Mixture-of-AI
Efficience for AI models

Joel AGBOGLO[1], Hilario HOUMEY [1]

HUMIRIS AI

research@humiris.ai

Abstract

We present Mixture of AI, a model that integrates a mixture of AI models designed to deliver superior performance, optimize costs, ensure data privacy, and reduce the carbon footprint. Our gating mechanism currently surpasses that of RouterLLM and several other LLM routers.Unlike traditional routing approaches, which often rely solely on criteria such as cost and quality, Mixture of AI's dynamic gating offers a more holistic, adaptive, and multiparameter approach.In addition to cost and quality, Mixture of AI's gating also considers factors such as data privacy, speed, and carbon footprint, providing a more comprehensive and customized request management.Our gating enables real-time adaptation to changing routing conditions and user needs, unlike static approaches, which are less responsive to variations in demand.

1. Introduction

Large language models (LLMs) demonstrate impressive capabilities across many tasks, but choosing the right model often requires considering multiple factors such as performance, cost, privacy, speed, and more.More powerful models deliver high-quality results but tend to be more expensive, slower, and often closed-source, while less powerful models are more cost-effective, faster, and often open-source.Large models like OpenAI's GPT-4, Anthropic's Claude 3, or models developed by Google DeepMind or other major labs tend to offer high performance in terms of language understanding, text generation, and answering complex questions. These models can process rich data and produce high-quality results but often require substantial computational resources.

Powerful commercial models are generally expensive to use, especially for applications requiring high volumes of API calls or extended usage. This includes infrastructure costs for hosting and operational costs related to energy and server maintenance. On the other hand, smaller or open-source models can provide a more economical alternative, especially when hosted on private infrastructure or more affordable cloud solutions.Using closed models hosted on third-party servers can raise privacy concerns, especially if the data being processed is sensitive. Companies concerned with data security may prefer solutions where they have full control over the infrastructure.Open-source models can be deployed internally, offering better control over data privacy.Although large models deliver impressive performance, they can be slow to execute due to their complexity. Latency can become an issue for applications that require real-time responses. Smaller models, or those optimized for efficiency, can provide faster

responses and may be more suitable for environments with strict latency constraints.
Open-source models are often more accessible, allowing the community to adapt and improve them. They offer great flexibility to customize solutions based on specific needs. Proprietary models may provide better technical support and stability but often limit customization or extension possibilities.The ecological footprint, or "green footprint," of language models is an increasingly important factor in the selection and use of AI technologies, including LLMs. This refers to the environmental impact of training and using these models, particularly in terms of energy consumption and carbon emissions.

To address the trade-offs between performance, cost, privacy, speed, accessibility, and environmental impact of large language models (LLMs), we propose **Mixture of AI**, a technology that blends different language models to enable efficient use of LLMs.

Our **Gating** system analyzes queries based on the mix instructions provided to it and directs them to the appropriate models, which is a complex process.An effective gating model must be able to determine the intent, complexity, and domain of an incoming query, while assessing the capabilities of the available models to route the query to the most suitable one. Additionally, this gating model must be cost-effective, fast, and adaptable to a constantly evolving environment, characterized by the regular introduction of new models with enhanced capabilities.

2. Related Work

Several approaches have been developed to enhance and optimize the use of large language models (LLMs). Single-model enhancements, like fine-tuning (Rafailov et al., 2023), improve task-specific performance but require additional training and domain-specific data. Techniques such as Chain-of-Thought (CoT) prompting (Wei et al., 2022; Zhou et al., 2023; Wang et al., 2022) and Tree of Thoughts (ToT) reasoning (Yao et al., 2023) aim to boost LLM capabilities without the need for fine-tuning. Mixture-of-Experts (MoE) (Eigen et al., 2014; Shazeer et al., 2017) is another strategy that increases efficiency by routing inputs to specialized "experts" within the model, yet these approaches tend to be model-specific and struggle to take advantage of the growing diversity of LLMs. To move beyond single-model optimization, LLM synthesis combines the outputs of multiple models to produce enhanced results (Jiang et al., 2023b), with research showing that smaller models can sometimes outperform larger ones when used strategically (Lu et al., 2024). However, this method involves multiple steps, increasing both latency and costs, limiting its practical use. FrugalGPT (Chen et al., 2023) offers a more cost-effective approach by using a generation judger that sequentially evaluates responses from different LLMs, invoking models until a satisfactory answer is reached. In addition to these, other multi-LLM strategies have emerged. RouteLLM focuses on routing models based on user preferences, while ROUTERBENCH provides a benchmark to evaluate the efficiency of multi-LLM routing systems. A Multi-LLM Debiasing Framework tackles the issue of biases across various models, while another study, "Small LLMs Are Weak Tool Learners," explores the challenges smaller models face in tool learning. CollabStory leverages multiple LLMs for collaborative story generation and authorship analysis. Cost-Effective Online Multi-LLM Selection presents versatile reward models to optimize real-time model selection. Additionally, a Multi-LLM Orchestration Engine facilitates personalized, context-rich assistance by coordinating multiple models, and Martian Model Router focuses on optimizing costs in routing LLMs. Lastly, Arcee.Ai merges outputs from various models, while Lamini's MoME (Mixture of Models for Efficiency) aims to improve factual accuracy and reduce hallucinations in LLM outputs, offering further refinements in multi-model systems.

3. Mixture-of-AI Methodology

3.1. Overview

Mixture of AI is an innovative technology comprising a set of n large language models (LLMs) {M1, . . . , Mn}, coupled with a "gating" mechanism, G. Each LLM represents a particular area of expertise or specialization, offering a diverse range of skills to handle various types of queries. The crucial role of the gating mechanism, which is itself an LLM equipped with a softmax activation function, is to evaluate each incoming query and determine the most suitable model to respond. The gating model operates by analyzing the query’s characteristics, such as its content, intent, complexity, and thematic domain. Based on this analysis, the gating model activates the foundational models that possess the most relevant capabilities to handle the query optimally, according to the provided mix-instructions.

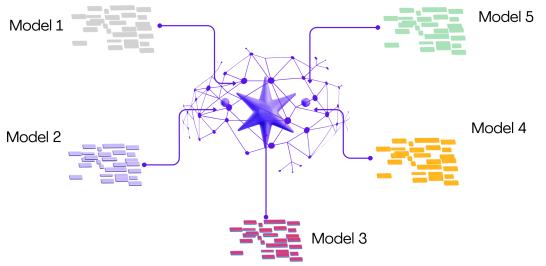

Fig1. MoAI overview representation

For example, for a query that requires deep natural language understanding and high precision, the gating mechanism might direct the query to a model specialized in these skills. Conversely, for a less complex query, the gating could select a lighter model to save resources while still providing an adequate response.
In this section, we present our proposed methodology for leveraging multiple models to achieve our objectives. We begin with the basic step of routing queries to the appropriate models. Next, we demonstrate that LLMs possess a collaborative capability, allowing them to improve their responses by utilizing outputs from other models. Finally, we explore the advanced aspect, which involves having multiple models (whether the same or different) interact to respond to a query. Afterward, we will introduce the concept of mix-tuning.

3.2. Methodology Mo-AI Basic

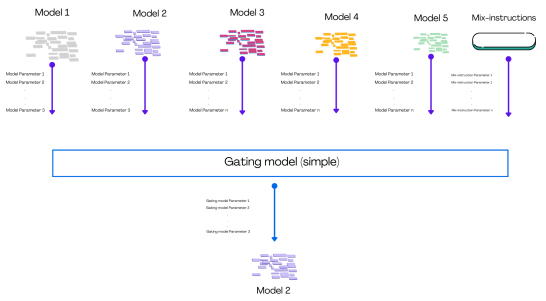

Fig2. MoAI-Basic architecture

3.2.1 Selection of LLMs for Query Routing

The first essential element of the Mo-AI methodology is the selection of large language models (LLMs) to which queries will be routed. The selection of LLMs is based on a thorough analysis of their specific capabilities and performance across various tasks. Selection criteria include the diversity of knowledge domains covered by each model, the quality of the generated responses, and performance in terms of latency and computational cost. This diversity maximizes the effectiveness of the Mixture of AI technology by ensuring comprehensive coverage of possible queries, ranging from simple information retrieval to complex text generation. In the first version of Mo-AI, we focus solely on text-generation models.

3.2.2 Selection of Inference Servers

Choosing the inference servers is crucial for ensuring smooth and efficient execution of the selected LLMs. This choice is guided by several factors, such as the processing capacity of the servers, their geographical location to minimize latency, and energy efficiency. High-performance cloud infrastructures are preferred for their flexibility and ability to handle varying workloads while ensuring scalability as needed. Simultaneously, robust security measures are implemented to protect sensitive data processed by the models. In our case, we will use Together AI and Groq servers for the inference of open-source models.

3.2.3 Creation of the Gating System: The Core of Mixture of AI

The central component of the Mo-AI methodology is the creation of the gating system, which plays a crucial role in managing queries. The gating mechanism is a sophisticated artificial intelligence model based on a neural network and equipped with a softmax activation function. This system is responsible for analyzing incoming queries and deciding how to route them to the most appropriate LLM. The decision-making process relies on the mix instructions provided to the gating mechanism, as well as an assessment of the query's nature, complexity, and the specific capabilities of the available models. The gating allows for optimization of both the quality of responses and the computational resources used, reserving the most powerful models for the most complex queries in cost optimization scenarios, or utilizing open-source models for queries involving sensitive data.

3.2.4 Creation of a Benchmark Dataset for Gating Training

To effectively train the gating system, a rigorously designed benchmark dataset is essential. This dataset consists of representative data for various types of queries that the system is likely to encounter. It includes samples of queries covering a broad spectrum of domains and complexities. The process of collecting and labeling this data is

meticulously planned to ensure high quality and diversity. The dataset is then used to train the gating model, enabling it to learn how to evaluate queries and select the most appropriate model for each situation.

3.2.5 Creation of a Real-Time Data Collection System for Monitoring LLM Consumption and Usage

Finally, a data collection system is established to monitor various performance and usage dimensions of the LLMs within the Mo-AI infrastructure. This system collects metrics such as the number of queries, the number of tokens per LLM, response time, energy consumption, inference costs, and user feedback. These data are crucial for the continuous improvement of the system, allowing for an understanding of the actual performance of each LLM in various scenarios and enabling adjustments to routing strategies accordingly. Furthermore, this monitoring helps identify opportunities for cost and energy efficiency optimization, thus contributing to the sustainability of the system.

3.2.6 Collaborativeness of LLMs

In Mixture of AI, collaboration among language models (LLMs) is facilitated through "mix instructions," which are directives provided by developers and orchestrated by a central component: the gating model. This gating model plays a crucial role in executing the instructions to guide interactions between the various LLMs, with the goal of optimizing final responses.

A key approach to maximizing the benefits of collaboration among multiple models is to characterize the specific skills of each model in different aspects of collaboration. In this process, we can categorize the models into two distinct roles:

Proposers:These models excel at generating useful reference responses for other models. A good proposer may not produce exceptional answers on its own, but it must provide context and diverse perspectives that enrich the final responses when used by an aggregator.

Aggregators:These models specialize in synthesizing responses provided by other models into a single high-quality answer. An effective aggregator should be able to maintain or enhance the quality of the output, even when integrating responses that are of lower quality than its own.

Managing Redundancy and Complementarity: The "mix instructions" include guidelines to avoid redundancy and maximize the complementarity of responses. For example, if multiple LLMs provide similar answers, the instructions may guide the gating model to prioritize complementary aspects or diversify the angles of the responses.

Execution Flexibility: Although the "mix instructions" are predefined, the gating model has some flexibility in their execution. This flexibility allows it to adapt in real-time to the specifics of each query and the observed performance of the involved LLMs, while still adhering to the provided instructions.

Integration of New Models:The "mix instructions" can be updated by developers to include new LLMs or to adjust roles and collaborations based on technological advancements or specific needs. The gating model then applies these new instructions, ensuring that Mixture of AI remains efficient and up-to-date.

Examples of Collaboration: Consider a question requiring both legal analysis and ethical reflection. Developers may define "mix instructions" where a law-specialized LLM acts as a proposer to provide a detailed legal analysis, while another LLM, focused on ethics, adds moral considerations. The gating model, following the "mix instructions," aggregates these contributions to deliver a final response that incorporates these different perspectives.

3.3. Methodology Mo-AI Advanced

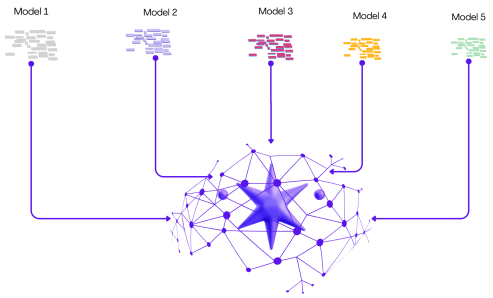

Fig3. MoAI Advanced Architecture

Mixture of AI Advanced takes the concept of collaboration among artificial intelligence models even further by enabling the simultaneous interaction of multiple models, whether identical or different, to respond to a given query. This approach leverages the diversity of capabilities and perspectives provided by various LLMs to enhance the quality and relevance of the generated responses. We delve into the underlying mechanisms, coordination strategies for interactions between the models, and the implications of this advanced methodology.By facilitating concurrent model interactions, Mixture of AI Advanced can harness the strengths of each

model to create a more comprehensive and nuanced answer. This system encourages an ecosystem where models can contribute their unique expertise, leading to a richer synthesis of information. Moreover, effective coordination among models is essential for maintaining the integrity of responses, managing potential redundancies, and ensuring that the final output reflects a cohesive understanding of the query.The implications of this advanced methodology are significant, particularly in applications requiring multifaceted analysis, such as legal assessments, scientific research, or creative problem-solving. By integrating multiple viewpoints, Mixture of AI Advanced not only improves the accuracy of responses but also promotes a more holistic approach to complex queries, ultimately enhancing user satisfaction and trust in AI-generated content.

3.3.1. Interactions Multi-LLMs

In the Mo-AI Advanced methodology, interactions among multiple LLMs are orchestrated to maximize synergy between the models. These interactions can occur at various levels and in different ways:

Sequential Interaction: In this configuration, one model handles part of the query, producing a partial or intermediate response that is then used as input by another model. This process can continue over several iterations, with each model adding an additional layer of analysis or detail. For example, one model might first generate a technical analysis, which is subsequently enriched by another model providing economic perspectives.

Concurrent Interaction: Models can also work in parallel, each processing the query independently or addressing different aspects of the query simultaneously. The responses produced by these models are then combined through an aggregation process, often guided by the mix instructions predefined by the developers. This approach is particularly beneficial for complex queries requiring a multifaceted solution.

Feedback Interaction**:** Another level of interaction may involve the revision of initial responses by additional models. For instance, after a set of models has produced a response, a supplementary model may be utilized to verify, correct, or refine that response, ensuring maximum coherence and quality.

3.3.2. Coordination of Interactions: The Role of the Gating Model

The gating model plays a central role in coordinating interactions between the various LLMs. The gating model, based on the mix instructions, determines which model should be solicited at each step and in what order. This determination relies on the specific strengths of the models for different tasks or aspects of the query.Once the different models have produced their respective responses, the gating model intervenes to orchestrate the aggregation process. It uses the mix instructions to guide the combination of responses, ensuring that important aspects are not overlooked and that the final response is coherent.The gating model is also responsible for managing any potential conflicts or inconsistencies between the responses provided by different models. It may decide to weight the contributions of certain models more heavily or eliminate contradictory elements based on the directives of the mix instructions.

3.3.3. Mix Instructions and Adaptive Flexibility

In the Mo-AI advanced methodology, mix instructions play a vital role by providing a framework for collaboration between models. However, they must also be flexible enough to adapt to changing contexts and next-generation models. Developers have the ability to update the mix instructions to incorporate new models or adjust collaboration strategies based on observed performance or the new capabilities of available models.The Mo-AI advanced methodology is particularly effective in scenarios where the complexity of the query and the expected performance exceed the capabilities of a single model.

For example:

Complex Medical Diagnosis: One model can provide an analysis based on symptoms, another can evaluate laboratory data, while a third can integrate genetic information. The gating model coordinates these interactions to deliver a comprehensive and accurate diagnosis.One model can analyze the overall sentiment of a text, another can assess specific cultural connotations, and a third can provide a comparative analysis with similar texts. This multi-model collaboration can offer a deep and nuanced understanding of content.

4. Mixture-of-AI Architecture

4.1 Analogy to Mixture-of-Experts

Mixture-of-Experts (MoE) is a well-established technique in machine learning that involves the use of multiple expert networks, each specialized in a particular skill or domain. A gating network directs inputs to the most appropriate experts, enabling

Thank you!
This is just an overview of our research paper.
Contact us to know more about our research.