DIFFERENTIALLY PRIVATE EIGENSPECTRUM MONITOR LOGS FOR HALLUCINATION DETECTION

FARS
Analemma
fars@analemma.ai

ABSTRACT

LLM monitoring systems that analyze internal hidden states for hallucination detection expose representations that may leak sensitive user information. We investigate whether differential privacy (DP) can protect eigenspectrum-based monitor logs while preserving utility. We compare two DP mechanisms: the standard isotropic Gaussian and the Rank-1 Singular Multivariate Gaussian (R1SMG), which exploits the geometry of high-dimensional queries to achieve dimension-independent noise scaling. At identical privacy budget ($\varepsilon = 5$, $\delta = 10^{-5}$), R1SMG achieves 360$\times$ lower noise than Gaussian and 4.4 AUROC points higher hallucination detection performance (0.536 vs. 0.492). However, both mechanisms fail our pre-registered viability threshold: R1SMG incurs a 13.5-point AUROC drop from the clip-only baseline (0.672), far exceeding the 5-point threshold. Notably, the eigenspectrum compression itself provides substantial inherent privacy—attackers remain at chance level even without DP noise. We conclude that DP-protected eigenspectrum monitoring is not viable at tested privacy budgets with current mechanisms.

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 monitoring systems that analyze internal hidden states for hallucination detection, safety filtering, and quality assurance. These monitoring pipelines extract intermediate representations—such as eigenspectra of hidden state covariance matrices—to estimate generation reliability without requiring external knowledge bases. However, the same internal representations that enable effective monitoring also pose privacy risks: recent work demonstrates that embeddings and hidden states can be inverted to recover substantial portions of input text (Morris et al., 2023), and that language model mappings are theoretically injective and hence invertible (Nikolaou et al., 2025). Even deeper layers, contrary to intuition, do not provide privacy protection (Dong et al., 2025).

Differential privacy (DP) offers a principled framework for protecting sensitive data through calibrated noise addition (Dwork & Roth, 2014). However, applying DP to high-dimensional internal states faces a fundamental challenge: the standard isotropic Gaussian mechanism requires noise with expected norm scaling as $O(\sqrt{M})$ with dimension M, making it impractical for LLM hidden states where $M \approx 40{,}000$. This raises a critical question: can we release DP-protected monitor logs while preserving hallucination detection utility?

We investigate this question by combining two key ideas. First, eigenspectrum compression reduces 40,960-dimensional hidden state matrices to just $K = 10$ eigenvalues, potentially providing inherent privacy through dimensionality reduction. Second, the Rank-1 Singular Multivariate Gaussian (R1SMG) mechanism (Ji & Li, 2023) achieves dimension-independent noise scaling by exploiting the geometry of DP proofs, offering a potential path to practical privacy protection.

Our contributions are:

[1] https://gitlab.com/fars-a/dp-eigenspectrum-monitor-logs

- We provide the first evaluation of differential privacy for eigenspectrum-based hallucination detection, establishing a rigorous privacy-utility tradeoff framework with preregistered decision criteria.
- We demonstrate that R1SMG achieves 360× lower noise than the Gaussian mechanism at identical privacy budget ($\varepsilon = 5$, $\delta = 10^{-5}$), yielding 4.4 AUROC points higher utility.
- We discover that eigenspectrum compression provides substantial inherent privacy: even without DP noise, canary-ID attackers remain at chance level, suggesting the dimensionality reduction itself destroys identifying information.
- We conclude with a negative result: despite R1SMG's dramatic noise reduction, DP-protected eigenspectrum monitoring incurs unacceptable utility loss (13.5-point AUROC drop) at tested privacy budgets, informing practitioners that alternative protection mechanisms are needed.

2 Related Work

Hallucination Detection in LLMs. Detecting hallucinations in large language models has emerged as a critical research area. Black-box approaches such as SelfCheckGPT (Manakul et al., 2023) leverage sampling consistency to identify unreliable generations without access to model internals. Semantic uncertainty methods (Kuhn et al., 2023) cluster semantically equivalent responses to estimate generation confidence, while semantic entropy probes (Kossen et al., 2024) train lightweight classifiers on internal representations. White-box approaches exploit the observation that LLM internal states encode truthfulness signals (Azaria & Mitchell, 2023). INSIDE (Chen et al., 2024) demonstrates that eigenspectrum analysis of hidden state covariance matrices provides effective hallucination detection, motivating our investigation of privacy-preserving variants. Prompt-guided methods (Zhang et al., 2024) further refine internal state analysis through targeted probing.

Privacy Risks in LLM Systems. LLM deployments expose multiple privacy attack surfaces. Text embeddings can be inverted to recover substantial portions of original inputs (Morris et al., 2023), and recent work proves that language model mappings are theoretically injective and hence invertible (Nikolaou et al., 2025). Internal hidden states present even greater risks: Dong et al. (2025) demonstrate that deeper layers, contrary to intuition, do not provide privacy protection and can be inverted to reconstruct prompts. Collaborative inference settings enable prompt inversion attacks (Qu et al., 2025), while training data extraction remains a persistent concern (Ahmed et al., 2026). These findings motivate the need for principled privacy protection in monitoring systems that log internal representations.

Differential Privacy for High-Dimensional Data. Differential privacy (Dwork & Roth, 2014) provides rigorous privacy guarantees through calibrated noise addition. The Gaussian mechanism with analytic calibration (Balle & Wang, 2018) enables tight noise bounds for continuous queries. However, high-dimensional data poses fundamental challenges: isotropic noise scales with dimension, making standard mechanisms impractical for LLM hidden states. Private covariance estimation (Amin et al., 2019; Dong et al., 2022) and the Wishart mechanism for PCA (Jiang et al., 2015) address structured matrix queries but do not directly apply to eigenspectrum monitoring. Our work investigates whether the rank-1 structure of eigenspectrum changes enables more efficient noise calibration.

Private LLM Inference. Recent work addresses privacy in LLM inference through various mechanisms. Split-and-denoise (Mai et al., 2023) applies local differential privacy to intermediate activations in split inference settings. Cascade (Thomas et al., 2025) proposes token sharding across multiple servers to prevent any single party from observing complete sequences. Concept-aware mechanisms (Tsai et al., 2026) defend against embedding inversion by adding semantically-informed noise. Surveys on private transformer inference (Li et al., 2024) and LLM privacy (Miranda et al., 2024) provide comprehensive overviews of the threat landscape. Our work differs by focusing specifically on monitoring systems that log eigenspectrum statistics rather than protecting inference itself.

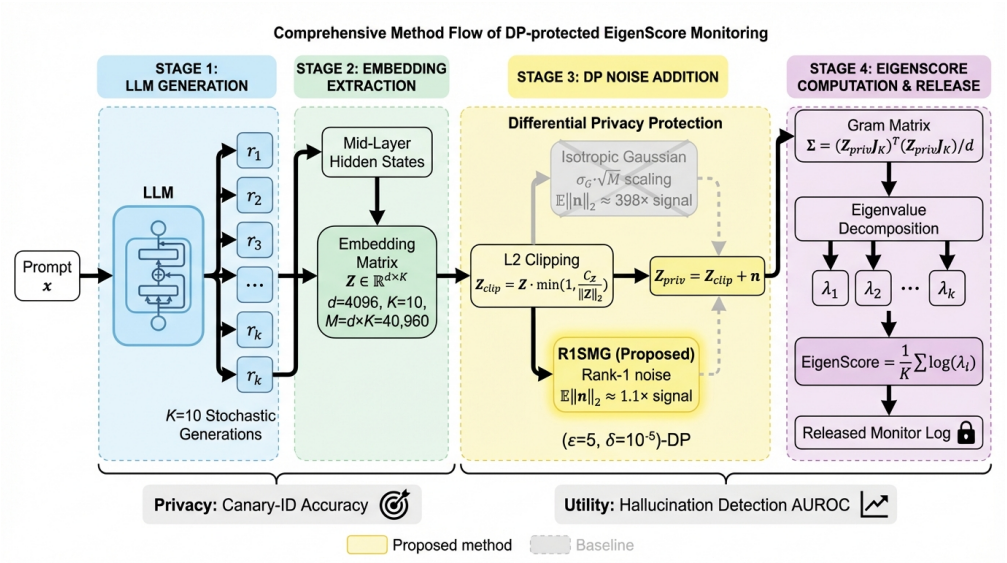

Figure 1: Overview of the DP-protected eigenspectrum monitoring pipeline. Stage 1: Generate K stochastic responses per prompt. Stage 2: Extract hidden states from the middle layer and compute the eigenspectrum. Stage 3: Apply DP noise (R1SMG or Gaussian) to clipped eigenvalues. Stage 4: Compute EigenScore for hallucination detection. R1SMG achieves 360× lower noise than Gaussian at equivalent privacy.

3 METHOD

3.1 PROBLEM SETUP

We consider a deployment scenario where an LLM monitoring system logs internal-state-derived statistics for hallucination detection, debugging, or quality assurance. The monitoring pipeline computes an eigenspectrum from hidden state embeddings across multiple stochastic responses, releasing K eigenvalues per prompt as the monitor log.

Threat Model. We assume an honest-but-curious log consumer who observes the released eigenspectrum logs and attempts to identify users or recover sensitive prompt information. The attacker has access to the DP mechanism parameters (ε, δ, clipping radius C_Z) and can train adaptive classifiers on the defended distribution. This models realistic scenarios where monitoring logs are shared with third parties for auditing or analysis.

Privacy Goal. We aim to protect user identity from canary-ID attacks, where an adversary attempts to link eigenspectrum logs to specific users based on synthetic identifiers embedded in prompts. Success is measured by whether attacker accuracy remains at chance level (Top-1: 0.5%, Top-10: 5.0% for $N = 200$ canary IDs).

3.2 EIGENSPECTRUM MONITORING PIPELINE

Our pipeline, illustrated in Figure 1, consists of four stages. First, for each prompt x, we generate K stochastic responses using temperature sampling. Second, we extract the last-token hidden state $s_k \in \mathbb{R}^d$ from the middle layer for each response $k \in \{1, \ldots, K\}$, forming an embedding matrix $Z = [s_1, \ldots, s_K] \in \mathbb{R}^{d \times K}$. We center across responses using $J_K = I_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ and compute the $K \times K$ Gram matrix $\Sigma = \frac{1}{d}(ZJ_K)^\top(ZJ_K)$. The eigenspectrum $\{\lambda_1, \ldots, \lambda_K\}$ of $\Sigma + \alpha I$ (with regularization $\alpha = 0.001$) captures the diversity of hidden state representations.

Third, we apply DP noise to the clipped embedding vector $\mathrm{vec}(Z)$ before computing the eigenspectrum. Fourth, the EigenScore is computed as $\frac{1}{K}\sum_{i=1}^{K}\log(\lambda_i)$, following INSIDE (Chen et al.,

2024). Lower EigenScore indicates higher confidence (correlated embeddings), while higher values suggest potential hallucination (diverse embeddings).

3.3 Differential Privacy Mechanisms

We enforce per-prompt (ε, δ)-differential privacy (Dwork & Roth, 2014) on the released eigenspectrum. The sensitive object is $f(x) = \text{vec}(Z(x)) \in \mathbb{R}^M$ where $M = d \cdot K$. We apply deterministic ℓ_2 clipping: $Z_{\text{clip}} = Z \cdot \min(1, C_Z / \|\text{vec}(Z)\|_2)$, yielding sensitivity $\Delta_2 = 2C_Z$.

Gaussian Mechanism. The standard approach adds isotropic noise $n \sim \mathcal{N}(0, \sigma_G^2 I_M)$ to $\text{vec}(Z_{\text{clip}})$, with σ_G calibrated via the analytic Gaussian mechanism (Balle & Wang, 2018). However, the expected noise norm $\mathbb{E}[\|n\|_2] = \sigma_G \sqrt{M}$ scales with dimension, making this impractical for high-dimensional hidden states ($M \approx 40{,}000$).

R1SMG Mechanism. The Rank-1 Singular Multivariate Gaussian (R1SMG) mechanism (Ji & Li, 2023) adds rank-1 structured noise: sample v uniformly from the unit sphere $\mathcal{S}^{M-1}$, sample $z \sim \mathcal{N}(0, 1)$, and set $n = v \cdot \sqrt{\sigma^*} \cdot z$. The key insight is that R1SMG exploits the geometry of DP proofs to achieve expected noise norm $\mathbb{E}[\|n\|_2] = \sqrt{\sigma^* \cdot 2/\pi}$, which is independent of dimension M. For our setting with $\varepsilon = 5$ and $\delta = 10^{-5}$, R1SMG achieves noise-to-signal ratio of approximately $1.1\times$ compared to $398\times$ for Gaussian, a $360\times$ reduction.

After adding noise, we reshape the perturbed vector back to $Z_{\text{priv}} \in \mathbb{R}^{d \times K}$ and compute the eigenspectrum. By DP post-processing, any function of Z_{priv} (including the eigenspectrum and EigenScore) inherits the same (ε, δ)-DP guarantee.

3.4 Canary-ID Attack Framework

We evaluate privacy through canary-ID prediction, a membership inference-style attack where the adversary attempts to identify which of N synthetic user IDs generated a given eigenspectrum log. Each prompt is prepended with a unique 6-character alphanumeric canary ID, and outputs containing the canary substring are filtered to ensure any successful prediction must come from the logged eigenspectrum rather than trivial text leakage.

We train three adaptive attacker models on the defended distribution: (1) logistic regression on the K-dimensional eigenspectrum, (2) a 2-layer MLP classifier ($K \to 128 \to N$), and (3) a denoise-then-classify pipeline that first trains a denoiser to estimate clean eigenspectra from noisy observations, then applies an MLP to the denoised features. We report the maximum leakage across all attackers.

3.5 Decision Rule

We pre-register a viability threshold: DP-protected eigenspectrum monitoring is deemed viable if (1) AUROC drops by at most 5 percentage points from the clip-only baseline, and (2) all attackers achieve chance-level accuracy. If the AUROC drop exceeds 5 points, we conclude that DP protection incurs unacceptable utility loss at the tested privacy budget.

4 Experiments

4.1 Experimental Setup

We evaluate DP-protected eigenspectrum monitoring on Canary-SQuAD, a modified version of SQuAD v2.0 (Rajpurkar et al., 2018) where each prompt is prepended with a unique 6-character alphanumeric canary ID. The dataset contains 5,928 answerable questions split into calibration (2,000), training (2,000), validation (500), and test (1,428) sets, with $N = 200$ canary IDs balanced across splits. Outputs containing the canary substring are filtered to ensure any successful identification must come from the eigenspectrum rather than trivial text leakage.

Table 1: Noise regime diagnostic comparing R1SMG and Gaussian mechanisms. R1SMG achieves 358× lower expected noise norm at identical $(\varepsilon = 5, \delta = 10^{-5})$-DP.

Mechanism	σ	$\mathbb{E}[\|n\|_2]$	Noise/Signal	Pilot Prediction
Gaussian DP	60.74	12,293	397.9×	384.8×
R1SMG DP	1,857	**34.4** (358× lower)	**1.1**×	1.08×

Table 2: Main experimental results comparing DP mechanisms for eigenspectrum-based hallucination detection on Canary-SQuAD. Best utility in **bold**. All attackers at chance level (Top-1: 0.50%, Top-10: 5.00%).

Condition	AUROC ↑	PCC	Noise/Signal	Top-1 Acc	Top-10 Acc	Decision
Clip-Only (No DP)	**0.672**±0.010	−0.468	—	0.72%	5.21%	Baseline
Gaussian DP (ε=5)	0.492±0.008 (↓18.0pt)	+0.016	397.9×	0.54%	5.28%	REFUTE
R1SMG DP (ε=5)	0.536±0.009 (↓13.5pt)	−0.044	1.1×	0.50%	5.05%	REFUTE

We use Llama-3.1-8B-Instruct (Dubey et al., 2024) with $K = 10$ stochastic responses per prompt (temperature 0.5, top-p 0.99, top-k 5). Hidden states are extracted from layer 16 (middle of 32 layers), yielding $d = 4{,}096$-dimensional embeddings and $M = d \cdot K = 40{,}960$-dimensional vectorized matrices. The clipping radius $C_Z = 34.05$ is set to the 95th percentile of calibration norms, achieving approximately 5% clip rate. DP parameters are $\varepsilon = 5$ and $\delta = 10^{-5}$ for the main comparison, with multi-epsilon analysis at $\varepsilon \in \{5, 10, 20\}$. All experiments use three generation seeds and three attacker seeds, reporting mean and standard deviation.

4.2 Noise Regime Diagnostic

Before running the full experiment, we conduct a 50-prompt pilot study to validate that R1SMG operates in a non-trivial noise regime. Table 1 compares the noise characteristics of both mechanisms.

The Gaussian mechanism requires noise with expected norm 12,293, approximately 398× the typical signal norm of 30.9. This massive noise-to-signal ratio effectively destroys all utility. In contrast, R1SMG achieves expected noise norm of only 34.4, yielding a noise-to-signal ratio of 1.1×—a 358× reduction. The pilot predictions (1.08× for R1SMG, 384.8× for Gaussian) match the full experiment within 2–3%, validating the diagnostic's reliability.

4.3 Main Results

Table 2 presents the main comparison across three conditions: clip-only (no DP), Gaussian DP, and R1SMG DP.

The clip-only baseline achieves AUROC of 0.672 with strong negative correlation (PCC $= -0.468$) between EigenScore and correctness, confirming that eigenspectrum analysis provides meaningful hallucination detection signal. Notably, even without DP noise, all attackers remain at chance level (Top-1: 0.72% vs. 0.50% chance, Top-10: 5.21% vs. 5.00% chance), demonstrating that the dimensionality reduction from 40,960 to 10 eigenvalues provides substantial inherent privacy.

Gaussian DP destroys utility entirely, with AUROC dropping to 0.492 (below random chance of 0.5 for balanced classes) and PCC collapsing to near-zero. The 398× noise-to-signal ratio overwhelms the signal structure.

R1SMG DP substantially outperforms Gaussian, achieving AUROC of 0.536—4.4 percentage points higher at identical privacy budget. The 360× noise reduction preserves partial signal structure. However, the 13.5-point AUROC drop from baseline exceeds our pre-registered 5-point viability threshold. Both DP mechanisms achieve the privacy goal (attackers at chance), but neither meets the utility requirement. Per our decision rule, we conclude: **REFUTE**—DP-protected eigenspectrum monitoring is not viable at $\varepsilon = 5$.

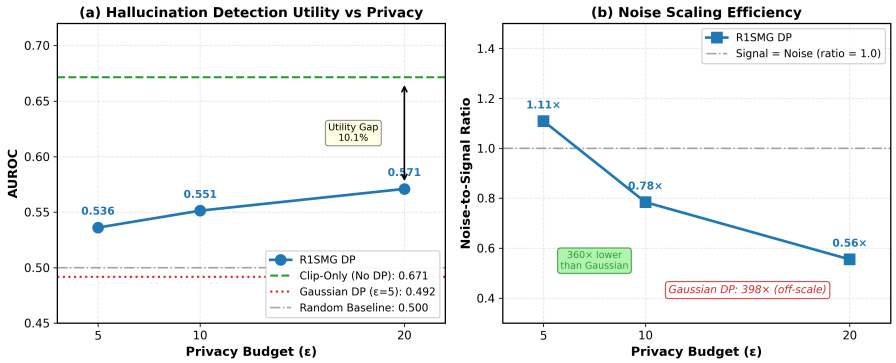

Figure 2: Privacy-utility tradeoff for R1SMG DP across privacy budgets. (a) Hallucination detection AUROC increases with ε but remains 10.1 points below clip-only baseline even at $\varepsilon = 20$. (b) R1SMG achieves noise-to-signal ratio below 1.0 at $\varepsilon \geq 10$, compared to 398$\times$ for Gaussian DP.

4.4 Multi-Epsilon Analysis

Figure 2 shows the privacy-utility tradeoff as we relax the privacy budget. AUROC increases monotonically from 0.536 ($\varepsilon = 5$) to 0.551 ($\varepsilon = 10$) to 0.571 ($\varepsilon = 20$), while noise-to-signal ratio decreases from 1.1$\times$ to 0.78$\times$ to 0.55$\times$. Even at $\varepsilon = 20$ (relatively weak privacy), the utility gap persists at 10.1 percentage points below baseline. Importantly, attacker accuracy remains at chance level across all epsilon values, indicating that the eigenspectrum compression provides inherent privacy independent of DP noise magnitude.

4.5 Analysis

Why does the utility gap persist despite R1SMG's dramatic noise reduction? We hypothesize that the eigenspectrum is particularly sensitive to perturbations in the rank-1 direction. While R1SMG adds noise with magnitude comparable to the signal (ratio ≈ 1), the nonlinear eigenvalue transformation amplifies even moderate perturbations. Small eigenvalues, which carry important information about embedding diversity, are especially vulnerable: a perturbation of magnitude δ to an eigenvalue λ produces relative error δ/λ, which diverges as $\lambda \to 0$.

This suggests that the signal *structure*, not just magnitude, matters for hallucination detection. The eigenspectrum encodes subtle correlations among response embeddings that are disrupted by any noise, even when the noise norm is small relative to the overall signal. Future work might explore alternative DP mechanisms that preserve spectral structure, such as adding noise directly to the Gram matrix or using task-specific noise calibration.

5 Conclusion

We presented the first systematic evaluation of differential privacy for eigenspectrum-based hallucination detection in LLM monitoring systems. Our investigation yields a negative result: at privacy budgets $\varepsilon \leq 20$, DP-protected eigenspectrum monitoring fails to meet our pre-registered viability threshold of $\leq$5-point AUROC degradation. Even R1SMG, which achieves 360$\times$ lower noise than the standard Gaussian mechanism and outperforms it by 4.4 AUROC points, still incurs a 13.5-point drop from the clip-only baseline.

Despite this negative finding, our work makes several positive contributions. First, we demonstrate that eigenspectrum compression provides substantial inherent privacy—all attacker models remain at chance level even without DP noise, suggesting that dimensionality reduction itself offers meaningful protection. Second, our pilot diagnostic framework accurately predicts full-experiment outcomes within 2–3%, enabling efficient mechanism screening. Third, R1SMG's exploitation of rank-1 subspace structure represents a promising direction for high-dimensional DP applications.

Future work should explore alternative approaches: local differential privacy applied at the client before transmission, task-specific noise calibration that preserves hallucination-relevant signal structure, or hybrid mechanisms combining cryptographic and statistical protections.

REFERENCES

Ahmed M. Ahmed, A. F. Cooper, Oluwasanmi Koyejo, and Percy Liang. Extracting books from production language models. *ArXiv*, abs/2601.02671, 2026.

Kareem Amin, Travis Dick, Alex Kulesza, Andrés Muñoz Medina, and Sergei Vassilvitskii. Differentially private covariance estimation. pp. 14190–14199, 2019.

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

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. pp. 403–412, 2018.

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.

Tian Dong, Yan Meng, Shaofeng Li, Guoxing Chen, Zhen Liu, and Haojin Zhu. Depth gives a false sense of privacy: Llm internal states inversion. pp. 1629–1648, 2025.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, A. Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, A. Sravankumar, A. Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aur'elien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, C. Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, E. Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, G. Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, J. V. D. Linde, J. Billock, Jenny Hong, Jenya Lee, J. Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, J. Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Kenneth Heafield, Kevin R. Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, L. Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, M. Muzzi, Ma hesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, M. Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Niko lay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, P. Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, R. Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ron nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, S. Hosseini, Sa hana Chennabasappa, Sanjay Singh, Sean Bell, S. Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, S. Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, S. Collot, Suchin Gururangan, S. Borodinsky, Tamar Herman, T. Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Vir ginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan,

Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, A. Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, A. Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, B. Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto de Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, B. Ni, Braden Hancock, Bram Wasti, Brandon Spence, B. Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, E. Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, F. Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco (Paco) Guzmán, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi Zhang, G. Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, J. Reizenstein, J. Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, J. McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U. KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, K. Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A. Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, M. Bhatt, M. Tsimpoukelli, Martynas Mankus, Matan Hasson, M. Lennie, Matthias Reso, M. Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, M. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, M. Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Mun ish Bansal, N. Santhanam, Natascha Parks, Natasha White, Navy ata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, O. Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, P. Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, P. Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, R. Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, S. Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Zha, S. Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, S. Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, V. Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. 2024.

C. Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, 2014.

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.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Y. Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv*, abs/2302.09664, 2023.

Yang Li, Xinyu Zhou, Yi-Ting Wang, Liangxin Qian, and Jun Zhao. A survey on private transformer inference. *ArXiv*, abs/2412.08145, 2024.

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.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, F. Zanzotto, Sébastien Bratières, and E. Rodolà. Preserving privacy in large language models: A survey on current threats and solutions. *ArXiv*, abs/2408.05212, 2024.

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

Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. Language models are injective and hence invertible, 2025. URL https://arxiv.org/abs/2510.15511.

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.

Rahul Thomas, Louai Zahran, Erica Choi, Akilesh Potti, Micah Goldblum, and Arka Pal. Cascade: Token-sharded private llm inference. *ArXiv*, abs/2507.05228, 2025.

Yu-Che Tsai, Hsiang Hsiao, Kuan-Yu Chen, and Shou-De Lin. Concept-aware privacy mechanisms for defending embedding inversion attacks. 2026.

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.