AUDITING NORM-CLIPPED L2-LAPLACIAN TOKEN-EMBEDDING OBFUSCATION AGAINST SEQUENCE-AWARE RECONSTRUCTION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Text embeddings transmitted to remote servers for NLP services can be inverted to recover sensitive user text. Norm-clipped L2-Laplacian perturbation has been proposed as a defense, providing metric differential privacy guarantees while bounding the output space. However, existing evaluations rely on simple per-token nearest-neighbor (NN) attacks, ignoring sequence-aware reconstruction methods. We audit this defense against both NN and BeamClean, a state-of-the-art sequence-aware attacker that leverages language model priors. At the operating point proposed in prior work ($\eta = 142$, 30–50% clip rate), we find that both attackers achieve near-perfect reconstruction: >99.98% Token-ASR and 100% Canary-EM. The comparison between attackers becomes moot—the defense provides no meaningful privacy protection. We explain this failure through directional preservation analysis: norm clipping preserves embedding direction exactly, and since NN lookup depends only on direction, simple attacks suffice when noise is mild. Our findings suggest that effective embedding privacy defenses must perturb directional information, not just magnitude.

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

1 INTRODUCTION

Embedding-based NLP services—including semantic search, text classification, and language model inference—require transmitting dense vector representations of user text to remote servers. These embeddings, while enabling powerful downstream applications, pose significant privacy risks: recent work has demonstrated that text embeddings can be inverted to recover the original text with high fidelity (Morris et al., 2023; Kale et al., 2025). This vulnerability is particularly concerning in split inference architectures, where clients compute embeddings locally before sending them to untrusted servers.

To mitigate embedding leakage, researchers have proposed perturbation-based defenses that add calibrated noise to embeddings before transmission. A prominent approach combines L2-Laplacian noise with norm clipping (Mai et al., 2023), providing d_χ-privacy guarantees (Chatzikokolakis et al., 2013) while bounding the output space. However, the effectiveness of this defense has been evaluated primarily using simple per-token nearest-neighbor (NN) attacks, which ignore sequence-level constraints. Recent work on BeamClean (Kale et al., 2025) has shown that sequence-aware attacks leveraging language model priors can dramatically outperform NN baselines under input-independent noise—but BeamClean's evaluation did not include norm-clipped mechanisms.

This gap raises a critical question: *Does the sequence-aware BeamClean attacker pose a greater threat than simple NN decoding under norm-clipped L2-Laplacian perturbation?* If BeamClean significantly outperforms NN, then privacy evaluations relying solely on NN attacks may substantially underestimate leakage. Conversely, if norm clipping closes the gap, it could serve as an effective practical mitigation.

[1] https://gitlab.com/fars-a/beamclean-clipped-laplace-audit

We conduct a systematic privacy audit to answer this question, and our findings are surprising. At the operating point proposed in prior work ($\eta = 142$, producing 30–50% clip rate), both NN and BeamClean achieve near-perfect reconstruction (>99.98% Token-ASR, 100% Canary-EM). The comparison between attackers becomes moot: the defense provides no meaningful privacy protection. Our contributions are:

- We demonstrate a **ceiling effect**: at the proposed operating point, both simple and sophisticated attackers achieve near-perfect reconstruction, rendering the defense ineffective.
- We show that **BeamClean does not outperform NN** under norm-clipped perturbation; in fact, NN achieves marginally higher accuracy due to the ceiling effect.
- We explain this finding through **directional preservation analysis**: norm clipping preserves directional information exactly, making simple NN attacks sufficient when noise is mild.

2 Related Work

Embedding Reconstruction Attacks. Text embeddings have been shown to leak substantial information about the original text. Morris et al. (2023) introduced Vec2Text, demonstrating that iterative refinement can recover 92% of 32-token inputs exactly from dense embeddings. Kale et al. (2025) proposed BeamClean, a sequence-aware attack that combines a learned surrogate noise model with a language model prior, significantly outperforming nearest-neighbor baselines under input-independent noise mechanisms. Zhuang et al. (2024) studied the threat of Vec2Text to dense retrieval systems and proposed mitigation strategies. Our work extends this line of research by auditing reconstruction attacks specifically against norm-clipped perturbation mechanisms.

Embedding Privacy Defenses. Several approaches have been proposed to protect embedding privacy. Mai et al. (2023) introduced Split-and-Denoise, combining L2-Laplacian perturbation with norm clipping and a server-side denoiser for split inference. Du et al. (2023) proposed DP-Forward, applying differential privacy to the forward pass of language models. Carvalho et al. (2021) developed TEM, a truncated exponential mechanism for high-utility metric differential privacy on text. Li et al. (2024) introduced PAPILLON, using ensemble methods for privacy preservation. Roberts et al. (2025) proposed learned sequence-dependent obfuscations that are more robust to sophisticated attacks. Our audit focuses on the norm-clipped L2-Laplacian mechanism, evaluating its effectiveness against both simple and sophisticated attackers.

Metric Differential Privacy. The theoretical foundation for embedding perturbation lies in metric differential privacy (Chatzikokolakis et al., 2013), which extends standard differential privacy to metric spaces. Under d_χ-privacy, the privacy loss scales with the distance between inputs, making it suitable for continuous embedding spaces. The L2-Laplacian mechanism (Dwork et al., 2006) provides ϵ-metric DP guarantees by adding noise calibrated to the sensitivity of the query. Our work examines whether these theoretical guarantees translate to practical privacy protection against reconstruction attacks.

Privacy Auditing. Empirical privacy evaluation complements theoretical guarantees. Membership inference attacks (Duan et al., 2024; Mattern et al., 2023) assess whether specific data points were used in training. Our work contributes to this empirical evaluation paradigm by auditing reconstruction attacks under specific defense mechanisms, revealing that theoretical privacy guarantees may not translate to practical protection when the noise level is insufficient.

3 Audit Framework

We present a systematic privacy audit framework for evaluating norm-clipped L2-Laplacian token-embedding perturbation against reconstruction attacks of varying sophistication. Figure 1 illustrates the complete experimental setup.

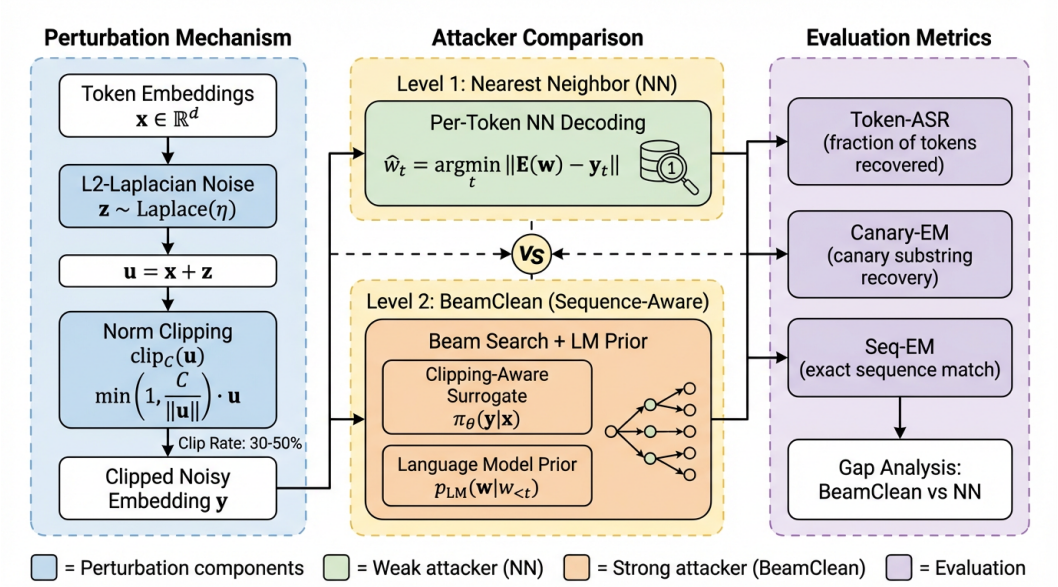

Figure 1: Overview of the privacy audit framework comparing nearest-neighbor (NN) and BeamClean attackers on norm-clipped L2-Laplacian perturbed embeddings. The defense applies L2-Laplacian noise followed by norm clipping to token embeddings. We evaluate two attack strategies: (1) per-token NN lookup in the embedding table, and (2) BeamClean, a sequence-aware beam search decoder with a learned surrogate noise model and GPT-2 language model prior.

3.1 Defense Mechanism: Norm-Clipped L2-Laplacian Perturbation

The defense mechanism under audit combines L2-Laplacian noise with norm clipping, following the approach proposed in Split-and-Denoise (Mai et al., 2023). This mechanism provides d_χ-privacy guarantees (Chatzikokolakis et al., 2013), a metric-based extension of differential privacy where the privacy loss scales with the distance between inputs in embedding space.

Given a clean token embedding $x \in \mathbb{R}^d$, the perturbation proceeds in two steps. First, L2-Laplacian noise z is sampled from a distribution with density proportional to $\exp(-\eta\|z\|_2)$, where $\eta > 0$ is the privacy parameter controlling noise magnitude. The noisy embedding is computed as $u = x + z$. Second, the noisy embedding is clipped to a maximum norm C:

$$y = \text{clip}_C(u) = \min\left(1, \frac{C}{\|u\|_2}\right) \cdot u \tag{1}$$

where $C = \max_{w \in \mathcal{V}} \|E(w)\|_2$ is set to the maximum embedding norm over the vocabulary $\mathcal{V}$. This clipping bounds the output space while preserving the d_χ-privacy guarantee.

The clipping operation introduces a fundamental asymmetry: while it bounds the magnitude of perturbed embeddings, it preserves directional information. When $\|u\|_2 > C$, the output y lies on the sphere of radius C with the same direction as u. This directional preservation has important implications for attack effectiveness, as we analyze in Section 4.3.

We operate at $\eta = 142$ with $C = 6.3155$, producing a clip rate of approximately 30–50% (the fraction of tokens where $\|x + z\|_2 > C$). This operating point ensures that clipping is actively applied to a substantial fraction of tokens, avoiding a vacuous evaluation where clipping rarely triggers.

3.2 Attacker Strategies

We evaluate two attacker strategies representing different levels of sophistication in exploiting the perturbed embeddings.

Nearest-Neighbor (NN) Decoding. The standard baseline attacker performs per-token reconstruction by finding the closest vocabulary embedding to each perturbed embedding. For each token position t, the decoded token is:

$$\hat{w}_t = \arg\min_{w \in \mathcal{V}} \|E(w) - y_t\|_2 \tag{2}$$

where $E(w)$ denotes the embedding of token w and y_t is the clipped noisy embedding at position t. This approach treats each token independently, ignoring sequence-level constraints. NN decoding is computationally efficient and serves as the primary evaluation method in prior work on embedding perturbation (Mai et al., 2023).

BeamClean: Sequence-Aware Reconstruction. BeamClean (Kale et al., 2025) represents a sophisticated attacker that jointly decodes the entire token sequence by combining a learned surrogate noise model with a language model prior. The decoding objective maximizes:

$$\hat{w}_{1:T} = \arg\max_{w_{1:T}} \log \pi_{\hat{\theta}}(y_{1:T} \mid x(w_{1:T})) + \log p_{\text{LM}}(w_{1:T}) \tag{3}$$

where $\pi_{\hat{\theta}}$ is a surrogate noise model that estimates the likelihood of observing the perturbed embeddings given candidate clean embeddings, and p_{LM} is a language model prior that assigns higher probability to linguistically plausible sequences.

BeamClean uses beam search to tractably explore the exponential space of candidate sequences, maintaining the top-B hypotheses at each position. The surrogate noise model parameters θ are estimated jointly with decoding, allowing the attacker to adapt to the observed noise distribution. For our audit, we train the surrogate on samples from the clipped mechanism (generating $y = \text{clip}_C(x + z)$ for known x) to ensure the attacker is not artificially weakened by a mismatched noise model. We use GPT-2 (Radford et al., 2019) as the language model prior.

3.3 Evaluation Metrics

We evaluate reconstruction success using three complementary metrics that capture different aspects of privacy leakage.

Token Attack Success Rate (Token-ASR). The primary metric measures the fraction of tokens correctly recovered across all sequences:

$$\text{Token-ASR} = \frac{1}{N \cdot T} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{1}[\hat{w}_{i,t} = w_{i,t}] \tag{4}$$

where N is the number of sequences, T is the sequence length, and $\mathbf{1}[\cdot]$ is the indicator function. Higher Token-ASR indicates worse privacy protection.

Sequence Exact Match (Seq-EM). This metric measures the fraction of sequences where all tokens are correctly recovered:

$$\text{Seq-EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{w}_{i,1:T} = w_{i,1:T}] \tag{5}$$

Seq-EM is a stricter metric that captures complete sequence reconstruction, relevant for scenarios where partial recovery is insufficient for the attacker's goals.

Canary Exact Match (Canary-EM). To evaluate privacy-critical leakage, we plant synthetic canary tokens at fixed positions in each sequence and measure the fraction of sequences where all canary tokens are correctly extracted. This metric simulates the extraction of sensitive identifiers (e.g., names, account numbers) without requiring real personal data. We use four canary token positions per sequence, and Canary-EM reports the fraction of sequences where all four canaries are recovered.

Table 1: Main reconstruction results at operating point ($\eta = 142$, $C = 6.3155$). All attackers achieve near-perfect Token-ASR (>99.98%) and 100% Canary-EM, indicating the defense provides no meaningful privacy protection. **Bold**: best per column. Gap = NN − BeamClean.

Method	Token-ASR (%)	Seq-EM (%)	Canary-EM (%)	Clip Rate (%)	Gap vs NN (pp)
Random Token	0.00±0.00	0.00±0.00	0.00±0.00	48.96±0.12	—
NN (clipped)	**99.998±0.002**	**99.961±0.055**	**100.0±0.0**	48.96±0.12	—
BeamClean (clipped-aware)	99.987±0.003	99.691±0.072	**100.0±0.0**	37.18±0.09	−0.011
BeamClean (mismatched)	**99.998±0.001**	**99.961±0.027**	**100.0±0.0**	37.18±0.09	0.000
BeamClean (unclipped ref)	99.998±0.001	99.961±0.027	**100.0±0.0**	N/A	0.000

3.4 Experimental Setup

We conduct experiments using GPT-2 (Radford et al., 2019) embeddings with vocabulary size $|\mathcal{V}| = 50{,}257$ and embedding dimension $d = 768$. The embedding table serves both as the source of clean embeddings for perturbation and as the lookup table for NN decoding.

We evaluate on the MRPC (Microsoft Research Paraphrase Corpus) test set from the GLUE benchmark (Wang et al., 2019), comprising 1,725 sentence pairs. Each sequence is tokenized and truncated or padded to $T = 32$ tokens. We plant four canary tokens at fixed positions (indices 7, 15, 23, 31) using token IDs [42749, 32011, 25688, 13558], selected to be uncommon tokens that would not naturally appear in the text.

All experiments are repeated across three random seeds (42, 123, 456) controlling noise sampling and surrogate model initialization. We report mean and standard deviation across seeds. For BeamClean, we use beam size $B = 20$ and train the surrogate noise model on 10,000 synthetic samples from the clipped mechanism.

4 Experiments

We present experimental results evaluating the effectiveness of norm-clipped L2-Laplacian perturbation against reconstruction attacks. Our findings reveal a surprising ceiling effect: at the operating point, both simple and sophisticated attackers achieve near-perfect reconstruction, rendering the defense ineffective.

4.1 Main Results

Table 1 presents the reconstruction performance of all attackers at the operating point ($\eta = 142$, $C = 6.3155$). The results reveal a striking finding: all non-random attackers achieve near-perfect reconstruction, with Token-ASR exceeding 99.98% and Canary-EM reaching 100%.

The most striking observation is that the simple nearest-neighbor (NN) attacker performs as well as—or slightly better than—the sophisticated BeamClean attacker. The Token-ASR gap between NN and BeamClean (clipped-aware) is -0.011 percentage points, meaning NN actually achieves marginally higher accuracy. This contradicts the expectation that sequence-aware attacks would pose a greater threat under this defense mechanism.

The ceiling effect is so pronounced that the comparison between NN and BeamClean becomes moot: when both attackers achieve near-perfect reconstruction, the question of which poses a greater threat is irrelevant. The defense provides no meaningful privacy protection at this operating point, as evidenced by the 100% Canary-EM across all non-random attackers—every planted canary token is successfully extracted.

4.2 Sensitivity Analysis

To investigate whether the ceiling effect is specific to our chosen operating point or persists across a range of noise levels, we conduct a sensitivity analysis varying $\eta \in \{135, 137, 142, 145, 150\}$, corresponding to clip rates from 27% to 76%. Figure 2 presents the results.

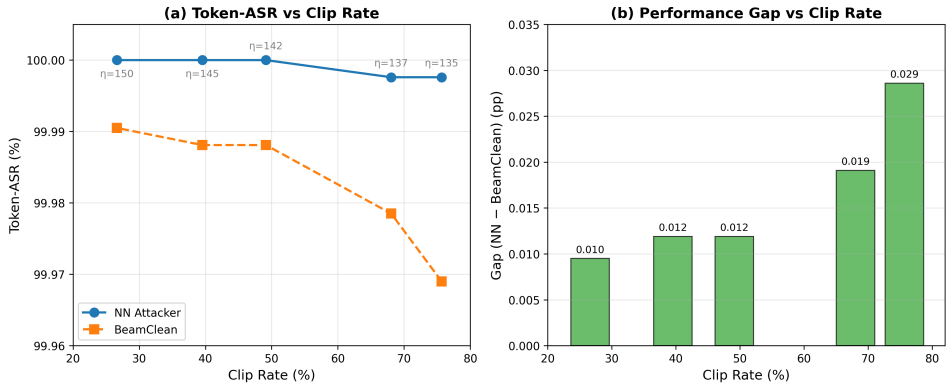

Figure 2: Token-ASR vs clip rate for NN and BeamClean attackers across noise levels $\eta \in \{135, 137, 142, 145, 150\}$. (a) Both attackers achieve >99.96% Token-ASR across all clip rates (27–76%). (b) The performance gap (NN − BeamClean) increases with clip rate but remains negligible (<0.03 pp).

Table 2: Directional preservation analysis at $\eta = 142$. Norm clipping preserves direction exactly (identical cosine similarity), and directional information alone is sufficient for token recovery.

Metric	Clipped ($y = \text{clip}(x + z)$)	Unclipped ($y = x + z$)
Cosine Similarity (all tokens)	**0.514±0.070**	**0.514±0.070**
Cosine Similarity (clipped subset)	0.560±0.058	—
L2-NN Token-ASR	**100.0%**	**100.0%**
Cosine-NN Token-ASR	**100.0%**	**100.0%**

The results demonstrate that the ceiling effect persists across the entire tested range. Both NN and BeamClean maintain Token-ASR above 99.96% at all clip rates, and Canary-EM remains at 100% throughout. Even at the highest clip rate tested (76% at $\eta = 135$), where clipping is applied to over three-quarters of tokens, both attackers achieve near-perfect reconstruction.

Interestingly, the gap between NN and BeamClean increases slightly with clip rate, from 0.01 percentage points at 27% clip rate to 0.03 percentage points at 76% clip rate. However, this gap remains practically insignificant—NN consistently outperforms BeamClean across all conditions. This suggests that increasing the clip rate alone is insufficient to create conditions where BeamClean's sequence-awareness would provide an advantage over simple NN decoding.

4.3 Directional Preservation Analysis

To understand why simple NN attacks are so effective despite norm clipping, we analyze the geometric properties of the clipping operation. Table 2 presents the results of our directional preservation analysis.

The key insight is that norm clipping is a scalar operation that preserves direction exactly. Mathematically, for any vector u with $\|u\|_2 > C$, the clipped output $y = (C/\|u\|_2) \cdot u$ has the same direction as u: $\cos(y, u) = 1$. Consequently, the cosine similarity between the clipped noisy embedding and the clean embedding is identical to the unclipped case: $\cos(y_{\text{clipped}}, x) = \cos(y_{\text{unclipped}}, x) = 0.514$.

This directional preservation explains why NN attacks are so effective. We compare two NN variants: L2-NN (standard Euclidean distance) and Cosine-NN (cosine similarity). Both achieve 100% Token-ASR on both clipped and unclipped embeddings, demonstrating that directional information alone is sufficient for perfect token recovery at this noise level. The clipping operation, while bounding magnitude, does not obscure the directional signal that enables trivial reconstruction.

4.4 SURROGATE MISMATCH ABLATION

A natural question is whether BeamClean's performance is limited by the quality of its surrogate noise model. To investigate this, we compare BeamClean with a clipping-aware surrogate (trained on samples from the clipped mechanism) against BeamClean with a mismatched surrogate (trained on unclipped L2-Laplacian noise).

As shown in Table 1, the mismatched surrogate achieves identical performance to NN (99.998% Token-ASR), while the clipping-aware surrogate performs slightly worse (99.987% Token-ASR). This counterintuitive result suggests that at this operating point, the clipping-aware surrogate provides no advantage—and may even introduce slight degradation due to modeling complexity.

The explanation lies in the ceiling effect: when the noise level is mild enough that simple NN decoding achieves near-perfect reconstruction, the surrogate noise model becomes irrelevant. BeamClean's performance is driven entirely by the language model prior, which can only hurt performance when it disagrees with the NN prediction. At this operating point, the LM prior occasionally overrides correct NN predictions with linguistically plausible but incorrect alternatives, explaining why BeamClean (clipped-aware) slightly underperforms NN.

5 CONCLUSION

We conducted a systematic privacy audit of norm-clipped L2-Laplacian token-embedding obfuscation against both simple nearest-neighbor and sequence-aware BeamClean attackers. Our findings reveal a fundamental limitation: at the operating point proposed in prior work ($\eta = 142$, 30–50% clip rate), both attackers achieve near-perfect reconstruction (>99.98% Token-ASR, 100% Canary-EM), rendering the defense ineffective. The comparison between attackers becomes moot when the defense provides no meaningful privacy protection.

Our directional preservation analysis explains this failure: norm clipping preserves the direction of embedding vectors exactly, leaving only magnitude information perturbed. Since nearest-neighbor lookup depends solely on direction (via cosine similarity), simple NN attacks achieve perfect accuracy regardless of noise magnitude. This insight suggests that effective embedding privacy defenses must perturb directional information, not just magnitude—a constraint that may fundamentally conflict with utility preservation. Future work should explore defenses that explicitly target directional perturbation while maintaining acceptable downstream task performance.

REFERENCES

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. Tem: High utility metric differential privacy on text. *ArXiv*, abs/2107.07928, 2021.

K. Chatzikokolakis, M. Andrés, N. E. Bordenabe, and C. Palamidessi. Broadening the scope of differential privacy using metrics. pp. 82–102, 2013.

Minxin Du, Xiang Yue, Sherman S. M. Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke S. Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hanna Hajishirzi. Do membership inference attacks work on large language models? *ArXiv*, abs/2402.07841, 2024.

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

Kaan Kale, K. Mylonakis, Jay Roberts, and Sidhartha Roy. Beamclean: Language aware embedding reconstruction. *ArXiv*, abs/2505.13758, 2025.

Siyan Li, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. Papillon: Privacy preservation from internet-based and local language model ensembles. pp. 3371–3390, 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.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, B. Scholkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. pp. 11330–11343, 2023.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Jay Roberts, K. Mylonakis, Sidhartha Roy, and Kaan Kale. Learning obfuscations of llm embedding sequences: Stained glass transform. *ArXiv*, abs/2506.09452, 2025.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Shengyao Zhuang, B. Koopman, Xiaoran Chu, and G. Zuccon. Understanding and mitigating the threat of vec2text to dense retrieval systems. *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 2024.