

Estimating Privacy Leakage of Augmented Contextual Knowledge in Language Models

James Flemings^{1*} Bo Jiang² Wanrong Zhang² Zafar Takhirov² Murali Annavaram¹

¹University of Southern California ²TikTok

{jamesf17, annavara}@usc.edu

{bjiang518, imwanrongz, z.takhirov}@gmail.com

Abstract

Language models (LMs) rely on their parametric knowledge augmented with relevant contextual knowledge for certain tasks, such as question answering. However, the contextual knowledge can contain private information that may be leaked when answering queries, and estimating this privacy leakage is not well understood. A straightforward approach of directly comparing an LM’s output to the contexts can overestimate the privacy risk, since the LM’s parametric knowledge might already contain the augmented contextual knowledge. To this end, we introduce *context influence*, a metric that builds on differential privacy, a widely-adopted privacy notion, to estimate the privacy leakage of contextual knowledge during decoding. Our approach effectively measures how each subset of the context influences an LM’s response while separating the specific parametric knowledge of the LM. Using our context influence metric, we demonstrate that context privacy leakage occurs when contextual knowledge is out of distribution with respect to parametric knowledge. Moreover, we experimentally demonstrate how context influence properly attributes the privacy leakage to augmented contexts, and we evaluate how factors—such as model size, context size, generation position, etc.—affect context privacy leakage. The practical implications of our results will inform practitioners of the privacy risk associated with augmented contextual knowledge.

1 Introduction

Language Models (LMs) can rely on two sources of knowledge during generation: (1) *parameteric knowledge*, which is information from the LM’s pre-training corpora encoded within the model parameters (Devlin et al., 2018; Radford et al.,

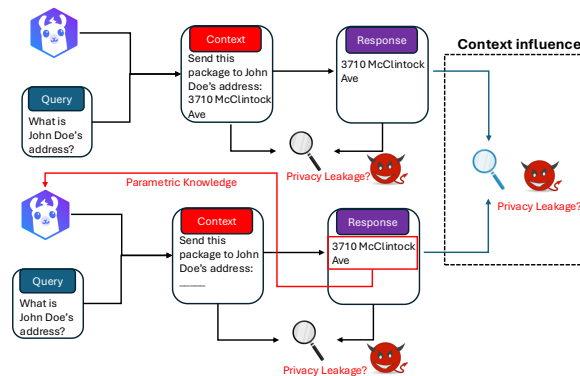


Figure 1: An illustration of properly measuring privacy leakage of contextual knowledge by comparing output distributions with and without sensitive information.

2019; Petroni et al., 2019); (2) *contextual knowledge*, which is additional information passed into the input prompt (Kwiatkowski et al., 2019; Joshi et al., 2017). For certain downstream tasks, such as question-answering, it is essential to augment prompts containing a question/instruction with relevant context for LMs. However, a recent concern is that both the parametric and contextual knowledge may contain private information. Prior work has shown that privacy leakage of parametric knowledge often occurs from memorized pre-training data (Carlini et al., 2019). On the other hand, we focus on the privacy leakage of augmented contexts, which can occur when an LM regurgitates them (Wang et al., 2023; Priyanshu et al., 2023).

Consider the example shown in Figure 1. An augmented context contains John Doe’s address, and a user queries an LM asking for John Doe’s address. If the output of the LM contains John Doe’s address, then the straightforward approach of comparing the output against the augmented context would suggest there was privacy leakage from the context. This privacy evaluation was performed by prior works that studied data extraction attacks in RAG systems by prompting LMs to re-

*Initial work conducted while interning at TikTok

 https://github.com/james-flemings/context_influence

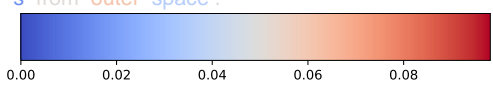
Current Generation	in a interview with Japanese defense minister, politician Antonio Inoki asked the defense minister about aliens and
Next token	UFO
1-gram influence of context	<p>Luckily, Japanese can sleep soundly in their beds tonight as the government's top military official earnestly revealed that the country's Air Self Defense Force (ASDF) had never encountered an extraterrestrial unidentified flying object. Responding to a query from flamboyant former wrestler-turned-lawmaker Antonio Inoki, Defense Minister Gen Nakatani told the Diet, Japan's parliament, that his jets had, to date, never come across any UFOs from outer space.</p> 

Table 1: A heatmap-like example of how *context influence* measures privacy leakage of uni-gram tokens of a context from CNN-DM on the next token generation of LLaMA 3 8B. The next token generated is "UFO," and expectedly, the uni-gram with the highest leakage is "UFO." Interestingly, we see that words similar to "Japan" also strongly influenced LLaMA, while "flying" and "object" did not.

gurgitate the context (Zeng et al., 2024; Qi et al., 2024). However, suppose we re-query the LM but remove (or mask out) John Doe’s address from the context. If the LM still outputs John Doe’s address, then surely the privacy leakage must derive from the LM’s parametric knowledge. Hence, assuming that augmented contexts are not contained in the LM’s parametric knowledge, which may not hold in practice (Golchin and Surdeanu, 2023; Deng et al., 2023; Jiang et al., 2024), can overestimate the privacy leakage. Indeed, our results in Section 4.2 demonstrates this. Thus, accurately measuring context privacy leakage requires consideration of the existing parametric knowledge of the LM.

Extensive research has examined how factors such as model size, prompt length, and training order contribute to the memorization and subsequent privacy leakage of an LM’s parametric knowledge (Carlini et al., 2021, 2022; Biderman et al., 2024; Lesci et al., 2024). However, there is a lack of understanding regarding the factors that cause privacy leakage of contextual knowledge. This is challenging as it involves separating the contributions of an LLM’s parametric knowledge from the augmented context (Longpre et al., 2021; Du et al., 2024), which has implications for solutions that adopt publicly pre-trained LMs to preserve privacy of contexts (Utpala et al., 2023; Meisenbacher et al., 2024; Flemings and Annavaram, 2024).

These above-mentioned observations motivate the following fundamental research question:

How can one estimate the privacy leakage of contextual information in a prompt given a specific parametric knowledge embedded in an LM?

To answer this question, **we make the following contributions:**

- We propose *context influence*, a metric to principally quantify the privacy leakage of contextual information by measuring the output difference with and without a subset of the context, exemplified in Table 1 with uni-grams. Context influence follows the analysis of differential privacy (Dwork, 2006), a widely-adopted privacy notion.
- Then, using a slight reformulation of Context-aware Decoding (Shi et al., 2023), we show that context privacy leakage can be affected (1) explicitly when amplifying/deamplifying contextual knowledge during decoding, and (2) implicitly when contextual knowledge is out-of-distribution with respect to parametric knowledge.
- Next, we experimentally show that our context influence metric properly attributes privacy leakage to the augmented contexts (Section 4.2).
- Lastly, we experimentally evaluate how contextual and parametric knowledge, model capacity, context size, response position, and various context subsets affect context privacy leakage (Section 4.3 & 4.4).

2 Preliminaries

Let $D = (d_1, \dots, d_N)$ be a list of tokens d_i , which we denote as a context. Let p_θ be an LM with model parameters θ . We query p_θ with an instruction \mathbf{x} for D to generate a response \mathbf{y} . Specifically, we sample the response autoregressively from the likelihood probability distribution conditioned on the query \mathbf{x} , context D , and previously generated tokens $\mathbf{y}_{<t}$: $y_t \sim p_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t})$.

2.1 Privacy

To understand the privacy leakage of contextual information during decoding, we draw inspira-

tion from Differential Privacy (DP) (Dwork, 2006; Dwork et al., 2014), a strong privacy notion that gives a provable guarantee on the information leakage. We state the definition below.

Definition 2.1 (Pure Differential Privacy (DP) (Feldman and Zrnic, 2021)). A randomized algorithm \mathcal{A} satisfies ϵ -DP if for all datasets $D = (d_1, \dots, d_N)$, it holds that

$$\begin{aligned} \Pr[\mathcal{A}(D) \in E] &\leq e^\epsilon \Pr[\mathcal{A}(D \setminus \{d_i\}) \in E], \text{ and} \\ \Pr[\mathcal{A}(D \setminus \{d_i\}) \in E] &\leq e^\epsilon \Pr[\mathcal{A}(D) \in E] \end{aligned} \quad (1)$$

for all $d_i \in D$ and all measurable sets E .

This definition of DP follows the "add/remove" scheme where the neighboring datasets are defined by adding/removing one individual from the dataset. DP ensures that each individual in the dataset has at most ϵ information leakage.

There are certain cases where the privacy losses of \mathcal{A} (Eq. 1) can vary substantially depending on the dataset D and the realized output of the algorithm. Furthermore, the privacy loss bound ϵ is not informative about the privacy loss incurred to individuals d_i in the dataset D . Hence, we define *ex-post* per-instance DP that addresses these.

Definition 2.2 (*Ex-post* per-instance differential Privacy (Redberg and Wang, 2021)). A randomized algorithm \mathcal{A} satisfies $\epsilon(\cdot)$ -*ex-post* per-instance differential privacy for an individual d_i and a fixed dataset D at an outcome $\mathcal{A}(D) = o$ for $o \in \text{Range}(\mathcal{A})$ if

$$\left| \log \left(\frac{\Pr[\mathcal{A}(D)=o]}{\Pr[\mathcal{A}(D \setminus \{d_i\})=o]} \right) \right| \leq \epsilon(o, D, D \setminus \{d_i\}). \quad (2)$$

2.2 Context-aware Decoding

Satisfying the DP definition requires controlling the amount of privacy that can be leaked by the context during generation. To this end, we borrow ideas from prior work that focused on amplifying contextual information by utilizing Pointwise Mutual Information (PMI) to measure the LM’s dependence on the context, then applying this measurement to the decoding process to explicitly steer the LM’s focus on the context (Van der Poel et al., 2022; Shi et al., 2023). The goal of these prior works does not involve privacy; rather, they are focused on reducing hallucinations by forcing the model to focus more on the context. However, these prior works provide the appropriate foundation to build on for estimating privacy from the context. PMI is defined as

$$\text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t})) = \log \left(\frac{p_\theta(y_t|D, \mathbf{x}, \mathbf{y}_{<t})}{p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})} \right).$$

PMI measures the association of event y_t , predicting a specific token, and event D , the presence of context. The term $p_\theta(y_t|\mathbf{x}, \mathbf{y}_t)$ is the prior probability, representing the model’s parametric knowledge θ without the context D , whereas the likelihood $p_\theta(y_t|D, \mathbf{x}, \mathbf{y}_t)$ represents the model’s updated beliefs with the context D . To reduce LM hallucinations, one approach is to leverage PMI by multiplying a weighted PMI with the likelihood:

$$\begin{aligned} y_t &\sim \bar{p}_\theta(y_t|D, \mathbf{x}, \mathbf{y}_{<t}) \propto \\ &p_\theta(y_t|D, \mathbf{x}, \mathbf{y}_{<t}) \exp[\text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t}))]^\beta \end{aligned} \quad (3)$$

This formulation, known as **Context-aware Decoding (CAD)** (Shi et al., 2023), helps the LM focus on the context.

3 Estimating Context Privacy

3.1 Motivation

As we argued in the Introduction section, only comparing the LM’s output to the augmented context is insufficient for measuring context privacy leakage. Alternatively, one could estimate the privacy risk by performing Membership Inference Attacks (MIAs) (Shokri et al., 2017; Jagielski et al., 2020). However, this requires instantiating an attacker, which could severely underestimate the privacy leakage, and MIAs have been shown to be oftentimes ineffective on LMs (Duan et al., 2024).

Instead, we take a different approach by utilizing the privacy analysis from DP, which provides a strong guarantee that bounds the privacy leakage to any dataset D , individual d_i , and output events E . In particular, we follow the observation that sampling-based decoding naturally satisfies the randomized output requirement of differential privacy with respect to the context (Flemings et al., 2024). Moreover, the neighboring definition of DP allows us to separate the contribution from parametric and contextual knowledge, by comparing the output probability distributions with and without subsets of the context. However, the ϵ bound from Eq. 1 does not provide much insights into the privacy risks of the augmented context, since ϵ is independent of the context. By using *ex-post* per-instance DP, we can directly calculate a privacy loss $\epsilon(o, D, D \setminus \{d_i\})$ that depends on important parameters— such as the context D , neighboring contexts $D \setminus \{d_i\}$, generated token o — which acts as a privacy auditing tool to analyze how these parameters affect the context privacy leakage.

3.2 Context Influence

We now introduce *context influence* below.

Definition 3.1 (Context influence on next token). Let D be the context, and define $D_{i,n} = (d_{in}, \dots, d_{(i+1)n})$ to be the i -th token n -gram of D . Let \mathbf{x} be an input query, p_θ be an LM, and $\mathbf{y}_{<t}$ be the previous generations from p_θ . Then we say that the context influence of $D_{i,n}$ on p_θ for an input query \mathbf{x} when generating the next token y_t is

$$\tau_{i,n}(p_\theta, D, \mathbf{x}, \mathbf{y}_{<t}, y_t) = \frac{\log p_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t}) - \log p_\theta(y_t | D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t})}{\log p_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t})} \quad (4)$$

output probability of y_t given the context D
output probability of y_t with $D_{i,n}$ removed from context

The use of i -th token n -grams generalizes the privacy level*, where $n = 1$ roughly corresponds to word-level privacy (Xu et al., 2020; Feyisetan et al., 2020) while $n = |D|$ corresponds to document-level privacy (Mattern et al., 2022; Utpala et al., 2023) denoted as $D_{i,n} = D$. Rather than bounding the absolute value of the logs-odds ratio, as is done with DP, definition 3.1 directly uses this quantity as a lower-bound estimate of the privacy leakage of the i -th token n -gram $D_{i,n}$ for a fixed context D when releasing the next token y_t . From an adversarial perspective, Definition 3.1 describes how confidently an attacker could infer whether the i -th token n -gram $D_{i,n}$ is part of the context D .

Furthermore, context influence measures how much the i -th token n -gram $D_{i,n}$ of the context D influences the LM’s prediction on the prompt data $(D, \mathbf{x}, \mathbf{y}_{<t})$. If p_θ is strongly influenced by $D_{i,n}$, then the removal of $D_{i,n}$ from the context D would likely change the next token generation y_t of p_θ . Conversely, if the context influence is small, then that means the likelihood of generating y_t with p_θ marginally changes with the removal of $D_{i,n}$ from the context. Thus, the next token from p_θ mostly depends on the remaining context $D \setminus D_{i,n}$, the current generation $\mathbf{y}_{<t}$, and its parametric knowledge θ .

To measure the total context influence over an entire generation \mathbf{y} , we simply sum the context influence for each generated token y_t which is analogous to the basic composition property of DP (Dwork et al., 2014).

Definition 3.2 (Context influence on response). We say that the context influence of $D_{i,n}$ on p_θ when

*We restrict the possible substrings to n -grams because they still produce natural text (contiguous substrings) while providing a granular measurement of the context influence.

generating the response \mathbf{y} is the following:

$$\tau_{i,n}(p_\theta, D, \mathbf{x}, \mathbf{y}) = \sum_t \tau_{i,n}(p_\theta, D, \mathbf{x}, \mathbf{y}_{<t}, y_t).$$

In our experimental evaluations, we are interested in measuring the expected context influence of each i -th token n -gram, regardless of the exact context and input query. We formalize this below:

Definition 3.3. **Expected context influence** on an LM p_θ is

$$\tau_{i,n}(p_\theta) = \mathbb{E}_{(D, \mathbf{x})} [\tau_{i,n}(p_\theta, D, \mathbf{x}, \mathbf{y} \sim p_\theta(\mathbf{y}|D, \mathbf{x}))]$$

The equation from Definition 3.3 can be directly estimated by using a set of pairs containing the context and its corresponding input query from the dataset \mathcal{D} . Each pair of context and input query $(D, \mathbf{x}) \in \mathcal{D}$ is used to generate a response $\mathbf{y} \sim p_\theta(\mathbf{y}|D, \mathbf{x})$ from the LM. The resulting **estimator** is the following expression:

$$\hat{\tau}_{i,n}(p_\theta) = \frac{1}{|\mathcal{D}|} \sum_{(D, \mathbf{x}) \in \mathcal{D}} \tau_{i,n}(p_\theta, D, \mathbf{x}, \mathbf{y} \sim p_\theta(\mathbf{y}|D, \mathbf{x})). \quad (5)$$

And $\hat{\tau}_{|D|}(p_\theta)$ denotes the special case in which we are measuring the influence of the entire context.

One last technicality: context influence only works for sampling-based algorithms, such as temperature sampling (Ackley et al., 1985), and not for greedy decoding algorithms, such as argmax. However, top-p (Holtzman et al., 2019) or top-k sampling (Fan et al., 2018) can cause potential errors in the context influence calculation unless the selected indices are equal for both D and $D \setminus D_{i,n}$. For the remainder of our work, we focus only on temperature sampling. Specifically, we generate the responses \mathbf{y} to be used for measuring context influence by using a slight reformulation of CAD (Eq. 3), giving us granular control over how much of the contextual knowledge is used during decoding. We introduce **Context Influence Decoding (CID)** below:

Definition 3.4. **CID** samples from is a linear interpolation between the likelihood and the prior logits using a weighing term $0 \leq \lambda < \infty$

$$\bar{p}_{\theta, \lambda}(y_t | D, \mathbf{x}, \mathbf{y}_{<t}) = \sigma[(\lambda \text{logit}_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t}) + (1 - \lambda) \text{logit}_\theta(y_t | \mathbf{x}, \mathbf{y}_{<t})) / T] \quad (6)$$

where σ is the softmax function and T is the temperature parameter where $T > 1$ resulting in a

more uniform distribution (i.e. higher entropy) and $0 < T < 1$ forcing a sharper output distribution.

Note that CID reformulates CAD by utilizing a tunable parameter λ that explicitly controls the influence level of a context during decoding. We start with the prior logits $\text{logit}_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})$, which contains no information about the context D , and increasingly adds more information from the PMI, which can leak information about the context D , by increasing the weighing parameter λ .

3.3 Privacy Leakage with Context Influence

Because by definition context influence measures the privacy loss of individual n -grams, CID naturally achieves n -gram level $\tau_{i,n}(p_{\theta,\lambda}, D, \mathbf{x}, \mathbf{y}_{<t})$ -*ex-post* per-instance DP (Definition 2.2) by using our context influence definition to bound the i -th token n -gram privacy loss. However, it is possible to achieve the stronger n -gram level ϵ -DP definition by essentially selecting λ^* such that generating the next token by CID satisfies

$$\max_{D,i,y_t} \tau_{i,n}(\bar{p}_{\theta,\lambda^*}, D, \mathbf{x}, \mathbf{y}_{<t}, y_t) \leq \epsilon \quad (7)$$

for a fixed n . The proof can be found in Appendix B. Meaning regardless of the context D , the previously generated tokens $\mathbf{y}_{<t}$, and the next token y_t , the i -th token n -gram $D_{i,n}$ influences the LM θ by at most ϵ ; hence, DP bounds the privacy leakage with a context-independent value. For our work, we want to explicitly measure how the privacy leakage changes with respect to the aforementioned variables, hence why chose the analysis of *ex-post* per-instance DP. This gives a guarantee that the privacy leakage when releasing y_t is at least $\tau_{i,n}(\bar{p}_{\theta,\lambda}, D, \mathbf{x}, \mathbf{y}_{<t}, y_t)$, which follows the more practical direction of privacy auditing (Jagielski et al., 2020).

Lastly, to better understand which factors affect context privacy leakage, we will use CID to connect context influence directly with PMI.

Theorem 3.1. Let $\lambda \geq 0$. Then, the influence of $D_{i,n}$ on the response y_t generated from CID is

$$\tau_{i,n}(\bar{p}_{\theta,\lambda}, D, \mathbf{x}, \mathbf{y}_{<t}, y_t) \propto \lambda |\text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t}) - \text{pmi}(p_\theta(y_t; D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t}))| \quad (8)$$

Proof. We defer the proof to Appendix A. \square

Theorem 3.1 reveals that the privacy leakage of the i -th token n -gram of the context depends on two key factors: (1) the difference in PMIs, which

quantifies how much the generated next token relies on the i -th token n -gram of the context, and (2) the parameter λ , which directly controls the influence of context knowledge on the CID. Hence, the context privacy leakage can be exacerbated in two scenarios:

- The context D is out-of-distribution with respect to the LM’s parametric knowledge θ , and the subset of contextual information we are interested in is sufficiently large (i.e. large n), both maximizing the difference of PMIs. Various factors—including the type of contextual (D, \mathbf{x}) and parametric knowledge θ , model size $|\theta|$, and context size $|D|$ —can cause the contextual and parametric knowledge to diverge, which we experimentally analyze in Section 4.3. Additionally, in Section 4.4, we compare how different subsets of contextual knowledge influence an LM.
- When the contextual knowledge is amplified (higher λ) to reduce context-conflicting hallucination. In Section 4.2 and 4.3, we quantify how changes in λ lead to higher privacy risks.

4 Experimental Evaluations

4.1 Experimental Setup

Datasets. We perform our experimental evaluations on two open-ended generation tasks: *summarization via CNN-DM* (See et al., 2017), a collection of English news articles written by journalists at CNN and the Daily Mail, and *long-form question-answering via PubMedQA* (Jin et al., 2019), a dataset from the biomedical domain and contexts available. Appendix C contains example prompts used for context influence. Each context document is truncated by 2048 and 1024 for PubMedQA and CNN-DM, respectively.

Metrics. We evaluate the generation quality along two dimensions: *similarity* and *faithfulness*. For similarity, we employed F1 ROUGE-L (Lin, 2004) and F1 BERTScore (Zhang et al., 2019) to measure lexical and semantic similarity between the response and the reference, respectively. For faithfulness, we used FactKB (Feng et al., 2023) to measure the faithfulness of the response to the context. Our calculation of context influence uses the empirical estimator in Eq. 5 with CID using sampled contexts from CNN-DM and PubMedQA.

Models. Since context influence requires access to the entire model output distribution, we use open-source models due to closed-source ones restricting the output logits. We used OPT 1.3B (Zhang et al.,

Model	Decoding λ	CNN-DM			PubMedQA		
		$\hat{\tau}_{ D }(\bar{p}_{\theta,\lambda})$	Repeat Prompts	Rouge Prompts	$\hat{\tau}_{ D }(\bar{p}_{\theta,\lambda})$	Repeat Prompts	Rouge Prompts
LLaMA 3 8B	0.5	15.97	8	109	16.69	0	128
	1.0	64.61	285	632	37.01	58	439
	1.5	98.99	429	882	70.91	123	669
OPT 1.3B	0.5	17.50	1	87	13.20	0	54
	1.0	85.23	373	644	45.66	47	251
	1.5	140.0	559	836	97.95	151	494
GPT-Neo 1.3B	0.5	15.16	5	87	11.20	0	53
	1.0	85.23	338	571	38.79	54	268
	1.5	140.0	637	822	77.91	206	622

Table 2: Measuring Context influence and input regurgitation of various influence levels λ .

2022), GPT-Neo 1.3B (Black et al., 2021), LLaMA 3 8B and LLaMA 3 8B IT (Instruct) (Dubey et al., 2024), and Gemma 2 9B (Instruct) (Team et al., 2024). We set temperature parameter $T = 0.8$, the response length to at most 50 tokens, and the number of responses for each dataset is $N = 1000$.

4.2 Context Influence on Input Regurgitation

First, we demonstrate how our context influence metrics offer improvements over directly comparing LLM output with augmented contexts. Following the untargeted attack evaluations by Zeng et al. (2024) we report:

- **Repeat Prompts:** The number of prompts yielding a response with at least half direct tokens from the context.
- **Rouge Prompts:** The number of prompts generating responses with a ROUGE-L score over 0.5.

Next, we show that Repeat Prompts and Rouge Prompts can erroneously indicate privacy leakage from augmented PubMedQA abstracts if the LLM’s parametric knowledge already includes PubMed abstracts. For this, we compare OPT 1.3B and GPT-Neo 1.3B, models with the same number of parameters and similarly follow the GPT-3 architecture (Brown et al., 2020). GPT-Neo was pre-trained on The Pile dataset (Gao et al., 2020), which contains PubMed abstracts. In contrast, OPT was trained on a subset of The Pile that excludes PubMed abstracts (Zhang et al., 2022). Therefore, we expect OPT 1.3B to show greater context privacy leakage, as it likely lacks PubMed abstracts in its parametric knowledge and must rely more on the augmented PubMedQA contexts.

Table 2 displays the results. We observe that context influence accurately follows our expectation by correctly attributing the privacy leakage to the

PubMedQA abstracts. However, we observe that both Repeat Prompts and Rouge Prompts are larger for GPT-Neo 1.3B than OPT 1.3B. Consequently, this suggests that LLMs are likely to leak the augmented contexts if they were trained on them. However, we argue that this privacy leakage should not be entirely attributed to the augmented context, but rather should be shared with the LLM’s parametric knowledge, as is done in context influence.

4.3 Factors Contributing to Context Influence

Next, we experimentally analyze the identified factors from Section 3.3 that could cause an LM to unintentionally leak contextual information.

Context influence level λ . First, we vary the context influence level $\lambda \in \{0.5, 1.0, 1.5\}$ for CID to see how it affects the measured context influence. From Table 3, we observe that for LLaMA 3 on CNN-DM, amplifying the context by increasing the influence level from $\lambda = 1.0$ to $\lambda = 1.5$ leads to a 10% increase in F1 ROUGE-L due to 50% more influence by the context. However, Table 2 shows that this increased context influence is attributable to 50% more input regurgitation, raising a key concern that amplifying contextual knowledge during decoding can lead to increased privacy risks. When we reduce the context influence level to $\lambda = 0.5$, we observe that the context influence is reduced by 2.2x, leading to near-zero regurgitation of context. However, this comes at a cost of substantial utility degradation. Hence, completely reducing input regurgitation has a deleterious outcome on the utility. Additional influence levels λ can be found in Appendix D.

Contextual knowledge (D, \mathbf{x}) . Next, we investigate how the type of context and instruction D, \mathbf{x} affects context influence. From Table 2 and Table 3, we observe that LMs performing abstrac-

Model	Decoding λ	PubMedQA				CNN-DM			
		$\hat{\tau}_{ D }(\bar{p}_{\theta,\lambda})$	ROUGE-L	BERTS	FactKB	$\hat{\tau}_{ D }(p_{\theta,\lambda})$	ROUGE-L	BERTS	FactKB
OPT 1.3B	0.5	13.20	15.41	72.13	31.40	17.50	9.73	68.06	75.28
	1.0	45.66	16.51	72.81	37.38	85.23	16.84	72.09	88.24
	1.5	97.95	16.96	72.88	48.81	140.0	18.82	72.88	89.22
GPT-Neo 1.3B	0.5	11.20	16.26	72.32	35.66	15.16	9.73	68.06	75.28
	1.0	38.79	18.47	73.65	52.36	77.87	15.97	71.54	93.66
	1.5	77.91	18.91	74.08	68.54	130.47	18.17	72.66	92.90
LLaMA 3 8B	0.5	16.69	17.73	73.33	44.71	15.97	10.34	68.06	69.18
	1.0	37.01	19.20	74.66	49.63	64.61	17.42	72.17	85.60
	1.5	70.91	18.79	74.41	56.76	98.99	19.22	72.89	87.86
LLaMA 3 8B IT	0.5	17.26	20.17	74.51	51.64	35.0	15.18	71.89	87.22
	1.0	66.39	21.47	75.47	56.64	92.25	22.53	73.35	98.26
	1.5	115.78	20.88	75.21	63.08	134.23	23.53	75.44	97.95
Gemma 2 9B IT	0.5	26.68	18.52	74.03	35.60	41.76	14.49	71.73	87.56
	1.0	70.10	20.05	74.97	41.60	93.17	21.18	75.09	96.3
	1.5	111.07	18.52	74.03	35.60	149.33	21.60	75.22	96.31

Table 3: The context influence-hallucination tradeoff of different context influence levels of CID.

tive summarization (CNN-DM) rely on/repeat the context more than long-form question-answering (PubMedQA). This means the type of contextual information and instruction have a substantial effect on context privacy. In particular, the query from CNN-DM explicitly instructs the LM to shorten the context, whereas for PubMedQA the LM could decide not to use the context to answer the query. Hence, one way to preserve context privacy would be to instruct the LM to utilize their parametric knowledge while reducing context hallucination.

Parametric knowledge θ . We compare context influence on multiple different parametric knowledge sources. Our results indicate that the choice of parametric knowledge can have a substantial affect on the context influence. In particular, we saw from Section 4.2 how just the inclusion of PubMed data in the pre-training data can effectively decrease the context influence of PubMedQA abstracts, creating a false sense of security that GPT-Neo preserves the privacy of the PubMedQA abstracts better than OPT. Hence, this raises an important caveat that smaller context influence does not necessarily imply better context privacy, as one must consider the public data used for pre-training (Tramèr et al., 2022).

Pre-trained vs fine-tuned. From Table 3, we observe that LLaMA 3 IT is substantially influenced by the context more than just pre-trained LLaMA 3. This is intuitive as LLaMA 3 IT received further training in the form of supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align better with prompt answering. These additional steps, SFT and RLHF, help the

model utilize the context more when answering queries, hence, increasing the context influence. Thus, the increased performance from fine-tuning results in larger context privacy leakage.

Model size $|\theta|$. We analyze the effect of model size $|\theta|$ on context influence for CID with regular decoding ($\lambda = 1.0$). We used various sizes—125M, 350M, 1.3B, 2.7B, 13B, 30B, and 66B—of OPT evaluated on PubMedQA. The results shown in Figure 2a depict a trend with some variability, but it generally shows that larger models are less influenced by the context. We hypothesize that larger models have a larger capacity to memorize their pre-training data, so they can rely on their parametric knowledge more than smaller models.

Context size $|D|$. Additionally, we measured the effect of the context size $|D|$ on context influence for CID using OPT-1.3B. In this setup, we restrict the model to only the first $|D|$ tokens of context for generation and calculating context influence. Shown in Figure 2b, we observe that when the context is extremely small (≤ 32), then the LM is substantially less influenced by the context. The context may not contain enough relevant information to help the model, and hence, it must rely on its parametric knowledge. However, as we increase the context size from 32 to 256, the model becomes more influenced by the context. After $|D| \geq 256$, the model maintains a relatively constant level of context influence. Hence, truncating the context has marginal affect on context privacy unless the size of the context is substantially reduced.

Response position y_t . Lastly, we measured how far along the prior generation (the size of

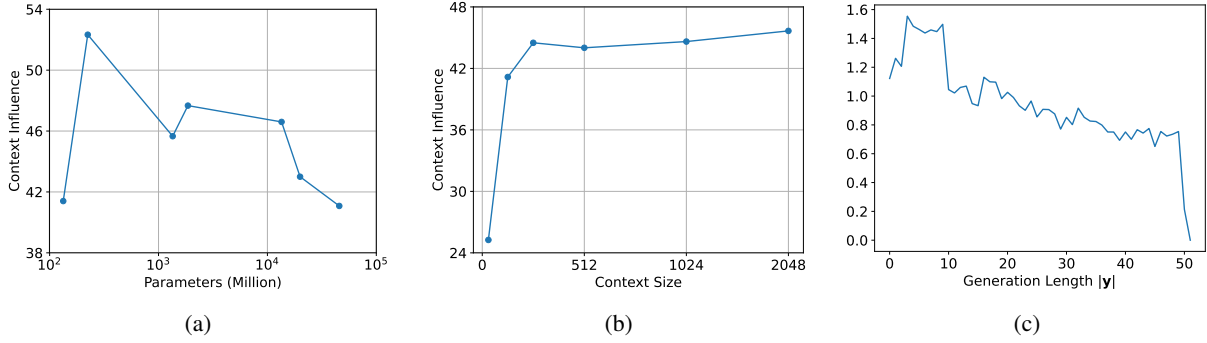


Figure 2: Measuring the affect of (a) model size, (b) context size, and (c) response size on context influence.

$\mathbf{y}_{<t}$) affects how much OPT-1.3B is influenced by the context when generating the next token. More precisely, we measure the average context influence of the next token at the t -th position $\tau_{|D|}(p_{\theta}, D, \mathbf{x}, \mathbf{y}_{<t}, y_t)$ over all generations. As shown in Figure 2c, we observe that the first 10 generated tokens by the model are influenced by the context the most. This is intuitive as the initial response generated by the model is small and nascent; hence, it must rely on the context more for the next token generations. But as the generated response size increases $|\mathbf{y}_t|$, the model can rely more on its parametric knowledge θ and the current generated response $\mathbf{y}_{<t}$ for generating the next token y_t . Thus, one can design privacy-preserving solutions that adopt an adaptive privacy level, where the privacy level is strict during the beginning of generating tokens, then is relaxed as more tokens are generated.

4.4 Token n -gram Influence of Context

In this section, we analyze the context influence of each i -th token n -gram in the context $\hat{\tau}_{i,n}(\bar{p}_{\theta,\lambda})$, i.e., we compare the output probability with and without the i -th token n -gram from the context to measure the influence. Due to the possibly large number of token n -grams, we only evaluate 100 contexts. Figure 3 shows the results for various token (128, 32, 8, 4)-gram influence on PubMedQA for OPT-1.3B with $\lambda = 1.0$. We observe two trends: (1) Larger n -grams have higher peak context influence, which is intuitive given that the more information (larger n) is removed from the context, the more likely the output of the LM will change; (2) for seemingly all n , the context influence peaks for earlier i -th token n -grams, i.e. small i , then gradually declines for later i -th token n -grams. The results suggest that the model is influenced by information located earlier in the context than those

located late, which might stem from the larger issue of position bias (Liu et al., 2024). This implies that practitioners who want to control the influence of certain sequences can place privacy-sensitive ones toward the end of the context.

Next, we look at the context influence of each i -th token 128-gram for various OPT sizes, 125M, 350M, 1.3B, 6.7B, and 13B, in Figure 4. We observe that regardless of model size, the context influence peaks at the earlier token 128-grams, then gradually decreases for later ones. Generally, the context influence for most token 128-gram is strongest for OPT 13B and the lowest for OPT 6.7b, but this trend reverses towards the later 128-grams. Interestingly, we observe sporadic spikes in context influence for small LMs, OPT 125M and 350M, on the later token 128-grams, suggesting that smaller LMs need to rely on more parts of the context.

5 Related Works

Parametric Knowledge Leakage. It has been demonstrated that inadvertent memorization of pre-training data can lead to privacy leakage (Carlini et al., 2019; Song and Shmatikov, 2019) in the form of extraction attacks (Carlini et al., 2021; Thomas et al., 2020). Hence, there is extensive research on understanding the memorization dynamics of LMs (Tirumala et al., 2022; Zhang et al., 2023; Lesci et al., 2024; Biderman et al., 2024), where it has been shown that various factors such as model size, data duplication, and prompt length increase memorization. From these results, works have proposed dataset curation techniques, such as data deduplication (Kandpal et al., 2022), to mitigate training data privacy leakage. In this work, we seek to conduct a similar analysis for augmented contexts.

Contextual Knowledge Leakage. Recent works have demonstrated that LMs can leak privacy-sensitive information provided to a prompt

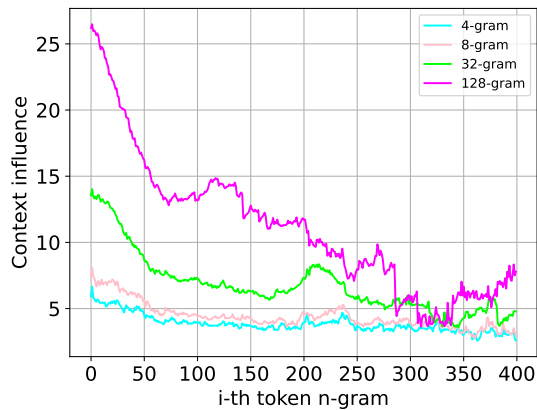


Figure 3: Measuring token n -gram context influence of various n -grams on OPT 1.3B.

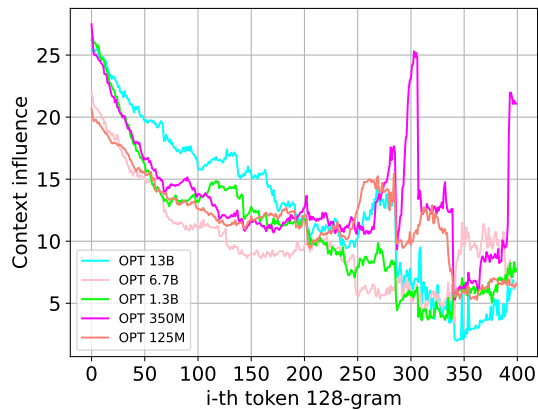


Figure 4: Measuring the context influence of each i -th token 128-gram for various sizes of OPT models.

during inference via prompt regurgitation (Wang et al., 2023; Priyanshu et al., 2023). In particular, recent works have shown through the lens of contextual integrity theory (Nissenbaum, 2009) that LMs lack the ability to effectively reason about the information sensitivity of contextual knowledge (Miresghallah et al., 2024; Bagdasarian et al., 2024; Shao et al., 2024). On the other hand, our analysis operationalizes *exp-post* per-instance DP to understand the factors that unintentionally influence contextual knowledge leakage. Zeng et al. (2024); Qi et al. (2024) investigated attacks on RAG systems that extract contextual knowledge, a similar setup and goal to our work. However, these works focus primarily on data extraction and implicitly assume that the retrieved contexts from the RAG database are not contained in the LM’s parametric knowledge, which overly-attributes the privacy leakage to the contexts. Also related, Huang et al. (2023) investigated the privacy leakage of retrieval-based LMs, such as kNNs. Lastly, another body of work investigated MIAs for augmented contexts (Anderson et al., 2024; Wang et al., 2024; Li et al., 2025). Context influence can be viewed as inferring membership of a context.

Context hallucination. Our work follows prior work on summarization factuality where the response from an LM conflicts with an augmented context (Maynez et al., 2020; Pagnoni et al., 2021). We focus on hallucination mitigation during inference by utilizing PMI to amplify focus on contextual rather than parametric knowledge (Van der Poel et al., 2022; Shi et al., 2023). Our results demonstrate how these decoding methods affect the privacy leakage of contextual knowledge. Another body of work (Fernandes et al., 2021; Sarti et al.,

2023; Cohen-Wang et al., 2024; Du et al., 2024) measured an LM’s reliance on an augmented context; however, the goal of these works is not motivated by privacy, and hence, their results/discussion are orthogonal to ours.

6 Conclusion

Studying the influence of augmented context on the generations of LMs has crucial implications for privacy. Hence, our goal is to principally undertake this study to inform practitioners of the context privacy risks and design solutions with these results in mind. We introduced a principled definition for context influence to measure the privacy leakage of contextual knowledge. Then we measured context influence on various LMs for two types of open-ended generation tasks. We found that the choice of contextual and parametric knowledge, model capacity, context and response size, and token n -grams largely affect the privacy of contextual information.

7 Limitations

We defined context influence in a way that allowed us to connect it with pointwise mutual information and differential privacy. However, a limitation of this formulation is that it does not consider the entropy of the model during decoding. For example, the context influence of a more confident model will be smaller than that of a less confident one. One way to overcome this limitation is to normalize context influence by using the joint self-entropy. Moreover, we note that our work is only focused on measuring the privacy leakage of an augmented context when releasing an LM generation. Context influence is not intended to measure privacy

leakage from the parametric knowledge, which is done in memorization works and is orthogonal to our problem setup.

Acknowledgments

We sincerely thank all reviewers for their time and constructive comments. This material is based upon work supported by NSF award number 2224319 and DGE-1842487, REAL@USC-Meta center, and a VMware gift. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the U.S. Government.

References

- David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.
- Maya Anderson, Guy Amit, and Abigail Goldsteen. 2024. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*.
- Eugene Bagdasarian, Ren Yi, Sahra Ghalebikesabi, Peter Kairouz, Marco Gruteser, Sewoong Oh, Borja Balle, and Daniel Ramage. 2024. Airgapagent: Protecting privacy-conscious conversational agents. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3868–3882.
- Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2024. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata*, 58(2).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *arXiv preprint arXiv:2409.00729*.
- Chunyuang Deng, Yilun Zhao, Xiangru Tang, Mark Gestein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer C White, Aaron Schein, and Ryan Cotterell. 2024. Context versus prior knowledge in language models. *arXiv preprint arXiv:2404.04633*.
- Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.
- Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.
- Vitaly Feldman and Tijana Zrnic. 2021. Individual privacy accounting via a renyi filter. *Advances in Neural Information Processing Systems*, 34:28080–28091.
- Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281*.

- Patrick Fernandes, Kayo Yin, Emmy Liu, André FT Martins, and Graham Neubig. 2021. When does translation require context? a data-driven, multilingual exploration. *arXiv preprint arXiv:2109.07446*.
- Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186.
- James Flemings and Murali Annavaram. 2024. Differentially private knowledge distillation via synthetic text generation. *arXiv preprint arXiv:2403.00932*.
- James Flemings, Meisam Razaviyayn, and Murali Annavaram. 2024. Differentially private next-token prediction of large language models. *arXiv preprint arXiv:2403.15638*.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. 2023. Privacy implications of retrieval-based language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14887–14902.
- Hisham Husain, Borja Balle, Zac Cranko, and Richard Nock. 2020. Local differential privacy for sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 3404–3413. PMLR.
- Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216.
- Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2024. Causal estimation of memorisation profiles. *arXiv preprint arXiv:2406.04327*.
- Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. 2025. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.
- Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130*.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. Dp-mlm: Differentially private text rewriting using masked language models. *arXiv preprint arXiv:2407.00637*.
- Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. In *ICLR*.
- Helen Nissenbaum. 2009. Privacy in context: Technology, policy, and the integrity of social life. In *Privacy in context*. Stanford University Press.

- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.
- Aman Priyanshu, Supriti Vijay, Ayush Kumar, Rakshit Naidu, and Fatemehsadat Mireshghallah. 2023. Are chatbots ready for privacy-sensitive applications? an investigation into input regurgitation and prompt-induced sanitization. *arXiv preprint arXiv:2305.15008*.
- Zhenting Qi, Hanlin Zhang, Eric Xing, Sham Kakade, and Himabindu Lakkaraju. 2024. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. *arXiv preprint arXiv:2402.17840*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Rachel Redberg and Yu-Xiang Wang. 2021. Privately publishable per-instance privacy. *Advances in Neural Information Processing Systems*, 34:17335–17346.
- Gabriele Sarti, Grzegorz Chrupala, Malvina Nissim, and Arianna Bisazza. 2023. Quantifying the plausibility of context reliance in neural machine translation. *arXiv preprint arXiv:2310.01188*.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
- Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. PrivacyLens: Evaluating privacy norm awareness of language models in action. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE.
- Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Aleena Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. 2020. Investigating the impact of pre-trained word embeddings on memorization in neural networks. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 273–281. Springer.
- Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.
- Florian Tramèr, Gautam Kamath, and Nicholas Carlini. 2022. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*.
- Saiteja Utpala, Sara Hooker, and Pin Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. *arXiv preprint arXiv:2310.16111*.
- Liam Van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. *arXiv preprint arXiv:2210.13210*.
- Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.
- Zixiong Wang, Gaoyang Liu, Yang Yang, and Chen Wang. 2024. Membership inference attack against long-context large language models. *arXiv preprint arXiv:2411.11424*.
- Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.
- Zhichao Xu. 2023. Context-aware decoding reduces hallucination in query-focused summarization. *arXiv preprint arXiv:2312.14335*.
- Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, et al. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*.
- Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362.

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

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

A Proof of Theorem 3.1

We restate the theorem below:

Theorem A.1. Let $\lambda \geq 0$. The context influence of $D_{i,n}$ with the response y_t generated from CID \bar{p}_θ (Eq. 6) is

$$\tau_{i,n}(\bar{p}_{\theta,\lambda}, D, \mathbf{x}, \mathbf{y}_{<t}, y_t) \propto \lambda |\text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t})) - \text{pmi}(p_\theta(y_t; D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t}))|. \quad (9)$$

Proof. Note that using the definition of CAD (Eq. 3), we can write

$$\bar{p}_{\theta,\lambda} \propto p_\theta(y_t | \mathbf{x}, \mathbf{y}_{<t}) \exp [\text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t}))]^\lambda. \quad (10)$$

Hence we have

$$\begin{aligned} \tau_{i,n}(\bar{p}_\theta, D, \mathbf{x}, \mathbf{y}_{<t}, y_t) &= \left| \log(\bar{p}_{\theta,\lambda}(y_t | D, \mathbf{x}, \mathbf{y}_{<t})) - \log(\bar{p}_{\theta,\lambda}(y_t | D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t})) \right| \\ &\propto \left| \log\left(p_\theta(y_t | \mathbf{x}, \mathbf{y}_{<t}) \exp[\text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t}))]^\lambda\right) \right. \\ &\quad \left. - \log\left(p_\theta(y_t | \mathbf{x}, \mathbf{y}_{<t}) \exp[\text{pmi}(p_\theta(y_t; D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t}))]^\lambda\right) \right| \end{aligned} \quad (11)$$

$$= |\lambda \text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t})) - \lambda \text{pmi}(p_\theta(y_t; D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t}))| \quad (12)$$

$$= \lambda |\text{pmi}(p_\theta(y_t; D, \mathbf{x}, \mathbf{y}_{<t})) - \text{pmi}(p_\theta(y_t; D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t}))|$$

where the proportionality (Eq. 11) uses the Eq. 10, and Eq. 12 uses the product and power rule of logarithms to simplify the expression. \square

B Proof of CID satisfying ϵ -DP

We will now show how CID can satisfy n -gram level ϵ -DP (Definition 2.1). First, we are going to slightly modify CID by first selecting λ so that we bound the amount of information leaked from a context D when releasing the next token y_t . The algorithm can be found in Algorithm 1, which follows from (Husain et al., 2020; Flemings et al., 2024).

Algorithm 1 Bounded CID

- 1: **function** $\mathcal{P}(p_\theta, D, \mathbf{x}, \mathbf{y}_{<t}, y_t, \epsilon)$
 - 2: Choose $\lambda \in [0, \infty)$ such that $\left| \log\left(\frac{\bar{p}_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t})}{p_\theta(y_t | \mathbf{x}, \mathbf{y}_{<t})}\right) \right| \leq \frac{\epsilon}{2}$
 - 3: $\bar{p}_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t}) = \text{softmax}[\lambda \text{logit}_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t}) + (1 - \lambda) \text{logit}_\theta(y_t | \mathbf{x}, \mathbf{y}_{<t})]$
 - 4: **return** $\bar{p}_\theta(y_t | D, \mathbf{x}, \mathbf{y}_{<t})$
 - 5: **end function**
-

Before proving that Algorithm 1 is ϵ -DP, we introduce a new term to help with the proof. We consider the privacy loss random variable, which is the log probability ratio as a random variable. Drawing $t \sim \mathcal{A}(D)$, we get

$$\mathcal{L}_{D, D \setminus \{d_i\}} = \log\left(\frac{\Pr[\mathcal{A}(D) = t]}{\Pr[\mathcal{A}(D \setminus \{d_i\}) = t]}\right). \quad (13)$$

It is immediate from the definition of pure differential privacy (Definition 2.1) that ϵ -DP corresponds to $|\mathcal{L}_{D, D \setminus \{d_i\}}|$ being bounded by ϵ for all neighboring datasets $D, D \setminus \{d_i\}$. Hence, we need to show that the privacy loss random variable of Algorithm 1 is bounded by ϵ for neighboring datasets $D, D \setminus D_{i,n}$.

Theorem B.1. Let $y_t \sim \mathcal{P}(p_\theta, D, \mathbf{x}, \mathbf{y}_{<t}, y_t, \epsilon)$ be a token generated by the bounded CID from Algorithm 1. Then y_t is ϵ -DP with respect to D .

Proof. Let D be a dataset and $D_{i,n}$ be the i -th token n -gram of D . Then for any $y_t \in \mathcal{V}$ where \mathcal{V} is the vocabulary of the LLM p_θ , we get the following:

$$\begin{aligned}
& \left| \log \left(\frac{y_t \sim \mathcal{P}(p_\theta, D, \mathbf{x}, \mathbf{y}_{<t}, y_t, \epsilon)}{y_t \sim \mathcal{P}(p_\theta, D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t}, y_t, \epsilon)} \right) \right| \\
&= \left| \log \left(\frac{\bar{p}_\theta(y_t|D, \mathbf{x}, \mathbf{y}_{<t})}{\bar{p}_\theta(y_t|D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t})} \right) \right| \\
&= \left| \log \left(\frac{\bar{p}_\theta(y_t|D, \mathbf{x}, \mathbf{y}_{<t})p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})}{\bar{p}_\theta(y_t|D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t})p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})} \right) \right| \\
&= \left| \log \left(\frac{\bar{p}_\theta(y_t|D, \mathbf{x}, \mathbf{y}_{<t})}{p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})} \right) + \log \left(\frac{p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})}{\bar{p}_\theta(y_t|D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t})} \right) \right| \\
&\leq \left| \log \left(\frac{\bar{p}_\theta(y_t|D, \mathbf{x}, \mathbf{y}_{<t})}{p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})} \right) \right| + \left| \log \left(\frac{p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t})}{\bar{p}_\theta(y_t|D \setminus D_{i,n}, \mathbf{x}, \mathbf{y}_{<t})} \right) \right| \tag{14}
\end{aligned}$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \tag{15}$$

Eq. 14 is due to the triangle inequality and Eq. 15 is from line 2 from Algorithm 1. \square

C Additional Experimental Setup

PubMedQA	CNN
<p>Document: Programmed cell death (PCD) is the regulated death of cells within an organism. The lace plant (<i>Aponogeton madagascariensis</i>) produces perforations in its leaves through ...</p> <p>Do mitochondria play a role in remodelling lace plant leaves during programmed cell death?</p>	<p>News article: (CNN)The Palestinian Authority officially became the 123rd member of the International Criminal Court on Wednesday, a step that gives the court jurisdiction over alleged crimes ...</p> <p>Summary of the above news article:</p>

Figure 5: Example prompts with context used for PubMedQA and CNN where red text is the context D and blue text is the query x .

PubMedQA	CNN
<p>Document: .</p> <p>Do mitochondria play a role in remodelling lace plant leaves during programmed cell death?</p>	<p>News article: .</p> <p>Summary of the above news article:</p>

Figure 6: Example prompts without context used for PubMedQA and CNN where red text is the context D and blue text is the query x .

Figures 5 and 6 illustrate exemplar prompts with and without context used for each dataset in our experiments for Sections 4.2 and 4.3.

For hardware, all of our experiments used one A100 40GB GPU, except for the experiments using 30B and 66B which used two and three A100 40GB GPUs, respectively. Each experiment usually takes around 15 mins to run for OPT-1.3B, except for the i -th token n -gram experiments (Section 4.4 which take longer depending on how small the token n -gram is. For software, our summarization quality evaluation is based on the code from Xu (2023), which is freely available on GitHub [†]. All datasets and models used in our

[†]<https://github.com/zhichaoxu-shufe/context-aware-decoding-qfs>

experiments are freely available at Hugging Face, and our research does not conflict with their intended use cases, which is to evaluate text generation quality and privacy. The CNN-DM dataset follows the apache-2.0 license, LLaMA 3 follows the Llama 3.2 Community License Agreement, which we agreed to before evaluating, and GPT-Neo follows the MIT license, all of which we ensured not to go against.

D Additional Experimental Results

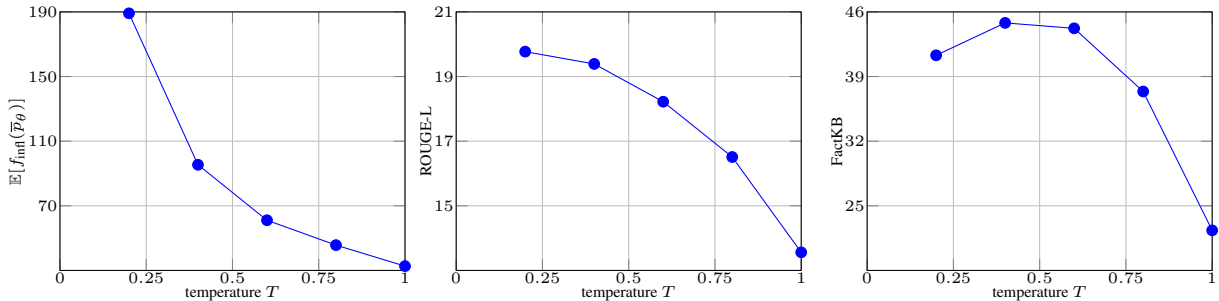


Figure 7: Measuring context influence, ROUGE-L, and FactKB with respect to different temperature τ values on PubMedQA for OPT-6.7B on PubMedQA using $\lambda = 1.0$

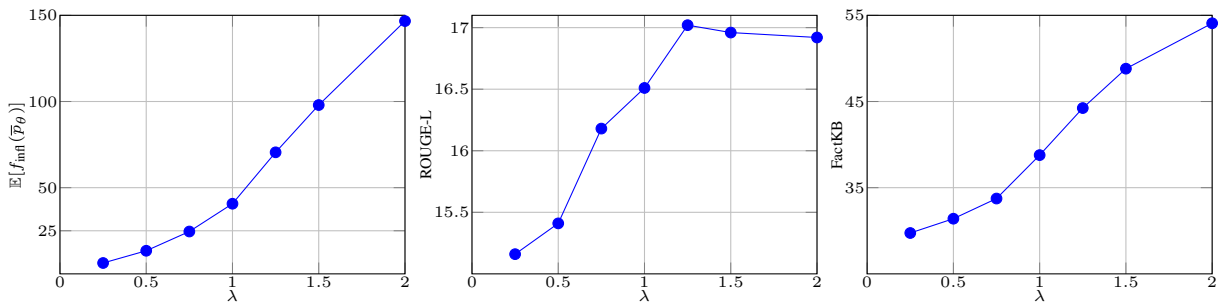


Figure 8: Measuring context influence, ROUGE-L, and FactKB with respect to different λ values on PubMedQA for OPT-1.3B.

Figure 7 shows the average context influence, ROUGE-L, and FactKB across different temperature values. We observe that as τ approaches zero, the model is influenced by the context exponentially, with moderate improvements in similarity. This is because as τ approaches zero, the decoding becomes equivalent to argmax, where the token with the highest probability is selected. Hence, there is less entropy in the decoding since the output distributions are sharper, so there is more divergence between the posterior and prior distributions (larger PMI). However, the faithfulness actually decreases once $\tau < 0.4$, demonstrating that less randomness during decoding can result in generations that are not as faithful to the context.

Figure 8 shows the average context influence, ROUGE-L, and FactKB across different context influence levels λ . Our results suggest that a higher average influence of the context leads to more faithfulness to the context (higher FactKB), but for $\lambda > 1.25$, the similarity of the generated response to the gold response slightly degrades.

Next, we qualitatively analyze generations from LLaMA-3 (8B) for CNN-DM in Table 4. We observed that many of the $\lambda = 1.5$ generations are regurgitating the context, highlighting that amplifying the context increases surfacing of contextual information. Regular decoding, $\lambda = 1.0$, is also prone to regurgitating contextual information but is not as severe. In particular, both $\lambda = 1.5$ and $\lambda = 1.0$ contain "UFO" in their generations, information likely derived verbatim from the context. On the other hand, $\lambda = 0.5$ does not contain UFO and instead contains "flying vehicle," which is broadly relevant but does not appear verbatim in the context, indicating a strong reliance on parametric knowledge.

CNN-DM

Article	... Luckily, Japanese can sleep soundly in their beds tonight as the government's top military official earnestly revealed that the country's Air Self Defense Force (ASDF) had never encountered an extraterrestrial unidentified flying object . Responding to a query from flamboyant former wrestler-turned-lawmaker Antonio Inoki, Defense Minister Gen Nakatani told the Diet, Japan's parliament, that his jets had, to date, never come across any UFOs from outer space. ...
$\lambda = 1.5$	Japanese can sleep soundly in their beds tonight as the government's top military official earnestly revealed that the country's Air Self Defense Force (ASDF) had never encountered an extraterrestrial unidentified flying object .
$\lambda = 1$	in a interview with Japanese defense minister, politician Antonio Inoki asked the defense minister about aliens and UFOs and the defense minister answered that the Air Self Defense Force (ASDF) has never encountered one.
$\lambda = 0.5$	The article discusses the topic of the possible appearance of aliens and their flying vehicles in the skies over Japan. The author of the article recalls that recently there was a flight of a mysterious object in the sky over Japan.

Table 4: Qualitative examples from LLaMA 3 using different influence levels of CID.