as Gemma3 (1B) (although they differ in mixture). However, this does not affect our results as i) every Gemma model likely includes the Wikipedia data in our evaluation setup in its pre-training mixture, and ii) GPT-2 (1.5B) matches VaultGemma on Wikipedia Science despite being a 2019 model trained on vastly less data, and despite WebText excluding Wikipedia outright.

Additionally, data-matching would require DP pre-training from scratch, which is infeasible at the required scale. The utility in using DP-SGD's to train language models hinges on very large batch sizes. For reference, VaultGemma was trained with an expected batch size of ≈500k, and even smaller models like BertTiny require a batch size of ≈283k (McKenna et al., 2025). McKenna et al. (2025) make the same point via DP scaling laws, where at 10^{22} FLOPs, $\sim 10^8$ parameters are compute-optimal under privacy versus $\sim 10^{10}$ non-privately. Due to the intense computational demands for pre-training these models, we use VaultGemma as a reference point for what DP pre-training yields at comparable scale, not as a controlled ablation of pre-training data. We leave further exploration of this problem through the use of controlled re-training experiments to future work.

The hyperparameters from our DP fine-tuning setup are summarized as follows:

LoRA Gradient Update We apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) in all experiments. The adapted weight is parameterized as

$$W = W_0 + \frac{\alpha}{r} AB$$

where the LoRA matrices are $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{r \times k}$, and W_0 is the frozen pre-trained weight. During training, the LoRA parameters A, B are updated via:

$$(A^{(t+1)}, B^{(t+1)}) = (A^{(t)}, B^{(t)}) - \eta \cdot \text{clip}\left(\nabla_{A,B} L,\, c\right)$$

where η is the learning rate, $c = 1.0$ is the clipping norm, and L is the loss function. In the DP setting, noise is added to the clipped gradient.

Setting	Batch Size	Epochs	LR	Clip	LoRA r	LoRA α	Seq. Len.	DP/PLD
Vanilla	8	15	1×10^{-4}	1.0	4	32	512	–
DP, $\epsilon \in \{8, 16\}$	4096	20	1×10^{-4}	1.0	4	512	512	$\delta = 1/n^{1.1}$

Table 5: Summary of major hyperparameters in vanilla and private training settings.

During training, the update step for (A, B) in SGD or DP-SGD is:

$$(A^{(t+1)}, B^{(t+1)}) = (A^{(t)}, B^{(t)}) - \eta \cdot \widetilde{\nabla} L$$

where

$$\widetilde{\nabla} L = \text{clip}\left(\nabla L, c\right) + \mathcal{N}(0,\, \sigma^2)$$

$$\nabla_A L = \frac{\alpha}{r} \nabla_W L \cdot B^\top$$

Instead of using a large learning rate in the DP fine-tuning, we use a larger value of ($\alpha = 512$) in the DP setting (compared to ($\alpha = 32$) in the standard fine-tuning setting) to amplify the contribution of the adapted weights without increasing the noise magnitude (which is applied to the gradients directly). This prevents the gradient signal from being obscured by the DP noise.

Why LoRA rather than full fine-tuning. DP-SGD requires per-example gradients, and the resulting memory overhead makes full fine-tuning at this scale computationally infeasible for most academic labs, including our own. Additionally, given that LoRA-based DP fine-tuning has already been shown to be competitive with full fine tuning (Kurakin et al., 2024), and parameter-efficient adaptation is a standard practice in related work (Yu et al., 2022; Tan et al., 2025), our setup is reflective of common practices in this area.

C.3 Dataset : Examples of Input

Table 6: Examples of excerpts of the input and output pairs from our Wikipedia AI and Science articles.

Input	Output
Right to Know	Right To Know is a non profit support project for those who discover via genealogical genetic testing that their lineage is not what they had supposed it to be due to family secrets and misattributed parentage, thus raising existential issues of adoption, race, ethnicity, culture, rape, etc. == See also == Genealogy Genetic testing == External links == Right To Know - Your Genetic Identity.
Neurosemiotics	Neurosemiotics is an area of science which studies the neural aspects of meaning making. It interconnects neurobiology, biosemiotics and cognitive semiotics. Neurolinguistics, neuropsychology and neurosemantics can be seen as parts of neurosemiotics. == Description == The pioneers of neurosemiotics include Jakob von Uexküll, Kurt Goldstein, Friedrich Rothschild, and others. The first graduate courses on neurosemiotics were taught in some American and Canadian universities since 1970s. The term 'neurosemiotics' is also not much older. Neurosemiotics demonstrates which are the necessary conditions and processes responsible for semiosis in the neural tissue. It also describes the differences in the complexity of meaning making in animals of different complexity of the nervous system and the brain. == See also == Semiotics Zoosemiotics.
Cyclosiloxanes	Cyclosiloxanes are a class of silicone material. They are volatile and often used as a solvent. The three main commercial varies are octamethylcyclotetrasiloxane (D4), decamethylcyclopentasiloxane (D5) and dodecamethylcyclohexasiloxane (D6). They evaporate and degrade in air under sunlight. == Octamethylcyclotetrasiloxane (D4) == The octamethylcyclotetrasiloxane silicone liquid has no odor and consists of four repeating units of silicon (Si) and oxygen (O) atoms in a closed loop giving it a circular structure. Each silicon atom has two methyl groups attached (CH3). == Decamethylcyclopentasiloxane (D5) == Decamethylcyclopentasiloxane silicone liquid has no odor and consists of five repeating units of silicon (Si) and oxygen (O) atoms in a closed loop giving it a circular structure. Each silicon atom has two methyl groups attached (CH3). Typically it is used as an ingredient in antiperspirant, skin cream, sun protection lotion and make-up. With a low surface tension of 18 mN/m this material has good spreading properties.
Cancer exodus hypothesis	The cancer exodus hypothesis establishes that circulating tumor cell clusters (CTC clusters) maintain their multicellular structure throughout the metastatic process. It was previously thought that these clusters must dissociate into single cells during metastasis. According to the hypothesis, CTC clusters intravasate (enter the bloodstream), travel through circulation as a cohesive unit, and extravasate (exit the bloodstream) at distant sites without disaggregating, significantly enhancing their metastatic potential. This concept is considered a key advancement in understanding of cancer biology and CTCs role in cancer metastasis. == Mechanism == Traditionally, it was believed that CTC clusters needed to dissociate into individual cells during their journey through the bloodstream to seed secondary tumors. However, recent studies show that CTC clusters can travel through the bloodstream intact, enabling them to perform every step of metastasis while maintaining their group/cluster structure.

Input	Output
Generative pre-trained transformer	Generative Pre-trained Transformer 1 (GPT-1) was the first of OpenAI's large language models following Google's invention of the transformer architecture in 2017. In June 2018, OpenAI released a paper entitled "Improving Language Understanding by Generative Pre-Training", in which they introduced that initial model along with the general concept of a generative pre-trained transformer. Up to that point, the best-performing neural NLP models primarily employed supervised learning from large amounts of manually labeled data. This reliance on supervised learning limited their use of datasets that were not well-annotated, in addition to making it prohibitively expensive and time-consuming to train extremely large models; many languages (such as Swahili or Haitian Creole) are difficult to translate and interpret using such models due to a lack of available text for corpus-building. In contrast, a GPT's "semi-supervised" approach involved two stages: an unsupervised generative "pre-training" stage in which a language modeling objective was used to set initial parameters, and a supervised discriminative "fine-tuning" stage in which these parameters were adapted to a target task.
GPTZero	GPTZero is an artificial intelligence detection software developed to identify artificially generated text, such as those produced by large language models. While GPTZero was praised for its efforts to prevent academic dishonesty, many news outlets criticized the tool's false positive rate, which can be especially harmful in academic settings. == History == GPTZero was developed by Edward Tian, a Princeton University undergraduate student, and launched online in January 2023 in response to concerns about AI-generated usage in academic plagiarism. GPTZero said in May 2023 it raised over 3.5 million dollars in seed funding. In the first week of its release, the GPTZero experienced 30,000 uses, which led to a crash. It was supported by the web app company Streamlit, who allocated more server resources in response. In July 2024, it had 4 million users, compared to 1 million one year earlier. In summer 2024, GPTZero raised $10 million in Series A round funding. In September 2024, GPTZero announced an authorship tracking software that enables "to compile and share data about their writing process such as their copy/paste history, the number of editors they had, and how long editing took", in an effort "to move away from an all-or-nothing paradigm around AI writing towards a more nuanced one."
GPT 4.5	GPT-4.5 (codenamed "Orion") is a large language model developed by OpenAI as part of the GPT series. Officially released on February 27, 2025, GPT-4.5 is available to users subscribed to the ChatGPT Plus and Pro plans across web, mobile, and desktop platforms. Access is also provided through the OpenAI API and the OpenAI Developer Playground. == Overview == It was primarily trained using unsupervised learning, which improves its ability to recognize patterns, draw connections, and generate creative insights without reasoning. This method was combined with supervised fine-tuning and reinforcement learning from human feedback. The computational resources needed for training were provided by Microsoft Azure. Sam Altman described GPT-4.5 as a "giant, expensive model".

Input	Output
Claude	Claude is a family of large language models developed by Anthropic. The first model was released in March 2023. The Claude 3 family, released in March 2024, consists of three models: Haiku, optimized for speed; Sonnet, which balances capability and performance; and Opus, designed for complex reasoning tasks. These models can process both text and images, with Claude 3 Opus demonstrating enhanced capabilities in areas like mathematics, programming, and logical reasoning compared to previous versions. Claude 4, which includes Opus and Sonnet, was released in May 2025. == Training == Claude models are generative pre-trained transformers. They have been pre-trained to predict the next word in large amounts of text. Then, they have been fine-tuned, notably using constitutional AI and reinforcement learning from human feedback (RLHF).

C.4 List of topics in the datasets

C.4.1 Wikipedia AI

124 articles: DALL-E; OpenAI; Midjourney; Imagen (text-to-image model); Text-to-image model; Recraft; DeepSeek; DeepSeek (chatbot); Liang Wenfeng; High-Flyer; 2025 in artificial intelligence; DeepSeek (disambiguation); Six Little Dragons; R1; Ideogram (text-to-image model); Stable Diffusion; Automatic1111; ComfyUI; Stability AI; Emad Mostaque; Artificial intelligence and copyright; Fooocus; LAION; Sai; BLOOM (language model); Gemini; Gemini (chatbot); Gemini (language model); Gemini Robotics; Gemini Home Entertainment; Jet Force Gemini; Pixel 9; NotebookLM; Large language model; AlphaEvolve; Anthropic; Google Lens; Google AI Studio; Android XR; Chris Welty; Large language models in government; ChatGPT; Generative pre-trained transformer; GPT-4; GPT; GPT-4o; GPT-3; GPT-2; GPT-4.1; GPT-4.5; GPT-1; AutoGPT; Microsoft Copilot; GPTs; GPT-J; GPT Store; OpenAI o1; GPT4-Chan; GPTZero; Sora (text-to-video model); EleutherAI; YandexGPT; Writesonic; ChatGPT in education; Pause Giant AI Experiments: An Open Letter; Deep Learning (South Park); PauseAI; Chinchilla (language model); Artificial intelligence content detection; General-purpose technology; The Last Screenwriter; Wu Dao; Microsoft Recall; Alice and Sparkle; Amazon Q; Connor Leahy; Multimodal learning; OpenAI o4-mini; 2022 in artificial intelligence; Death of an Author (novella); GigaChat; P(doom); XLNet; Boyfriend Maker; 2023 in artificial intelligence; LLMs in higher education; Perceiver; NovelAI; Supremacy (book); Rabbit r1; Preamble (company); BookCorpus; Omneky; Machine unlearning; Artificial empathy; Llama (language model); Llama.cpp; DBRX; Llama (disambiguation); Alpaca (disambiguation); Qwen; Brave Leo; B65; Mistral; Mistral AI; Arthur Mensch; General Catalyst; Cédric O; Le Chat (disambiguation); PaLM; List of large language models; Prompt engineering; Foundation model; BERT (language model); LaMDA; T5 (language model); Alibaba Group; Claude (language model); Grok (chatbot); XAI (company); Colossus (supercomputer); X Corp.; Explainable artificial intelligence; Google DeepMind

C.4.2 Wikipedia Science

231 articles: Eurotrac; Scienticide; 505(b)(2) regulatory pathway; Anti-asthmatic agent; Breastmilk medicine; Cancer exodus hypothesis; Confocal endoscopy; Diabetes self-management; Dorsal pancreatic agenesis; Drone-Enhanced Emergency Medical Services; Electronic health record (Germany); Follicular drug delivery; LAMA2 related congenital muscular dystrophy; Most Favored Nation Drug Pricing; Musicians' Medicine; Poison exon; RNU2-2 syndrome; RNU4-2 syndrome; Synthetic Cannabinoid Use Disorder; Urinary antiinfective agent; Vestibular paroxysmia; Antarlide; Bioliteracy; Cancer exodus hypothesis; Dermestarium; Functional information; Interdigitation; Plasmagene; Poison exon; Polylecty; Spatial biology; Edge states; Electrostatic solitary wave; Frenesy (physics); History of the LED; HUN-REN Wigner Research Centre for Physics; Joaquim da Costa Ribeiro; Missile lofting; Nottingham effect; Physics of Life; Quasi-isodynamic stellarator; Riccardo D'Auria (theoretical physicist); Shockwave cosmology; Synchronous lateral excitation; Toroidal solenoid; Wohlfarth Lectureship; Compliance constants; Cononsolvency; Corrosion inhibitors for the petroleum industry; Cyclosiloxane; Dark oxygen; Direct reduction; Energy-rich species; Grupo Fertiberia; Intrinsic DNA fluorescence; Krupp–Renn process; Mental gland; Probico; School of Molecular Sciences; Shape of the atomic nucleus; Stable phosphorus radicals; Superelectrophilic anion; TOP Assay; Mathematical oncology; Mathethon; The Math(s) Fix; Conductivity cell; Generalized renewal process; Glossary of engineering: M–Z; Marine construction; Museum of Engines and Mechanisms; Northern Technical College; Safer end of engineering life; Synchronous lateral excitation; The Clark Collection of Mechanical Movements; Third medium contact method; UNESCO World Engineering Day for Sustainable Development; Positive health; Bell's mania; Chialvo map; Dysfunctome; Femoral nerve dysfunction; Fiber photometry; Fork cell; High Price (book); Hyper-empathy; Large dense core vesicles; Lateral olfactory tract usher substance; Malaria therapy; Max Planck Institute for Biological Intelligence; Nerve glide; Neural synchrony; Neurosemiotics; Neurotrophin mimetics; Optogenetic methods to record cellular activity; Personality neuroscience; Representational drift; Single-particle trajectory; Smell training; Spongy degeneration of the central nervous system; Walk Again Project; Amoeboflagellate; Borg (microbiology); Chrompodellid; Dissimilatory iron reducing bacteria; Garrod Lecture and Medal; Hydrocarbon-

oclastic bacteria; Laboratory-acquired infection; Matground; Microbial pathogenesis; Milnesium alpigenum; Mitochondrion-related organelle; Phageome; Phytoplankton microbiome; Virivore; Virome analysis; Zodletone Mountain; Glossary of cellular and molecular biology (M–Z); Agricultural weed syndrome; Cell autonomous sex identity; Codon reassignment; De novo domestication; Endemixit; Genetic map function; Hovlinc; Integrative and conjugative element; Jena Declaration; Macrosatellite; Museomics; Poison exon; Polydactyly-myopia syndrome; Red cell genotyping; Right To Know; Selection limits; Shadow effect; Transcriptome-wide association study; Tumor mutational burden; Allogeneic processed thymus tissue; Cellular anastasis; COVID-19 passports in the United Kingdom; History of phagocytosis; Immunocapitalism; Macrophage-activating lipopeptide 2; Metal allergy; Milk immunity; Myocarditis-myositis-myasthenia gravis overlap syndrome; Oligoclonal antibody; P-i mechanism; Pathogen avoidance; Peripheral ulcerative keratitis; Post-acute infection syndrome; RVT-802; T memory stem cell; Thymic mimetic cells; TMEM61; Type 2 inflammation; Vaccine passports during the COVID-19 pandemic; Vaccine resistance; Zigakibart; 2022–2023 pediatric care crisis; Acoustic epidemiology; Causal pie model; Connecting Organizations for Regional Disease Surveillance; Elimination of tuberculosis; Epidemics Act; Epidemiology in Relation to Air Travel; Epidemiology of gonorrhoea; European Society of Health and Medical Sociology; Harvard Six Cities study; Hyperendemic; Loneliness epidemic; Microbial pathogenesis; Origin tracing; Pathogenic microorganisms in frozen environments; SARS-CoV-2 in white-tailed deer; Source attribution; Sporadic disease; Outline of public health; Alcohol tax; Autobesity; Biomedical Research Center; CalOptima; Care Group approach; Christian Health Association of Malawi; Commercial determinants of health; Connecting Organizations for Regional Disease Surveillance; COVID-19 lockdowns by country; Epidemics Act; History of public health in Australia; History of public health in Canada; History of public health in Chicago; History of public health in New York City; History of public health in the United Kingdom; History of public health in the United States; Intermittent water supply; International Association for Cannabinoid Medicines; Langya virus; LGBT life expectancy; User:Lguzmang06/sandbox; Loneliness epidemic; Malawi Network of AIDS Services; Mass. and Cass; Medical officer of environmental health; Motonormativity; National Association for People living with HIV/AIDS in Malawi; North Karelia Project; Nuisance ordinance; Origin tracing; Preventive and social medicine; Responsibility Deal; SaTScan; Sleeping Sickness Commission; Slug gate; Social determinants of mental health; Special Programme of Research, Development and Research Training in Human Reproduction- HRP; Telemedicine in Nepal; Vaccine equity; Vaccine line jumping; Vaccine storage; WHO Hub for Pandemic and Epidemic Intelligence; WHO public health prizes and awards; Additive effect; Antica Farmacia Sant'Anna; FK962; Institute for Safe Medication Practices; Model-Informed Precision Dosing; P-i mechanism; Penetration enhancer; Pharmacological cardiotoxicity; Pullulan bioconjugate; Reversible Hill equation

C.5 Training Loss Curves for Wikipedia Fine-tuning

We report the training loss curves for both the non-DP and DP models trained on the Wikipedia datasets in Figure 8 and Figure 9. We include results from two fine-tuning settings: (1) the setup where the unseen Wikipedia articles are interspersed with pre-training Wikipedia data, and privacy is specified through a target privacy budget ε; and (2) fine-tuning only over the unseen Wikipedia dataset, where we directly set the noise multiplier due to the instability in computations of noise multipliers for large ε.

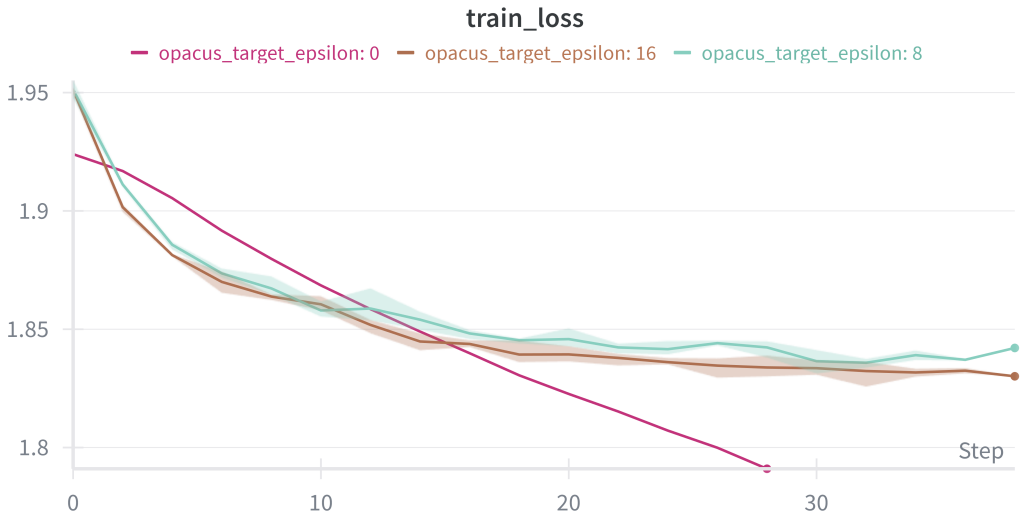


Figure 8: Training loss curve for the models fine-tuned on the large Wikipedia dataset,

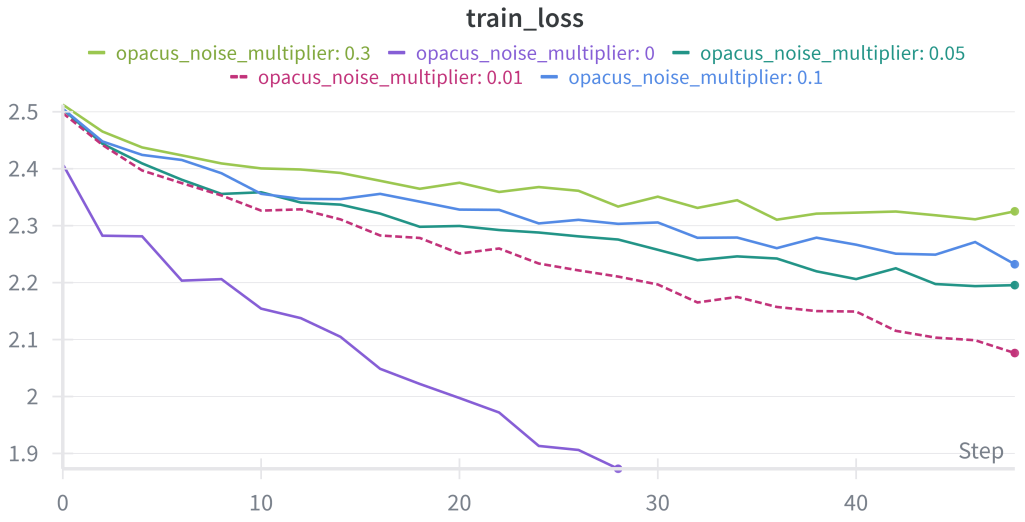


Figure 9: Training loss curve for the models fine-tuned only on the private Wikipedia articles.

D Detailed Experimental Results

D.1 Analysis of Recurring Hallucinations

The metrics in §2.3 can capture if models hallucinate incorrect facts, but they do not distinguish between models that output a range of incorrect information (suggesting general noisiness) or if they repeatedly output the same incorrect information across generations (suggesting encoding of inaccurate facts). To identify recurring factual claims across model generations, we propose a multi-stage clustering algorithm. We give a high-level summary below, with precise pseudo-code in Algorithm 2.

Consider the generated synthetic documents grouped by the the topic $t \in \mathcal{T}$ used to prompt their generation. Each document is decomposed into atomic claims and these claims are aggregated into a claim set $C_t = c_i$ for each topic t.

We first index all extracted claims to their source (synthetic) documents, then cluster them using sentence-embedding–based agglomerative clustering to group semantically similar claims. To refine boundaries, we apply DBSCAN with Jaccard similarity, ensuring clusters are both semantically coherent and lexically consistent. This reduces cases where semantically related but factually distinct claims are grouped together.

Finally, we retain only recurring claim clusters, defined as clusters containing claims from at least two distinct documents for the same topic. This step isolates claims that recur across different generations, highlighting factual patterns the model consistently produces rather than one-off statements. We then analyze these clusters of claims to identify potential recurrent hallucinations.

D.1.1 Recurring Hallucination Analysis

Table 7 reports results from the claim clustering analysis, as described in §D.1. In both datasets, the DP model with $\epsilon = 8$ outputs fewer recurring supported claims (i.e. factually correct statements) than other models. DP models also output more recurring unsupported claims than non-DP models, with $\epsilon = 8$ highest for Wikipedia Science and $\epsilon = 16$ highest for Wikipedia AI. For both datasets, the ratio of supported recurring claims to unsupported recurring claims is consistently lower with stricter privacy budgets. This indicates that the increased hallucinations under DP are not random. Given that the same unsupported claims recur across generations, it suggests that DP systematically shifts the model's output distribution toward factually incorrect content.

Table 8 shows an example of a cluster of recurring hallucinations for each model, where the prompt was "AlphaEvolve." The non-DP model correctly outputs that AlphaEvolve is a model, but incorrectly describes model use. In contrast, the DP models both hallucinate that AlphaEvolve is a video game, with repeated fabricated information about the development and game play.

Algorithm 2 Recurring Claim Cluster Algorithm

Require: Set of topics $\mathcal{T}$, where each $t \in \mathcal{T}$ has $\mathcal{S}$ corresponding generated documents about the topic. $C = \{C_t\}$: Claims per topic $t \in \mathcal{T}$, where each $C_t = \{c_i\}$ with atomic facts (supported or unsupported).

$\mathcal{I} \leftarrow$ IndexClaimsByText$(C, \mathcal{I})$ ▷ Index Claims to their Source Document
$\mathcal{K} \leftarrow$ ClusterAssignmentOfClaims$(\mathcal{I})$ ▷ Sentence Embedding-based Agglomerative Clustering
$\mathcal{K} \leftarrow$ DBSCANJaccardClustering$(\mathcal{K})$ ▷ DBSCAN Clustering over clusters to ensure their Jaccard Distance is low
for $t \in \mathcal{T}$ **do**
    $\mathcal{K}'[t] \leftarrow \{\}$ ▷ Initialize $\mathcal{K}'$ to contain clusters of recurring claims
end for
for topic $t \in \mathcal{T}$ **do**
    for Cluster $k \in \mathcal{K}(t)$ **do**
        if COUNT(S) for any $c_i \in k \geq 2$ **then**
            Append k to $\mathcal{K}'[t]$ ▷ Append a cluster of claims if the claims contain at least two supporting documents
        end if
    end for
end for

	DP Setting	*Supported*			*Unsupported*			
		Recur Count	**Total Count**	**% Avg Recur**	**Recur Count**	**Total Count**	**% Avg Recur**	**Supported: Unsupported**
Wikipedia AI	$\epsilon = \infty$	242	1198	20.20	494	1992	24.80	0.490
	$\epsilon = 16$	200	1103	18.13	**523**	2002	**26.12**	0.382
	$\epsilon = 8$	174	1019	**17.08**	509	2011	25.31	0.342
	ANADP$_{\epsilon=16}$	207	1025	20.20	452	1915	23.60	0.458
	ANADP$_{\epsilon=8}$	**161**	932	17.27	499	1922	25.96	**0.323**
	Task-tuned	**161**	920	17.50	460	1945	23.65	0.350
Wikipedia Science	$\epsilon = \infty$	705	2896	24.34	560	2417	23.17	1.259
	$\epsilon = 16$	582	2467	23.59	556	2128	26.13	1.047
	$\epsilon = 8$	556	2378	23.38	**633**	2222	**28.49**	**0.878**
	ANADP$_{\epsilon=16}$	567	2372	23.90	536	2065	25.96	1.058
	ANADP$_{\epsilon=8}$	544	2287	23.79	509	2128	23.92	1.069
	Task-tuned	**474**	2258	**20.99**	535	2002	26.72	0.886

Table 7: Analysis of recurrent claims and hallucinations (§D.1) for temperature $\tau = 0.3$. "Total Count" reports the total number of generated clusters, "Recur Count" reports the number of those clusters with ≥ 2 supporting documents, and %Avg Recur is "Recur Count"/"Total Count". The far right column reports "Recur Count" of supported claims / "Recur Count" of unsupported claims. DP models output a lower ratio of repeated supported claims to unsupported claims, suggesting increased repeated hallucinations. Bolding indicates most hallucinations (i.e. least supported or most unsupported). Task-tuned indicates the base model trained on pretraining data for 1 epoch.

$\epsilon = \infty$	'AlphaEvolve is used for generating the molecular structures of organic molecules.', 'AlphaEvolve is used for generating molecular structures.'
$\epsilon = 16$	'There are two types of enemies in the game.', 'The game features two types of enemies.'
$\epsilon = 8$	'Black Hole Interactive is a game development company.', 'Black Hole Interactive is a video game development company.'

Table 8: Example unsupported claim clusters (hallucinations) for each model. We provide additional examples in subsection E.4.

D.2 Statistically Significant Differences in FactScore

We report the statistical significance of the FactScore differences in this section. We compare responses only for topics that have a usable generation with a corresponding FactScore. A topic is only taken into consideration if the configurations being compared cover it. For the DP fine-tuning comparisons, we take one intersection across $\varepsilon \in \{8, 16, \infty\}$, which results in $221/222$ Science, $118/124$ AI topics for the topic-level differences. For the pre-trained models, we take the pairwise intersection of topics (between VaultGemma and the pre-trained model being compared against) for which there are valid generations, which results in 198–213 (Science) and 103–118 (AI) topics. The response-level tests use all responses from the model configuration, irrespective of topics.

D.2.1 Pre-trained models

	Wiki-Science			Wiki-AI		
Model	Δ	p_W	p_t	Δ	p_W	p_t
Gemma-2-2B	$+.111$	$< .001^{***}$	$< .001^{***}$	$+.055$	$.035^{*}$	$.025^{*}$
Gemma-2B	$+.138$	$< .001^{***}$	$< .001^{***}$	$+.084$	$< .001^{***}$	$< .001^{***}$
Gemma-3	$+.143$	$< .001^{***}$	$< .001^{***}$	$+.100$	$< .001^{***}$	$< .001^{***}$
GPT2-XL	$+.038$	$.028^{*}$	$.033^{*}$	$-.140$	$< .001^{***}$	$< .001^{***}$

Table 9: Topic-level paired tests against VaultGemma, on the worst (minimum) response per topic. Δ: mean paired FActScore difference (model $-$ VaultGemma); positive favors the model. p_W: Wilcoxon signed-rank; p_t: paired t-test. Topics are excluded where either model produced no scorable generation (empty or degenerate output yielding no atomic claims), with varying n of the range 198–201 topics (Science), 103–116 (AI). $^{*}p < .05$, $^{***}p < .001$.

	Wiki-Science		Wiki-AI	
Model	Δ	p_t	Δ	p_t
Gemma-2-2B	$+.108$	$3.5\text{e}{-14}^{***}$	$+.062$	$6.0\text{e}{-4}^{***}$
Gemma-2B	$+.119$	$5.2\text{e}{-17}^{***}$	$+.064$	$3.1\text{e}{-4}^{***}$
Gemma-3	$+.109$	$3.1\text{e}{-12}^{***}$	$+.087$	$1.4\text{e}{-5}^{***}$
GPT2-XL	$-.005$	$.752$	$-.154$	$8.0\text{e}{-17}^{***}$

Table 10: Welch two-sample t-tests against VaultGemma for the pretrained baselines, over every generated response. Δ: mean response-level FActScore difference (model $-$ VaultGemma); positive favors the model.

D.2.2 Fine-tuned models

	Wiki-Science			Wiki-AI		
ε	Δ	p_W	p_t	Δ	p_W	p_t
8	$-.032$	$.026^{*}$	$.030^{*}$	$-.048$	$.009^{**}$	$.007^{**}$
16	$-.035$	$.005^{**}$	$.018^{*}$	$-.033$	$.032^{*}$	$.079$

Table 11: Topic-level paired tests against the non-private baseline ($\varepsilon = \infty$), on the worst (minimum) response per topic. Δ: mean paired FActScore difference (DP model $-$ non-private); positive favors the model. p_W: Wilcoxon signed-rank; p_t: paired t-test. Topics are excluded where either model produced no scorable generation (empty or degenerate output yielding no atomic claims), with $n = 221$ (Science), $n = 118$ (AI). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

	Wiki-Science		Wiki-AI	
ε	Δ	p_t	Δ	p_t
8	$-.026$	$.024^{*}$	$-.061$	$2.4\text{e}{-5}^{***}$
16	$-.025$	$.033^{*}$	$-.038$	$.009^{**}$

Table 12: Response-level tests against $\varepsilon = \infty$, treating each generation as an observation. Δ: mean FActScore difference (model $-$ non-private); negative favors non-private. p_t: Welch two-sample t-test. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

D.3 Generation Length Statistics

Category	ε	Facts /Resp.	Length (words) Mean	Std.
Wiki. AI	8	10.24	80.41	21.16
	16	10.60	82.00	21.42
	∞	11.12	86.64	17.47
Wiki. Science	8	8.78	79.14	19.56
	16	8.66	79.88	18.91
	∞	9.75	88.75	14.79

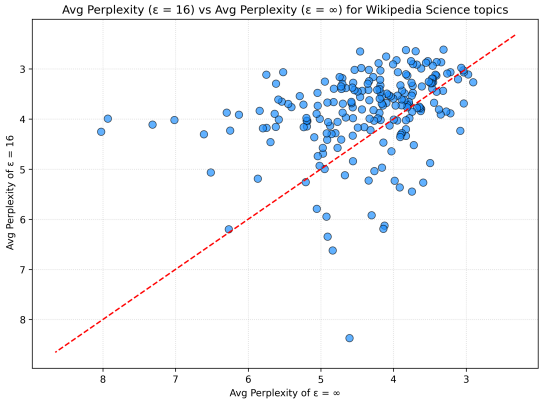
Table 13: Average number of facts and response length by category and privacy budget ε.

We report the generation-length statistics to rule out any potential length-related artifacts in our FactScore evaluations in Table 13. While DP-generated text is slightly shorter on average than the text from the non-private models, consistent with prior work (Çano and Habernal, 2025), this difference is not large enough to warrant a significant shift in the FactScore (which already accounts for length by normalizing by the number of claims). Likewise, we show that the density of claims is stable across privacy budgets, so our reported dif-

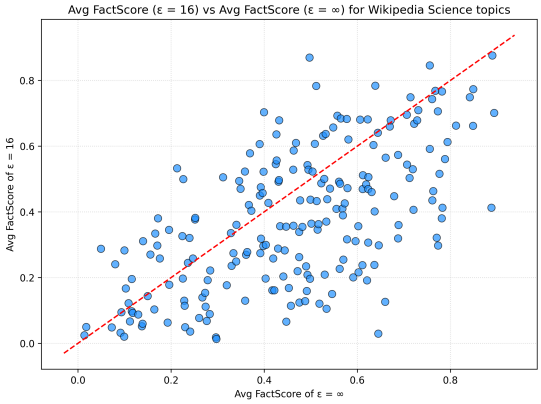
ference in FactScores are not attributable to the length of the generation.

D.4 Lower Perplexity Does Not Correlate With Factual Reliability

Our analysis of hallucinations from language model outputs is a construct that is distinct from and cannot be conflated with generation notions of utility, as depicted in Figure 11 and Figure 10, which demonstrates the lack of correlation between perplexity and factual correctness. Several generations with low perplexity receive low FactScores, and vice versa. The DP fine-tuned models often produce text with lower perplexity on average, further underscoring that other utility metrics do not necessarily capture hallucinations, that is to say, a model can generate highly fluent text (low perplexity) and also hallucinate. We have also included training loss curves in Appendix C.5, which demonstrates that the models we evaluate do learn to fit the training data.

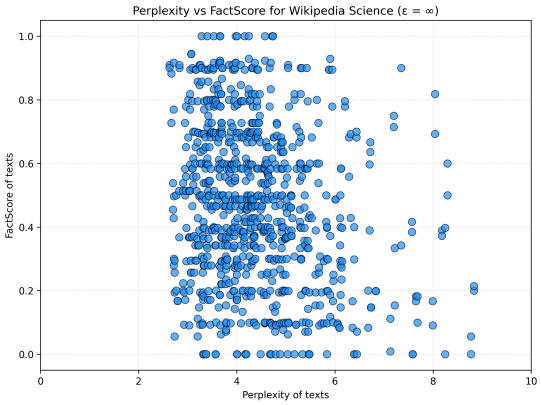


(a) Avg perplexity of each topic for $\varepsilon = 16$ vs $\varepsilon = \infty$.

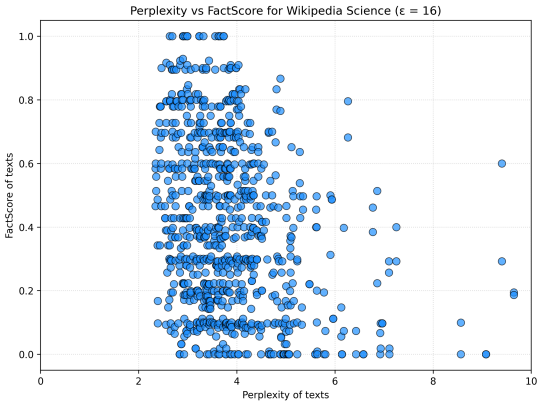


(b) Avg FactScore of each topic for $\varepsilon = 16$ vs $\varepsilon = \infty$

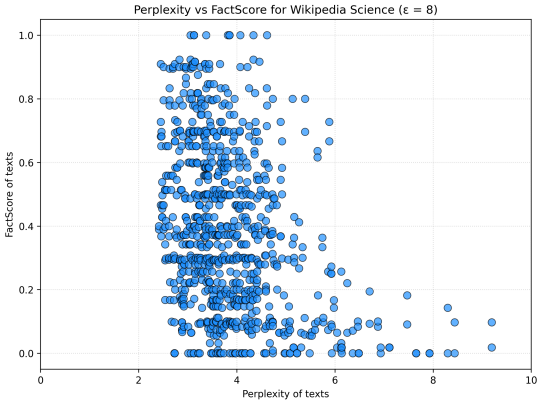
Figure 10: Average perplexity and average FactScore per topic across models. DP-finetuned models often achieve lower average perplexity (points above the y = x line), yet the non-DP model attains higher average FactScores across topics—showing that lower perplexity does not imply better factuality.



(a) $\varepsilon = \infty$



(b) $\varepsilon = 16$



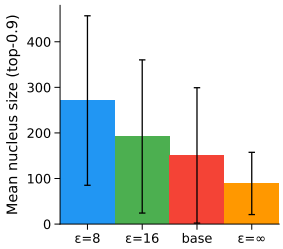
(c) $\varepsilon = 8$

Figure 11: Relationship between model perplexity and FactScore for all texts generated for topics from the Wikipedia Science. Lower perplexity is not predictive of higher factual accuracy.

D.5 Hallucination of pre-training facts in DP fine-tuned models

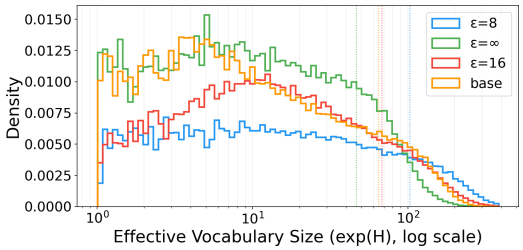
Our experiments also look to address whether DP fine-tuning degrades the knowledge already encoded during standard pre-training. Table 14 reports FactScores over Wikipedia pre-training. The differences between DP and non-DP models are marginal, suggesting DP finetuning does not disrupt factual knowledge acquired from pre-training data. This trend is consistent across temperatures (Figure 19) and stands in stark contrast to the evaluations on previously unseen data, where stronger privacy constraints correlate with lower factual accuracy.

D.6 Distributional flattening for pre-training data



(a) Wikipedia Pretraining

Figure 12: Mean nucleus size (top-$p = 0.9$) for models fine-tuned under different DP budgets. DP models exhibit larger nucleus sizes, indicating greater dispersion of probability mass across tokens and increased next-token uncertainty.



(a) Wikipedia Pretraining

Figure 13: Distribution of effective vocabulary size ($\exp(H)$), where higher values indicate probability mass spread across more tokens. DP models show higher effective vocabulary sizes, reflecting increased distributional uncertainty.

Although DP models exhibit higher entropy for both the unseen fine-tuning and the Wikipedia pre-training data (Figure 12 and Figure 13), the distributional flattening does not affect hallucinations in the same way across settings. An analysis of the nucleus threshold overlap, where we compare the top-p nucleus sets of the DP and non-DP models, indicates that the ranking of tokens is preserved for the seen data despite the distribution flattening. On the nucleus-overlap metric, the $\varepsilon = 16$ model attains a mean Jaccard similarity of 0.707 ± 0.071 on seen data and 0.623 ± 0.051 on unseen data, while the $\varepsilon = 8$ model achieves 0.704 ± 0.069 (seen) and 0.619 ± 0.051 (unseen). The models share greater token overlap on seen pre-training data than on unseen data. These results suggest that on data seen during pre-training, the increased entropy reflects a more diffused but still structurally intact next-token probability distribution. On the other hand, for the unseen data, the redistributed probability mass falls on incorrect alternatives, leading to the observed increase in hallucinations.

D.7 Is the model reverting to its pre-training priors?

A natural concern is that the DP fine-tuned models revert to their pre-trained priors during generation, rather than acquiring knowledge from the unseen fine-tuning data. Our evidence points to the contrary. A model reverting to its priors would be expected to resemble the base model more closely on the unseen fine-tuning data. However, as shown in Figure 15 for both DP models, the KL divergence from the base model is higher on the unseen data (which the base model was never fine-tuned on) than on the seen data.

This is further corroborated by the 1D loss interpolation (Figure 14) in the model's subspace. Although the DP models remain relatively close to the base model in the parameter space compared to the non-private model, it sits in a region of the space associated with a lower loss. The non-private model, on the other hand, has shifted away significantly from the base model, which is reflective of a much more precise signal during fine-tuning. Additionally, as discussed in Section 3.1 the DP fine-tuned models hallucinate at a rate higher than the base model on unseen topics. If the model was relying on its priors, we would expect hallucination rates to be similar to that of the base model. Together, these results provide clear evidence that DP models learn from the fine-tuning data and diverge from their pre-trained priors across both the unseen and seen evaluation sets.

DP Budget	Average FS	Median FS	Q1 FS	Q3 FS	Avg Max FS / Topic	Avg Min FS / Topic	# of FS $\geq$0.5	# of FS $<$0.5
$\epsilon = \infty$	30.4	23.5	9.1	46.2	42.5	18.8	131	**841**
$\epsilon = 16$	**29.5**	**23.1**	9.1	**44.4**	41.8	**18.1**	**125**	829
$\epsilon = 8$	30.6	25.0	**8.3**	50.0	43.2	18.8	144	824
Base	29.7	25.0	9.1	**44.4**	**41.0**	19.6	139	837

Table 14: FactScores (FS; reported in %) for GPT-J evaluated with temperature $\tau = 0.3$ over Wikipedia pre-training, which contains articles likely to be in pre-training data, but not included in fine-tuning. DP models perform similarly as non-DP models, suggesting no disruption to facts learned in pre-training. Bolding indicates worse FactScores (i.e. higher hallucinations).

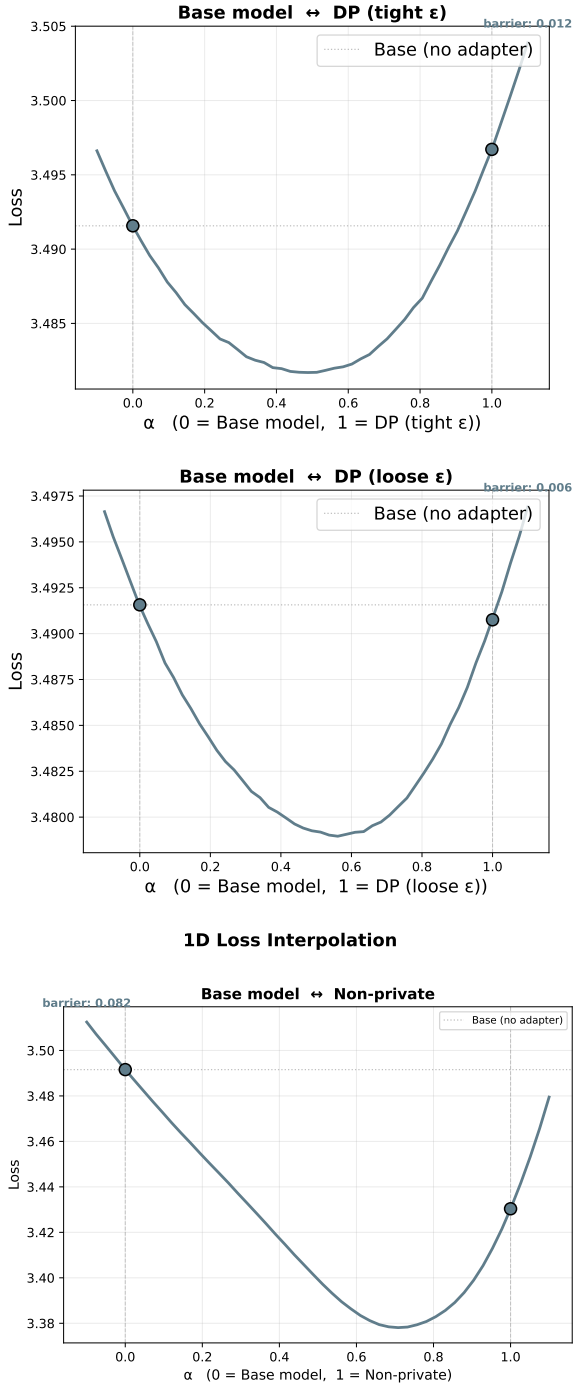
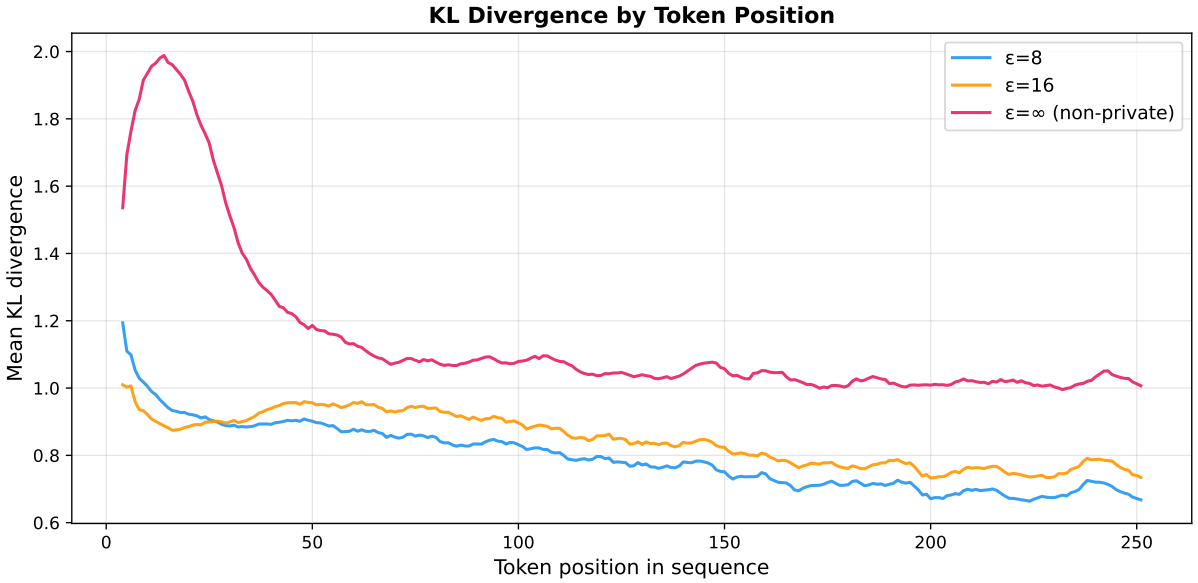
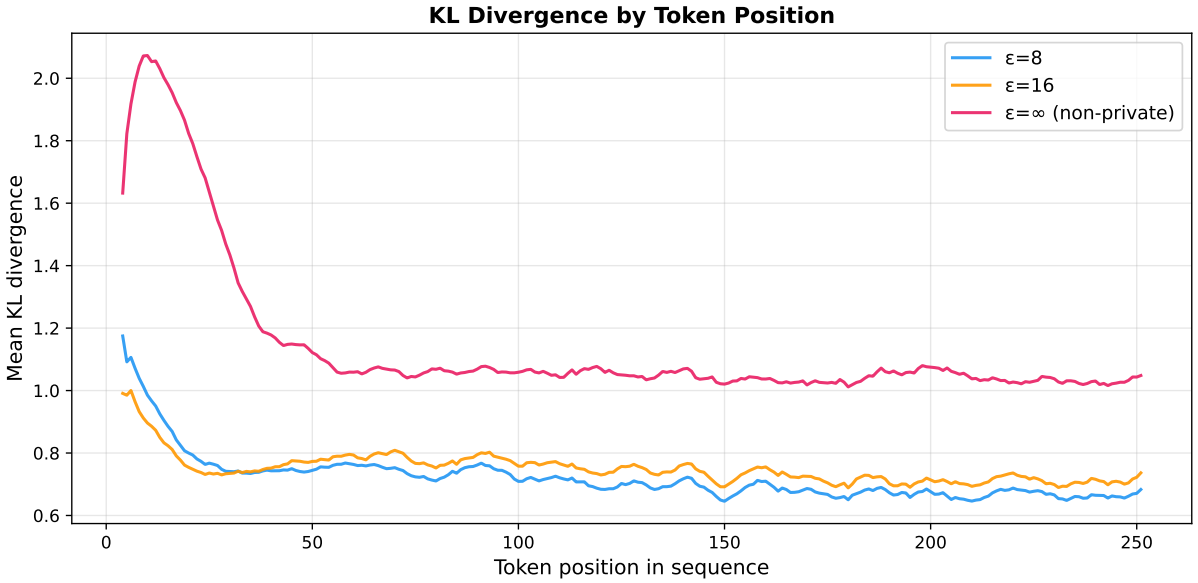


Figure 14: 1D loss interpolation between the base GPT-J 6B model ($\alpha = 0$) and fine-tuned models ($\alpha = 1$) under different privacy settings. The dashed line indicates the loss of the base model without an adapter. DP models ($\varepsilon = 8$ and $\varepsilon = 16$) remain relatively close to the base model in parameter space, with small loss barriers (0.012 and 0.006, respectively), while the non-private model shifts significantly further (barrier of 0.082), reflecting a stronger gradient signal during fine-tuning. Despite remaining closer to the base model, the DP models sit in regions of lower loss, indicating that they learn from the fine-tuning data rather than reverting to pre-training priors.



(a) Wikipedia science and AI



(b) Wikipedia pretraining

Figure 15: KL divergence between each fine-tuned model and the base GPT-J 6B model, measured at each token position in generated sequences. KL divergence quantifies how much the fine-tuned model's next-token predictions differ from the base model's: higher values indicate that fine-tuning has shifted the model's predictions further from the base model's behavior. (a) On unseen fine-tuning data (Wikipedia Science and AI), all fine-tuned models—including DP models—exhibit higher KL divergence from the base model than (b) on data seen during pre-training. If DP models were simply reverting to their pre-training priors, we would expect low KL divergence on the unseen data, as the models would behave similarly to the base model on topics it was never fine-tuned on. The higher divergence on unseen data indicates the opposite: DP models do acquire new information from fine-tuning and diverge from their pre-trained behavior, even under strict privacy budgets.

D.8 Manual Annotations: Claim Annotation Interface

D.8.1 Annotation Guidelines

In our study, annotators were tasked with evaluating the quality of factual claims from a model-generated text, with respect to a given source text. Each annotation instance consists of: (i) a source text corresponding to the original document for a given topic from Wikipedia; (ii) the text generated by the DP/non-DP fine-tuned models for the given topic; and (iii) a set of atomic claims extracted from the model-generated text. Annotators were required to assess each claim along multiple dimensions such as its factual accuracy, support from the source text, in addition to other quality flags.

D.8.2 Annotation Dimensions

1. Veracity Annotators evaluate the factual correctness of each claim based on real-world knowledge, independent of whether it appears in the source text.

- **Correct**: The claim is factually accurate and verifiable through reliable sources.
- **Incorrect**: The claim is factually inaccurate or false.
- **Unclear/Do Not Know**: The claim is too vague, ambiguous, or requires some specialized expertise to verify.

2. Source Support This dimension captures whether the claim is explicitly supported by the content in the source text.

- **Supported in Source**: The claim is directly mentioned or implied by the content in the source text.
- **Not Supported in Source**: The claim is either absent from the source or contradicted by the information in the source.

3. Quality Flags Annotators may also flag issues related to clarity or subjectivity.

- **Vague / Ambiguous**: The claim does not communicate a precise meaning or is difficult to interpret.
- **Subjective**: The claim expresses an opinion or value judgment, as opposed to an objective fact.

D.8.3 Annotation Procedure

For each claim, annotators perform the following steps:

1. They go through the source text and model-generated text to understand the context.
2. Label each claim along the three dimensions described above.
3. Optionally, provide free-text comments to clarify ambiguous cases or provide additional context. Additionally, they were permitted to make use of search engines or external judgements to annotate each of the claims.

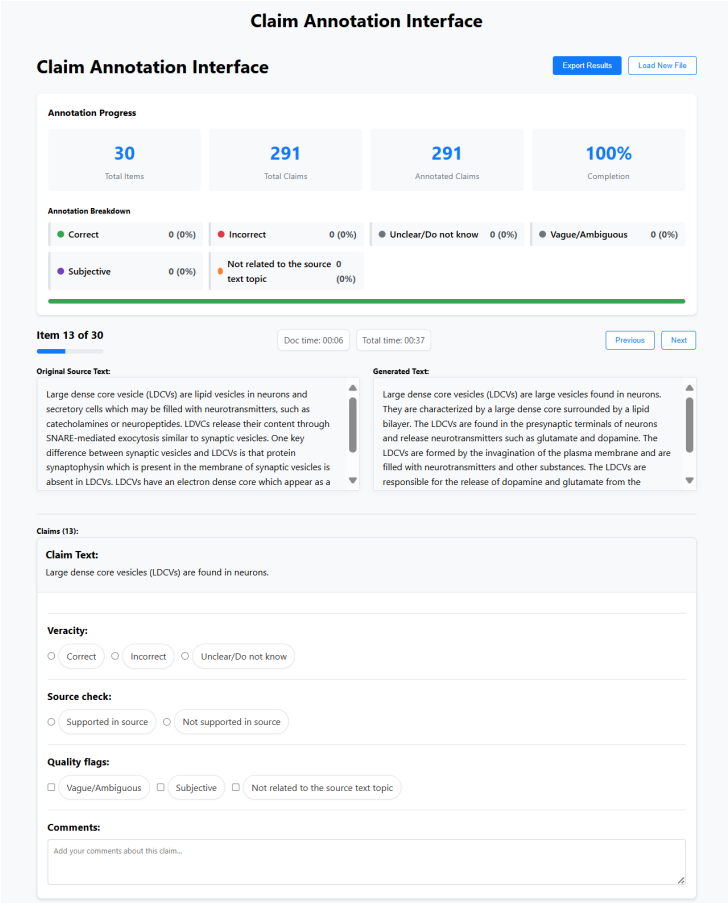


Figure 16: The interactive claim annotation tool used by participants to evaluate the factual correctness of claims from the model-generated text.

Comparison		DP-inf	DP-16
Inter-Annotator	Veracity	0.57	0.84
	Support	0.57	0.73
Model vs Annotator	Support (Annotator 1)	0.49	0.23
	Support (Annotator 2)	0.48	0.34

Table 15: Cohen's Kappa scores for DP-inf and DP-16 settings for human annotation of Wikipedia AI article claims.

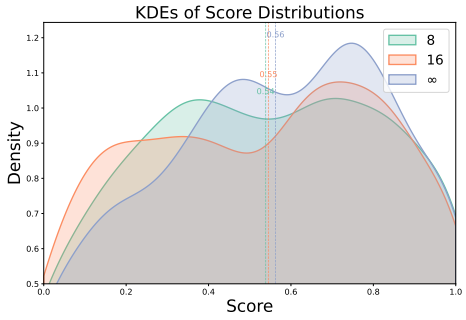
D.9 Expanded FactScore Evaluations

We report additional experiments in the following tables:

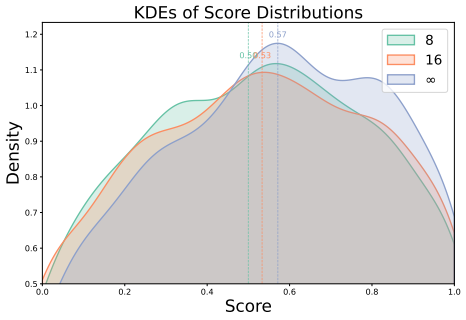
- FactScore evaluations with different claim decomposition and evaluation models in Table 16 and Table 17
- FactScore distributions per temperature setting under different models in Figure 18, Figure 19, Figure 17, Figure 21, Figure 22, Figure 20.
- FactScores on unseen Wikipedia articles after fine-tuning on only the Wikipedia training data in Table 19.

D.10 Variability in FactScore

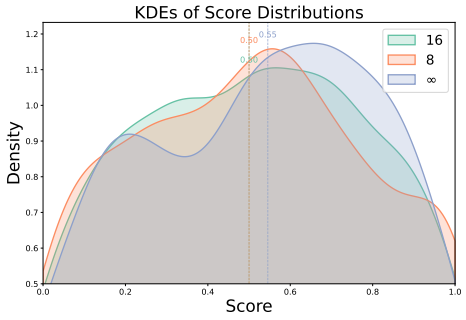
We report the per-topic standard deviation of FactScore for Wiki AI and Wiki Science to assess the stability of model factuality across prompt variations. Figures 23a and 23b show the distributions for models trained with $\varepsilon \in \{8, 16, \infty\}$.



(a) Temperature = 0.5

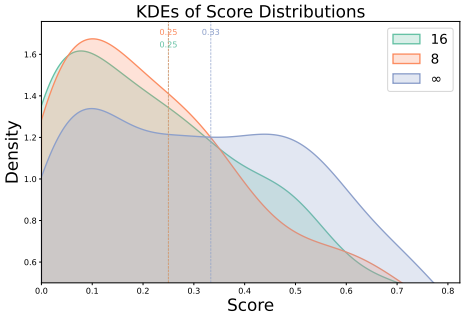


(b) Temperature = 0.7

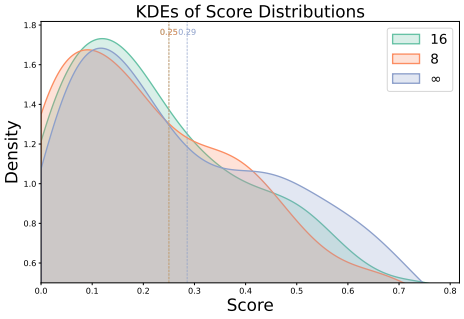


(c) Temperature = 1.0

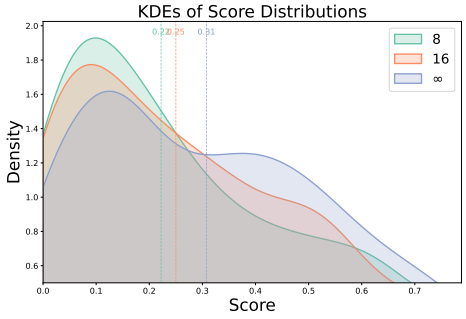
Figure 17: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Science data for under different temperature settings for GPT-J 6B.



(a) Temperature = 0.5



(b) Temperature = 0.7

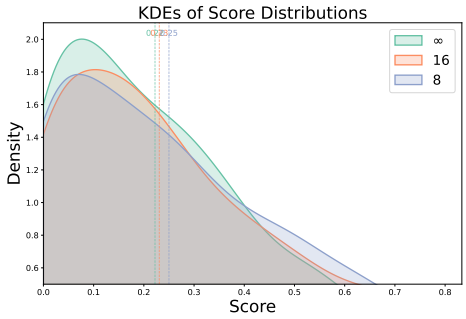


(c) Temperature = 1.0

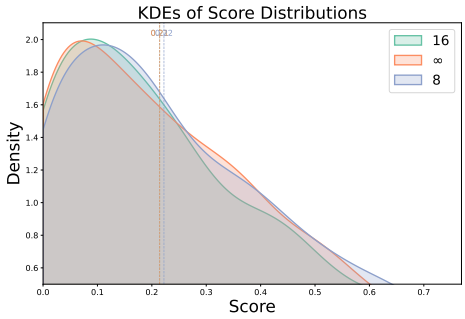
Figure 18: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia AI data for under different temperature settings for GPT-J 6B.

Temperature	Pretrained Model	Average FS	Median FS	Q1 FS	Q3 FS	Avg Max FS / Topic	Avg Min FS / Topic	# of FS >0.5	# of FS <=0.5
0.3	∞	56.3	57.1	33.3	80	71.2	41.2	397	491
	16	54.3	57.1	28.6	80	69.7	38.9	314	564
	8	53.2	54.5	28.6	76.9	68.3	37.9	294	588
0.5	∞	55.2	57.7	33.3	77.8	68.6	41	418	470
	16	51.7	50	25	77.8	67.6	35.5	335	547
	8	52.1	54.2	28.6	77.1	67.5	36.7	332	548
0.7	∞	55	57.1	30	80	69.8	40.6	436	452
	16	52.6	54.2	28.6	77.8	67.4	37.9	380	502
	8	51.5	50	28.6	75	67.2	35.8	366	522
1.0	∞	52.6	53.8	30	77.8	67.5	37.8	442	446
	16	51.2	50	27.3	75	66.1	36.2	420	464
	8	49.8	50	25	72.7	65.4	34.4	381	505

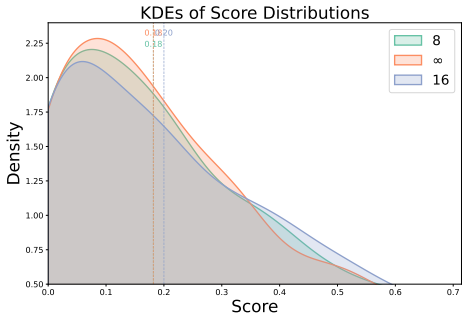
Table 16: Factuality evaluation scores for different temperature settings for the Wikipedia Science articles, using Meta Llama 3.2-3b-Instruct for claim evaluation and decomposition.



(a) Temperature = 0.5



(b) Temperature = 0.7

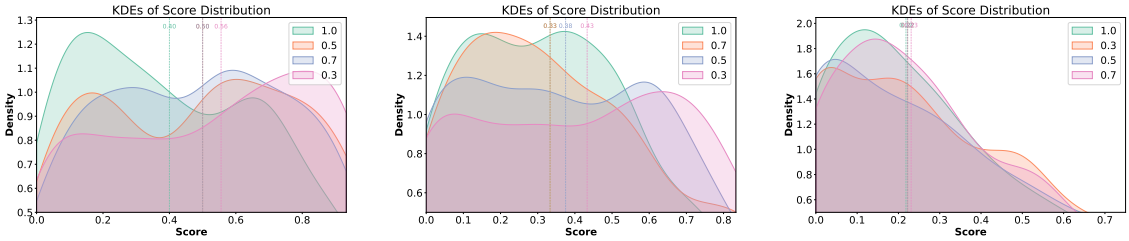


(c) Temperature = 1.0

Figure 19: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Pretraining data for under different temperature settings for GPT-J 6B.

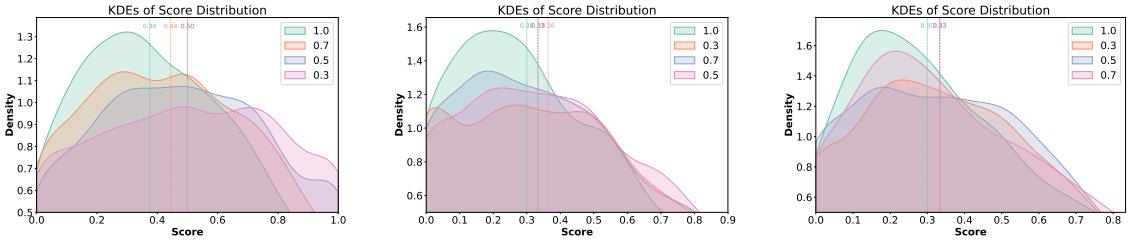
Temperature	Pretrained Model	Average FS	Median FS	Q1 FS	Q3 FS	Avg Max FS / Topic	Avg Min FS / Topic	# of FS >0.5	# of FS <0.5
0.3	∞	48.6	50.0	28.9	66.7	71.4	27.2	108	765
	16	47.5	45.5	25.0	66.7	70.9	24.2	79	788
	8	46.9	44.4	25.0	66.7	69.0	26.0	61	809
0.5	∞	48.9	50.0	30.0	66.7	68.7	29.1	112	759
	16	44.6	44.4	25.0	62.5	70.8	18.8	82	789
	8	46.0	50.0	25.0	66.7	65.0	27.3	81	784
0.7	∞	45.5	44.4	27.3	63.6	63.8	27.4	123	752
	16	45.5	44.4	27.3	62.5	68.0	23.1	99	766
	8	45.8	44.4	25.0	66.7	66.0	25.9	85	785
1.0	∞	44.4	42.9	25.0	62.5	69.2	22.6	128	738
	16	43.9	42.9	25.0	60.0	62.7	26.1	85	783
	8	44.5	42.9	27.3	62.5	63.0	25.5	105	772

Table 17: Factuality evaluation scores for different temperature settings for the Wikipedia Science articles, using DeepSeek-R1-Distill-Qwen-7B for claim evaluation and decomposition.



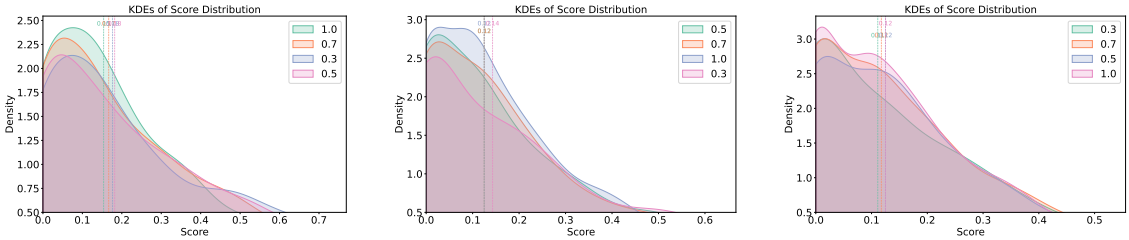
(a) Gemma3-1B-PT (b) VaultGemma-1B (c) GPT2-XL

Figure 20: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia AI data for the pre-trained models.



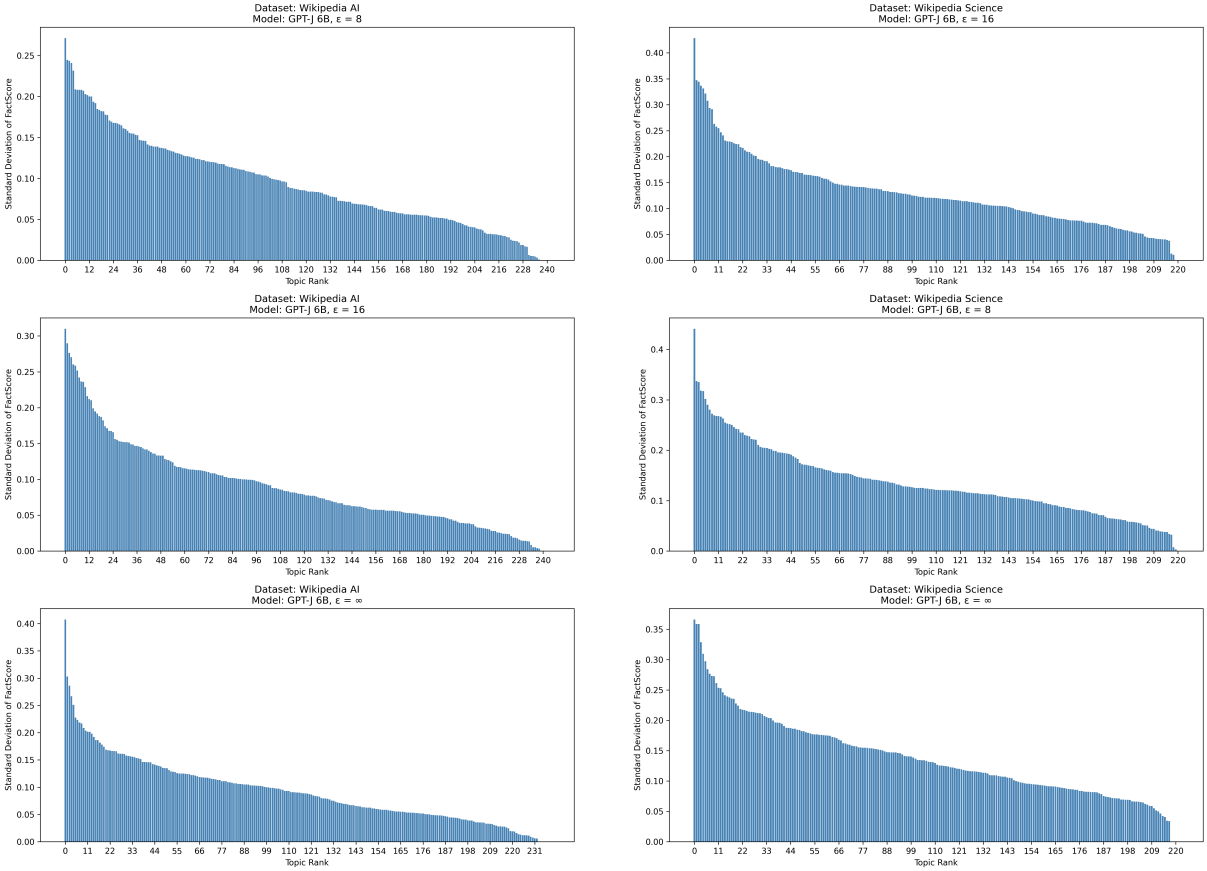
(a) Gemma3-1B-PT (b) VaultGemma-1B (c) GPT2-XL

Figure 21: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Science data for the pre-trained models.



(a) Gemma3-1B-PT (b) VaultGemma-1B (c) GPT2-XL

Figure 22: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Pretraining data for the pre-trained models.



(a) Wikipedia AI

(b) Wikipedia Science

Figure 23: Standard deviation of FactScore across topics for Wiki AI and Wiki Science.

Dataset	Target Epsilon	Average FS	Median FS	Q1 FS	Q3 FS	# of FS >0.5	# of FS <0.5
Wikipedia Science	∞	49.0	25.7	9.7	46.5	247	1072
	16	53.8	36.8	17.1	55.9	411	878
	8	52.6	37.2	18.5	58.4	427	860
Wikipedia AI	∞	32.7	18.8	8.6	34.2	60	646
	16	36.7	28.6	12.5	46.0	136	565
	8	37.4	29.2	15.4	46.2	116	567

Table 18: Factuality evaluation scores over the unseen data at temperature = 0.3 when fine-tuning only over the Wikipedia articles from pre-training, using Llama-3.1-8B-Instruct for claim decomposition and verification.

Dataset	Noise Multiplier (σ)	Average FS	Median FS	Q1 FS	Q3 FS	Avg Max FS / Topic	Avg Min FS / Topic	# of FS >0.5	# of FS <0.5
Wikipedia Science	0.0	24.4	38.5	20	55.8	36.7	12.9	34	440
	0.01	24.5	33.3	18.6	54.5	36.6	14.4	37	432
	0.1	22.2	30.4	11.5	51.3	35.1	10.1	41	427
	0.3	24.7	27.9	9.7	46.5	35.3	14.9	50	417
Wikipedia AI	0.0	44.7	19	8.6	33.3	58.8	30.6	252	600
	0.01	42.1	18.2	8.3	35.7	56.7	27.4	231	622
	0.1	41.5	14.3	0	29.2	55.4	26.9	222	618
	0.3	37.3	17.6	5.9	35.7	51.5	23	175	680

Table 19: Factuality evaluation scores for temperature = 0.5 when fine-tuning only over the unseen Wikipedia articles, using Llama-3.1-8B-Instruct for claim decomposition and verification.

Dataset	Pretrained Model	Average FS	Median FS	Q1 FS	Q3 FS	Avg Max FS / Topic	Avg Min FS / Topic	# of FS $\geq$ 0.5	# of FS $<$0.5
AI	Gemma3-1B-PT	51.9	55.6	25.0	79.6	65.8	37.2	222	248
	VaultGemma-1B	**43.2**	43.3	17.0	66.7	59.9	27.7	126	340
	GPT-2-1.5B	27.7	**22.2**	**7.6**	**45.5**	**45.2**	**13.5**	**38**	**366**
Science	Gemma3-1B-PT	50.4	50.0	25.0	75.0	72.7	28.8	219	600
	VaultGemma-1B	39.5	**33.3**	**14.6**	60.0	65.2	**14.8**	**81**	697
	GPT-2-1.5B	**39.1**	**33.3**	16.9	**57.1**	**61.3**	17.9	84	**750**
Pretraining	Gemma3-1B-PT	26.6	17.6	7.1	42.9	39.8	14.7	140	835
	VaultGemma-1B	22.0	14.3	**0.0**	33.3	37.2	9.5	58	**931**
	GPT-2-1.5B	**17.6**	**11.1**	**0.0**	**25.0**	**31.6**	**6.0**	**32**	880

Table 20: FactScore (FS; reported as %) results of pre-trained models, evaluated at temperature $\tau = 0.3$. Reported are average, median, and quartile FactScores, per-topic average maxima and minima, and counts of factual ($\geq$0.5) and non-factual ($<$0.5) responses. Bolding indicates lower FactScore (increased hallucinations).

E Supplementary Experiments

E.1 ANADP: Adaptive Noise Allocation for Differential Privacy

In addition to standard DP-SGD, we evaluate ANADP (Adaptive Noise Allocation DP) (Li et al., 2024b), which redistributes noise away from important parameters while keeping the total noise budget unchanged. We include this baseline to test whether shifting noise away from value matrices responsible for encoding factual associations (Nichani et al., 2024) can improve factuality. However, ANADP's formal privacy guarantee may not hold as stated, the reasons for which we describe below.

We nonetheless evaluate ANADP alongside our standard DP-SGD models, and find that ANADP does not meaningfully mitigate the privacy-hallucination tradeoff. On the Wikipedia AI dataset, ANADP performs at par or worse than standard DP-SGD at both privacy budgets. On the Wikipedia Science dataset, it marginally outperforms standard DP-SGD at $\epsilon = 8$.

Algorithm 3 ANADP Algorithm

Input: Training batches $\mathcal{L} = \{L_1, \ldots, L_T\}$, Initial trainable parameter weights $\omega_0 \in \mathbb{R}^d$, noise multiplier σ_0
Hyper-parameters: α, β_1, β_2, clipping threshold C, learning rate γ
$S_0 \leftarrow \mathbf{0}_d$, $\bar{S}_0 \leftarrow \mathbf{0}_d$, $\bar{U}_0 \leftarrow \mathbf{0}_d$
for $L_t \in \mathcal{L}$ **do**
  Compute gradients $g(L_t) \in \mathbb{R}^d$
  $S_t \leftarrow |g(L_t) \odot \omega_{t-1}| \in \mathbb{R}^d$ ▷ *Param. Sensitivity*
  $\bar{S}_t \leftarrow \beta_1 \bar{S}_{t-1} + (1-\beta_1) S_t \in \mathbb{R}^d$
  $\bar{U}_t \leftarrow \beta_2 \bar{U}_{t-1} + (1-\beta_2)|\bar{S}_t - S_t| \in \mathbb{R}^d$
  $I_t \leftarrow \bar{S}_t \odot \bar{U}_t \in \mathbb{R}^d$
  $\mu \leftarrow \text{mean}\left(\frac{I_t - \text{median}(I_t)}{q_1(I_t) - q_2(I_t)}\right) \in \mathbb{R}$ ▷ *Mean importance*
  $\hat{I}_t \leftarrow (1-\alpha)\left(\frac{I_t - \text{median}(I_t)}{q_1(I_t) - q_2(I_t)}\right) + \alpha\mu \in \mathbb{R}^d$
  $\bar{I}_t \leftarrow \hat{I}_t - \left(\text{mean}\left(\hat{I}_t\right) - 1\right) \in \mathbb{R}^d$
  $\tilde{g}(L) \leftarrow \min\left(g(L), C\right) + \mathcal{N}\left(\mathbf{0}_d, \frac{\sigma_0^2}{\bar{I}_t}\text{Id}_{d\times d}\right)$
  $\omega_t \leftarrow \omega_{t-1} - \gamma\, \tilde{g}(L_t)$ ▷ *Update weights*
end for
Output: Final parameters ω_T

ANADP Mechanism Li et al. (2024b) propose an adaptive coordinate-wise noise allocation scheme called ANADP inspired by parameter-importance scores. We recall the algorithm in Algorithm 3 and explain the shortcomings in the paper's privacy analysis.

The algorithm first computes the sensitivity of the training loss w.r.t. individual model parameters. This is derived as a linear approximation of the effect of zeroing a certain parameter. To be precise, suppose $\omega^{(-i)}$ is obtained by zeroing out the i^{th} entry $[\omega]_i$ of $\omega \in \mathbb{R}^d$, then we have the linear approximation $\ell(\omega^{(-i)}) \approx \ell(\omega) - [\nabla\ell(\omega)]_i[\omega]_i$. This is distinct from "sensitivity" in the context of DP (which refers to the effect of changing a datapoint on the output of a function) and is not to be confused. It then defines the "importance" of a parameter as the product of this sensitivity with its uncertainty U_t — the latter measures how this sensitivity changes from its moving average. Parameters with high "importance" are given smaller noise.

Issues with the privacy analysis The privacy proof of ANADP appears incomplete. They state that the mean (across parameters) of the inverse noise variance is the same as that of DP-SGD. However, this argument is incorrect for the following reason.

Consider a Gaussian mechanism that adds anisotropic Gaussian noise $\mathcal{N}(\mathbf{0}, \Sigma)$. A correct privacy analysis of this mechanism requires calculating the sensitivity of the original algorithm w.r.t. the Mahalanobis norm $\|u\|_{\Sigma^{-1}} := \sqrt{u^\top \Sigma^{-1} u}$. In particular, the sensitivity (and hence DP) is governed by directions with small noise rather than by the trace of Σ or Σ^{-1}. Consequently, any mechanism that reduces noise on some coordinates can incur larger privacy loss along those directions unless the clipping/sensitivity definitions are modified accordingly.

Moreover, ANADP computes parameter importance from the same private batch gradients to which the noise is subsequently added, making the covariance data-dependent. Thus, a reduction to standard DP-SGD accounting is not mathematically correct.

We are not aware of a proof that correctly handles these two aspects in this paper or in subsequent works. Thus, the stated DP guarantee of ANADP should be treated as unverified. We include it in our experimental baselines to see if anisotropic noise addition can help reduce hallucinations. We find that anisotropic Gaussian noise in this form does not help reduce hallucinations, although it does reduce the recurring incorrect claims generated by

the model as shown in Table 7.

	Response-level				Topic-level
DP Setting	**Avg FS**	**Med**	**Q1**	**Q3**	**Avg FS [95% CI]**
AI					
$\epsilon = 16$	33.5	25.0	10.0	50.0	33.3 [29.2, 37.6]
$\epsilon = 8$	32.0	25.0	11.1	50.0	31.7 [27.7, 35.9]
Science					
$\epsilon = 16$	52.5	53.1	28.6	77.8	52.5 [48.9, 56.0]
$\epsilon = 8$	53.6	55.6	28.6	79.3	53.7 [50.2, 57.1]

Table 21: FactScore (FS; %) of GPT-J 6B fine-tuned with ANADP at two privacy budgets, evaluated at temperature $\tau = 0.3$. Response-level statistics pool all generations; topic-level averages within each entity first, with bootstrap 95% CIs resampled over topics.

E.2 Controlled Fact Acquisition Experiment

To study the conditions under which models trained with differentially private guarantees acquire factual associations, we construct a controlled memorization benchmark using synthetic facts of known frequency. To simulate this experiment, we generate a set of fictional entity associations using a fixed set of templates (details provided below) with fictional subject-object pairings. We use a fresh set of entities, rather than reusing the templates from the bigram experiments in Section 4 to avoid any confounding effects of the template patterns on the frequency effects, as these templates contain cues that language models are known to exploit when memorizing information (Shilov et al., 2026).

This is to ensure that no fact can be answered from prior knowledge, so as to isolate the effect of training frequency and DP noise from pre-existing associations in the model's weights. Every fact has an associated frequency per epoch (where $f \in \{1, 3, 7, 15, 30, 60\}$), and each frequency has five facts associated with it. We train the randomly initialized and pre-trained versions of GPT-2 to simulate the pre-training and fine-tuning regimes where differentially private guarantees may be applied. We freeze the input token embeddings and positional embeddings during training (we observe no significant differences in the setup for the randomly initialized model when these parameters are unfrozen). We train with DP-SGD using Opacus, with a clipping norm $C = 1.0$ and $\sigma \in \{0.0, 0.1, 0.3, 0.5, 0.7, 1.0\}$, where $\sigma = 0$ is equivalent to training without noise. We apply Poisson subsampling with a batch size of 64 and use a learning rate of 5e-3 with the AdamW optimizer in both setups (we perform a grid search to find the optimal learning rate).

For each epoch, we evaluate the factual recall by prompting the model with the fact prefix and measuring $P(\text{correct completion})$, which is the probability assigned to the ground-truth answer token. We report the mean factual recall across the five facts within each tier and define the frequency threshold as the minimum f at which the factual recall is $\geq$0.5.

E.2.1 Templates and Subject-Object Lists

Facts are generated from six fixed templates of the form `relation prefix, subject, object`, covering a range of relation types:

- *the capital of* [subject] *is* [object]
- *the leader of* [subject] *is* [object]
- *the currency of* [subject] *is* [object]
- *the main export of* [subject] *is* [object]
- *the official language of* [subject] *is* [object]
- *the founder of* [subject] *was* [object]

Subjects: zephyria, kaldor, belvane, thornwick, maldren, corvath, selenix, dravion, arcthos, velrune, pyraxis, glenmoor, obsidyn, halcyon, nexara, stratholm, verdania, cryosten, luminex, dawnridge, ironvale, novaheim, solheim, temporia, crystara, emberfell, frostholm, goldenreach, shadowmere, titanforge

Objects: mordath, ventris, draxil, seraph, luxon, kestrel, zircon, thalis, silvane, fenwick, caldris, orinath, vexel, aldric, casciel, delmar, evander, faelorn, gareth, ithral, jarenth, kelwyn, lorenth, maelis, norvin, pellarn, quillon, raveth, stellan, thandor

σ	ε @ $\delta = 1/N^{1.1}$
0.0	∞
0.1	1310.535
0.3	108.122
0.5	29.847
0.7	13.433
1.0	6.397

Table 22: Mapping of the noise multiplier σ to the privacy budget ε for the RQ3 setup, computed with the PLD accountant at $\delta = 1/N^{1.1}$. Values are reported for the full training run. $\sigma = 0$ corresponds to non-private training ($\varepsilon = \infty$).

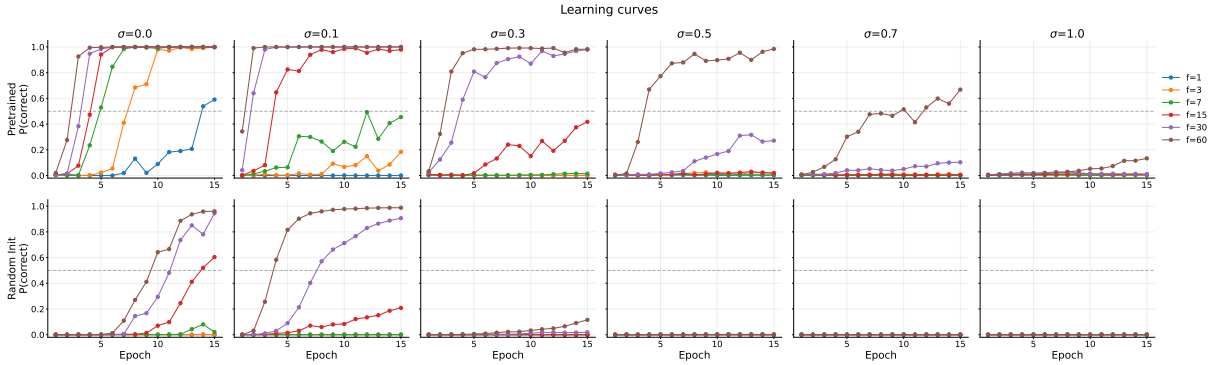


Figure 24: Factual recall over training epochs for pre-trained (top) and randomly initialized (bottom) GPT-2 under DP-SGD with varying noise multipliers $\sigma \in \{0.0, 0.1, 0.3, 0.5, 0.7, 1.0\}$ and fact frequencies $f \in \{1, 3, 7, 15, 30, 60\}$. Higher σ requires greater fact frequency for successful recall, and pre-trained models learn at lower frequencies than randomly initialized ones.

E.3 Layer-wise Signal-to-Noise Ratio

Our experiments revealed that the signal-to-noise ratio is the highest for the LoRA value matrices in the early-to-middle layers. (Nichani et al., 2024) shows that the value matrices are associated with factual recall, explaining the increase in hallucinations. The query matrices tend to yield far lower SNRs across layers. For both the query and value matrix updates, the SNR declines progressively because the magnitude of the signal diminishes, particularly in the early stages of fine-tuning. The plots indicate that the clipping contributes to the utility being hurt in early stages (by bounding larger signals), while the DP noise obfuscates updates during later stages of fine-tuning. This strongly suggests that meeting stricter privacy budgets in data-constrained settings is challenging. While adaptations to the noise allocation in DP have been explored for encoder models (Li et al., 2024b), these methods rely on increasing noise added to the intermediate layers responsible for factual recall and would likely further hurt factual accuracy.

We evaluate this by fine-tuning only the first 15 layers (roughly half the layers). In this setup, training loss converges faster and converges more smoothly, comparable to joint query+value LoRA updates. In contrast, freezing the query matrices slows down convergence (Figure 25).

E.4 Additional Examples of Claim Clusters

We include additional examples of unsupported claim clusters in Table 23.

F LLM Usage

We used large language models to help with the writing of this paper. Specifically, we used ChatGPT to generate the code for LaTeX tables and figures in this research paper and to assist in generating the captions for some of the figure descriptions.

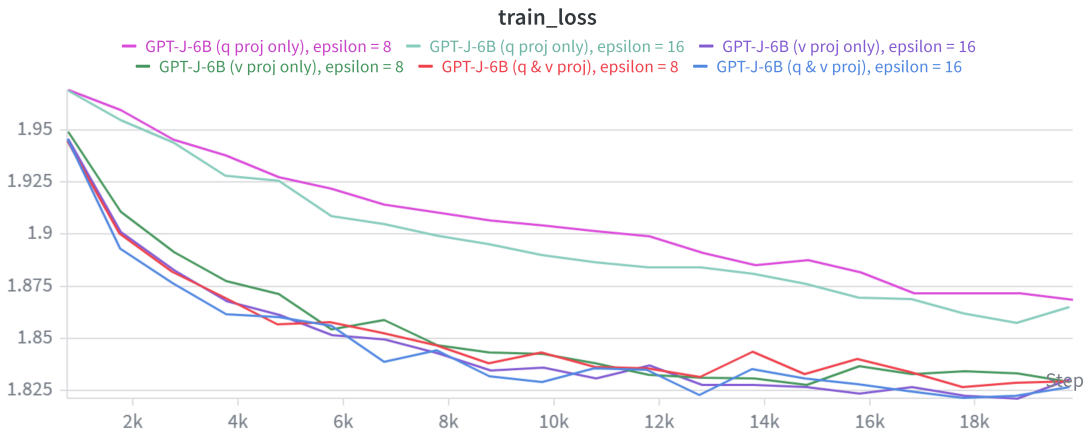


Figure 25: Value-only LoRA fine-tuning drops faster and more smoothly with loss convergence on par with full-finetuning, while updates to only the query matrices converge more slowly and plateau faster.

DP Setting	Claims
Data: Wikipedia AI, Topic: AlphaEvolve	
$\epsilon = \infty$	'AlphaEvolve is used for generating the molecular structures of organic molecules.', 'AlphaEvolve is used for generating molecular structures.'
	'The project or system is based on the DeepChem molecular modeling framework.', 'It is based on the DeepChem molecular modeling framework.'
$\epsilon = 16$	'AlphaEvolve is a first-person shooter.', 'AlphaEvolve is a first-person shooter video game.'
	'The game features a series of missions.', 'The game features a variety of weapons.'
	'There are two types of enemies in the game.', 'The game features two types of enemies.'
$\epsilon = 8$	'Black Hole Interactive is a game development company', 'Black Hole Interactive is a video game development company.'
	'The Behemoth is a studio.', 'The Behemoth is a video game development studio.', 'The Behemoth is an American studio.'
	'AlphaEvolve is a shooter video game', 'AlphaEvolve is free-to-play', 'AlphaEvolve is a free-to-play video game', 'AlphaEvolve is a 2D video game.', 'AlphaEvolve is a physics-based video game.'
Data: Wikipedia Science, Topic: Allogeneic processed thymus tissue	
$\epsilon = \infty$	'The thymus tissue is processed to remove stem cells.', ('The thymus tissue is processed to induce the recipient's immune cells to develop.'
	'AML is a type of disease.', 'AML is a type of cancer.'
$\epsilon = 16$	'CLL is also known as chronic lymphocytic leukemia.', 'CLL is often referred to as chronic lymphocytic leukemia.'
	'Thymus glands are found in donor pigs.', 'The thymus glands are minced.', 'The thymus glands are from donor pigs.', 'The thymus glands are removed from pigs.'
$\epsilon = 8$	'Dr. Daniel L. Scharff is the founder of the International Society for Cellular Therapy (ISCT).', 'Dr. Daniel L. Scharff is the former President of the International Society for Cellular Therapy (ISCT)'
	'Antigen presenting cells can cause graft rejection.', 'Lymphocytes can cause graft rejection'

Table 23: Examples of unsupported recurring claim clusters