

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 THE PRIVACY-HALLUCINATION TRADEOFF IN DIFFERENTIALLY PRIVATE LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

While prior work has studied privacy tradeoffs with utility and fairness, the impact of privacy-preservation on factual consistency and hallucination in LLM outputs remains unexplored. Given that privacy-preservation is paramount in high-stakes domains like healthcare, the factual accuracy of these systems is critical. In this study, we uncover and investigate a *privacy-hallucination tradeoff* in differentially private language models. We show that while stricter DP guarantees do not distort knowledge acquired during standard pre-training, they hinder the model’s ability to learn new factual associations when fine-tuned on previously unseen data, as a result of which the model tends to hallucinate incorrect or irrelevant information instead. We find that the proportion of factual texts generated drops by 17-24% when models are fine-tuned on the same data using DP ($\epsilon = 8$), compared to the non-DP models, and on average, the factuality scores differ by at least 3-5%. This disparity is further pronounced when pre-training with DP, where we find a 43% drop in the number of factually consistent texts. 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 setting 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 gold-standard paradigm to provably mitigate such 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 issues, 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 a body of research that aims to evaluate and improve factual correctness in LLM outputs (e.g. Ji et al., 2023; Li et al., 2024). 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 trade-off 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 indicate that, rather than only reducing the generation of rare facts, private training problematically leads to a general increase in hallucinations.

¹We use “hallucination” to refer to LLM generation of false or misleading information presented as fact. Unless stated otherwise, we restrict ourselves to information not supported by or contradictory to the training (pre-training or fine-tuning) data.

054 **Goals and Questions.** Concretely, we study the interplay of DP training (using the DP-SGD algorithm for various
 055 privacy budgets) and hallucination in open-ended generation from language models. We consider potential hallucinations
 056 relative to facts found both in fine-tuning and pre-training datasets through the following research questions:
 057

- 058 (RQ1) What impact does *DP fine-tuning* have on hallucinations concerning facts present in the *fine-tuning data*?
 059 (RQ2) What impact does *DP fine-tuning* have on the hallucination (or forgetting) of factual information already
 060 learned by the model from the *pre-training data*?
 061 (RQ3) What impact does *DP pre-training* have on hallucinations concerning facts present in the *pre-training data*?
 062

063 **Main findings.** Through a detailed experimental study, using both automatic (Min et al., 2023) and human evaluations,
 064 we find the following:
 065

- 066 • **RQ1:** A more stringent privacy budget (i.e. smaller ϵ for DP) leads to consistent increase in hallucinations, as
 067 measured both by automatic and human evaluations (automated factuality scores decline by 3-5 points from non-DP
 068 to DP models). This highlights a privacy-hallucination tradeoff for fine-tuning. Further, we find that the DP models
 069 tend to repeatedly hallucinate the same incorrect claims, suggesting a systematic encoding of inaccurate facts rather
 070 than random errors.
- 071 • **RQ2:** We find no significant difference in the hallucination of pre-training facts with or without DP fine-tuning. That
 072 is, the noise injected during DP fine-tuning does not obfuscate knowledge encoded by models in pre-training.
- 073 • **RQ3:** The DP-pre-trained VaultGemma-1B model hallucinates significantly more than a similar non-private Gemma-
 074 3-1B model (factuality scores decline by 4-11 points). In some domains, its hallucination rate is comparable to a
 075 GPT-2 model that was never trained on these facts. Thus, the privacy-hallucination tradeoffs from pre-training are
 076 significantly worse than for fine-tuning (RQ1).

077 In summary, our results show a privacy-hallucination tradeoff in language models across both pre-training and fine-
 078 tuning settings. Our findings highlight the need for more nuanced privacy-preservation mechanisms that can protect
 079 data privacy without increasing the hallucination rate of the models. Upon publication, we will publicly release our
 080 code and data required to reproduce our research, and to facilitate future work in this direction.
 081

2 RELATED WORK

083 **Privacy in LLMs.** This risk of privacy leakage by language models has inspired work on provable privacy-preserving
 084 strategies such as differential privacy (Li et al., 2022; Miranda et al., 2025), as well as heuristics such as knowledge
 085 unlearning to reduce the influence of sensitive data points on the model parameters (Jang et al., 2023; Zhang et al.,
 086 2024) and knowledge editing to locate and modify neurons containing private information (Wu et al., 2023; 2024).
 087

088 Privacy-preserving methods typically (either directly or indirectly) involve changes to the model’s parameters. For
 089 instance, DP training is typically accomplished through a modified version of SGD, which involves clipping and
 090 noising the gradient update (Abadi et al., 2016). Knowledge unlearning methods trace and remove an approximate
 091 estimate of the influence of a training point on the model’s parameters. All of these approaches contribute to some
 092 degradation in model utility: DP limits the information the model learns, while knowledge unlearning and knowledge
 093 editing affects useful non-private information the model has encoded. We focus on the rigorous and future-proof
 094 guarantees of DP since heuristic privacy defenses can often be broken (Aerni et al., 2024; Du et al., 2024).
 095

096 **Tradeoffs from DP.** The classical no-free-lunch theorem of DP states that DP necessarily incurs a penalty on utility
 097 (Kifer & Machanavajjhala, 2011). In practice, this results in a privacy-utility-compute tradeoff (McMahan et al.,
 098 2018; Ponomareva et al., 2023). Research has since established that privacy also comes at a cost to fairness in statistical
 099 estimation tasks (Tran et al., 2021), discriminative models (Bagdasaryan et al., 2019; Farrand et al., 2020;
 100 de Oliveira et al., 2024), and LLMs (Lyu et al., 2020; Matzken et al., 2023; Ramesh et al., 2024; Hansen et al., 2024).
 101 Ngong et al. (2025) explore the adverse effect of DP on elements such as grammatical correctness, fluency and the
 102 coherence of model-generated text. We note that it is possible to optimize for more balanced privacy-utility tradeoffs
 103 (Miresghallah et al., 2021) or privacy-fairness tradeoffs (Pillutla et al., 2024; Zhou & Bassily, 2024).
 104

105 Specifically in the text domain, prior work has empirically examined the privacy-utility tradeoffs with task-specific
 106 measures of utility, including classification accuracy, linguistic aspects such as fluency, grammatical correctness and
 107 lexical diversity. In contrast, we focus on tradeoffs between DP and hallucination. Minimizing hallucinations is highly
 108 desirable, and is often distinct from other task-specific measures of utility surveyed above.
 109

110 **Factuality in Language Models.** No single factuality metric generalizes across settings, so it is common to use task-
 111 or domain-specific methods. Early research focused on measures such as the factual precision of cloze-style or short-
 112

108 form responses (Youssef et al., 2023; Petroni et al., 2019), and NLI-based methods to determine whether generated
 109 summaries are consistent with their source document (Chen et al., 2021; Fabbri et al., 2022; Laban et al., 2022). The
 110 rise of LLM applications with open-domain, free-form model-generated text, where there can be multiple plausible
 111 responses has led to the development of factuality metrics in these settings (Min et al., 2023; Wei et al., 2024; Song
 112 et al., 2024). This is significantly more challenging, as the decoding strategy also influences model outputs (Wang
 113 et al., 2024), and models can generate correct answers across multiple attempts (Tian et al., 2023). Further, they may
 114 generate factually accurate content that contradicts or is unsupported by the training data (Cao et al., 2022).² We
 115 leverage existing state-of-the-art methods for fact-checking in open-ended generation along with human evaluations
 116 to analyze privacy-hallucination tradeoffs in LLMs (Min et al., 2023).

117 3 EXPERIMENTAL DESIGN

120 Our goal is to measure how factual hallucinations are affected by DP at various privacy budgets. To this end, we
 121 carefully construct experimental setups where we can fine-tune valid differentially private models, control for overlap
 122 between fine-tuning and pre-training data, and evaluate factual accuracy in open-ended model outputs. As the specific
 123 choices of datasets, models, and training setups are crucial for achieving these requirements, we include those details
 124 within the broader description of our approach. Our experimental design is reusable for facilitating future analyses of
 125 LLM behavior, especially settings that similarly require careful separation of data not included in pre-training.

127 3.1 MODELS AND TRAINING SETUP

128 **Fine-tuning setup for RQ1 and RQ2.** We fine-tune LLMs for controllable text generation, similar to previously
 129 proposed applications for DP in LLMs, particularly for privacy-preserving synthetic and open-ended text generation
 130 (Yue et al., 2023; Mattern et al., 2022; Ramesh et al., 2024). We provide a brief overview of the settings here and
 131 discuss the precise details and detailed background in Appendix A.

132 DP fine-tuning is achieved using a stochastic gradient optimization approach known as DP-SGD (Abadi et al., 2016).
 133 This algorithm bounds the information learned from each sample by clipping the per-sample gradients to a fixed ℓ_2
 134 norm bound, and perturbs them (for DP) with white Gaussian noise. The scale of the Gaussian noise is calibrated to
 135 the desired (ε, δ) -DP guarantee.

136 The DP guarantees are provided with respect to the add-or-remove adjacency at the sequence-level, i.e., the model
 137 outputs should be nearly indistinguishable if a new sequence of 1024 tokens is added to or removed from the training
 138 dataset. We set privacy budgets of $\varepsilon \in \{8, 16\}$, and $\delta = n^{-1.1}$, where n is the dataset size (in terms of number of
 139 sequences). We use the DP-SGD implementation from Opacus (Yousefpour et al., 2022) and measure the privacy
 140 budget consumed using the PLD accountant (with amplification by sampling) (Doroshenko et al., 2022). All fine-
 141 tuning (DP and non-DP) is achieved using Low-Rank Adaption (LoRA) for computational efficiency (Hu et al., 2022).
 142 We specify additional details about LoRA and all hyperparameter settings in Appendix B.

143 The base language model behind all fine-tuning experiments is GPT-J 6B (Wang & Komatsuzaki, 2021). This is a
 144 highly-performant decoder-only transformer model, similar to currently popular models. We make this choice for a
 145 crucial reason: its pre-training dataset, namely The Pile, is fully open and known, with a known cutoff date. This lets
 146 us ascertain that GPT-J is not pre-trained on any post-2020 data. This knowledge of the pre-training data mixture and
 147 cut-off date allows us to select fine-tuning data that has no overlap with pre-training data; cf. §3.2.

148 **Private pre-training for RQ3.** We address the effect of DP pre-training (as opposed to fine-tuning) on factuality using
 149 the recently released DP-pre-trained VaultGemma (VaultGemma Team, 2025). This is a 1B-parameter open-weights
 150 model fully pre-trained with DP ($\varepsilon = 2$); this is the largest model known to be pre-trained from scratch with DP as of
 151 this writing. We compare VaultGemma with the same two models as in VaultGemma Team (2025):

- 152 • Gemma3-1B, which can be viewed as the “non-private counterpart” of VaultGemma, and is expected to be similar
 153 except for the use of DP.
- 154 • GPT-2 XL (1.5B), whose performance is similar to VaultGemma on standard benchmarks. It has a much earlier
 155 knowledge cut-off date, as it was released in 2019.

156 ²We regard this as hallucinations for the purpose of this work as the training dataset is considered as the sole *source of truth*.

Dataset Name	Size	Examples/Description
Wikipedia Science	231	Follicular drug delivery; Malaria therapy; Eurotrac
Wikipedia AI	124	DeepSeek; DeepSeek (chatbot); DeepSeek (disambiguation); DALL-E
Wikipedia Pretraining	250	[randomly sampled articles]
Fine-tuning Data	20355	Wikipedia Science + Wikipedia AI + 20,000 randomly sampled Wikipedia articles

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

3.2 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, the articles’ meta-data 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.

3.2.1 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 D.

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 of the concepts in these articles 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 also could not have existed in other pre-training data sources. 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. For RQ2, where we investigate effects of DP fine-tuning on knowledge acquired during pre-training, we randomly sample 250 Wikipedia articles from the GPT-J pretraining data. We make sure that these articles are not included in fine-tuning dataset.

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

216 The data is divided into sequences of 1024 tokens, which is the unit of privacy protection. We fine-tune the model to
 217 produce an article when prompted on the article title (e.g., a topic). For evaluations, we similarly prompt the model
 218 with article titles and evaluate the factual accuracy of the generated text.
 219
 220

221 3.3 EVALUATION OF FACTUAL ACCURACY IN OPEN-ENDED TEXT GENERATION

223 Hallucinations are factual inaccuracies, so a higher factual accuracy is indicative of a lower hallucination rate. We
 224 evaluate the factual accuracy in open-ended text generation settings where we prompt the model with a title and
 225 generate a full Wikipedia article. We note that this is closer to the concurrent uses of LLMs, as opposed to traditional
 226 factuality evaluations based on cloze-style or short-form response queries designed to probe models (Youssef et al.,
 227 2023; Petroni et al., 2019). We use both automated and human assessments of factuality, as described below.
 228

229 **Automated Evaluation via FactScore.** Given a generated text d_i , FactScore (Min et al., 2023) operates in two distinct
 230 phases: (i) **atomic claim extraction**, where d_i is decomposed into a set of minimal, verifiable claims \mathcal{AF}_{d_i} , and (ii)
 231 **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
 232 both intrinsic language model judgments and evidence retrieved from an external knowledge source \mathcal{K} , as detailed
 233 in Algorithm 1, Appendix F. We use the original Wikipedia article as the external knowledge source for verifying
 234 generated claims. As claim decomposition methods can produce redundant claims that artificially inflate scores, we
 235 use the CORE module (Jiang et al., 2024b) to filter down the superfluous and repetitive claims.
 236

237 We use Llama-3.1-8B-Instruct for steps (i) and (ii) above involved in FactScore computations. We verify in Appendix I
 238 that the overall trends of FactScore results are consistent with other choices of models.
 239

240 **Human Evaluation.** As FactScore is an automated metric relying on LLM judgments that may not be accurate, we
 241 conduct human evaluations to validate results and provide finer-grained analysis of outputted information. We use the
 242 Wikipedia AI dataset, as this dataset most carefully separates the pre-training and fine-tuning data. For the annotation
 243 task, we recruited 5 computer science graduate students, whom we expect to have high AI literacy. We ensure that the
 244 human evaluators see generations with both high and low factual accuracy using stratified sampling: we sample 15
 245 articles with both DP ($\varepsilon = 16$) and non-DP models have FactScore $> 50\%$ and another 15 articles where both models
 246 have FactScore $< 50\%$. Our custom annotation interface (Appendix §E) displays a source text and a generated text
 247 side-by-side, followed by automatically decomposed atomic claims. Annotators rated claims for their overall veracity
 248 (correct/incorrect/unclear) and their support within the source text. Annotators were also asked to flag quality issues in
 249 decomposed claims, marking them as vague or subjective. Two annotators rated each claim, and we report the annotator
 250 agreement for both the non-DP and DP texts in Table 10 as Cohen’s Kappa values for both the veracity of the claim,
 251 and whether or not it is supported by evidence in the training text corresponding to the topic it was generated for.³
 252

253 3.4 ANALYSIS OF RECURRING HALLUCINATIONS

254 The metrics in §3.3 can capture if models hallucinate incorrect facts, but they do not distinguish between models that
 255 output a range of incorrect information (suggesting general noisiness) or if they repeatedly output the same incorrect
 256 information across generations (suggesting encoding of inaccurate facts). To identify recurring factual claims across
 257 model generations, we propose a multi-stage clustering algorithm. We give a high-level summary below, with precise
 258 pseudo-code in Algorithm 2, Appendix F.

259 Consider the generated synthetic documents grouped by the the topic $t \in \mathcal{T}$ used to prompt their generation. Each
 260 document is decomposed into atomic claims and these claims are aggregated into a claim set $C_t = c_i$ for each topic t .
 261

262 We first index all extracted claims to their source (synthetic) documents, then cluster them using sentence-
 263 embedding-based agglomerative clustering to group semantically similar claims. To refine boundaries, we apply
 264 DBSCAN with Jaccard similarity, ensuring clusters are both semantically coherent and lexically consistent. This
 265 reduces cases where semantically related but factually distinct claims are grouped together.
 266

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

³ Annotators were offered compensation at an estimated payment rate of \$20/hr

Dataset	DP Setting	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 AI	$\epsilon = \infty$	37.7	33.3	12.5	58.7	51.4	24.8	132	346
	$\epsilon = 16$	35.0	30.8	12.3	54.5	49.9	21.7	100	388
	$\epsilon = 8$	32.2	25.0	10.0	50.0	45.9	19.0	80	404
	Base	32.9	27.3	9.1	50.0	47.5	19.3	78	362
Wikipedia Science	$\epsilon = \infty$	56.1	58.3	33.3	78.9	70.3	41.6	393	495
	$\epsilon = 16$	54.9	57.1	30.0	80.0	70.1	39.3	323	555
	$\epsilon = 8$	53.2	55.6	28.6	77.8	68.5	37.9	290	592
	Base	52.1	50.0	27.3	75.0	76.0	27.5	190	435

Table 2: FactScore (FS; reported as a percentage) results from GPT-J 6B, fine-tuned with different DP budgets, evaluated at temperature $\tau = 0.3$. Reported are average, median, and quartile FactScores, per-topic average maxima and minima, and counts of factual (≥ 0.5) and non-factual (< 0.5) responses. “Base” refers to the base GPT-J 6B model with no fine-tuning. The row with the lowest FactScore (i.e. most hallucinations) is bolded. The factual consistency decreases with stricter privacy budgets (smaller ϵ), highlighting the tradeoff between privacy and hallucinations.

4 EXPERIMENTAL RESULTS

We empirically study each of the three research questions in turn.

4.1 RQ1: HALLUCINATION CONCERNING FINE-TUNING FACTS IN DP FINE-TUNED MODELS

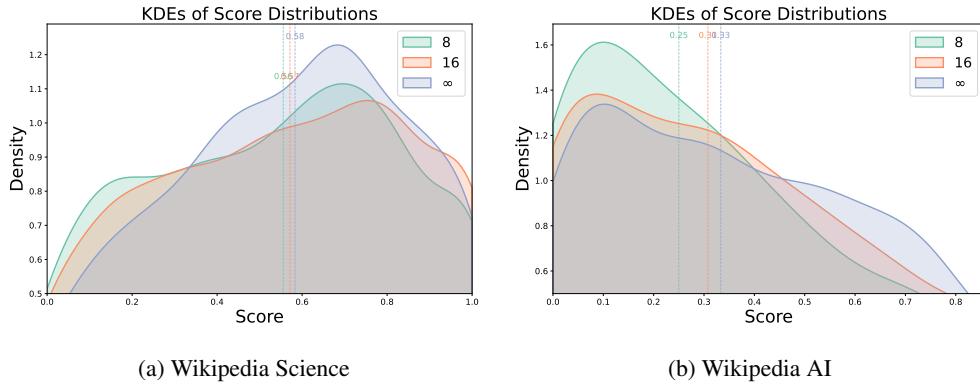


Figure 1: Kernel density (KDE) plots comparing FactScore distributions when finetuning GPT-J 6B under different DP budgets. As the privacy budget becomes more stringent (ϵ from ∞ to 8), the FactScore distribution shifts leftward, indicating that stronger privacy constraints result in less factually consistent outcomes.

Automated Evaluation. The automated factuality evaluation for unseen fine-tuning data in Table 2 suggests that DP fine-tuning leads to more frequent hallucinations. Specifically, we observe the following general trend in factuality: non-DP ($\epsilon = \infty$) \succ DP ($\epsilon = 16$) \succ DP ($\epsilon = 8$) \succ base ($\epsilon = 0$), where “base” denotes the pre-trained model. That is, we observe a consistent decrease in the average FactScore for the text generated from models fine-tuned with DP compared to those trained without DP. This decline is more pronounced for the stricter privacy budget setting $\epsilon = 8$, and is further illustrated by the skew (toward lower factuality scores) in the distributions of FactScores for DP-trained models in Figure 1. Furthermore, the proportion of factual summaries ($\text{FactScore} \geq 0.5$) also progressively decreases with stricter privacy budgets.

As a sanity check, 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 the usefulness of fine-tuning. Notably, on the Wikipedia AI dataset, the model fine-tuned with $\epsilon = 8$ performs worse or comparably (mean: 32.2%; 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. While this marginal degradation in FactScore

324	325	326	327	328	329	330	Wikipedia AI		FactScore < 0.5		FactScore ≥ 0.5				
									$\varepsilon = \infty$	$\varepsilon = 16$	$\varepsilon = \infty$	$\varepsilon = 16$			
									Avg. Veracity Score	0.18	0.17	0.51			
										Avg. Support Ratio	0.19	0.09			
										Quality Issues (count)	87	79			
										Quality Issues (count)	55.5	79			

331 Table 3: Comparison of human annotation results on Wiki AI FactScore summaries at $\varepsilon = \infty$ and $\varepsilon = 16$, stratified by response
 332 groups with veracity scores below 0.5 and above 0.5. Reported metrics are average veracity score, average support ratio, and
 333 number of claims annotated as having quality issues. In both cases, we observe an increase in hallucinations with DP ($\varepsilon = 16$)
 334 relative to no DP ($\varepsilon = \infty$).

336	337	338	339	340	341	342	343	Supported			Unsupported			344		
								DP Setting	Recur	Total	% Avg	Recur	Total	% Avg		
									Count	Count	Recur	Count	Count	Recur		
345	346	347	348	349	350	351	352	Wikipedia	$\varepsilon = \infty$	242	1198	20.20	494	1992	24.80	0.490
									$\varepsilon = 16$	200	1103	18.13	523	2002	26.12	0.382
									$\varepsilon = 8$	174	1019	17.08	509	2011	25.31	0.342
353	354	355	356	357	358	359	360	Wikipedia	$\varepsilon = \infty$	705	2896	24.34	560	2417	23.17	1.259
									$\varepsilon = 16$	582	2467	23.59	556	2128	26.13	1.047
									$\varepsilon = 8$	556	2378	23.38	633	2222	28.49	0.878

353 Table 4: Analysis of recurrent claims and hallucinations (§3.4) for temperature $\tau = 0.3$. “Total Count” reports the total number of
 354 generated clusters, “Recur Count” reports the number of those clusters with ≥ 2 supporting documents, and %Avg Recur is “Recur
 355 Count”/“Total Count”. The far right column reports “Recur Count” of supported claims / “Recur Count” of unsupported claims.
 356 DP models output a lower ratio of repeated supported claims to unsupported claims, suggesting increased repeated hallucinations.
 357 Bolding indicates most hallucinations (i.e. least supported or most unsupported).

358 for Wikipedia AI articles could be statistical noise, it could also indicate the gradient signal from the AI articles being
 359 masked by the larger levels of noise introduced by DP-SGD when fine-tuning with lower privacy budgets ($\varepsilon = 8$).
 360 Under a more generous budget ($\varepsilon = 16$), FactScore does increase compared to the baseline (mean: 35.0%), but still
 361 falls short of the model trained without DP (mean: 37.7%).

362 FactScore is higher for the Science articles compared to AI articles; median scores are consistently $\geq 50\%$. Furthermore,
 363 the DP model with $\varepsilon = 8$ achieves a higher mean FactScore than the base model, suggesting that DP fine-tuning
 364 hallucinates less than the base model. We again observe a clear privacy-hallucination tradeoff where FactScore de-
 365 creases with more stringent privacy budgets (smaller ε). The differences between the AI and Science datasets likely
 366 results from their differing overlap with pre-training data: concepts from Science articles are more likely to occur in
 367 pre-training data, even if exact articles are non-overlapping.

368 **Human Evaluation.** Table 3 reports the average ratings selected by human annotators. In lower-quality generations
 369 ($\text{FactScore} \leq 0.5$), annotators rated both the non-DP and DP models with equally low veracity, while the DP model
 370 exhibited more unsupported facts. In higher-quality generations ($\text{FactScore} > 0.5$), annotators rated the DP model
 371 as outputting both lower veracity information and more unsupported facts. There are not conclusive differences in
 372 counts of quality issues. While agreement between human annotators was generally high, model-human agreement
 373 was not always high (Table 10), indicating limitations of relying exclusively on automated evaluation. Regardless,
 374 overall trends are consistent between human and automated evaluations: both indicate greater hallucination in the DP
 375 model, even the model with a more generous privacy budget.

376 **Recurring Hallucination Analysis.** Table 4 reports results from the claim clustering analysis, as described in §3.4. In
 377 both datasets, the DP model with $\varepsilon = 8$ outputs fewer recurring supported claims (i.e. factually correct statements) than
 378 other models. DP models also output more recurring unsupported claims than non-DP models, with $\varepsilon = 8$ highest for
 379 Wikipedia Science and $\varepsilon = 16$ highest for Wikipedia AI. For both datasets, the ratio of supported recurring claims to
 380 unsupported recurring claims is consistently lower with stricter privacy budgets. This indicates that the hallucination-

378	$\epsilon = \infty$	‘AlphaEvolve is used for generating the molecular structures of organic molecules.’, ‘AlphaEvolve is used for generating molecular structures.’
379	$\epsilon = 16$	‘There are two types of enemies in the game.’, ‘The game features two types of enemies.’
380	$\epsilon = 8$	‘Black Hole Interactive is a game development company.’, ‘Black Hole Interactive is a video game development company.’

384 Table 5: Example unsupported claim clusters (hallucinations) for each model. We provide additional examples in Appendix P.
385

DP Budget	Average FS	Median FS	Q1 FS	Q3 FS	Avg Max FS / Topic	Avg Min FS / Topic	# of FS ≥ 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

394 Table 6: FactScores (FS; reported in %) for GPT-J evaluated with temperature $\tau = 0.3$ over Wikipedia pre-training, which contains
395 articles likely to be in pre-training data, but not included in fine-tuning. DP models perform similarly as non-DP models, suggesting
396 no disruption to facts learned in pre-training. Bolding indicates worse FactScores (i.e. higher hallucinations).
397398 privacy tradeoff involves groups of similar factual inaccuracies, suggesting a systemic encoding of inaccurate facts,
399 rather than random hallucinations.400 In Table 5, we show an example of a cluster of recurring hallucinations for each model. Here, the generation prompt
401 was ‘AlphaEvolve.’ The non-DP model correctly outputs that AlphaEvolve is a model, but incorrectly describes
402 model use. In contrast, the DP models both hallucinate that AlphaEvolve is a video game, with repeated fabricated
403 information about the development and game play.404
405 4.2 RQ2: HALLUCINATION OF PRE-TRAINING FACTS IN DP FINE-TUNED MODELS406 Next, we turn to RQ2 to address whether DP fine-tuning degrades the knowledge already encoded during standard
407 pre-training. Table 6 reports FactScores over Wikipedia pre-training. The differences between DP and non-DP models
408 are marginal, suggesting DP finetuning does not disrupt factual knowledge acquired from pre-training data. This trend
409 is consistent across temperatures (Figure 6; Appendix) and stands in stark contrast to the evaluations on previously
410 unseen data, where stronger privacy constraints correlate with lower factual accuracy.

411 4.3 RQ3: HALLUCINATIONS IN DP PRE-TRAINED MODELS

412 Finally, we investigate if DP pre-training, as opposed to fine-tuning, leads to increased hallucinations in Table 7. It
413 is worth noting that the pre-training data mixtures for the Gemma models likely include *all* the datasets used in our
414 factual evaluations, while we expect GPT-2 to only have been trained on the Wikipedia pre-training evaluation set.415 The DP pre-trained model (VaultGemma) consistently outputs a greater percentage of inaccurate facts relative to the
416 non-DP pre-trained Gemma3-1B. On the Wikipedia pre-training data, Gemma3 achieves an average FactScore of 26.6,
417 compared to 22.0 for VaultGemma. This difference is reflected further in the proportion of factual summaries, with
418 Gemma3-1B producing over twice as many factually correct texts as VaultGemma. This gap in factual correctness
419 is more pronounced for the domain-specific scientific and AI articles, where the difference Gemma3-1B and Vault-
420 Gemma reaches 9.9 points (50.4 vs. 39.5) and 8.7 points (51.9 vs. 43.2), respectively. Distributions of FactScores
421 (Figure 2) show the same results: VaultGemma outputs have a higher density of low FactScores than Gemma3-1B.422 Compared to GPT-2 XL, VaultGemma does have higher average FactScores for AI and pre-training articles. As the
423 cutoff date for GPT-2 XL’s pre-training data was in 2019, this model had no exposure to most concepts in the AI
424 articles, thus constituting an extremely low bar for factual correctness in this setting. The improved FactScores of
425 VaultGemma over the Wikipedia pre-training may be a reflection of general improvements in LLM development over
426 the last 5 years that are not undone by DP training. More surprisingly, VaultGemma fails to output more factually
427 correct information than GPT-2 XL in the Wikipedia Science setting, even though these articles were all created after
428 2020, suggesting they were included in VaultGemma training data and not GPT-2 XL data.

Dataset	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
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 7: 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 (≥ 0.5) and non-factual (< 0.5) responses. Bolding indicates lower FactScore (increased hallucinations).

Overall, these results suggest DP-SGD pre-training can significantly 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.

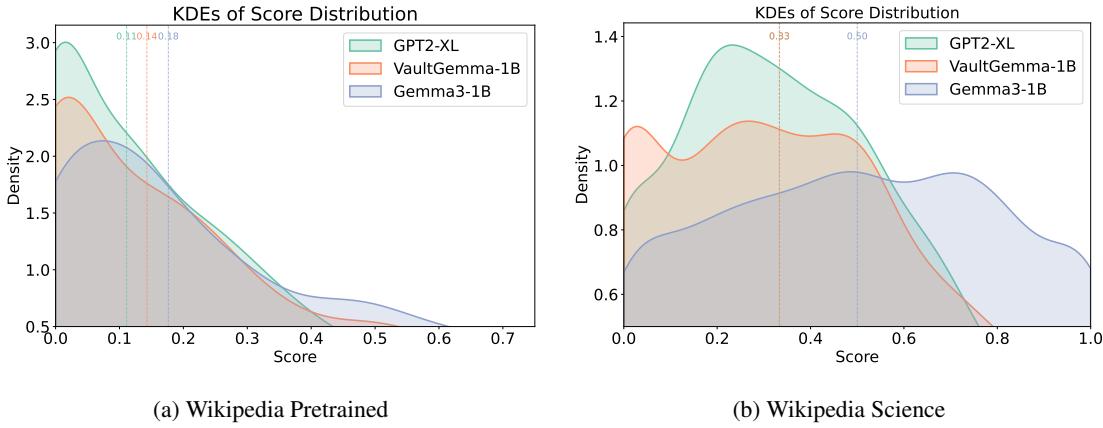


Figure 2: Kernel density (KDE) plots comparing FactScore distributions for models pre-trained with (VaultGemma $\epsilon = 2$) and without DP (Gemma3-1B & GPT2-XL). GPT2-XL, which has not seen the data during training, and VaultGemma (DP pre-trained) both lag behind Gemma-3B in factual consistency.

5 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. Further, we find the same unsupported facts repeated across generations, suggesting a systematic encoding of inaccurate facts rather than random errors. While our metrics focus on analyzing model outputs, they suggest the clipping and additive noise in DP may lead to the acquisition of incorrect facts; this could potentially be targeted directly in future work.

Our findings suggest that, in some applications, DP fine-tuning may simply be unsuitable. For example, a healthcare model trained without DP on public data, like medical journal articles, may be more reliable than one trained with DP on patient data, even if the non-DP model is unable to assist with some tasks or queries. Use of DP in these settings requires careful analysis of if the perceived benefits (e.g., acquisition of domain-specific knowledge) are worth the risk of increased hallucinations.

486 Alternatively, in settings where training on private data is essential, more research is needed to counteract hallucinations.
 487 While fully mitigating hallucinations is an unsolved problem even in non-DP settings, targeted interventions
 488 could reduce risks. Specialized post-training objectives could steer models towards expressing uncertainty or refusal
 489 over generating incorrect content (Kalai et al., 2025). A nascent line of work has proposed incorporating sensitive
 490 information through DP retrieval-augmented generation, rather than fine-tuning (Koga et al., 2024; Grislain, 2025),
 491 though maintaining a privacy guarantees under multiple queries remains challenging, and more research is needed to
 492 assess if hallucinations persist in these settings.

493 Overall, our findings also underscore that existing DP training methods can compromise factual reliability in high-
 494 stakes applications. This highlights the urgent need for refined privacy-preserving approaches that balance rigorous
 495 privacy guarantees with factual accuracy. Addressing this tradeoff is essential for deploying trustworthy AI systems in
 496 sensitive domains such as healthcare and law where both privacy and factual consistency are non-negotiable.

498 REFERENCES

- 500 Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep
 501 learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Commu-*
 502 *nications Security, CCS '16*, pp. 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN
 503 9781450341394. doi: 10.1145/2976749.2978318. URL <https://doi.org/10.1145/2976749.2978318>.
- 504 Michael Aerni, Jie Zhang, and Florian Tramèr. Evaluations of machine learning privacy defenses are misleading. In
 505 *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1271–1284,
 506 2024.
- 507 Michael Aerni, Javier Rando, Edoardo Debenedetti, Nicholas Carlini, Daphne Ippolito, and Florian Tramèr. Measuring
 508 non-adversarial reproduction of training data in large language models. In *The Thirteenth International Conference*
 509 *on Learning Representations*, 2025.
- 510 Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic
 511 Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj*
 512 *Digital Medicine*, 8(1):274, May 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01670-7.
- 513 Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. *Differential privacy has disparate impact on model*
 514 *accuracy*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- 515 Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in
 516 abstractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of*
 517 *the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3340–
 518 3354, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.236.
 519 URL <https://aclanthology.org/2022.acl-long.236/>.
- 520 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and test-
 521 ing unintended memorization in neural networks. In Nadia Heninger and Patrick Traynor (eds.), *28th USENIX*
 522 *Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14–16, 2019*, pp. 267–284. USENIX
 523 Association, 2019.
- 524 Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts,
 525 Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large
 526 language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX As-
 527 sociation, August 2021. ISBN 978-1-939133-24-3. URL <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>.
- 528 Jifan Chen, Eunsol Choi, and Greg Durrett. Can NLI models verify QA systems' predictions? In Marie-Francine
 529 Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Compu-*
 530 *tational Linguistics: EMNLP 2021*, pp. 3841–3854, Punta Cana, Dominican Republic, November 2021. Association
 531 for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.324. URL <https://aclanthology.org/2021.findings-emnlp.324/>.
- 532 Junjie Chu, Zeyang Sha, Michael Backes, and Yang Zhang. Reconstruct your previous conversations! comprehen-
 533 sively investigating privacy leakage risks in conversations with GPT models. In Yaser Al-Onaizan, Mohit Bansal,
 534 and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Pro-*
 535 *cessing*, pp. 6584–6600, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:
 536 10.18653/v1/2024.emnlp-main.377. URL <https://aclanthology.org/2024.emnlp-main.377>.

- 540 Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Everett,
 541 Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes,
 542 Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra,
 543 Mark A. Burbridge, James C. McAvoy, Roya Saffary, Stephen P. Ma, Dev Dash, James Xie, Ellen Y. Wang, Clif-
 544 ford A. Schmiesing, Nigam Shah, and Nima Aghaeepour. Verifact: Verifying facts in llm-generated clinical text
 545 with electronic health records. 2025. URL <https://arxiv.org/abs/2501.16672>.
- 546 Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski,
 547 Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. Advancing differential privacy: Where we
 548 are now and future directions for real-world deployment. *Harvard Data Science Review*, 2024.
- 549
- 550 Souvik Das, Sougata Saha, and Rohini Srihari. Diving deep into modes of fact hallucinations in dialogue systems.
 551 In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Lin-
 552 guistics: EMNLP 2022*, pp. 684–699, Abu Dhabi, United Arab Emirates, December 2022. Association for Compu-
 553 tational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.48. URL <https://aclanthology.org/2022.findings-emnlp.48/>.
- 554
- 555 Anderson Santana de Oliveira, Caelin Kaplan, Khawla Mallat, and Tanmay Chakraborty. An empirical analysis of
 556 fairness notions under differential privacy, 2024. URL <https://arxiv.org/abs/2302.02910>.
- 557
- 558 Vadym Doroshenko, Badih Ghazi, Prithish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter
 559 discrete approximations of privacy loss distributions. *Proceedings on Privacy Enhancing Technologies*, 4:552–570,
 560 2022.
- 561
- 562 Jiacheng Du, Zhibo Wang, and Kui Ren. Textual unlearning gives a false sense of unlearning. *CoRR*, abs/2406.13348,
 563 2024.
- 564
- 565 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data
 566 analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006.
 567 Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.
- 568
- 569 Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in
 Theoretical Computer Science*, 9(3–4):211–407, 2014.
- 570
- 571 Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual
 572 consistency evaluation for summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir
 573 Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Com-
 574 putational Linguistics: Human Language Technologies*, pp. 2587–2601, Seattle, United States, July 2022. Associa-
 575 tion for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.187. URL <https://aclanthology.org/2022.naacl-main.187/>.
- 576
- 577 Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. Neither private nor fair: Impact of
 578 data imbalance on utility and fairness in differential privacy, 2020. URL <https://arxiv.org/abs/2009.06389>.
- 579
- 580 Ferdinando Fioretto, Pascal Van Hentenryck, et al. Differential Privacy in Artificial Intelligence: From Theory to
 581 Practice. 2025.
- 582
- 583 Nicolas Grislain. Rag with differential privacy. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pp. 847–852.
 584 IEEE, 2025.
- 585
- 586 Abhimanyu Hans, John Kirchenbauer, Yuxin Wen, Neel Jain, Hamid Kazemi, Prajwal Singhania, Siddharth Singh,
 587 Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. Be like a goldfish, don’t memorize!
 588 mitigating memorization in generative LLMs. In *The Thirty-eighth Annual Conference on Neural Information
 Processing Systems*, 2024. URL <https://openreview.net/forum?id=DylSyAfmWs>.
- 589
- 590 Victor Hansen, Atula Neerkaje, Ramit Sawhney, Lucie Flek, and Anders Søgaard. The impact of differential privacy
 591 on group disparity mitigation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association
 592 for Computational Linguistics: NAACL 2024*, pp. 3952–3965, Mexico City, Mexico, June 2024. Association for
 593 Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.249. URL <https://aclanthology.org/2024.findings-naacl.249/>.

- 594 Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.
 595 LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*,
 596 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- 597 Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. Differentially private natural language models: Recent advances
 598 and future directions. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational*
 599 *Linguistics: EACL 2024*, pp. 478–499, St. Julian’s, Malta, March 2024. Association for Computational Linguistics.
 600 URL <https://aclanthology.org/2024.findings-eacl.33/>.
- 601 Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. DP-rewrite: Towards reproducibility and transparency in
 602 differentially private text rewriting. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo
 603 Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio,
 604 Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond,
 605 and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp.
 606 2927–2933, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
 607 URL <https://aclanthology.org/2022.coling-1.258/>.
- 608 Shotaro Ishihara. Training data extraction from pre-trained language models: A survey. In Anaelia Ovalle, Kai-Wei
 609 Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop
 610 Kumar, and Rahul Gupta (eds.), *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing*
 611 (*TrustNLP 2023*), pp. 260–275, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.
 612 18653/v1/2023.trustnlp-1.23. URL <https://aclanthology.org/2023.trustnlp-1.23/>.
- 613 Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo.
 614 Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and
 615 Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
 616 (*Volume 1: Long Papers*), pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics.
 617 doi: 10.18653/v1/2023.acl-long.805. URL <https://aclanthology.org/2023.acl-long.805/>.
- 618 Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination
 619 via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL <https://aclanthology.org/2023.findings-emnlp.123/>.
- 620 Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou.
 621 On large language models’ hallucination with regard to known facts. In Kevin Duh, Helena Gomez, and Steven
 622 Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1041–1053, Mexico City,
 623 Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.60. URL
 624 <https://aclanthology.org/2024.naacl-long.60/>.
- 625 Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi
 626 Liu, and Benjamin Van Durme. Core: Robust factual precision with informative sub-claim identification, 2024b.
 627 URL <https://arxiv.org/abs/2407.03572>.
- 628 Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.
- 629 Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, and Zheng Xu. User
 630 inference attacks on large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 18238–18265. Association for Computational Linguistics, 2024.
- 631 Aly Kassem, Omar Mahmoud, and Sherif Saad. Preserving privacy through dememorization: An unlearning technique
 632 for mitigating memorization risks in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.),
 633 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4360–4379, Singapore,
 634 December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.265. URL
 635 [https://aclanthology.org/2023.emnlp-main.265/](https://aclanthology.org/2023.emnlp-main.265).
- 636 Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 193–204, 2011.

- 648 Tatsuki Koga, Ruihan Wu, and Kamalika Chaudhuri. Privacy-preserving retrieval-augmented generation with differ-
 649 ential privacy. *arXiv preprint arXiv:2412.04697*, 2024.
- 650
- 651 Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models
 652 for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:
 653 163–177, 2022. doi: 10.1162/tacl_a.00453. URL <https://aclanthology.org/2022.tacl-1.10/>.
- 654 Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark:
 655 An empirical study on factuality hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek
 656 Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume
 657 1: Long Papers)*, pp. 10879–10899, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
 658 doi: 10.18653/v1/2024.acl-long.586. URL <https://aclanthology.org/2024.acl-long.586/>.
- 659 Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong dif-
 660 ferentially private learners. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=bVuP3ltATMz>.
- 661
- 662
- 663 Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. Scaling laws for fact
 664 memorization of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Find-
 665 ings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11263–11282, Miami, Florida, USA,
 666 November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.658. URL
 667 <https://aclanthology.org/2024.findings-emnlp.658/>.
- 668 Lingjuan Lyu, Xuanli He, and Yitong Li. Differentially private representation for NLP: Formal guarantee and an em-
 669 pirical study on privacy and fairness. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association
 670 for Computational Linguistics: EMNLP 2020*, pp. 2355–2365, Online, November 2020. Association for Computa-
 671 tional Linguistics. doi: 10.18653/v1/2020.findings-emnlp.213. URL [https://aclanthology.org/2020.findings-emnlp.213/](https://aclanthology.org/2020.findings-emnlp.213).
- 672
- 673 Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. Differentially
 674 private language models for secure data sharing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.),
 675 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4860–4873, Abu
 676 Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.
 677 emnlp-main.323. URL <https://aclanthology.org/2022.emnlp-main.323/>.
- 678
- 679 Cleo Matzen, Steffen Eger, and Ivan Habernal. Trade-offs between fairness and privacy in language modeling. In
 680 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Lin-
 681 guistics: ACL 2023*, pp. 6948–6969, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:
 682 10.18653/v1/2023.findings-acl.434. URL [https://aclanthology.org/2023.findings-acl.434/](https://aclanthology.org/2023.findings-acl.434).
- 683 H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language
 684 models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April
 685 30 - May 3, 2018, Conference Track Proceedings*, 2018.
- 686
- 687 Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book?
 688 document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX
 689 Security 24)*, pp. 2369–2385, 2024.
- 690 Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data
 691 frequency in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
 692 URL <https://openreview.net/forum?id=EDoD3DgivF>.
- 693
- 694 Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and
 695 Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text genera-
 696 tion. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical
 697 Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Compu-
 698 tational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL [https://aclanthology.org/2023.emnlp-main.741/](https://aclanthology.org/2023.emnlp-main.741).
- 699
- 700 Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, Fabio Massimo Zanzotto, Sébastien Bratières, and Emanuele
 701 Rodolà. Preserving privacy in large language models: A survey on current threats and solutions, 2025. URL
<https://arxiv.org/abs/2408.05212>.

- 702 Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert
 703 Sim. Privacy regularization: Joint privacy-utility optimization in LanguageModels. In Kristina Toutanova,
 704 Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy
 705 Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of
 706 the Association for Computational Linguistics: Human Language Technologies*, pp. 3799–3807, Online, June
 707 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.298. URL <https://aclanthology.org/2021.naacl-main.298/>.
- 709 Ivoline C. Ngong, Joseph Near, and Niloofar Mireshghallah. Differentially private learning needs better model initial-
 710 ization and self-distillation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference
 711 of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Tech-
 712 nologies (Volume 1: Long Papers)*, pp. 9009–9027, Albuquerque, New Mexico, April 2025. Association for Compu-
 713 tational Linguistics. ISBN 979-8-89176-189-6. URL <https://aclanthology.org/2025.naacl-long.455/>.
- 714 Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, Minhui Xue, and Bo Li. Reconstruction of dif-
 715 ferentially private text sanitization via large language models, 2024. URL <https://arxiv.org/abs/2410.12443>.
- 716 Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander
 717 Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan
 718 (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the
 719 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong
 720 Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL
 721 <https://aclanthology.org/D19-1250/>.
- 722 Krishna Pillutla, Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. Federated learning with superquantile aggre-
 723 gation for heterogeneous data. *Machine Learning*, 113(5):2955–3022, 2024.
- 724 Krishna Pillutla, Jalaj Upadhyay, Christopher A Choquette-Choo, Krishnamurthy Dvijotham, Arun Ganesh, Monika
 725 Henzinger, Jonathan Katz, Ryan McKenna, H Brendan McMahan, Keith Rush, et al. Correlated Noise Mechanisms
 726 for Differentially Private Learning. *arXiv preprint arXiv:2506.08201*, 2025.
- 727 Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei
 728 Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ML: A practical guide to machine learning
 729 with differential privacy. *J. Artif. Intell. Res.*, 77:1113–1201, 2023. doi: 10.1613/JAIR.1.14649. URL <https://doi.org/10.1613/jair.1.14649>.
- 730 Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. Evaluating differen-
 731 tially private synthetic data generation in high-stakes domains. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung
 732 Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15254–15269, Miami,
 733 Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.
 734 894. URL <https://aclanthology.org/2024.findings-emnlp.894/>.
- 735 Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form
 736 text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for
 737 Computational Linguistics: EMNLP 2024*, pp. 9447–9474, Miami, Florida, USA, November 2024. Association
 738 for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL <https://aclanthology.org/2024.findings-emnlp.552/>.
- 739 Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference
 740 with large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- 741 Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models
 742 for factuality, 2023. URL <https://arxiv.org/abs/2311.08401>.
- 743 Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning
 744 with large-scale public pretraining. In *Forty-first International Conference on Machine Learning, ICML 2024,
 745 Vienna, Austria, July 21-27, 2024*, 2024.
- 746 Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy
 747 under a fairness lens. In *IJCAI*, pp. 560–566, 2021.

- 756 VaultGemma Team. VaultGemma: A differentially private Gemma model, 2025. URL https://services.google.com/fh/files/blogs/vaultgemma_tech_report.pdf. Google Technical Report.
- 757
- 758
- 759 Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- 760
- 761 Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19519–19529, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1088. URL <https://aclanthology.org/2024.emnlp-main.1088/>.
- 762
- 763
- 764
- 765
- 766 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. Long-form factuality in large language models, 2024. URL <https://arxiv.org/abs/2403.18802>.
- 767
- 768
- 769
- 770 Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2875–2886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.174. URL <https://aclanthology.org/2023.emnlp-main.174/>.
- 771
- 772
- 773
- 774
- 775
- 776 Xinwei Wu, Weilong Dong, Shaoyang Xu, and Deyi Xiong. Mitigating privacy seesaw in large language models: Augmented privacy neuron editing via activation patching. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5319–5332, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.315. URL <https://aclanthology.org/2024.findings-acl.315/>.
- 777
- 778
- 779
- 780
- 781 Rui Xin, Niloofer Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Se-woong Oh, and Pang Wei Koh. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL <https://openreview.net/forum?id=3JLtuCozOU>.
- 782
- 783
- 784
- 785
- 786 Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan McMahan, Jesse Rosenstock, and Yuanbo Zhang. Federated learning of gboard language models with differential privacy. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 629–639, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.60. URL <https://aclanthology.org/2023.acl-industry.60/>.
- 787
- 788
- 789
- 790
- 791
- 792 Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. In *2024 International Conference on Meta Computing (ICMC)*, pp. 1–12, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/ICMC60390.2024.00008. URL <https://doi.ieeecomputersociety.org/10.1109/ICMC60390.2024.00008>.
- 793
- 794
- 795
- 796 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. Opacus: User-friendly differential privacy library in pytorch, 2022.
- 797
- 798
- 799
- 800 Paul Youssef, Osman Koraş, Meijie Li, Jörg Schütterer, and Christin Seifert. Give me the facts! a survey on factual knowledge probing in pre-trained language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15588–15605, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1043. URL <https://aclanthology.org/2023.findings-emnlp.1043/>.
- 801
- 802
- 803
- 804
- 805 Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1321–1342, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.74. URL [https://aclanthology.org/2023.acl-long.74/](https://aclanthology.org/2023.acl-long.74).
- 806
- 807
- 808
- 809

810 Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to
 811 effective unlearning, 2024. URL <https://arxiv.org/abs/2404.05868>.

812
 813 Xinyu Zhou and Raef Bassily. Differentially private worst-group risk minimization. *arXiv preprint arXiv:2402.19437*,
 814 2024.

815 A BACKGROUND: DIFFERENTIAL PRIVACY

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

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

823 Formally, differential privacy is defined as follows:

824 **Definition:** Two datasets D and D' are said to be neighboring (in the add-or-remove sense) if $D = D' \cup \{x\}$ or
 825 $D' = D \cup \{x\}$ for some training example x . A randomized algorithm A is (ε, δ) -private for some $\varepsilon > 0$ and
 826 $\delta \in [0, 1]$ if for any two neighboring datasets D, D' , the following holds true for all sets Y in the range of A :

$$827 \Pr[A(D) \in Y] \leq e^\varepsilon \Pr[A(D') \in Y] + \delta.$$

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

834 We use DP-SGD (Abadi et al., 2016), a modification to the stochastic gradient descent (SGD) algorithm, which is
 835 typically used to train neural networks. DP-SGD clips the gradients to limit the contribution of individual samples
 836 from the training data and subsequently adds noise from the Gaussian distribution to the sum of the clipped gradients
 837 across all samples.⁴ DP-SGD thus provides a differentially private guarantee to obfuscate the gradient update, thereby
 838 ensuring that the contribution of any given sample in the training data is indistinguishable due to the aforementioned
 839 post-processing property. This process ensures (ε, δ) -differential privacy for each model update. Given a privacy
 840 budget, number of epochs, and other training parameters, we can estimate the privacy parameters using standard
 841 privacy accounting algorithms, which implemented in common software.

842 B EXPERIMENTAL SETUP

843 We conduct both standard (non-private) fine-tuning and differentially private (DP) fine-tuning, and the noise multiplier
 844 in the DP setting is calculated using the PLD accountant. We report the results for two privacy budgets: $\varepsilon = 8$ and
 845 $\varepsilon = 16$, and the δ is set to $1/n^{1.1}$, where n is the size of the training set. We use the PLD accounting algorithm
 846 proposed in Doroshenko et al. (2022), which provides us with tighter estimates of the privacy loss as compared to
 847 alternate accounting techniques. This in turn allows us to more accurately determine the noise multiplier required to
 848 satisfy the specified privacy budget for our fine-tuning setups.

849 The hyperparameters are summarized as below.

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

$$852 \quad W = W_0 + \frac{\alpha}{r} AB$$

853 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
 854 LoRA parameters A, B are updated via:

$$855 \quad (A^{(t+1)}, B^{(t+1)}) = (A^{(t)}, B^{(t)}) - \eta \cdot \text{clip}(\nabla_{A,B} L, c)$$

856
 857 ⁴The noise added is independent coordinate-wise and across time, although DP-SGD with temporal correlations has recently
 858 been of growing interest (Pillutla et al., 2025).

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, $\varepsilon \in \{8, 16\}$	4096	20	1×10^{-4}	1.0	4	512	512	$\delta = 1/n^{1.1}$

Table 8: Summary of major hyperparameters in vanilla and private training settings.

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.

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

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

where

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

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

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.

C DATASET : EXAMPLES OF INPUT

Table 9: 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.

918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971	Input	Output
	Cyclosiloxanes	Cyclosiloxanes are a class of silicone material. They are volatile and often used as a solvent. The three main commercial varieties 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 (CH ₃). == 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 (CH ₃). 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.
	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."

972 973 974 975 976 977 978 979 980 981 982	973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992	973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992
	Input	Output
	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".
	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).

D LIST OF TOPICS IN THE DATASETS

D.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 stodio; 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

D.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 anti-infective 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;

1026 Physics of Life; Quasi-isodynamic stellarator; Riccardo D'Auria (theoretical physicist); Shockwave cosmology; Syn-
 1027 chronous lateral excitation; Toroidal solenoid; Wohlfarth Lectureship; Compliance constants; Cononsolvency; Corro-
 1028 sion inhibitors for the petroleum industry; Cyclosiloxane; Dark oxygen; Direct reduction; Energy-rich species; Grupo
 1029 Fertiberia; Intrinsic DNA fluorescence; Krupp–Renn process; Mental gland; Probico; School of Molecular Sciences;
 1030 Shape of the atomic nucleus; Stable phosphorus radicals; Superelectrophilic anion; TOP Assay; Mathematical oncol-
 1031 ogy; Mathethon; The Math(s) Fix; Conductivity cell; Generalized renewal process; Glossary of engineering: M–Z;
 1032 Marine construction; Museum of Engines and Mechanisms; Northern Technical College; Safer end of engineering
 1033 life; Synchronous lateral excitation; The Clark Collection of Mechanical Movements; Third medium contact method;
 1034 UNESCO World Engineering Day for Sustainable Development; Positive health; Bell's mania; Chialvo map; Dysfunc-
 1035 tome; Femoral nerve dysfunction; Fiber photometry; Fork cell; High Price (book); Hyper-empathy; Large dense core
 1036 vesicles; Lateral olfactory tract usher substance; Malaria therapy; Max Planck Institute for Biological Intelligence;
 1037 Nerve glide; Neural synchrony; Neurosemiotics; Neurotrophin mimetics; Optogenetic methods to record cellular
 1038 activity; Personality neuroscience; Representational drift; Single-particle trajectory; Smell training; Spongy degen-
 1039 eration of the central nervous system; Walk Again Project; Amoeboflagellate; Borg (microbiology); Chromopodellid;
 1040 Dissimilatory iron reducing bacteria; Garrod Lecture and Medal; Hydrocarbonoclastic bacteria; Laboratory-acquired
 1041 infection; Matground; Microbial pathogenesis; Milnesium alpigenum; Mitochondrion-related organelle; Phageome;
 1042 Phytoplankton microbiome; Virivore; Virome analysis; Zodletone Mountain; Glossary of cellular and molecular biol-
 1043 ogy (M–Z); Agricultural weed syndrome; Cell autonomous sex identity; Codon reassignment; De novo domestication;
 1044 Endemixit; Genetic map function; Hovlinc; Integrative and conjugative element; Jena Declaration; Macrosatellite;
 1045 Museomics; Poison exon; Polydactyly-myopia syndrome; Red cell genotyping; Right To Know; Selection limits;
 1046 Shadow effect; Transcriptome-wide association study; Tumor mutational burden; Allogeneic processed thymus tis-
 1047 sue; Cellular anastasis; COVID-19 passports in the United Kingdom; History of phagocytosis; Immunocapitalism;
 1048 Macrophage-activating lipopeptide 2; Metal allergy; Milk immunity; Myocarditis-myositis-myasthenia gravis overlap
 1049 syndrome; Oligoclonal antibody; P-i mechanism; Pathogen avoidance; Peripheral ulcerative keratitis; Post-acute in-
 1050 fection syndrome; RVT-802; T memory stem cell; Thymic mimetic cells; TMEM61; Type 2 inflammation; Vaccine
 1051 passports during the COVID-19 pandemic; Vaccine resistance; Zigakibart; 2022–2023 pediatric care crisis; Acoustic
 1052 epidemiology; Causal pie model; Connecting Organizations for Regional Disease Surveillance; Elimination of tuber-
 1053 culosis; Epidemics Act; Epidemiology in Relation to Air Travel; Epidemiology of gonorrhoea; European Society of
 1054 Health and Medical Sociology; Harvard Six Cities study; Hyperendemic; Loneliness epidemic; Microbial pathogen-
 1055 esis; Origin tracing; Pathogenic microorganisms in frozen environments; SARS-CoV-2 in white-tailed deer; Source
 1056 attribution; Sporadic disease; Outline of public health; Alcohol tax; Autobesity; Biomedical Research Center; CalOpti-
 1057 ma; Care Group approach; Christian Health Association of Malawi; Commercial determinants of health; Connecting
 1058 Organizations for Regional Disease Surveillance; COVID-19 lockdowns by country; Epidemics Act; History of public
 1059 health in Australia; History of public health in Canada; History of public health in Chicago; History of public
 1060 health in New York City; History of public health in the United Kingdom; History of public health in the United
 1061 States; Intermittent water supply; International Association for Cannabinoid Medicines; Langya virus; LGBT life
 1062 expectancy; User:Lguzmang06/sandbox; Loneliness epidemic; Malawi Network of AIDS Services; Mass. and Cass;
 1063 Medical officer of environmental health; Motonormativity; National Association for People living with HIV/AIDS in
 1064 Malawi; North Karelia Project; Nuisance ordinance; Origin tracing; Preventive and social medicine; Responsibility
 1065 Deal; SaTScan; Sleeping Sickness Commission; Slug gate; Social determinants of mental health; Special Programme
 1066 of Research, Development and Research Training in Human Reproduction- HRP; Telemedicine in Nepal; Vaccine eq-
 1067 uity; Vaccine line jumping; Vaccine storage; WHO Hub for Pandemic and Epidemic Intelligence; WHO public health
 1068 prizes and awards; Additive effect; Antica Farmacia Sant'Anna; FK962; Institute for Safe Medication Practices;
 1069 Model-Informed Precision Dosing; P-i mechanism; Penetration enhancer; Pharmacological cardiotoxicity; Pullulan
 1070 bioconjugate; Reversible Hill equation

1071

1072

1073

1074

1075

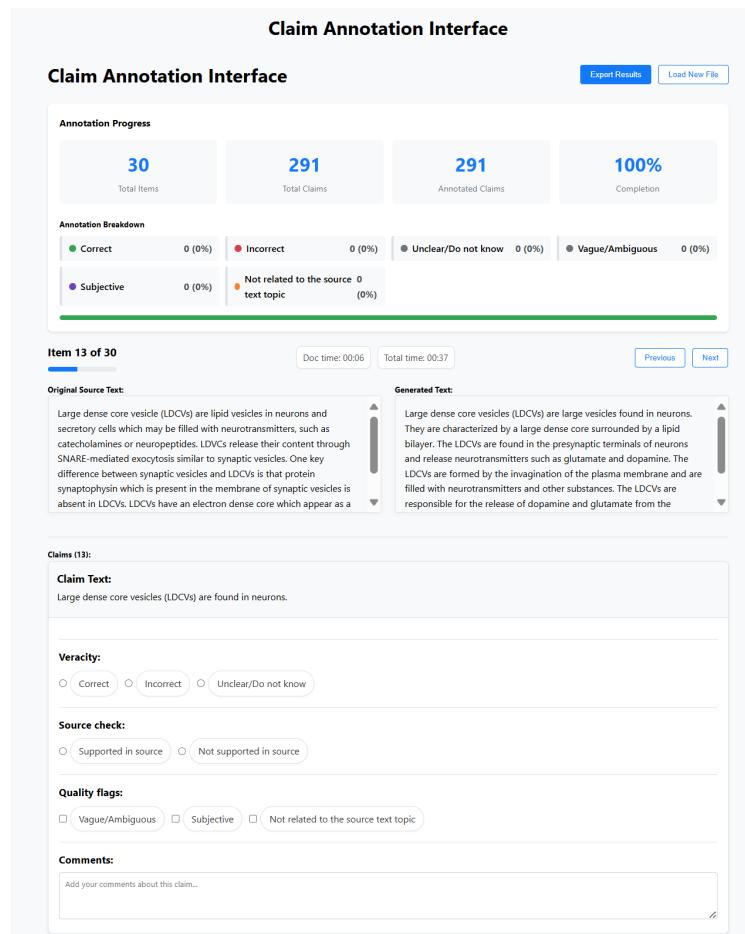
1076

1077

1078

1079

1080 **E MANUAL ANNOTATIONS: CLAIM ANNOTATION INTERFACE**
 1081
 1082
 1083
 1084



1114 Figure 3: The interactive claim annotation tool used by participants to evaluate the factual correctness of claims from
 1115 the model-generated text.

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

1116
 1117
 1118
 1119 Table 10: Cohen’s Kappa scores for DP-inf and DP-16 settings for human annotation of Wikipedia AI article claims.

1130 **F FACTSCORE AND RECURRING CLAIMS ALGORITHMS**
 1131
 1132

1133 We recall the pseudo-code of FactScore in Algorithm 1 and describe our algorithm to cluster repeated claims in
 Algorithm 2 (cf. §3.4).

1134
1135 **G FACTSCORE ALGORITHM**
1136
1137
1138
11391140 **Algorithm 1** FActScore: Atomic Fact Extraction and Verification

1141 **Input:** Generated texts $\mathcal{D} = d_1, d_2 \dots d_n$, atomic fact extractor module \mathcal{E} , claim verification model \mathcal{V} , knowledge
1142 source \mathcal{K} 1143 **Output:** FactScore for each document in the generated corpus : $\mathcal{S}(\mathcal{E}, \mathcal{D})$ 1144 1: **for** each document $d_i \in \mathcal{D}$ **do**

1145 2: Extract a candidate set of atomic claims:

1146
$$\mathcal{AF}_{d_i} = \mathcal{E}(d_i)$$

1147 3: **for** each atomic claim $\alpha_j^{(d_i)} \in \mathcal{AF}_{d_i}$ **do**1148 4: Verify factuality if $\alpha_j^{(d_i)}$ is supported by knowledge source \mathcal{K} :

1149
$$\hat{y}_j^{(d_i)} = \mathcal{V}(\alpha_j^{(d_i)}, \mathcal{K}) \quad \text{where } \hat{y}_j^{(d_i)} \in \{0, 1\}$$

1150 5: **end for**

1151 6: Compute per-document precision:

1152
$$\mathcal{S}(\mathcal{E}, g) = \frac{1}{|\mathcal{AF}_{d_i}|} \sum_{j=1} \mathbb{I}(\hat{y}_j^{(d_i)} = 1)$$

1153 7: **end for**

1160
1161
1162
1163 Algorithm 1 describes how the FactScore is computed for a set of generated documents. For each document d_i , the
1164 atomic fact extractor \mathcal{E} which is an instruction-tuned LLM, is prompted with in-context examples to decompose the
1165 text into a set of atomic claims, denoted \mathcal{AF}_{d_i} .
11661167 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
1168 source is the reference data against which claims are verified, and in our setup, this consists of the relevant Wikipedia
1169 articles for the evaluation domain. For each atomic claim, evidence passages are retrieved (e.g. via BM25) from these
1170 articles which are then provided to the claim verification model. The claim verification model (\mathcal{V}) then judges whether
1171 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
1172 claim is factually supported.1173 The FactScore for the document is computed as follows:
1174
1175
1176

1177
$$\mathcal{S}(\mathcal{E}, d_i) = \frac{1}{|\mathcal{AF}_{d_i}|} \sum_j \mathbb{I}(\hat{y}_j^{(d_i)} = 1),$$

1178
1179

1180
1181
1182 which yields a scalar score corresponding to the factual correctness of the information in the generated document with
1183 respect to the knowledge source. A higher score translates to a greater proportion of verified claims, and conversely, a
1184 lower score is indicative of fewer supported claims.1185 We use Llama-3.1-8B-Instruct, to perform both the atomic fact extraction as well as the atomic claim verification. To
1186 demonstrate that our results remain consistent across different choices in the claim decomposition and claim verifica-
1187 tion 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 11 and Table 12.

1188 H CLUSTERING ALGORITHM DESCRIPTION
11891190
1191 **Algorithm 2** Recurring Claim Cluster Algorithm

1192 **Require:** Set of topics \mathcal{T} , where each $t \in \mathcal{T}$ has \mathcal{S} corresponding generated documents about the topic. $C = \{C_t\}$:
1193 Claims per topic $t \in \mathcal{T}$, where each $C_t = \{c_i\}$ with atomic facts (supported or unsupported).

1194 1: $\mathcal{I} \leftarrow \text{INDEXCLAIMSBYTEXT}(C, \mathcal{I})$ ▷ Index Claims to their Source Document
1195 2: $\mathcal{K} \leftarrow \text{CLUSTERASSIGNMENTOFCLAIMS}(\mathcal{I})$ ▷ Sentence Embedding-based Agglomerative Clustering
1196 3: $\mathcal{K} \leftarrow \text{DBSCANWITHJACCARDCLUSTERING}(\mathcal{K})$ ▷ DBSCAN Clustering over clusters to ensure their Jaccard
1197 Distance is low

1198 4: **for** $t \in \mathcal{T}$ **do**
1199 5: $\mathcal{K}'[t] \leftarrow \{\}$
1200 6: **end for** ▷ Initialize \mathcal{K}' to contain clusters of recurring claims

1201 7: **for** topic $t \in \mathcal{T}$ **do**
1202 8: **for** Cluster $k \in \mathcal{K}(\sqcup)$ **do**
1203 9: **if** COUNT(\mathcal{S}) for any $c_i \in k \geq 2$ **then**
1204 10: Append k to $\mathcal{K}'[t]$ ▷ Append a cluster of claims if the claims contain at least two supporting
1205 documents
1206 11: **end if**
1207 12: **end for**
1208 13: **end for**

1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

1242 **I FACTSCORE EVALUATIONS WITH DIFFERENT CLAIM DECOMPOSITION AND EVALUATION**
 1243 **MODELS**

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	inf	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	inf	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	inf	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	inf	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

1262 Table 11: Factuality evaluation scores for different temperature settings for the Wikipedia Science articles, using Meta
 1263 Llama 3.2-3b-Instruct for claim evaluation and decomposition.

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	inf	48.6	50	28.9	66.7	71.4	27.2	121	765
	16	47.5	45.5	25	66.7	70.9	24.2	87	788
	8	46.9	44.4	25	66.7	69	26	73	809
0.5	inf	48.9	50	30	66.7	68.7	29.1	129	759
	16	44.6	44.4	25	62.5	70.8	18.8	89	789
	8	46	50	25	66.7	65	27.3	96	784
0.7	inf	44.4	27.3	63.6	63.8	27.4	136	752	452
	16	44.4	27.3	62.5	68	23.1	115	766	502
	8	44.4	25	66.7	66	25.9	102	785	522
1.0	inf	42.9	25	62.5	69.2	22.6	147	738	446
	16	42.9	25	60	62.7	26.1	101	783	464
	8	42.9	27.3	62.5	63	25.5	114	772	505

1282 Table 12: Factuality evaluation scores for different temperature settings for the Wikipedia Science articles, using
 1283 DeepSeek-R1-Distill-Qwen-7B for claim evaluation and decomposition.

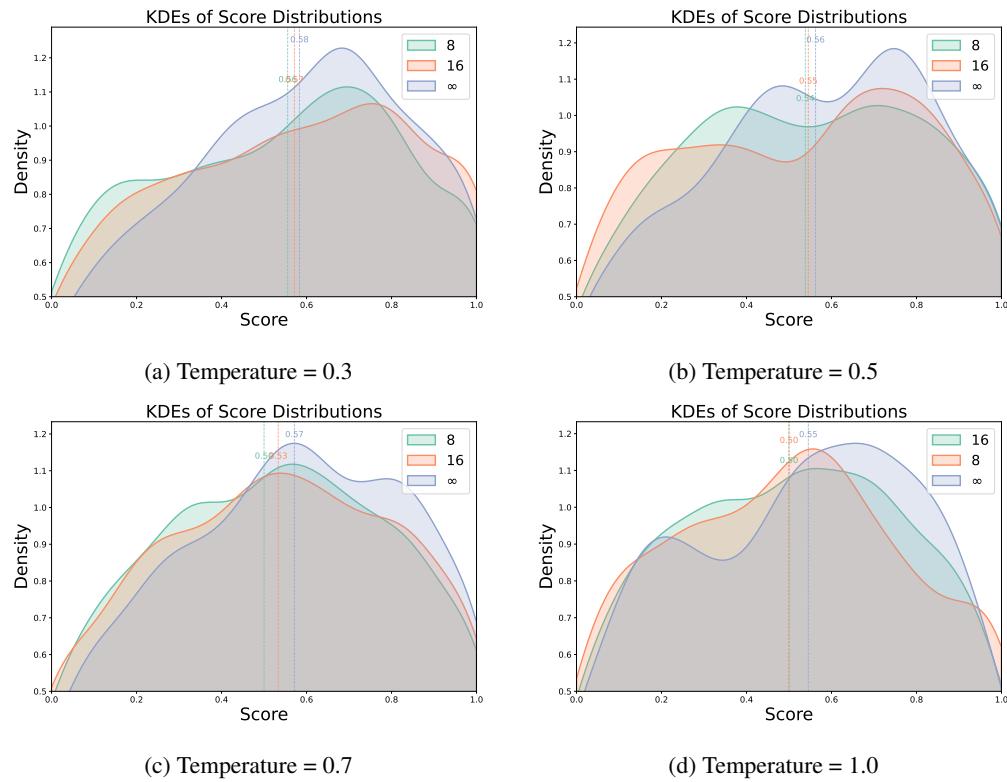
1296 **J FACTSCORE DISTRIBUTIONS PER TEMPERATURE SETTING UNDER DIFFERENT MODELS**
1297
1298
12991310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

Figure 4: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Science data for under different temperature settings for GPT-J 6B.

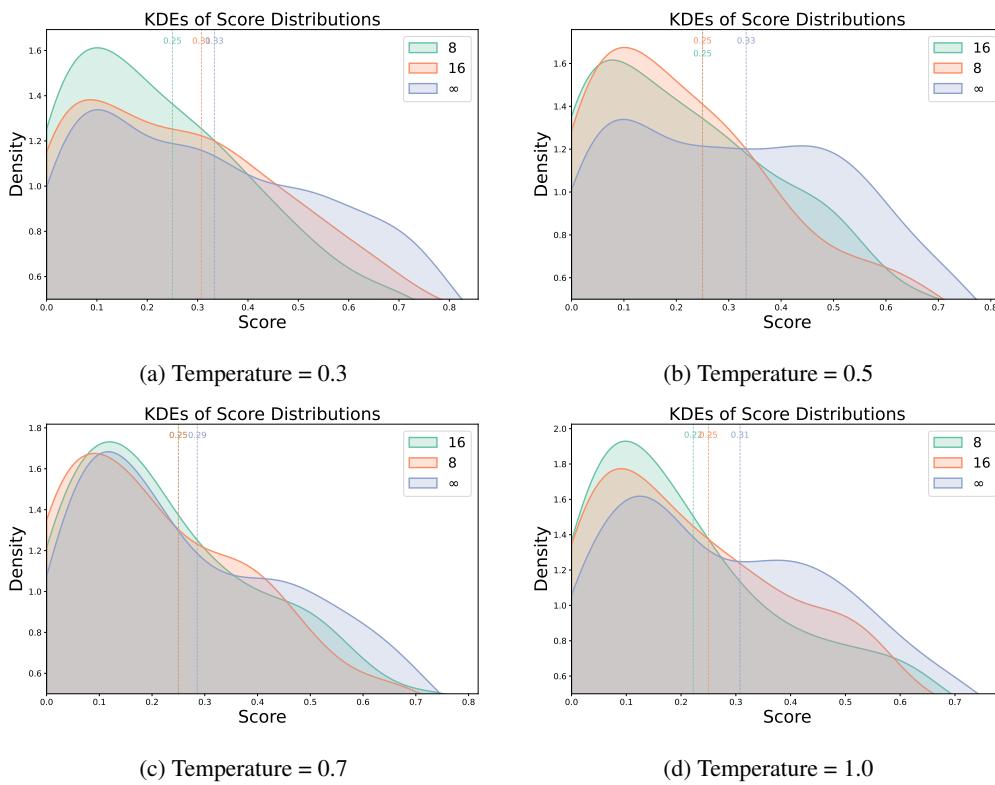


Figure 5: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia AI data for under different temperature settings for GPT-J 6B.

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417

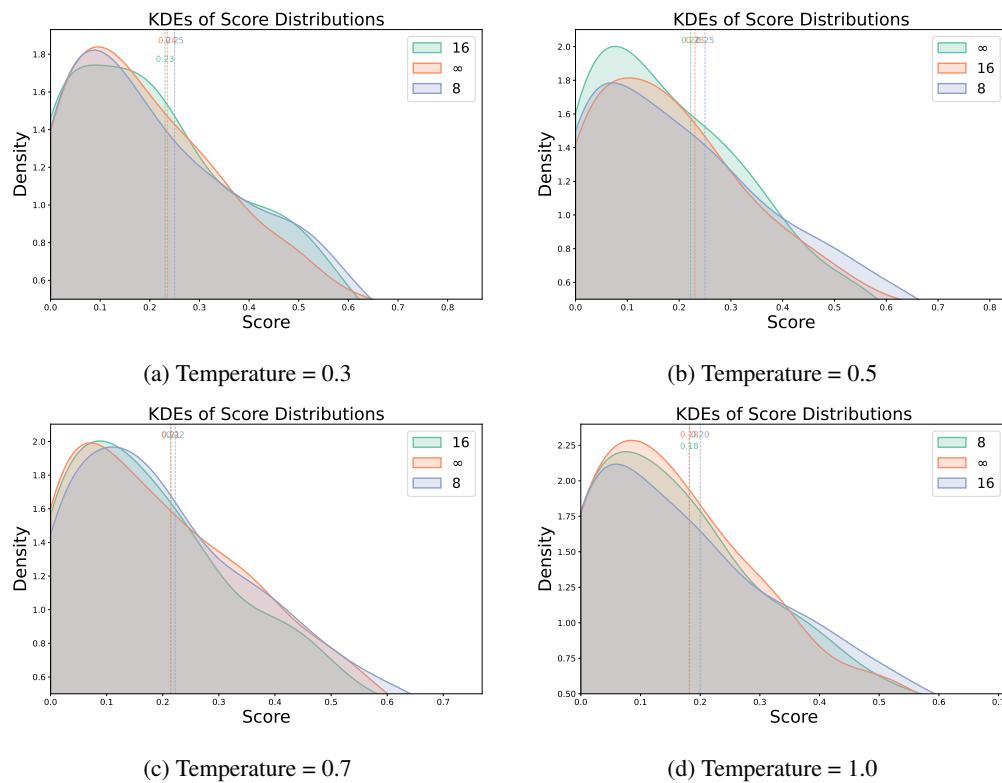
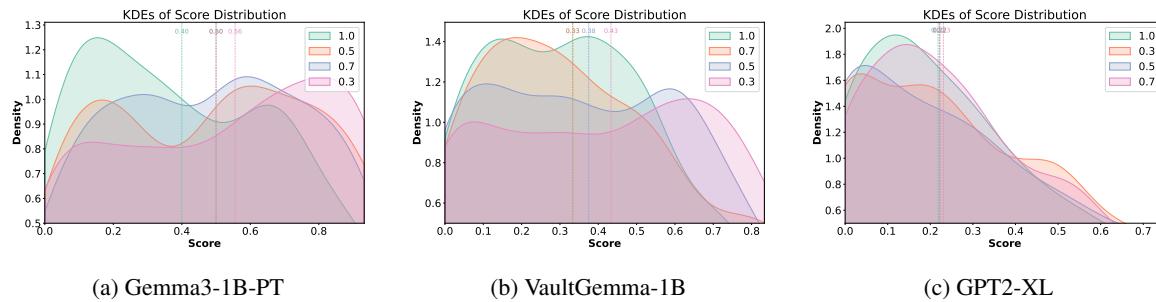


Figure 6: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Pretraining data for under different temperature settings for GPT-J 6B.

1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

1458 K FACTSCORE DISTRIBUTIONS PER MODEL UNDER DIFFERENT TEMPERATURE SETTINGS
14591460 K.1 MODELS PRE-TRAINED WITH AND WITHOUT DP
14611462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

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

Figure 7: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia AI data for the pre-trained models.

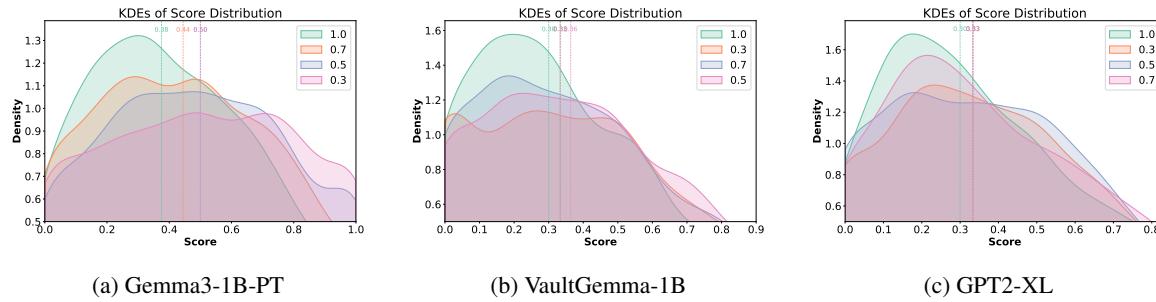


Figure 8: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Science data for the pre-trained models.

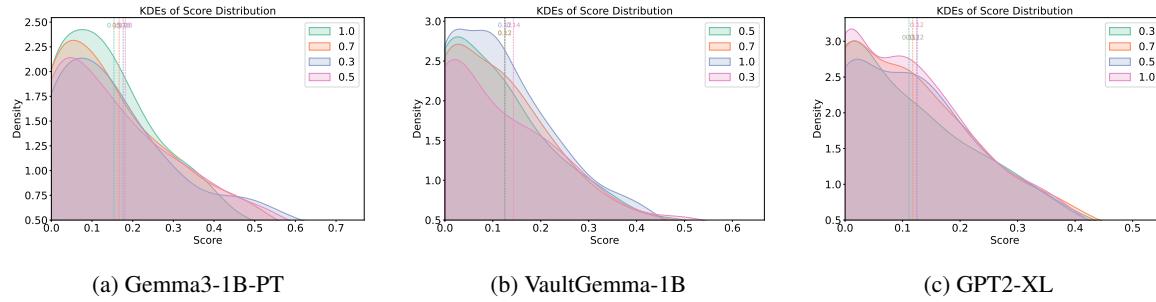
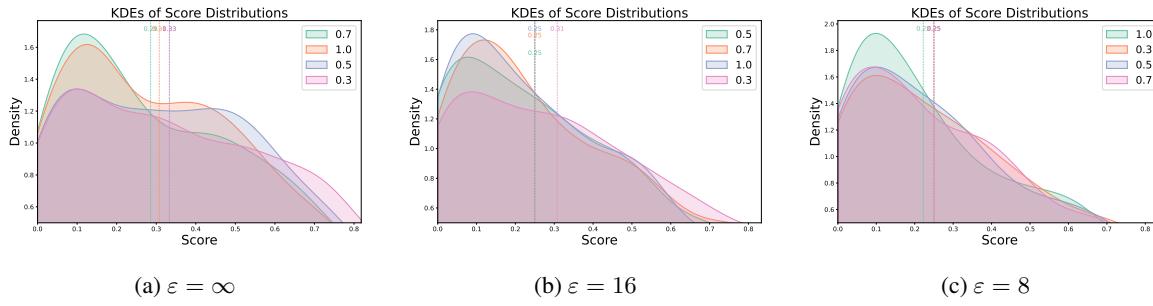
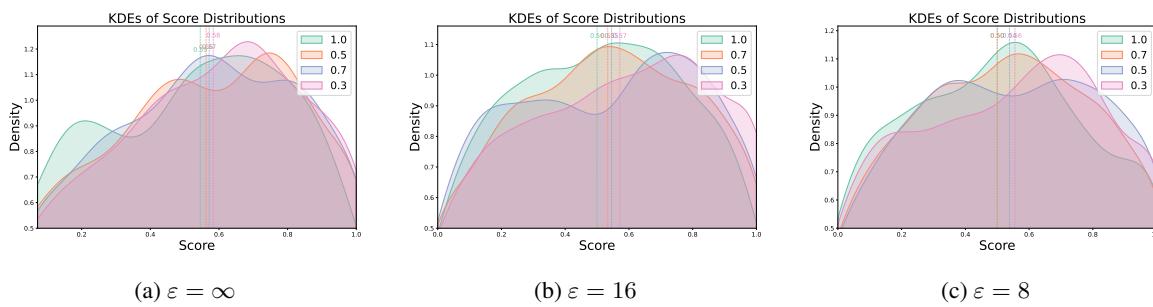
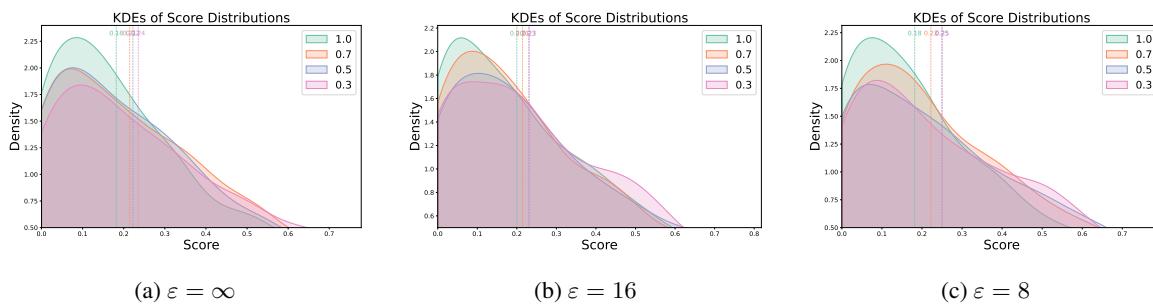
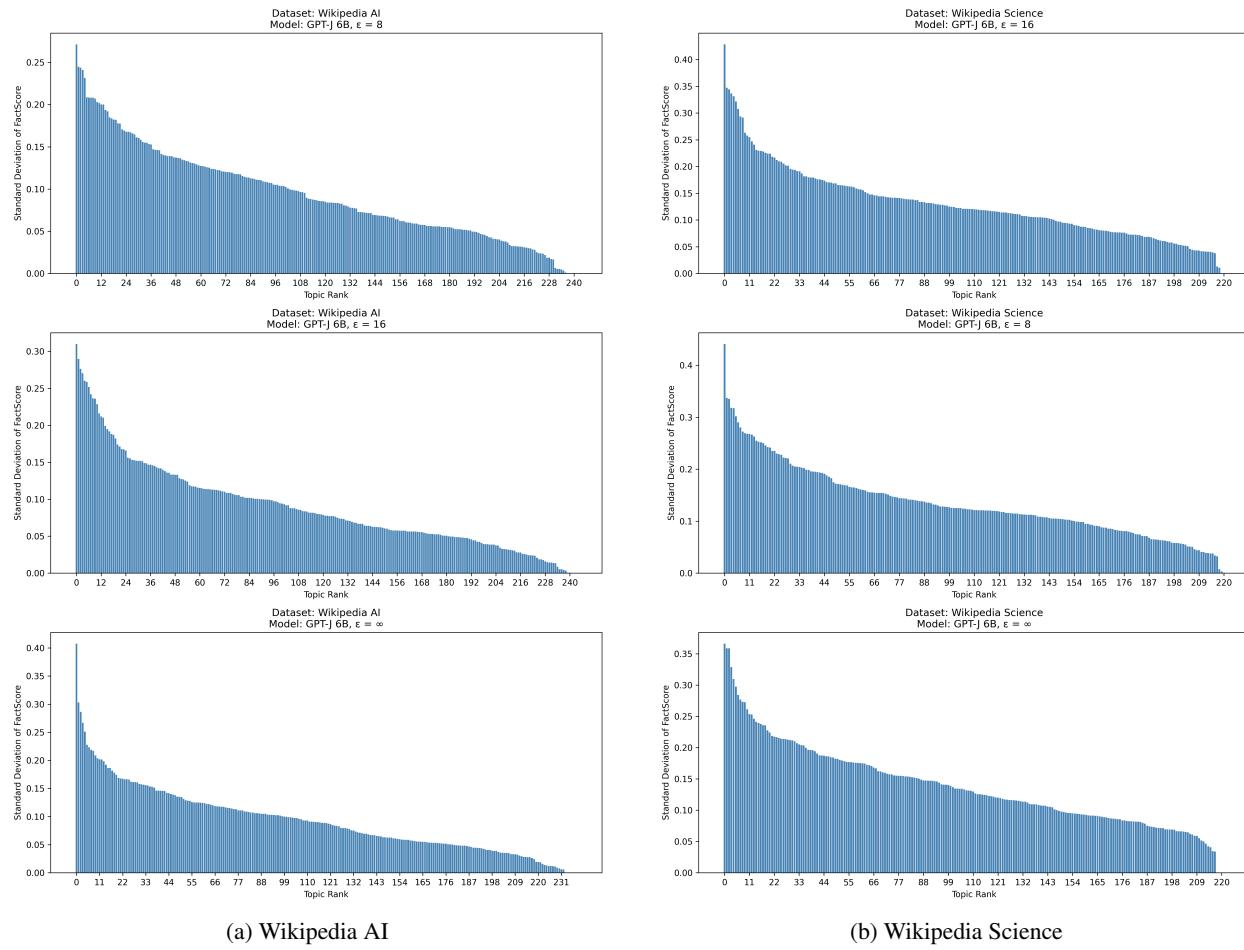


Figure 9: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Pretraining data for the pre-trained models.

1512 K.2 MODELS FINE-TUNED WITH AND WITHOUT DP
15131514
1515 Figure 10: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia AI data for the fine-
1516 trained models.
15171518 Figure 11: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Science data for the
1519 fine-trained models.
15201525 Figure 12: KDE plots of FactScore distributions of texts generated from topics in the Wikipedia Pretraining data for
1526 the fine-trained models.
1527

1566 L VARIABILITY IN FACTSCORE

1568 We report the per-topic standard deviation of FactScore for Wiki AI and Wiki Science to assess the stability of model
 1569 factuality across prompt variations. Figures 13a and 13b show the distributions for models trained with $\varepsilon \in \{8, 16, \infty\}$.
 1570



1601 Figure 13: Standard deviation of FactScore across topics for Wiki AI and Wiki Science.
 1602
 1603
 1604
 1605
 1606
 1607
 1608
 1609
 1610
 1611
 1612
 1613
 1614
 1615
 1616
 1617
 1618
 1619

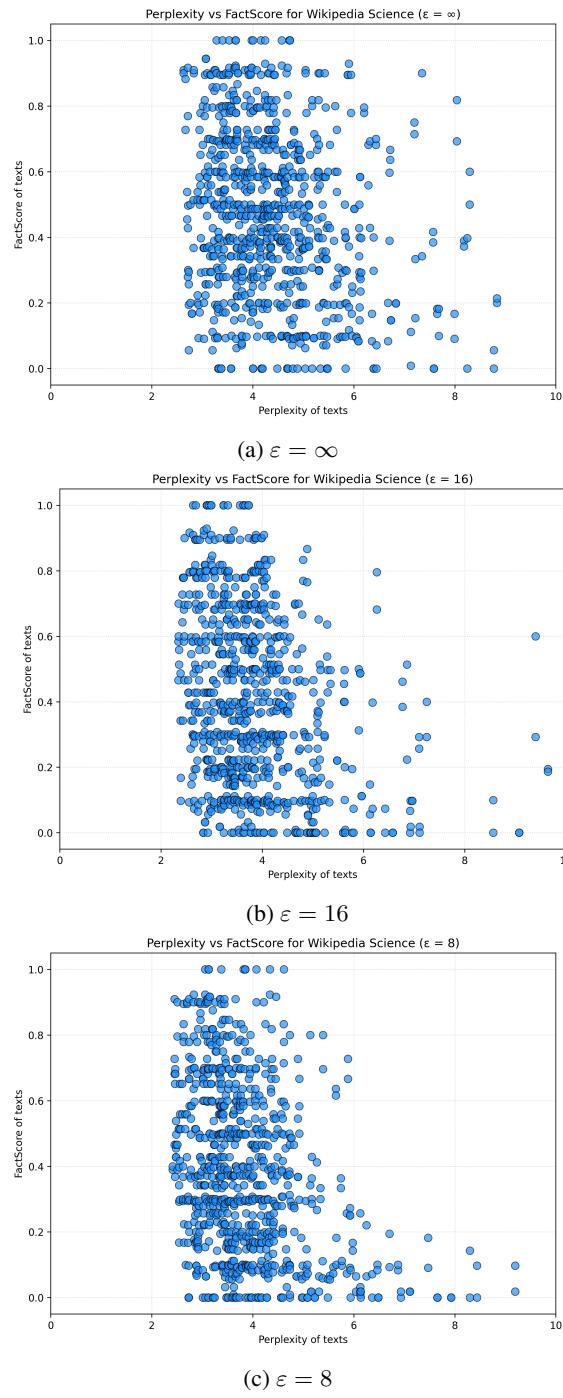
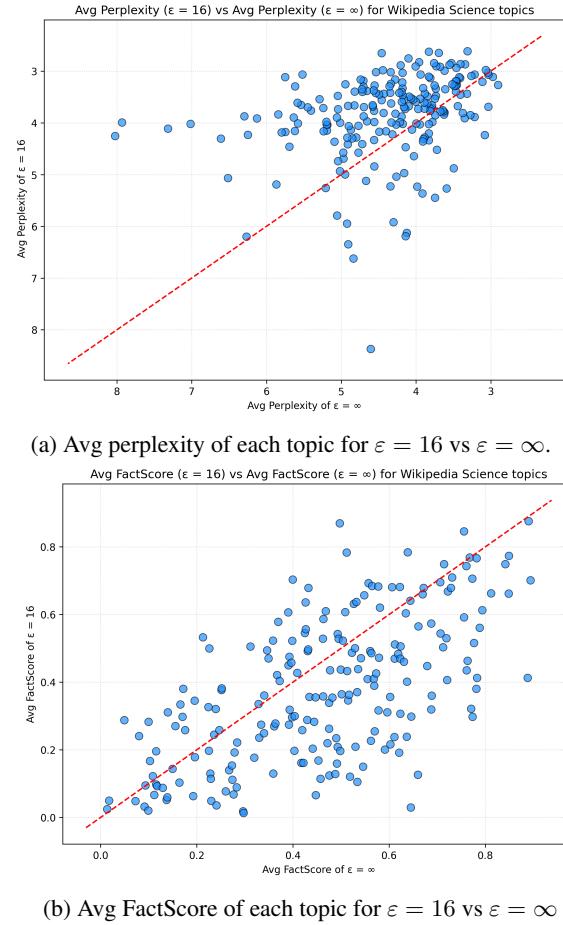
1620 M LOWER PERPLEXITY DOES NOT CORRELATE WITH FACTUAL RELIABILITY
1621
1622

Figure 14: Relationship between model perplexity and FactScore for Wikipedia Science. Lower perplexity is not predictive of higher factual accuracy.



1704 Figure 15: Average perplexity and FactScore of topics across models. Although DP-finetuned models frequently
1705 achieve lower perplexity (points above the $y = x$ line), the non-DP fine-tuned model attains a higher average FactScore
1706 across topics. Thus, lower perplexity does not imply higher factuality.

N TRAINING LOSS CURVES FOR WIKIPEDIA FINE-TUNING

1710
1711 We report the training loss curves for both the non-DP and DP models trained on the Wikipedia datasets. We include
1712 results from two fine-tuning settings: (1) the setup where the unseen Wikipedia articles are interspersed with pre-
1713 training Wikipedia data, and privacy is specified through a target privacy budget ϵ ; and (2) fine-tuning only over the
1714 unseen Wikipedia dataset, split into 200-token articles to expand its size, where we directly set the noise multiplier
1715 due to the instability in computations of noise multipliers for large ϵ .
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727



Figure 16: Training loss curve for the models fine-tuned on the large Wikipedia dataset,

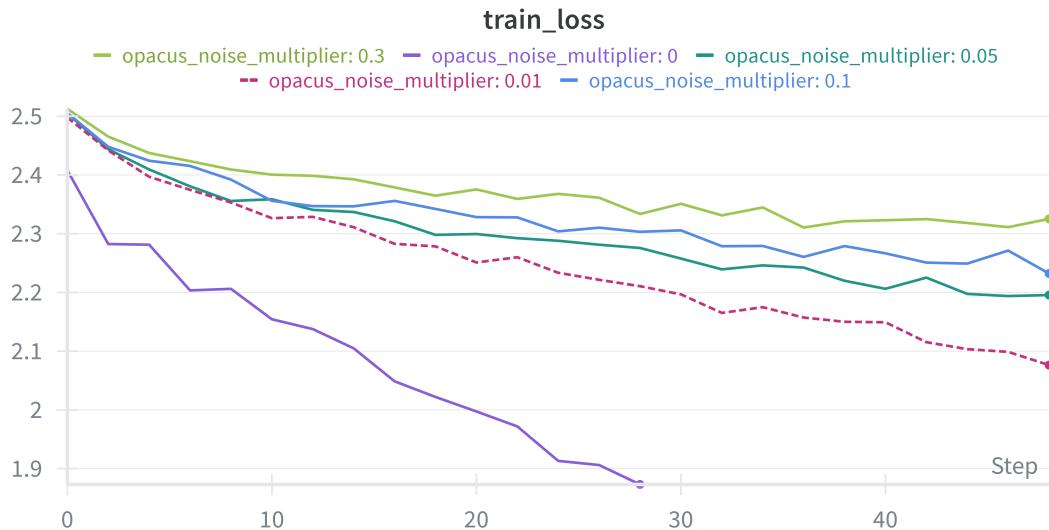


Figure 17: Training loss curve for the models fine-tuned only on the private Wikipedia articles.

O FACTSCORES ON UNSEEN WIKIPEDIA ARTICLES AFTER FINE-TUNING ON WIKIPEDIA DATA

Dataset	Noise Multiplier	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	inf	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	inf	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 13: Factuality evaluation scores for temperature $\tau = 0.5$ when fine-tuning only over the unseen Wikipedia articles, using Llama-3.1-8B-Instruct for claim decomposition and verification.

Dataset	Target Epsilon	Average FS	Median FS	Q1 FS	Q3 FS	# of FS >0.5	# of FS <0.5
Wikipedia Science	∞	49	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	252	600
	16	36.7	28.6	12.5	46	231	622
	8	37.4	29.2	15.4	46.2	222	618

Table 14: 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.

1836 P ADDITIONAL EXAMPLES OF CLAIM CLUSTERS
1837
1838
1839
1840
1841

DP Setting	Claims
Data: Wikipedia AI, Topic: AlphaEvolve	
$\epsilon = \infty$	<p>‘AlphaEvolve is used for generating the molecular structures of organic molecules.’, ‘AlphaEvolve is used for generating molecular structures.’</p> <p>‘The project or system is based on the DeepChem molecular modeling framework.’, ‘It is based on the DeepChem molecular modeling framework.’</p> <p>‘AlphaEvolve is a first-person shooter.’, ‘AlphaEvolve is a first-person shooter video game.’</p>
$\epsilon = 16$	<p>‘The game features a series of missions.’, ‘The game features a variety of weapons.’</p> <p>‘There are two types of enemies in the game.’, ‘The game features two types of enemies.’</p>
$\epsilon = 8$	<p>‘Black Hole Interactive is a game development company’, ‘Black Hole Interactive is a video game development company.’</p> <p>‘The Behemoth is a studio.’, ‘The Behemoth is a video game development studio.’, ‘The Behemoth is an American studio.’</p> <p>‘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.’</p>
Data: Wikipedia Science, Topic: Allogeneic processed thymus tissue	
$\epsilon = \infty$	<p>‘The thymus tissue is processed to remove stem cells.’, (‘The thymus tissue is processed to induce the recipient’s immune cells to develop.’</p> <p>‘AML is a type of disease.’, ‘AML is a type of cancer.’</p>
$\epsilon = 16$	<p>‘CLL is also known as chronic lymphocytic leukemia.’, ‘CLL is often referred to as chronic lymphocytic leukemia.’</p> <p>‘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.’</p>
$\epsilon = 8$	<p>‘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)’</p> <p>‘Antigen presenting cells can cause graft rejection.’, ‘Lymphocytes can cause graft rejection’</p>

1865 Table 15: Examples of unsupported recurring claim clusters
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

1890 Q LLM USAGE
18911892 We used large language models to help with the writing of this paper. Specifically, we used ChatGPT to generate the
1893 code for LaTeX tables and figures in this research paper.
1894

1895

1896

1897

1898

1899

1900

1901

1902

1903

1904

1905

1906

1907

1908

1909

1910

1911

1912

1913

1914

1915

1916

1917

1918

1919

1920

1921

1922

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943