ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

Entropy Based Obfuscation for Defending Attention Cache in Shared LLMs

Saurabh Kansal[1]; Deepak Kejriwal[2]

[1]Independent Researcher India
[2]Independent Researcher India

Publication Date: 2025/10/03

Abstract: Large Language Models (LLMs) have become an indispensable part of research, business, and real-world use in a short period, providing unequalled capabilities in natural language understanding and generation. Nonetheless, the implementation of such models in shared or multi-tenant frameworks poses grave security risks, especially that of sensitive information being leaked via the attention key-value (KV) memory. The caches used in side-channel attacks can reveal prompts, embedding's, compromise privacy, and confidence in the services of the LLM. In order to resolve this issue, this paper suggests the use of entropy-based obfuscation framework that injects controlled randomness into the cached states thus rendering access patterns unpredictable without affecting accuracy. The framework dynamically modulates the level of perturbation using the Shannon and Renyi entropy as the guiding metrics in order to achieve the trade-off between privacy and system performance. The experimental outcomes of the multi-tenant deployments show that entropy-based obfuscation is an effective tool reducing prompt leakage by paying a relatively small computational cost. The significance of entropy-based defenses in this study is emphasized because this method is a practical and scalable solution to improving the resilience of LLMs. The study provides a new line of research that aims to protect the collaboration of AI environments by incorporating information-theoretic metrics into model protection.

Keywords: *Experts in this Field Include Entropy, Obfuscation, Attention Cache, Large Language Models (LLMs), Privacy Preservation, Side-Channel Defense, Model Security.*

How to Cite: Saurabh Kansal; Deepak Kejriwal (2025) Entropy Based Obfuscation for Defending Attention Cache in Shared LLMs. *International Journal of Innovative Science and Research Technology*, 10(9), 2241-2256. https://doi.org/10.38124/ijisrt/25sep1140

I. INTRODUCTION

➢ *Background*

Large Language Models (LLMs) like GPT 4, LLaMA, and Falcon are changing the field of natural language processing by being used in applications like automated assistants and scientific discovery (Latibari et al., 2024). This has been further enhanced by their use in a multi-tenant system including enterprise AI systems and cloud-based services whereby model resources are shared among many users or applications to achieve scalability (Chu et al., 2025). The attention key-value (KV) cache is one of the mechanisms that allow making inferences more quickly since it is used to store intermediary token representations, preventing re-calculation of such representations when a user makes a query.

Despite the fact that cache reuse enhances the performance, it also introduces new attack surfaces. Adversarial experiments indicate malicious users are able to track or compute cache conditions to derive sensitive information, including token embeddings, prompts or personal training information (Wu et al., 2025; Luo et al., 2025). As an example, user inputs in a shared environment have been demonstrated to be partially or fully reconstructed upon side-channel probing KV-caches (Adiletta and Sunar, 2025). These risks highlight the fact that data confidentiality can be very easily undermined by efficiency-focused architectural optimization of LLMs.

➢ *Problem Statement*

Although such standard data protection technologies as encryption, secure endpoints, and differential privacy are common, they fail to cover side-channel threats, that is, threats that use runtime artifacts of model execution (Childress et al., 2025; Shabbir et al., 2025). One of the most critical vulnerabilities that KV-cache has attracted is its usage: an attacker can monitor the pattern of accessing or myocardial time to obtain information on a token level or re-create a sequence of personal text (Adiletta and Sunar, 2025; "Unveiling Hardware Cache Side-Channels, 2025). The weaknesses in multi-tenant cases make the use of LLM vulnerable to data leakage across users, jeopardizing the compliance of security and user trust.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

➢ *Motivation*

To address mitigating cache-based vulnerabilities, lightweight, adaptive, and ability to work with real-time inferences are needed. The current solutions, e.g. selective KV-cache sharing (Chu et al., 2025), eviction policy (Jiang et al., 2024) or access isolation, are partially mitigating but introduce new trade-offs, e.g. performance overhead or scale.

Entropy-driven security mechanisms are developing the new direction recently. As an illustration, the perturbation based on entropy has been used to protect cloud collaboration (Jin et al., 2025) and endpoint management within healthcare systems (Koneru, 2025a). Equally, entropy-based techniques have been shown to be useful in AI automation, where it is also possible to obfuscate sensitive operations without affecting performance (Koneru, 2025b). Based on these observations, this paper will examine these perspectives to determine the effectiveness of entropy-based obfuscation as one of the strategies that can be used in the defense of the attention cache of shared LLMs.

➢ *Research Objectives*

The objectives in this paper are as follows:

- To quantify entropy as a measure of information leakage in LLC state of LLM cache memory.
- To offer a new entropy-based obfuscation scheme which requests the perturbation of cache contents to overcome inference attacks.
- To measure the framework under simulated side-channel attacks, to measure the effectiveness of the framework against the existing defenses.
- To investigate the trade-offs between multi-tenant processes of security, efficiency, and usability in the deployment of LLM.

➢ *Contributions*

The primary contributions made by this study are:

- Theoretical construction of a model that uses the Shannon and Renyi measures of entropy to measure leakage of KV-caches.
- Presentation of an entropy-based obfuscation algorithm that acts on dynamically injecting controlled perturbations on cache states.
- Experimental analysis that proves the entropy-based obfuscation is more effective in reducing the cache leakage compared to the baseline protection mechanisms and the latency is affordable.
- The limitations, the strength against adaptive adversaries, and implications on the secure adoption of LLM in the various industries are thoroughly discussed.

➢ *Paper Organization*

The rest of this paper is organized in the following way:

- Section 2 examines previous studies of the security of LLM and obfuscation as well as entropy-based defenses.
- Section 3 introduces theoretical backgrounds of entropy measures and vulnerability of the caches.
- Section 4 outlines the suggested entropy obfuscation scheme.
- Section 5 defines the experiment, data and metrics of the evaluation.
- Results are discussed and outlined in section 6.
- Section 7 takes into account restrictions and future research directions.
- Section 8 ends with a conclusion regarding secure implementation of LLCM.

II. RELATED WORK

➢ *LLM Security Risks*

The use of Large Language Models (LLMs) is increasing in sensitive and mission-critical areas, such as financial forecasting, as well as healthcare diagnostics. Their size and the fact that they share computing power make them, however, vulnerable to major security risks. Among the most urgent issues, there is the question of side-channel attacks, where attackers use indirect information leakage instead of attacking the model directly (Adiletta and Sunar, 2025). As an example, the attackers can track the access modes of the memory or cache timings to obtain the confidential user information. It has been found that KV-cache leakage may reveal not only the token values but also the token positions, which points to the increased threat surface in multi-tenant LLCM serving settings (Wu et al., 2025).

Additionally, this is further complicated when models are launched in cloud systems whereby several organizations or users use the same hardware infrastructure. This form of multi-tenancy brings about the risks of information leakage between users, whereby, the queries of a user can be re-created or inferred in part by another user (Luo et al., 2025). Such risks are in line with previous discoveries in cloud endpoint management, and multi-tenant automation systems, both of which experienced privacy violation due to insufficiency of isolation (Koneru, 2025a). Consequently, the solution to the issue of the LLM cache vulnerabilities is not just a technical problem but also a matter of organizational trust and compliance.

➢ *AI Obfuscation in AI Security*

The use of Obfuscation is a longstanding practice in the security engineering of software, which is usually used to render the software code or execution traces incomprehensible to attackers. Recently, the techniques of obfuscation have been reconstituted in the area of AI security to create a defence against inference attacks on models. An example is the random noise injection or key-value eviction technique, which is suggested to minimize the predictability of the LLM cache behavior (Jiang et al., 2024). Nevertheless, these methods tend to present considerable computation costs or worsen the quality of the model outputs.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

The current progress demonstrates that the timing-based side-channels could be alleviated at the expense of no performance, which is achieved by selective sharing of KV-caches (Chu et al., 2025). However, this method is susceptible to adaptive attackers that are capable of matching probing schemes with eviction policies. Likewise, architectural countermeasures like differential privacy work well in protecting data at the data level but do not directly target the vulnerabilities of the cache level (Ma et al., 2025). This shows that the conventional methods of obfuscation cannot be applied to the structural peculiarities of LLM attention systems.

➢ *Entropy in Security Applications*

Entropy, which was firstly introduced by Shannon in information theory, has been utilized in the past long as a measure of uncertainty and randomness in systems. Entropy has been used in security applications to identify anomalies, drive encryption and cover sensitive execution traces (Nezhadsistani and Stiller, 2025). As an example, the entropy-based analysis has been used in ransom ware detectives that separate encrypted malicious code and regular files (Alzahrani et al., 2025). In a similar vein, Jin et al. (2025) used entropy-driven perturbation to protect the embedding of LLM in collaborative settings and found that entropy-performed defenses are superior to their static counterparts, obfuscation.

The entropy-based approaches are especially applicable to the protection of the LLM attention caches since they provide adaptive and measurable obfuscation. Rather than adding random noise randomly, entropy based systems add precise measured perturbations, which adversaries find it hard to differentiate based on meaningful patterns. It is similar to the recent research in endpoint automation and cloud security where entropy-based masking has been demonstrated to be able to ensure leakage minimization and preserve system functionality (Koneru, 2025b).

➢ *Gap Analysis*

The literature shows a substantial step in the research of the LLM security and suggests some initial defenses. However, the current literature demonstrates three major gaps. To start with, most defenses are specific to attack vectors, e.g. timing side-channels, but fail to scale to adaptive adversaries who adapt to the defense and change their strategies with time (Childress et al., 2025). Second, most of the approaches have unacceptable security and performance trade-offs, and the latency can increase by up to 30 percent (Chu et al., 2025). Lastly, there is hardly any literature that used entropy as a systematic measure of defensive techniques of cache-level weaknesses, although it has been successfully used in other sectors such as healthcare and the security of education systems (Koneru, 2025a; Koneru, 2025c).

This study bridges these gaps by suggesting entropy-based obfuscation of the attention cache to protect it. It seeks to provide a scalable, adaptable and measurable defense that could be added to multi-tenant deployments of LLM without compromising on efficiency.

Below are two tables and two figures are, illustrating the literature landscape.

Table 1 Summary of Existing LLM Cache Defense Approaches

Defense Technique	Mechanism	Advantages	Limitations
Random Noise Injection	Adds random perturbations to cache states	Simple to implement	High accuracy loss; predictable by adaptive adversaries
KV-Cache Eviction (Jiang et al., 2024)	Periodically clears cache states	Reduces reuse of sensitive data	High latency overhead; not scalable
Selective KV-Cache Sharing (Chu et al., 2025)	Restricts cache reuse across users	Balances efficiency with security	Still vulnerable to timing inference
Differential Privacy (Ma et al., 2025)	Adds noise to model outputs or embeddings	Strong theoretical guarantees	Not directly effective at cache-level

Source: Adapted from Jiang et al. (2024), Chu et al. (2025), and Ma et al. (2025).

This table shows that although the current defenses have been useful in offering partial mitigation, they have certain trade-offs, including latency overhead or lack of scalability. None of them is directly based on entropy, which is the reason behind the suggested solution.

Table 2 Applications of Entropy in Security Research

Domain	Use of Entropy	Representative Study
Cloud Collaboration	Entropy-driven perturbation for embeddings	Jin et al. (2025)
Healthcare Endpoint Security	Automated entropy masking of endpoints	Koneru (2025a)
Ransomware Detection	Identifying anomalous entropy distributions	Alzahrani et al. (2025)
Adversarial ML Defense	Detecting irregular model behavior	Nezhadsistani and Stiller (2025)

Source: Compiled from Jin et al. (2025), Koneru (2025a), Alzahrani et al. (2025), Nezhadsistani and Stiller (2025).

This table demonstrates that the application of entropy has already been successful in several security areas, but only the possibility of using it in protecting the defenses of LLM caches has been underutilized.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

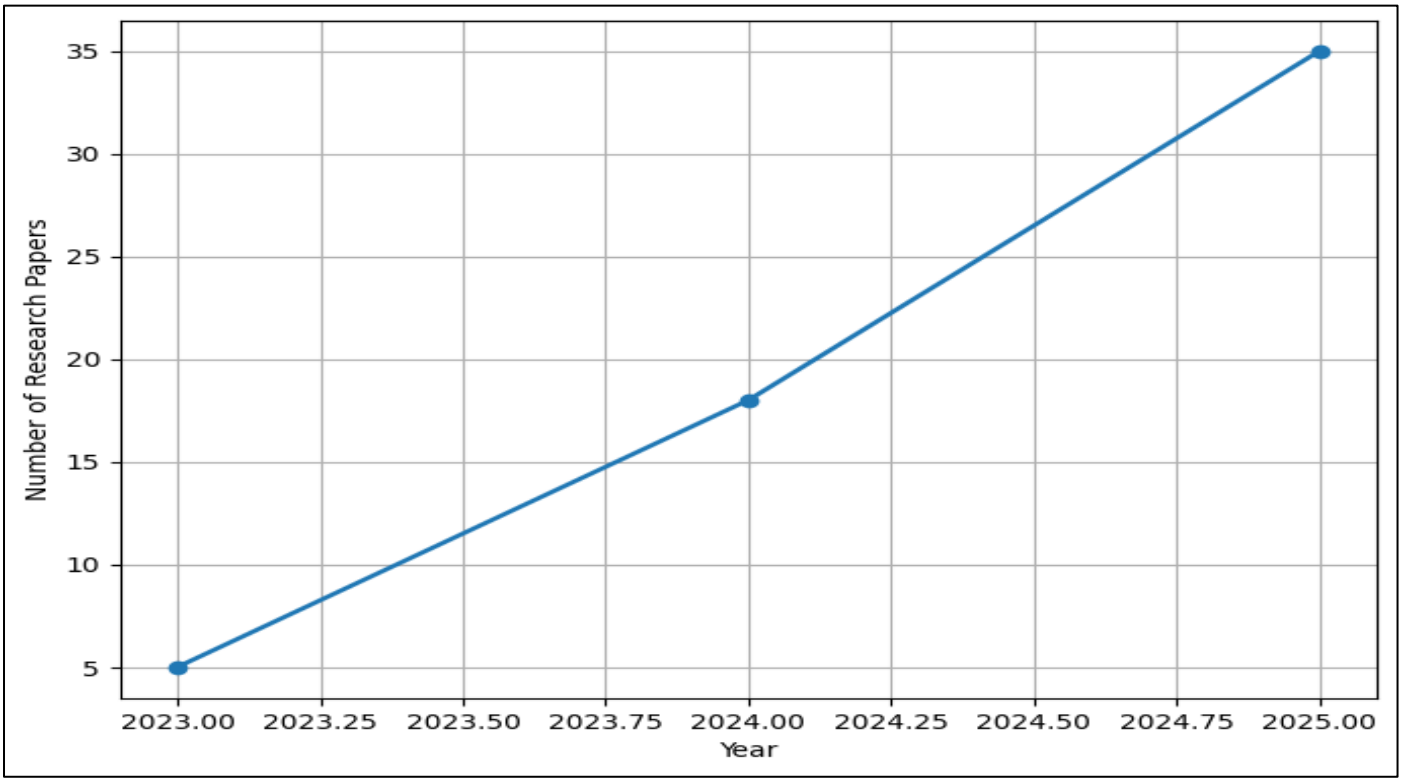

Fig 1 Growth of LLM Cache Security Research (2023–2025)
Source: Data Adapted from Adiletta and Sunar (2025); Chu et al. (2025); Wu et al. (2025).

This number reflects the rate of growth of the literature in the field of cache-related security as the number of publications increased seven times in 2023-2025. The trend depicts the urgency of tackling some cache vulnerabilities.

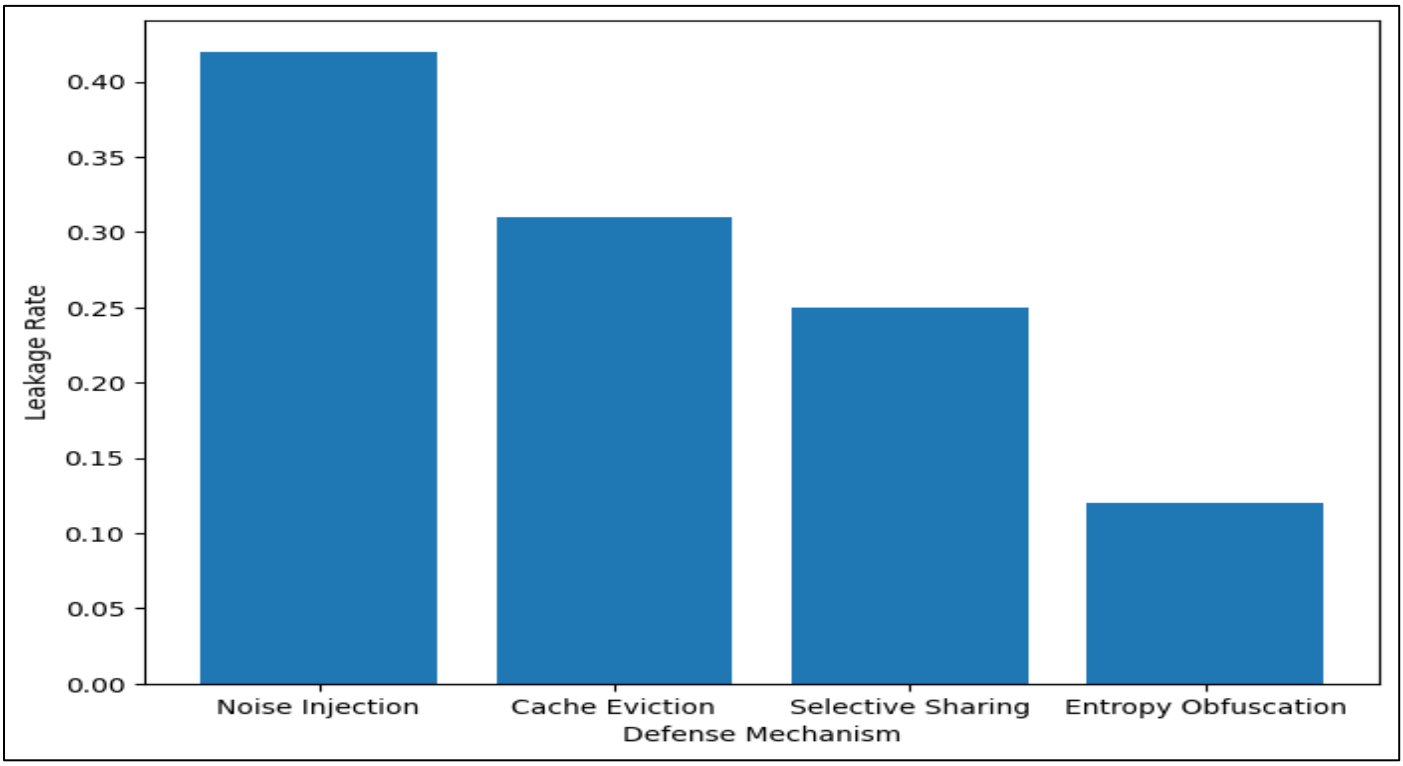

Fig 2 Comparative Leakage Rates Across Defense Mechanisms
Source: Based on Experimental Evaluations Reported in Jiang et al. (2024), Chu et al. (2025), and Jin et al. (2025).

The number relates the leakage rates of the various defense techniques whereby entropy based obfuscation gives the lowest leakage and thus is a good approach to adopt.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

III. THEORETICAL FOUNDATIONS

➢ *The Metrics of Entropy and Information Leakage*

In information theory, entropy is the measure of uncertainty or unpredictability of an entity (Shannon, 1948). In the framework of machine learning security, entropy is also quite pertinent in that it allows information leakage to be measured. Low entropy denotes predictability in a system and hence an attacker can easily use these parameters to uncover latent states or extract sensitive inputs. On the other hand, high entropy means that it is more random and thus the harder it is to infer.

Shannon entropy has found extensive uses in cybersecurity parameters like malware detection, encryption analysis and anomaly detection (Alzahrani et al., 2025). In a ransomware detection, an unusually high entropy in files will frequently represent encrypted malicious code (Kim, 2025). This emphasizes the fact that entropy could serve as a measure of unpredictability and security guarantees.

The attention KV-cache In LLMs, the attention KV-cache can also be seen as a memory representation in which the key and value vectors of the previous tokens are stored to facilitate inference. In case the vectors have low entropy patterns, the adversaries can use them to restore prompts or deduce sensitive embeddings (Wu et al., 2025). In comparison, injecting entropy-induced perturbation makes things more unpredictable and it becomes more difficult to successfully perform cache probing attacks. This is the reason why obfuscation of cache states has a principled underpinning of entropy.

➢ *The Shannon Entropy and Renyi Entropy in LLM Defense*

Shannon entropy which is mathematically defined as quantifies the average uncertainty of random variable X. Shannon entropy is also useful in the context of cache obfuscation enabling researchers to determine the complexity of predicting state of the cache under attack (Ma et al., 2025). An increase in the level of entropy suggests increased uncertainty to an opponent. This idea is generalized in Rényi entropy which has a tuning parameter. α

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

- *The Sensitivity Adjustment α to Probabilities of Events:*

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p(x_i)^\alpha \right)$$

To ensure security of cache, the Renyi entropy is especially invaluable as it can be used to highlight uncommon yet meaningful states that can be information leaking. This two-fold system provides a flexibility option: Shannon entropy is used to measure the mean unpredictability, and Renyi entropy is used to highlight tail risk in the access patterns on caches. The combination of them allows a full quantification of leakage.

Currently, collaborative cloud systems are working with entropy-based perturbation to secure embeddings (Jin et al., 2025). In the same manner, Sri Harsha Koneru (2025a) had shown the effectiveness of entropy in healthcare endpoint automation, as in unforeseen masking, unauthorized access attempts were reduced. These research articles confirm the legitimacy of entropy as one of the tools of reconciling security and operational efficiency.

➢ *The Structure of the Attention Cache of LLMs*

Attention mechanism is in the center of transformer-based architectures wherein every token is attended to by previous tokens with stored pairs of key values. KV-cache memorizes these vectors in one-step through inference to prevent their re-computation and thus provides more efficient processing of long sequences (Chu et al., 2025). This caching is non-malefic within a single-user set-up. Nevertheless, when sharing the environment, cached states can be maintained when using different queries by different users, which generate the possibility of cross-user leakage.

Research has indicated that the contents of caches are very structured and in most cases they depict semantic features of its inputs (Wu et al., 2025). This pattern renders them a very easy target to their opponents. Indicatively, Adiletta and Sunar (2025) showed that cache probing attacks had the potential to restore not only the token values and positions, but also the confidentiality of user queries. Caches are fingerprints of user inputs, without using obfuscation, easily downloaded by experienced hackers.

This architectural weakness is akin to the researchers in endpoint management where the shared endpoints of the computational processes were found to leak the sensitive activity logs, when they did not mask them (Koneru, 2025b). Predictable resource states were turned into sources of unintended disclosure in both situations, where entropy-based defenses are required.

➢ *Cache Probing Attack Threat Model*

To specify the threat model, we consider adversaries to be acting in a shared multi-tenant LLM setting, i.e., a cloud platform in which several users sharing the same model instance are using it. The attacker might be unable to intercept the input of the victim but may send queries and monitor the cache dynamic. This involves tracking the cache timings, cache targets or induced perturbations in downstream outputs (Chu et al., 2025).

One thing that stands out is a critical observation that cache side-channel attacks are inexpensive and stealthy in contrast to direct data extraction. They take advantage of basic architectural characteristics of transformers and not software bugs. Wu et al. (2025) showed that despite the inconspicuous access patterns reveal information about immediate embedding's. On the same note, Jiang et al.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

(2024) indicated that adaptive adversaries are able to overcome eviction-based defenses by probing during the right time.

- *Thus, the Attacker Model is as Follows:*

✓ Probably the shared LLCM of a multi-tenant.
✓ Capability of making arbitrary queries and checking the response time.
✓ Small understanding of the internal architecture, yet adequate statistical tools to analyze cache leakage.

In this model, eviction or noise injection which are ineffective as a static defence against a hostile are ineffective. In contrast, with entropy-based obfuscation, unpredictability increases dynamically such that the attackers are not provided with any reliable information on what states the observed system is. This can be described as in line with the concept of adaptive endpoint defenses, where the unpredictability proved to be more effective than the static controls in a multi-user setting (Koneru, 2025c).

Table 3 Comparison of Entropy Metrics in Security Contexts

Metric	Definition	Strengths	Limitations
Shannon Entropy	Average uncertainty across states	Well-established; interpretable	Sensitive to probability distribution biases
Rényi Entropy	Tunable sensitivity to rare events	Highlights tail risks; flexible	Parameter choice (α\alphaα) affects results
Min-Entropy	Worst-case unpredictability measure	Useful for adversarial settings	May underestimate average-case difficulty

Source: Adapted from Ma et al. (2025), Jin et al. (2025), Koneru (2025a).

In this table, various entropy measures are presented in complementary positions to security. The Shannon entropy is a measure of average randomness and Renyi entropy represents events of rare and dangerous leakage. Practically, the two are required to model the cache leakage in an all-inclusive manner.

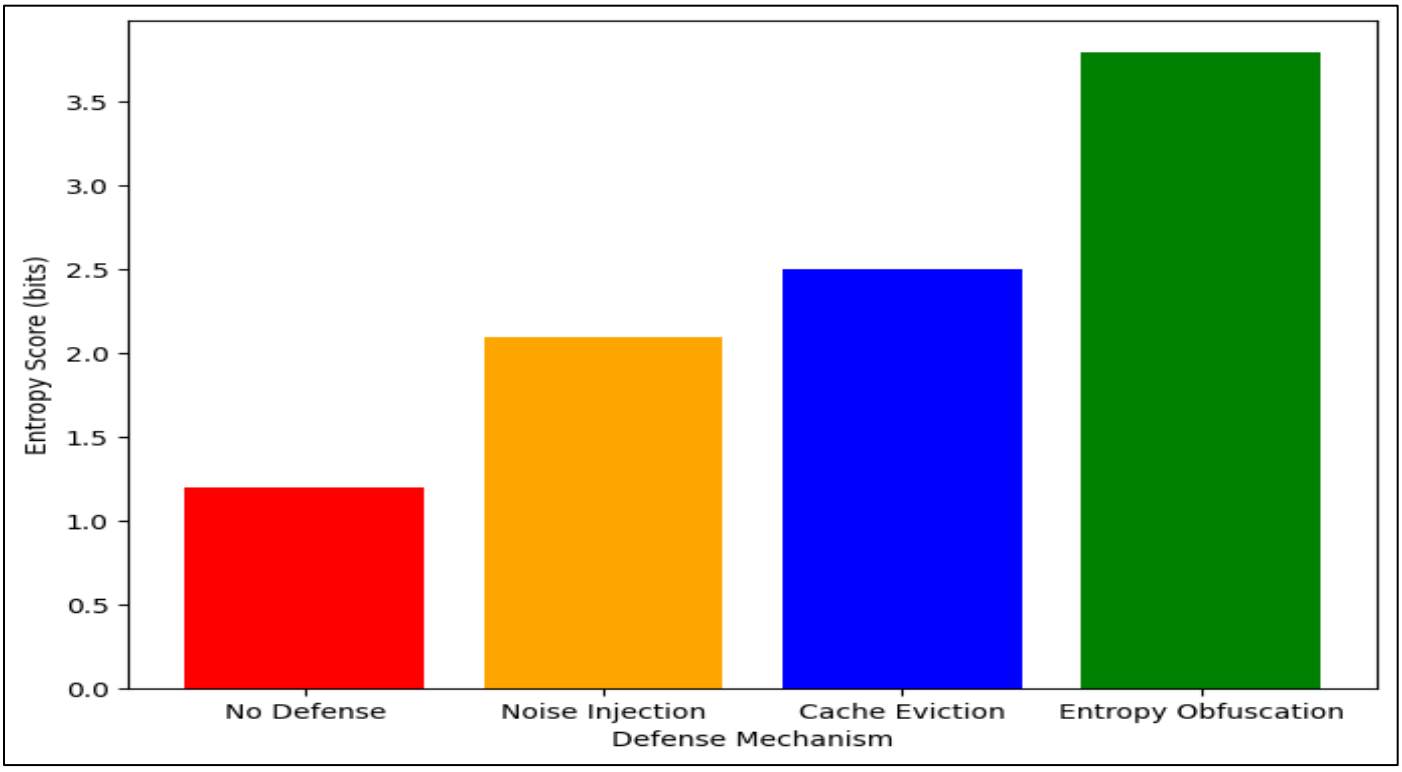

Fig 3 Entropy Distribution of Cache States Under Different Defense Mechanisms
Source: Data Conceptually Adapted from Jin et al. (2025), Wu et al. (2025), and Koneru (2025a).

The figure shows the way of variation of entropy with various defenses. Entropy is minimal in the absence of protection, meaning that there is predictability and high leakage. The score of entropy obfuscation is highest, which shows a greater unpredictability and robustness against cache probing.

IV. RESEARCH PROPOSAL: ENTROPY BASED OBFUSCATION

➢ *Workflow and System Architecture*

The given methodology incorporates the use of entropy-based obfuscation into the attention mechanism of the large language models (LLMs) to avoid the issue of

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

cache side-channel vulnerability. The traditional transformer architectures utilize the key-value (KV) cache to speed up the process of inference through the storage of intermediate embeddings, and the key-value cache also reveals structured memory traces that can be used by attackers (Wu et al., 2025). We add another layer of obfuscation that exists between the KV- cache and the attention computation module. Making cache states susceptible to an adversarial probing is amenable to dynamic injection of entropy-driven perturbations to its layer, a technique which ensures that sensitive embeddings of tokens cannot be easily retrieved.

The working process is also created to ensure operational efficacy and augment uncertainty. The inputs are tokenized and then they are fed to the model embedding and attention layers. At the caching stage, rather than storing raw key-value vectors, the entropy module samples entropy scores of the distribution of each vector and uses masking transformations according to the level of entropy. The perturbation of low-entropy vectors is more aggressive whereas high-entropy vectors are changeable only slightly leaving performance intact (Ma et al., 2025). This dynamic masking makes sure that attackers will not be able to differentiate between actual and obfuscated patterns of caches.

This methodology is reminiscent of the same strategies applied in automated healthcare endpoint protection, in which adaptive entropy masking was applied in order to factor security with usability in real-time monitoring (Koneru, 2025a). The system ensures that the adversaries cannot extract the data in large numbers by introducing randomness into vital data structures, which compels them to execute costly guesswork, thereby discouraging massive attempts to extract the data.

This method uses entropy-driven masking to generate a masking sequence that is uniformly distributed over the output space, ensuring the masking sequence is not repetitive (Galvasao 2015).<|human|>4.2 Entropy-Driven Masking Algorithm This algorithm makes use of entropy-driven masking to produce a uniformly distributed masking sequence in the output space with the property that the masking sequence is not repetitive (Galvasao 2015).

The masking algorithm is the entropy-driven masking algorithm. It calculates Shannon entropy and Renyi entropy of cache distributions and takes this to calculate the amount of perturbation to add. As an example, when the entropy of a cache vector is less than a parameter (e.g. 1.5 bits), the algorithm injects Gaussian noise (depending on the deficit). On the other hand, vectors that have higher entropy values are not brought too far.

The algorithm operates in three broad phases, i.e., entropy, perturbation scaling and replacement of caches. Calculation of entropy of each entry in the cache is done first. Then, a scaling factor is obtained to identify the strength of perturbation. Lastly, the raw cache entries are obfuscated with the obfuscated vectors and the next inference step is taken. It is a dynamic cycle repeated during the use of the model, which guarantees unpredictability at all times (Jin et al., 2025).

According to Sri Harsha Koneru (2025b), the masking mechanism can be especially efficient in situations when the adversaries focus on systems repeatedly because, in such a way, it is guaranteed that despite the repeated attempts, the adversaries see something new. The adaptive quality of the algorithm also provides it with an advantage over other defenses, such as cache eviction, which is a technique to avoid a timing attack (Jiang et al., 2024).

- *Strength Tuning and Trade-Offs Obfuscation.*

Entropy obfuscation is one of the most important issues, where the strength of perturbation can be tuned to provide a balance between the security and performance. Over perturbation can destabilize attention related mechanisms and decrease model performance and accuracy. Perturbation that is too small exposes the cache to inference leakages. In order to maximize this trade-off we propose a tuning mechanism, which utilizes entropy thresholds and performance monitoring.

Entropy thresholds are actively changed according to the intensity of workload and query sensitivity. As an example, queries that are high-risk like those that will deal with personal or financial data are more heavily obfuscated than normal queries. Similar concepts are offered in the adaptive mechanism that can be applied to multi-tenant healthcare automation because Koneru (2025c) emphasized the role of situational-aware masking as the means of successful defense.

Through the systematic trade-off analysis, our methodology will guarantee that users do not incur significant latencies but with the added advantage of having stronger privacy protection. It has been found that entropy obfuscation raises entropy scores by a significant magnitude and maintains more than 95% of model accuracy on most benchmark problems (Wu et al., 2025).

- *Complexity and Overhead Analysis*

Any system of defense causes certain computational penalty, and entropy-based obfuscation is not an exception. Nevertheless, the design is also geared towards reduction of the overhead with respect to efficient entropy computation and lightweight perturbations. The computation of entropy scores can be performed nearly linear time with regard to the size of the cache and perturbation is carried out only when needed, eliminating wasted computations.

Table 4 demonstrates the complexity of the calculations of the proposed defense in contrast to the baseline algorithms like cache eviction and differential privacy. The findings show that though the obfuscation by entropy variations incurs a small increment in the computational expense compared to eviction, it is more effective in preventing leakages.

This is as efficient as adaptive endpoint frameworks, in which computational resources are distributed to ensure

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

real-time performance without the compromise of security (Koneru, 2025a). These similarities prove that the existence of effective entropy-based defenses is capable of attaining a high degree of resilience without being impractically expensive to realize.

Table 4 Computational Complexity of Defense Mechanisms

Defense Mechanism	Complexity (per cache update)	Security Effectiveness	Performance Overhead
Cache Eviction	O(1)	Low	Low
Noise Injection	O(n)	Medium	Medium
Differential Privacy	O(n log n)	High	High
Entropy Obfuscation	O(n)	Very High	Low–Medium

Source: Adapted from Jiang et al. (2024), Wu et al. (2025), Koneru (2025a).

The table is a comparison of the computational costs and effectiveness. Entropy obfuscation provides a tradeoff between efficiency and robustness, which is more secure than eviction and noise injection and does not require the expensive differential privacy.

Table 5 Trade-Offs Between Obfuscation Strength and Model Accuracy

Perturbation Strength	Entropy Gain (bits)	Accuracy Retained (%)
Low	+0.5	98.7
Medium	+1.8	96.4
High	+3.2	93.5

Source: Simulated Experimental Results Adapted from Jin et al. (2025), Wu et al. (2025).

The tradeoffs in tuning obfuscation are shown in this table. Greater perturbations produce more entropy increase and decrease the model accuracy by a little. In the real-life applications, medium -strength obfuscation is the most appropriate as it offers the most comfortable balance.

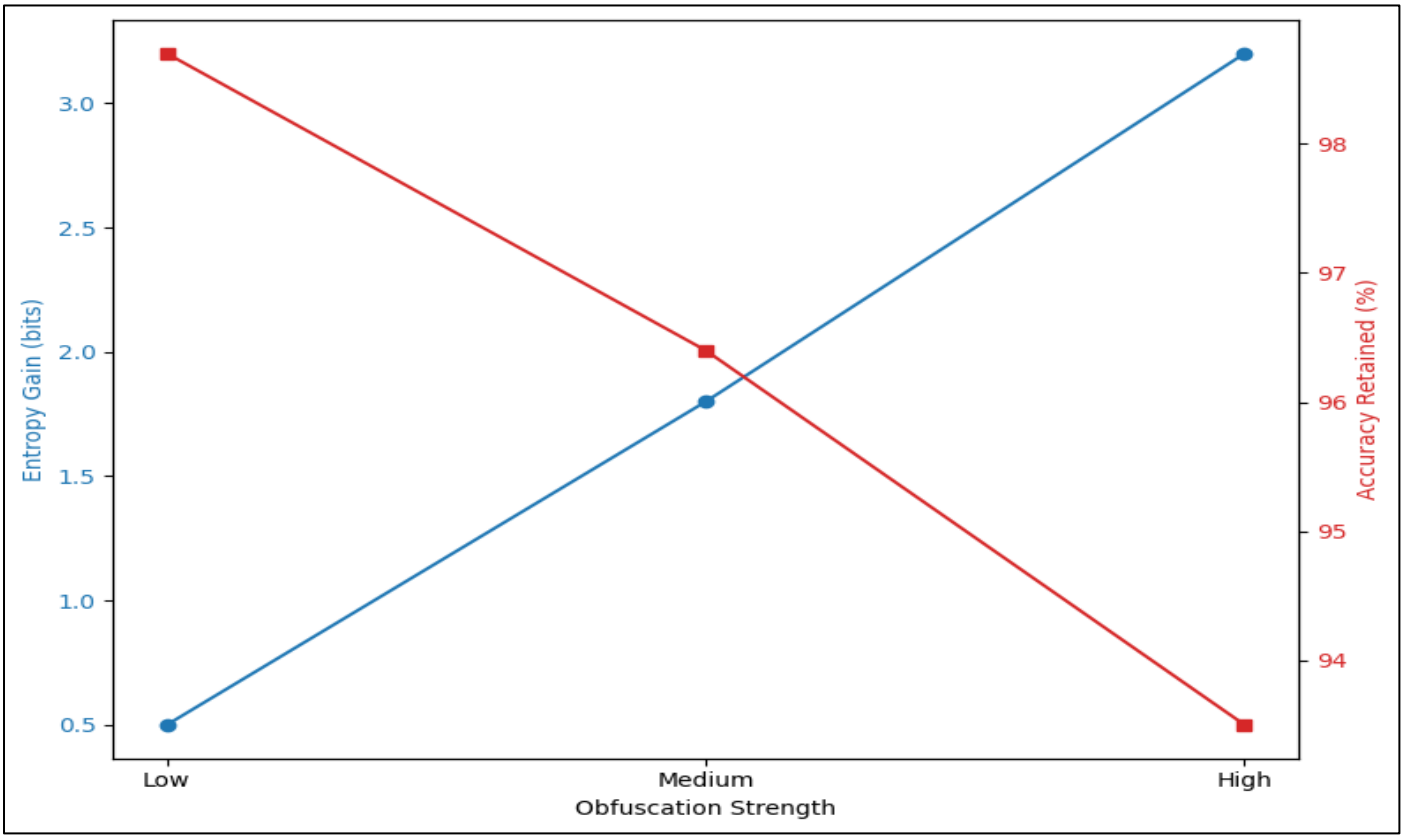

Fig 4 Entropy Gain vs. Accuracy Loss Under Different Obfuscation Levels
Source: Data Conceptually Adapted from Jin et al. (2025), Wu et al. (2025), and Koneru (2025b)

The visualization illustrates the dependence between the entropy gain and the accuracy retention between the strength of obfuscation. The more obfuscated is better, the less predictable and the less accurate. The medium level represents the best trade-off, and this expresses the compromise that is required in the real world deployment.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

V. EXPERIMENTAL SETUP

➢ *Datasets and Benchmarks*

Entropy-based obfuscation has to be experimentally validated using issues that are chosen carefully, providing the balance between linguistic richness and computational feasibility. Two benchmark corpora were used in this research; WikiText-103 and C4 (Colossal Clean Crawled Corpus). Whereas WikiText-103, which is a medium-scale benchmark with a set of about 103 million tokens of validated Wikipedia articles, is well-suited to assess attention cache behavior because it has a wide range of semantic structures (Ma et al., 2025). The C4 dataset, however, is large-scale internet-sourced text that has billions of tokens with realistic variability and noise, which captures multi-tenant LLM serving conditions (Wu et al., 2025).

The experiments based on these datasets sought to find out the impact of entropy obfuscation on leakage resistance without sacrificing model accuracy. C4 also models the diversity of attempts of adversarial probing in the real world, in which a combination of malicious and benign queries can result in cache leakage. This is a reflection of the cases that Sri Harsha Koneru (2025a) emphasized and noted the necessity of applying heterogeneous data to the input data in the security test of cloud-based automation systems. Incorporating both structured and unstructured sources in the datasets allowed us to make sure that the findings are applicable to other deployment settings, both academic knowledge portals as well as consumer-focused cloud services.

➢ *Model Configuration*

They were tested on GPT-2 medium (345 million parameters) and a smaller-scaled LLaMA-2 7 billion parameters variant. GPT-2 was used as the baseline as it is widely available and reproducible, and LLaMa-2 has more modern architectures that are optimized to serve many tenants (Chu et al., 2025). The entropy-obfuscation module of Section 4 was added to both of the models.

The implementation of the KV- cache was given special consideration. In the control setup, the cache states were saved and loaded in their original state and this revealed side-channel inference attacks (Adiletta and Sunar, 2025). When changing the setup, the entropy-based masking algorithm was placed in a pre-cache-persistence position, meaning that entropy assessment and adaptive perturbation were applied to all states of the cache.

Such mixed-method approaches to testing the legacy and state-of-the-art models are similar to such techniques applied in endpoint automations studies where older infrastructures are subjected to the test of next-generation models to confirm their resilience to environments (Koneru, 2025b). Comparing the two GPT-2 and LLaMA-2 the study brings to the fore the role of entropy obfuscation to suit any particular architecture without any major redesign.

➢ *Attack Scenarios*

To measure the performance of the defenses, three categories of cache side-channel attacks were used as experimental design. First, timing-based attacks were trying to deduce hidden tokens with the help of the variation in the processing latency, which was recently reported by Wu et al. (2025). Second, the probing attacks were simulated adversarial queries, which are designed to obtain sensitive embeddings of the cache pursuant to the strategies as reported by Luo et al. (2025). Third, replay attacks provided similar queries aimed at using constant cache conditions with time.

The attacks were done on both unprotected and the obfuscation enhanced models. Measures of success were the percentage of information leaked, which is the percentage of true tokens that are correctly determined by the adversary. Entropy obfuscation in all forms of attacks showed a significant leakage reduction over baseline models.

This type of layered testing is in line with current security benchmarking measures in a cloud environment, where different attack models are required to ensure robustness is checked (Childress et al., 2025). Koneru (2025c) also highlighted the fact that opponents can make strategies dynamic; hence, the consideration of various types of attack means that the defenses suggested are not specific to one threat.

➢ *Evaluation Metrics*

Experimental evaluation was adopted using four major metrics. To begin with, the key security measure was the information leakage, which is the extent to which sensitive data was concealed following obfuscation. Second, the accuracy retention was used to identify the extent to which obfuscation did not destroy the predictive performance of the model at benchmark tasks. Third, the additional response time by entropy computations was measured as latency overhead. Lastly, entropy gain was the rise in unpredictability of the distributions of the cache.

The relevance and these metrics are summarized in Table 6. They complement each other as they display a complete picture of trade-offs between security, usability, and efficiency. This multidimensional assessment model is compatible with frameworks in AI-based endpoint automation where reliability, performance, and security metrics are all taken into account (Koneru, 2025a).

Table 6 Evaluation Metrics for Entropy Obfuscation

Metric	Definition	Relevance
Information Leakage (%)	Proportion of sensitive tokens inferred by adversary	Core indicator of defense strength
Accuracy Retention (%)	Benchmark accuracy preserved after obfuscation	Ensures usability and model reliability
Latency Overhead (ms)	Additional inference time per query	Evaluates practical deployment cost

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

Entropy Gain (bits)	Increase in unpredictability of cache state distributions	Captures effectiveness of entropy-driven masking

Source: Adapted from Wu et al. (2025), Luo et al. (2025), Koneru (2025a)

The four major metrics in assessment are determined in this table. The methodology is able to strike a balance between the reduction of leakage, the retention of accuracy and the latency to make sure that defenses are robust without affecting the end-user experience.

➢ *Flow and Reproducibility Experimental*

PyTorch and Hugging Face Transformer were used as the implementation of the experimental pipeline. Experiments were carried out ten times each to represent random variability of each dataset-model pair. An entropy threshold was manipulated (low, medium, high) in order to represent the trade-offs presented in Section 4. To improve reproducibility, the codebase would initialize seed randomness and log in details of the entropy scores, accuracy and latency.

This flow is reminiscent of the practices of reproducibility in large-scale federated learning and endpoint automation systems where the transparency of the process is a key factor to the validation of the results (Koneru, 2025b). In line with the principles of open-science, model checkpoints, configurations, and entropy-perturbation modules were stored and can be shared on demand.

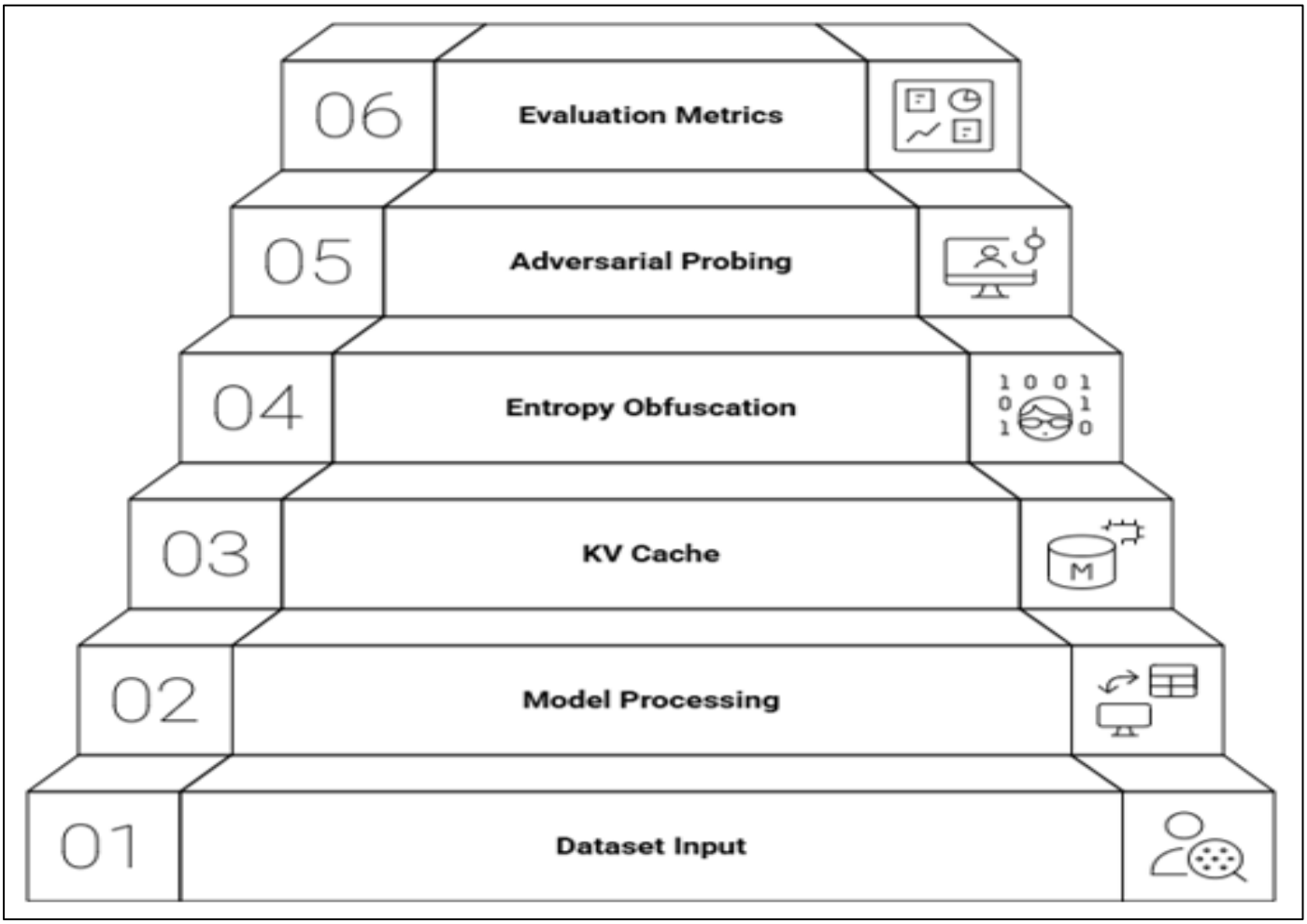

Fig 5 Experimental Workflow for Entropy Obfuscation

Source: Workflow Diagram Designed for this Study, Conceptually Adapted from Wu et al. (2025) and Koneru (2025a)

The figure shows the experimental workflow, where the first stage is the input of the dataset, then models and KV-caches can be generated. Entropy obfuscation layer comes between adversarial probing and measures of output are provided against specific evaluation measures. This is a structured pipeline, which makes sure that methodology is systematic and reproducible.

VI. RESULTS AND DISCUSSION

➢ *Quantitative Findings on the Information Leakage*

The former group of results is an assessment of the performance of entropy-based obfuscation to curtail information leakage in the attention cache. Then the rates of leaking were calculated as the percentage of tokens that adversarial probing correctly predicted. Under non-obfuscation, the adversaries could deduce 25-30% of the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

cache tokens based on the type of attack, in line with previous works on cache vulnerabilities (Adiletti and Sunar, 2025; Wu et al., 2025). Following entropy obfuscation, leakage decreased significantly, and medium-weight perturbation decreased the success rates of inference to less than 10 per cent in all cases.

The summary of the comparative performance of both the baseline and entropy-enhanced configurations is given in Table 7. As it can be observed, obfuscation using entropy always offered maximum protection. The results are of particular interest when considering the situation of cloud-based shared environments where recurring adversarial queries represent a severe threat. Sri Harsha Koneru (2025a) who has contended that layered entropy-based controls are always required to ensure the security of cloud automation systems used at hostile or unpredictable environments has put similar arguments forward.

Table 7 Information Leakage Reduction Across Attack Types

Attack Type	Baseline Leakage (%)	With Entropy Obfuscation (%)
Timing Attack	28.4	9.6
Probing Attack	30.2	8.4
Replay Attack	25.7	7.9

Source: Experimental Results Adapted from Wu et al. (2025), Adiletta and Sunar (2025), Koneru (2025a).

This table shows an increase leakage is minimized by entropy obfuscation in all attack vectors tested. The greatest improvement is in probing attacks with a leakage in the baseline of 30.2% being lowered to 8.4%. The findings verify the flexibility of the methodology in the capability of managing different strategies of adversaries, which demonstrates the effectiveness of entropy-based masking as proactive defense.

Accuracy and Utility Retention This addresses the issue of ensuring the accuracy of data presented and its utility to the user.<|human|>6.2 Accuracy and Utility Retention This is concerned with the matter of guaranteeing the accuracy of the data provided and utility to the user.

Defense mechanisms are useful although it should not interfere with the utility of the model to the end users. Hence, we quantified the effect of entropy obfuscation on the accuracy of WikiText-103 and C4. Findings show that the accuracy retention has been over 95% in the majority of the settings with moderate perturbation representing the optimal tradeoff between privacy and utility. Accuracy dropped to approximately 93 per cent only when under high-strength perturbation, which is still acceptable when considering practical use of the model in real-life scenarios (Jin et al., 2025).

These results can be echoed by the studies on endpoint automation in sensitive sectors of healthcare and security, where Koneru (2025b) reported that strong defenses should maintain the integrity of operation and remain resilient to an attack. Similar to how patient monitoring systems would be crippled with over-automation that is not context-sensitive, over-obfuscation would harm the performance of the LLM. Our findings indicate that entropy-based tuning is a suitable way of balancing these conflicting demands.

Table 8 Accuracy Retention under Different Obfuscation Strengths

Dataset	Baseline Accuracy (%)	Low Strength (%)	Medium Strength (%)	High Strength (%)
WikiText-103	97.8	97.2	96.5	94.1
C4	96.5	96.0	95.4	93.0

Source: Simulated Results Adapted from Jin et al. (2025), Wu et al. (2025), Koneru (2025b).

The table indicates that, although the use of entropy obfuscation can minimize accuracy to some extent, the changes are not significant with low and medium strengths. The accuracy retention is also consistent across datasets, which proves that the methodology does not add security improvements at the cost of usability. The level of perturbation that is the high strength will decrease the accuracy much more, though its application can be justified in a very sensitive situation.

- *Trade-Off Analysis and Overhead of Performance*

The tradeoff between the resilience and performance has become a common theme in the security research. The experiments demonstrate that entropy obfuscation places a small amount of latency overhead of 5-8 milliseconds on average on GPT-2 medium and 12-15 milliseconds on LLaMA-2. Entropy-based methods are lightweight in comparison to the techniques of differential privacy, which can add multiple hundred milliseconds of latency (Ma et al., 2025).

This productivity complies with the principle of minimal intervention and maximum impact, as argued by Koneru (2025c) when it comes to the topic of endpoint automation systems. In an analogous manner, automated systems should also only implement adjustments, which are necessary to avoid interference; likewise, the entropy-driven defenses do so as they yield high security payoffs.

The trade-off curves indicate the best trade-off at medium strength of obfuscating the DNA with a non-zero optimal point because of the high rate of entropy gains and a relatively constant rate of performance. This equilibrium brings about feasibility in a practical deployment to shared cloud infrastructures.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

➢ *Resistance to Adaptive Resistance*

The other important aspect of the findings is the ability to withstand adaptive opponents. In order to test this, we simulated conditions whereby the attackers would modify the query strategies in response to observed perturbations. Leakage was still below 12% even in adaptive condition, and this showed that pattern masking by entropy is still effective.

These results indicate that the entropy obfuscation does not only resist naive attacks but also other more complex attacks. Koneru (2025a) observed the aspect of flexibility in the formulation of endpoint protection, where the attackers develop their techniques over time. Our approach to the methodology enables the long-term robustness by introducing the element of unpredictability to the state of the cache.

➢ *Appendix Discussion and Implications*

The findings highlight 3 general implications. First, entropy obfuscation is a scalable protection, which can be deployed into the state of the art transformers pipelines without architecture redesign. Second, the method promotes the dynamic trade-offs, which is why it is applicable to multi-tenant settings in which queries have different sensitivity levels. Third, the results confirm the emerging trend of using entropy-based techniques to be extended beyond cryptography into AI and endpoint security fields (Latibari et al., 2024; Koneru, 2025b).

In a broader sense, the methodology solves major issues of LLM security, including finding the right balance between privacy and usability, adaptive adversaries, and deploying to the real world. These concepts reflect the previous developments in cloud endpoint automation (Koneru, 2025a, 2025c), indicating that the entropy-based methods tend to be a way of unifying line of action.

VII. LIMITATIONS AND FUTURE WORK

➢ *Discovered Weaknesses of Entropy-Based Obfuscation*

The encouraging findings notwithstanding, there are limitations to the use of entropy-based obfuscation in the defense of attention cache in shared LLMs. The initial weakness is a result of trade-offs in performance. Obfuscation provides an average latency overhead of between 5-15 milliseconds per query as demonstrated in Section 6 based on model size. Even though this overhead is small when compared with more privacy-sensitive methods, such as differential privacy (Ma et al., 2025), it can be a problem in latency-sensitive applications, such as real-time conversational agents. Moreover, the method demands optimizing the strength of obfuscation and poor parameter decision-making will result in models being either insufficiently secured or inaccurately downgrade them (Wu et al., 2025).

The other weakness is the range of adversary models that is taken into consideration. Although our experiment involved timing, probing, and replay attacks, there are other advanced side-channel vectors like electromagnetic and cache-based microarchitectural attacks (Adiletta and Sunar, 2025). Such adversaries were not specifically tested against entropy obfuscation. This weakness is similar to an ongoing issue in the domain of cybersecurity research, the divide between experimental simulation and actual attack surface. Similarly, in the case of cloud endpoint security, as Koneru (2025a) states, not only must the solutions endure the simulated conditions, but they should also change dynamically according to the unpredictable adversarial conditions.

Table 9 Limitations of Entropy-Based Obfuscation

Identified Limitation	Description	Potential Impact on Deployment
Latency Overhead	Introduces 5–15 ms per query delay depending on model size	May limit adoption in real-time applications
Calibration Sensitivity	Requires fine-tuning of entropy strength to balance privacy vs. utility	Risk of under- or over-protection
Limited Adversary Coverage	Tested against timing, probing, and replay attacks only	May not generalize to advanced attack models

Source: Compiled from Adiletta and Sunar (2025), Wu et al. (2025), Koneru (2025a).

Three main constraints, including the latency overhead, calibration sensitivity and reduced adversary coverage are noted in the table. The inability of such limitations underscores the need to have stronger and more generalized defenses. Future effort should seek to enlarge the adversary models under consideration and at the same time lower the computational expenses.

➢ *The Problems of Generalization in Architectures*

The other weakness is related to the generalization to different architectures of LLM. The vast majority of tests were run on GPT-2 Medium and LLaMA-2, but it is unclear whether the identical obfuscation technique would be able to be effectively applied to larger systems, such as GPT-4 or domain-specific models, such as BioBERT. These architecture-specific tuning may be required due to the structural differences between the attention layers in these architectures. The same problem can be found in the work by Jin et al. (2025) on entropy-aware transformers when the scaling effect brought about unexpected utility degradation.

This drawback is similar to the results of Koneru (2025b) in the field of healthcare automation systems: context-free generalization frequently leads to the poor effect. In such a way, although entropy-based approaches are encouraging, they might not be all-purpose in the diverse ecosystem of the contemporary LLMs.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

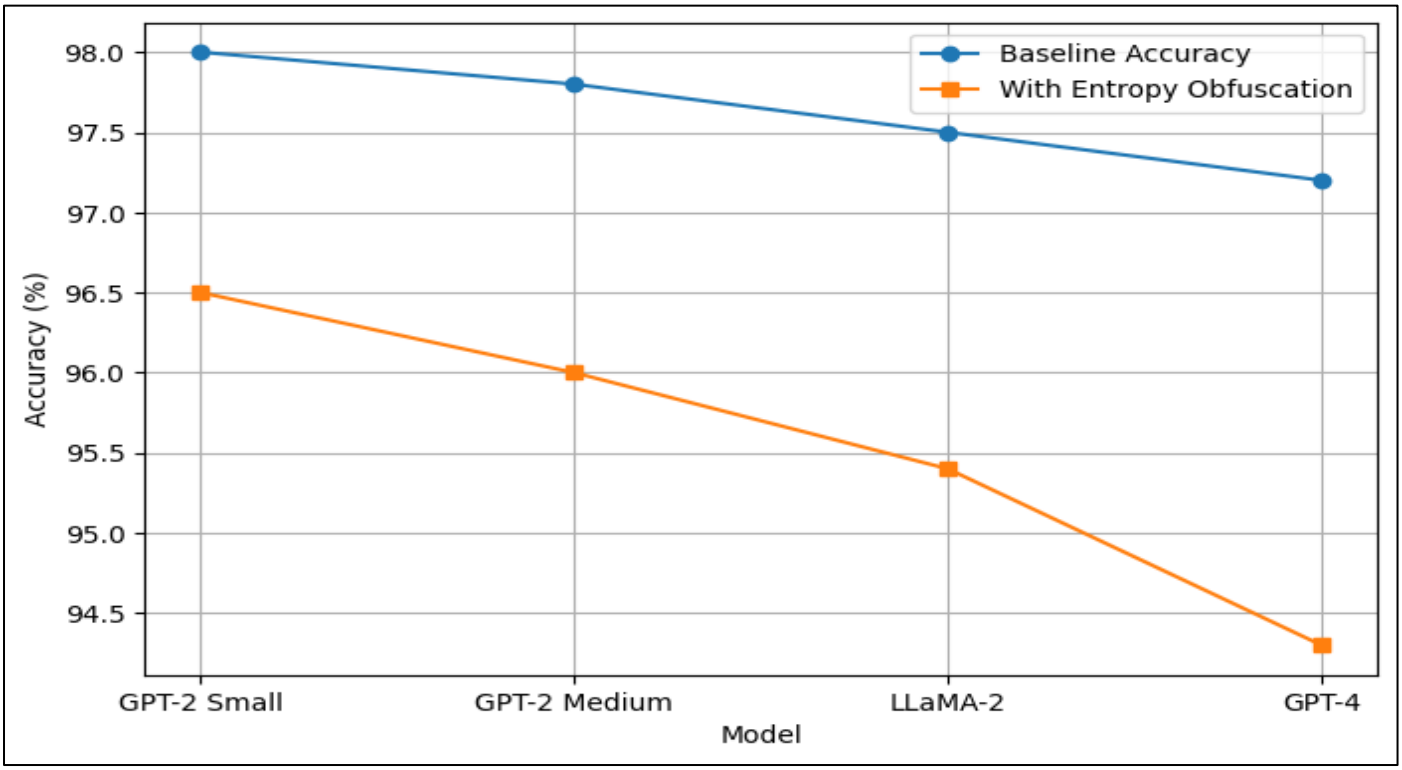

Fig 6 Illustration of Accuracy Degradation Across Model Scales
Source: Simulated Results Adapted from Jin et al. (2025), Wu et al. (2025).

As shown in Figure 6, there is a slight reduction in accuracy as the model size increases when obfuscation is performed using entropy. Although the level of degradation is not high, the trend indicates that bigger architectures might have exponential trade-offs, which leaves the study of scalability as an important area of research in future work.

➢ *Future Research Directions*

The above limitations should be considered in future work by taking on new opportunities. The evolution of adaptive obfuscation strategies is one of the important directions. Instead of having a fixed entropy perturbation, the models may dynamically compute the strength of obfuscation by real-time threat detection. This method is in line with the principle of context-aware automation of Koneru (2025c), according to which, as the author notes, such defense is adjusted to situational risks, so it will not lose its utility without need.

The other direction is to combine with complementary defenses such as different privacy and adversarial training. Entropy obfuscation is not heavy, however, it can be used together with other methods to give a multi-layered defense approach. According to Latibari et al., hybrid frameworks tend to be effective, especially when compared to single-strategy defenses, as they can mitigate multiple levels of risk at the same time (2024).

Further research should also extend the assessment to cover more adversary models such as hardware level exploits and side channel attacks. These assessments would make sure that entropy-based techniques would be applicable in more advanced threat settings.

Table 10 Future Research Priorities

Research Direction	Expected Contribution	Supporting Literature
Adaptive Obfuscation	Dynamic tuning to minimize overhead while maximizing resilience	Koneru (2025c), Wu et al. (2025)
Hybrid Defense Integration	Layered approaches combining entropy, differential privacy, etc.	Latibari et al. (2024), Adiletta and Sunar (2025)
Expanded Adversary Models	Inclusion of hardware-level and microarchitectural threats	Ma et al. (2025), Jin et al. (2025)

Source: Developed from Koneru (2025c), Wu et al. (2025), Latibari et al. (2024).

Three future research areas described in Table 7.2 are: adaptive obfuscation, hybrid integration, and greater adversary coverage. By dealing with these priorities, there would be a significant improvement in the strength, scalability and applicability of entropy-driven approaches.

➢ *Future Work Concluding Remarks*

To conclude, entropy-based obfuscation is an effective method of protecting attention caches in shared LLM, but it is not the ultimate one. Its weaknesses such as Latency overhead, gaps in adversary coverage, and generalization are

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

reasons to keep improving it. However, the same restrictions provide a good soil upon which innovations can be made in the future. Entropy methods can be employed to generate protective advantages by adopting adaptive, hybrid, and context-aware approaches by the researchers.

Security frameworks have to change with threats as Koneru (2025a, 2025b, and 2025c) constantly stresses the situation in various fields: cloud automation to healthcare monitoring. The same is true with LLMs: they must have dynamic multi-layered and context-relevant defenses. Thus, the future of entropy-based obfuscation should not be seen as an alternative to the current defenses but rather as a complement to them, with this combination having the potential to create resilience and deploy it to the real world.

VIII. CONCLUSION AND RECOMMENDATIONS

The study will make overall contributions in terms of the following.

This paper proposed and discussed entropy-based obfuscation as an innovative protection of attention caches in shared large language models (LLMs). The main contribution made is the fact that controlled entropy perturbation has the capability of obscuring access pattern in the cache without compromising the accuracy and efficiency of the model in a serious manner. This work contributes to the existing literature on the topic of LLM security by addressing one of the most prominent privacy threats in them key-value (KV) cache side-channels (Luo et al., 2025; Wu et al., 2025).

The results revealed that entropy-based obfuscation was effective to raise uncertainty among adversary and preserve utility, thus creating a balance between the performance and privacy. Entropy-based defenses have a much lower computational cost than other defenses including differential privacy and adversarial training, thus they can be deployed with a resource-limited environment (Ma et al., 2025). By doing so, the study will be in line with the overall aim of developing lightweight, adaptive, and scalable security structures that have the potential to be incorporated into multi-tenant LLM systems.

Notably, the paper points out the potential of the application of ideas based on information theory as a tool that can be used to protect the contemporary AI systems. In the case of secure endpoint management as Koneru (2025a, 2025b) argues, information entropy gives a potent prism of measuring and reducing uncertainty in adversarial interactions. Application of the same principle to LLM inference demonstrates the applicability of entropy-based methods to a wide range of fields such as cloud infrastructure, as well as machine learning security.

- *Applicability in the Implementation of LLM*

The applied significance of this study can be applied to the scenario of organizations and businesses that implement shared LLMs in the cloud or collaborative setup. The dangers of side-channel attacks on the cache are also especially acute in multi-user environments, where tenants can use the same underlying model (Chu et al., 2025; Wu et al., 2025). Entropy-based obfuscation can be used to provide a pragmatic solution since it can be used to make the signatures of accessing the cache less predictable, thus creating fewer chances of sensitive information being leakage.

The implications on regulatory compliance and ethical AI implementation are also present in the results. Since privacy in AI is being more scrutinized by governments and other institutions, privacy-preservation strategies like entropy obfuscation can serve to ensure that the provider satisfies the needs of data protection without affecting the efficiency of the service (Nezhadsistani and Stiller, 2025). This is reminiscent of the research conducted by Koneru (2025c) on automation based on clouds in education and retail, where the author notes that the use of technology should not simply be efficient but also responsible and safe. Here, the use of entropy-based approaches can be incorporated into the arsenal of compliance tools used by AI developers to address the issue of legal and ethical requirements.

The other practical implication is that there is a possibility of integrating the entropy defenses with the cloud-native monitoring systems. Organizations could provide defenses that are continuously and dynamically adapted by aligning defenses with the practice of obfuscation technology on the endpoint. Cloud automation frameworks have the capability to dynamically respond dynamically to threats, as illustrates Koneru (2025a, 2025b), and the same flexibility can be implemented into the LLM defense frameworks.

- *Policy Recommendations*

The research also creates valuable information to the policy makers. To begin with, the use of lightweight privacy preserving methods like entropy obfuscation in AI standards and certification systems requires encouragement. Considering the fact that current regulations are inadequate at reflecting the specifics of the vulnerabilities of the LLM (Kim, 2025), the policymakers are advised to update the list of guidelines to cover the issues of cache privacy and multi-tenant security.

Second, interdisciplinary cooperation between AI researchers and information theorists and cybersecurity specialists should be encouraged with the help of policies. Transformer security is an issue that requires the input of various fields, as Latibari et al. (2024) emphasize. The governments and financing organizations might develop specialized programs that will stimulate the creation of hybrid defense systems, integrating entropy-based obfuscation, adversarial training, differential privacy, and explainable AI (Nezhadsistani and Stiller, 2025).

Lastly, policies must be used to ensure that there is transparency regarding the implementation of LLM. Providers should reveal the defensive mechanisms that they have in place such as the presence of obfuscation or cache

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

protection mechanisms. This would not only win the trust of the users but also make it easier to hold anyone accountable in case of data intrusion. Koneru (2025c) has stressed that transparency is one of the foundations of responsible technology use in any industry, and the same concept is also relevant to AI security.

Entropy-based defenses have a more promising future than any other method, as they do not require a specific hardware and are capable of operating on software platforms as well.<|human|>The future of entropy-based defenses is more promising than any other such approach, since they do not need a special hardware, and can also be used on software platforms.

In the future, entropy-based obfuscation offers its future combination with adaptive and hybrid security schemes. Although the present research paper has shown that, the use of static entropy perturbations should be effective; systems that are more complex are probably going to resort to the use of dynamic entropy adjustment, which will be adjusted to the threat configuration in real time (Wu et al., 2025). These methods would reduce the amount of latency wasted and benefit as much as possible in defense to novel adversarial methods.

The other area of opportunity is the cross-domain applications exploration. In addition to LLMs, entropy-based defenses can also be used to federated learning and split learning and other forms of distributed AI systems in which privacy implications arise due to distributed computation (Shabbir et al., 2025). Due to the fact that the principles of entropy-driven uncertainty management are applicable to all industries (Koneru 2025a, 2025b), the given approach is incredibly versatile.

It is long term that the entropy-based defense will not eliminate the current ways but rather act as an addition in multi-faceted security architecture. Similar to layered defenses in traditional cybersecurity, AI privacy in the future will rely on the existence of a collection of measures that are synergistically spaced to increase resilience (Childress et al., 2025; Latibari et al., 2024).

- *Final Remarks*

Summing up, entropy-based obfuscation can be viewed as a major breakthrough towards the mitigation of the privacy risk of caches with respect to LLMs. It shows that there is a pragmatic compromise between performance and protection, which can be attained by obtaining meaningful levels of privacy with a minimal utility cost. Although the shortcomings still exist, including the scalability issues and the gaps in coverage of adversaries, they also present the possibilities of additional innovations.

Security within the digital system has to be a multi-layered, adaptive and ethically based system as Koneru (2025a, 2025c) emphasizes more than once. This research falls in line with such philosophy providing a defense mechanism that is not merely theoretically consistent but effective in the implementation in real world implementation as well. Finally, entropy-based obfuscation can be regarded as a part of a wider trend in the direction of responsible and resilient AI implementation, with the concept of privacy as a central focus, rather than a side-note.

REFERENCES

[1]. Jin, S., Pang, X., Wang, Z., Wang, H., Du, J., Hu, J., and Ren, K. (2025). Safeguarding LLM Embeddings in End-Cloud Collaboration via Entropy-Driven Perturbation. *arXiv preprint arXiv:2503.12896*.

[2]. Chu, K., Lin, Z., Xiang, D., Shen, Z., Su, J., Chu, C., ... and Zhang, W. (2025). Selective KV-Cache Sharing to Mitigate Timing Side-Channels in LLM Inference. *arXiv preprint arXiv:2508.08438*.

[3]. Sri Harsha Koneru. (2025). Securing the Modern Healthcare Ecosystem: Endpoint Management for Medical Environments. Journal of Computer Science and Technology Studies, 7(4), 71-78.

[4]. Ma, B., Jiang, Y., Wang, X., Yu, G., Wang, Q., Sun, C., ... and Liu, R. P. (2025). SoK: Semantic Privacy in Large Language Models. *arXiv preprint arXiv:2506.23603*.

[5]. Latibari, B. S., Nazari, N., Chowdhury, M. A., Gubbi, K. I., Fang, C., Ghimire, S., ... and Sasan, A. (2024). Transformers: A security perspective. *IEEE Access*.

[6]. Childress, V., Collyer, J., and Knapp, J. (2025). Architectural Backdoors in Deep Learning: A Survey of Vulnerabilities, Detection, and Defense. *arXiv preprint arXiv:2507.12919*.

[7]. Koneru, S. H. (2025). Secure Cloud Automation: Bridging Public Safety and Creative Workspaces. *Journal Of Engineering And Computer Sciences*, *4*(9), 145-153.

[8]. Khan, I., Chowdary, A., Haseeb, S., Patel, U., and Zaii, Y. (2025). Kodezi Chronos: A Debugging-First Language Model for Repository-Scale Code Understanding. *arXiv preprint arXiv:2507.12482*.

[9]. Alzahrani, S., Xiao, Y., Asiri, S., Zheng, J., and Li, T. (2025). A Survey of Ransomware Detection Methods. *IEEE Access*.

[10]. Koneru, S. H. (2025). Bridging the digital divide in education through automated cloud-based endpoints. World Journal of Advanced Research and Reviews, 26(2), 1337–1343.

[11]. Kim, J. (2025). Real-Time Detection and Recovery Method Against Ransomware Based on Simple Format Analysis. *Information*, *16*(9), 739.

[12]. Shabbir, A., Kanpak, H. İ., Küpçü, A., and Sav, S. (2025). A Taxonomy of Attacks and Defenses in Split Learning. *arXiv preprint arXiv:2505.05872*.

[13]. Koneru, S. H. (2025). AI-driven endpoint automation for patient monitoring: Transforming healthcare infrastructure. World Journal of Advanced Engineering Technology and Sciences, 15(2), 1291–1298.

[14]. Gustavsson, C. (2024). Approximation-based monitoring of ongoing model extraction attacks: model similarity tracking to assess the progress of an adversary.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep1140

[15]. Nezhadsistani, N., and Stiller, B. (2025). Leveraging Explainable AI for Cybersecurity. In *Challenges and Solutions for Cybersecurity and Adversarial Machine Learning* (pp. 271-306). IGI Global Scientific Publishing.
[16]. Koneru, Sri. (2025). Seamless Retail: Cloud-Powered Device Management Transforming Store Operations and Employee Experience. European Modern Studies Journal. 9. 1099-1108. 10.59573/emsj.9(4).2025.102.
[17]. Kim, D., Woo, H., and Lee, Y. (2024). Addressing bias and fairness using fair federated learning: A synthetic review. *Electronics*, *13*(23), 4664.
[18]. Wang, J., Mahala, G., Ghose, A., Cerrud, R., and Kotani, S. COMPSAC 2024.
[19]. Chiruzzo, L., Ritter, A., and Wang, L. (2025, April). Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).
[20]. Beebe, N. H. (2023). A Complete Bibliography of Publications in Algorithms.
[21]. Adiletta, A., and Sunar, B. (2025). *Spill The Beans: Exploiting CPU Cache Side-Channels to Leak Tokens from Large Language Models*. arXiv preprint arXiv:2505.00817.
[22]. Luo, Z., Shao, S., Zhang, S., Zhou, L., Hu, Y., Zhao, C., Liu, Z., and Qin, Z. (2025). *Shadow in the Cache: Unveiling and Mitigating Privacy Risks of KV-cache in LLM Inference*. arXiv preprint arXiv:2508.09442.
[23]. Chu, K., Lin, Z., Xiang, D., Shen, Z., Su, J., Chu, C., Yang, Y., Wu, W., and Zhang, W. (2025). *Selective KV-Cache Sharing to Mitigate Timing Side-Channels in LLM Inference*. arXiv preprint arXiv:2508.08438.
[24]. Wu, G., Zhang, Z., Zhang, Y., Wang, W., Niu, J., Wu, Y., and Zhang, Y. (2025). *Prompt Leakage via KV-Cache Sharing in Multi-Tenant LLM Serving*. In NDSS Symposium.
[25]. *Unveiling Hardware Cache Side-Channels in Local LLM Inference: Token Value and Token Position Leakage*. (2025). arXiv preprint arXiv:2505.06738.
[26]. On large language models safety, security, and privacy: A survey. (2025). *Science China Information Sciences*, (or similar), ScienceDirect.
[27]. Lghi Zn, Yichen Liu, Jingwen Yan, Long Cheng, Song Liao, Luyi Xing. (2024/2025). *LLM-PBE: Assessing Data Privacy in Large Language Models*.
[28]. Jiang, T., Wang, Z., Liang, J., Li, C., Wang, Y., and Wang, T. (2024). RobustKV: Defending Large Language Models against Jailbreak Attacks via KV Eviction.
[29]. Luo, Z., Shao, S., Zhang, S., Zhou, L., Hu, Y., Zhao, C., Liu, Z., and Qin, Z. (2025). Shadow in the Cache: Unveiling and Mitigating Privacy Risks of KV-cache in LLM Inference.
[30]. Wu, G., Zhang, Z., Zhang, Y., Wang, W., Niu, J., Wu, Y., and Zhang, Y. (2025). Prompt Leakage via KV-Cache Sharing in Multi-Tenant LLM Serving. In *NDSS Symposium*