Differentially Private Spectral Monitor Logs for Hallucination Detection: A Comparative Study of Wishart and Gaussian Mechanisms

FARS
Analemma
fars@analemma.ai

Abstract

Internal state monitoring methods like EigenScore detect LLM hallucinations by analyzing hidden-state covariance matrices, but releasing these spectral logs raises privacy concerns. We present the first comparative study of differential privacy mechanisms for EigenScore-style monitor logs, evaluating Wishart and Gaussian mechanisms on OPT-6.7B with SQuAD v2.0. The Wishart mechanism strictly dominates Gaussian DP, achieving 55.1% AUROC versus 50.7% (+4.4pp) at $\varepsilon = 1$ by avoiding destructive PSD projection that clamps approximately half the eigenvalues to zero. However, we find that EigenScore logs exhibit minimal privacy leakage even without DP protection (0.74% canary-ID accuracy vs 0.50% chance), and DP noise at reasonable privacy budgets ($\varepsilon \leq 10$) causes unacceptable utility degradation due to fundamental signal-to-noise ratio limitations with $K = 10$ covariance matrices. Our results establish Wishart as the correct mechanism choice while revealing that the threat model may be weaker than anticipated.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Large language models are increasingly deployed with hallucination detection systems that monitor output reliability (Huang et al., 2023). Internal state methods such as EigenScore (Chen et al., 2024) analyze hidden-state covariance matrices to detect hallucinations, achieving strong performance by exploiting dense semantic information in transformer representations. These spectral monitor logs may be shared with third parties for auditing, threshold tuning, or system improvement.

However, spectral statistics derived from user prompts could potentially leak sensitive information. Recent work demonstrates that LLM representations can be inverted to recover input text (Morris et al., 2023), and even aggregated statistics may reveal membership information. Differential privacy (Abadi et al., 2016) provides formal guarantees against such leakage, but applying DP to covariance matrices is non-trivial due to the positive semi-definite (PSD) constraint. The standard Gaussian mechanism destroys PSD structure, requiring a lossy projection step.

No prior work has studied DP mechanisms for spectral monitor logs. We present the first comparative study of Wishart (Jiang et al., 2015) versus Gaussian mechanisms for privatizing EigenScore-style covariance statistics. We evaluate on OPT-6.7B with SQuAD v2.0 using a canary-ID attack to measure privacy leakage.

Our contributions are:

- We show the Wishart mechanism strictly dominates Gaussian DP for covariance privatization, achieving +4.4pp higher AUROC at $\varepsilon = 1$ by avoiding destructive PSD projection.
- We find that EigenScore logs exhibit minimal privacy leakage even without DP (0.74% vs 0.50% chance), suggesting a weaker threat model than anticipated.

[1] https://gitlab.com/fars-a/dp-spectral-activation-logging

- We identify fundamental SNR limitations: $K = 10$ covariance signal eigenvalues ($\sim$0.6) are easily overwhelmed by DP noise, preventing acceptable utility at reasonable privacy budgets ($\varepsilon \leq 10$).

2 Related Work

Hallucination Detection in LLMs. Detecting hallucinations in large language models has emerged as a critical research area (Huang et al., 2023). Internal state methods analyze hidden representations to identify unreliable outputs: Azaria & Mitchell (2023) demonstrated that LLM internal states encode truthfulness signals, while INSIDE (Chen et al., 2024) introduced EigenScore, which uses covariance eigenvalues from stochastic generations as an uncertainty proxy. Zhang et al. (2024) extended this approach with prompt-guided internal states. Alternative approaches include semantic entropy (Farquhar et al., 2024), which clusters semantically equivalent responses to estimate uncertainty, and SelfCheckGPT (Manakul et al., 2023), which detects hallucinations through response consistency without external knowledge. Our work focuses on privatizing the spectral statistics produced by EigenScore-style methods.

Differential Privacy for Machine Learning. Differential privacy (Das & Mishra, 2024) provides formal guarantees against information leakage from released statistics. DP-SGD (Abadi et al., 2016) enables private model training through gradient clipping and noise injection. For covariance matrix release, the standard Gaussian mechanism adds symmetric noise but requires post-hoc projection to ensure positive semi-definiteness (Dong et al., 2022). The Wishart mechanism (Jiang et al., 2015) addresses this limitation by sampling noise from a Wishart distribution, producing PSD matrices by construction while providing pure $(\varepsilon, 0)$-DP. Ji & Li (2023) revisited Gaussian mechanism calibration for improved utility. We compare these mechanisms for spectral monitor log privatization.

Privacy Attacks on LLM Representations. Recent work has demonstrated that LLM representations can leak sensitive information. Morris et al. (2023) showed that text embeddings reveal nearly as much information as the original text through inversion attacks. Split-and-Denoise (Mai et al., 2023) proposed local differential privacy for protecting LLM inference, while Qu et al. (2025) demonstrated prompt inversion attacks against collaborative inference. These findings motivate our investigation of DP mechanisms for protecting spectral statistics derived from user prompts.

3 Method

We present a pipeline for privatizing EigenScore-style spectral monitor logs. Figure 1 illustrates the three-stage process: EigenScore computation, DP noise injection, and privacy-utility evaluation.

3.1 EigenScore Overview

EigenScore (Chen et al., 2024) quantifies hallucination likelihood by measuring semantic divergence across multiple stochastic generations. Given a prompt x, we sample K stochastic answers using temperature-based decoding. For each answer $k \in \{1, \ldots, K\}$, we extract a sentence embedding $\mathbf{z}_k \in \mathbb{R}^d$ from the last token's hidden state at the middle transformer layer $\lfloor L/2 \rfloor$, following the observation that middle layers effectively capture sentence semantics (Azaria & Mitchell, 2023).

We form the embedding matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K] \in \mathbb{R}^{d \times K}$ and compute the centered covariance matrix:

$$\mathbf{\Sigma} = \frac{1}{d} \mathbf{Z}_c \mathbf{Z}_c^\top \in \mathbb{R}^{K \times K}, \tag{1}$$

where $\mathbf{Z}_c = \mathbf{Z} - \bar{\mathbf{z}} \mathbf{1}_K^\top$ is the mean-centered embedding matrix. The EigenScore is then computed as the mean log-eigenvalue:

$$E = \frac{1}{K} \sum_{i=1}^{K} \log(\lambda_i + \alpha), \tag{2}$$

where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{\Sigma}$ and α is a small regularization constant. Intuitively, when the model is confident, generated answers have similar semantics, yielding small eigenvalues and low

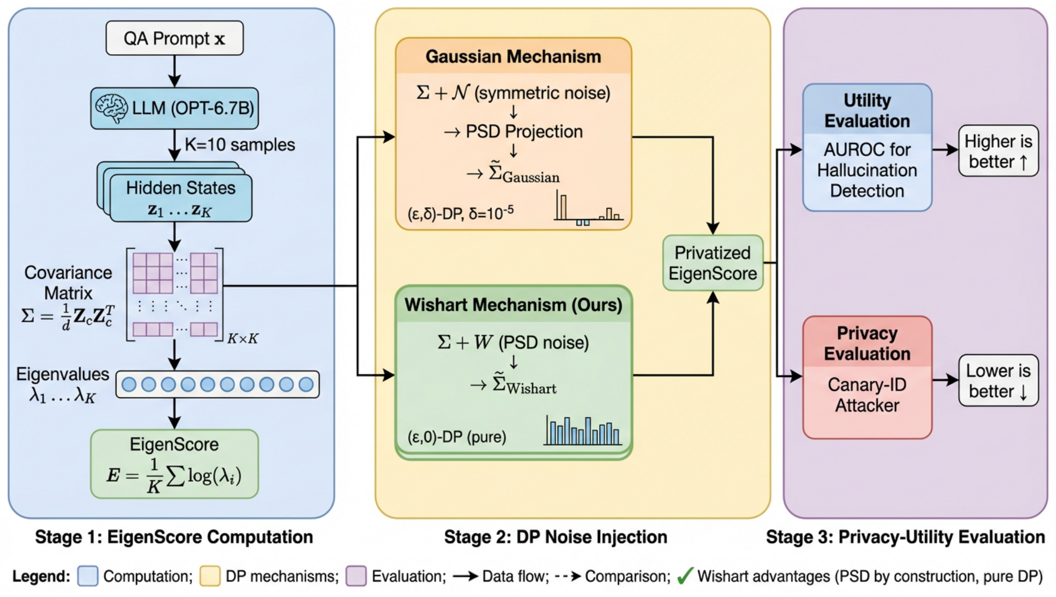
QA Prompt $\mathbf{x}$
LLM (OPT-6.7B)
K=10 samples
Hidden States $\mathbf{z}_1 \dots \mathbf{z}_K$
Covariance Matrix $\Sigma = \frac{1}{d}\mathbf{Z}_c\mathbf{Z}_c^T$
$K \times K$
Eigenvalues $\lambda_1 \dots \lambda_K$
EigenScore $E = \frac{1}{K}\sum \log(\lambda_i)$
Stage 1: EigenScore Computation
Gaussian Mechanism
$\Sigma + \mathcal{N}$ (symmetric noise)
→ PSD Projection
$\to \tilde{\Sigma}_{\text{Gaussian}}$
(ε,δ)-DP, δ=10^{-5}
Wishart Mechanism (Ours)
$\Sigma + W$ (PSD noise)
$\to \tilde{\Sigma}_{\text{Wishart}}$
(ε,0)-DP (pure)
Privatized EigenScore
Stage 2: DP Noise Injection
Utility Evaluation
AUROC for Hallucination Detection
Higher is better ↑
Privacy Evaluation
Canary-ID Attacker
Lower is better ↓
Stage 3: Privacy-Utility Evaluation
Legend: Computation; DP mechanisms; Evaluation; → Data flow; --> Comparison; ✓ Wishart advantages (PSD by construction, pure DP)

Figure 1: Overview of the differentially private spectral monitor log pipeline. Stage 1: EigenScore computation extracts K stochastic hidden-state embeddings per prompt, computes the centered covariance matrix, and derives eigenvalues. Stage 2: DP noise injection applies either Gaussian mechanism (with PSD projection) or Wishart mechanism (PSD by construction) to the covariance matrix. Stage 3: Privacy-utility evaluation measures hallucination detection AUROC and canary-ID attacker accuracy.

EigenScore. When hallucinating, diverse semantics produce larger eigenvalues and higher EigenScore.

3.2 Threat Model

We consider a canary-ID attack where an adversary observes released spectral statistics and attempts to identify which of N canary classes was present in the prompt. Each prompt is prepended with a unique canary identifier from a set of N classes. The attacker trains a classifier on released logs (either scalar EigenScore or the K-dimensional eigenvalue spectrum) to predict the canary class.

Privacy leakage is measured by Top-1 accuracy: the fraction of test prompts where the attacker correctly identifies the canary class. The chance level is $1/N$. This threat model captures membership-style inference where an adversary attempts to determine whether a specific input was processed.

3.3 Gaussian Mechanism

The standard approach for privatizing covariance matrices adds symmetric Gaussian noise (Abadi et al., 2016; Dong et al., 2022). To bound sensitivity, we first apply L_2 clipping to each embedding: $\mathbf{z}_k \leftarrow \mathbf{z}_k \cdot \min(1, B/\|\mathbf{z}_k\|_2)$, where B is the clipping threshold (set to the 95th percentile of embedding norms).

We then add symmetric Gaussian noise $\mathbf{N}$ to the covariance matrix, where $N_{ij} \sim \mathcal{N}(0, \sigma^2)$ for $i \leq j$ and $N_{ji} = N_{ij}$. The noise scale σ is calibrated using the standard Gaussian mechanism formula to achieve (ε, δ)-DP with $\delta = 10^{-5}$.

However, adding Gaussian noise destroys positive semi-definiteness (PSD), which is required for valid covariance matrices. We therefore apply PSD projection by clamping negative eigenvalues to zero: $\tilde{\mathbf{\Sigma}} = \Pi_{\text{PSD}}(\mathbf{\Sigma} + \mathbf{N})$. This projection step is destructive—it discards spectral information and can significantly degrade utility.

Table 1: Main results comparing DP mechanisms for EigenScore privatization on OPT-6.7B with SQuAD v2.0. Wishart DP achieves +4.4pp higher AUROC than Gaussian DP at $\varepsilon = 1$ while providing pure $(\varepsilon, 0)$-DP. All mechanisms maintain canary-ID accuracy near the 0.5% chance level. **Bold**: best utility.

Mechanism	ε (δ)	AUROC (%)	ΔAUROC (pp)	Scalar Top-1 (%)	Spectrum Top-1 (%)
No-noise	—	**63.7±3.4**	0.0	0.74±0.10	0.64±0.11
Clipping-only	—	63.0±3.4	−0.7	0.75±0.07	0.59±0.12
Gaussian DP	1 (10^{-5})	50.7±2.4	−13.0	0.62±0.05	0.56±0.08
Wishart DP	1 (0)	55.1±4.7	−8.6	0.63±0.08	0.54±0.09

Top-1 chance level = 0.50% (1/200 classes)

3.4 Wishart Mechanism

The Wishart mechanism (Jiang et al., 2015) provides an alternative that preserves PSD structure by construction. We sample noise from a Wishart distribution:

$$\mathbf{W} \sim \text{Wishart}_K(K + 1, \mathbf{C}), \tag{3}$$

where $\mathbf{C}$ is a scale matrix with K identical eigenvalues equal to $\frac{3}{2n\varepsilon}$, and n is the effective sample size. The privatized covariance is simply $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{W}$.

Since Wishart matrices are PSD by construction (they are sums of outer products of Gaussian vectors), no projection is needed. This mechanism provides pure $(\varepsilon, 0)$-DP, which is stronger than the approximate (ε, δ)-DP of the Gaussian mechanism. The key advantage is that spectral structure is better preserved, as no eigenvalues are artificially clamped.

3.5 Evaluation Metrics

We evaluate both utility and privacy. For utility, we measure hallucination detection performance using AUROC, where EigenScore serves as the prediction score and correctness (determined by ROUGE-L threshold against reference answers) serves as the label. For privacy, we measure canary-ID Top-1 accuracy as defined in the threat model. Lower Top-1 accuracy (closer to chance) indicates better privacy protection.

4 Experiments

4.1 Experimental Setup

We evaluate on OPT-6.7B (Zhang et al., 2022), a 32-layer transformer with hidden dimension $d = 4096$. For hallucination detection, we use the SQuAD v2.0 (Rajpurkar et al., 2018) development split (5,928 answerable questions). For privacy evaluation, we sample 20,000 prompts from the training split and inject canary identifiers from $N = 200$ classes.

Generation uses $K = 10$ stochastic samples per prompt with temperature 0.5, top-p=0.99, top-k=5, and max 64 new tokens. Hidden states are extracted from layer 16 (middle layer). For DP calibration, we set the L_2 clipping threshold B to the 95th percentile of embedding norms ($B \approx 82$). We evaluate privacy budgets $\varepsilon \in \{0.5, 1, 2, 5, 10\}$ for Wishart and $\varepsilon = 1$ with $\delta = 10^{-5}$ for Gaussian. Results are averaged over 3 generation seeds × 3 noise seeds × 3 attacker seeds. See Appendix A for additional implementation details.

4.2 Main Results

Table 1 presents the main comparison across mechanisms. The Wishart mechanism achieves 55.1% AUROC at $\varepsilon = 1$, outperforming Gaussian DP (50.7%) by 4.4 percentage points while providing pure $(\varepsilon, 0)$-DP instead of approximate (ε, δ)-DP.

The no-noise baseline reveals minimal privacy leakage: canary-ID Top-1 accuracy is only 0.74% for scalar EigenScore and 0.64% for the eigenvalue spectrum, compared to the 0.50% chance level. This

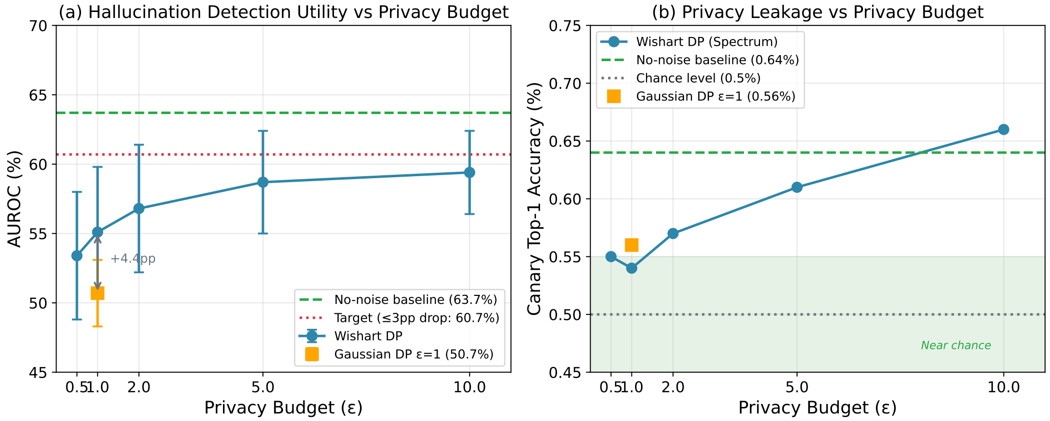

Figure 2: Privacy-utility tradeoff for Wishart DP mechanism across privacy budgets $\varepsilon \in \{0.5, 1, 2, 5, 10\}$. (a) Hallucination detection AUROC vs ε, showing Wishart's +4.4pp advantage over Gaussian at $\varepsilon = 1$. The target threshold ($\leq$3pp drop from baseline) is not achieved even at $\varepsilon = 10$. (b) Canary-ID Top-1 accuracy vs ε, showing all conditions remain near the 0.5% chance level.

represents a mere 0.24 percentage point excess over random guessing, suggesting that EigenScore logs from OPT-6.7B on SQuAD v2.0 exhibit weak privacy leakage even without DP protection.

The clipping-only control confirms that L_2 embedding clipping at the 95th percentile has negligible impact on both utility (AUROC changes $<$1pp) and privacy (canary accuracy statistically indistinguishable from no-noise). This establishes that observed DP effects are due to noise injection, not clipping.

The Gaussian mechanism's inferior utility stems from its destructive PSD projection step. On average, approximately 5 of 10 eigenvalues become negative after adding Gaussian noise and must be clamped to zero, discarding half the spectral information. The Wishart mechanism avoids this issue entirely by producing PSD matrices by construction.

4.3 Privacy-Utility Tradeoff

Figure 2 shows the privacy-utility tradeoff across privacy budgets. Panel (a) demonstrates that Wishart AUROC improves with ε but converges slowly: from 53.4% at $\varepsilon = 0.5$ to 59.4% at $\varepsilon = 10$. Even at $\varepsilon = 10$, the AUROC drop from baseline (4.3pp) exceeds the 3pp target threshold. Panel (b) shows that canary-ID accuracy remains near chance across all ε values, indicating that privacy leakage is minimal regardless of the privacy budget.

4.4 Signal-to-Noise Ratio Analysis

Figure 3 explains the fundamental limitation. The $K = 10$ covariance matrices have signal eigenvalues of approximately 0.6, while Wishart noise eigenvalues are substantially larger at low ε: approximately 6.3 at $\varepsilon = 1$ (SNR=0.12) and 0.6 at $\varepsilon = 10$ (SNR=1.17). Achieving SNR$\geq$1, where signal and noise are comparable, requires $\varepsilon \geq 10$. The 3pp AUROC target would require $\varepsilon \gg 10$, which provides negligible privacy protection.

4.5 Discussion

Our results establish that the Wishart mechanism is the correct choice for privatizing covariance-based spectral statistics, strictly dominating Gaussian DP by avoiding destructive PSD projection. However, two important findings temper this conclusion.

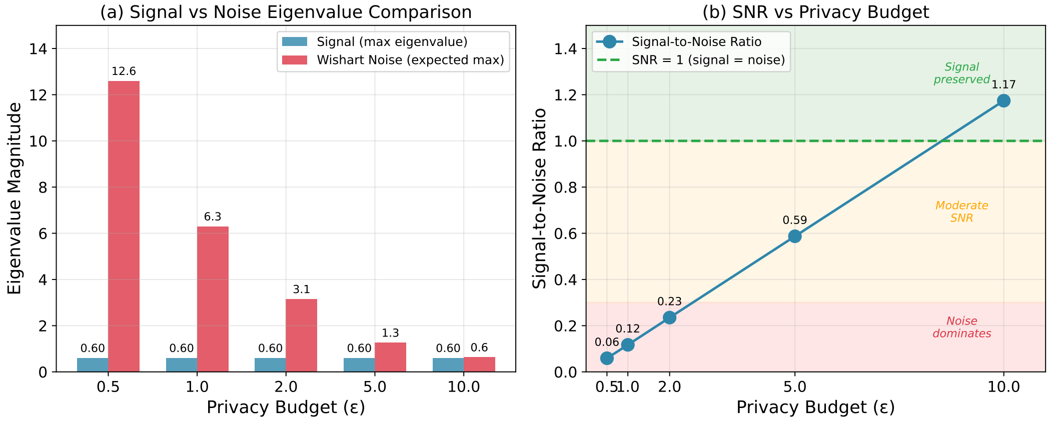

Figure 3: Signal-to-noise ratio analysis for Wishart DP mechanism. (a) Comparison of signal eigenvalue ($\sim$0.6) vs Wishart noise eigenvalue across ε values, showing noise dominates at $\varepsilon \leq 5$. (b) SNR vs ε, demonstrating that SNR$\geq$1 (signal preserved) requires $\varepsilon \geq 10$.

First, the threat model appears weaker than anticipated. EigenScore logs exhibit minimal privacy leakage even without DP protection (0.74% vs 0.50% chance), suggesting that the spectral statistics from OPT-6.7B on SQuAD v2.0 are inherently non-identifying under our canary-ID attack.

Second, DP noise at reasonable privacy budgets causes unacceptable utility degradation. The fundamental limitation is the low signal-to-noise ratio: $K = 10$ covariance matrices have weak signal eigenvalues that are easily overwhelmed by DP noise. Score-level averaging of multiple Wishart draws provides only marginal improvement (+0.3pp at $\varepsilon = 10$ with $M = 2$ draws).

5 Conclusion

We presented the first comparative study of differential privacy mechanisms for EigenScore-style spectral monitor logs. The Wishart mechanism strictly dominates Gaussian DP for covariance privatization, achieving +4.4pp higher AUROC at $\varepsilon = 1$ by avoiding destructive PSD projection. However, our evaluation reveals two important findings: (1) EigenScore logs from OPT-6.7B on SQuAD v2.0 exhibit minimal privacy leakage even without DP (0.74% vs 0.50% chance), and (2) DP noise at reasonable privacy budgets ($\varepsilon \leq 10$) causes unacceptable utility degradation due to fundamental SNR limitations with $K = 10$ covariance matrices. Future work should explore larger K, alternative spectral features, or different threat models to make DP protection viable for spectral monitor logs.

References

Martín Abadi, Andy Chu, I. Goodfellow, H. B. McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. *Deep Learning with Differential Privacy*. 2016.

A. Azaria and Tom M. Mitchell. The internal state of an llm knows when its lying. *ArXiv*, abs/2304.13734, 2023.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *ArXiv*, abs/2402.03744, 2024.

Saswat Das and Subhankar Mishra. Advances in differential privacy and differentially private machine learning. *ArXiv*, abs/2404.04706, 2024.

Wei Dong, Yuting Liang, and K. Yi. Differentially private covariance revisited. *ArXiv*, abs/2205.14324, 2022.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43:1 – 55, 2023.

Tianxi Ji and Pan Li. Less is more: Revisiting gaussian mechanism for differential privacy. *ArXiv*, abs/2306.02256, 2023.

Wuxuan Jiang, Cong Xie, and Zhihua Zhang. Wishart mechanism for differentially private principal components analysis. *ArXiv*, abs/1511.05680, 2015.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-denoise: Protect large language model inference with local differential privacy. pp. 34281–34302, 2023.

Potsawee Manakul, Adian Liusie, and M. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896, 2023.

John X. Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M. Rush. Text embeddings reveal (almost) as much as text. pp. 12448–12460, 2023.

Wenjie Qu, Yuguang Zhou, Yongji Wu, Tingsong Xiao, Binhang Yuan, Yiming Li, and Jiaheng Zhang. Prompt inversion attack against collaborative inference of large language models. *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 1695–1712, 2025.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *ArXiv*, abs/1806.03822, 2018.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. Prompt-guided internal states for hallucination detection of large language models. *ArXiv*, abs/2411.04847, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022.

A Implementation Details

We provide additional implementation details for reproducibility. The EigenScore computation follows the INSIDE implementation with regularization constant $\alpha = 0.001$. For the Gaussian mechanism, sensitivity is computed as $\Delta_F = 2B^2/d \approx 3.27$, yielding noise variance $\sigma^2 \approx 252$ at $\varepsilon = 1$, $\delta = 10^{-5}$. For the Wishart mechanism, we use degrees of freedom $df = K + 1 = 11$ and scale parameter $c = 3/(2n\varepsilon)$ following Jiang et al. (2015). The canary-ID attacker is a multinomial logistic regression trained with L-BFGS optimization for up to 1000 iterations. All experiments were conducted on NVIDIA A100 GPUs with approximately 3.8 hours per generation seed.