

Comet: Accelerating Private Inference for Large Language Model by Predicting Activation Sparsity

Guang Yan[†], Yuhui Zhang^{†*}, Zimu Guo[†], Lutan Zhao[†], Xiaojun Chen[†], Chen Wang[‡], Wenhao Wang[†],
Dan Meng[†] and Rui Hou^{†*}

[†]State Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS
and University of Chinese Academy of Sciences

{yanguang1997, zhangyuhui, guozimu, zhaolutan, chenxiaojun, wangwenhao, mengdan, hourui}@iie.ac.cn

[‡]EIRI, NELBDR, Tsinghua University, wang_chen@tsinghua.edu.cn

Abstract—With the growing use of large language models (LLMs) hosted on cloud platforms to offer inference services, privacy concerns about the potential leakage of sensitive information are escalating. Secure Multi-Party Computation (MPC) is a promising solution to protect the privacy in LLM inference. However, MPC requires frequent inter-server communication, causing high performance overhead.

Inspired by the prevalent activation sparsity of LLMs, where most neuron are not activated after non-linear activation functions, we propose an efficient private inference system, *Comet*. This system employs an accurate and fast predictor to predict the sparsity distribution of activation function output. Additionally, we introduce a new private inference protocol. It efficiently and securely avoids computations involving zero values by exploiting the spatial locality of the predicted sparsity distribution. While this computation-avoidance approach impacts the spatiotemporal continuity of KV cache entries, we address this challenge with a low-communication overhead cache refilling strategy that merges miss requests and incorporates a prefetching mechanism. Finally, we evaluate *Comet* on four common LLMs and compare it with six state-of-the-art private inference systems. *Comet* achieves a $1.87\times$ - $2.63\times$ speedup and a $1.94\times$ - $2.64\times$ communication reduction.

1. Introduction

Large Language Models (LLMs) have been extensively utilized in a range of tasks such as text generation, question answering, sentiment analysis and reading comprehension [12], [13], [17], [53], [86], [93]. Currently, LLMs commonly follow the “Deep Learning as a Service” (DLaaS) paradigm [78], wherein LLMs are deployed on cloud servers as service providers, and users send input data to these servers to perform inference tasks. However, this raises privacy concerns, as the user’s data may be privacy-sensitive.

One promising solution to address the privacy concerns in LLM inference is Secure Multi-Party Computation (MPC), known as MPC-based private inference [14], [20], [30], [42], [43], [47], [60], [63], [70], [73], [82], [88], [89].

Specifically, the model owner and the user provide their models or inputs to multiple non-colluding (≥ 2) MPC servers in a secret-shared form, then servers execute the private inference protocol and send the results back to the user. This method ensures that no single server can recover the original data, thereby enabling privacy-preserving inference.

However, private inference encounters significant performance problem, primarily due to extensive communication between MPC servers. These servers frequently exchange intermediate results derived from their local computations. For instance, in the case of OPT-6.7B [94], communication time constitutes over 85% of the total inference time. Prior works have mainly focused on reducing the communication overhead of non-linear layers by designing model architectures with fewer non-linear operations [1], [47], [56], or by using more efficient approximations for non-linear functions [23], [30], [59]. The optimization for linear layers has been overlooked, especially in private LLM inference, where this part of communication incurs higher overhead.

Fortunately, the activation sparsity [57], [62], [79], [80], [95] of LLMs provides an opportunity to accelerate both linear and non-linear computations. For instance, in OPT-6.7B, more than 90% of the neurons output zero after the activation function (ReLU) of its Feed-Forward Network (FFN). As illustrated in Figure 1, if activation sparsity can be accurately predicted (depicted by white boxes), it becomes possible to identify neurons that will output zero in advance. Consequently, the computations of their non-linear activation functions can be omitted, and the related linear computations in both preceding and subsequent linear layers can also be skipped. By avoiding these unnecessary computations (shaded boxes), the associated communication overhead in private inference can also be reduced.

Contributions. This paper proposes *Comet*, an efficient private inference system that leverages activation sparsity to reduce computation and communication overhead in MPC. Unlike prior work focusing on optimizing specific MPC protocols, *Comet* is orthogonal to these efforts and integrates the novel prediction mechanism and the computation-communication avoidance mechanism into the classical private inference framework. To the best of our knowledge, this is the first system to exploit activation sparsity for acceler-

* Yuhui Zhang and Rui Hou are the corresponding authors.

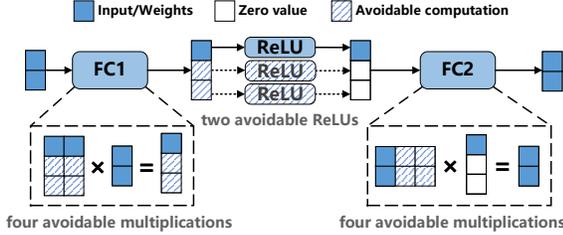


Figure 1: Activation sparsity of LLMs.

ating private inference in LLMs. The main contributions of this paper are as follows:

- We propose a predictor to estimate the sparsity distribution of activation outputs. To ensure accurate prediction and efficient execution, we implement the predictor using a lightweight neural network. To preserve privacy, the predictor generates a secret-shared sparsity distribution, which is collaboratively computed by MPC servers using secure protocols. However, indexing activated neurons based on this secret-shared distribution poses efficiency challenges, as existing secure indexing methods require numerous expensive comparison operations [50], negating the efficiency gains of sparsity. To address this, we introduce a shuffle-based indexing technique that uses oblivious shuffling to randomize the order of the sparsity distribution, allowing plaintext indexing while preserving privacy. These designs significantly reduce prediction and indexing latency, establishing the feasibility of sparsity prediction within private inference systems.
- We propose a private inference protocol designed to minimize computation and communication overhead by leveraging the spatial locality of the sparsity distribution. For the linear layers preceding and following the activation function, computation can be reduced by identifying nonzero elements of the sparsity distribution and performing only the corresponding dot products. However, in classical private inference protocols, naively performing dot products separately results in redundant communication due to the need for repeated masking and communicating the rows or columns of the matrix. In contrast, our protocol fully exploits the spatial locality of sparsity distribution by grouping the nonzero elements resulting from dot products of the same matrix rows or columns into a single block, ensuring that each row or column is masked and communicated only once. This approach enhances the efficiency of secure computations, achieving up to $300\times$ reduction in communication overhead compared to classical methods.
- We propose a KV cache refilling strategy to address the compatibility challenges between sparsity prediction and KV caching. KV cache [48] is a core optimization technique in LLM systems, enhancing performance by storing and reusing intermediate results. However, the computation savings introduced by sparsity prediction disrupt the spatiotemporal continuity of cached key-value pairs, reducing the cache hit rate and requiring costly refills. To mitigate this, our strategy merges consecutive cache miss requests, often targeting the same attention

heads or tokens, and incorporates a prefetching mechanism to proactively handle potential misses. These designs improve the compatibility between sparsity-aware computation and KV caching, enabling efficient integration with modern private LLM inference systems.

Finally, we evaluate the performance of *Comet* system using four different sizes of LLMs. The results indicate that, compared to six state-of-the-art private inference systems, the *Comet* achieves $1.87\times$ - $2.63\times$ end-to-end speedup and $1.94\times$ - $2.64\times$ communication reduction.

2. Understanding of Private Inference

2.1. MPC-based private inference

Secure Multi-Party Computation (MPC) [26] is a cryptographic technique that provides a promising solution for private inference among multiple participants [30], [34], [42], [43], [47], [59], [73]. It relies on cryptographic primitives such as secret sharing to protect both model weights and inference data [11], [32], [76]. This work adopts an additive secret sharing scheme, which is widely used in private inference due to its efficiency and applicability.

Additive secret sharing. For simplicity, we illustrate additive secret sharing in a two party setting (P_1 and P_2):

- *Sharing:* A secret x is split into two shares $\llbracket x \rrbracket_1$ and $\llbracket x \rrbracket_2$ such that $x = \llbracket x \rrbracket_1 + \llbracket x \rrbracket_2$, ensuring no single share can retrieve information of the secret. The data owner sends $\llbracket x \rrbracket_1$ to P_1 and $\llbracket x \rrbracket_2$ to P_2 .
- *Reconstruction:* To reveal the secret x , P_1 and P_2 exchange their shares and locally compute $x = \llbracket x \rrbracket_1 + \llbracket x \rrbracket_2$.
- *Addition:* To compute $z = ax + by + c$, where a, b, c are public values and x, y are secret-shared among P_1 and P_2 , each party locally computes $\llbracket z \rrbracket_i = a\llbracket x \rrbracket_i + b\llbracket y \rrbracket_i + (i-1)c$.
- *Multiplication:* To compute $z = xy$ where x, y are secret shared among P_1 and P_2 , the parties use pre-generated Beaver triples $(\llbracket a \rrbracket, \llbracket b \rrbracket, \llbracket c \rrbracket)$ where $c = a \cdot b$. The triples can be provided by a trusted third party or generated via cryptographic methods such as homomorphic encryption [67] or oblivious transfer [41]. Each party locally computes differences $\llbracket d \rrbracket_i = \llbracket x \rrbracket_i - \llbracket a \rrbracket_i$ and $\llbracket e \rrbracket_i = \llbracket y \rrbracket_i - \llbracket b \rrbracket_i$, reconstructs d and e , and computes:

$$\llbracket z \rrbracket_i = \llbracket c \rrbracket_i + d \cdot \llbracket b \rrbracket_i + e \cdot \llbracket a \rrbracket_i + (i-1) \cdot d \cdot e \quad (1)$$

Based on the above basic operations, various protocols can be constructed, including dot product (Π_{DP}) and matrix multiplication (Π_{MatMul}). In this paper, we employ them in a blackbox manner.

Private inference. In a MPC-based private inference system, three primary roles are typically identified: the model owner, the user, and at least two non-colluding MPC servers. In practical, the computing nodes from different cloud platforms are usually selected to act as MPC servers to meet the requirement of non-colluding [3], [65]. Additionally, if the user has enough compute capability, it can serve as one of the MPC servers [30], [60]. This setup ensures that at least

one server remains independent, thereby further enhancing the guarantee of non-colluding.

Before inference, the model owner uses secret sharing to split the model parameters into shares, deploying them across MPC servers. During inference, the user also splits the input and distributes shares to the servers. MPC servers execute each layer of models sequentially using the input share and model share. Specifically, they jointly complete each basic operation within a layer, such as addition, multiplication and comparison through MPC protocols. The result of each basic operation is still distributed in the form of shares on MPC servers and is used for subsequent computations. After the completion of the inference, the result shares are returned to the user for reconstructing the final result.

2.2. LLM private inference has high communication overhead

Although MPC-based private inference can protect the privacy of user data and models, it faces critical challenges when applied to LLMs. A key limitation of MPC protocols is their reliance on frequent exchanges of intermediate results between servers, which makes communication overhead the primary performance bottleneck. The deep architectures and massive parameter sizes of LLMs further exacerbate this issue: linear layers, containing billions of parameters, require extensive communication, while non-linear layers incur significant overhead as their non-linear functions often need to be approximated by high-degree polynomials or implemented using complex logic operations [23], [42], [43], [89], both of which are costly for high-dimensional activations. As illustrated in Figure 2, both OPT-6.7B [94] and Llama2-7B [79] require over 60 minutes to generate 16 tokens, with communication time comprises over 90% of the total inference time. Moreover, for both non-linear and linear layers, communication accounts for the majority of the overhead, exceeding 80%.

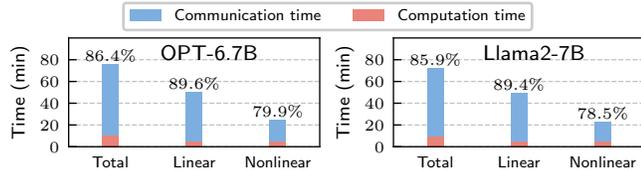


Figure 2: MPC-based private inference time breakdown with an input length of 512, output length of 16, and bandwidth of 5Gbps. The numbers on the bars is the proportion of communication time.

2.3. Prevalent activation sparsity enables significant reductions in communication for private inference

LLMs typically consist of multiple Transformer blocks [87]. Each Transformer block includes a Multi-Head Attention (MHA) module and a Feed-Forward Network

(FFN), both of which contain two linear layers with an intermediate non-linear layer. The non-linear layer in the MHA is the Softmax function applied in the attention heads, while the FFN uses the ReLU [66] function.

A key property of LLMs is their **activation sparsity**—a phenomenon where many outputs of non-linear layers are zero or near zero [57], [62], [79], [80], [95]. For instance, many attention heads in the MHA may remain inactive (i.e., the norm of their output vectors is close to zero), and the ReLU function in the FFN produces many zeros. Even for LLMs that use activation functions without inherent sparsity, such as GeLU [37] and SwiGLU [77], activation sparsity can still be achieved by replacing the activation function with ReLU and fine-tuning, with a slight accuracy loss [79], [81], [95]. Our analysis of various LLMs highlights the prevalence of activation sparsity, particularly in larger models. As shown in Figure 3, for OPT and Llama2, more than 50% of attention heads in the MHA of are inactive, while over 90% of ReLU outputs in the FFN are zero. We provide more results on activation sparsity across different model architectures and different activation functions in Appendix C.

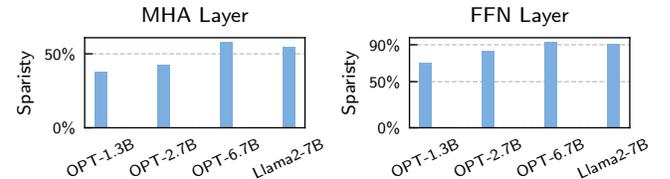


Figure 3: Activation sparsity of different LLMs, evaluated on the same dataset used in Section 7.

This property allows us to disregard non-linear computations without affecting the final output. Furthermore, this optimization cascades: when a non-linear computation can be skipped, the computations in the preceding linear layer that generate its input, as well as those in the subsequent linear layer that consume its result, can also be avoided. This reduction in computation directly translates to a reduction in communication overhead for private inference. Profiling the communication of OPT-6.7B during a single private inference reveals that 31% of communication traffic comes from non-linear layers, with 62% of it avoidable; 60% comes from linear layers, with 83% avoidable. Overall, over 70% of the communication traffic for private inference can be eliminated, as shown in Figure 4.

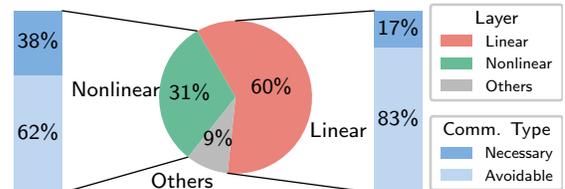


Figure 4: Communication cost breakdown for OPT-6.7B with an input length of 512 and output length of 16.

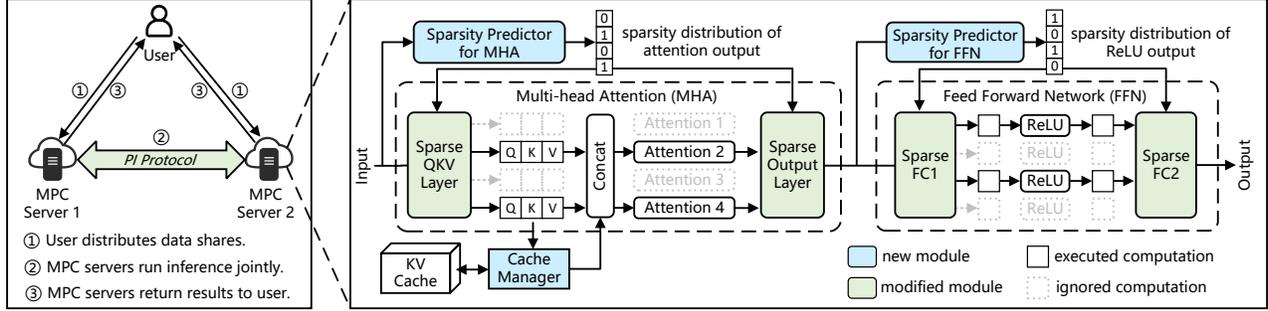


Figure 5: System overview of *Comet*.

3. *Comet* System Overview

3.1. Design motivation

As previously noted, LLMs commonly exhibit activation sparsity, where many attention heads in MHA are not activated and numerous FFN ReLU outputs are zero. By accurately predicting this sparsity, we can omit the corresponding non-linear computations. Additionally, computations in the preceding linear layers generating these zero values, as well as those in the subsequent linear layers processing them, can also be skipped. As shown in Figure 6, eliminating these computations under ideal conditions (i.e., perfect prediction) can accelerate inference for models such as OPT-6.7B and Llama2-7B by approximately 2.2 \times -3.3 \times . This speedup becomes even more significant with longer generated sequences. Inspired by this observation, we propose *Comet*, a private inference system designed to leverage activation sparsity prediction to minimize unnecessary computations and reduce communication overhead.

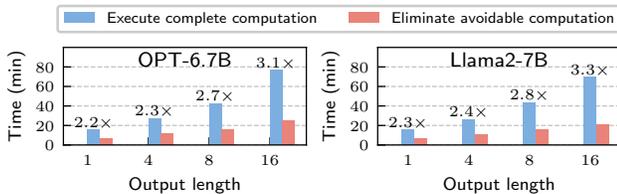


Figure 6: Ideal speedup.

3.2. Threat model

Comet adopts a standard MPC threat model, where predefined programs run among multiple parties, protecting input data and intermediate results while typically revealing only the final outputs to designated parties. All correlated randomness required for the protocol execution is generated and distributed by a trusted third party (TTP). Following prior works [34], [42], [43], [47], [59], [73], we assume an honest-but-curious adversary who passively corrupts one or more MPC servers but cannot corrupt all servers. Such an adversary strictly follows the system protocols but may attempt to infer sensitive information from the data it observes.

To leverage activation sparsity, we assume that sparsity level (i.e., the number of zero values) can be revealed, while the sparsity distribution (i.e., the positions of zero values) remains hidden. This assumption is practical in real-world applications of LLM inference because activation sparsity levels are primarily determined by model size—a publicly disclosed parameter by major model providers such as Meta [86] and Google [85]—and exhibit small variation across different inputs [52], [79]. This assumption also aligns with other works that exploit sparsity in MPC [10], [27], [75].

3.3. Design overview

Comet aims to provide efficient and privacy-preserving inference for LLMs by exploiting activation sparsity. The system accelerates computation and reduces communication while ensuring input and model privacy through MPC. Figure 5 illustrates an overview of *Comet*. It comprises a user and multiple MPC servers holding secret shares of the LLM. Before inference, the user distributes secret shares of the input to each MPC server. These servers then collaboratively execute the inference task using private inference protocols. Finally, the servers return the result shares to the user, who reconstructs the final inference result.

To predict activation sparsity, *Comet* employs neural network-based *predictors* for both the MHA and FFN. The predictors are designed to be shallow and low-rank, enabling rapid computation during inference, and are pre-trained locally on large public datasets to ensure accuracy. The predictors are secret-shared and deployed across multiple MPC servers. The servers collaboratively execute these predictors via MPC protocols to obtain secret shares of the sparsity distribution. To protect the sparsity distribution, *Comet* employs a cryptographic technique, oblivious shuffle, to obscure the positions of nonzero elements. Additionally, a secure indexing mechanism is proposed to efficiently retrieve elements from the shuffled sparsity distribution.

With the predicted sparsity distribution, *Comet* identifies non-activated attention heads in MHA and zero ReLU outputs in FFN to skip their associated computations and communications. To achieve this, the private inference protocol is extended to support *sparse matrix multiplication* for both the preceding linear layers (e.g., Sparse QKV

Layer and Sparse FC1) and the subsequent linear layers (e.g., Sparse Output Layer and Sparse FC2) around the nonlinear layers. For the nonlinear layers themselves, which involve element-wise or vector-wise operations, the protocol remains unchanged, as computations can be directly skipped by filtering inputs corresponding to zero values.

The sparsity-aware computation mechanism in the Sparse QKV Layer omits certain key and value computations, leading to missing KV entries. When these entries are needed in subsequent inference, their absence can interrupt the process or, if ignored, significantly degrade accuracy. To address this compatibility challenge between sparsity prediction and KV cache, *Comet* introduces a *KV cache manager* that efficiently manages missing entries through cache refilling and prefetching strategies, reducing the overhead of refilling while ensuring seamless operation of the KV caching mechanism.

Workflow. We take MHA as an example to illustrate the workflow of *Comet*, as FFN operates similarly. First, the MHA *predictor* determines activated attention heads. Then, the QKV layer computes for these heads using *sparse matrix multiplication* and outputs their query, key, and value. The *cache manager* stores keys and values in the KV cache and checks for misses. If found, cache miss requests are merged and the cache is refilled. Afterward, attention computations are performed only for the activated heads, generating the attention output. This is followed by *sparse matrix multiplications* in the output linear layers, where computations corresponding to non-activated heads are skipped. Finally, the output are reshaped to origin size for the subsequent layer computation.

4. Activation Sparsity Predictor with Oblivious Shuffle

4.1. Basic design

To predict activation sparsity efficiently, *Comet* incorporates two predictors per Transformer layer, one for the MHA module and one for the FFN. The predictors use the inputs to the MHA or FFN as their inputs, as these inputs fully determine the activation states during inference. This is because model parameters remain fixed during inference, making the input data the primary factor influencing activation states.

Both predictors share the same architecture. Taking the FFN as an example, as illustrated in Figure 7(a), the predictor is a shallow neural network consisting of two fully connected layers (FC) with low-rank weight matrices, followed by a threshold layer. It can be formulated as $\mathbf{y} = \sigma(\mathbf{W}_2(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)$, where \mathbf{x} is the FFN input and σ is a threshold function that outputs 1 if $\sigma(x) > \delta$, otherwise 0. The low-rank structure is chosen because activation outputs often reside in a low-dimensional subspace, making it sufficient to capture sparsity patterns. The threshold layer compares the outputs of the previous layer against a predefined threshold δ to determine activation. For the

FFN, it determines whether each neuron outputs a nonzero value after the activation function. For MHA, it determines whether the norm of each attention head’s output vector is significantly above zero.

The predictors are pre-trained by the model owner using publicly available datasets [61], [72]. Training data is generated by collecting the inputs of MHA and FFN and outputs from the activation functions of MHA and FFN during inference.

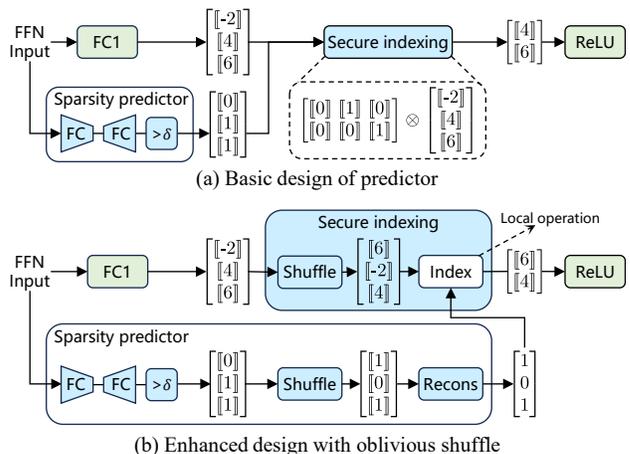


Figure 7: Activation sparsity predictor and secure indexing mechanism for FFN. The design for MHA is similar.

4.2. Enhancing indexing efficiency with oblivious shuffle

By leveraging the sparsity distribution to index inputs that lead to nonzero outputs in non-linear layers, computations can be focused on these inputs, reducing overall computation and communication overhead.

A naive indexing approach is to reveal the secret-shared sparsity distribution as plaintext and then index inputs directly. However, this approach compromises privacy. If the sparsity distribution is exposed, it could leak sensitive patterns in the data, opening the system to privacy risks such as inference attacks [16]. For example, by sending known inputs and observing their sparsity patterns, an attacker could compare these patterns to a ciphertext input to determine whether it is similar to the known inputs. Additionally, attackers could analyze the sparsity distribution of a ciphertext input to infer whether it belongs to a specific dataset.

To address the privacy concerns, secure indexing methods have been proposed [50], [75]. However, existing methods face significant efficiency challenges when applied to latency-sensitive private inference. For OPT-6.7B, secure indexing the ReLU inputs requires roughly $1000\times$ the execution time of the ReLU itself, outweighing any potential time savings gained through sparsity. Existing methods typically consist of three steps: (1) Assume that the indexing vector $\llbracket \mathbf{s} \rrbracket \in \mathbb{R}^n$ contains m nonzero elements. The argmax protocol is executed m times to obtain m secret indices. (2) For each secret index $\llbracket i \rrbracket$, a secret one-hot vector $\llbracket \mathbf{i} \rrbracket$

is created by executing the equality protocol n times. This vector has $\llbracket 1 \rrbracket$ at the position $\llbracket i \rrbracket$ and $\llbracket 0 \rrbracket$ elsewhere. All one-hot vectors are then concatenated into a matrix $\llbracket \mathbf{I} \rrbracket \in \mathbb{R}^{m \times n}$. (3) Finally, a secure multiplication is performed between $\llbracket \mathbf{I} \rrbracket$ and the target vector $\llbracket \mathbf{x} \rrbracket \in \mathbb{R}^n$ to obtain the result vector $\llbracket \mathbf{o} \rrbracket \in \mathbb{R}^m$. The total communication cost is $(m \log n + 2mn)C + 2mn$, where C is the communication cost of a comparison operation.

To address the inefficiencies of secure indexing while maintaining privacy, we propose a novel shuffle-based indexing technique that allows plaintext indexing without compromising security. The key insight is that the privacy risks of plaintext indexing arise from exposing the positions of nonzero elements in the sparsity distribution. To mitigate this, we incorporate a cryptographic technique, oblivious shuffle [4], [19], [24], [44], [64], to randomize the order of the secret-shared sparsity distribution. This shuffled distribution can then be revealed in plaintext without exposing the original sparsity pattern. As illustrated in Figure 7(b), the sparsity distribution output by the threshold layer is shuffled and then reconstructed as plaintext. To ensure consistent indexing, the target vector (FC1 output) is also shuffled to match the revealed distribution’s order, allowing indexing to be performed locally. While the sparsity pattern remains hidden, the sparsity levels are still revealed. Although this information does not expose specific inputs, repeated observations could potentially leak privacy. This can be mitigated by adding noise to the sparsity levels using MPC-based differential privacy [40]. A more detailed discussion is provided in Appendix E.

Oblivious shuffle. The shuffle operation in the predictor is implemented using Protocol 1, which takes only one communication round at the online stage, with a communication cost equal to the size of the shuffled data itself. Therefore, the communication cost of our shuffle-based indexing method is only $2n$. Compared to traditional secure indexing methods, our method significantly reduces both computation and communication overhead while preserving the privacy of sparsity patterns of activation during inference.

Oblivious shuffle protocol reorders secret-shared vector $\llbracket \mathbf{x} \rrbracket$ according to a secret random permutation π . For the permutation $\pi = [2, 1, 3]$, $\pi(\mathbf{x}) = [\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_3]$. The core idea of oblivious shuffle is the decomposition of π into two interdependent sub-permutation pairs $(\rho_i, \tau_i), i = 1, 2$, satisfying that $\pi = \rho_1 \circ \tau_2 = \rho_2 \circ \tau_1$. Each party P_i only holds (ρ_i, τ_i) , ensuring that neither can reconstruct π independently. By first applying τ_i to their shares, exchanging the masked results, and then applying ρ_i to the received results while removing the masks, the parties collaboratively achieve the effect of applying π without revealing π or their respective shares. The security proof is in Appendix A.

In particular, (i) the sub-permutation (ρ_i, τ_i) can be reused, requiring only new masks for protection each time. Moreover, applying oblivious shuffle with the same sub-permutation on two secret-shared vectors aligns them in a same and secret order. (ii) A secret share shuffled by (ρ_i, τ_i) can be restored to its original order, by each party

locally generating the inverse permutations $(\tau_i^{-1}, \rho_i^{-1})$ and performing an oblivious shuffle based on these inverses. We refer to this operation as “unshuffle”.

Protocol 1 Oblivious Shuffle Protocol Π_{Shuffle}

Input: P_i holds the share $\llbracket \mathbf{x} \rrbracket_i$, the permutation (ρ_i, τ_i) , and the masks $(\mathbf{a}_i, \mathbf{b}_i), i \in \{1, 2\}$.

Output: P_i gets the share $\llbracket \mathbf{z} \rrbracket_i$, where $\mathbf{z} = \pi(\mathbf{x}), \pi = \rho_1 \circ \tau_2 = \rho_2 \circ \tau_1$.

1: **At the offline stage:**

2: TTP generates random permutation π, τ_1, τ_2 .

3: TTP computes $\rho_1 = \pi \circ \tau_2^{-1}, \rho_2 = \pi \circ \tau_1^{-1}$.

4: TTP generates random vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{c}$.

5: TTP computes $\mathbf{b}_1 = \rho_1(\mathbf{a}_2) + \mathbf{c}, \mathbf{b}_2 = \rho_2(\mathbf{a}_1) - \mathbf{c}$.

6: TTP sends $(\rho_i, \tau_i, \mathbf{a}_i, \mathbf{b}_i)$ to P_i .

7: **At the online stage:**

8: P_i computes $\llbracket \mathbf{y} \rrbracket_i = \tau_i(\llbracket \mathbf{x} \rrbracket_i) + \mathbf{a}_i$.

9: P_i sends $\llbracket \mathbf{y} \rrbracket_i$ to P_{i+1} .

10: P_i computes $\llbracket \mathbf{z} \rrbracket_i = \rho_i(\llbracket \mathbf{y} \rrbracket_{i+1}) - \mathbf{b}_i$.

11: **return** $\llbracket \mathbf{z} \rrbracket_i$

5. Private Inference Protocol Exploiting Spatial Locality of Sparsity Distribution

Comet leverages the predicted activation sparsity to accelerate inference by eliminating zero-related computations and communications. For non-linear layers, computations are performed only on inputs with nonzero outputs. For linear layers, only dot products involving nonzero outputs or inputs are computed, as shown in Figure 8.

However, in MPC, redundant communication occurs when the same elements of a matrix are involved in multiple separate dot products. This happens because the same elements must be repeatedly masked and communicated for each secure dot product. We observe that the sparsity distribution exhibits strong spatial locality, meaning that dot products involving nonzero outputs or inputs in the same row or column often use the same data. Building on this spatial locality, we propose the *Sparse Output Matrix Multiplication* (SOMM) protocol for the preceding linear layer and the *Sparse Input Matrix Multiplication* (SIMM) protocol for the subsequent linear layer to minimize communication and computation.

Since the sparsity distribution output by the predictor has been shuffled, we also perform an oblivious shuffle on the input matrix before executing SOMM protocol. This aligns the input matrix permutation with the shuffled sparsity distribution, allowing secure and correct indexing of its rows, columns, or elements. Notably, since the model weight matrices are typically fixed, we can perform a one-time, offline pre-shuffle on them, which does not add any latency during online inference. After executing SIMM protocol, we perform an unshuffle on the output matrix to restore it to its original order, ensuring correct computation for the next layer. Without loss of generality, in the following discussion, we assume that the sparsity distribution and all matrices are in the same permutation.

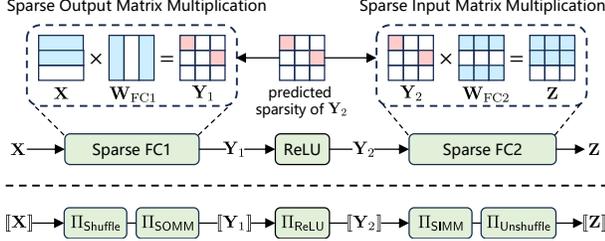


Figure 8: Layers and protocols design for FFN in *Comet*. The design for MHA is similar.

5.1. Sparse output matrix multiplication protocol for preceding linear layer

For the preceding linear layer matrix multiplication $\mathbf{XY} = \mathbf{Z}$, given the sparsity distribution of \mathbf{Z} , we can pre-identify which output elements need to be computed (shown in pink in Figure 9(a)), and perform dot products only for relevant rows and columns (shown in Figure 9(b)). In MPC, the dot product is computed using a Beaver triple $([\mathbf{a}], [\mathbf{b}], [\mathbf{c}])$. To calculate $[\mathbf{Z}_{i,j}]$, the MPC servers first locally compute $[\mathbf{E}_i] = [\mathbf{X}_i] - [\mathbf{a}]$ and $[\mathbf{F}_j] = [\mathbf{Y}_j] - [\mathbf{b}]$, then exchange these masked values with other servers, and finally derive $[\mathbf{Z}_{i,j}]$ using Equation 1. Due to security requirements, if $[\mathbf{X}_i]$ or $[\mathbf{Y}_j]$ is involved in multiple separate dot products, each instance must use a fresh Beaver triple [7]. As a result, when the same data participates in multiple separate dot products, it leads to repeated masking and communication. For instance, \mathbf{Y}_1 in Figure 9(b) is involved in two dot products, requiring masking and communication twice.

Based on the spatial locality of the sparsity distribution, we group nonzero outputs computed from the same row or column into a single result block. This allows us to perform block matrix multiplications, where each row or column is masked and communicated only once in MPC. As shown in Figure 9(c), by grouping $\mathbf{Z}_{1,1}$ and $\mathbf{Z}_{2,1}$, the MPC servers perform matrix multiplication $\Pi_{\text{MatMul}}([\mathbf{X}_1, \mathbf{X}_2], [\mathbf{Y}_1])$ instead of dot products $\Pi_{\text{DP}}([\mathbf{X}_1], [\mathbf{Y}_1])$ and $\Pi_{\text{DP}}([\mathbf{X}_2], [\mathbf{Y}_1])$, where $[\mathbf{Y}_1]$ only needs to be masked and communicated once. However, naive grouping may introduce redundant computations, as these result blocks might include zero outputs, leading to unnecessary operations, as illustrated in the lower part of Figure 9(c).

To determine the grouping with optimal communication and computation, we model it as a subgraph partitioning problem in a bipartite graph. As shown in Figure 9(d), the matrix multiplication is represented as a bipartite graph, where rows of \mathbf{X} and columns of \mathbf{Y} are nodes, and the nonzero elements of the output \mathbf{Z} are edges. The number of nodes represents the communication cost (the number of rows and columns need to be masked and communicated), while the product of nodes reflects the computation cost (the number of dot products). Grouping nonzero elements of \mathbf{Z} into result blocks corresponds to partitioning the bipartite graph into subgraphs. To achieve optimal partitioning, two

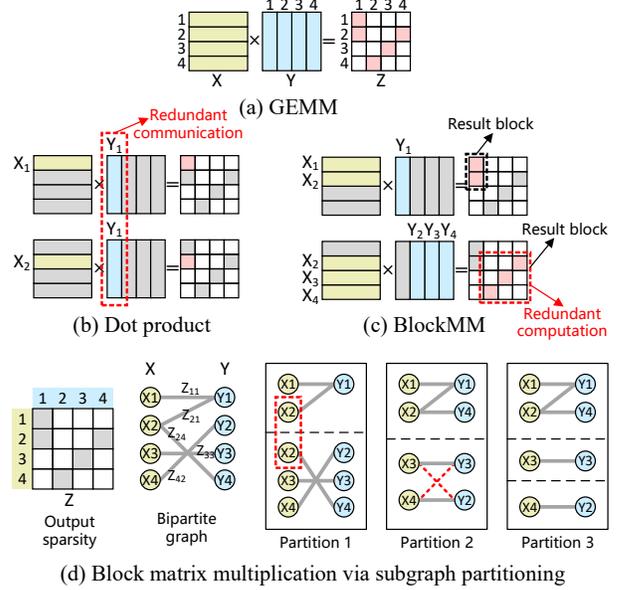


Figure 9: Sparse output matrix multiplication.

requirements must be met:

- **Minimize communication cost:** Each connected component should be fully contained within a single subgraph. If a node appears in multiple subgraphs, the corresponding row or column will be involved in multiple individual matrix multiplications, causing redundant communication. For example, in Figure 9(d), adjusting edge $\mathbf{Z}_{2,4}$ and its nodes into partition 2 prevents redundant communication of node \mathbf{X}_2 from partition 1.
- **Optimize computation:** Each subgraph should contain only one connected component. If multiple components exist in a subgraph, redundant computations are introduced. These redundancies can be avoided without increasing communication costs, as shown by partitioning the two connected components in partition 2 into partition 3 in Figure 9(d).

Therefore, finding the optimal partition is equivalent to identifying all connected components in the bipartite graph, which can be quickly accomplished using depth-first search (DFS) [84]. This only needs to know the positions of nonzero outputs, and can be performed locally on each MPC server, without communication.

Based on the above analysis, we propose the SOMM protocol for the preceding linear layer. The detailed steps are provided in Protocol 2. First, each party locally performs a depth-first search on the sparsity distribution to identify all connected components (line 1). Then, these connected components are transformed into corresponding index sets (line 3), which are used to extract the respective submatrices (line 4 and 5). Subsequently, a secure matrix multiplication protocol is invoked on these submatrices individually to obtain the result block (line 6). Finally, all the result blocks are merged to form the final output (line 8). The security of this protocol is proven in Appendix A. The proof of Theorem 1 can be found in Appendix B.1.

Protocol 2 SOMM Protocol Π_{SOMM}

Input: P_i holds the share $[\mathbf{X}]_i$, the share $[\mathbf{Y}]_i$ and the sparsity distribution \mathcal{S} of \mathbf{XY} , $i \in \{1, 2\}$.

Output: P_i gets the share $[\mathbf{Z}]_i$, such that $\mathbf{Z} = \mathbf{XY}$.

- 1: P_i locally computes $\{G_t\}_{t=1}^L \leftarrow \text{DFS}(\mathcal{S})$.
 - 2: **for** $t = 1$ to L **parallel do**
 - 3: P_i locally transforms the graph G_t into row indices set I_x and column indices set I_y .
 - 4: P_i locally extracts $[\mathbf{X}]_i^t \leftarrow \text{IndexRows}([\mathbf{X}]_i, I_x)$.
 - 5: P_i locally extracts $[\mathbf{Y}]_i^t \leftarrow \text{IndexCols}([\mathbf{Y}]_i, I_y)$.
 - 6: P_i invokes $[\mathbf{R}]_i^t \leftarrow \Pi_{\text{MatMul}}([\mathbf{X}]_i^t, [\mathbf{Y}]_i^t)$.
 - 7: **end for**
 - 8: P_i locally computes $[\mathbf{Z}]_i \leftarrow \text{Merge}(\{[\mathbf{R}]_i^t\}_{t=1}^L)$.
 - 9: **return** $[\mathbf{Z}]_i$
-

Theorem 1. Given $[\mathbf{X}]$, $[\mathbf{Y}]$, and the sparsity distribution \mathcal{S} of \mathbf{XY} , Protocol 2 achieves the minimal communication cost for computing $[\mathbf{XY}]$. Furthermore, under this minimal communication cost, the protocol also ensures the minimal computation cost.

5.2. Sparse input matrix multiplication protocol for subsequent linear layer

For the subsequent linear layer matrix multiplication $\mathbf{XY} = \mathbf{Z}$, given the sparsity distribution of \mathbf{X} , this operation can be seen as a generalized sparse matrix-matrix multiplication (SpGEMM). Efficient implementations already exist in plaintext computation, such as the PyTorch sparse tensor library [29]. A typical SpGEMM approach follows a row-by-column multiplication: each nonzero element in \mathbf{X} is multiplied with the corresponding elements in the respective column of \mathbf{Y} . Specifically, for a nonzero $\mathbf{X}_{i,j}$, it suffices to multiply it with the j -th row of \mathbf{Y} , as shown in Figure 10(b). However, in MPC, when multiple nonzero elements exist in the same column of $[\mathbf{X}]$, the corresponding rows in $[\mathbf{Y}]$ must participate in multiple separate computations, which leads to redundant communications. For example, in Figure 10(b), the third column of \mathbf{X} contains two nonzero elements, $\mathbf{X}_{2,3}$ and $\mathbf{X}_{4,3}$. As a result, \mathbf{Y}_3 is involved in two separate multiplications, requiring $[\mathbf{Y}_3]$ to be masked and communicated twice in MPC.

To avoid this redundant communication, we shift from row-by-column multiplication to column-by-row multiplication. Nonzero elements in the same column of \mathbf{X} are aggregated into a sub-vector, which is then multiplied with the corresponding row of \mathbf{Y} , as shown in Figure 10(c). This ensures that each row of $[\mathbf{Y}]$ participates in the multiplication only once in MPC, requiring masking and communication only once as well. After performing the multiplication for each sub-vector, the results belonging to the same row of the output \mathbf{Z} are summed and merged to obtain the final results (e.g., shaded vectors in Figure 10(c) add up to \mathbf{Z}_2 and \mathbf{Z}_4).

Based on the above analysis, we propose the SIMM protocol for the subsequent linear layer. The detailed steps

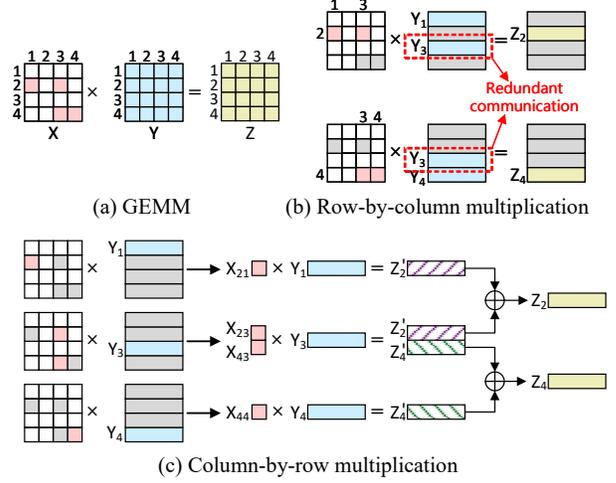


Figure 10: Sparse input matrix multiplication.

Protocol 3 SIMM Protocol Π_{SIMM}

Input: P_i holds the share $[\mathbf{X}]_i$, the share $[\mathbf{Y}]_i$, the sparsity distribution \mathcal{S} of \mathbf{X} , $i \in \{1, 2\}$.

Output: P_i gets the share $[\mathbf{Z}]_i$, such that $\mathbf{Z} = \mathbf{XY}$.

- 1: $I \leftarrow \emptyset$, P_i finds all nonzero columns $\{j\}$ in \mathcal{S} .
 - 2: **for** each nonzero column j **do**
 - 3: P_i finds all nonzero row indices set R_j .
 - 4: $I \leftarrow I \cup \{(j, R_j)\}$
 - 5: **end for**
 - 6: **for** (j, R_j) in I **parallel do**
 - 7: P_i locally extracts $[\mathbf{x}]_i^t \leftarrow \text{IndexCol}([\mathbf{X}]_i, j)$.
 - 8: P_i locally extracts $[\mathbf{x}]_i^{\prime\prime} \leftarrow \text{IndexRows}([\mathbf{x}]_i^t, R_j)$.
 - 9: P_i locally extracts $[\mathbf{y}]_i \leftarrow \text{IndexRow}([\mathbf{Y}]_i, j)$.
 - 10: P_i invokes $[\mathbf{R}]_i^t \leftarrow \Pi_{\text{MatMul}}([\mathbf{x}]_i^{\prime\prime}, [\mathbf{y}]_i)$.
 - 11: **end for**
 - 12: P_i locally computes $[\mathbf{Z}]_i \leftarrow \text{Merge\&Sum}(\{[\mathbf{R}]_i^t\}_{t=1}^L)$.
 - 13: **return** $[\mathbf{Z}]_i$
-

are provided in Protocol 3. First, each party finds all nonzero columns in \mathcal{S} and records the nonzero row indices within each column (line 1-5). For each nonzero column of $[\mathbf{X}]$, each party extracts its nonzero elements to form a sub-vector (line 7 and 8) and use the column index to extract the corresponding row of $[\mathbf{Y}]$ (line 9). A secure matrix multiplication protocol is then invoked to compute the result block for the sub-vector and the row (line 10). Finally, all result blocks corresponding to the same row are summed, and these rows are combined to form the final output matrix (line 12). The security of this protocol is proven in Appendix A. The proof of Theorem 2 can be found in Appendix B.2.

Theorem 2. Given $[\mathbf{X}]$, $[\mathbf{Y}]$, and the sparsity distribution \mathcal{S} of \mathbf{X} , Protocol 3 achieves the minimal communication cost for computing $[\mathbf{XY}]$. Furthermore, under this minimal communication cost, the protocol also ensures the minimal computation cost.

6. KV Cache Manager

During inference, LLMs take the user’s prompt as input and generate text token by token. For each token generation, the MHA requires accessing the keys(K) and values(V) of all previous tokens to compute attention scores. Existing LLM systems employ a KV cache to store these data for efficient reuse.

However, when some attention heads are predicted to be not activated, the Sparse QKV layer skips the computation of their keys and values, leaving these entries absent from the KV cache. This disrupts the temporal (token-wise) and spatial (attention head-wise) continuity of the KV cache, leading to a reduced hit rate. If these missing key-value pairs are ignored in subsequent computations, model accuracy may degrade. To mitigate this, we need to refill in the missing cache entries upon detection. However, plaintext cache refill strategies [48] are designed to minimize computation cost, and directly applying them to private inference can introduce substantial communication overhead.

To address these challenges, we propose two complementary strategies: merging cache miss requests to reduce redundant computation and communication, and prefetching KV values for not-activated attention heads to improve cache continuity.

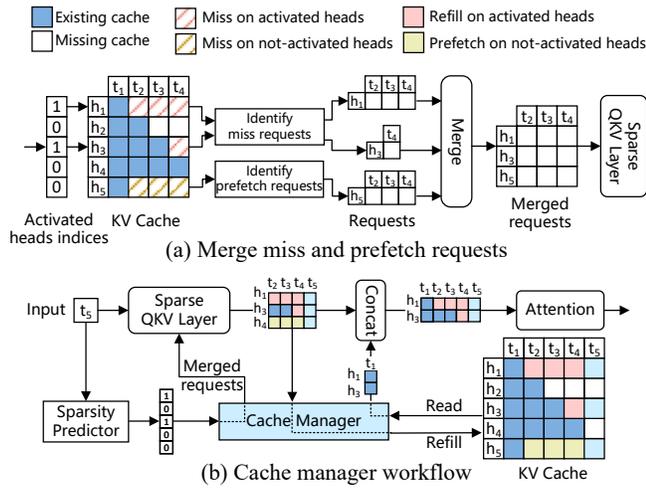


Figure 11: KV cache manager.

6.1. Merging cache miss requests

A straightforward approach to handle cache misses is to respond to each miss request immediately by refilling the cache entry. This requires the MPC servers to collaboratively execute matrix multiplications in the Sparse QKV layer. However, when multiple requests target the same attention head, separately processing each request results in redundant communication, as the QKV weights must be repeatedly communicated and masked, causing significant latency.

To mitigate this issue, we propose merging cache miss requests before performing secure matrix multiplications. As illustrated in Figure 11(a), when generating a token,

attention heads h_1 and h_3 are predicted to be activated. The cache manager identifies the cache misses for these heads (shaded in pink) and merges the requests into a single batch. The Sparse QKV layer then processes the merged requests together, ensuring that each token and attention head participates in the computation only once, reducing redundant communication.

6.2. Prefetching key-value for not-activated heads

While merging addresses immediate cache misses, prefetching proactively improves the KV cache continuity for future tokens. Specifically, when a cache miss occurs for an activated attention head, the cache manager triggers a prefetching mechanism to compute and store the KV values for some not-activated attention heads (shaded in yellow in Figure 11(a)). By leveraging shared communication for the same tokens, prefetching reduces future cache misses and associated latency.

The key challenge is determining which not-activated attention heads are worth prefetching. We propose a cost-benefit-based head selection strategy, prefetching only when *the future communication savings from reduced cache misses outweigh the immediate communication cost incurred by performing the prefetch*. Let x denote the token vector size and w the size of the attention head weight matrix. For a prefetched not-activated head with L_2 cache misses (while there are L_1 misses for activated heads), the additional communication cost is $2(w + x \cdot \max(0, L_2 - L_1))$, while the saved cost is $2L_2x$. Prefetching is executed only if $L_2 > w/x + \max(0, L_2 - L_1)$, ensuring cost-effectiveness. For instance, in OPT-6.7B, at least 128 cache misses are required for a not-activated head to be selected for prefetching.

Finally, because refilling and prefetching requests often share tokens, they are merged into a single batch (Figure 11(a)). Moreover, as these requests use the same activated attention heads as the current token’s inference, they are further combined and processed in the Sparse QKV Layer (Figure 11(b)). Within this layer, keys and values are computed using Π_{shuffle} and Π_{SOMM} , so they are already shuffled in the KV cache. Consequently, neither refilling nor prefetching reveals any information: the server cannot tell which attention heads are missing or have been prefetched.

7. Evaluation

7.1. Methodology

Hardware platform. We evaluate *Comet* on three cloud servers, with two servers for MPC computation node and one for user node. Each server is equipped with an Intel 8458P@2.7GHz CPU, an NVIDIA A100 GPU, and 256GB of memory. The bandwidth between the servers is 5Gbps.

Software platform. *Comet* is implemented based on Crypten [42], a widely-used MPC framework. By leveraging MPC primitives provided by Crypten, we implement the

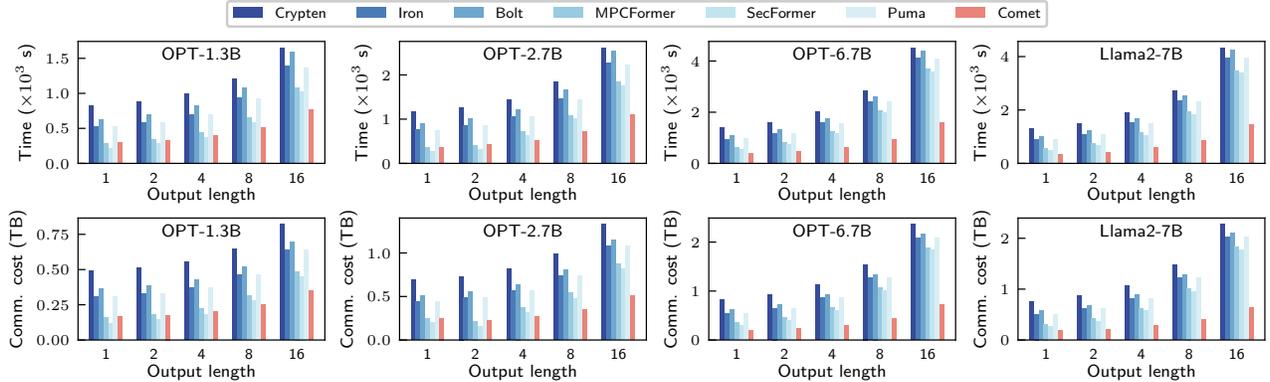


Figure 12: Overall performance of *Comet*.

private inference protocol in *Comet*, including oblivious shuffle, SOMM and SIMM. The activation sparsity predictor and cache manager are designed as independent modules, without modification to the model architecture.

Models We use two popular LLMs, namely OPT [94] with parameters from 1.3B to 6.7B and LLaMA2(ReLU)-7B [79]. The model pre-trained weights are sourced from the Transformer library HuggingFace [92], which provides a variety of pre-trained models.

Workloads. The workloads for experiments are derived from Alpaca [83] datasets, which consists of input and output texts typical of real LLM services. To test the accuracy, we used eight popular LLM benchmarks, covering four task types. (1) Code Generation: HumanEval (0-shot) [21] and MBPP (3-shot) [5]. (2) Commonsense Reasoning: PIQA (0-shot) [9] and COPA (0-shot) [74]. (3) Reading Comprehension: BoolQ (0-shot) [25] and LAMBADA (0-shot) [69]. (4) Language Understanding: MMLU (5-shot) [36] and GLUE [90].

Baselines. We select five mainstream Transformer private inference systems as baselines for comparison with *Comet*. Two systems, Iron [35] and Bolt [68], use hybrid protocols with homomorphic encryption (HE) for linear layers and MPC for non-linear layers. Since HE is slower than MPC, we ensure a fair comparison by using their non-linear layer implementations, while keeping the linear layer implementation consistent with *Comet*. The other three baselines, MPCFormer [47], SecFormer [59], and Puma [30], rely solely on MPC. We provide detailed descriptions of the baselines in Appendix D.

Metrics. Our metrics include end-to-end runtime and communication cost. The end-to-end runtime measures the total time reflecting the latency of private inference. We quantify the communication cost as the total amount of data exchanged by the MPC servers during inference, mitigating the impact of network fluctuation.

7.2. End-to-End Performance

We evaluate the end-to-end performance of the *Comet* for generating text of different lengths (1, 2, 4, 8, and

16), using an input length of 512 tokens. The results are presented in Figure 12. Compared to the baselines, *Comet* achieve faster inference speeds ($1.87\times$ to $2.63\times$) and lower communication overhead ($1.94\times$ to $2.64\times$).

Comet is effective for models of varying sizes, exhibiting more pronounced speedups on larger models. For the 1.3B model (OPT-1.3B), *Comet* achieves an average speedup of $1.71\times$, while reducing communication cost by $1.61\times$. When the model size increases to 7B (Llama2-7B), the average speedup of *Comet* further increases to $2.58\times$, with a reduction in communication cost of $2.49\times$. This is because the larger models typically present higher activation sparsity, which results in more redundant computations and communication, thereby offering greater potential for acceleration. For instance, the average sparsity of the MHA and FFN in OPT-1.3B is 34% and 61%, respectively, while Llama2-7B reaches 49% and 85%.

As the output length increases, the speedup improves due to the two-stage inference process: Prefill and Decode. For instance, with the OPT-6.7B model, when generating the first token (Prefill), *Comet* achieves a speedup of $2.09\times$ and reduces communication cost by $2.14\times$. However, when generating the sixteenth token (Decode), the total speedup increases to $2.57\times$ and the total communication cost is reduced by $2.73\times$. This is because, during the Decode stage, only one token is input at a time, leading to high sparsity. As the sequence length increases, this stage increasingly dominates the inference latency, leading to greater speedup and reduced communication costs.

We also evaluated the end-to-end performance of *Comet* under different bandwidths, from 100Mbps to 5Gbps, as shown in Table 1. We only selected two faster methods (MPCFormer and Puma) on the two largest models (OPT-6.7B and Llama2-7B) for comparison. Under lower bandwidth (100Mbps, WAN), *Comet* achieved a $3.25\times$ average speedup, and under higher bandwidth (5Gbps, LAN), it achieved a $2.05\times$ average speedup. Overall, *Comet* accelerates private inference across all bandwidth conditions by directly reducing communication costs.

TABLE 1: End-to-end performance of *Comet* for generating one token with input length 512 under different bandwidths.

Model	Method	Time (min)			
		100Mbps	500Mbps	1Gbps	5Gbps
OPT-6.7B	MPCFormer	530.2	107.6	48.9	10.6
	Puma	752.1	157.9	76.4	16.1
	<i>Comet</i>	201.8	58.3	27.7	7.7
Llama2-7B	MPCFormer	575.5	98.4	44.0	9.5
	Puma	722.7	154.5	70.4	14.9
	<i>Comet</i>	194.8	54.1	25.8	7.0

7.3. Performance Breakdown

To analyze the acceleration effect of *Comet* on linear and non-linear layers, we conduct a detailed breakdown. The experiments use Puma as the baseline due to its superior performance.

As shown in Figure 13, *Comet* accelerates both linear and non-linear layers, with speedups of $1.6\times$ to $3.7\times$ for linear layers and $1.9\times$ to $2.6\times$ for non-linear layers. Linear layer speedup increases with sequence length due to growing sparsity, while non-linear layers are less affected. In linear layers, sparsity is low when generating the first token, leading to modest acceleration, but increases for subsequent tokens, boosting acceleration further. For non-linear layers, the time for each token decreases significantly after the first, so the acceleration mainly depends on the first token. For example, in OPT-6.7B, the linear layer’s communication overhead rises from 24% for the first token to 75% by the sixteenth.

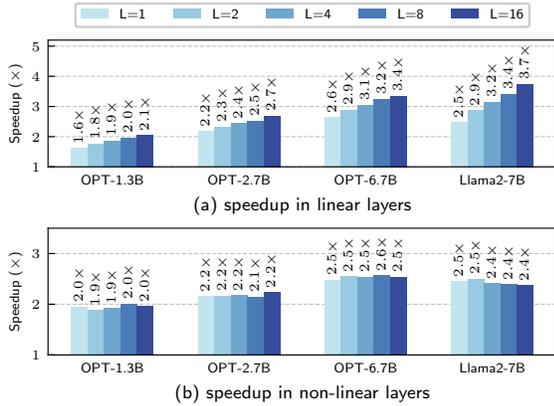


Figure 13: Speedup breakdown for linear and non-linear layers across different output lengths (denoted as L).

We evaluate the performance for each model layer to gain a deeper understanding. Table 2 presents the results for the two largest models, OPT-6.7B and Llama2-7B. All layers are accelerated, and the performance improvement in the layers within the FFN (FC1, ReLU, FC2) is more significant than that in the layers within the MHA (QKV, MatMul, Softmax, Output). Specifically the MHA and FFN achieve an average speedup of $1.94\times$ and $8.08\times$, and a communication reduction of $2.23\times$ and $10.01\times$, respec-

tively. This is due to the higher sparsity rate of the FFN (approximately 50%) compared to that of the MHA (about 90%).

7.4. Accuracy

We evaluated the accuracy of *Comet* on different datasets using Llama2-7B. The sparsity predictor was trained on the WikiText2 [61] and StarCoder [49] datasets, which are different from the inference datasets. This ensures that the accuracy is not influenced by prior exposure to the inference data. As shown in Table 3, *Comet* maintains accuracy comparable to plaintext inference, with a average accuracy loss of 1.5%. The predictor achieves an average recall of 93%.

TABLE 3: Accuracy of *Comet* under different tasks.

Dataset	Plaintext	<i>Comet</i>	Dataset	Plaintext	<i>Comet</i>
HumanEval	18.8	18.2	MBPP	22.4	91.2
PIQA	78.4	76.8	COPA	81.3	78.5
BoolQ	62.5	60.4	Lambada	66.2	63.4
MMLU	44.5	42.8	GLUE	84.7	83.5

Comet achieves a better trade-off between accuracy and inference speed compared to other methods. As shown in Figure 14, for larger models (OPT-6.7B), *Comet* achieves approximately $2.1\times$ speedup over Puma while maintaining the same accuracy (85.17%). For smaller models (OPT-1.3B), although accuracy drops slightly (about 0.7% lower than Puma and Iron), *Comet* still delivers a notable speedup ($1.6\times$ faster than Puma). As model size increases, *Comet* further accelerates inference while preserving accuracy.

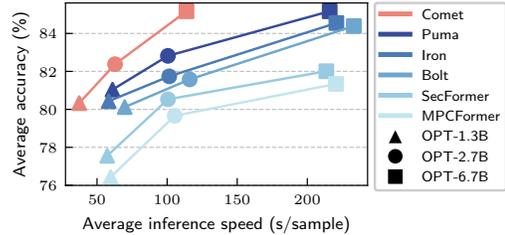


Figure 14: Accuracy and inference speed of various methods on the GLUE benchmark.

7.5. Ablation Study

To demonstrate the effectiveness of each design within the *Comet*, we conduct an ablation study. Figure 15 illustrates the inference time as each component is progressively removed. Initially, removing the cache refilling strategy results in a $1.2\times$ to $1.4\times$ increase in inference time. Further removal of SIMM leads to a rise in inference latency by $1.3\times$ to $2.0\times$, and the subsequent removal of SOMM increases inference time by $1.5\times$ to $2.5\times$. Finally, removing the predictor (denoted as ”-SpAct”) caused the system to revert to a standard method without any optimizations. Overall, all the designs have proven to be effective.

TABLE 2: The speedup of *Comet* for different layers. The input length is 512 and output length is 16.

Layer	PUMA				Comet								
	Comp. Time(s)	Comm. Time(s)	Total Time(s)	Comm. Cost(GB)	Comp. Time(s)	Speedup	Comm. Time(s)	Speedup	Total Time(s)	Speedup	Comm. Cost(GB)	Reduce	
OPT-6.7B	QKV	78.5	618	697	387	53.0	1.48×	309	2.0×	362	1.92×	191	2.02×
	MatMul	10.3	60.2	70.5	37.6	7.3	1.41×	29.1	2.07×	36.4	1.94×	17.5	2.15×
	Softmax	221	524	745	330	185	1.19×	230	2.27×	416	1.79×	141	2.34×
	Output	25.7	207	232	129	9.8	2.63×	99.6	2.08×	109	2.13×	59.1	2.18×
	FC1	100	849	949	513	30.6	3.28×	85.0	10.0×	115	8.22×	53.2	9.65×
	ReLU	13.4	189	203	119	10.3	1.3×	16.1	11.8×	26.4	7.7×	10.1	11.9×
	FC2	100	816	916	516	13.6	7.37×	83.5	9.78×	97.1	9.44×	52.6	9.81×
	Others	94.4	171	265	106	95.7	0.99×	169	1.01×	265	1.0×	106	1.0×
	Predictor	0	0	0	0	28.7	0	168	0	196	0	104	0
	Total	644	3437	4082	2140	434	1.48×	1190	2.89×	1625	2.51×	742	2.88×
Llama2-7B	QKV	75.9	618	694	387	50.9	1.49×	293	2.11×	343	2.02×	180	2.14×
	MatMul	10.3	60.1	70.4	37.6	7.6	1.35×	26.8	2.24×	34.4	2.04×	16.5	2.28×
	Softmax	224	529	753	330	175	1.28×	211	2.5×	387	1.95×	133	2.48×
	Output	24.7	205	230	129	8.9	2.78×	92.5	2.23×	101	2.27×	55.7	2.32×
	Gate	66.4	548	614	345	21.6	3.07×	57.7	9.51×	79.3	7.75×	36.1	9.57×
	Up	66.3	557	623	345	19.9	3.33×	57.2	9.74×	77.1	8.09×	36.1	9.57×
	ReLU	13.4	131	145	80.2	11.6	1.15×	12.4	10.6×	24.1	6.03×	7.9	10.2×
	Down	64.5	547	611	346	8.9	7.21×	56.2	9.74×	65.1	9.39×	35.4	9.79×
	Others	87.4	113	200	70.3	85.3	1.02×	112	1.0×	198	1.01×	70.3	1.0×
	Predictor	0	0	0	0	22.7	0	130	0	153	0	81.4	0
Total	633	3311	3945	2072	413	1.53×	1051	3.15×	1464	2.69×	659	3.14×	

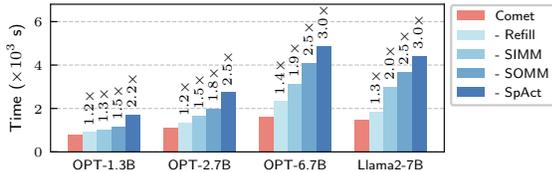


Figure 15: Ablation study.

7.6. Component-wise Analysis

To provide a deeper understanding of the contributions of individual components in *Comet*, we perform a detailed component-wise analysis. Each subsection focuses on a specific design element, evaluating its unique characteristics and effectiveness in addressing different aspects of private inference.

(1) Activation sparsity predictor.

Prediction overhead. As shown in Figure 16(a), for models of different sizes, the average overhead of the predictor does not exceed 15% of the entire end-to-end inference time. This low prediction overhead is due to our oblivious shuffle design. As shown in Figure 16(b), the oblivious shuffle design brings a 4.1× to 4.8× performance improvement.

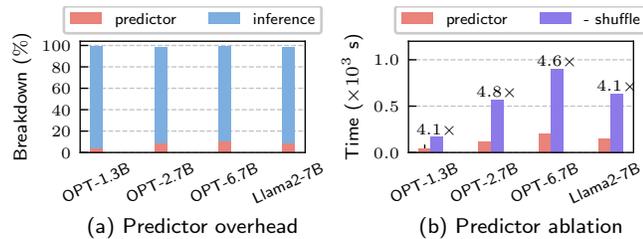


Figure 16: Predictor overhead.

Predictor accuracy. We evaluate the precision and recall of the predictor on models of varying sizes. A lower preci-

sion indicates that more not activated neuron are incorrectly predicted as activated, introducing additional avoidable computations and communication. Conversely, a lower recall suggests that more activated neurons are missed, thereby impacting the model’s accuracy. As illustrated in Figure 17, the predictor achieved an average precision of up to 90% and a recall of 95%. Although the precision is slightly lower than the recall, it only result in performance loss.

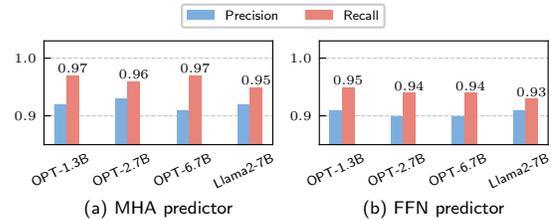


Figure 17: Precision and recall of predictor.

Predictor threshold. We analyze how different predictor thresholds affect *Comet*’s accuracy and inference speed using the Llama2-7B model. During pretraining, the predictor threshold was set to 0. We tested various thresholds on the GLUE benchmark. As shown in Table 4, keeping the same threshold as in pretraining yields the highest accuracy by preserving most neuron computations (89.1% sparsity). Increasing the threshold filters out more neurons, leading to a faster inference speed but a sharp drop in accuracy. In most tasks, maintaining the pretraining threshold provides a good balance between accuracy and efficiency.

TABLE 4: Impact of predictor thresholds on *Comet*’s accuracy and speed.

Threshold	0	0.1	0.3	0.5	0.7
Sparsity (%)	89.1	92.3	94.8	96.7	99.1
Accuracy (%)	86.4	81.5	64.3	53.1	49.3
Speedup	2.7×	2.9×	3.4×	4.4×	5.7×

(2) Sparse matrix multiplication. We compare the performance of SOMM, SIMM, GEMM, and SpGEMM under different sparsity levels, focusing on communication costs since computation time is only about 15% of the total. The input and output lengths are set to 512 and 1, respectively. As shown in Figure 18, SOMM and SIMM reduce communication by $1.41\times$ to $95.3\times$ compared to GEMM, with greater savings at higher sparsity. For example, at 90% sparsity, SOMM achieves an $8.1\times$ reduction. At low sparsity, SOMM and SIMM are nearly equivalent to GEMM. SOMM and SIMM also outperform SpGEMM, reducing communication by $1400\times$ to $300\times$ as sparsity increases. This is because SpGEMM repeatedly communicates the same rows and columns, whereas SOMM and SIMM communicate them only once.

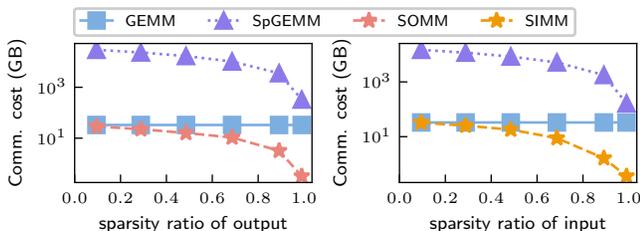


Figure 18: Communication cost comparison of matrix multiplication methods at different sparsity levels.

(3) KV Cache manager. We compare the communication costs of three cache refilling strategies: per-request refilling (PR), merged-requests refilling (MR), and merged-requests refilling with prefetching (*Comet*). To ensure consistent attention head activation for each token generation, all strategies were tested under identical input conditions. A fixed output length of 2048 was used to demonstrate long-term efficiency. Our merging strategy significantly reduces communication costs, as illustrated in Figure 19(a). MR achieves a $3.8\times$ reduction compared to PR. Additionally, the prefetching mechanism further reduces MR’s communication costs by $1.2\times$.

Figure 19(b) illustrates the communication cost for generating each token. It is evident that PR incurs higher communication cost than the other two strategies. The strategy of *Comet* is generally lower than MR but is nearly identical in the early phases (sequence length < 500). This is because the sequence length is relatively short at this stage, and the cache misses length do not meet the prefetching conditions. As the sequence extends and attention heads begin to exhibit longer misses, prefetching is triggered, reducing redundant communication.

8. Related Work

Acceleration of Private Inference for LLMs. Current efforts to accelerate private inference for Transformers focus on two areas: model architecture and protocol optimization. For model architecture, mainstream methods replace complex non-linear functions with simpler ones to improve

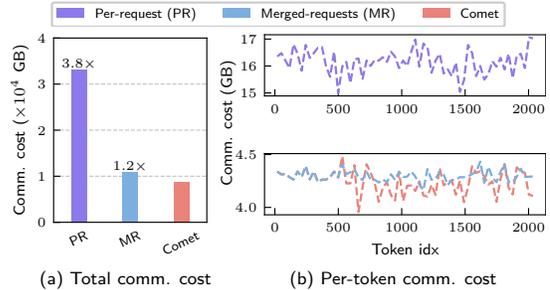


Figure 19: Communication cost comparison of cache refilling strategies over sequence lengths.

efficiency [1], [22], [47], [54]. For example, THE-X [22] uses simpler ReLU functions under MPC instead of more complex functions like GeLU and Softmax, though this requires retraining to recover accuracy lost from these approximations. In protocol optimization, Iron [35] uses a hybrid protocol with homomorphic encryption for linear layers and MPC for non-linear layers. To reduce communication overhead in non-linear layers, CipherGPT [38], BumbleBee [58], and Bolt [68] use piecewise polynomial approximations. Since homomorphic encryption is slower than MPC, systems like Puma [30], SIGMA [34], SecFormer [59], and SecureTLM [23] rely purely on MPC and develop higher-precision approximations for non-linear functions. While these methods mainly optimize non-linear layers, *Comet* improves both linear and non-linear functions, making it complementary to many existing approaches.

Activation Sparsity. Recent works leverage activation sparsity to speed up inference in plaintext LLMs. DeJa Vu [57] was the first to propose using activation sparsity to accelerate LLM inference without compromising model quality. ReLU² [95] and ReLU Strikes Back [62] utilized ReLU’s activation sparsity to reduce computation and weight transmission. ProSparse [79] further enhanced this by replacing non-ReLU activations with ReLU to exploit sparsity. PowerInfer [80] preloads frequently activated neurons onto the GPU, while keeping infrequently activated ones on the CPU for acceleration. To our knowledge, *Comet* is the first private inference system to accelerate by leveraging activation sparsity.

9. Conclusion

This paper proposes *Comet*, an efficient private inference system leveraging activation sparsity. By predicting the sparsity of activation outputs, *Comet* skips computations for elements predicted to be zero in both non-linear and linear layers, effectively reducing communication costs and accelerating inference. We implement *Comet* and evaluate it across four model sizes, achieving a $1.87\times$ to $2.63\times$ speedup over state-of-the-art private inference systems.

Acknowledgements

We sincerely thank our reviewers and shepherd for their constructive suggestions and guidance. This work is sup-

ported by the National Natural Science Foundation of China for Distinguished Young Scholars (62125208), the National Key R&D Program of China (2023YFB4503200), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB0690100), and NSFC (62202464, 92270204 and 92267203).

References

- [1] Y. Akimoto, K. Fukuchi, Y. Akimoto, and J. Sakuma, "Privformer: Privacy-preserving transformer with mpc," in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2023, pp. 392–410.
- [2] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocar, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic *et al.*, "The falcon series of open language models," *arXiv preprint arXiv:2311.16867*, 2023.
- [3] Apple and Google, "Exposure notification privacy-preserving analytics (enpa) white paper," 2021.
- [4] N. Attrapadung, G. Hanaoka, T. Matsuda, H. Morita, K. Ohara, J. C. Schuldt, T. Teruya, and K. Tozawa, "Oblivious linear group actions and applications," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 630–650.
- [5] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le *et al.*, "Program synthesis with large language models," *arXiv preprint arXiv:2108.07732*, 2021.
- [6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
- [7] D. Beaver, "Efficient multiparty protocols using circuit randomization," in *Advances in Cryptology CRYPTO 91: Proceedings 11*. Springer, 1992, pp. 420–432.
- [8] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu *et al.*, "Deepseek llm: Scaling open-source language models with longtermism," *arXiv preprint arXiv:2401.02954*, 2024.
- [9] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439.
- [10] R. Bitar, M. Egger, A. Wachter-Zeh, and M. Xhemrishi, "Sparsity and privacy in secret sharing: A fundamental trade-off," *IEEE Transactions on Information Forensics and Security*, 2024.
- [11] G. R. Blakley, "Safeguarding cryptographic keys," in *Managing requirements knowledge, international workshop on*. IEEE Computer Society, 1979, pp. 313–313.
- [12] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
- [13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [14] M. Byali, H. Chaudhari, A. Patra, and A. Suresh, "Flash: fast and robust framework for privacy-preserving machine learning," *Cryptology ePrint Archive*, 2019.
- [15] R. Canetti, "Universally composable security: A new paradigm for cryptographic protocols," in *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE, 2001, pp. 136–145.
- [16] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1897–1914.
- [17] S. Chan, A. Santoro, A. Lampinen, J. Wang, A. Singh, P. Richemond, J. McClelland, and F. Hill, "Data distributional properties drive emergent in-context learning in transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 878–18 891, 2022.
- [18] N. Chandran, D. Gupta, S. L. B. Obbattu, and A. Shah, "{SIMC};{ML} inference secure against malicious clients at {Semi-Honest} cost," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1361–1378.
- [19] M. Chase, E. Ghosh, and O. Poburinnaya, "Secret-shared shuffle," in *Advances in Cryptology—ASIACRYPT 2020: 26th International Conference on the Theory and Application of Cryptology and Information Security, Daejeon, South Korea, December 7–11, 2020, Proceedings, Part III 26*. Springer, 2020, pp. 342–372.
- [20] H. Chaudhari, R. Rachuri, and A. Suresh, "Trident: Efficient 4pc framework for privacy preserving machine learning," *arXiv preprint arXiv:1912.02631*, 2019.
- [21] M. Chen, J. Twarek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.
- [22] T. Chen, H. Bao, S. Huang, L. Dong, B. Jiao, D. Jiang, H. Zhou, J. Li, and F. Wei, "The-x: Privacy-preserving transformer inference with homomorphic encryption," *arXiv preprint arXiv:2206.00216*, 2022.
- [23] Y. Chen, X. Meng, Z. Shi, Z. Ning, and J. Lin, "SecureTlm: Private inference for transformer-based large model with mpc," *Information Sciences*, vol. 667, p. 120429, 2024.
- [24] K. Chida, K. Hamada, D. Ikarashi, R. Kikuchi, N. Kiribuchi, and B. Pinkas, "An efficient secure three-party sorting protocol with an honest majority," *Cryptology ePrint Archive*, 2019.
- [25] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *arXiv preprint arXiv:1905.10044*, 2019.
- [26] R. Cramer, I. B. Damgård *et al.*, *Secure multiparty computation*. Cambridge University Press, 2015.
- [27] J. Cui, C. Chen, L. Lyu, C. Yang, and W. Li, "Exploiting data sparsity in secure cross-platform social recommendation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 524–10 534, 2021.
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [29] P. Documentation, "torch.sparse.mm," Online, 2024, accessed: March 12, 2025. [Online]. Available: <https://pytorch.org/docs/stable/generated/torch.sparse.mm>
- [30] Y. Dong, W.-j. Lu, Y. Zheng, H. Wu, D. Zhao, J. Tan, Z. Huang, C. Hong, T. Wei, and W. Cheng, "Puma: Secure inference of llama-7b in five minutes," *arXiv preprint arXiv:2307.12533*, 2023.
- [31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [32] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game, or a completeness theorem for protocols with honest majority," in *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, 2019, pp. 307–328.
- [33] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

- [34] K. Gupta, N. Jawalkar, A. Mukherjee, N. Chandran, D. Gupta, A. Panwar, and R. Sharma, "Sigma: Secure gpt inference with function secret sharing," *Cryptology ePrint Archive*, 2023.
- [35] M. Hao, H. Li, H. Chen, P. Xing, G. Xu, and T. Zhang, "Iron: Private inference on transformers," *Advances in neural information processing systems*, vol. 35, pp. 15 718–15 731, 2022.
- [36] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.
- [37] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [38] X. Hou, J. Liu, J. Li, Y. Li, W.-j. Lu, C. Hong, and K. Ren, "Ciphergpt: Secure two-party gpt inference," *Cryptology ePrint Archive*, 2023.
- [39] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. I. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.
- [40] P. Kairouz, S. Oh, and P. Viswanath, "Secure multi-party differential privacy," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [41] M. Keller, E. Orsini, and P. Scholl, "Mascot: faster malicious arithmetic secure computation with oblivious transfer," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 830–842.
- [42] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten, "Crypten: Secure multi-party computation meets machine learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4961–4973, 2021.
- [43] N. Kumar, M. Rathee, N. Chandran, D. Gupta, A. Rastogi, and R. Sharma, "Cryptflow: Secure tensorflow inference," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 336–353.
- [44] S. Laur, J. Willemson, and B. Zhang, "Round-efficient oblivious database manipulation," in *Information Security: 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings 14*. Springer, 2011, pp. 262–277.
- [45] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," 2023.
- [46] R. Lehmkuhl, P. Mishra, A. Srinivasan, and R. A. Popa, "Muse: Secure inference resilient to malicious clients," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2201–2218.
- [47] D. Li, R. Shao, H. Wang, H. Guo, E. P. Xing, and H. Zhang, "Mpcformer: fast, performant and private transformer inference with mpc," *arXiv preprint arXiv:2211.01452*, 2022.
- [48] H. Li, Y. Li, A. Tian, T. Tang, Z. Xu, X. Chen, N. Hu, W. Dong, Q. Li, and L. Chen, "A survey on large language model acceleration based on kv cache management," *arXiv preprint arXiv:2412.19442*, 2024.
- [49] R. Li, L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. Marone, C. Akiki, J. Li, J. Chim *et al.*, "Starcoder: may the source be with you!" *arXiv preprint arXiv:2305.06161*, 2023.
- [50] X. Li, R. Dowsley, and M. De Cock, "Privacy-preserving feature selection with secure multiparty computation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6326–6336.
- [51] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 1, pp. 5–22, 2022.
- [52] Z. Li, C. You, S. Bhojanapalli, D. Li, A. S. Rawat, S. J. Reddi, K. Ye, F. Chern, F. Yu, R. Guo *et al.*, "The lazy neuron phenomenon: On emergence of activation sparsity in transformers," *arXiv preprint arXiv:2210.06313*, 2022.
- [53] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.
- [54] Z. Liang, P. Wang, R. Zhang, N. Xu, S. Zhang, L. Xing, H. Bai, and Z. Zhou, "Merge: Fast private text generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 18, 2024, pp. 19 884–19 892.
- [55] Y. Lindell, "How to simulate it—a tutorial on the simulation proof technique," *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, pp. 277–346, 2017.
- [56] X. Liu and Z. Liu, "Llms can understand encrypted prompt: Towards privacy-computing friendly transformers," *arXiv preprint arXiv:2305.18396*, 2023.
- [57] Z. Liu, J. Wang, T. Dao, T. Zhou, B. Yuan, Z. Song, A. Shrivastava, C. Zhang, Y. Tian, C. Re *et al.*, "Deja vu: Contextual sparsity for efficient llms at inference time," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 137–22 176.
- [58] W.-j. Lu, Z. Huang, Z. Gu, J. Li, J. Liu, K. Ren, C. Hong, T. Wei, and W. Chen, "Bumblebee: Secure two-party inference framework for large transformers," *Cryptology ePrint Archive*, 2023.
- [59] J. Luo, Y. Zhang, J. Zhang, X. Mu, H. Wang, Y. Yu, and Z. Xu, "Secformer: Towards fast and accurate privacy-preserving inference for large language models," *arXiv preprint arXiv:2401.00793*, 2024.
- [60] K. Maeng and G. E. Suh, "Accelerating relu for mpc-based private inference with a communication-efficient sign estimation," *Proceedings of Machine Learning and Systems*, vol. 6, pp. 128–147, 2024.
- [61] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," *arXiv preprint arXiv:1609.07843*, 2016.
- [62] I. Mirzadeh, K. Alizadeh, S. Mehta, C. C. Del Mundo, O. Tuzel, G. Samei, M. Rastegari, and M. Farajtabar, "Relu strikes back: Exploiting activation sparsity in large language models," *arXiv preprint arXiv:2310.04564*, 2023.
- [63] P. Mohassel and P. Rindal, "Aby3: A mixed protocol framework for machine learning," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 35–52.
- [64] M. Movahedi, J. Saia, and M. Zamani, "Secure multi-party shuffling," in *Structural Information and Communication Complexity: 22nd International Colloquium, SIROCCO 2015, Montserrat, Spain, July 14-16, 2015. Post-Proceedings 22*. Springer, 2015, pp. 459–473.
- [65] MPC Alliance, "Powering secure computation, together," <https://www.mpcalliance.org/>, 2023, accessed: 2023.
- [66] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
- [67] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *EUROCRYPT*, 1999, pp. 223–238.
- [68] Q. Pang, J. Zhu, H. Möllering, W. Zheng, and T. Schneider, "Bolt: Privacy-preserving, accurate and efficient inference for transformers," *Cryptology ePrint Archive*, 2023.
- [69] D. Paperno, G. Kruszewski, A. Lazaridou, Q. N. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández, "The lambda dataset: Word prediction requiring a broad discourse context," *arXiv preprint arXiv:1606.06031*, 2016.
- [70] A. Patra and A. Suresh, "Blaze: blazing fast privacy-preserving machine learning," *arXiv preprint arXiv:2005.09042*, 2020.
- [71] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [72] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [73] M. S. Riazzi, C. Weinert, O. Tkachenko, E. M. Songhori, T. Schneider, and F. Koushanfar, "Chameleon: A hybrid secure computation framework for machine learning applications," in *Proceedings of the 2018 Asia conference on computer and communications security*, 2018, pp. 707–721.

- [74] M. Roemmele, C. A. Bejan, and A. S. Gordon, “Choice of plausible alternatives: An evaluation of commonsense causal reasoning.” in *AAAI spring symposium: logical formalizations of commonsense reasoning*, 2011, pp. 90–95.
- [75] P. Schoppmann, A. Gascón, M. Raykova, and B. Pinkas, “Make some room for the zeros: Data sparsity in secure distributed machine learning,” in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 1335–1350.
- [76] A. Shamir, “How to share a secret,” *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.
- [77] N. Shazeer, “Glu variants improve transformer,” *arXiv preprint arXiv:2002.05202*, 2020.
- [78] J. Soifer, J. Li, M. Li, J. Zhu, Y. Li, Y. He, E. Zheng, A. Oltean, M. Mosyak, C. Barnes *et al.*, “Deep learning inference service at microsoft,” in *2019 USENIX Conference on Operational Machine Learning (OpML 19)*, 2019, pp. 15–17.
- [79] C. Song, X. Han, Z. Zhang, S. Hu, X. Shi, K. Li, C. Chen, Z. Liu, G. Li, T. Yang *et al.*, “Prospare: Introducing and enhancing intrinsic activation sparsity within large language models,” *arXiv preprint arXiv:2402.13516*, 2024.
- [80] Y. Song, Z. Mi, H. Xie, and H. Chen, “Powerinfer: Fast large language model serving with a consumer-grade gpu,” *arXiv preprint arXiv:2312.12456*, 2023.
- [81] Y. Song, H. Xie, Z. Zhang, B. Wen, L. Ma, Z. Mi, and H. Chen, “Turbo sparse: Achieving llm sota performance with minimal activated parameters,” *arXiv preprint arXiv:2406.05955*, 2024.
- [82] S. Tan, B. Knott, Y. Tian, and D. J. Wu, “Cryptgpu: Fast privacy-preserving machine learning on the gpu,” in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 1021–1038.
- [83] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, “Stanford alpaca: An instruction-following llama model,” 2023.
- [84] R. Tarjan, “Depth-first search and linear graph algorithms,” *SIAM journal on computing*, vol. 1, no. 2, pp. 146–160, 1972.
- [85] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, “Gemini: Open models based on gemini research and technology,” *arXiv preprint arXiv:2403.08295*, 2024.
- [86] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [87] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [88] S. Wagh, D. Gupta, and N. Chandran, “Securenn: 3-party secure computation for neural network training,” *Proceedings on Privacy Enhancing Technologies*, 2019.
- [89] S. Wagh, S. Tople, F. Benhamouda, E. Kushilevitz, P. Mittal, and T. Rabin, “Falcon: Honest-majority maliciously secure framework for private deep learning,” *arXiv preprint arXiv:2004.02229*, 2020.
- [90] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “Glue: A multi-task benchmark and analysis platform for natural language understanding,” *arXiv preprint arXiv:1804.07461*, 2018.
- [91] H. Wang, S. Ma, R. Wang, and F. Wei, “Q-sparse: All large language models can be fully sparsely-activated,” *arXiv preprint arXiv:2407.10969*, 2024.
- [92] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.
- [93] S. Yuan, J. Chen, Z. Fu, X. Ge, S. Shah, C. R. Jankowski, Y. Xiao, and D. Yang, “Distilling script knowledge from large language models for constrained language planning,” *arXiv preprint arXiv:2305.05252*, 2023.
- [94] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [95] Z. Zhang, Y. Song, G. Yu, X. Han, Y. Lin, C. Xiao, C. Song, Z. Liu, Z. Mi, and M. Sun, “Relu² wins: Discovering efficient activation functions for sparse llms,” *arXiv preprint arXiv:2402.03804*, 2024.
- [96] Z. Zhang, Z. Liu, Y. Tian, H. Khaitan, Z. Wang, and S. Li, “R-sparse: Rank-aware activation sparsity for efficient llm inference,” in *The Thirteenth International Conference on Learning Representations*, 2025.

Appendix A. Security Proof

The security of *Comet* adheres to the standard simulation-based security definition in MPC: a protocol is considered secure if there exists a polynomial-time simulator \mathcal{S} that can construct a simulated world where the view of an adversary \mathcal{A} is computationally indistinguishable from the real-world view [55]. Based on Universal Composability (UC) framework [15], to prove the security of *Comet*, it suffices to prove the security of its three core protocols: Oblivious Shuffle, SOMM, and SIMM.

Theorem 3. *The protocols oblivious shuffle, SOMM and SIMM are secure against the honest-but-curious adversary.*

Oblivious Shuffle. The adversary’s view in the oblivious shuffle protocol is $\text{view}_{\mathcal{A}} = \{\rho, \tau, \mathbf{a}, \mathbf{b}, \llbracket \mathbf{x} \rrbracket, \tau(\llbracket \mathbf{x} \rrbracket), \llbracket \mathbf{y} \rrbracket, \rho(\llbracket \mathbf{y} \rrbracket), \llbracket \mathbf{z} \rrbracket\}$. where ρ and τ are random permutations, and \mathbf{a}, \mathbf{b} are uniformly random vectors. The secret-shared variables $\llbracket \mathbf{x} \rrbracket$ and $\llbracket \mathbf{y} \rrbracket$, as well as their shuffled results $\tau(\llbracket \mathbf{x} \rrbracket)$ and $\rho(\llbracket \mathbf{y} \rrbracket)$, are uniformly random due to the properties of secret sharing and random permutations. The output $\llbracket \mathbf{z} \rrbracket$, computed as the sum of two secret-shared vectors, also retains uniform randomness. A simulator \mathcal{S} can replicate $\text{view}_{\mathcal{A}}$ by generating these components independently, ensuring that the simulated view is computationally indistinguishable from the real view. Thus, this protocol is secure against the honest-but-curious adversary.

SOMM and SIMM. The adversary’s view in SOMM protocol is $\text{view}_{\mathcal{A}} = \{\llbracket \mathbf{X} \rrbracket, \llbracket \mathbf{Y} \rrbracket, \mathcal{S}, \llbracket \mathbf{X} \rrbracket^t, \llbracket \mathbf{Y} \rrbracket^t, \llbracket \mathbf{R} \rrbracket^t, \llbracket \mathbf{Z} \rrbracket\}$. The secret-shared input $\llbracket \mathbf{X} \rrbracket, \llbracket \mathbf{Y} \rrbracket$ are uniformly random due to the properties of secret sharing. \mathcal{S} is a plaintext sparsity distribution obtained by applying an oblivious shuffle to a secret-shared 0-1 sparsity distribution, where the randomness of the shuffle ensures that \mathcal{S} is uniformly random. $\llbracket \mathbf{X} \rrbracket^t$ is obtained by performing local indexing operations on the secret-shared input, which do not alter their uniform randomness. $\llbracket \mathbf{R} \rrbracket^t$ is generated using a proven secure multiplication protocol, ensuring that it is also uniformly random. The output $\llbracket \mathbf{Z} \rrbracket$ is obtained by merged from secret-shared matrices, also retains uniform randomness. Thus, the SOMM protocol is secure against the honest-but-curious adversary. The security proof of SIMM follows the same

reasoning as SOMM, as both protocols share a similar structure and rely on the same security guarantees.

Appendix B. Optimality Proof

B.1. The proof of Theorem 1

We model the matrix multiplication $\mathbf{XY} = \mathbf{Z}$ using a bipartite graph $G = (X, Y, Z)$, where nodes $x \in X$ represent the row vectors of \mathbf{X} , nodes $y \in Y$ represent the column vectors of \mathbf{Y} , and edges $z \in Z$ correspond to nonzero elements in \mathbf{Z} , computed as the dot product of x and y . Since zero elements in \mathbf{Z} do not induce edges in G , some nodes in X and Y may be isolated, meaning they do not contribute to computation or communication. Without loss of generality, we focus on the reduced subgraph $G^r = (X^r, Y^r, Z)$, where X^r and Y^r include only nodes that participate in at least one multiplication.

The communication cost of an MPC matrix multiplication is given by $\mathcal{C}_1(G) = 2(|X| + |Y|)$, representing the number of row and column vectors communicated. The computation cost is $\mathcal{C}_2(G) = 3|X||Y|$, representing the total number of local dot product operations performed. For simplicity, we omit the coefficients. To achieve the minimal communication cost and, under this condition, the minimal computation cost, the problem can be formulated as finding a subgraph partition G_1, G_2, \dots, G_k of G^r such that

$$\begin{aligned} \min_{G_1, G_2, \dots, G_k} \quad & \mathcal{S}_2 = \sum_{i=1}^k \mathcal{C}_2(G_i) \\ \text{s.t.} \quad & \mathcal{S}_1 = \sum_{i=1}^k \mathcal{C}_1(G_i) \text{ is minimized,} \\ & \forall i \neq j, Z_i \cap Z_j = \emptyset. \end{aligned}$$

\mathcal{S}_1 achieves its minimum when each connected component of G^r is fully assigned to a single subgraph. Since

$$\sum \mathcal{C}_1(G_i) = \sum |X_i| + \sum |Y_i| \geq (|X^r| + |Y^r|),$$

the minimum is exactly the total number of nodes. This minimum is achieved only if each node is assigned to a single subgraph, meaning that the connected component containing the node must be entirely within one subgraph. Otherwise, some nodes would be assigned to multiple subgraphs, increasing \mathcal{C}_1 .

Under the condition that each connected component is assigned to a single subgraph (minimizing \mathcal{S}_1), the \mathcal{S}_2 is minimized by further ensuring that each subgraph contains exactly one connected component. Consider a subgraph $G' = (X', Y', Z')$ containing multiple connected components $G_i = (X_i, Y_i, Z_i)$ and $G_j = (X_j, Y_j, Z_j)$. The computation cost for G' is:

$$\begin{aligned} \mathcal{C}_2(G') &= |X'| |Y'| = |X_i + X_j| \cdot |Y_i + Y_j| \\ &\geq |X_i| |Y_i| + |X_j| |Y_j| = \mathcal{C}_2(G_i) + \mathcal{C}_2(G_j) \end{aligned}$$

Thus, splitting G' into G_i and G_j reduces the computation cost. Repeating this argument, the optimal partition occurs when each connected component forms its own subgraph.

Finding all connected components of G^r can be achieved using Depth-First Search (DFS). The SOMM protocol uses DFS to identify connected components, performs matrix multiplications for each component, and merges the results. Thus, Protocol 2 achieves both minimal communication cost and computation optimality, completing the proof.

B.2. The proof of Theorem 2

Communication minimization. In Protocol 3, we adopt a column-by-row multiplication strategy. Each nonzero element in \mathbf{X} is assigned to at most one column subvector, and each row of \mathbf{Y} participates in at most one multiplication. Therefore, every nonzero element of \mathbf{X} and every row of \mathbf{Y} are masked and communicated exactly once, achieving minimal communication cost.

Computation optimality. While MPC inherently decomposes one matrix multiplication into multiple local multiplications, the dimensions of these local multiplications remain consistent with those of the original matrices. Therefore, the total computation cost of MPC matrix multiplication can be directly represented by the number of scalar multiplications in the original matrix multiplication. Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p}$. Define R_i as the number of nonzero elements in the i -th row of \mathbf{X} and C_j as the number of nonzero elements in the j -th column of \mathbf{X} . The total number of nonzero elements is $N = \sum_{i=1}^m R_i = \sum_{j=1}^n C_j$. Under a row-by-column approach, each nonzero element participates in p scalar multiplications, leading to $\sum R_i \cdot p = N \cdot p$. Under a column-by-row approach, each nonzero element also contributes to p scalar multiplications, giving $\sum C_j \cdot p = N \cdot p$. Hence, either method produces $N \cdot p$ scalar multiplications, which matches the lower bound for computing \mathbf{XY} .

Since the column-by-row approach simultaneously achieves this computational optimum and incurs the minimal communication overhead, Protocol 3 is optimal in both communication and computation, completing the proof.

Appendix C. Generality of *Comet*

Architectures. We evaluated *Comet* on various Transformer architectures, including encoder-only (ViT [31], Bert [28]), encoder-decoder (T5 [72]), and decoder-only (OPT [94]), all of which use ReLU. As shown in Table 5, compared to Puma [30], *Comet* achieves a $1.25 \times - 2.84 \times$ speedup in TTFT (Time To First Token).

Activation functions. We extend *Comet* beyond ReLU-based LLMs by evaluating models that use GeLU [37] and SwiGLU [77], two of the most widely adopted activation functions in modern LLMs. Specifically, GeLU is used in models such as GPT2 [71], Falcon [2], and BLOOM [45], while SwiGLU is used in Llama2 [86], Llama3 [33], Mixtral [39], Qwen1.5 [6], and DeepSeek [8].

TABLE 5: Performance across different model architectures.

Model	ViT-Base	ViT-Large	Bert-Base	Bert-Large	T5-Base	T5-Large	OPT-6.7B	OPT-13B
Sparsity (%)	82.2	86.7	85.2	87.3	92.1	94.7	90.1	93.5
TTFT (min)	0.85	2.12	1.34	2.23	1.97	3.72	6.5	10.4
Speedup	1.55×	1.64×	1.73×	1.97×	1.85×	2.24×	2.57×	2.84×

TABLE 6: Performance across different activation functions.

Model	GPT2-XL	Falcon-7B	Bloom-7B	Llama2-7B	Llama3-8B	Mixtral-7B	Qwen1.5-7B	DeepSeek-7B
Activation	GeLU	GeLU	GeLU	Swish	Swish	Swish	Swish	Swish
Sparsity (%)	84.3	94.9	88.2	89.4	90.7	90.2	86.2	87.4
Accuracy deviation (%)	-2.1	-0.3	+0.5	-1.0	-0.4	+0.3	-0.5	-0.6
Speedup	1.82×	3.12×	2.52×	2.69×	2.75×	2.31×	2.53×	2.62×

Although these activation functions do not inherently exhibit sparsity, prior works [62], [79], [81], [91], [95], [96] have shown that replacing them with ReLU-family functions (e.g., ReLU [66], ReLU² [95], shiftedReLU [62], dReLU [81]) followed by fine-tuning enables activation sparsity without significant accuracy loss. We applied this ReLUfication strategy to eight different LLMs, as shown in Table 6. Each model achieved an 80%-94% sparsity ratio, with an average accuracy drop of less than 0.6% on the GLUE benchmark. On these sparsified LLMs, *Comet* achieves a 1.82×-3.12× speedup over Puma.

Appendix D. MPC-based Private Inference Systems

MPCFormer accelerates Transformer private inference by replacing costly Softmax and GeLU functions with simple quadratic approximations. Since these approximations degrade model accuracy, knowledge distillation is applied to restore performance. While effective in reducing MPC overhead, this approach introduces additional training steps.

SecFormer, like MPCFormer, replaces Softmax with a quadratic approximation but further improves efficiency by designing a high-precision polynomial for GeLU and an MPC-friendly LayerNorm computation. To compensate for approximation errors, it also requires fine-tuning the model.

Puma enhances MPC efficiency by designing high-precision polynomial approximations for Softmax and GeLU, ensuring a slight accuracy loss. It also introduces secure protocols for Embedding and LayerNorm computations. Unlike other approaches, it achieves near plaintext-level accuracy in encrypted inference without requiring any model fine-tuning.

The above works adopt semi-honest security. Some, like MUSE [46] and SIMC [18], achieve malicious security by ensuring protocol termination upon an attack, preventing sensitive information leakage. However, this significantly reduces inference efficiency. Enhancing malicious security in private LLM inference remains a key direction for future research.

Appendix E. Impact of Revealing Sparsity Levels

Comet leverages activation sparsity to accelerate private inference. Compared to existing private inference systems, the only additional information it reveals is the activation sparsity levels. Although there are currently no known effective attacks targeting sparsity levels, this information could introduce potential privacy risks. For example, in a semi-honest setting, an adversary could issue repeated inference requests for different inputs to map the relationship between input and sparsity level. In a malicious setting, a model provider could design a predictor that exhibits specific sparsity levels on some layers for certain inputs. Notably, the risks discussed here are not unique to *Comet*; these inference attacks [16] and backdoor attacks [51] remain open challenges even for state-of-the-art private inference systems.

Fortunately, differential privacy (DP) [40] under MPC can effectively mitigate risks of revealing sparsity levels. Specifically, before revealing the shuffled sparsity distribution $\llbracket s \rrbracket$, the MPC servers collaboratively generate a secret-shared 0-1 perturbation vector $\llbracket p \rrbracket$, where the number of ones is determined via MPC-based DP. Then, by applying an XOR operation between $\llbracket s \rrbracket$ and $\llbracket p \rrbracket$, some “0” are flipped to “1”, hiding the exact sparsity level. Importantly, this transformation does not turn “1” into “0”, ensuring that all activated neurons are still computed, preserving model accuracy. We evaluate the performance of *Comet* with DP. For Llama2-7B, even under strong DP guarantees ($\epsilon = 0.01$), *Comet* achieves 71.1% sparsity and a 2.01× speedup over Puma.

TABLE 7: The performance of *Comet* with DP under different privacy budget.

Privacy budget (ϵ)	w/o DP	0.5	0.1	0.05	0.01
Sparsity (%)	89.4	87.7	85.9	79.9	71.7
Speedup	2.69×	2.63×	2.52×	2.34×	2.01×

Appendix F. Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

F.1. Summary

The authors propose *Comet*, an MPC-based private inference system for LLMs that exploits activation sparsity to reduce communication and computation overhead. The authors introduce new protocols to securely skip zero-valued neurons and develop a KV-cache management strategy compatible with sparse inference. Extensive experiments demonstrate significantly reduced inference time (1.87x-2.63x) and communication (1.94x-2.64x) compared to state-of-the-art private LLM inference systems, with further evidence showing generalization across different activation functions, architectures, and tasks.

F.2. Scientific Contributions

- Provides a Valuable Step Forward in an Established Field
- Creates a new tool to enable Future Science

F.3. Reasons for Acceptance

- 1) The paper addresses a timely challenge in secure inference for LLMs and notably improves over state-of-the-art in speed and communication.
- 2) The proposed method is presented comprehensively, and the authors provide an extensive evaluation (including WAN settings, different activation functions, and various LLM architectures).
- 3) The paper contains many interesting and novel ideas on how to speed up inference.