EDEN: Efficient Decoding Methods for Language Models on Encrypted Data

Tara Patel[1], Padmini Saroj[1], Brad Wilmange[2]
[1]Rutgers University
[2]Cohere, Inc

***Abstract*—Inference with large language models (LLMs) over encrypted data promises strong confidentiality for user inputs, intermediate features, and outputs, yet the decoding loop remains the bottleneck for practicality[1]. Even if one assumes the existence of an upstream mechanism that produces logits under privacy-preserving computation—via homomorphic encryption (HE), secure multiparty computation (MPC), functional encryption, or trusted execution—the act of turning a large, encrypted logit vector into the next token by sampling or search typically requires operations that are ill-suited to privacy-preserving substrates: softmax normalization is nonlinear and numerically sensitive; top-k and nucleus filtering involve comparisons and conditional pruning over vocabularies with $V \approx 10^5$; and sampling demands generation of high-quality randomness and inverse-CDF selections without revealing intermediate structure. Naïvely porting plaintext decoders either leaks through access patterns and partial revelations or incurs prohibitive latency through deep Boolean circuits or high-degree polynomial approximations. This paper presents EDEN, a family of decoding methods that preserves statistical fidelity while substantially reducing cryptographic and communication overhead. The design is guided by three principles. First, replace normalization with *lazy* or *relative* schemes that select tokens based on approximate order or truncated neighborhoods, proving bounds on the total variation distance (TV) between sampled distributions and the true softmax while avoiding global divides[2]. Second, exploit *SIMD packing* and structured quantization to compress logits into a few ciphertexts or secret shares so that comparisons and reductions operate in parallel across entire vocabularies, amortizing the cost of bootstrapping or interactive rounds. Third, construct *oblivious sampling* mechanisms that consume correlated randomness established at session setup, enabling inverse transform or alias sampling with constant rounds and without revealing thresholds, indices, or heap shapes. Concretely, EDEN introduces CKKS-based polynomial surrogates for the log-sum-exp with Chebyshev control of approximation error; a two-stage encrypted selection that first identifies a coarse top-$K^\star$ via blockwise maxima in approximate HE, then refines to the exact top-k inside a small secure-comparison enclave or MPC gadget; and an encrypted alias method whose tables are synthesized under encryption and updated incrementally across decoding steps so that sampling costs remain constant in V. We formalize a leakage profile that exposes only the final token and a bounded-time success/failure bit for rejection sampling, and we prove that, under standard IND-CPA or simulation-based MPC assumptions, EDEN does not reveal rank, threshold, or neighborhood membership beyond what is implied by the output token. We analyze asymptotic and concrete costs and show, via model-agnostic derivations, parameter regimes where decoding is dominated by a handful of batched comparisons and low-degree polynomial evaluations rather than full softmaxes and global sorts. The result is a principled path to practical, privacy-preserving LLM decoding that composes with any upstream private inference mechanism and yields end-to-end confidentiality without surrendering quality or introducing brittle numerical heuristics.**

I. Introduction

The decoding loop transforms model logits into text by selecting the next token according to a decision rule such as greedy argmax, temperature-scaled sampling[3], nucleus filtering, or beam search. In privacy-preserving inference, where inputs and intermediate tensors are encrypted or secret-shared, this loop is both the last mile for utility and a primary source of leakage risk and computational overhead. Comparisons, branching, and heap updates are expensive in HE and require interactive rounds in MPC; numeric stability is delicate when approximations replace exponentials and divisions; and partial revelation of ranks or thresholds can leak information even if the final token remains hidden until reveal. The challenge is to design decoders whose primitives map cleanly onto encrypted arithmetic and secure comparison, while preserving output distributions to within provable error and keeping latency competitive with plaintext systems once constant factors from cryptography are amortized.[4]

II. Current Landscape of Research

Existing work on private inference has focused on linear and convolutional layers, low-degree activation approximations, and quantization that reduces HE depth or MPC rounds[5]. CryptoNets and successors demonstrated inference with HE for small networks; Gazelle, Delphi, and related systems hybridized linear layers in HE with non-linear layers in garbled circuits to balance cost[5][3]. Confidential computing enclaves have enabled trusted decoders but inherit a large attack surface and are sensitive to side channels and provisioning. For transformers, recent prototypes move a subset of attention and MLP blocks into HE or MPC, yet most either terminate in plaintext decoding or rely on enclaves for the entire head, which collapses the trust model. Within cryptography, top-k under encryption has been addressed through oblivious selection networks[2], bitonic sorts, or approximate order statistics; however, general-purpose sorting networks scale poorly with vocabulary size and create deep circuits with many bootstraps in FHE or many rounds in MPC. Softmax under encryption is often replaced by polynomial surrogates for exponentials and reciprocal approximations via Newton–Raphson, but global normalization still requires logarithmic-depth reductions and

division. Sampling has received less attention: inverse-CDF sampling needs cumulative sums and binary search on encrypted CDFs, while the alias method demands oblivious table construction and index selection, both expensive to realize generically[6]. Prior LLM-specific systems thus leave a gap at decoding: either one accepts enclave trust, or one tolerates prohibitive overheads and leaky access patterns.

III. Proposed Solution

EDEN decomposes decoding on encrypted logits into three tightly coupled subroutines—lazy normalization, two-stage encrypted selection, and oblivious alias sampling—implemented over packing-friendly homomorphic arithmetic and, where exact comparisons are required, constant-round secure comparison. The design is parameterized by a modulus chain and scaling schedule for approximate HE, a candidate bandwidth $K^\star \ll V$, and a quantization grid that controls sampling error. Throughout, $\boldsymbol{z} \in \mathbb{R}^V$ denotes the plaintext logits, $\hat{\boldsymbol{z}}$ their centered form, V the vocabulary size, k the top-k or the refinement width, and B a known bound on dynamic range after centering. The encrypted representation is a vector of CKKS ciphertexts $\mathbf{ct} = \mathrm{Enc}_{\mathsf{CKKS}}(\boldsymbol{z})$ with ring $R_q = \mathbb{Z}_q[X]/(X^N+1)$, $N = 2^n$, and scale $\Delta = 2^\sigma$.

The first subroutine replaces explicit softmax with lazy normalization. Let $m = \max_i z_i$. Computing m under CKKS is realized as an approximate reduction. Two schemes are supported. The first uses a smoothed maximum via log-sum-exp,

$$\tilde{m}(\tau) = \mu + \tau \log\left(\sum_{i=1}^{V} \exp\left(\frac{z_i - \mu}{\tau}\right)\right), \qquad \mu = \mathrm{mean}(\boldsymbol{z}), \tag{1}$$

where $\tau > 0$ is a temperature for which the approximation error satisfies $0 \le \tilde{m}(\tau) - m \le \tau \log V$. The exponential is approximated on $[-B/\tau, 0]$ by a degree-d Chebyshev series $E_d(x) = \sum_{\ell=0}^{d} c_\ell T_\ell(\xi x)$ with affine rescaling ξ into $[-1,1]$. The second scheme uses the p-norm surrogate $m \le \|\boldsymbol{z}\|_p \le V^{1/p} m$ with p a power of two to enable repeated squaring; the relative error is bounded by $V^{1/p} - 1$ and decreases monotonically in p. Both schemes use SIMD packing to reduce the depth of the tree reduction to $O(\log V)$ without bootstrapping as long as the multiplicative depth of E_d or repeated squaring fits the CKKS level budget. With $\tilde{m}$ in hand, centered logits are $\hat{z}_i = z_i - \tilde{m}$ and satisfy $\hat{z}_i \in [-B, 0]$ with overwhelming probability after a per-step clamp that preserves ordering.

The second step of lazy normalization produces approximate unnormalized weights $\tilde{w}_i = E_d(\hat{z}_i)$ with uniform supremum error $\varepsilon_e = \sup_{x\in[-B,0]} |E_d(x) - e^x|$. The normalizer $\tilde{Z} = \sum_i \tilde{w}_i$ is computed by packed reductions. For top-p nucleus filtering without revealing thresholds or ranks, EDEN estimates the quantile θ_p that satisfies $\sum_{i:\tilde{w}_i \ge \theta_p} \tilde{w}_i \approx p\tilde{Z}$ via encrypted histograms. Let $\{c_b\}_{b=0}^{M}$ be bin boundaries covering $[-B, 0]$ and define smooth bin indicators $\chi_b(x)$ with $\sum_b \chi_b(x) = 1$ such that $\chi_b(x) \approx \mathbb{1}\{x \in [c_{b-1}, c_b)\}$. A CKKS-friendly construction uses an odd polynomial approximation to the hyperbolic tangent to emulate a smoothed step,

$$\chi_b(x) \approx \frac{1}{2}\left(\tanh\big(\alpha(x - c_{b-1})\big) - \tanh\big(\alpha(x - c_b)\big)\right), \tag{2}$$

where $\tanh$ is realized by its minimax polynomial on $[-B, B]$ and α tunes the transition width. Encrypted bin masses $h_b = \sum_i \tilde{w}_i \, \chi_b(\hat{z}_i)$ are accumulated by slot-wise additions; the encrypted cumulative $H_b = \sum_{j \ge b} h_j$ yields an index $\hat{b}$ with $H_{\hat{b}} \approx p\tilde{Z}$ obtained by a logarithmic encrypted search that uses a polynomial comparator $\mathrm{cmp}(x, y) \approx \frac{1}{2}(1 + \mathrm{sgn}(x - y))$ where sgn is approximated by a high-slope odd polynomial. The nucleus set proxy is then $\mathcal{N}_p = \{i : \hat{z}_i \ge c_{\hat{b}}\}$ and is represented as a mask $\mathbf{m}$ computed by applying the same comparator to each $\hat{z}_i$ against $c_{\hat{b}}$ within packed slots. The approximate top-$K^\star$ screening uses either the same histogram with a higher quantile or a blockwise maxima cascade: partition the packed logits into blocks of width W, compute per-block maxima with a $O(\log W)$ reduction, replicate block maxima back, and mark elements whose gap to the block maximum is below a tuned margin δ determined by the polynomial approximation error and the dynamic range.

Two-stage encrypted selection[6] refines the coarse candidate set to an exact argmax or top-k. Restricting to the candidate indices S with $|S| = K^\star$, EDEN executes an exact comparison network either in a TFHE bitwise comparator or in an honest-majority, constant-round MPC gadget. In the TFHE realization, each fixed-point candidate $\hat{z}_i$ is quantized to b bits and encrypted bitwise as torus LWE samples; the comparator is a ripple-carry subtractor with complexity $O(b)$ bootstraps per comparison and depth $O(b)$, but an argmax tree on $K^\star$ elements can be realized with $K^\star - 1$ comparators and depth $\lceil \log_2 K^\star \rceil$, so the overall bootstrapped depth scales as $O(b \log K^\star)$. In the MPC realization with arithmetic and Boolean sharing, a constant-round argmax protocol based on pairwise compare-and-swap with oblivious conditional moves achieves round complexity $O(1)$ and communication $O(K^\star \log K^\star)$ words, relying on OT extensions or preprocessed multiplication triples. Both variants output an encrypted one-hot vector $\mathbf{o} \in \{0,1\}^{K^\star}$ or the index encoded as a secret-shared integer. If top-k is required, EDEN uses a bitonic selection network of size $O(K^\star \log^2 K^\star)$ and depth $O(\log^2 K^\star)$ to avoid full sorting, which remains practical for $K^\star$ in the hundreds.

Oblivious alias sampling produces a draw from the refined categorical distribution with constant rounds and no access-pattern leakage. Let $\tilde{\boldsymbol{p}} = \{\tilde{p}_i\}_{i\in S}$ be the approximate probabilities on the refined set after optional temperature scaling $\tilde{z}_i \leftarrow \tilde{z}_i / T$ and repetition or frequency penalties implemented as affine transforms on logits. The alias method requires arrays $(\boldsymbol{a}, \boldsymbol{q})$ with $\sum_{i\in S} \tilde{p}_i = 1$ such that for a uniform bucket $j \in S$ and $u \sim \mathrm{Unif}[0,1)$, the sample is j if $u < q_j$ and a_j otherwise. EDEN synthesizes $(\boldsymbol{a}, \boldsymbol{q})$ under encryption by solving the linear feasibility system

$$q_i + \sum_{j : a_j = i} (1 - q_j) = |S| \cdot \tilde{p}_i, \qquad 0 \le q_i \le 1, \quad a_j \in S, \tag{3}$$

using a parallelized version of the classical two-list construction. Define scaled quotas $s_i = |S| \cdot \tilde{p}_i$ and index sets $L = \{i \in S : s_i < 1\}$ and $H = \{i \in S : s_i \geq 1\}$. In plaintext the algorithm repeatedly pairs an $i \in L$ with a $j \in H$, sets $q_i = s_i$, $a_i = j$, updates $s_j \leftarrow s_j - (1 - q_i)$, and moves j between sets if s_j crosses 1. Under encryption, EDEN executes this as a sequence of masked scans. Represent L and H by encrypted indicator vectors and compute pairings by stable assignments P derived from cyclic shifts in SIMD slots, which avoid data-dependent branching. Updates become

$$q_i \leftarrow s_i, \qquad a_i \leftarrow \mathrm{sel}(H, P(i)), \qquad s_{P(i)} \leftarrow s_{P(i)} - (1 - q_i), \tag{4}$$

where sel is an encrypted gather realized through slot rotations and masked additions. The process converges in at most $|S|-1$ rounds, but EDEN compresses rounds by grouping transfers via prefix sums over deficit and surplus vectors, which is valid because the invariant $\sum_{i \in L}(1 - s_i) = \sum_{j \in H}(s_j - 1)$ is maintained at each step. The final arrays are encrypted in CKKS; for the TFHE or MPC comparator required during sampling, $\boldsymbol{q}$ is re-encoded into the corresponding domain. Sampling consumes two correlated random variables per step: a bucket J uniform on S and $U \sim \mathrm{Unif}[0, 1)$. In HE-only deployments the client seeds a session PRF and provides encryptions $\mathrm{Enc}(J)$ and $\mathrm{Enc}(U)$; in MPC deployments parties derive J and U from shared seeds. The decision bit is computed as $b = \mathbb{1}\{U < q_J\}$ by an encrypted comparator, and the output index is $I = (1 - b) \cdot a_J + b \cdot J$. Only I is revealed; no heap shape, threshold, or pairwise comparison leaves the encrypted domain.

Session-to-session efficiency follows from incremental updates. Let $\boldsymbol{z}^{(t)}$ be logits at step t and $\delta_i = z_i^{(t+1)} - z_i^{(t)}$. The unnormalized weights update multiplicatively as $w_i^{(t+1)} = w_i^{(t)} \exp(\delta_i - \log \sum_j w_j^{(t)}(\exp(\delta_j) - 1))$. For small $\|\delta\|_\infty$ a first-order expansion yields

$$\frac{w_i^{(t+1)}}{w_i^{(t)}} \approx 1 + \delta_i - \sum_{j \in S} \tilde{p}_j^{(t)} \delta_j, \tag{5}$$

so scaled quotas $s_i^{(t)} = |S|\tilde{p}_i^{(t)}$ update as $s_i^{(t+1)} \approx s_i^{(t)} + |S|\tilde{p}_i^{(t)}\left(\delta_i - \sum_j \tilde{p}_j^{(t)} \delta_j\right)$. This is a rank-one correction that can be applied to $(\boldsymbol{a}, \boldsymbol{q})$ by a bounded number of local transfers. EDEN maintains a slack vector $\boldsymbol{\sigma}^{(t)} = s^{(t)} - \mathbf{1}$ and applies a few rounds of the encrypted transfer scan only to indices where $|\sigma_i^{(t+1)}| > \eta$, with η a fixed slack threshold that trades tiny TV error for significant speedup; the error is bounded by the total slack deficit resolved in later steps and is dominated by the polynomial and quantization errors.

The transformations that users expect—temperature, nucleus mass, and penalties—map to local encrypted arithmetic. Temperature rescales centered logits as $\hat{z}_i \leftarrow \hat{z}_i / T$, which multiplies the Chebyshev polynomial's argument but is absorbed by building E_d for $[-B/T, 0]$. A repetition penalty modifies logits by a convex combination with a count statistic $c_i^{(t)}$ of previously emitted tokens, realized as $\hat{z}_i \leftarrow \hat{z}_i - \lambda f(c_i^{(t)})$ for a sublinear f, which is an encrypted table lookup implemented via piecewise polynomials under CKKS or via circuit selection in MPC. Frequency penalties that depend on entire prefixes are handled by maintaining encrypted counters in parallel with decoding and applying the same local transform just before the selection subroutines; these transforms commute with lazy normalization since centering cancels constants.

The security boundary is formalized by an ideal functionality[**?**] $\mathcal{F}_{\mathrm{dec}}$ that, on input logits $\boldsymbol{z}$, policy Π (temperature, nucleus mass, penalties), and randomness r, outputs a token index $I \sim \mathrm{Dec}(\boldsymbol{z}, \Pi; r)$ and nothing else. A real-world adversary $\mathcal{A}$ interacting with EDEN's protocols for lazy normalization, selection, and sampling has a view consisting of HE ciphertexts or MPC transcripts and the revealed token I. For any $\mathcal{A}$ there exists a simulator $\mathcal{S}$ that, given only I, produces a computationally indistinguishable view under the IND-CPA security of CKKS for approximate arithmetic, the LWE hardness of TFHE bootstrapping for comparison, and the simulation-based security of the MPC gadget, because the only data-dependent branches are encoded as arithmetic on ciphertexts, histograms are computed under encryption[7], and the only revelation is I. The correlated randomness for alias sampling is generated either by encrypting PRF outputs under the client's key or via MPC's standard randomness generation, so timing and acceptance bits are hidden inside encrypted comparisons. Consequently, the leakage profile reduces to the unavoidable revelation of I and coarse-grained, policy-independent timing variations due to fixed circuit depth.

Concrete parameter selection[8] follows from error and noise-budget accounting. Given a target total variation distance ϵ_{TV}, select degree d such that the minimax Chebyshev error on $[-B, 0]$ satisfies $\varepsilon_e \leq \epsilon_{\mathrm{TV}}/4C$, choose histogram bin count M and slope α so that bin leakage contributes at most $\epsilon_{\mathrm{TV}}/4$, set the alias quantization grid Q with $1/Q \leq \epsilon_{\mathrm{TV}}/4$, and pick $K^\star$ so that the probability that a true top-k element is screened out due to approximation margins is at most $\epsilon_{\mathrm{TV}}/4$. With these, the composed TV error obeys $\|\tilde{\boldsymbol{p}} - \boldsymbol{p}\|_{\mathrm{TV}} \leq \epsilon_{\mathrm{TV}}$ by triangle inequality. The CKKS scale Δ and modulus chain $\{q_\ell\}$ are provisioned so that the multiplicative depth of E_d, histogram polynomials, and a constant number of rescalings fit within the noise budget without bootstrapping; this is feasible with $d \in [5, 9]$ for $B \in [8, 16]$ at $\Delta \approx 2^{40}$ and $N \in \{2^{14}, 2^{15}\}$ for V up to 10^5 with packing. The refinement stage determines the remaining latency: for TFHE with gate bootstrapping time τ_g and $b \in [16, 24]$, the argmax tree latency is approximately $\tau_g\, b \lceil \log_2 K^\star \rceil$, while for MPC with LAN-level latency the constant-round protocol remains sub-millisecond for $K^\star \leq 512$. In both cases, the amortized per-step cost of alias-table maintenance is dominated by a handful of slot rotations and additions over $K^\star$ slots and is negligible relative to the encrypted exponential evaluation.

In summary, EDEN's proposed solution realizes a statistically faithful and privacy-preserving decoder by centering logits with provably controlled approximations, restricting exact comparisons to a tiny, encrypted candidate set, and sampling from an encrypted alias structure synthesized and

maintained without revealing indices or thresholds. The protocol composes with any upstream private inference mechanism that emits encrypted logits and provides a tunable Pareto front among approximation degree, candidate bandwidth, and ciphertext modulus that aligns total variation error with cryptographic noise budgets and end-to-end latency targets.

IV. Analysis

The statistical fidelity of EDEN follows from approximation bounds and stability of selection under perturbations. Let $p_i = \exp(z_i)/\sum_j \exp(z_j)$ and $\tilde{p}_i$ be the EDEN sampling distribution after lazy normalization and alias sampling. Suppose the exponential surrogate $E_d(\cdot)$ on $[-B, 0]$ satisfies $\sup_{x\in[-B,0]} |E_d(x) - e^x| \le \epsilon_e$, and the max approximation $\tilde{m}$ satisfies $|\,\tilde{m} - \max(\boldsymbol{z})\,| \le \epsilon_m$. Then, after centering, the unnormalized weights incur a multiplicative error bounded by $(1 \pm \eta)$ with $\eta \le c_1\epsilon_e + c_2\epsilon_m$ for constants determined by the dynamic range. By Lipschitz continuity of normalization in the simplex interior, the induced total variation distance satisfies $\|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|_{\mathrm{TV}} \le C\eta$ for a constant C that depends on B and V, while top-p truncation introduces an additional δ equal to the tail mass discarded; EDEN chooses the approximate threshold so that the realized tail mass deviates from p by at most $O(\eta)$. Two-stage selection does not bias the distribution, because the exact argmax or refined top-k is computed in the second stage[**?**]; the only risk is false negatives during coarse screening, which we bound probabilistically by ensuring that the screening margin exceeds the approximation error with high probability. For alias sampling, quantization to a Q-level grid yields L_1 error at most $1/Q$, and the use of session PRFs and encrypted comparisons preserves distributional correctness conditioned on the quantized probabilities. Composing these errors yields a TV bound $\le C\eta + 1/Q$, which can be made arbitrarily small for moderate degree d and grid size Q.

The cryptographic cost decomposes into polynomial evaluation[**?**][9], packed reductions, a constant number of bootstraps or MPC rounds in the refinement stage, and constant-time sampling. CKKS packing allows evaluating E_d simultaneously on blocks of logits with complexity $O(d \cdot \lceil V/W \rceil)$ ciphertext multiplications where W is the SIMD width. Approximate maxima use tree reductions with depth $O(\log V)$ but no bootstrapping when modulus is provisioned accordingly; relinearization and rescaling are bounded by the multiplicative depth of E_d. Refinement over $K^\star$ candidates requires $O(K^\star \log K^\star)$ comparisons in TFHE with bootstrapping per gate, or $O(1)$ rounds and $O(K^\star \log K^\star)$ communication in MPC; with $K^\star$ in the hundreds, this term dominates neither latency nor bandwidth. Alias-table synthesis consists of encrypted scans and pointwise operations over the refined set and is reused across steps via rank-one updates whose cost is linear in $K^\star$. The leakage profile of EDEN is the minimal one for any correct decoder: the final token and a per-step timing budget. By avoiding access-pattern revelation for heap updates, threshold locations, or rank vectors, and by simulating coarse screening inside encryption, EDEN aligns with standard simulation-based definitions: an adversary learns nothing beyond what can be inferred from outputs and allowed timing side channels[**?**].

V. Future direction

Several extensions can further reduce overheads or strengthen assurances[**?**]. One direction is to integrate formal optimizer-aware bounds on logit dynamics so that alias-table updates can be predicted or batched for many steps, reducing even the amortized cost. Another is to exploit structured sparsity in vocabularies and subword statistics to precompute encrypted cluster-level thresholds that accelerate coarse screening. A third direction is to co-design upstream private inference so that the final transformer[2] block emits encrypted low-dimensional sketches from which EDEN can reconstruct the same top-$K^\star$ set with fewer polynomial evaluations. On the security front, it is attractive to replace the refinement enclave with constant-round MPC robust to dropouts, thereby eliminating trusted hardware assumptions entirely; similarly, verifiable randomness beacons can replace session PRFs when regulatory or multi-party governance[10] is required. Finally, end-to-end proofs that combine the approximation error of private inference layers with EDEN's decoding error would offer distributional guarantees for the entire encrypted generation process and guide principled selection of approximation degrees and grid sizes[11].

VI. Conclusion

Decoding has been the missing piece in practical privacy-preserving language-model inference. By rethinking normalization, selection, and sampling around cryptographic strengths, EDEN makes it possible to turn encrypted logits into tokens with bounded statistical error and competitive latency, while revealing only the information inherent in the output[7]. The combination of lazy normalization with polynomial surrogates, two-stage encrypted selection that reserves exact comparisons for a tiny candidate set, and oblivious alias sampling with correlated randomness yields a decoder that composes with any upstream private inference substrate. The analysis clarifies how approximation and quantization errors translate into TV bounds and why the cost scales with the refined candidate size rather than the full vocabulary[12]. These properties suggest that encrypted decoding need not be a showstopper for confidential LLM services and that principled algorithm–cryptography co-design can make full end-to-end encrypted generation viable in multi-tenant clouds and regulated environments.

References

[1] M. Avitan, M. Baruch, N. Drucker, I. Zimerman, and Y. Goldberg, "Efficient decoding methods for language models on encrypted data," *arXiv preprint arXiv:2509.08383*, 2025.

[2] D. Rho, T. Kim, M. Park, J. W. Kim, H. Chae, E. K. Ryu, and J. H. Cheon, "Encryption-friendly llm architecture," *arXiv preprint arXiv:2410.02486*, 2024.

[3] E. Chou, J. Beal, D. Levy, S. Yeung, A. Haque, and L. Fei-Fei, "Faster cryptonets: Leveraging sparsity for real-world encrypted inference," *arXiv preprint arXiv:1811.09953*, 2018.

[4] Z. Ruoyan, Z. Zhongxiang, and B. Wankang, "Practical secure inference algorithm for fine-tuned large language model based on fully homomorphic encryption," *arXiv preprint arXiv:2501.01672*, 2025.
[5] L. Folkerts, C. Gouert, and N. G. Tsoutsos, "Redsec: Running encrypted dnns in seconds." *IACR Cryptol. ePrint Arch.*, vol. 2021, p. 1100, 2021.
[6] S. Ashe and H. Ramachandra, "The effect of continuous encryption of data in cloud native architecture," in *2024 IEEE Cloud Summit*. IEEE, 2024, pp. 163–169.
[7] A. Knapp, M. Linga, and S. Yen, "Scaling policy assignment in containerized environment."
[8] T. Schneider, H.-C. Wang, and H. Yalame, "He-securenet: An efficient and usable framework for model training via homomorphic encryption," *Cryptology ePrint Archive*, 2025.
[9] F. Alder, N. Asokan, A. Kurnikov, A. Paverd, and M. Steiner, "S-faas: Trustworthy and accountable function-as-a-service using intel sgx," in *Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2019, pp. 185–199.
[10] M. P. Andersen, J. Kolb, K. Chen, G. Fierro, D. E. Culler, and R. Katz, "Democratizing authority in the built environment," *ACM Transactions on Sensor Networks (TOSN)*, vol. 14, no. 3-4, pp. 1–26, 2018.
[11] R. Bahmani, F. Brasser, G. Dessouky, P. Jauernig, M. Klimmek, A.-R. Sadeghi, and E. Stapf, "Cure: A security architecture with customizable and resilient enclaves," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 1073–1090.
[12] T. Xu, L. Wu, R. Wang, and M. Li, "Privcirnet: Efficient private inference via block circulant transformation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 111 802–111 831, 2024.