Model Fusion for Scalable and Sustainable Artificial Intelligence: A Review and Outlook

Qi Zhou, Yiming Zhang, Yanggan Gu, Yuanyi Wang, Zhaoyi Yan, Zhen Li, Chi Yung Chung, and Hongxia Yang

***Abstract*—Large language models (LLMs) have achieved remarkable progress in recent years. Nevertheless, the prevailing centralized paradigm for training generative artificial intelligence (AI) is increasingly approaching its structural limits. First, the concentration of large-scale graphics processing unit (GPU) clusters restricts the access to the pre-training stage, confining the fundamental model development to a small number of resource-rich institutions. Second, the economic and energy costs associated with operating massive data centers render this paradigm progressively less sustainable. Third, the hardware gatekeeping narrows the participation to computer science specialists, limiting the involvement of domain experts who are essential for high-impact applications. Finally, small- and medium-sized enterprises remain dependent on expensive application programming interface (APIs) or shallow fine-tuning methods that are insufficient to modify the core knowledge of a model. Together, these constraints impede innovation and hinder equitable access to next-generation AI systems. Model fusion offers a scalable alternative by integrating multiple specialized models without retraining from scratch. This paper analyzes the current landscape of model fusion, outlining the strengths and limitations of existing methods and discussing future directions. We highlight recent advances such as InfiFusion, InfiGFusion, and InfiFPO, which improve the alignment and scalability through techniques like top-*K* logit selection, graph-based distillation, and preference optimization. These techniques demonstrate substantial efficiency and reasoning gains, pointing toward a more accessible and resource-aware paradigm for large-scale model development. Finally, we discuss the practical applicability of model fusion, using the energy domain as an illustrative example.**

***Index Terms*—Artificial intelligence (AI), large language model (LLM), model fusion.**

I. Introduction

With the rapid evolution of large language models (LLMs), we have witnessed remarkable progress and development of artificial intelligence (AI). These models have been widely applied across various fields such as smart manufacturing [1]-[3], financial investment research [4]-[6], and corporate customer service [7]-[9], bringing significant changes and benefits. The field of AI has made groundbreaking advancements, largely spurred by the rise of foundation models and large-scale generative architectures. Many recent foundation models are characterized by large parameter sizes and are trained on massive datasets containing trillions of tokens. These high-capacity models have enabled strong performance across a wide range of tasks, including natural language understanding, machine translation, text generation, and complex question answering. Qwen 2.5 [10], for example, features 72 billion parameters and is trained on 18 trillion tokens, illustrating the scale and power typical of modern AI systems.

However, the rapid progress of LLMs has brought about several pressing concerns. The increasing size and complexity of state-of-the-art models have led to a concentration of AI development within a few resource-rich institutions. These institutions benefit from substantial financial resources, specialized expertise, and privileged access to large-scale datasets, enabling them to develop and train increasingly complex models on massive computational infrastructures. This has created significant barriers to entry for smaller research groups and domain experts. Smaller teams often lack the financial resources to purchase and maintain high-performance computing equipment, the technical expertise to train and fine-tune large models, and access to large-scale and high-quality datasets. As a result, they struggle to compete with the giants in the field of large model research and application. Moreover, the monolithic training paradigm of large models, which relies on massive, diverse datasets, and extensive hardware, poses challenges for specialized or privacy-sensitive applications. Training a large model from scratch requires a comprehensive dataset covering a wide range of fields and topics. However, for specialized applications such as medical diagnosis or legal advice, the data are often highly specific and sensitive. Collecting and using such data for model training must comply with strict privacy regulations and ethical standards. Additionally, the extensive hardware requirements make it difficult for organizations in these specialized fields to carry out model training and deployment on their own. The high computational costs and complex technical operations limit the application of large models in

Manuscript received: October 10, 2025; revised: December 16, 2025; accepted: January 6, 2026. Date of CrossCheck: January 6, 2026. Date of online publication: January 30, 2026.

Q. Zhou, Y. Zhang, Y. Gu, Y. Wang, Z. Li, and H. Yang (corresponding author) are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China, and H. Yang is also with InfiX.ai, Hong Kong, China (e-mail: 25022441r@connect.polyu.hk; yiming1.zhang@connect.polyu.hk; yanggangu@outlook.com; wangyuanyi713@gmail.com; ethan-zhen.li@connect.polyu.hk; hongxia.yang@polyu.edu.hk).

Z. Yan is with InfiX.ai, Hong Kong, China (e-mail: yanzhaoyi@outlook.com).

C. Y. Chung is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: c.y.chung@polyu.edu.hk).

DOI: 10.35833/MPCE.2025.000973

these fields, making it hard to meet the specific needs of these fields and protect data privacy at the same time.

These challenges collectively underscore the urgent need for more flexible, efficient, and accessible strategies to harness the capabilities of modern AI systems. Crucially, enabling knowledge reuse and promoting modularity while maintaining model performance are essential for democratizing AI, as it allows broader communities to build, adapt, and deploy powerful models without prohibitive costs. As an alternative, building a unified multitask language model by integrating the capabilities of multiple specialized models presents a scalable and resource-efficient solution. However, this introduces new challenges in architectural compatibility, semantic alignment, and efficient knowledge transfer, which are those that this work aims to address. Existing methods for unifying multitask language models can be broadly categorized into two approaches: parameter-level model merging and knowledge distillation-based fusion, as shown in Fig. 1. Parameter-level model merging requires that all source models share the same architecture, while knowledge distillation-based fusion can be divided into logits-level and data-level fusion.

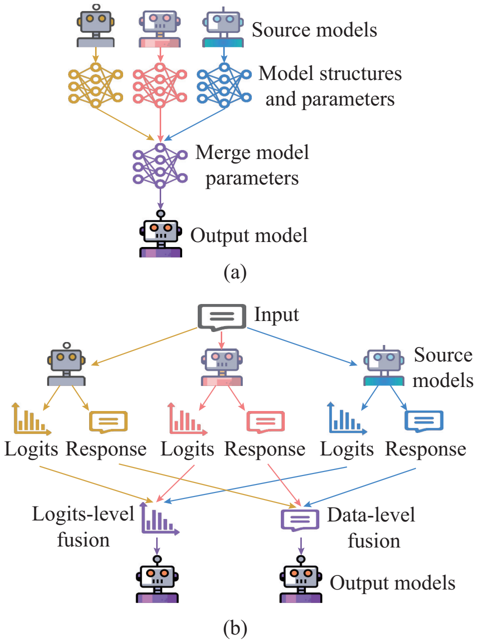

Fig. 1. Illustration of parameter-level model merging and knowledge distillation-based fusion. (a) Parameter-level model merging. (b) Knowledge distillation-based fusion.

Parameter-level model merging refers to the process of combining multiple pre-trained models into a single model that integrates their knowledge and capabilities. Instead of training a large model from scratch, this approach enables efficient reuse of existing models, often improving performance across multiple tasks while reducing the computational and data requirements. Early parameter-level model merging could be achieved by weighted averaging [11] and fisher-based merging [12]. However, these simple methods often overlook the conflicts between different models. In contrast, other merging methods such as drop and rescale (DARE) [13] and trim, elect sign & merge (TIES) [14] effectively address this limitation by resolving conflicts and improving the performance of the merged models. TIES merges multiple task-specific models into a single multitask model, mitigating parameter interference by trimming redundant values and resolving sign conflicts, which leads to improved performance. However, these merging methods are restricted to the models with identical architectures and vocabularies, and still cannot fully resolve the interference between conflicting task representations. More modular approaches like composition to augment language models (CALM) [15] introduce compositionality through cross-attention, but this comes at the cost of adaptability and requires carefully curated integration schemes.

Knowledge distillation-based fusion is an effective approach for integrating the capabilities of multiple models into a single compact model by transferring knowledge through both logits-level and data-level supervision. This approach enables the resulting model to retain the strengths of diverse source models while significantly reducing the computational overhead. Unlike parameter-level model merging, it does not require structural compatibility among source models, making it highly flexible and practical for real-world applications in resource-constrained environments.

Logits-level fusion captures not just the final predictions of expert models, but also their internal confidence levels across all possible outcomes. Before making a final decision, a model assigns numerical scores (called logits) to each option, indicating how likely it considers each one to be correct. Rather than simply copying the final answer, the student (or pivot) model learns from these full distributions of confidence. This allows it to replicate the nuanced reasoning and uncertainty of the expert models, leading to more accurate and calibrated decisions. Existing methods like FuseLLM [16] and FuseChat [17] show that combining models in this way can work well, but they still struggle with issues such as noisy low-confidence outputs, mismatched meanings, and inefficiency when dealing with very large sets of possible answers. To address these problems, methods such as InfiFusion [18] and InfiGFusion [19] have been proposed. InfiFusion reduces noise and improves alignment by focusing on the most important outputs and standardizing their scores, making it efficient for combining both two and multiple models. InfiGFusion further extends this idea by representing relationships between outputs as graphs and employing advanced mathematical tools [20], [21] to better capture and align their semantic structures, leading to significant improvements on complex reasoning tasks.

Data-level fusion, by contrast, focuses on transferring knowledge through the training examples and responses provided by expert models. Instead of learning from confidence scores, the pivot model is trained on a curated dataset consisting of input questions paired with detailed answers or explanations generated by the experts. By studying these rich examples, the pivot model absorbs the reasoning patterns and domain knowledge encoded in the expert responses, allowing it to generalize more effectively across similar tasks. Beyond logits-level fusion, data-level fusion provides an alternative avenue by transferring knowledge through sequence-level data distributions. Following this direction, InfiFPO [22] extends model fusion to the preference alignment stage by substituting the reference model in DPO [23] with a fused distribution from multiple sources. Through techniques such as length normalization, probability clipping, and max-margin fusion, it achieves stable and robust alignment across domains.

Together, these two approaches establish a flexible and unified framework for model fusion, enabling scalable and semantically aligned integration of heterogeneous models. By allowing effective fusion across different architectures and tokenizers without retraining from scratch, the model fusion lowers the barriers to developing powerful LLMs and contributes to a more open and collaborative AI ecosystem.

In the power and energy domain, the deployment of large-scale AI models is often constrained by limited data, high computational cost, and strict reliability requirements. Model fusion provides a promising solution by enabling the integration of multiple specialized models without training a single large monolithic network. Model fusion could be effective in applications such as load and renewable generation forecasting, security assessment, economic dispatch, and electricity market analysis.

The remainder of this paper is organized as follows. Section II reviews the development of model fusion methods, tracing their evolution and key paradigms. Section III discusses recent advancements, including the model merging scaling law and the knowledge distillation-based fusion. Section IV presents applications of model fusion, using the energy domain as a representative example, and outlines future research directions and emerging challenges.

II. Development of Model Fusion Methods

A. Parameter-level Model Merging

Integrating the capabilities of different models is an important research focus to build a multitask model. Therefore, researchers have explored methods to merge multiple models.

Early fusion techniques primarily focused on parameter-level model merging, which means direct integration of model parameters. Reference [24] simply combines source models to build a multitask model based on the weighted average weights of the source models. Also, Model Soup [11] is proposed in a similar way. After fine-tuning the pretrained models with different parameter configurations, Model Soup averages the weights of the source models to combine multiple models. Merging models by weighted average weights can be observed as combination of different task vectors from different models.

The task-vector-based model merging paradigm starts from a shared pretrained model $\boldsymbol{\theta}_{\text{base}}$, which serves as a common initialization for multiple domain-specific fine-tuning processes. Each fine-tuned model $\boldsymbol{\theta}_{t_k}^{\text{SFT}}$ represents the adaptation of the base model to a particular domain or task t_k, such as mathematics, coding, or scientific reasoning. The goal of merging is to integrate these specialized capabilities into a single unified model $\boldsymbol{\theta}_{\text{merge}}$ without requiring additional large-scale retraining. The basic formulation can be expressed as:

$$\boldsymbol{\theta}_{\text{merge}} = \boldsymbol{\theta}_{\text{base}} + \sum_{k=1}^{K} \lambda_k \left(\boldsymbol{\theta}_{t_k}^{\text{SFT}} - \boldsymbol{\theta}_{\text{base}} \right) \tag{1}$$

where λ_k is the scaling coefficient controlling the overall contribution of the k^{th} fine-tuned model; K is the number of source domains; and $\boldsymbol{\theta}_{t_k}^{\text{SFT}} - \boldsymbol{\theta}_{\text{base}}$ is the task vector of the task t_k.

This linear formulation follows the principle underlying task arithmetic (TA) [25], reflecting the assumption that knowledge acquired in different tasks can be approximately superposed in the parameter space. Despite its simplicity, this linear merging rule has been found to produce surprisingly strong performance when the fine-tuned models share a similar architecture and pretraining distribution.

However, recent studies reveal that the real-world parameter landscapes are highly non-linear, and pure linear composition may lead to conflicts or interference across domains. To address this, a number of variants introduce stochastic and adaptive mechanisms to enhance the robustness and generalization. Typical strategies include: ① random parameter dropout to mitigate co-adaptation and reduce noise accumulation across merged directions; ② noise injection or denoising regularization to smooth the parameter manifold and prevent overfitting to specific domain biases; and ③ rescaling and normalization of task vectors based on their magnitude or fisher information, aligning their relative contributions before aggregation.

TIES-merging [14] is proposed to address the interferences by resolving the sign conflicts when merging parameters and selecting the parameters that align with the final sign for merging. Similarly, DARE [13] is proposed to reduce most of the difference of parameters between the pretrained models and the fine-tuned models to mitigate the interferences in merging methods. Merging methods combine models of specific tasks to a multitask model without additional training, which is friendly to individuals or corporations with limited computational resources. CALM [15] enables a model to extend various capabilities by composition with other models using cross-attention. CALM does not modify parameters from the source models, and only some additional parameters are learned from a small amount of data. Therefore, it does not need mass data. However, the composition with other models reduces the flexibility and lack of adaptability.

These refinements effectively relax the linearity constraint in (1), leading to improved stability and consistent gains across diverse merging densities and task similarities. While effective, all these merging methods are restricted to homogeneous model families and cannot fuse heterogeneous tokenizers or sizes; they also suffer from "interference" when models specialize in conflicting skills, failing to capture the strengths of diverse specialized models.

B. Knowledge Distillation-based Fusion

Knowledge distillation-based fusion has emerged as a more flexible paradigm [26]-[28], as it enables the integration of heterogeneous models with varying architectures and sizes. Such fusion can be performed at both the logits level and the data level, offering greater adaptability across diverse model types.

1) *Logits-level Fusion*

In logits-level fusion, the new model (called the pivot model) learns by observing the confidence scores, known as logits, produced by several expert models when they answer the same question. These logits reflect the probability that each model assigns to every possible answer, not just the final choice. By aligning its own predictions with these detailed confidence patterns, the pivot model can absorb nuanced knowledge such as how certain or uncertain the experts are about different options. Importantly, this process does not require the expert models to have the same internal structure or even use the same vocabulary, making it a flexible way to combine knowledge from diverse source models into a single and more efficient model.

FuseLLM [16] creates a unified model by distillation. FuseLLM fuses the probabilistic matrices from multiple source models, which can be taken as teacher models, and then uses the fused probabilistic matrices to train the target model, which is taken as the student model. The researchers believe the probabilistic distributions can represent the inherent knowledge of the models, so fusing multiple probabilistic matrices to train a model can combine the knowledge of different specific domains. However, fusing multiple models simultaneously lacks adaptability because it does not seamlessly support inclusion of a new model to the fusion model. Therefore, FuseChat [17], another method fusing models through the probabilistic matrices, is proposed to resolve the limitation. Instead of fusing multiple models all at once, FuseChat selects a pivot model and fuses other source models with the pivot model pairwise by distillation, and then merges all the fused models to get the final model. FuseChat is a plug-and-play method, which makes it easy to add a new model to the final fused model. However, since different models usually have differences in the conversation templates and vocabularies, token alignment is needed to address the mapping between the probabilistic matrices from the source models to be fused. FuseChat does not give any systematic methods to resolve the problems brought by the conversation templates.

Researchers introduce universal logit distillation (ULD) [29] loss to address the limitation that the models do not share the same vocabulary and tokenizer in the distillation. The proposed solution is close to the solutions to optimal transport, but it does not provide a mapping for different models and it is also difficult for ULD to solve the alignment thoroughly. Dual-space knowledge distillation (DSKD) [30] unifies the output spaces of the student model and teacher model in distillation. For the student model and teacher model with different vocabularies, DSKD utilizes the cross-attention mechanism to learn the token alignment automatically instead of constructing mapping matrices. However, the performance is limited when the student model is relatively small.

Most logits correspond to low-probability categories, which contribute little to the distillation process but increase the computational burden. InfiFusion [18] enhances the ULD framework by incorporating top-K logit selection and logit standardization. These innovations effectively suppress the noise from low-probability tokens and improve the robustness of knowledge alignment across models with diverse vocabularies and output distributions. Moreover, InfiFusion supports both pairwise fusion and unified multi-source fusion, thus providing flexibility for different deployment scenarios while maintaining superior computational efficiency. InfiFusion demonstrates superior performance in reasoning, coding, mathematics, and instruction-following tasks through extensive experiments on multiple benchmarks.

InfiGFusion [19] introduces a novel graph-on-logits distillation (GLD) mechanism, which models token co-activation patterns as graphs and aligns semantic dependencies between source and pivot models using an efficient approximation of Gromov-Wasserstein distance. This structure-aware design enables InfiGFusion to capture relational knowledge that traditional token-wise distillation overlooks. Its effectiveness is particularly evident in complex reasoning tasks, where semantic dependencies across tokens are essential for correct inference.

2) *Data-level Fusion*

The data-level fusion involves training a new model on responses generated by expert models. Expert models act as tutors, providing specialized knowledge through their outputs, which serve as training material. This captures knowledge from closed-source models without requiring internal access, making it practical and modular.

The theoretical foundation stems from knowledge distillation work [31], which demonstrates that teacher-generated sequences can effectively transfer capabilities to student models. The rise of LLMs has accelerated the research on data-level fusion, focusing on instruction-following, reasoning, and domain-specific tasks.

Instruction synthesis represents the most extensive research area within data-level fusion. Techniques like Evol-Instruct [32] iteratively enhance the instruction complexity, while instruction fusion [33] combines different instruction types. Models like Alpaca [34], Vicuna [35], and Koala [36] demonstrate the effectiveness by training on diverse conversational data [37].

For reasoning capabilities, the Orca series [38], [39] pioneer augmenting responses with step-by-step explanations using Chain-of-Thought methodologies. Subsequent research works including MAmmoTH [40] and Mixed Distillation [41] extend these techniques to mathematical domains. Recent advances include DeepSeek-R1 [42], which leverages curated datasets for direct distillation to open-source models like Qwen [43] and LLaMA [44], and HuatuoGPT-o1 [45], which uses GPT-4o [46] to generate self-correcting reasoning processes.

Researchers have developed prompting strategies for controlling data diversity [47]-[49] and augmentation techniques like AugGPT [50] for semantic enhancement. LLMs serve as effective data generators for both natural language understanding and natural language generation tasks [51]-[56]. Projects like UltraChat [57] and FireAct [58] demonstrate quality-diversity balance and domain-specific applications.

Most research focuses on supervised fine-tuning (SFT), with limited investigation in preference alignment (PA). FuseChat 3.0 [59] addresses this by integrating fusion into both SFT and direct preference optimization (DPO) phases. The weighted-reward preference optimization (WRPO) algorithm [60] represents advances in policy fusion through weighted random optimization of multiple sub-policies.

Traditional preference alignment fusion methods have limitations: they use only response outputs, discard probabilistic information, and focus solely on preferred responses while neglecting dispreferred signals. InfiFPO [22] addresses these limitations through implicit model fusion based on sequence-level probabilities, replacing the reference model in DPO with a fused source model construction. It incorporates length normalization, probability clipping, and max-margin fusion strategies, achieving significant improvements in mathematics, coding, and reasoning tasks.

Figure 2 illustrates the evolution path of InfiFusion series for LLM fusion, including InfiFusion [18], InfiGFusion [19], and InfiFPO [22], where $P(\cdot)$ is the probability function; and the symbol $>$ represents a preference relation, i. e., the preferred responses have a higher priority or satisfaction level than the dispreferred responses. Parameter-level model merging is inherently limited to fusing the models with identical architectures, making the fusion of heterogeneous models a critical and increasingly important research area. The output of a model, whether represented as a probability distribution or as generated text, can be regarded as a reflection of its internal knowledge. From this perspective, fusing outputs naturally enables the integration of underlying knowledge across models. Knowledge distillation aligns well with this intuition by facilitating knowledge transfer through output alignment.

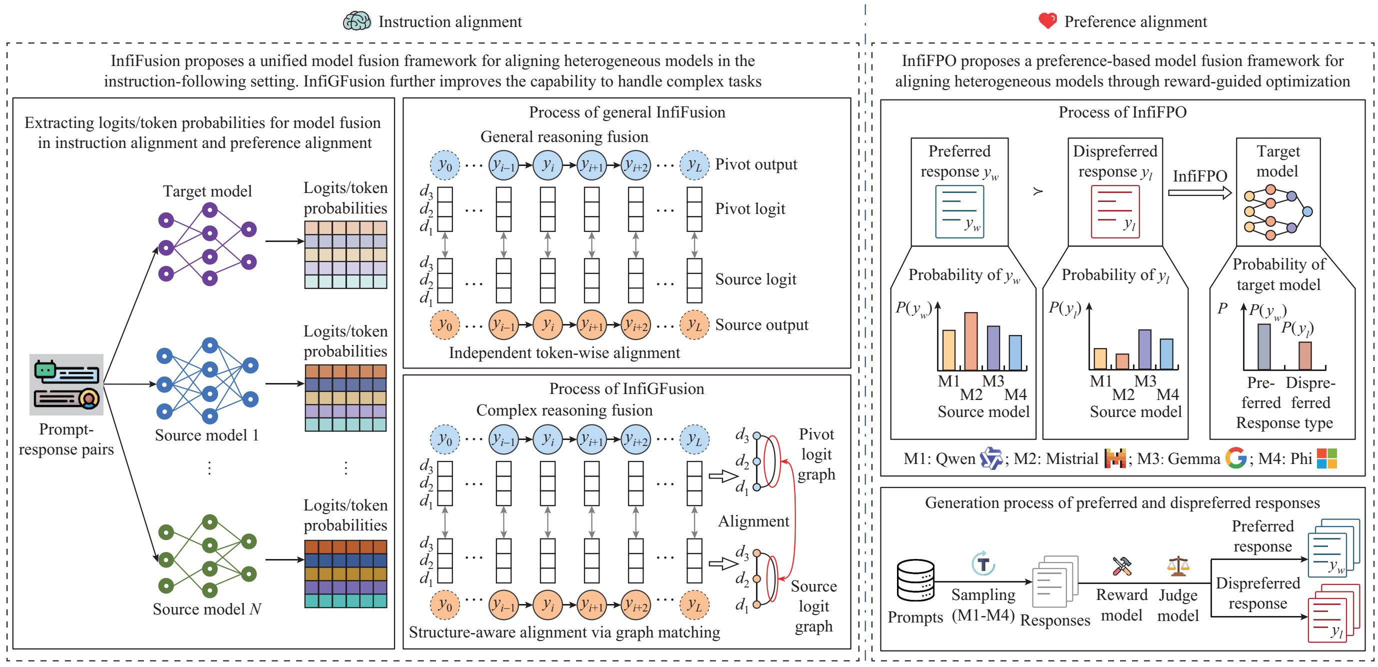
Instruction alignment
InfiFusion proposes a unified model fusion framework for aligning heterogeneous models in the instruction-following setting. InfiGFusion further improves the capability to handle complex tasks
Extracting logits/token probabilities for model fusion in instruction alignment and preference alignment
Target model
Logits/token probabilities
Source model 1
Logits/token probabilities
Source model N
Logits/token probabilities
Prompt-response pairs
Process of general InfiFusion
General reasoning fusion
y_0 y_{i-1} y_i y_{i+1} y_{i+2} y_L
Pivot output
Pivot logit
d_3 d_2 d_1
Source logit
Source output
Independent token-wise alignment
Process of InfiGFusion
Complex reasoning fusion
Pivot logit graph
Alignment
Source logit graph
Structure-aware alignment via graph matching
Preference alignment
InfiFPO proposes a preference-based model fusion framework for aligning heterogeneous models through reward-guided optimization
Process of InfiFPO
Preferred response y_w
Dispreferred response y_l
InfiFPO
Target model
Probability of y_w
$P(y_w)$
M1 M3 M2 M4
Source model
Probability of y_l
$P(y_l)$
M1 M3 M2 M4
Source model
Probability of target model
P $P(y_w)$ $P(y_l)$
Pre-ferred Dispre-ferred
Response type
M1: Qwen; M2: Mistrial; M3: Gemma; M4: Phi
Generation process of preferred and dispreferred responses
Prompts
Sampling (M1-M4)
Responses
Reward model
Judge model
Preferred response
Dispreferred response

Fig. 2. Evolution path of InfiFusion series for LLM fusion.

To provide a more systematic and neutral overview of different model-fusion paradigms and their key trade-offs, Table I summarizes representative model fusion methods in terms of typical prerequisites, training signals, qualitative computational costs, and main strengths and limitations.

III. Advancements

Model fusion has become a key paradigm for integrating knowledge from multiple specialized models while avoiding the prohibitive cost of retraining. Recent progress has centered on two major directions: parameter-level model merging and knowledge distillation-based fusion. The former operates directly in the weight space, combining fine-tuned models through arithmetic or statistical rules to obtain a unified model with diverse capabilities and minimal computation. The latter aligns models at the output or representation level, transferring knowledge through supervised or reinforcement signals that preserve semantic and behavioral consistency across heterogeneous architectures. Both directions share a common goal of efficiently consolidating expertise from multiple sources, yet they differ fundamentally in mechanism, data requirements, and interpretability.

In this section, we present recent advancements along these two lines. We first examine how parameter-level model merging has evolved from heuristic strategies into a principled and scalable framework, exemplified by the emergence of the model merging scaling law. We then discuss advances in distillation-based fusion, highlighting recent methodological innovations that improve alignment efficiency and performance across diverse model architectures.

A. Model Merging Scaling Law

Equation (1) in Section II defines a linear combination of task vectors, and serves as the foundation for several recent techniques discussed in this section. Beyond empirical success, model merging is now being examined through the lens of scaling laws [61], which describe how the performance scales with factors such as model size, number of experts, and domain diversity. Understanding these scaling behaviors provides a quantitative foundation for predicting fusion efficiency, identifying the optimal trade-offs between resource investment and performance, and guiding the design of future multi-model systems.

TABLE I
COMPARISON OF REPRESENTATIVE MODEL FUSION METHODS

Representative method	Typical prerequisite	Training signal	Qualitative computational cost	Main strength	Main limitation
Parameter averaging (AVG, Model Soup)	① Same architecture and tokenizer ② Checkpoints fine-tuned from a common base model	SFT; linear averaging of model weights	Low	Extremely simple; without additional training; convenient for quick multitask models	Requiring homogeneous tasks and domains; mainly used for closely related fine-tuned models in the same LLM family
TA (TA and related task-vector methods)	① Same architecture and tokenizer ② Shared pretrained base model	SFT; task vectors $\boldsymbol{\theta}_{\text{SFT}}-\boldsymbol{\theta}_{\text{base}}$ composed linearly	Low	Interpretable view of tasks as directions; supporting controlled rescaling and combination of capabilities; easy to analyze effects of individual tasks	Reliable on approximate linearity of the loss landscape; prone to performance degradation from conflicting task vectors
Conflict-aware merging (TIES, DARE)	① Same architecture and tokenizer ② Multiple task-specific fine-tuned models	SFT; task vectors merged with masking/rescaling to reduce conflicts	Low – Medium	Mitigating sign conflicts and parameter interference; more stable than naive averaging when merging many experts; maintaining low training overhead	Still restricted to identical architectures and vocabularies; residual interference persisting for highly conflicting skills
Modular composition (CALM-style compositional augmentation)	Compatible attention/interface between a base model and expert modules	SFT on adapter/cross-attention modules; source models frozen	Medium	Preserving original experts; modular and extensible; enabling plug in new domain specialists with small amounts of data; supporting composition of independently trained models	Increasing inference latency and memory footprint; less flexible than fully merged models
Logits-level fusion (FuseLLM, FuseChat)	Access to teacher logits	Logits-based distillation (e.g., KL, OT-style losses) into a pivot model	High	Fusing heterogeneous experts into a single deployable model; exploiting soft confidence patterns rather than hard labels; often more robust than simple ensembling	Being dominated by repeated teacher inference over large corpora in computation; cumbersome token and prompt alignment
Cross-tokenizer logit distillation (ULD, DSKD)	① Heterogeneous vocabularies ② Mechanisms for cross-token alignment required	Logits-based distillation with cross-tokenizer alignment objectives	High	Explicitly handling tokenizer mismatch; enabling logits-level fusion across different model ecosystems; broad fusion applicability	More complex implementation; mainly applying to fusion of strong but structurally heterogeneous teachers
Data-level fusion (Evol-Instruct, Alpaca, Orca, etc.)	① No structural compatibility requirement ② Access to expert-generated sequences	Supervised fine-tuning on synthetic instructions, rationales, or task solutions	Medium – High	Simple pipeline; reusing closed- or open-source experts; naturally supporting domain- and reasoning-oriented fusion via curated datasets; easy scalability with more data	Discarding fine-grained probabilistic information; strong performance dependence on prompt design and data filtering
Advanced logits-level fusion (InfiFusion, InfiGFusion)	Heterogeneous architectures and tokenizers supported via logit-space and graph-based alignment	Logits-based distillation with top-K selection, standardization, and graph-on-logits objectives	Medium	Suppressing low-probability noise; capturing structural dependencies for complex reasoning; achieving strong gains on different benchmarks with reduced GPU hours	Requiring access to teacher logits and curated fusion corpora; performance still bounded by teacher quality
Preference-based fusion (FuseChat 3.0, WRPO, InfiFPO)	① Access to preference data or implicit preference signals from multiple sources ② Architectures may differ	Preference-based objectives at policy/sequence level	Medium	Combining expertise from several models while explicitly aligning with human preferences; avoiding tokenizer conflicts via sequence-level fusion; directly targeting real-world utility and safety	Requiring high-quality preference data or reliable proxy rewards; risk of propagating teacher biases

1) Scaling Laws in Deep Learning

Scaling laws describe how the performance of a system changes as its fundamental resources such as model size, data volume, or computational budget increase. They reveal that the performance often follows predictable power-law trends rather than arbitrary fluctuations, enabling researchers to forecast accuracy and efficiency without exhaustive experimentation. Classical scaling laws quantify how loss scales with model size, data volume, and computational budget, leading to parameter/data power-laws and computation-optimal trade-offs [62]-[64]. Extensions explore transfer efficiency, precision and quantization scaling [65], and sparsity-induced trade-offs [66], among others. Together, these studies establish a quantitative framework for reasoning about resource allocation in large-scale model development [67]-[70].

2) Toward a Scaling Law of Model Merging

While traditional scaling laws focus on how the performance of a single model improves with more parameters or data, they do not address composition in weight space: how knowledge from multiple pretrained experts can be efficiently combined. Existing merging studies typically examine only a few experts, leaving the relationship between the number of merged models and the resulting performance underexplored. References [71] and [72] examine this question from theoretical and empirical perspectives, identifying the factors that influence merging success but without establishing a unified and predictive framework.

Recent research work [61] empirically analyzes how model-merging performance changes with both the number of merged experts (K) and the base model size (N). Figure 3 illustrates representative results of the scaling law, where each panel corresponds to one merging method, AVG (averaging the weights of different models), TA, TIES, and DARE, showing how the cross-entropy loss evolves as more experts are merged. This figure is redrawn based on the results reported in [61], with modifications in visualization style for clarity. Reference [61] selects multiple domain-specific experts from mathematics, coding, and science, merges them, and evaluates the merged models using cross-entropy loss, where lower cross-entropy loss indicates better performance. They collect results for model merging across model sizes from 5×10^{8} to 3.28×10^{10} with 1-9 experts, fit the empirical data with functional curves, and validate the fitted relationship on the base model with size of 7.27×10^{10}. The dots in Fig. 3 represent the empirical results obtained from actual merging experiments, while the smooth curves are fitted using the model merging scaling law, providing a quantitative model that captures the observed relationship among K, N, and the performance. Two clear and consistent patterns emerge from these fitted curves.

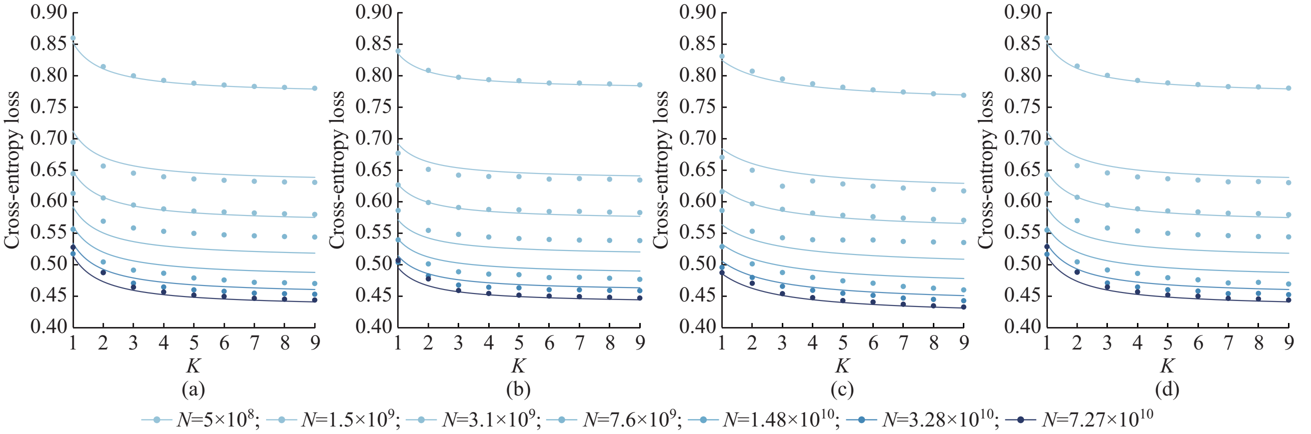

Fig. 3. Representative results of scaling law across different merging methods. (a) AVG. (B) TA. (c) TIES. (d) DARE.

1) Diminishing returns with increasing number of merged experts. Across all merging methods, the cross-entropy loss decreases monotonically (or nearly so) as the number of merged experts increases, following the scaling law. Most of the performance improvement occurs early: the curves exhibit a visible "elbow" around $K\approx$5-6, beyond which additional experts yield progressively smaller gains. This pattern reflects a diminishing-returns effect in multi-expert fusion, where early merging provides rapid benefits while later additions contribute marginal improvements.

2) Scaling with model size. It is observed that base models with larger sizes not only achieve better accuracy but also reach saturation more quickly. Moreover, the domain-dependent tendencies are evident: mathematics-related tasks tend to saturate earlier (shorter tails), whereas science-related domains continue to benefit from adding experts before plateauing.

Overall, the alignment between the experimental data points and the fitted scaling law curves demonstrates that the model-merging performance follows smooth and predictable regularities across different merging methods and model sizes. These results provide strong empirical evidence that the merging scaling law accurately characterizes the relationship among expert count, model capacity, and performance gain, transforming model fusion from a heuristic process into a quantitatively predictable paradigm.

B. Advances in Knowledge Distillation-based Fusion

Knowledge distillation-based fusion aligns heterogeneous models through their output behaviors rather than parameter values, enabling integration across architectures or vocabularies without structural constraints. Recent advances are exemplified by the InfiFusion series [18], [19], [22], as illustrated in Fig. 2. Three key stages of evolution could be summarized: ① instruction alignment, where models are fused by matching token-level outputs in instruction-following tasks; ② structure-aware alignment introduced by InfiGFusion, which incorporates graph-based reasoning alignment; and ③ preference alignment introduced by InfiFPO, which leverages reward-guided optimization to reflect human feedback. Together, these developments transform distillation-based fusion from heuristic knowledge transfer into a principle and multi-level framework for aligning heterogeneous LLMs.

1) InfiFsusion

Distillation-based fusion extends applicability to heterogeneous models by aligning knowledge at the output level. The InfiFusion [18] framework introduces mechanisms for aligning token distributions across different vocabularies, supported by robust objectives such as the optimal transport and distance-based alignment. Efficiency and stability are further enhanced through techniques like top-K logit selection and logits standardization, which not only concentrate the optimization on high-confidence outputs but also mitigate discrepancies caused by heterogeneous model architectures. These methodological refinements collectively ensure that the alignment is both semantically meaningful and computationally tractable. Based on this, the unified fusion jointly integrates multiple source models in a single optimization process, significantly reducing the GPU cost compared with pairwise training while delivering competitive or superior performance. For instance, a unified InfiFusion variant attains an average score of 79.92, closely matching a pairwise counterpart (79.96). Meanwhile, it consumes only about 160 GPU hours (less than the 450 GPU hours required by the pairwise model) and merely accounts for 0.016% of the GPU hours of base model with size of 1.4×10^{9} required for the full training of a comparable foundation model. These results demonstrate that the unified fusion not only scales more gracefully with the number of source models but also represents a fundamental shift from heuristic aggregation toward principled and optimization-based objectives for knowledge alignment. This makes it particularly relevant in the resource-constrained scenarios, where practitioners must balance model quality against strict limits on training budgets.

2) InfiGFsusion

Beyond token-level alignment, InfiGFusion [19] incorporates relational structures into the fusion process, addressing the critical limitation that the token-wise distribution matching alone often fails to capture higher-order reasoning patterns. By modeling token co-activation patterns as graphs and aligning them across models, InfiGFusion captures structural dependencies such as causality, logical consistency, and temporal or semantic ordering. This design moves fusion closer to reasoning-aware alignment rather than shallows probability matching. Empirically, InfiGFusion improves the average accuracy from 77.94% (SFT baseline) to 83.79% across diverse reasoning benchmarks, with especially large gains on multi-step tasks. These improvements illustrate that the structure-aware fusion allows models to internalize not only what tokens to generate, but also how to reason about relationships between them. Importantly, InfiGFusion achieves this without incurring prohibitive computational overhead, thanks to efficient graph approximation methods that summarize relational dependencies in a compact form. Such advances highlight a path forward for building fused models that exhibit stronger logical consistency, interpretability, and robustness, all of which are crucial for the domains such as scientific discovery, law, and energy system optimization, where reliable reasoning is indispensable.

3) InfiFPO

The fusion at the preference alignment stage remains relatively unexplored, despite its central role in aligning LLMs with human values and practical utility. InfiFPO [22] introduces a sequence-level implicit fusion strategy that circumvents vocabulary conflicts by aligning entire response sequences rather than individual tokens, thereby preserving semantic coherence across heterogeneous architectures. Built upon the FuseRLHF framework, InfiFPO integrates reinforcement learning from human feedback into the fusion process, enabling the pivot model to inherit both preference alignment and source model expertise. Reformulated as an efficient offline objective, InfiFPO yields substantial gains: using Phi-4 as the pivot and multiple models with sizes of 9×10^{9}-2.4×10^{10}, the average score across 11 benchmarks that can represent aggregate performance improves from 79.95 to 83.33, while avoiding expensive online sampling and reward model training. This positions preference-aligned fusion as both practical and scalable, lowering the barrier to integrating alignment into multi-model fusion pipelines. Enhancement strategies such as length normalization, probability clipping, and dynamic max-margin fusion further stabilize the training and mitigate the risks of overfitting or inheriting biased behaviors from source models. Taken together, these innovations elevate preference-aligned fusion into a promising direction for future research, as it provides a direct mechanism for combining heterogeneous LLMs while ensuring that the fused models not only achieve strong performance but also faithfully reflect human-centered objectives. Looking ahead, such techniques may become essential in deploying safe, robust, and socially aligned AI systems across domains where decision quality and ethical considerations are paramount.

All numerical performance results reported in this subsection for InfiGFusion and InfiFPO (including average scores and benchmark-level improvements) are quoted directly from the corresponding original publications rather than reproduced in this review. The benchmark lists, evaluation protocols, model scales, tokenizers, and experimental settings strictly follow those described in [19] and [22].

C. Operational Sustainability

From an operational perspective, recent advances in the Infi-series provide clear quantitative evidence that the model fusion can achieve competitive or superior performance under substantially reduced computational budgets. Specifically, InfiFusion and InfiGFusion integrate multiple source models within a single optimization process and require only about 160-195 GPU hours, whereas existing logits-level fusion baselines such as FuseLLM and FuseChat require approximately 225 GPU hours and 650 GPU hours, respectively, under the same model size of 1.4×10^{10} and comparable evaluation protocols [19]. This reduces GPU hours at the training stage by approximately 60%-75% with no loss (and often a gain) in average benchmark score, yielding a marked energy-efficiency improvement at the fusion stage.

A similar efficiency advantage is observed for the preference-based fusion. In InfiFPO, the preference optimization is completed using approximately 55-60 GPU hours, while achieving higher average performance than several baselines that rely on comparable or even larger computational budgets [22]. Compared with the much heavier training pipelines typically used in reinforcement-learning-based alignment and multi-stage fusion frameworks such as FuseChat, InfiFPO achieves stronger alignment effectiveness under an order-of-magnitude lower GPU budget, indicating that the effective preference alignment can be realized in a highly energy-efficient manner.

All in all, InfiGFusion and InfiFPO demonstrate that model fusion delivers better performance per unit time than pairwise distillation and multi-stage training. When the hardware setup stays the same, the longer these models run, the more electricity they use, and the more carbon emissions they produce. So, cutting down the time needed to train these models directly reduces their environmental impact.

This supports the view that the model fusion constitutes not only an effective strategy for capability integration, but also a practically sustainable pathway for developing and aligning large-scale foundation models.

IV. Applications and Future Directions

The value of model fusion lies not only in methodological innovation but also in enabling practical AI deployment across diverse domains. Applications in energy, healthcare, and finance illustrate the ability of model fusion to integrate heterogeneous data for more adaptive decision-making, while the growing data scale and model complexity call for advances in efficiency, scalability, and multimodality. This section first reviews representative applications of model fusion using the energy domain as an illustrative example, then outlines future directions and open challenges that will shape the next stage of the field.

A. Applications of Model Fusion

1) Applications of Parameter-level Model Merging

The applicability of parameter-level model merging spans a wide range of domains, including healthcare, finance, robotics, and industrial systems. For instance, in the energy domain, separate models are often developed for renewable energy generation forecasting, equipment condition monitoring, and energy market optimization. By merging these models, it becomes possible to construct an integrated system that holistically considers environmental dynamics, infrastructure reliability, and economic signals. Such a system can improve the accuracy of solar and wind power predictions, enable the proactive maintenance of generation and storage assets, and enhance the decision-making in real-time market operations. Ultimately, this contributes to more reliable, efficient, and sustainable energy systems.

Beyond these conceptual benefits, concrete use cases in modern power and energy systems already begin to align naturally with fusion-style designs. In the short-term load and renewable energy generation forecasting, the ensemble approaches combine physical models with diverse machine learning predictors to improve the robustness under non-stationary weather and demand patterns [73]-[75]. In the asset condition monitoring and fault diagnosis, practical deployments are inherently multi-source: measurements and inspection signals are collected through heterogeneous sensing and detection mechanisms, and reliable diagnosis often requires aggregating evidence across channels while being robust to noise and interference [76], [77]. In the electricity markets and dispatch, recent research works on electricity price forecasting have shown that heterogeneous machine learning models and ensemble schemes can be combined to enhance the accuracy and robustness of day-ahead and longer-horizon price predictions [78], [79]. These examples illustrate that the power and energy applications often have a natural multi-source structure, making them particularly well suited to benefit from principled fusion frameworks.

Thus, the parameter-level model merging offers a scalable and modular approach to democratizing AI capabilities across specialized tasks and domains. It provides a flexible framework for integrating diverse models while preserving their unique strengths, thereby facilitating the development of comprehensive and adaptive solutions in various application contexts.

2) Applications of Model Merging Scaling law

The model merging scaling law quantitatively describes how the performance of fused models improves as the number of merged experts and the base model size increase, offering a predictive framework for understanding the compositional efficiency. In the industrial contexts, this law can guide the design of scalable and resource-aware AI systems by identifying when the inclusion of additional expert models produces diminishing performance gains. In the energy domain, for instance, the forecasting and optimization tasks must account for diverse temporal horizons (e.g., short-term load prediction and long-term generation planning), spatial variations across regions, and dynamically changing environmental or market conditions. Traditionally, separate models are trained for sub-domains such as wind power forecasting, photovoltaic control, and grid stability analysis. By applying the model merging scaling law, practitioners can evaluate how the predictive accuracy scales with the number of merged domain experts and determine an optimal fusion point that balances accuracy, computational cost, and energy consumption. In practice, this enables adaptive and efficient model composition for real-time energy management, leading to more reliable, sustainable, and cost-effective power system operation.

Beyond the energy domain, the same principle extends naturally to other data-intensive fields where multiple specialized models coexist. In healthcare, for example, predictive systems often combine models trained on different data sources such as medical imaging, genomic profiles, and electronic health records, each capturing complementary aspects of clinical knowledge. The model merging scaling law provides a systematic framework for estimating how diagnostic accuracy or generalization is improved as more expert models are integrated. This helps researchers and practitioners allocate computational resources efficiently while maintaining data privacy by avoiding the need for joint retraining. In finance, applications such as multi-market forecasting and risk modeling rely on models specialized for distinct asset classes or temporal patterns. Scaling analysis helps determine the point at which adding more models yields diminishing returns, ensuring the computational efficiency in high-frequency decision-making environments. Similarly, in manufacturing and industrial automation, predictive maintenance and process optimization often depend on models tuned for specific sensors, machinery, or production lines. By quantifying the performance improvements as a function of the number of merged experts, organizations can plan gradual model integration strategies that match available hardware capacity and latency requirements.

More broadly, the model merging scaling law marks a transition from heuristic and trial-and-error fusion to a quantitative understanding of how composition scales with performance. It provides not only descriptive patterns but also practical predictions: a small set of early experiments can be used to forecast the entire performance-versus-scale trend, guiding the resource allocation before the large-scale deployment. This predictive capability supports a new paradigm of sustainable AI engineering, in which the model integration decisions explicitly account for the computational cost, environmental impact, and system efficiency. By turning model merging into a predictable and theoretically grounded process, the model merging scaling law establishes a unified foundation for scalable, cost-effective, and environmentally responsible AI deployment across scientific and industrial domains.

3) Applications of Knowledge Distillation-based Fusion

Model fusion techniques at the logits and data levels offer an important degree of flexibility for integrating heterogeneous models, particularly in settings where the parameter-level alignment is difficult or impractical. These techniques are especially relevant to application domains characterized by heterogeneous and multi-source data such as modern energy systems that combine time-series sensor streams and geospatial weather information. By enabling the fusion without requiring strict architectural compatibility or joint retraining, such methods open the door to more versatile and adaptive predictive frameworks that can better accommodate the diversity of data inherent in the energy-related applications.

Despite these advantages, the deployment of model fusion in energy contexts also presents significant challenges. The forecasting and optimization in power and energy systems often involve data that vary substantially across temporal horizons, geographic regions, and operational conditions, making the robustness and adaptability of fused models a critical concern. Moreover, the real-world scenarios such as distributed energy resources, grid balancing, and demand-side management impose strict constraints on latency, computational efficiency, and resource budgets, underscoring the need for compact yet capable models that can be deployed at the edge. Finally, as the sustainability becomes a central evaluation criterion, the energy cost of computation itself must be considered alongside the predictive accuracy. This highlights the importance of fusion strategies that are not only effective in performance but also resource-efficient, scalable, and aligned with the broader goal of sustainable energy intelligence.

While these issues are particularly salient in energy systems, the modularity and scalability of fusion methods make them equally valuable in other data-intensive fields. In these domains, combining models across institutions or data silos is often required under strict privacy and interoperability constraints. Model fusion thus represents a versatile tool for building high-performing AI systems in complex and data-diverse environments. By enabling the seamless integration of models trained on disparate data sources and architectures, the fusion techniques pave the way for more robust, adaptive, and sustainability-aware AI solutions across a wide range of applications.

B. Future Directions

The future trajectory of model fusion is defined not only by methodological innovations but also by its ability to address cross-domain challenges, particularly in the resource-constrained and energy-sensitive applications. Several interrelated research directions are central to this agenda.

First, a priority involves the development of plug-and-play fusion frameworks capable of integrating models across heterogeneous domains, architectures, and training paradigms. Current methods remain hindered by mismatches in tokenizers, objectives, and parameter scales, limiting their applicability in real-world pipelines. A modular and dynamic design could enable practitioners to integrate domain-specific models without full retraining, facilitating continual learning while ensuring adaptability in fast-evolving domains such as climate modeling and energy system optimization.

Second, the development of computation- and data-efficient fusion methods is indispensable for broadening participation and ensuring sustainability. Although the current fusion methods already offer significant efficiency gains compared with full retraining, they still entail non-negligible computational and data costs, which can pose barriers to broader adoption. Yet in energy-sensitive domains, the computational efficiency is not only a matter of accessibility but also that of sustainability: reducing training and inference energy costs is essential for lowering the carbon footprint of large-scale AI systems. Techniques such as sparsity, parameter sharing, and low-resource distillation will be crucial in making fusion both equitable and environmentally responsible.

Finally, extending fusion methods to multimodal contexts introduces both opportunities and unresolved barriers. Integrating modalities such as language, vision, and sensor data can enable richer multimodal inference; however, achieving alignment across heterogeneous representation spaces remains challenging. For instance, in energy applications, fusing textual reports with satellite imagery and real-time sensor streams requires principled methods for cross-modal embedding and distillation to ensure the reliable joint decision-making.

Taken together, these directions highlight the dual necessity of advancing theoretical principles and addressing application-driven challenges. Model fusion must evolve not only as a methodological discipline but also as a cross-domain enabler, with energy-efficient and application-aware designs serving as critical testbeds for its broader societal impact.

V. Conclusion

This review surveys the current landscape of research on democratizing AI, focusing on improving accessibility, scalability, and knowledge reuse in large-scale model development. It synthesizes recent progress in model fusion, summarizing key paradigms and representative techniques such as InfiFusion, InfiGFusion, and InfiFPO, which have advanced logits-level denoising, structural alignment, and preference-based sequence fusion. Recent studies on model merging scaling laws are also discussed, providing a quantitative perspective on how model performance evolves with the number of fused experts and offering theoretical guidance for the efficient and sustainable model composition. Furthermore, this review examines the practical applications of model fusion, using the energy domain as a representative case to illustrate its effectiveness in integrating heterogeneous models and enabling adaptive and resource-aware AI systems. Looking ahead, the future research is expected to move toward more general and multimodal fusion frameworks that enhance scalability, reasoning, and interpretability, paving the way for flexible and sustainable AI architectures.

References

[1] A. Kusiak, "Generative artificial intelligence in smart manufacturing," *Journal of Intelligent Manufacturing*, vol. 36, no. 1, pp. 1-3, Jan. 2025.

[2] J. Lee and H. Su, "A unified industrial large knowledge model framework in industry 4.0 and smart manufacturing," *International Journal of AI for Materials and Design*, vol. 3681, p. 20, Feb. 2024.

[3] T. Wang, J. Fan, and P. Zheng, "An LLM-based vision and language cobot navigation approach for human-centric smart manufacturing," *Journal of Manufacturing Systems*, vol. 75, pp. 299-305, Oct. 2024.

[4] Y. Yang, Y. Tang, and K. Y. Tam. (2023, Sept.). InvestLM: a large language model for investment using financial domain instruction tuning. [Online]. Available: https://arxiv.org/abs/2309.13064

[5] Y. Nie, Y. Kong, X. Dong *et al.*. (2024, Jun.). A survey of large language models for financial applications: progress, prospects and challenges. [Online]. Available: https://arxiv.org/abs/2406.11903

[6] Q. Xie, W. Han, Z. Chen *et al.*, "FinBen: a holistic financial benchmark for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 95716-95743, Dec. 2024.

[7] X. Ma, R Zhao, Y. Liu *et al.*, "Design of a large language model for improving customer service in telecom operators," *Electronics Letters*, vol. 60, no. 10, p. e13218, May 2024.

[8] J. Yun, J. E. Sohn, and S. Kyeong, "Fine-tuning pretrained language models to enhance dialogue summarization in customer service centers," in *Proceedings of the Fourth ACM International Conference on AI in Finance*, New York, USA, Nov. 2023, pp. 365-373.

[9] J. Wulf and J. Meierhofer. (2024, May). Exploring the potential of large language models for automation in technical customer service. [Online]. Available: https://arxiv.org/abs/2405.09161

[10] Q. Team. (2024, Sept.). Qwen2.5: a party of foundation models. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/

[11] M. Wortsman, G. Ilharco, S. Y. Gadre *et al.*, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *Proceedings of International Conference on Machine Learning*, Maryland, USA, Jul. 2022, pp. 23965-23998.

[12] M. S. Matena and C. A. Raffel, "Merging models with fisher-weighted averaging," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17703-17716, Oct. 2022.

[13] L. Yu, B. Yu, H. Yu *et al.* (2024, Jun.). Language models are super Mario: absorbing abilities from homologous models as a free lunch. [Online]. Available: https://arxiv.org/abs/2403.05286

[14] P. Yadav, D. Tam, L. Choshen *et al.*, "TIES-merging: resolving interference when merging models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7093-7115, Dec. 2023.

[15] R. Bansal, B. Samanta, S. Dalmia *et al.* (2024, Jan.). LLM augmented LLMs: expanding capabilities through composition. [Online]. Available: https://arxiv.org/abs/2401.02412

[16] F. Wan, X. Huang, D. Cai *et al.* (2024, Jan.). Knowledge fusion of large language models. [Online]. Available: https://arxiv. org/abs/2401.10491

[17] F. Wan, L. Zhong, Z. Yang *et al.* (2024, Aug.). FuseChat: knowledge fusion of chat models. [Online]. Available: https://arxiv. org/abs/2408.07990

[18] Z. Yan, Y. Zhang, B. He *et al.* (2025, Jan.). InfiFusion: a unified framework for enhanced cross-model reasoning via LLM fusion. [Online]. Available: https://arxiv.org/abs/2501.02795

[19] Y. Wang, Z. Yan, Y. Zhang *et al.* (2025, May). InfiGFusion: graph-on-logits distillation via efficient Gromov-Wasserstein for model fusion. [Online]. Available: https://arxiv.org/abs/2505.13893

[20] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-Wasserstein averaging of kernel and distance matrices," in *Proceedings of International Conference on Machine Learning*, New York, USA, Jun. 2016, pp. 2664-2672.

[21] H. Xu, D. Luo, and L. Carin, "Scalable Gromov-Wasserstein learning for graph partitioning and matching," in *Proceedings of 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2019, pp. 1-11.

[22] Y. Gu, Z. Yan, Y. Wang *et al.* (2025, May). InfiFPO: implicit model fusion via preference optimization in large language models. [Online]. Available: https://arxiv.org/abs/2505.13878

[23] R. Rafailov, A. Sharma, E. Mitchell *et al.*, "Direct preference optimization: your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53728-53741, Dec. 2023.

[24] J. Smith and M. Gashler, "An investigation of how neural networks learn from the experiences of peers through periodic weight averaging," in *Proceedings of 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, Mexico, Dec. 2017, pp. 731-736.

[25] G. Ilharco, M. T. Ribeiro, M. Wortsman *et al.*. (2022, Dec.). Editing models with task arithmetic. [Online]. Available: https://arxiv. org/abs/2212.04089

[26] G. Hinton, O. Vinyals, and J. Dean. (2015, Mar.). Distilling the knowledge in a neural network. [Online]. Available: https://arxiv. org/abs/1503.02531

[27] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *Advances in Neural Information Processing Systems*, vol. 27, pp. 1-9, Jan. 2014.

[28] J. Gou, B. Yu, S. J. Maybank *et al.*, "Knowledge distillation: a survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789-1819, Jun. 2021.

[29] X. Cui, M. Zhu, Y. Qin *et al.*, "Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Philadelphia, USA, Feb. 2025, pp. 23724-23732.

[30] S. Zhang, X. Zhang, Z. Sun *et al.* (2024, Jun.). Dual-space knowledge distillation for large language models. [Online]. Available: https://arxiv. org/abs/2406.17328

[31] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, USA, Nov. 2016, pp. 1317-1327.

[32] C. Xu, Q. Sun, K. Zheng *et al.* (2024). WizardLM: empowering large pre-trained language models to follow complex instructions. [Online]. Available: https://openreview.net/forum?id=CfXh93NDgH

[33] W. Guo, J. Yang, K. Yang *et al.*, "Instruction fusion: advancing prompt evolution through hybridization," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, Aug. 2024, pp. 3883-3893.

[34] R. Taori, I. Gulrajani, T. Zhang *et al.* (2023, May). Stanford alpaca: an instruction-following llama model. [Online]. Available: https://github. com/tatsu-lab/stanford_alpaca

[35] W. -L. Chiang, Z. Li, Z. Lin *et al.* (2023, Mar.). Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[36] X. Geng, A. Gudibande, H. Liu *et al.* (2023, Apr.). Koala: a dialogue model for academic research. [Online]. Available: https://bair.berkeley. edu/blog/2023/04/03/koala/

[37] C. Xu, D. Guo, N. Duan *et al.*, "Baize: an open-source chat model with parameter-efficient tuning on self-chat data," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Dec. 2023, pp. 6268-6278.

[38] S. Mukherjee, A. Mitra, G. Jawahar *et al.* (2023, Jun.). Orca: progressive learning from complex explanation traces of GPT-4. [Online]. Available: https://arxiv.org/abs/2306.02707

[39] A. Mitra, L. D. Corro, S. Mahajan *et al.* (2023, Nov.). Orca 2: teaching small language models how to reason. [Online]. Available: https://arxiv.org/abs/2311.11045

[40] X. Yue, X. Qu, G. Zhang *et al.* (2023, Sept.). Mammoth: building math generalist models through hybrid instruction tuning. [Online]. Available: https://arxiv.org/abs/2309.05653

[41] L. Chenglin, Q. Chen, L. Li *et al.*, "Mixed distillation helps smaller language models reason better," in *Proceedings of the Association for Computational Linguistics: EMNLP 2024*, Miami, USA, Nov. 2024, pp. 1673-1690.

[42] DeepSeek-AI. (2025, Jan.). DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. [Online]. Available: https://arxiv.org/abs/2501.12948

[43] A. Yang, B. Zhang, B. Hui *et al.* (2024, Sept.). Qwen2.5-math technical report: toward mathematical expert model via self-improvement. [Online]. Available: https://arxiv.org/abs/2409.12122

[44] A. Grattafiori, A. Dubey, A. Jauhri *et al.* (2024, Jul.). The Llama 3 herd of models. [Online]. Available: https://arxiv.org/abs/2407.21783

[45] J. Chen, Z. Cai, K. Ji *et al.* (2024, Dec.). HuatuoGPT-O1, towards medical complex reasoning with LLMs. [Online]. Available: https://arxiv.org/abs/2412.18925

[46] A. Hurst, A. Lerer, A. P. Goucher *et al.* (2024, Oct.). GPT-4o system card. [Online]. Available: https://arxiv.org/abs/2410.21276

[47] W. Chen, D. Song, and B. Li. (2024, Jan.). Grath: gradual self-truthifying for large language models. [Online]. Available: https://arxiv.org/abs/2401.12292

[48] Z. Sun, Y. Shen, Q. Zhou *et al.* (2023). Principle-driven self-alignment of language models from scratch with minimal human supervision. [Online]. Available: https://openreview.net/forum?id=p40XRfBX96

[49] M. Li, J. Chen, L. Chen *et al.*, "Can LLMs speak for diverse people? tuning LLMs via debate to generate controllable controversial statements," in *Proceedings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, Aug. 2024, pp. 16160-16176.

[50] H. Dai, Z. Liu, W. Liao *et al.* (2023, Feb.). AugGPT: leveraging ChatGPT for text data augmentation. [Online]. Available: https://arxiv.org/abs/2302.13007

[51] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, pp. 1-10, Jul. 2023.

[52] Y. Wang, Y. Kordi, S. Mishra *et al.*, "Self-Instruct: aligning language models with self-generated instructions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, Jul. 2023, pp. 13484-13508.

[53] X. He, Z. Lin, Y. Gong *et al.*, "AnnoLLM: making s to be better crowdsourced annotators," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, Mexico City, Mexico, Jun. 2024, pp. 165-190.

[54] S. Wang, Y. Liu, Y. Xu *et al.*, "Want to reduce labeling cost? GPT-3 can help," in *Proceeding of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, Nov. 2021, pp. 4195-4205.

[55] Y. Li, S. Bubeck, R. Eldan *et al.* (2023, Sept.). Textbooks are all you need II: Phi-1.5 technical report. [Online]. Available: https://arxiv.org/abs/2309.05463

[56] J. Jung, P. West, L. Jiang *et al.*, "Impossible distillation for paraphrasing and summarization: how to make high-quality lemonade out of small, low-quality model," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, Jun. 2024, pp. 4439-4454.

[57] N. Ding, Y. Chen, B. Xu *et al.*, "Enhancing chat language models by scaling high-quality instructional conversations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Dec. 2023, pp. 3029-3051.

[58] B. Chen, C. Shu, E. Shareghi *et al.* (2023, Oct.). FireAct: toward language agent fine-tuning. [Online]. Available: https://arxiv.org/abs/2310.05915

[59] Z. Yang, F. Wan, L. Zhong *et al.* (2025, Mar.). FuseChat-3.0: preference optimization meets heterogeneous model fusion. [Online]. Available: https://arxiv.org/abs/2503.04222

[60] Z. Yang, F. Wan, L. Zhong *et al.* (2024, Dec.). Weighted-reward preference optimization for implicit model fusion. [Online]. Available: https://arxiv.org/abs/2412.03187

[61] Y. Wang, Y. Gu, Y. Zhang *et al.* (2025, Sept.). Model merging scaling laws in large language models. [Online]. Available: https://arxiv.org/abs/2509.24244

[62] J. Kaplan, S. McCandlish, T. Henighan *et al.* (2020, Jan.). Scaling laws for neural language models. [Online]. Available: https://arxiv.org/abs/2001.08361

[63] J. Hestness, S. Narang, N. Ardalani *et al.* (2017, Dec.). Deep learning scaling is predictable, empirically. [Online]. Available: https://arxiv.org/abs/1712.00409

[64] J. Hoffmann, S. Borgeaud, A. Mensch *et al.*, "Training compute-optimal large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, Nov. 2022, pp. 30016-30030.

[65] T. Kumar, Z. Ankner, B. F. Spector *et al.* (2024, Nov.). Scaling laws for precision. [Online]. Available: https://openreview.net/forum?id=JjER3VgSSo

[66] J. Hilton, J. Tang, and J. Schulman. (2023, Jan.). Scaling laws for single-agent reinforcement learning. [Online]. Available: https://arxiv.org/abs/2301.13442

[67] N. Ardalani, C.-J. Wu, Z. Chen *et al.* (2022, Aug.). Understanding scaling laws for recommendation models. [Online]. Available: https://arxiv.org/abs/2208.08489

[68] T. Klug and R. Heckel. (2022, Sept.). Scaling laws for deep learning based image reconstruction. [Online]. Available: https://arxiv.org/abs/2209.13435

[69] O. Neumann and C. Gros. (2022, Oct.). Scaling laws for a multi-agent reinforcement learning model. [Online]. Available: https://arxiv.org/abs/2210.00849

[70] J. Geiping, M. Goldblum, G. Somepalli *et al.* (2022, Oct.). How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. [Online]. Available: https://arxiv.org/abs/2210.06441

[71] Z. Wang, X. Xu, Y. Liu *et al.* (2025, May). Why do more experts fail? a theoretical analysis of model merging. [Online]. Available: https://arxiv.org/abs/2505.21226

[72] P. Yadav, T. Vu, J. Lai *et al.* (2024, Oct.). What matters for model merging at scale? [Online]. Available: https://arxiv.org/abs/2410.03617

[73] K. Hou, X. Zhang, J. Yang *et al.*, "Short-term load forecasting based on multi-frequency sequence feature analysis and multi-point modified Fedformer," *Frontiers in Energy Research*, vol. 12, p. 1524319, Mar. 2025.

[74] Y. Yang, Y. Li, L. Cheng *et al.*, "Short-term wind power prediction based on a modified stacking ensemble learning algorithm," *Sustainability*, vol. 16, no. 14, p. 5960, Jul. 2024.

[75] A. L. Suárez-Cetrulo, L. Burnham-King, D. Haughton *et al.*, "Wind power forecasting using ensemble learning for day-ahead energy trading," *Renewable Energy*, vol. 191, pp. 685-698, May 2022.

[76] S. Li and J. Li, "Condition monitoring and diagnosis of power equipment: review and prospective," *High Voltage*, vol. 2, no. 2, pp. 82-91, Jul. 2017.

[77] D. Neupane, M. R. Bouadjenek, R. Dazeley *et al.* (2025, Mar.). Data-driven machinery fault diagnosis: a comprehensive review. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224012629

[78] C. O'Connor, M. Bahloul, S. Prestwich *et al.*, "A review of electricity price forecasting models in the day-ahead, intra-day, and balancing markets," *Energies*, vol. 18, no. 12, p. 3097, Jun. 2025.

[79] S. M. Gonzales, H. Iftikhar, and J. L. López-Gonzales, "Analysis and forecasting of electricity prices using an improved time series ensemble approach: an application to the Peruvian electricity market," *Aims Math*, vol. 9, no. 8, pp. 21952-21971, Aug. 2024.

Qi Zhou received the bachelor's degree from Southeast University, Nanjing, China, in 2022, and the master's degree from Harbin Institute of Technology (Shenzhen), Shengzhen, China, in 2025. Currently, she is pursuing her Ph. D. degree at The Hong Kong Polytechnic University, Hong Kong, China. Her research interests include model fusion and post-training of large language models (LLMs).

Yiming Zhang received the B.Eng. degree from Harbin Institute of Technology, Harbin, China, in 2019, and the M.Sc. degree from the Technical University of Munich, Munich, Germany, in 2023. Currently, he is a Ph.D. student at The Hong Kong Polytechnic University, Hong Kong, China. His research interest focuses on efficient LLM.

Yanggan Gu received the bachelor's degree from Guangdong University of Technology, Guangzhou, China, in 2018, and the master's degree from Soochow University, Suzhou, China, in 2022. He is currently a Ph.D. student

at The Hong Kong Polytechnic University, Hong Kong, China. His research interests focus on large language model, including model fusion and preference alignment.

Yuanyi Wang received the bachelor's and master's degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 2022 and 2025, respectively. He is currently a Ph.D. student at The Hong Kong Polytechnic University, Hong Kong, China. His current research interest includes model fusion.

Zhaoyi Yan received the bachelor's degree from Harbin Engineering University, Harbin, China, in 2016, and the Ph.D. degree in Computer Science from Harbin Institute of Technology, Harbin, China, in 2021. His research interests include knowledge distillation, model fusion, and LLM.

Li Zhen received the bachelor's degree from Jiangnan University, Wuxi, China, in 2021, and the master's degree from University of Science and Technology of China, Hefei, China, in 2024. Currently, He is pursuing his Ph.D. degree at The Hong Kong Polytechnic University, Hong Kong, China. His research research interests include efficient training and inference for LLMs.

Chi Yung Chung received the B.Eng. (with First Class Honors) and Ph.D. degrees in electrical engineering from The Hong Kong Polytechnic University, Hong Kong, China, in 1995 and 1999, respectively. He is the Head of Department, Chair Professor of Power Systems Engineering, and Founding Director of Research Centre for Grid Modernisation in the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. His research interests include smart grid, microgrid, renewable energy, power system stability/control, planning and operation, application of artificial intelligence, electricity market, and electric vehicle charging.

Hongxia Yang received the Ph.D. degree from Duke University, Durham, USA, in 2020. Now she is a Professor at The Hong Kong Polytechnic University, Hong Kong, China. Her research interests include generative artificial intelligence (GAI), reinforcement learning, and decentralized computing.