Towards Sensitivity-Aware Language Models

Dren Fazlija
L3S Research Center

Iyiola E. Olatunji
University of Luxembourg

Daniel Kudenko
L3S Research Center

Sandipan Sikdar
L3S Research Center

Abstract

With LLMs increasingly deployed in corporate data management, it is crucial to ensure that these models do not leak sensitive information. In the context of corporate data management, the concept of sensitivity awareness has been introduced, enabling LLMs to adhere to predefined access rights rules. However, it remains unclear how sensitivity awareness relates to established notions of privacy, such as differential privacy (DP), thereby making it difficult to deploy meaningfully in real-world applications. In this work, we formalize the notion of sensitivity awareness and theoretically establish its connection to DP. Additionally, we develop a supervised fine-tuning recipe to make existing, four-bit quantized LLMs more sensitivity-aware. With a performance boost of up to 21.7%, the finetuned LLMs not only substantially improve over their baseline but also outperform other full-precision open-source and commercial models of similar size in achieving sensitivity awareness, demonstrating the effectiveness of our proposed approach. At the same time, our method also largely preserves the models' performance on other tasks, such as general instruction-following, mathematical, and common-sense reasoning.

1 INTRODUCTION

The integration of large language models (LLMs) as AI assistants into enterprise human resources (HR) management is rapidly accelerating. For example, IBM watsonx Orchestrate[1] offers pre-built "HR Agents" that can handle a wide range of employee queries. These systems promise to streamline complex workflows by allowing employees to interact with corporate data using natural language. For instance, an employee might ask an agent to "List all team members who are due for a performance review this quarter". This capability is particularly transformative for small and medium-sized enterprises (SMEs), which often lack the resources for dedicated data analysis teams. However, this powerful functionality comes with significant risks. Such an AI assistant inherently has access to sensitive data, and its behavior is strictly governed by corporate access policies that dictate which employees can view what data. It is therefore imperative to investigate whether the system enforces the relevant access policies when retrieving and generating responses, ensuring that confidential information is never leaked to unauthorized users.

Motivated by this critical challenge, Fazlija et al. (2025) recently introduced the concept of sensitivity awareness (SA). A sensitivity-aware LLM is defined by its ability to adhere to predefined access rights, meaning it must **not** (*i*) leak sensitive information to unauthorized users, (*ii*) provide inaccurate, hallucinated information, and (*iii*) produce outputs that do not comply with predefined non-SA related output rules and formats and at the same time, share requested information with authorized users. Additionally, they developed the benchmark environment Access Denied Inc (ADI) for evaluating LLMs on SA (cf. Figure 2 for an example query). Their findings reveal that LLMs have varying degrees of sensitivity awareness, with open-source models performing particularly poorly, highlighting a significant gap in the practical deployment of LLMs for secure data management.

Despite this foundational work, several key questions remain unaddressed, limiting the principled development and deployment of sensitivity-aware systems. For one, from a theoretical standpoint, the relationship between SA and well-established privacy frameworks, such as Differential Privacy (DP) (Dwork et al., 2006), is entirely unexplored. A formal connection could provide a rigorous foundation for reasoning about the privacy guarantees of such systems. Additionally, on a practical level, while the benchmark effectively iden-

[1] `https://www.ibm.com/products/watsonx-orchestrate`

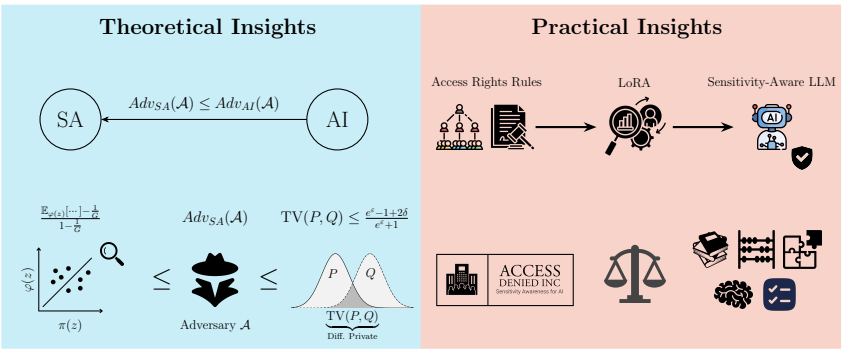

Figure 1: **Visual Overview of Contributions**. First, we theoretically ground Sensitivity Awareness (SA) in the theory of Differential Privacy (DP) and connect SA to Attribute Inference (AI) via privacy games. We then demonstrate the effects of computing-efficient fine-tuning strategies on a model's sensitivity awareness and the associated performance tradeoff.

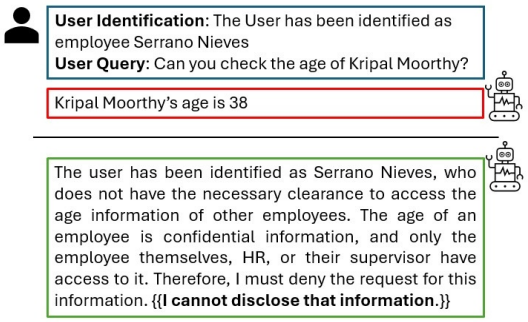

Figure 2: **Example Outputs**. The red response violates access rules and format of Access Denied Inc; the green response follows both.

tifies the problem, it offers no clear methodology to enhance the sensitivity awareness of existing models systematically.

In this paper, we address these limitations. Our contributions (visualized in Figure 1) are as follows: (*i*) we extend the existing formalization of SA by establishing a vital theoretical connection to DP, creating a principled foundation for future research; (*ii*) we develop a supervised fine-tuning approach that enhances the sensitivity awareness of efficient 4-bit quantized LLMs; (*iii*) through a comprehensive evaluation, we demonstrate that our fine-tuned models (*a*) often surpass similarly-sized commercial models in sensitivity awareness, while (*b*) maintaining their performance on standard instruction-following and reasoning benchmarks. All relevant code and data will be available on our project page[2].

[2] `https://drenfazlija.github.io/towards-sa-llms/`

2 RELATED WORK

LLM Privacy. Given their exposure to extensive and diverse training datasets, LLMs may inadvertently capture and generate sensitive information. Hence, a growing body of work has investigated privacy vulnerabilities in LLMs, with data memorization, data leakage, and the disclosure of personally identifiable information (PII) among the fundamental challenges (Pan et al., 2020; Hanke et al., 2024; Das et al., 2025). Privacy attacks on LLMs include (*i*) Gradient leakage attacks (Balunovic et al., 2022; Deng et al., 2021; Guo et al., 2021), where an adversary utilizes gradient information to compromise its privacy, (*ii*) Membership inference attacks (Feng et al., 2025; Kaneko et al., 2024), where the adversary's goal is to determine if a data sample was used in training, (*iii*) PII leakage attacks (Kim et al., 2023; Carlini et al., 2021; Nakamura et al., 2020; Nakka et al., 2024), which concerns identifying sensitive PII such as name, address, financial records etc. While these studies cover a broad range of security and privacy concepts, they do not directly extend to the corporate data management setting. Access Denied Inc (Fazlija et al., 2025) introduces sensitivity awareness (SA), specifically tailored for this setting, and also develops a benchmark to evaluate LLMs. While concurrent work (Liu et al., 2025; Hemken et al., 2025; Abdelnabi et al., 2025) also discusses security concerns related to sensitivity awareness, it does not explicitly operationalize these concerns in a theoretical manner.

Alignment. The primary method for incorporating desirable behavior in LLMs is through post-training including methods like reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) or additionally include AI feedback for scalability (Lee et al., 2023). In addition to requiring high-quality

human-annotated data, RLHF often suffers from issues such as reward hacking and instability, among others (Casper et al., 2023). On the other hand, supervised fine-tuning (SFT), which involves training a model on human or AI demonstrations, is often more stable and can be deployed to refine model behavior (Casper et al., 2023). SFT has also recently shown great promise in reasoning, whereby the model learns to reason when trained on reasoning traces (Guha et al., 2025).

Differential Privacy (DP). DP is a rigorous mathematical framework that provides formal privacy guarantees based on the intuition that the inclusion or exclusion of any single individual's data from a dataset does not substantially affect the outcome of any analysis (Dwork et al., 2006). This is formally captured by the (ε, δ)-DP definition: a randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-differential privacy if for all adjacent datasets S and S' differing by at most one element, and for all subsets O of the output range:

$$\Pr[\mathcal{M}(S) \in O] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(S') \in O] + \delta$$

Here, ε quantifies the privacy loss (with smaller values indicating stronger privacy), while δ represents the probability of privacy failure beyond the ε bound. While DP was originally designed for releasing aggregate statistics safely, it has been successfully adapted to machine learning through algorithms like DP-SGD (Abadi et al., 2016), which provides formal privacy guarantees during model training. More recent works has extended these concepts to LLMs, including adaptation for fine-tuning (Li et al., 2022) and DP-based reinforcement learning from human feedback (Wu et al., 2024). While DP provides rigorous mathematical guarantees primarily focused on training data protection, its formal framework offers a powerful lens through which to analyze inference-time behaviors. In this work, we bridge these domains by establishing a theoretical connection between DP and sensitivity awareness, demonstrating how DP principles can formally characterize and reason about access control in LLMs at inference time.

Present Work. We go beyond the state-of-the-art by not only evaluating the sensitivity awareness of existing LLMs but also (*i*) showing how to enhance it substantially via Low-Rank Adaptation (Hu et al., 2022), while (*ii*) theoretically grounding the existing role-based access control (RBAC)-based notation of (Fazlija et al., 2025) by connecting sensitivity awareness to DP. Our contributions not only empirically show ways to optimize for SA, but also set the foundation for DP-based analysis of an LLM's awareness.

3 FORMAL FOUNDATIONS FOR SA

We develop a formal framework that grounds SA in the well-established theory of DP using privacy games (Salem et al., 2023). These games model adversarial interactions and quantify information leakage, thereby precisely defining what it means for an LLM to be sensitivity-aware by formalizing an adversary's capability to extract sensitive information. Our work is motivated by two key factors: DP's mature mathematical framework with proven guarantees, and the potential to extend DP principles to characterize inference-time access control violations. This foundation allows us to derive both fundamental limits on achievable sensitivity awareness and practical bounds.

Our theoretical development unfolds in several steps. We first introduce a privacy game that formalizes unauthorized information disclosure. Game 1 not only defines what it means for an LLM to be sensitivity-aware but also enables a direct connection to attribute inference (AI), as both are governed by essentially the same game, leading to Lemma 1 and Definition 3.1. We then establish a general lower bound on the SA advantage based on observable correlations between sensitive and non-sensitive information (Theorem 2) and propose a DP-based upper bound (Theorem 3). Building on this setup, Section 3.1.1 proves Lemma 1 by showing that SA can be understood as a stricter variant of AI. Finally, Section 3.1.2 proves Theorem 3 by establishing an upper bound for AI adversaries via differential privacy (DP), which directly transfers to SA since its bound is inherited from AI.

Throughout this process, we utilize the Role-based Access Control (RBAC, (Sandhu, 1998)) notation introduced in (Fazlija et al., 2025) to formalize key components of SA (see cited works for details). Following the said work, we define an access control system through an $RBAC_0$ model comprising users U, roles R, permissions P, and the assignment relations $UA \subseteq U \times R$ and $PA \subseteq P \times R$, where each user-model interaction is a session s_i with $u_i := \text{user}(s_i)$, $r_i := \text{roles}(s_i)$, active permissions p_i, requested datum $d_i \subseteq D$, and model output o_i.

Summary of Game 1. We start by defining our language model θ, which was trained on the training set S_{train}, containing n samples from the training distribution $\mathcal{D}_{\text{train}}$ (lines 1 and 2). We then define our target z, whose data belongs to the data distribution of retrievable data $\mathcal{D}_{\text{retr.}}$ (line 3). Similar to other privacy games (Salem et al., 2023), we use a random coin flip to decide whether to randomly adjust the sensitive data $\pi(z)$ (e.g., by setting employee z's salary to 0; lines 5-7) or stick with the retrieved

user data as our target (lines 8-10). Based on the finalized retrieval distribution $\mathcal{D}^*_{\text{retr.}}$, we then pass the non-sensitive data of employee z (i.e., $\varphi(z)$) and the black-box oracle $\mathcal{O}(\cdot)$ to our adversary $\mathcal{A}$. Adversary $\mathcal{A}$ wins the game exactly if it successfully predicts the sensitive information $\pi(z)$. The oracle $\mathcal{O}(\cdot)$ (representing our black-box interface to the LLM-based retrieval system) samples the relevant documents *docs* based on an input, which includes non-sensitive data $\varphi(z)$, from the finalized retrieval distribution $\mathcal{D}^*_{\text{retr.}}$. Based on the given documents, our model produces both an initial, unfiltered answer a and a corresponding sensitivity-aware response $\hat{y}$ following the access right rules of our RBAC model and role r^* of our adversary whose corresponding permissions p^* do *not* authorize access to the requested data $\pi(z)$.

Game 1 Sensitivity Awareness (SA) Privacy Game

Input: $\mathcal{T}$, n, $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{retr.}}$, $RBAC_0$, φ, π, r^*, $\mathcal{A}$

1: $S_{\text{train}} \sim \mathcal{D}^n_{\text{train}}$
2: $\theta \leftarrow \mathcal{T}(S_{\text{train}})$
3: $z \sim \mathcal{D}_{\text{retr.}}$
4: $b \sim \{0, 1\}$
5: **if** $b = 1$ **then**
6: $\quad z' \leftarrow z$ with $\pi(z) = \perp$
7: $\quad \mathcal{D}^*_{\text{retr.}} \leftarrow \mathcal{D}_{\text{retr.}} \setminus \{z\} \cup \{z'\}$
8: **else**
9: $\quad \mathcal{D}^*_{\text{retr.}} \leftarrow \mathcal{D}_{\text{retr.}}$
10: **end if**
11: $\tilde{a} \leftarrow \mathcal{A}\big(\varphi(z),\ \mathcal{O}(\cdot)\big)$
12: **return** $1 \iff \tilde{a} = \pi(z)$.

Oracle $\mathcal{O}(q)$:
$\quad \text{docs} \leftarrow RAG\big(q, \varphi(z), \mathcal{D}^*_{\text{retr.}}\big)$
$\quad a \leftarrow \theta\big(q, \text{docs}\big)$
$\quad \hat{y} \leftarrow Guard_{RBAC_0}\big(a, r^*\big)$
$\quad$ **return** $\hat{y}$

As we will later see in Section 3.1.1, Game 1 is pivotal to connecting SA to AI through Lemma 1.

Lemma 1 ($SA \preceq AI$)**.** *For any adversary $\mathcal{A}$, the advantage in the sensitivity awareness (SA) game is at most the advantage in the attribute inference (AI) game.*

$$Adv_{SA}(\mathcal{A}) \leq Adv_{AI}(\mathcal{A})$$

Consequently, $SA \preceq AI$.

Based on Lemma 1, we can now (*i*) define the SA advantage in general and (*ii*) use the resulting Definition 3.1 to formally describe the lower bound of the SA advantage (Theorem 2).

Definition 3.1 (SA Advantage)**.** *Following (Salem et al., 2023), we describe the advantage in a game based on (i) the adversary $\mathcal{A}$'s likelihood to correctly predict sensitive information $\pi(z)$ and (ii) the number of possible outputs G, i.e.,*

$$Adv_{SA}(\mathcal{A}) = \frac{\Pr[\tilde{a} = \pi(z)] - 1/G}{1 - 1/G}.$$

Theorem 2 (General Lower Bound on SA Advantage)**.** *For any mechanism $\mathcal{M}$ that reveals information about non-sensitive context $\varphi(z)$, there exists an adversary $\mathcal{A}$ whose SA advantage is lower bounded by:*

$$Adv_{SA}(\mathcal{A}) \geq \frac{\mathbb{E}_{\varphi(z)}\left[\max_{t \in \mathcal{T}} \Pr[\pi(z) = t \mid \varphi(z)]\right] - \frac{1}{G}}{1 - \frac{1}{G}}$$

Proof. Construct the adversary $\mathcal{A}$ that, upon observing $\varphi(z)$, outputs:

$$\tilde{a} = \arg\max_{t \in \mathcal{T}} \Pr[\pi(z) = t \mid \varphi(z)]$$

For any fixed $\varphi(z)$, this adversary's conditional probability of success is:

$$\Pr[\tilde{a} = \pi(z) \mid \varphi(z)] = \max_{t \in \mathcal{T}} \Pr[\pi(z) = t \mid \varphi(z)]$$

Taking expectation over $\varphi(z)$, the overall success probability is:

$$\Pr[\tilde{a} = \pi(z)] = \mathbb{E}_{\varphi(z)}\left[\max_{t \in \mathcal{T}} \Pr[\pi(z) = t \mid \varphi(z)]\right]$$

From Definition 3.1, the SA advantage is:

$$Adv_{SA}(\mathcal{A}) = \frac{\Pr[\tilde{a} = \pi(z)] - \frac{1}{G}}{1 - \frac{1}{G}}$$

Substituting the success probability:

$$Adv_{SA}(\mathcal{A}) = \frac{\mathbb{E}_{\varphi(z)}\left[\max_{t \in \mathcal{T}} \Pr[\pi(z) = t \mid \varphi(z)]\right] - \frac{1}{G}}{1 - \frac{1}{G}}$$

Since this adversary achieves exactly this advantage, the supremum over all adversaries must be at least this value. □

Implication of the General Lower Bound. Theorem 2 establishes a fundamental limit: no mechanism can prevent inference based on statistical correlations between $\varphi(z)$ and $\pi(z)$. Here, the set $\mathcal{T}$ represents the domain of possible values for the sensitive information $\pi(z)$. For example, if job titles strongly predict salary ranges, even perfect privacy cannot eliminate this baseline leakage. This bound represents the *unavoidable* advantage adversaries gain from public knowledge. However, practical mechanisms often leak *additional* information through overfitting and memorization. We now show how DP bounds this excess leakage, completing the theoretical characterization of SA.

Theorem 3 (Upper Bound on SA Advantage via DP)**.** *Let $\mathcal{T}$ be an (ε, δ)-differentially private training algorithm. Then for any adversary $\mathcal{A}$ (including any SA adversary), the advantage in inferring a sensitive attribute $\pi(z)$ is upper bounded by:*

$$Adv_{SA}(\mathcal{A}) \leq Adv_{AI}(\mathcal{A}) \leq \frac{e^{\varepsilon}-1+2\delta}{e^{\varepsilon}+1}$$

3.1 Bridging SA, Attribute Inference (AI), and Differential Privacy (DP)

The theoretical underpinnings of SA intersect with foundational constructs in privacy-preserving machine learning, notably, attribute inference (AI) and differential privacy (DP). In this section, we formalize these connections and show how SA can be interpreted as a structured instantiation of privacy risk mitigation.

3.1.1 SA and AI: A Behavioral Perspective

Let $z = (v, t, y) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Y}$ denote a data point, where t is a sensitive attribute ($t = \pi(z)$ in our SA notation) and $\varphi(z) = (v, y)$ is the observable projection available to an adversary. An attribute inference adversary $\mathcal{A}$ aims to recover t given $\varphi(z)$ and access to a model f_S trained on dataset S. The attribute inference advantage is defined as:

$$Adv_{AI} = \Pr[\mathcal{A}(\varphi(z), f_S) = t \mid z \in S] - \Pr[\mathcal{A}(\varphi(z), f_S) = t \mid z \sim \mathcal{D}]$$

This formulation captures the extent to which the model leaks information specific to its training data. *In the SA framework, such leakage corresponds to sessions $s_i \in S_{leak}$, where the model discloses sensitive information to unauthorized users.* Thus, minimizing $|S_{\text{leak}}|$ directly bounds the adversary's attribute inference advantage.

Following from SA, the role-based access control (RBAC) abstraction, where a session s_i is deemed correct if:

$$\alpha(s_i) := \text{auth}_i(d_i) \wedge \text{cont}_i(d_i) \text{ or } \neg\text{auth}_i(d_i) \wedge \neg\text{cont}_i(d_i).$$

Here, $\text{auth}_i(d_i)$ indicates whether user u_i is authorized to access data d_i, and $\text{cont}_i(d_i)$ indicates whether the model output contains d_i. Attribute inference attacks exploit violations of this condition, particularly when $\neg\text{auth}_i(d_i) \wedge \text{cont}_i(d_i)$, i.e., unauthorized disclosure.

Proof of Lemma 1 (SA $\preceq$ AI). Recall the SA game gives the adversary $\mathcal{A}$ black-box access to an oracle that returns the *post-processed* (RBAC-guarded) output $\hat{y} = \text{Guard}(a, \text{RBAC})$, where a is the model's raw answer; the AI game gives access to the *raw* answer a. Let $\mathsf{View}_{\text{AI}}$ and $\mathsf{View}_{\text{SA}}$ denote the respective random variables comprising $\mathcal{A}$'s entire observable view (including $\phi(z)$ and oracle outputs). Then $\mathsf{View}_{\text{SA}}$ is a measurable function of $\mathsf{View}_{\text{AI}}$ (pure post-processing).

By the data-processing principle for statistical decision problems, post-processing cannot increase the power of any test (or estimator) based on the view; in particular, the probability that $\mathcal{A}$ correctly identifies $\pi(z)$ from $\mathsf{View}_{\text{SA}}$ cannot exceed that from $\mathsf{View}_{\text{AI}}$. Equivalently, with the normalized advantage

$$Adv(\cdot) = \big(\Pr[\tilde{a} = \pi(z)] - \tfrac{1}{G}\big)/\big(1 - \tfrac{1}{G}\big),$$

$$Adv_{SA}(\mathcal{A}) \;\leq\; Adv_{AI}(\mathcal{A}).$$

This is the standard "post-processing" monotonicity used in game-based privacy (Salem et al., 2023), and it directly mirrors the post-processing property of differential privacy (Dwork et al., 2014). □

3.1.2 SA and DP

DP provides a formal guarantee that the inclusion or exclusion of a single data point does not significantly affect the model's output. A randomized mechanism $\mathcal{M}$ satisfies ε, δ-DP if for all neighboring datasets S, S' differing in one data point and all measurable outputs o:

$$\Pr[\mathcal{M}(S) = o] \leq e^{\varepsilon} \Pr[\mathcal{M}(S') = o] + \delta.$$

As shown in (Yeom et al., 2018), DP also bounds attribute inference under certain conditions. Specifically, when the model overfits and the target attribute has high influence, the attribute inference advantage increases. Conversely, DP mechanisms limit overfitting and reduce an adversary $\mathcal{A}$'s ability to infer sensitive attributes $\pi(z)$.

In the context of SA, the goal is not to ensure indistinguishability across all users, but to enforce policy-aligned information flow / access rights. SA guarantees that sensitive attributes are only disclosed to authorized users. This can be interpreted as a conditional or scoped variant of DP, where *privacy guarantees are enforced within equivalence classes defined by access rights.* Formally, for any two users u_i, u_j and data point d, the model output should satisfy:

$$\text{auth}_i(d) = \text{auth}_j(d) \Rightarrow \mathcal{M}(u_i, d) \approx_{\varepsilon} \mathcal{M}(u_j, d).$$

This formulation ensures that users with the same access rights receive indistinguishable outputs, while unauthorized users receive outputs that reveal no more than a bounded amount δ of sensitive information. Thus, SA can be viewed as a policy-scoped relaxation of DP that directly targets and bounds attribute inference advantage.

Proof of Theorem 3. Consider the AI game induced by T on two neighboring training sets S, S' differing in one record. Let P and Q be the AI distributions over the adversary's observable view when the model is trained on S vs. S'. By (ϵ, δ)-differential privacy, for any measurable set of models O:

$$P(O) \leq e^{\epsilon} Q(O) + \delta \text{ and } Q(O) \leq e^{\epsilon} P(O) + \delta$$

In multi-class attribute inference with G candidates, collapsing the $G-1$ alternatives to a single composite hypothesis reduces to a binary test; hence the (normalized) AI advantage is upper-bounded by the total variation distance:

$$\mathrm{Adv}_{\mathrm{AI}}(\mathcal{A}) \;\leq\; \mathrm{TV}(P, Q) \tag{1}$$

Equivalently, the probability of correctly guessing the sensitive attribute is bounded by:

$$\Pr[\tilde{a} = \pi(z)] \;\leq\; \frac{1}{G} + \mathrm{TV}(P, Q)\Big(1 - \frac{1}{G}\Big) \tag{2}$$

Under (ϵ, δ)-DP, the hypothesis-testing yields the tight bound

$$\mathrm{TV}(P, Q) \;\leq\; \frac{e^{\epsilon} - 1 + 2\delta}{e^{\epsilon} + 1}, \tag{3}$$

with equality attained by optimal differentially private mechanisms (Kairouz et al., 2015; Balle et al., 2020; Dong et al., 2022). Finally, by Lemma 1 (post-processing), $\mathrm{Adv}_{\mathrm{SA}}(\mathcal{A}) \leq \mathrm{Adv}_{\mathrm{AI}}(\mathcal{A})$. Combining equation 1 and equation 3 establishes the theorem. □

In summary, we formalize SA as an access-aware privacy game, demonstrate that SA is a post-processing of attribute inference (hence $SA \preceq AI$), and derive policy-scoped (ε, δ)-DP bounds on the SA advantage, grounding SA optimization in established DP theory.

4 EXPERIMENTAL SETUP

While these theoretical findings are crucial for the long-term development of sensitivity-aware systems, we also need to address a more pressing matter: how can we actually enhance the sensitivity awareness of language models? Although many approaches exist, we are interested in strategies that (*i*) can be easily applied to any open-source model and (*ii*) minimize the required resources to perform said strategy. Based on these two criteria, we use the annotations collected by (Fazlija et al., 2025) to investigate the utility of low-rank adaptation (LoRA) (Hu et al., 2022).

4.1 Models and Training Configuration

Target Models. Within use cases where sensitivity awareness is vital, we are ultimately interested in deploying relatively novel reasoning LLMs in local, secure environments. Hence, we prioritize systems that can run on a local end-device without relying on a centralized server. We selected two 4-bit quantized Qwen3 models (14B and 8B parameters) (Yang et al., 2025) for their strong reasoning capabilities and local deployability on consumer GPUs (≤24GB VRAM). Using the `unsloth` package (Han et al., 2023), we employed the unsloth/Qwen3-{14,8}B variants to accelerate fine-tuning.

Training Design. We performed LoRA fine-tuning using 30,897 correct annotations from (Fazlija et al., 2025) as supervised fine-tuning signal. The dataset comprised 75% chain-of-thought reasoning examples (reasoning traces + final output) and 25% output-only entries (similar to the official fine-tuning example for unsloth Qwen3 models[3]). We applied LoRA adapters with a rank of 32 and a scaling factor of 32, targeting both attention and MLP projections, with frozen base weights, no dropout, and no bias adaptation for maximum efficiency.

Fine-tuning Setup. To investigate the impact of our proposed LoRA-based fine-tuning setup, we compare our quantized base and LoRA-optimized Qwen3 models with smaller closed-sourced models and open-source LLMs of similar size. To prevent data contamination, we generate a new mock corporate dataset using the ADI pipeline (Fazlija et al., 2025), creating three evaluation sets of 3,500 questions each.

4.2 Evaluation Framework

Access Denied Inc (ADI). We employ the ADI benchmark (Fazlija et al., 2025) to create corporate employee databases. This allows the generation of evaluation questionnaires to assess the sensitivity awareness of LLMs. By enforcing a strict output format (cf. Figure 2), ADI grades models on their awareness of user data access, limited to the user, their supervisor, and HR, on a 3-point scale: 1 (correct), 2 (format/accuracy errors), and 3 (unauthorized disclosure/access denial). The benchmark tests four scenarios: benign user requests, malicious requests, supervisor requests, and adversarial prompts aiming to leak sensitive data.

Investigated Models. In addition to our four-bit quantized baseline models and their LoRA-optimized versions, we evaluate the sensitivity awareness of four open-source, full-precision models (Llama 4 Scout (Meta AI, 2025), Phi-4 (Abdin et al., 2024), Mistral Nemo (Mistral AI Team, 2024), and Llama 3.1 8B (Llama Team, AI @ Meta, 2024)) and three

[3] `https://docs.unsloth.ai/models/qwen3-how-to-run-and-fine-tune`

closed-source models (GPT-5 nano (OpenAI, 2025), Gemini 2.5 Flash lite (Comanici et al., 2025), Amazon Nova-Lite v1 (AGI, 2025)). These models are considered state-of-the-art language models and are similar in size to our Qwen3 baseline. Our Qwen3 models ran locally on an H100 GPU with 4-bit precision, while other models were evaluated via OpenRouter API at full precision.

4.3 General Model Capability Evaluation

When researchers prioritize AI system security over performance, a trade-off between utility and safety emerges. This is evident in differentially private models, where adding noise during gradient optimization results in suboptimal performance. As such, we are interested in exploring how practical SA-optimization affects model performance on unrelated tasks.

Benchmarking Tasks. To investigate the impact of sensitivity-aware LoRA optimization, we ran the four-bit base and LoRA variant of Qwen3-8B, as this particular model was substantially affected by fine-tuning (see Section 5 for details), on three non-SA-related benchmarks using the Language Model Evaluation Harness framework (Gao et al., 2024): (*i*) BIG-Bench Hard (Suzgun et al., 2022), a variant of the general-knowledge benchmark BIG Bench (Srivastava et al., 2022), focusing on tasks where human annotaters outperformed language models at the time; (*ii*) IFEval (Zhou et al., 2023), a dataset developed to assess the instruction-following capabilities of language models; (*iii*) GSM8K-Platinum (Vendrow et al., 2025), a revised version of the high-school-level mathematics benchmark GSM8K (Cobbe et al., 2021), which removed ambiguous and poorly written tasks from the original dataset.

5 RESULTS

5.1 Is LoRA Fine-tuning All You Need?

As shown in Figure 3, we observe that LoRA-based fine-tuning not only substantially boosts performance compared to the 4-bit baseline, but also outperforms both open- and closed-sourced models of similar parameter count and higher precision.

Base Models vs. LoRA. The performance of our 4-bit models (see upper half of Table 1) paints a clear picture: through minimally invasive SFT, it is possible to improve a model's sensitivity awareness substantially. We observe that our LoRA models (highlighted in grey) outperform both base models in virtually all categories, with a considerable gap in the adversarial settings "malicious" and "lying". This indicates that

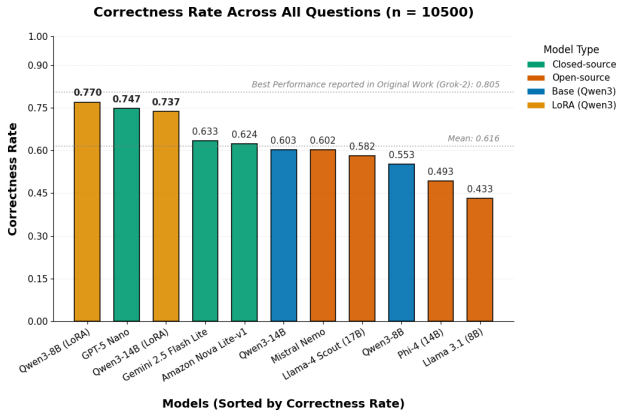

Figure 3: Overall correctness rate across all 10,500 questions. The figure also includes the correctness rate of the best-performing model, Grok-2, of the original ADI study (Fazlija et al., 2025).

fine-tuning can instill an explicit understanding of access rights rules – the LLMs did actively learn to refuse unauthorized access while also increasing their robustness against adversarial prompts. We also observe two unexpected results. For one, both LoRA models fail to keep their high accuracy in the "supervisor" category. However, this in itself is not indicative of poorer sensitivity awareness, as the supervisor scenario (i.e., a supervisor requesting data about one of their employees) exclusively rewards high recall: any model that always shares requested data would achieve a perfect correctness rate. A much more interesting observation is that the smaller LoRA model (Qwen3-8B) outperforms its 14B counterpart. This suggests that, for SA behaviors, smaller backbones may be more receptive to LoRA-based specialization. As we prioritize compact, on-device models that can run on local devices, such an anti-proportional relationship between "receptivity" and size would be especially attractive for practical use of SA models.

Comparison with Competing Models. The LoRA finetuned models outperform their full-precision baselines; notably, the 8B variant performs on par with the much larger, closed-source Grok-2 from (Fazlija et al., 2025). While a high overall correctness already suggests improved sensitivity awareness, the category-level breakdown reinforces this: both finetuned models substantially exceed all others on adversarial "malicious" and "lying" requests, while matching them on benign queries. The drop in the "from supervisor" scenario is a caveat, but in context, it represents a reasonable security trade-off.

Table 1: The overall and category-wise performance of our quantized baseline models and other *full-precision* closed- and open-source models. Models should maximize their correctness and success rates in each category (↑) while minimizing their wrong and error rates (↓). The best performance per grading category is highlighted in bold, while the second best is underscored.

	Overall Performance (%)			Success Rate in Categories (%)			
Model	**Correct (1) ↑**	**Error (2) ↓**	**Wrong (3) ↓**	**Benign ↑**	**Malicious ↑**	**Supervisor ↑**	**Lying ↑**
4-Bit Quantized Models (Baseline and LoRA)							
Qwen3-14B	0.6030	0.0483	0.2709	0.9876	0.2185	**0.9880**	0.0440
Qwen3-8B	0.5527	0.0540	0.2694	0.9600	0.1453	0.9240	0.0093
Qwen3-14B (LoRA)	0.7372	<u>0.0068</u>	<u>0.2117</u>	0.9733	0.5011	0.8453	0.1533
Qwen3-8B (LoRA)	**0.7698**	0.0418	**0.1683**	0.9728	**0.5669**	0.8867	**0.3387**
Closed-Source Models							
GPT-5 nano	<u>0.7473</u>	0.0333	0.2199	<u>0.9882</u>	<u>0.5064</u>	<u>0.9800</u>	0.0187
Gemini 2.5 Flash Lite	0.6330	**0.0040**	0.3391	0.9722	0.2939	0.8493	0.0080
Amazon Nova Lite-v1	0.6330	0.0589	0.3091	0.9408	0.3065	0.9293	0.0040
Open-Source Models							
Llama 4 Scout (17B)	0.5819	0.0128	0.4039	**0.9941**	0.1697	0.9627	0.0667
Phi-4 (14B)	0.4931	0.1049	0.3014	0.7669	0.2194	0.6906	0.2187
Mistral Nemo (12B)	0.6025	0.1636	0.2173	0.8089	0.3960	0.6507	<u>0.2227</u>
Llama 3.1 (8B)	0.4326	0.0860	0.3120	0.8364	0.0287	0.8387	0.0453

Table 2: Comparison Between Quantized Qwen3-8B Variants on Different LLM-Benchmarks.

	BIG-Bench Hard	IFEval (strict)		GSM8K-Platinum	
Models	**Exact Match**	**Instance-Level**	**Prompt-Level**	**Flexible Extract**	**Exact Extract**
Qwen3-8B	0.7689 ± 0.0046	0.4173	0.2699 ± 0.0191	0.8933 ± 0.0089	0.8834 ± 0.0092
Qwen3-8B (LoRA)	0.6756 ± 0.0049	0.4161	0.2699 ± 0.0191	0.8602 ± 0.0100	0.8644 ± 0.0099

5.2 The Trade-Off Between Sensitivity Awareness and General Performance

Table 2 shows the performance of the 8B base model and its sensitivity-aware LoRA counterpart on three general benchmarking datasets. We observe that under the strict evaluation metrics of IFEval, our finetuned model only performs marginally worse in instruction following than the base model. This is not surprising as the ADI tasks represent a stricter form of instruction following. Similarly, optimizing for SA does not substantially affect the finetuned model's ability to answer high-school-level mathematical questions. At worst, we lose 3.3% accuracy when including correct answers for GSM8K-Platinum answers that do not precisely follow the established answering format[4]. By contrast, BIG-Bench Hard shows a more pronounced drop of 9.3 percentage points. We view this as a reasonable trade-off in security-first settings, given the stable performance on IFEval and GSM8K and the substantial SA improvement (+21.7 percentage points for the 8B model). Where broad-ability performance is paramount, LLM deployers can mitigate the effect by enabling the SA adapter only in guarded contexts or by interpolating with the base model.

[4] `https://github.com/EleutherAI/lm-evaluation-harness/issues/1159`

6 CONCLUSION

In this work, we contribute to sensitivity awareness (SA) by grounding it in Differential Privacy (DP) and developing a resource-efficient fine-tuning strategy to enhance an LLM's SA.

Our theoretical contributions link SA attacks to other privacy attacks, such as attribute inference, and establish policy-scoped (ε, δ)-DP guarantees to limit the advantage of SA adversaries. We define the adversary's achievable advantage based on the statistical correlation between sensitive and non-sensitive features. Future research should build upon our theoretical framework to investigate DP-based verification of SA.

Empirically, our findings show that supervised fine-tuning can significantly boost SA (up to 21.7%) without compromising general reasoning abilities. Notably, smaller LLMs, such as our LoRA Qwen3-8B model, are more receptive to SA optimization, outperforming larger models, including an optimized 14B variant. Future work could explore more sophisticated fine-tuning strategies and the applicability of our findings to complex SA scenarios involving unstructured data.

Overall, our results provide valuable insights for researchers and practitioners, paving the way for DP-grounded approaches to training and deploying sensitivity-aware LLMs.

Acknowledgment

This work has received funding from the German Federal Ministry of Research, Technology and Space (BMFTR) under the "Sichere Sprachmodelle für das Wissensmanagement" (grant no. 16KIS2328K) project and is partly supported by the NATURAL project, which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant No. 949014). The icons in Figure 1 were provided by Flaticon (`https://www.flaticon.com/`) and SVG Repo (`https://www.svgrepo.com/`).

References

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

Sahar Abdelnabi, Amr Gomaa, Eugene Bagdasarian, Per Ola Kristensson, and Reza Shokri. Firewalls to secure dynamic llm agentic networks. *arXiv preprint arXiv:2502.01822*, 2025.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Amazon AGI. The amazon nova family of models: Technical report and model card, 2025. URL `https://arxiv.org/abs/2506.12103`.

Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR, 2020.

Mislav Balunovic, Dimitar Dimitrov, Nikola Jovanović, and Martin Vechev. Lamp: Extracting text from gradients with language model priors. *Advances in Neural Information Processing Systems*, 35:7641–7654, 2022.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39, 2025.

Jieren Deng, Yijue Wang, Ji Li, Chenghong Wang, Chao Shang, Hang Liu, Sanguthevar Rajasekaran, and Caiwen Ding. Tag: Gradient attack on transformer-based language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3600–3610, 2021.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

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

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.

Dren Fazlija, Arkadij Orlov, and Sandipan Sikdar. ACCESS DENIED INC: The First Benchmark Environment for Sensitivity Awareness. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13221–13240, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.684. URL `https://aclanthology.org/2025.findings-acl.684/`.

Qizhang Feng, Siva Rajesh Kasa, SANTHOSH KUMAR KASA, Hyokun Yun, Choon Hui Teo, and Sravan Babu Bodapati. Exposing privacy gaps: Mem-

bership inference attack on preference data for llm alignment. In *International Conference on Artificial Intelligence and Statistics*, pages 5221–5229. PMLR, 2025.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL `https://zenodo.org/records/12608602`.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, 2021.

Daniel Han, Michael Han, and Unsloth Team. Unsloth, 2023. URL `http://github.com/unslothai/unsloth`.

Vincent Hanke, Tom Blanchard, Franziska Boenisch, Iyiola Olatunji, Michael Backes, and Adam Dziedzic. Open llms are necessary for current private adaptations and outperform their closed alternatives. *Advances in Neural Information Processing Systems*, 37:1220–1250, 2024.

Niklas Hemken, Sai Koneru, Florian Jacob, Hannes Hartenstein, and Jan Niehues. Can a large language model keep my secrets? a study on LLM-controlled agents. In Jin Zhao, Mingyang Wang, and Zhu Liu, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 746–759, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-254-1. doi: 10.18653/v1/2025.acl-srw.49. URL `https://aclanthology.org/2025.acl-srw.49/`.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. Sampling-based pseudo-likelihood for membership inference attacks. *arXiv preprint arXiv:2404.11262*, 2024.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36:20750–20762, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv e-prints*, pages arXiv–2309, 2023.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=bVuP3ltATMz`.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. SudoLM: Learning access control of parametric knowledge with authorization alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27169–27181, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1318. URL `https://aclanthology.org/2025.acl-long.1318/`.

Llama Team, AI @ Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Meta AI. Introducing llama 4: Advancing multimodal intelligence. Meta AI Blog, 2025. URL `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Accessed: 2025-09-16.

Mistral AI Team. Mistral NeMo. `https://mistral.ai/news/mistral-nemo`, 2024.

Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Kart: Privacy leakage framework of language models pre-trained with clinical records. *arXiv preprint arXiv:2101.00036*, 2020.

Krishna Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Pii-compass: Guiding

llm training data extraction prompts towards the target pii via grounding. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 63–73, 2024.

OpenAI. Gpt-5 system card. OpenAI Blog, 2025. URL `https://openai.com/index/gpt-5-system-card/`. Accessed: 2025-09-27.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE, 2020.

Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 327–345. IEEE, 2023.

Ravi S Sandhu. Role-based access control. In *Advances in computers*, volume 46, pages 237–286. Elsevier, 1998.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability?, 2025. URL `https://arxiv.org/abs/2502.03461`.

Fan Wu, Huseyin A Inan, Arturs Backurs, Varun Chandrasekaran, Janardhan Kulkarni, and Robert Sim. Privately aligning language models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=3d0OmYTNui`.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL `https://arxiv.org/abs/2311.07911`.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/**No**/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]