

fMoE: Fine-Grained Expert Offloading for Large Mixture-of-Experts Serving

Hanfei Yu
Stevens Institute of Technology

Xingqi Cui*
Rice University

Hong Zhang
University of Waterloo

Hao Wang
Rutgers University

Hao Wang
Stevens Institute of Technology

Abstract

Large Language Models (LLMs) have gained immense success in revolutionizing various applications, including content generation, search and recommendation, and AI-assisted operation. To reduce high training costs, Mixture-of-Experts (MoE) architecture has become a popular backbone for modern LLMs. However, despite the benefits, serving MoE-based LLMs experience severe memory inefficiency due to sparsely activated experts. Recent studies propose to offload inactive experts from GPU memory to CPU memory to improve the serving efficiency of MoE models. However, they either incur high inference latency or high model memory footprints due to coarse-grained designs. To tame the latency-memory trade-off in MoE serving, we present *fMoE*, a fine-grained expert offloading system for MoE serving that achieves low inference latency with memory efficiency. We design *fMoE* to extract fine-grained expert selection patterns from MoE models and semantic hints from input prompts to efficiently guide expert prefetching, caching, and offloading decisions. *fMoE* is prototyped on top of HuggingFace Transformers and deployed on a six-GPU testbed. Experiments with open-source MoE models and real-world workloads show that *fMoE* reduces inference latency by 47% and improves expert hit rate by 36% over state-of-the-art solutions.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success in advancing Natural Language Processing (NLP) research and transforming various applications, including content generation [2, 6, 11, 40], search and recommendation [30, 59], and AI-assisted operations [22, 29, 34]. Given the high training costs, modern LLMs have returned to Mixture-of-Experts (MoE) architectures [1, 10, 21, 46, 53, 57] as their backbone implementations. Inside MoE models, each MoE

layer comprises a gating network and a collection of experts, with only a subset of experts being activated during computation. This sparse activation mechanism significantly reduces the number of floating point operations (FLOPs), enabling MoE-based LLMs to achieve substantially lower training costs compared to dense LLMs [10, 21, 57].

Despite the computational efficiency, MoE models exhibit substantial memory inefficiency during the serving phase. Though certain model parameters remain inactive during inference, they must still reside in GPU memory to allow for potential future activation. Expert offloading [4, 16, 47, 54] has emerged as a promising strategy to address this issue, which predicts inactive experts and transfers them to CPU memory while retaining only the necessary experts in GPU memory, reducing the overall model memory footprint.

However, existing expert offloading solutions struggle to effectively balance the *latency-memory trade-off* in MoE serving. These approaches either suffer from high inference latency [4, 47] or incur substantial model memory footprints [16, 54]. The key reason is that existing works track expert patterns and manage experts in *coarse granularity*. They fail to accurately identify and retain only the necessary experts in GPU memory during inference, resulting in frequent and costly on-demand expert loading [47], which severely degrades serving performance.

In this paper, we propose *fMoE*, a *fine-grained* expert offloading system that tames the latency-memory trade-off in MoE serving. To track and analyze MoE models’ expert selection behaviors in fine granularity, we propose a new data structure called *expert map*, which records the iteration-level probability distributions output by the gate network. *fMoE* uses historical expert maps for comparing expert trajectory similarity to guide offloading.¹ Apart from the expert map, *fMoE* is designed to track fine-grained input semantic embeddings from individual request prompts processed by the MoE model. Given the collected semantic-based and trajectory-based information, *fMoE* carefully searches the most accurate

*This work was performed when Xingqi Cui was a remote intern student advised by Dr. Hao Wang at the IntelliSys Lab of Stevens Institute of Technology.

¹In this paper, “trajectory” is defined as the collection of probability distributions over experts observed through layers.

expert map for guiding expert prefetching, caching, and offloading through inference iterations. In summary, we make the following contributions:

- We design *fMoE*, a **fine-grained** expert offloading system that achieves low inference latency while reducing model memory footprints.
- We propose a new data structure, expert map, that tracks fine-grained expert selection behaviors of MoE models. *fMoE* leverages input semantic embeddings to augment the expert map search for guiding expert offloading.
- We prototype *fMoE* on top of HuggingFace Transformers [52] and deploy it on a six-GPU testbed. Extensive experiments with open-source MoE models and real-world workloads show that *fMoE* reduces inference latency by 47% and improves expert hit rate by 36% compared to state-of-the-art solutions.

2 Background and Motivation

2.1 LLM Serving

Unlike traditional Deep Learning (DL) model inference, Large Language Model (LLM) serving consists of two consecutive stages: *prefill* and *decode*. Figure 1a illustrates the two stages when an LLM performs inference for a request prompt. In the prefill stage, the LLM first computes the intermediate key-value (KV) states of the prompt tokens, prefills the KV cache [3, 24, 27, 32, 61], and then generates the first answer token. In the decode stage, the LLM sequentially generates the answer to the prompt token-by-token in an auto-regressive manner, where tokens generated previously are used for generating the next token.

The two stages have their own unique characteristics. The prefill stage only requires one *iteration*, processing all tokens in parallel and generating the first answer token. The decode stage spans several iterations, generating one token per iteration until the answer is completed. Due to the different characteristics of the two stages, recent studies [38, 61] have identified that the prefill stage is compute-bounded, while the decode stage is considered memory-bounded. Therefore, people typically quantify the serving performance of LLM two stages using different metrics. For the prefill stage, Time-To-First-Token (TTFT) is commonly employed, which measures the latency from receiving the user request until generating the first answer token. For the decode stage, Tokens-Per-Second (TPS) or Time-Per-Output-Token (TPOT) is used to measure the generation rate of LLM serving.

2.2 MoE-based LLM Serving

By integrating MoE layers in Transformer blocks [51], MoE architectures [58] have emerged as a popular backbone for modern LLMs, such as Mixtral [21], Snowflake Arctic [46], and DeepSeek-MoE [10]. Figure 1a illustrates MoE-based

LLMs’ typical structures, where feed-forward network (FFN) modules are replaced by MoE layers. Each MoE layer consists of a gate network and a set of expert networks. Inside each Transformer block, the self-attention module first calculates the attentions [51] based on input hidden states, and then the gate network determines which expert(s) to activate for computing the output representations. Compared to traditional dense LLMs, MoE-based LLMs only activate a subset of parameters during training and inference, reducing computational overhead while delivering superior generation performance compared to dense LLMs with a comparable number of parameters [1, 10, 21, 46, 53, 57].

Despite the benefits of saving training computations, MoE-based LLM serving still suffers from GPU memory inefficiency as MoE inference requires loading all model parameters into GPU memory, including those inactive experts. For example, Mixtral-8×7B [21] and DeepSeek-MoE [10] have 72% and 83% inactive parameters during inference due to the sparsity of expert activation in MoE, leading to low memory efficiency and serving throughput. Therefore, to efficiently serve large MoE models, we must seek a solution to the memory inefficiency inherited from MoE architecture.

2.3 Latency-Memory Trade-Off

Recently, a few studies have been proposed to improve MoE-based LLM serving efficiency. Figure 2 describes the design space in MoE serving. Existing major studies can be categorized into two types: **Lossy serving** applies compression [39], pruning [26], and quantization [23] techniques to the original MoE models to reduce the serving memory requirements. However, this line of work achieves serving efficiency by sacrificing the generation quality. **Lossless serving** focuses on *offloading* model weights (parameters [4, 36] or experts [16, 47, 54]) that are sparsely utilized in temporal or spatial patterns from GPU memory to CPU memory, aiming to preserve reasonable inference latency. Specifically, expert offloading seeks to predict the activation of experts in advance, prefetching or caching only the necessary experts in GPU memory during inference. We opt for lossless serving to design *fMoE* because this line of methods avoids modifying models, hence assuring generation quality.

However, existing offloading solutions cannot achieve an optimal spot in the latency-memory trade-off when serving MoE-based LLMs. Figure 1b compares the performance (*i.e.*, inference latency and memory footprint) of existing state-of-the-art (SOTA) offloading solutions, which either provide low inference latency but suffer from large memory footprint (*e.g.*, No-offload and MoE-Infinity [54]), or vice versa (*e.g.*, ProMoE [47], Mixtral-Offloading [16], and DeepSpeed-Inference [4]).

The key reason behind this dilemma is that MoE-based decoder-only LLMs have balanced expert routing [47], leaving existing solutions hard to find effective patterns for guid-

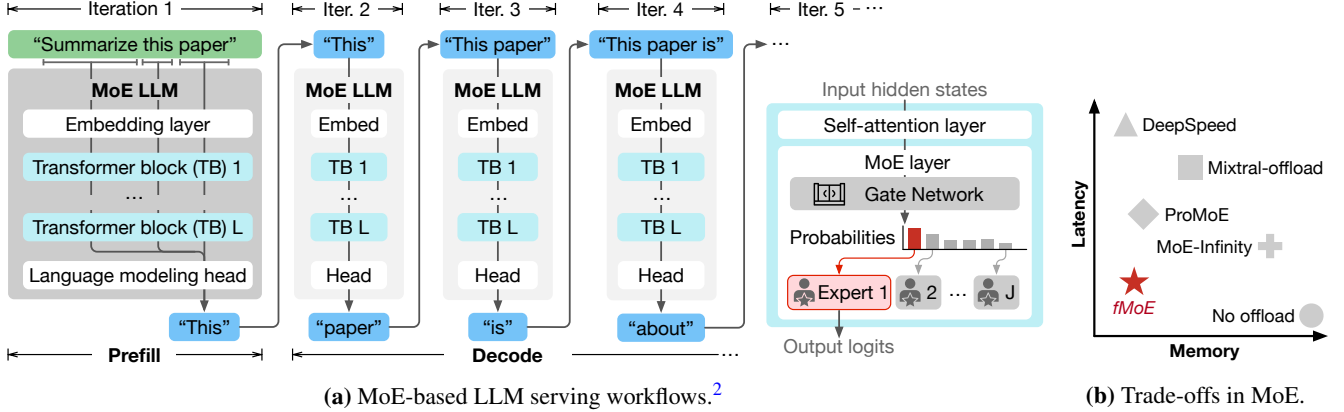


Figure 1: Mixture-of-Experts (MoE) Large Language Model (LLM) serving.

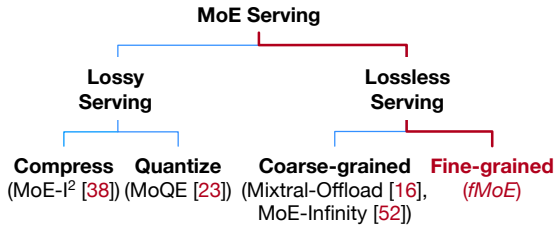


Figure 2: The design space of MoE-based LLM serving.

ing expert offloading. Existing research has identified two main reasons for this dilemma: First, most MoE-based LLMs are decoder-only architectures, which exhibit uniform expert activation patterns and low expert access skewness compared to encoder-decoder MoE LLMs [18, 47]. Second, recent MoE-based LLMs are trained with a unique load balancing loss [1, 10, 21, 46, 53], which enforces gate networks to balance the tokens routed to each expert within the same MoE layer, ensuring no experts are trivial throughout training. This balanced routing diminishes the predictability of expert patterns, thus making existing solutions ineffective.

2.4 Existing MoE Offloading Solutions

Existing expert offloading approaches [16, 54] rely on **coarse-grained** expert patterns, which are inefficient for guiding offloading. We define coarse-grained information as the expert patterns collected at the request level, where information is aggregated over multiple iterations of a request prompt. For example, MoE-Infinity [54] tracks request-level expert activations. Fine-grained information is defined as the expert patterns observed separately during each inference iteration. Figure 3a shows examples of coarse-grained and fine-grained expert activation heatmaps for Mixtral-8×7B [21]. The heatmap records the expert activations across 32 MoE layers, where each layer contains eight experts and activates two experts out of eight to compute representations. While fine-grained (iteration-level) heatmaps show clear expert acti-

vation patterns, the aggregated coarse-grained (request-level) heatmap diminishes predictability.

To demonstrate this point, we analyze the Shannon entropy [44] of expert activations per MoE layer for three popular MoE models. Entropy is an essential metric to quantify the uncertainty and unpredictability of variables in information theory. A balanced expert activation pattern (e.g., probability distribution [0.25, 0.25, 0.25, 0.25] of four experts) results in a high entropy, which indicates the pattern is more unpredictable and hard to select experts. Figure 3b presents the mean entropy computed per layer for three MoE models (Mixtral-8×7B [21], Qwen1.5-MoE [57], and Phi-3.5-MoE [1]) across two real-world datasets (LMSYS-Chat-1M [60] and ShareGPT [45]). Coarse-grained expert patterns have significantly higher entropy than fine-grained patterns, meaning that expert patterns in coarse granularity can be less effective for predictions. Figure 3c shows the mean entropy per layer through inference iterations. While the entropy is low at the beginning of inference, it gradually increases through iterations due to aggregating expert activation information, thus becoming more unpredictable.

In contrast to coarse-grained expert offloading solutions, we argue that expert offloading should be carefully guided by **fine-grained** designs: analyzing iteration-level patterns, understanding models’ expert selection preferences, and leveraging semantic characteristics of request prompts.

2.5 Problems of Coarse-Grained Offloading

Existing coarse-grained expert offloading solutions exhibit three problems:

1) Insufficient latency-memory trade-off. Existing solutions prefetch and offload experts in coarse granularity, either heavily focusing on reducing inference latency but incurring large memory footprint [54] or reducing memory footprint but severely increasing inference latency [4, 16].

2) Low expert hit rates. Existing solutions employ coarse-grained expert pattern tracking methods (e.g., Expert Activation Matrix in MoE-Infinity [54]), which produce ineffective

²For simplicity, we only show one request prompt in one batch.

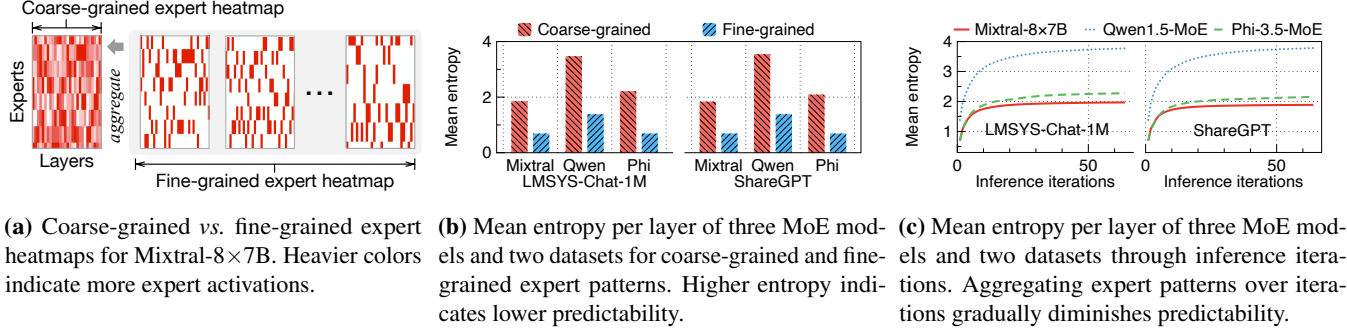


Figure 3: Expert pattern and predictability analysis in coarse granularity (request-level) and fine granularity (iteration-level).

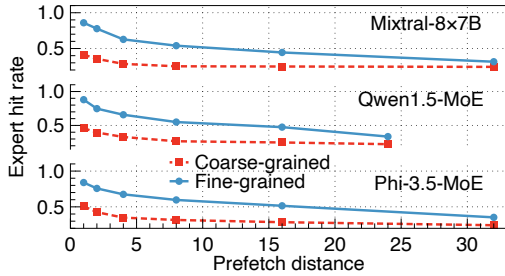


Figure 4: Expert hit rates of coarse-grained and fine-grained expert offloading designs when serving Mixtral-8x7B, Qwen1.5-MoE, and Phi-3.5-MoE at different prefetch distances, respectively.

expert patterns for guiding offloading decisions, leading to low expert hit rates and high inference latency.

3) Ignorance of MoE models’ and prompts’ heterogeneity. Existing solutions largely ignore the unique characteristics of different MoE models and input prompts and serve them in a one-fits-all manner [4, 16, 47, 54], which omits opportunities for fine-grained optimizations adaptive to heterogeneous models and prompts in MoE serving.

Figure 4 shows the expert hit rates of serving three popular MoE-based LLMs, Mixtral-8x7B [21], Qwen1.5-MoE [57], and Phi-3.5-MoE [1] using LMSYS-Chat-1M dataset [60] with coarse-grained and fine-grained expert offloading designs at different prefetch distances, respectively. Prefetch distance refers to the number of layers ahead that a prefetch instruction is issued before the target layer activates its experts. By leveraging fine-grained expert offloading, we can achieve significantly higher expert hit rates over coarse-grained methods and preserve better performance by adapting to varying prefetch distances.

3 *fMoE*’s Overview

3.1 Objectives and Challenges

fMoE is designed to achieve the following three goals:

Memory-efficient MoE serving with minimal inference latency. We have demonstrated that existing expert offloading solutions [16, 47, 54] fail to tame the latency-memory trade-off in MoE serving (§2.3). We aim to achieve both low memory footprint and inference latency by proposing fine-grained expert offloading.

Minimize expert miss due to mispredictions in expert prefetching. Expert prefetching, involving future expert activation predictions, is an essential step in expert offloading solutions. However, a recent study [47] has shown that *expert miss* due to mispredictions can cause high on-demand expert loading delay in inference. We should minimize expert miss and mitigate mispredictions in expert offloading.

Adapt to heterogeneous MoE models and prompts. MoE inference can serve heterogeneous models [10, 21, 46, 53, 57] with varying prompts [45, 60] in real-world scenarios. While existing solutions handle different models and prompts with a one-fits-all design, we should design our expert offloading to adapt to the heterogeneity in MoE serving.

We must address three critical challenges to realize the above objectives:

How to maximize expert hit rate when prefetching and offloading experts? Expert hit rate directly relates to the inference latency. With more experts being hit, fewer experts need to be loaded on demand. We propose a fine-grained expert offloading solution to achieve a high expert hit rate.

How to adapt to different MoE models and prompts? Heterogeneous MoE models and input prompts exhibit unique system and semantic characteristics. We should craft our solution with fine-grained optimizations to enable adaptivity.

How to avoid additional system overheads when managing experts? Our design must not introduce additional system overheads when serving existing MoE LLMs. We apply a series of system optimizations in *fMoE* to ensure serving efficiency and minimize additional overheads.

3.2 Architecture and Workflow

Figure 5 describes the architecture and workflow of *fMoE*, which consists of three main components:

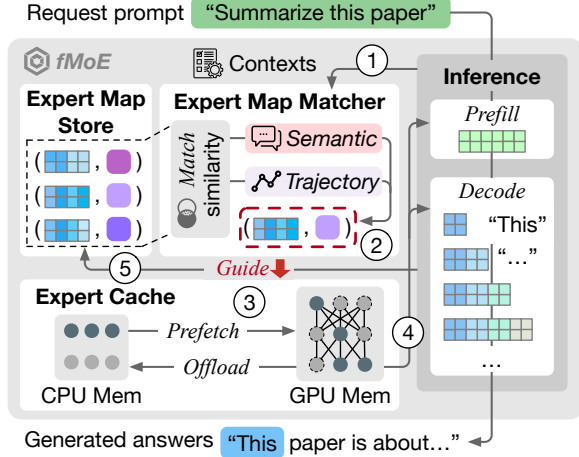


Figure 5: *fMoE*'s architecture and workflow.

- **Expert Map Store.** We record *expert maps*, a new data structure defined in *fMoE*, to track *fine-grained* expert activation patterns from historical request prompts. expert maps provide nuance expert selection preferences over existing coarse-grained expert tracking methods (e.g., Expert Activation Matrix in MoE-Infinity [54]). The Expert Map Store dynamically keeps the most useful and unique expert maps for real-time queries during inference.
- **Expert Map Matcher.** When a request prompt arrives, *fMoE* searches the Expert Map Store for appropriate expert maps to guide expert prefetching before inference. expert map search is guided by calculating similarity scores in two folds: *semantic* and *trajectory* similarity.
- **Expert Cache.** After receiving the matched expert maps, *fMoE* prefetches experts from CPU memory to GPU memory for performing computations in inference. *fMoE* evicts and offloads low-priority expert weights to CPU memory if exceeding Expert Cache capacity.

fMoE follows the five steps below to enable memory-efficient MoE serving with minimal inference latency:

Step ①: Inference context collection. Before every inference iteration, *fMoE* collects necessary *contexts*, such as semantic embeddings and previous expert activation trajectories (§4.1), and feeds them to the Expert Map Matcher for hybrid similarity matching.

Step ②: Expert map similarity matching. After receiving iteration-level contexts, the Expert Map Matcher finds and extracts the most similar expert maps by comparing the input context data with historical context data in the Expert Map Store (§4.2). The matched expert maps are forwarded to the Expert Cache to guide expert prefetching and offloading decisions.

Step ③: Guided expert prefetching and offloading. We dynamically compute expert selection thresholds to determine which expert(s) to prefetch and offload in the MoE model guided by the searched expert maps (§4.3). Then, *fMoE*

prefetches the expert weights from CPU to GPU memory and offloads cached experts from GPU to CPU when reaching the cache limit (§4.5).

Step ④: Expert serving. The whole inference process consists of one iteration in the Prefill stage and multiple iterations in the Decode stage. For each MoE layer in every iteration, *fMoE* directly serves the expert required by the gating network if the corresponding weights are available in the GPU memory (defined as an expert hit). Otherwise, *fMoE* on-demand loads the expert weights from CPU to GPU to perform lossless serving (defined as an expert miss).

Step ⑤: Expert map update. *fMoE* observes new expert maps produced after each iteration and updates them in the Expert Map Store (§4.4). When reaching the store capacity (e.g., 1K expert maps), *fMoE* deduplicates the Expert Map Store by identifying and dropping redundant expert maps to maintain diversity, maximizing the possibility of providing effective expert maps for any request prompts.

3.3 Problem Formulation

We consider serving an MoE-based LLM with L MoE layers on a GPU cluster, where each MoE layer has one gating network and J experts. The gating network of each layer selects top $K \in [1, J]$ experts for computation. The MoE model processes and generates answers for a workload consisting of W unique request prompts. Each request prompt $w \in [W]$ consists of multiple iterations processed during the prefill and decode stages, where $[W]$ is the request prompt collection. Let $E_{l,j}^{(i)}$ denote the j -th expert at the l -th layer in the i -th iteration, where $l \in [L]$, $j \in [J]$, and $i \in [w]$. During each iteration i , we can make at most $L \cdot J$ prefetching decisions. Let E_{cache}^i and $E_{activate}^i$ denote the set of cached experts and the set of activated experts for Iteration i , respectively. Hence, we represent the result of whether an expert $E_{l,j}^{(i)} \in E_{activate}^i$ is hit (served by E_{cache}^i) or miss (on-demand loading from CPU memory):

$$R_{l,j}^{(i)} = \begin{cases} 1, & \text{if } (E_{l,j}^{(i)} \in E_{activate}^i) \wedge (E_{l,j}^{(i)} \notin E_{cache}^i), \\ 0, & \text{otherwise,} \end{cases}$$

where $R_{l,j}^{(i)} = 1$ means $E_{l,j}^{(i)}$ is a miss. Since all experts in an MoE model are typically designed to have the same weight size, we assume experts' loading time T_e and memory footprint M_e are homogenous.³ Therefore, the total on-demand loading latency T is summed across all iterations for each expert during the inference process, i.e., $T := T_e \cdot \sum_{w \in [W]} \sum_{i \in [w]} \sum_{l \in [L]} \sum_{j \in [J]} R_{l,j}^{(i)}$.

Finally, employing the above definitions, we formulate the MoE expert offloading as an integer linear programming (ILP)

³We only consider selective experts. Some MoE models, such as Qwen1.5-MoE-A2.7B, have a few always-on experts that are not offloadable.

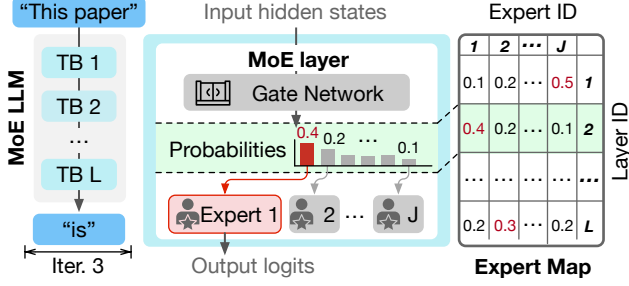


Figure 6: Expert selections tracked by an expert map.

optimization problem:

$$\min_{\{E_{l,j}^{(i)}\}} \left(T_e \cdot \sum_{w \in [W]} \sum_{i \in [w]} \sum_{l \in [L]} \sum_{j \in [J]} R_{l,j}^i \right) \quad (1)$$

$$\text{s.t. } |E_{cache}^{(i)}| \leq L \cdot J, \quad \forall i \in [w], \forall w \in [W], \quad (1)$$

$$|E_{activate}^{(i)}| = L \cdot K, \quad \forall i \in [w], \forall w \in [W], \quad (2)$$

$$|E_{cache}^{(i)}| \cdot M_e \leq M, \quad \forall i \in [w], \forall w \in [W]. \quad (3)$$

The objective is to minimize the on-demand loading latency (ideally $T = 0$ with perfect predictions) while limiting the total memory footprint of cached experts to satisfy the available GPU memory M . Constraint 1 denotes the total number of prefetched experts should not exceed the total number of all experts in the MoE model. Constraint 2 represents the total number of activated experts, which must be the same as the total number of top K experts summed across all L layers. Constraint 3 describes the total memory footprint of prefetched experts must be limited by the available GPU memory size. Note that solving the ILP problem is already NP-hard [9], while in reality, prefetching experts always have mispredictions that further complicate the problem. Therefore, we opt for a heuristic-based design for $fMoE$.

4 $fMoE$'s Design

4.1 Expert Maps

We propose a new data structure, *expert maps*, to track expert activation patterns with a fine granularity. Figure 6 depicts the structure of an expert map. During the i -th iteration, the l -th self-attention layer first calculates the attention states. The gate network receives attentions and computes a probability distribution $\mathbf{P}_l^{(i)} \in \mathbb{R}^J$ over all the experts at Layer l :

$$\mathbf{P}_l^{(i)} := \{p_{l,1}^{(i)}, \dots, p_{l,j}^{(i)}, \dots, p_{l,J}^{(i)}\}, \quad \sum_{j \in [J]} p_{l,j}^{(i)} = 1, \quad \forall p_{l,j}^{(i)} \geq 0.$$

Then, top $K \in [1, J]$ experts are selected from $\mathbf{P}_l^{(i)}$ to compute representations for Layer l . We collect the probability distributions $\mathbf{P}_l^{(i)}$ across all L layers to form the expert map of

Iteration i :

$$map_i := \{\mathbf{P}_1^{(i)}, \dots, \mathbf{P}_L^{(i)}\}, \quad l \in [L].$$

By tracking expert maps, we guide $fMoE$ to discover fine-grained expert patterns—the iteration-level expert selection preferences via probability distributions. Intuitively, analyzing probability distributions enables $fMoE$ to not only identify which experts are binarily selected or omitted, but also to assess the confidence or preference assigned to each expert from the perspective of the gate networks.

The design of expert maps has two key advantages over existing coarse-grained expert tracking methods (e.g., MoE-Infinity [54] tracks the request-level expert hit counts). *First*, existing works only focus on *aggregated* request-level expert activations, whereas an expert map tracks individual iterations with detailed expert selections. *Second*, existing works only record the expert hit counts, whereas we track detailed probability distributions. Note that expert maps can easily recover coarse-grained information by applying a top K selection operator to the probability distributions and aggregating expert counts over iterations, therefore generalizing to existing tracking methods. We evaluate $fMoE$ against other tracking methods to show the effectiveness of expert maps in §6.5.

4.2 Expert Map Search

When predicting and prefetching experts for MoE models, a *prefetch distance* is usually defined to avoid impacting inference latency [47]. Prefetch distance refers to the number of layers ahead that a prefetch instruction is issued before the target layer activates its experts, similar to the same term in memory prefetching [25]. Let d denote the prefetch distance of the MoE model to serve. Figure 7 shows that $fMoE$ employs two fine-grained search approaches to jointly match expert maps for guiding expert prefetching. Semantic search compares the input embeddings with historical embeddings to find expert maps with similar inputs, whereas trajectory search observes previous expert trajectories (i.e., probability distributions) and matches similar expert maps. We combine both semantic and trajectory features to improve $fMoE$'s map-matching and expert offloading accuracy. Two search approaches' effectiveness is evaluated in §6.5.

Semantic-based expert map search. For the initial layers $l \in [1, d]$, due to the prefetch distance d , existing solutions [16, 47, 54] cannot observe expert patterns for prediction and prefetching before the target layer is ready to activate experts. Thus, they usually define coarse-grained rules for prefetching initial layers. For example, MoE-Infinity [54] prefetches the most popular experts across all historical data points.

In contrast, $fMoE$ leverages semantic hints from the input prompt to search for the most useful expert maps, requiring zero knowledge from the expert activation patterns. When serving request prompts and recording their expert maps, we record the *semantic embeddings* for each inference iteration:

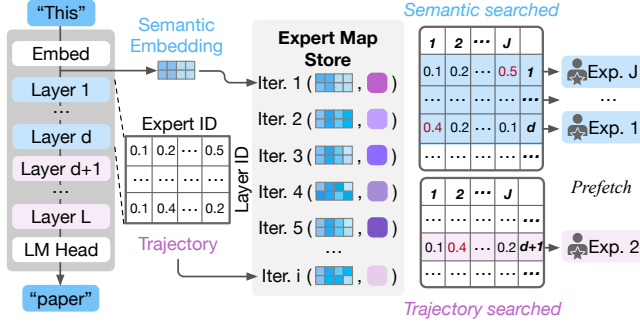


Figure 7: Semantic and trajectory expert map search.

Since existing MoE-based LLMs all contain an embedding layer for analyzing the semantic meaning of inputs, it's natural to extract the semantic embeddings using the output from the model's original embedding layer. For any input prompts, we compute pairwise cosine similarity $score^{sem} \in \mathbb{R}^{B \times C}$ between the semantic embedding $sem^{new} \in \mathbb{R}^{B \times h}$ and the collection of historical semantic embeddings $sem^{old} \in \mathbb{R}^{C \times h}$ in the Expert Map Store:

$$score_{x,y}^{sem} := \frac{sem_x^{new} \cdot sem_y^{old}}{\|sem_x^{new}\| \cdot \|sem_y^{old}\|}, \quad x \in [B], y \in [C], \quad (4)$$

where B is the batch size of input prompts, C is the Expert Map Store capacity, and h is the hidden dimension size. Then, for prompt x , the historical Iteration y with the highest score is selected. We use partial expert maps from the selected iteration, $\{\mathbf{P}_1^{(y)}, \dots, \mathbf{P}_d^{(y)}\} \in map_y^{old}$, to guide the expert prefetching for layers $l \in [1, d]$.

Trajectory-based expert map search. Unlike layers $l \in [1, d]$, we can observe expert probability trajectories of previous $(l-1)$ layers to search expert maps for layers $l \in [d+1, L]$. Similar to the semantic-based search, we compute pairwise cosine similarity $score^{traj} \in \mathbb{R}^{B \times C}$ between the observed trajectory, $map^{new} \in \mathbb{R}^{B \times (l-1)J}$, and the collection of historical expert maps, $map^{old} \in \mathbb{R}^{C \times (l-1)J}$, in the Expert Map Store:

$$score_{x,y}^{map} := \frac{map_x^{new} \cdot map_y^{old}}{\|map_x^{new}\| \cdot \|map_y^{old}\|}, \quad x \in [B], y \in [C]. \quad (5)$$

We select the historical iteration with the highest score. Then, we use $\mathbf{P}_{l+d}^y \in map_y^{old}$ from the selected expert map to guide the expert prefetching for the target Layer $l+d$, where d is the prefetch distance.

By combining the two expert map search methods, we carefully customize the map that guides expert prefetching for every inference iteration in MoE serving. With this design, expert map search introduces negligible overhead to the end-to-end inference latency, which we demonstrate in §6.7.

4.3 Expert Prefetching

Given the searched and customized expert map for a layer $l \in [L]$, we guide the expert prefetching in fine granularity.

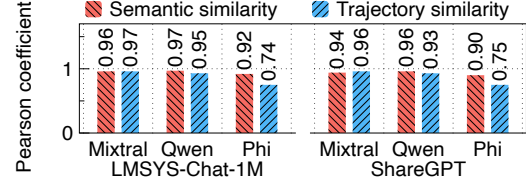


Figure 8: Pearson correlation coefficients between semantic and trajectory cosine similarity and expert hit rate across three MoE models and two datasets.

Similarity-aware expert selection. With the different contexts collected during iterations, expert maps searched by $fMoE$ also have varying similarity scores, which reflects the search confidence. To quantify the correlations between similarity score and expert hit rate, we run three MoE models (Mixtral-8×7B, Qwen1.5-MoE, and Phi-3.5-MoE) with two datasets (LMSYS-Chat-1M and ShareGPT) using the methodology described in §4.2. For each inference iteration, we compute the similarity scores, collect expert hit rates guided by the searched expert maps, and calculate the Pearson correlation coefficients [8]. Pearson coefficient is commonly used to measure correlations between variables, where a coefficient close to 1 indicates a strong positive correlation and a coefficient close to 0 means a weak correlation. Figure 8 shows the Pearson coefficient between similarity score and expert hit rate with three MoE models and two datasets. The results show that high similarity scores potentially relate to high expert hit rates if using the corresponding expert maps. Hence, we design $fMoE$'s expert prefetching to be similarity-aware.

For a layer $l \in [L]$ with a $score \in [-1, 1]$ to prefetch, we first dynamically compute an expert selection threshold $\delta_l \in [0, 1]$ given by

$$\delta_l := \text{Clip}(1 - score, 0, 1) = \max(0, \min(1 - score, 1)),$$

where $score$ is the cosine similarity score computed in Equations 4 and 5. Given searched \mathbf{P}_l , we find the set of experts to prefetch $E_{prefetch}$ by iteratively picking the expert with the highest probability from $\mathbf{P}_l = \{p_{l,1}, \dots, p_{l,j}, \dots, p_{l,J}\}$ until the summed probability of $E_{prefetch}$ exceeds δ_l :

$$\min_{\{E_{l,j}\}} |E_{prefetch}| \quad (6)$$

$$\text{s.t.} \quad \sum_{E_{l,j} \in E_{prefetch}} p_{l,j} \geq \delta_l, \quad j \in [J], \forall l \in [L], \quad (7)$$

$$|E_{prefetch}| \geq K, \quad K \leq [J], \quad (8)$$

where K is the number of experts needed to activate per layer (e.g., Mixtral-8×7B activates two experts per layer). Constraint 7 requires the total probability of selected experts to prefetch per layer to be greater than δ_l . Constraint 8 represents the minimum amount of selected experts must be larger than the number of experts to activate required by the MoE model. Intuitively, we assign higher δ to low-score expert maps so that more experts are prefetched to mitigate mispredictions

and assign lower δ for high-score expert maps to reduce memory footprint. Experts with higher probabilities are prioritized to be prefetched.

Asynchronous expert map matching and prefetching.

Existing studies [16, 54] predict and prefetch experts synchronously during inference, severely hindering the inference performance. For example, MoE-Infinity [54] cannot compute forward functions before finishing expert prediction and prefetching at every MoE layer [55]. To minimize the system overhead and inference latency, we decouple the map matching and expert prefetching from the inference process using an asynchronous Publisher-Subscriber architecture (Figure 7). The Expert Map Store is a message broker that keeps messages from both the inference process and the Expert Map Matcher. As the inference proceeds, *fMoE*’s inference process continuously publishes and writes the inference contexts (*i.e.*, semantic embeddings and expert probability distributions) to the Expert Map Store. At the same time, the Expert Map Matcher subscribes to the context data, matches expert maps based on new context data, and prefetches experts to the Expert Cache in an asynchronous manner.

4.4 Expert Map Store Management

Practically, we design *fMoE*’s Expert Map Store to maintain a capacity C for storing unique expert maps. To effectively guide inference across diverse prompts, it makes sense to identify and deduplicate redundant expert maps.

Expert map deduplication. Since *fMoE* uses two approaches (*i.e.*, semantic-based and trajectory-based) to compute similarity, we unify the two similarity scores to compute the pairwise redundancy scores between new iteration data and historical iteration data:

$$RDY_{x,y} := \frac{d}{L} \cdot score_{x,y}^{sem} + \frac{L-d}{L} \cdot score_{x,y}^{map}, \quad x \in [B], y \in [C],$$

where $score_{x,y}^{sem} \in \mathbb{R}^{B \times C}$ and $score_{x,y}^{map} \in \mathbb{R}^{B \times C}$ are semantic-based and trajectory-based pairwise similarity scores calculated from Equations 4 and 5, d is the prefetch distance, L is the total number of layers, B is the batch size of new interaction data, and C is the Expert Map Store capacity. Intuitively, as shown in Figure 7, the semantic-based and trajectory-based similarity scores contribute to the search expert map in proportion to $\frac{d}{L}$ and $\frac{L-d}{L}$, respectively. Therefore, we follow the same ratio to unify and compute the redundancy score. Whenever new iterations’ context data arrive at the Expert Map Store, we compute the pairwise redundancy score $RDY_{x,y}$ to determine which old iterations to drop. Hence, we update the old iterations y (columns in $RDY_{x,y}$) with new iterations x (corresponding rows in $RDY_{x,y}$) in the Expert Map Store.

Theoretical analysis. The expert map deduplication can be formulated as a Minimum Sphere Covering problem [17]. The objective is to minimize the total number of expert maps in the store, where each expert map is a vector representation of spheres, while maximizing the sphere coverage in

the expert activation space. Studies [15, 41] have proved that maintaining at least $2LJ$ expert maps guarantees a lower bound of 75% expert map similarity (*i.e.*, we can find an expert map that is at least 75% similar to any new iterations), and keeping $\frac{1}{2}LJ \ln(LJ)$ expert maps provides a lower bound of 98% similarity, where L and J are the numbers of layers and experts per layer in the MoE model, respectively. Given that modern MoE-based LLMs generally have $L \in [8, 128]$ and $J \in [24, 96]$, we can approximate the Expert Map Store’s maximal requirement to be less than 50K expert maps with 200 MB CPU memory [54].

4.5 Expert Caching and Eviction

Similar to existing expert offloading solutions [16, 47, 54], we design *fMoE* to maintain an Expert Cache on GPUs to reuse expert weights when serving different request prompts. Given matched expert maps from §4.2, we guide *fMoE*’s Expert Cache to compute two priority scores for individual experts: 1) a prefetching priority to decide the prefetching orders of experts in the searched maps, and 2) an eviction priority to determine the eviction orders of experts in the Expert Cache.

Expert prefetching priority. Recall the set of experts to prefetch $E_{prefetch}$ is determined in Equation 6. For each expert $E_{l,j} \in E_{prefetch}$, we define the prefetching priority to be

$$PRI_{l,j}^{prefetch} := \frac{p_{l,j}}{l - l_{now}}, \quad l \in [L], j \in [J],$$

where $p_{l,j}$ is the expert probability from the searched expert map, and l_{now} is the current layer that inference process stays at. Intuitively, experts with higher probability $p_{l,j}$ to be activated should be prefetched sooner, and experts that sit closer to the current layer (*i.e.*, smaller $l - l_{now}$) should also be prioritized.

Expert eviction priority. Similar to MoE-Infinity [54], *fMoE*’s expert caching is based on the least frequently used (LFU) algorithm. We integrate the searched map to jointly determine the eviction priority. For each expert $E_{l,j} \in E_{cache}$, we define the eviction priority to be

$$PRI_{l,j}^{evict} := \frac{1}{p_{l,j} \cdot freq_{l,j}}, \quad l \in [L], j \in [J],$$

where $freq_{l,j}$ is the cache visit frequency and $p_{l,j}$ is the probability from the searched map for an expert $E_{l,j} \in E_{cache}$. Intuitively, when reaching the Expert Cache limit, we want to first evict experts who are less frequently hit and have lower probabilities of being activated. Note that similar to existing works [47, 54], we do not consider the recent usage of experts as opposed to the classic least recently used (LRU) algorithm [16]. Since the expert usage is layer-wise sequential, *i.e.*, one layer following another, prioritizing recently used experts is against the nature of sequential forward computation in MoE serving.

MoE Models	Parameters (active / total)	Experts Per Layer (active / total)	Num. of Layers
Mixtral-8×7B [21]	12.9B / 46.7B	2 / 8	32
Qwen1.5-MoE [57]	2.7B / 14.3B	4 / 60	24
Phi-3.5-MoE [1]	6.6B / 42B	2 / 16	32

Table 1: Characteristics of three MoE models in evaluation.

On-demand expert loading. Mispredictions of expert prefetching lead to expert miss in the Expert Cache, as the MoE model cannot find available experts designated by the gate networks. Whenever an expert miss occurs, *fMoE* pauses all expert prefetching tasks and immediately loads missed experts from CPU to GPU memory for fast serving.

5 *fMoE*’s Implementation

We prototype *fMoE* on top of Huggingface Transformers framework [52] using MoE-Infinity codebase [55]. The implementation of *fMoE* is described as follows.

Expert Map Store is implemented in Python using PyTorch [37] and NumPy [19] libraries. We store all semantic embeddings and expert maps using `ndarrays` data structure for efficient array operations. The arrays are converted to tensors to compute similarity for expert map matching.

Expert Map Matcher is implemented in Python using PyTorch [37] and TorchMetrics [12] libraries. We implement the pairwise computations, including similarity (§4.2) and redundancy (§4.4) scores, using the Cosine Similarity interfaces in TorchMetrics. We use the Python multithreading library to implement the asynchronous expert map matching and expert prefetching, where the threads share the same memory space with the Expert Map Store for efficient reading and writing.

Expert Cache is implemented in C++ based on MoE-Infinity codebase [55]. The expert management in GPUs is implemented with the CUDA Runtime APIs [35]. We implement the caching logic of *fMoE* and fix critical bugs in the MoE-Infinity codebase to enable expert offloading. Same with MoE-Infinity, *fMoE* supports multi-GPU inference with expert parallelism, where the experts are mapped to different GPU devices for loading and offloading. We use a hash map to assign expert IDs to different GPUs and retrieve them during inference. The expert assignment follows a round-robin manner to balance the overall GPU load. Additionally, we use a multi-thread task pool in the GPU space to schedule and execute expert prefetching and on-demand loading tasks.

6 Evaluation

6.1 Experimental Setup

Testbed. We conduct all experiments on a six-GPU testbed, where each GPU is an NVIDIA GeForce RTX 3090 with 24 GB GPU memory. All GPUs are inter-connected using pairwise NVLinks and connected to the CPU memory using

PCIe 4.0 with 32GB/s bandwidth. Additionally, the testbed has a total of 32 AMD Ryzen Threadripper PRO 3955WX CPU cores and 480 GB CPU memory.

Models. We employ three popular MoE-based LLMs in our evaluation: Mixtral-8×7B [21], Qwen1.5-MoE [57], and Phi-3.5-MoE [1]. Table 1 describes the parameters, number of MoE layers, and number of experts per layer for the three models. Following the evaluation of existing works [47], we profile the models to set the optimal prefetch distance d to three before evaluation.

Datasets and traces. We employ two real-world prompt datasets commonly used for LLM evaluation: LMSYS-Chat-1M [60] and ShareGPT [45]. For most experiments, we split the sampled datasets in a standard 7:3 ratio, where 70% of the prompts’ context data (*i.e.*, semantic embeddings and expert maps) are stored in *fMoE*’s Expert Map Store, and 30% of the prompts are used for testing. For online serving experiments, we empty the Expert Map Store and use real-world LLM inference traces [38, 48] released by Microsoft Azure to set input and generation lengths and drive invocations.

Baselines. We compare *fMoE* against four SOTA MoE serving baselines: 1) **MoE-Infinity** [54] uses coarse-grained request-level expert activation patterns and synchronous expert prediction and prefetching for MoE serving. We prepare the expert activation matrix collection for MoE-Infinity before evaluation for a fair comparison. 2) **ProMoE** [47] employs a stride-based speculative expert prefetching approach for MoE serving. Since the codebase of ProMoE is not open-sourced and requires training predictors for each MoE model, we reproduced a prototype of ProMoE on top of MoE-Infinity in our best effort. 3) **Mixtral-Offloading** [16] combines a layer-wise speculative expert prefetching and a LRU-based expert cache. 4) **DeepSpeed-Inference** employs an expert-agnostic layer-wise parameter offloading approach, which uses pure on-demand loading and does not support prefetching. We implement the offloading logic of DeepSpeed-Inference in the MoE-Infinity codebase and add an expert cache for a fair comparison. We enable all baselines to serve MoE models from HuggingFace Transformer [52].

Metrics. Following the standard evaluation methodology of existing works [3, 47, 54, 61] on LLM serving, we report the performance of the prefill and decode stages separately. We measure Time-to-First-Token (TTFT) for the prefill stage and Time-Per-Output-Token (TPOT) for the decode stage. Additionally, we also report other system metrics, such as expert hit rate and overheads, for detailed evaluation.

6.2 Overall Performance

We first evaluate the performance of prefill and decode stages when running *fMoE* and other baselines with the three MoE models, where we measure Time-To-First-Token (TTFT) and Time-Per-Output-Token (TPOT) for each stage. Note that the inference latency with expert offloading tends to be higher

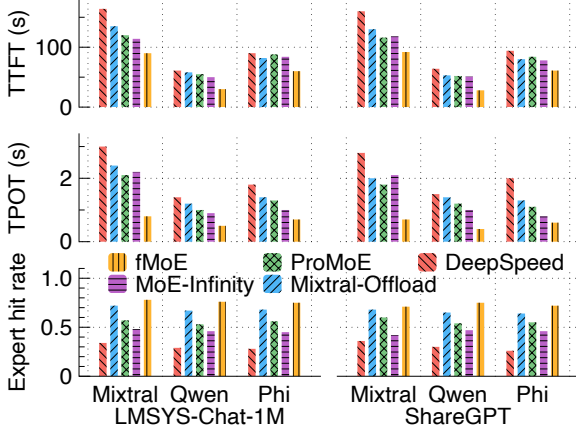


Figure 9: Overall performance of prefill and decode stages for *fMoE* and other four baselines.

than no offloading due to two reasons: 1) During inference, an excessive amount of parameters in MoE models are loaded and offloaded, which prolongs the inference latency. 2) All baselines and *fMoE* are implemented on top of the MoE-Infinity codebase [55], whose inference latency is inherently impacted by MoE-Infinity’s implementation. Nevertheless, comparing *fMoE* and baselines is fair with the same experimental setup.

Figure 9 shows the TTFT, TPOT, and expert hit rate of *fMoE* and other four baselines when serving three MoE models with LMSYS-Chat-1M and ShareGPT datasets, respectively. DeepSpeed has both the worst TTFT and TPOT due to expert-agnostic offloading and lacking expert prefetching. While Mixtral-Offloading, ProMoE, and MoE-Infinity perform better than DeepSpeed-Inference, they are underperformed by *fMoE* because of coarse-grained offloading designs. Compared to DeepSpeed-Inference, Mixtral-Offloading, ProMoE, and MoE-Infinity, our *fMoE* reduces the average TTFT by 44%, 35%, 33%, 30%, and reduces the average TPOT by 70%, 61%, 55%, 48%, across three MoE models. For expert hit rate, Mixtral-Offloading achieves a higher hit rate than the other three baselines because of its synchronous speculative prefetching with a prefetch distance of 1. However, due to synchronous prefetching, its TTFT and TPOT are worse than others except DeepSpeed-Inference. *fMoE* improves the average expert hit rate by 147%, 11%, 34%, and 63% over DeepSpeed-Inference, Mixtral-Offloading, ProMoE, and MoE-Infinity, respectively.

6.3 Online Serving Performance

Except for the offline evaluation (*i.e.*, Expert Map Store in full capacity before serving), we also evaluate *fMoE* against other baselines in online serving settings. We empty the Expert Map Store of *fMoE* and the expert activation matrix collection of MoE-Infinity for the online serving experiment. The request traces are derived from Azure LLM inference traces [38, 48],

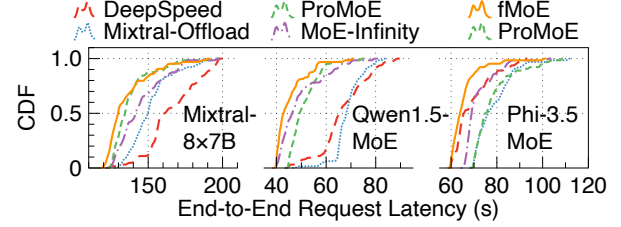


Figure 10: CDF of request latency for MoE online serving.

with 64 requests randomly sampled to drive LMSYS-Chat-1M prompts for each MoE model serving. To ensure consistency, *fMoE* and all baselines input and generate the exact number of tokens specified in the traces. Figure 10 illustrates the CDF of end-to-end request latency across three MoE models. The results demonstrate that *fMoE* significantly reduces overall request latency compared to other baselines in online serving scenarios.

6.4 Impact of Expert Cache Limits

We measure the TPOT of *fMoE* and other baselines by limiting the expert cache memory budget to investigate their performance in the latency-memory trade-off (§2.3). We mainly focus on TPOT to show the end-to-end performance impacted by varying cache limits. Figure 11 shows the TPOT of *fMoE* and other four baselines when serving three MoE models under different expert cache limits. We gradually increase the GPU memory allocated for caching experts from 6 GB to 96 GB while employing the same experimental setting in §6.2. Similarly, DeepSpeed-Inference has the worst TPOT due to being expert-agnostic. *fMoE* consistently outperforms Mixtral-Offloading, ProMoE, and MoE-Infinity under varying expert cache limits. Especially for limited GPU memory sizes (*e.g.*, 6GB), *fMoE* reduces the TPOT by 32%, 24%, 18%, and 18%, compared to DeepSpeed-Inference, Mixtral-Offloading, ProMoE, and MoE-Infinity, across three MoE models, respectively. With fine-grained expert offloading, *fMoE* significantly reduces the expert on-demand loading latency while maintaining a lower GPU memory footprint, therefore achieving a better spot in the latency-memory trade-off of MoE serving.

6.5 Ablation Study

We present the ablation study of *fMoE*’s design.

Effectiveness of expert map search. One of *fMoE*’s key designs is the expert map, which tracks expert selection preferences in fine granularity. We evaluate the effectiveness of the expert map against five expert pattern-tracking approaches as follows. 1) **Speculate**: speculative prediction used by Mixtral-Offloading [16] and ProMoE [47], 2) **Hit count**: request-level expert hit count used by MoE-Infinity [54], 3) **Map (T)**: expert map with only trajectory similarity search, 4) **Map (T+S)**: expert map with both trajectory and semantic similarity search, and 5) **Map (T+S+ δ)**: expert map with full features

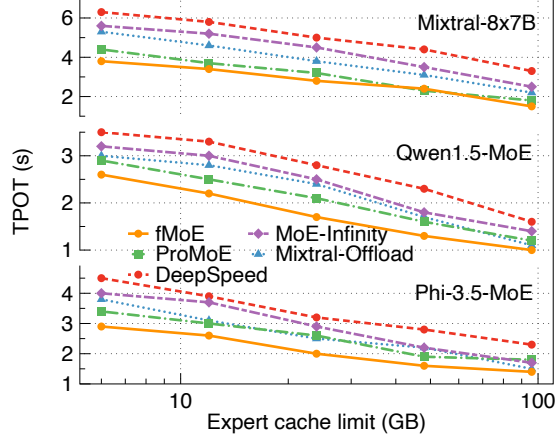
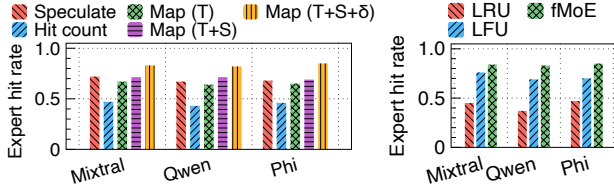


Figure 11: Performance of *fMoE* and other four baselines under varying expert cache limits.



(a) Expert pattern tracking approaches. (b) Prefetch and caching.

Figure 12: Ablation study of *fMoE*.

enabled, including trajectory and semantic similarity search (§4.2) and dynamic expert selection (§4.3). We implement the above methods in *fMoE*’s Expert Map Matcher for a fair comparison. Figure 12a shows the expert hit rate of the above expert pattern tracking methods. Speculative prediction is effective due to the widespread presence of residual connections in Transformer blocks. However, its effectiveness decreases drastically as prefetch distance increases [47]. The request-level expert activation count has the worst performance due to coarse granularity. As features are incrementally restored to *fMoE*’s expert map, the expert hit rate gradually increases, demonstrating its effectiveness.

Effectiveness of expert prefetching and caching. We evaluate *fMoE*’s expert prefetching and caching against two caching algorithms: 1) **LRU** used by Mixtral-Offloading [16] and 2) **LFU** used by MoE-Infinity [54]. Figure 12b depicts the expert hit rate of *fMoE* and two baselines. The results show that LRU performs poorly in expert offloading scenarios. Though LFU achieves a higher hit rate than LRU, *fMoE* surpasses both, achieving the highest expert hit rate.

6.6 Sensitivity Analysis

We analyze the sensitivity of three hyperparameters: prefetch distance of MoE models, the capacity of Expert Map Store, and inference batch size.

Prefetch distance of MoE models. Figure 13 shows the TTFT and TPOT of *fMoE* when serving three MoE models

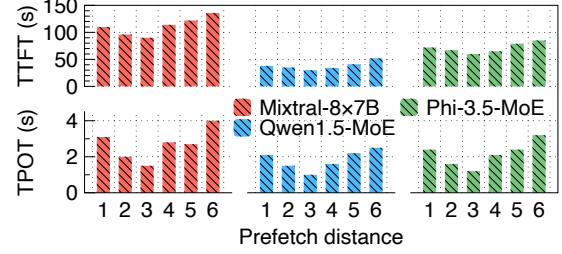
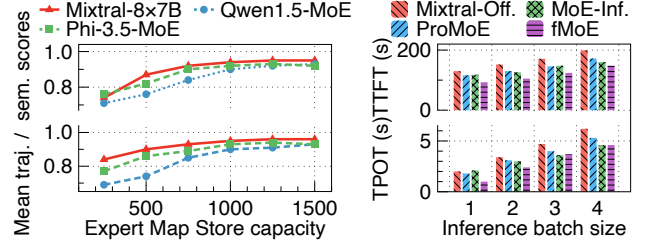


Figure 13: Performance of *fMoE* serving MoE models with different prefetch distances.



(a) Expert Map Store capacity.

(b) Inference batch size.

Figure 14: Sensitivity analysis of *fMoE*.

with different prefetch distances. We have demonstrated that the expert hit rate decreases when gradually increasing the prefetch distance (Figure 4). When the prefetch distance is small (< 3), *fMoE* cannot perfectly hide its system delay from the inference process, such as the map matching and expert prefetching, leading to the increase of inference latency. With larger prefetch distances (> 3), *fMoE* has worse expert hit rates that also degrade the performance. Therefore, we set the prefetch distance d to 3 for evaluating *fMoE*.

Capacity of Expert Map Store. We measure the mean semantic and trajectory similarity scores searched in *fMoE*’s expert map matching for MoE model serving. Figure 14a presents the mean semantic and trajectory similarity scores of *fMoE* with different Expert Map Store capacity sizes. Both semantic and trajectory similarity scores improve as the store capacity increases. While the similarity scores exhibit a significant increase with capacities below 1K, further capacity expansion yields diminishing similarity gains. To minimize *fMoE*’s memory overhead, we set *fMoE*’s Expert Map Store capacity to 1K in evaluation.

Inference batch size. We investigate the impact of inference batch size on *fMoE* and three baselines using Mixtral-8x7B with LMSYS-Chat-1M. Figure 14b presents the performance of *fMoE*, Mixtral-Offloading, ProMoE, and MoE-Infinity as the batch size increases from one to four. *fMoE* achieves the lowest TTFT and TPOT in most cases.

6.7 System Overheads

Latency overheads of *fMoE*’s operations. Figure 15 shows the latency breakdown of one inference iteration in *fMoE* when serving the three MoE models. We report any opera-

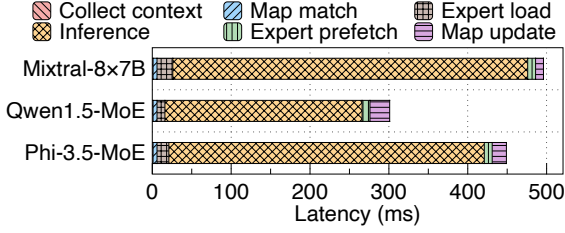


Figure 15: Latency breakdown of *fMoE*’s one inference iteration with three MoE models.

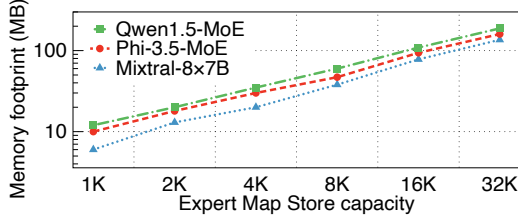


Figure 16: CPU memory footprint of *fMoE*’s Expert Map Store with different capacity.

tions of *fMoE* in §6.2 that may incur a significant latency delay, including context collection, map matching, expert on-demand loading, expert prefetching, and map update after the iteration completes. Qwen1.5-MoE has lower end-to-end iteration latency than Mixtral-8×7B and Phi-3.5-MoE because of significantly fewer parameters. Note that expert prefetching, map matching, and map update tasks are executed asynchronously, aside from the inference process. Hence, they do not contribute to the end-to-end iteration latency. Excluding three asynchronous tasks, the total delay incurred by other operations is consistently less than 30ms (5% of the iteration) across three MoE models, which is negligible compared to the inference latency.

Memory overheads of *fMoE*’s Expert Map Store. Figure 16 shows the CPU memory footprint of *fMoE*’s Expert Map Store when varying the store capacity from 1K to 32K maps. The memory needed to store expert maps for Qwen1.5-MoE is more than Mixtral-8×7B and Phi-3.5-MoE because it has more experts per layer over the other two models, which increases the map shape. Even for the largest capacity (32K), the Expert Map Store requires less than 200MB of memory to store the maps, which is trivial since modern GPU servers usually have abundant CPU memory (e.g., p4d.24xlarge on AWS EC2 [5] has over 1100 GB of CPU memory). In the evaluation, *fMoE*’s map store capacity with 1K maps is sufficient for maintaining performance (§6.6), resulting in minimal memory overhead.

7 Related Work

Lossless MoE serving. Recent studies on lossless MoE serving have been widely proposed. DeepSpeed inference [4] offload layer-wise parameters without expert awareness.

Mixtral-Offloading [16] employs LRU expert caching and introduces speculative prediction to enable expert prefetching. MoE-Infinity [54] proposes the request-level expert activation matrix to guide offloading in coarse granularity. Swap-MoE [42] maintains a set of critical experts in GPU memory and adjusts them based on workload changes to minimize offloading overhead. ProMoE [47] trains predictors per MoE layer to achieve high speculative prediction accuracy and low inference latency. Lina [28] moves infrequently used experts to host memory and focuses more on distributed MoE training. Liu *et al.* [31] serves MoE models on serverless computing by predicting expert patterns with black-box Bayesian techniques. Unlike existing coarse-grained offloading solutions, *fMoE* tracks fine-grained expert patterns from both trajectory and semantic aspects and outperforms SOTA baselines.

Lossy MoE serving. Except for lossless offloading, other works also propose lossy MoE serving. Expert pruning [13, 50] reduces memory usage by removing underutilized experts. Knowledge distillation [33, 43] produces compact sparse MoE models. ComPEFT [56] demonstrates expert compression without accuracy loss, while MC-SMoE [39] further decomposes merged experts into low-rank and structurally sparse alternatives. Hobbit [49] uses low precision to serve less-critical experts. However, lossy serving may impact the generation quality and is orthogonal to *fMoE*.

MoE refactorization. Some works propose to redesign and refactor the current MoE architecture. Pre-gated MoE [20] utilizes preemptive expert selection to eliminate the sequential dependencies between expert selection and execution. SiDA-MoE [14] proposes a sparsity-inspired, data-aware inference system that decouples the expert routing from inference. READ-ME [7] refactors pre-trained dense LLMs into specialized MoE models. Unlike the above works, *fMoE* requires zero training to serve open-source MoE models.

8 Conclusion

This paper proposes *fMoE*, a fine-grained expert offloading system for MoE serving that achieves low inference latency without incurring significant model memory footprints. *fMoE* tracks iteration-level expert probability distributions from the MoE model using expert map and analyzes input semantic embeddings from individual request prompts. Based on the input semantic and expert trajectory information, *fMoE* searches the most accurate expert map to carefully guide the expert prefetching, caching, and offloading decisions tailored to every inference iteration. *fMoE* is prototyped on top of HuggingFace Transformers and deployed to a six-GPU testbed. Extensive experiments with open-source MoE models and real-world workloads show that *fMoE* reduces inference latency by 47% and improves expert hit rate by 36% compared to state-of-the-art solutions.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2024.
- [4] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. DeepSpeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2022.
- [5] AWS. AWS EC2: Secure and Resizable Compute Capacity in the Cloud. <https://aws.amazon.com/ec2/>, 2006.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 2020.
- [7] Ruisi Cai, Yeonju Ro, Geon-Woo Kim, Peihao Wang, Babak Ehteshami Bejnordi, Aditya Akella, and Zhangyang Wang. Read-ME: Refactorizing LLMs as Router-Decoupled Mixture of Experts with System Co-Design. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- [8] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson Correlation Coefficient. *Noise Reduction in Speech Processing*, 2009.
- [9] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT press, 2022.
- [10] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv preprint arXiv:2401.06066*, 2024.
- [11] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. Neural Retrievers are Biased Towards LLM-Generated Content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.
- [12] Nicki Skaftø Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics: Measuring Reproducibility in Pytorch. *Journal of Open Source Software (JOSS)*, 2022.
- [13] Dmitry Lepikhin, Hyoungho Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, Zhifeng Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations (ICLR)*, 2021.
- [14] Zhixu Du, Shiyu Li, Yuhao Wu, Xiangyu Jiang, Jingwei Sun, Qilin Zheng, Yongkai Wu, Ang Li, Hai Li, and Yiran Chen. SiDA: Sparsity-Inspired Data-Aware Serving for Efficient and Scalable Large Mixture-of-Experts Models. *Proceedings of Machine Learning and Systems (MLSys)*, 2024.
- [15] Ilya Dumer. Covering Spheres with Spheres. *Discrete & Computational Geometry*, 2007.
- [16] Artyom Eliseev and Denis Mazur. Fast Inference of Mixture-of-Experts Language Models with Offloading. *arXiv preprint arXiv:2312.17238*, 2023.
- [17] D Jack Elzinga and Donald W Hearn. The Minimum Covering Sphere Problem. *Management Science*, 1972.
- [18] Vima Gupta, Kartik Sinha, Ada Gavrilovska, and Anand Padmanabha Iyer. Lynx: Enabling Efficient MoE Inference through Dynamic Batch-Aware Expert Selection. *arXiv preprint arXiv:2411.08982*, 2024.
- [19] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi,

- Christoph Gohlke, and Travis E. Oliphant. Array Programming with NumPy. *Nature*, 2020.
- [20] Ranggi Hwang, Jianyu Wei, Shijie Cao, Changho Hwang, Xiaohu Tang, Ting Cao, and Mao Yang. Pre-gated MoE: An Algorithm-System Co-Design for Fast and Scalable Mixture-of-Expert Inference. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, 2024.
- [21] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [22] Zhihan Jiang, Jinyang Liu, Zhuangbin Chen, Yichen Li, Junjie Huang, Yintong Huo, Pinjia He, Jiazhen Gu, and Michael R Lyu. LILAC: Log Parsing using LLMs with Adaptive Parsing Cache. *Proceedings of the ACM on Software Engineering*, 2024.
- [23] Young Jin Kim, Raffy Fahim, and Hany Hassan Awadalla. Mixture of Quantized Experts (MoQE): Complementary Effect of Low-bit Quantization and Robustness. *arXiv preprint arXiv:2310.02410*, 2023.
- [24] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PageAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.
- [25] Jaekyu Lee, Hyesoon Kim, and Richard Vuduc. When Prefetching Works, When It Doesn't, and Why. *ACM Transactions on Architecture and Code Optimization (TACO)*, 2012.
- [26] Jaeseong Lee, Aurick Qiao, Daniel F Campos, Zhewei Yao, Yuxiong He, et al. STUN: Structured-Then-Unstructured Pruning for Scalable MoE Pruning. *arXiv preprint arXiv:2409.06211*, 2024.
- [27] Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2024.
- [28] Jiamin Li, Yimin Jiang, Yibo Zhu, Cong Wang, and Hong Xu. Accelerating Distributed MoE training and inference with Lina. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, 2023.
- [29] Yichen Li, Yintong Huo, Renyi Zhong, Zhihan Jiang, Jinyang Liu, Junjie Huang, Jiazhen Gu, Pinjia He, and Michael R Lyu. Go Static: Contextualized Logging Statement Generation. *Proceedings of the ACM on Software Engineering*, 2024.
- [30] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. Data-efficient Fine-tuning for LLM-based Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.
- [31] Mengfan Liu, Wei Wang, and Chuan Wu. Optimizing distributed deployment of mixture-of-experts model inference in serverless computing. In *IEEE Conference on Computer Communications (INFOCOM)*, 2025.
- [32] Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, et al. CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving. In *Proceedings of the ACM SIGCOMM 2024 Conference*, 2024.
- [33] Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Man-deep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Ves Stoyanov. . Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*, 2021.
- [34] Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. Using an LLM to Help with Code Understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024.
- [35] NVIDIA. CUDA Runtime API :: CUDA Toolkit Documentation. <https://docs.nvidia.com/cuda/cuda-runtime-api/index.html>, 2024.
- [36] Ollama. Get Up and Running with Large Language Models. <https://ollama.com/>.
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems (NIPS)*, 2019.
- [38] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient Generative LLM Inference Using Phase Splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, 2024.

- [39] Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, Tianlong Chen. Merge, Then Compress: Demystify Efficient SMOE with Hints from Its Routing Policy. In *International Conference on Learning Representations (ICLR)*, 2024.
- [40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 2019.
- [41] Robert Alexander Rankin. On the Closest Packing of Spheres in N Dimensions. *Annals of Mathematics*, 1947.
- [42] Rui Kong, Yuanchun Li, Qingtian Feng, Weijun Wang, Xiaozhou Ye, Ye Ouyang, Linghe Kong, Yunxin Liu. SwapMoE: Serving Off-the-shelf MoE-based Large Language Models with Tunable Memory Budget. *arXiv preprint arXiv:2308.15030*, 2023.
- [43] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. *arXiv preprint arXiv:2201.05596*, 2022.
- [44] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.
- [45] ShareGPT. ShareGPT: Share Your Wildest ChatGPT Conversations. <https://sharegpt.com/>.
- [46] Snowflake. Snowflake Arctic: The Best LLM for Enterprise AI. <https://www.snowflake.com/en/data-cloud/arctic/>.
- [47] Xiaoni Song, Zihang Zhong, and Rong Chen. ProMoE: Fast MoE-based LLM Serving using Proactive Caching. *arXiv preprint arXiv:2410.22134*, 2024.
- [48] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, and Esha Choukse. DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency. In *International Symposium on High-Performance Computer Architecture (HPCA)*, 2025.
- [49] Peng Tang, Jiacheng Liu, Xiaofeng Hou, Yifei Pu, Jing Wang, Pheng-Ann Heng, Chao Li, and Minyi Guo. Hobbit: A mixed precision expert offloading system for fast moe inference. *arXiv preprint arXiv:2411.01433*, 2024.
- [50] Tianyu Chen, Shaohan Huang, Yuan Xie, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, Furu Wei. Task-specific expert pruning for sparse mixture-of-experts. *arXiv preprint arXiv:2210.17323*, 2022.
- [51] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [52] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. HuggingFace’s Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.
- [53] xAI. Announcing Grok. <https://x.ai/blog/grok>.
- [54] Leyang Xue, Yao Fu, Zhan Lu, Luo Mai, and Mahesh Marina. MoE-Infinity: Offloading-Efficient MoE Model Serving. *arXiv preprint arXiv:2401.14361*, 2024.
- [55] Xue, Leyang and Fu, Yao and Lu, Zhan and Mai, Luo and Marina, Mahesh. MoE-Infinity Codebase. <https://github.com/TorchMoE/MoE-Infinity>.
- [56] Prateek Yadav, Leshem Choshen, Colin Raffel, and Mohit Bansal. Compeft: Compression for communicating parameter efficient updates via sparsification and quantization. *arXiv preprint arXiv:2311.13171*, 2023.
- [57] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2024.
- [58] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2012.
- [59] Yuyue Zhao, Jiancan Wu, Xiang Wang, Wei Tang, Dingxian Wang, and Maarten de Rijke. Let Me Do It for You: Towards LLM Empowered Recommendation via Tool Learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.
- [60] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. *arXiv preprint arXiv:2309.11998*, 2023.
- [61] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. Distserve: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2024.