Engineering 44 (2025) 87–100

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Next Ten Years: Create a Better Future—Review

Knowledge-Empowered, Collaborative, and Co-Evolving AI Models: The Post-LLM Roadmap

Fei Wu [a,*], Tao Shen [a], Thomas Bäck [g], Jingyuan Chen [a], Gang Huang [h], Yaochu Jin [c], Kun Kuang [a], Mengze Li [f], Cewu Lu [b], Jiaxu Miao [e], Yongwei Wang [a], Ying Wei [a], Fan Wu [b], Junchi Yan [b], Hongxia Yang [d], Yi Yang [a], Shengyu Zhang [a], Zhou Zhao [a], Yueting Zhuang [a], Yunhe Pan [a,*]

[a] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[b] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[c] School of Engineering, Westlake University, Hangzhou 310024, China
[d] Department of Computing, The Hong Kong Polytechnic University, Hong Kong 999077, China
[e] School of Cyber Science and Technology, Sun Yat-Sen University, Shenzhen 518107, China
[f] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China
[g] Leiden Institute of Advanced Computer Science, Leiden University, Leiden 2333 CC, Netherlands
[h] College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 6 October 2024
Revised 9 November 2024
Accepted 8 December 2024
Available online 19 December 2024

Keywords:
Artificial intelligence
Large language models
Knowledge empowerment
Model collaboration
Model co-evolution

ABSTRACT

Large language models (LLMs) have significantly advanced artificial intelligence (AI) by excelling in tasks such as understanding, generation, and reasoning across multiple modalities. Despite these achievements, LLMs have inherent limitations including outdated information, hallucinations, inefficiency, lack of interpretability, and challenges in domain-specific accuracy. To address these issues, this survey explores three promising directions in the post-LLM era: knowledge empowerment, model collaboration, and model co-evolution. First, we examine methods of integrating external knowledge into LLMs to enhance factual accuracy, reasoning capabilities, and interpretability, including incorporating knowledge into training objectives, instruction tuning, retrieval-augmented inference, and knowledge prompting. Second, we discuss model collaboration strategies that leverage the complementary strengths of LLMs and smaller models to improve efficiency and domain-specific performance through techniques such as model merging, functional model collaboration, and knowledge injection. Third, we delve into model co-evolution, in which multiple models collaboratively evolve by sharing knowledge, parameters, and learning strategies to adapt to dynamic environments and tasks, thereby enhancing their adaptability and continual learning. We illustrate how the integration of these techniques advances AI capabilities in science, engineering, and society—particularly in hypothesis development, problem formulation, problem-solving, and interpretability across various domains. We conclude by outlining future pathways for further advancement and applications.

1. Introduction

In the field of artificial intelligence (AI) [1–4], large language models (LLMs) have revolutionized progress by achieving success across multimodal tasks. Models such as OpenAI o1 have demonstrated high capabilities in natural language understanding, generation, and reasoning. They have been instrumental in advancing applications ranging from conversational agents to complex problem-solving systems. Despite these achievements, LLMs have significant challenges that limit their effectiveness and applicability in certain domains:

- **LLMs suffer from inherent limitations that necessitate knowledge empowerment.** Their training on large-scale, unsupervised text corpora primarily results in models that encode knowledge implicitly within a vast number of parameters, leading to several issues including stale or outdated

* Corresponding authors.
E-mail addresses: wufei@zju.edu.cn (F. Wu), panyh@zju.edu.cn (Y. Pan).

https://doi.org/10.1016/j.eng.2024.12.008

information, hallucinations and inaccuracies, inability to reason over structured data, and lack of interpretability. These shortcomings highlight the need to integrate explicit knowledge sources to empower LLMs' factual accuracy, reasoning capabilities, and interpretability.

- **LLMs often struggle with efficiency and domain-specific accuracy**. LLMs require significant resources, making them impractical for certain applications and environments. Moreover, their decision-making processes can be opaque, limiting their interpretability. Model collaboration leverages the complementary strengths of various models, integrating the capabilities of large models with the efficiency and specialization of smaller or different functional models. This approach enhances performance, usability, and transparency, addressing the inherent limitations of LLMs.
- **LLMs exhibit limitations in adaptability and continual learning**. They are generally trained on static datasets and may not effectively incorporate new information or adapt to evolving tasks without extensive retraining. To overcome these issues, models need to mutually promote each other's learning processes in order to achieve co-evolution, which will enable them to stay current and effective in dynamic environments.

To address these challenges, integrating external knowledge—which provides semantically rich representations of entities and relationships—into LLMs has emerged as a promising direction [5–7]. This integration occurs through several technical approaches:

(1) **Integrating external knowledge into training objectives.** These methods design knowledge-aware loss functions by assigning higher masking probabilities to important entities or balancing token-level and entity-level losses.

(2) **Incorporating knowledge into LLM inputs.** These methods inject relevant subgraphs into input sequences using mechanisms such as visible matrices to mitigate knowledge noise.

(3) **Knowledge-empowered instruction tuning.** In this approach, models can fine-tune LLMs to comprehend external knowledge by converting them into natural language prompts and employing self-supervised tasks.

(4) **Retrieval-augmented knowledge fusion during inference.** These methods combine non-parametric retrieval modules with LLMs to dynamically fetch and incorporate pertinent knowledge information.

(5) **Knowledge prompting.** This approach transforms external knowledge into textual prompts for LLMs without retraining, although it often requires manual prompt engineering.

To improve efficiency and domain-specific accuracy, some works [8–10] have explored the collaborative interplay between LLMs and smaller models (SMs). In the context of this survey, we use the term "SMs" to refer to models that have significantly fewer parameters than LLMs and lack the latter's emerging properties. Model collaboration involves the interaction of AI models with varying architectures, sizes, and functionalities to enhance their overall performance. This approach allows models to combine their strengths—such as the efficiency of SMs with the powerful capabilities of larger models—to improve their accuracy, interpretability, and computational efficiency. Model collaboration can be categorized into strategies such as model merging and functional model collaboration. These methods enable the integration of diverse AI techniques, leading to better performance and adaptability across tasks.

To advance adaptability and continual learning, model co-evolution harnesses the mutual evolutionary processes between LLMs and SMs to enhance performance and computational efficiency in multimodal tasks. Model co-evolution refers to the simultaneous evolution of multiple models that influence each other's development over time while working together to solve complex and diverse tasks in various environments. In this dynamic process, models influence one another by sharing knowledge, parameters, and learning strategies, which helps them adapt to heterogeneous conditions such as different architectures, tasks, and data distributions. Through co-evolution, models can balance the need for specialization and generalization, making them more robust and efficient—particularly in decentralized and federated learning settings where privacy and resource constraints are critical.

Together, knowledge empowerment, collaboration, and co-evolution form an interconnected framework to enhance AI capabilities beyond individual models by achieving levels of reasoning, accuracy, and adaptability unattainable by isolated models. The functionality of each component may depend on the functionality of the others. Knowledge empowerment sometimes relies on collaboration to effectively integrate and utilize external knowledge sources. Collaboration acts as a catalyst for co-evolution, as interacting models influence each other's development. In turn, co-evolution enhances both knowledge empowerment and collaboration by fostering continuous adaptation and learning.

Furthermore, the ternary space made up of the cyberspace, physical world, and human society (CPH) has expanded the interplay among science, engineering, and society, leading to new dimensions of interaction and development. All these advancements are inseparable from the technologies of the post-LLM era—specifically, knowledge-empowered, collaborative, and co-evolving AI models—which further improve and facilitate these complex interactions. As depicted in Fig. 1, in the post-LLM era, the integration of such techniques has the potential to tackle complex challenges in hypothesis development, problem formulation, problem-solving, and interpretability. Hypothesis development now leverages domain-specific knowledge within AI models to improve accuracy and reliability. Problem formulation has advanced through the modeling of entities, environments, and laws using multi-agent systems, such as simulating personalized roles in educational settings to uncover pedagogical principles and employing physics-informed neural networks (PINNs) to incorporate physical laws for improved predictive accuracy in, for example, fluid mechanics and heat conduction [11,12]. In problem-solving, the shift from symbolic logic reasoners to large-scale neural networks has enabled models to either retrieve knowledge from databases or memorize and generate complete solutions, with collaborative agent systems enhancing mathematical problem-solving through the separation of computation and verification tasks. Interpretability has been improved by integrating standardized operating procedures (SOPs) into multi-agent workflows for better task decomposition and coordination, and through enhanced human–computer interaction, with multiple LLM-based agents collaborating via natural language and programming exchanges to refine software development processes.

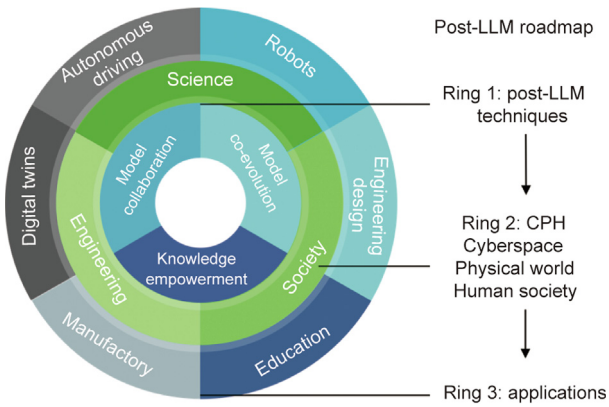

Fig. 1. Post-LLM roadmap.

In this light, this review aims to examine post-LLM techniques, addressing ongoing challenges in science, engineering, and society, and illuminate future pathways to further advance AI applications. Fig. 2 depicts the overall outline of this survey. In Section 2, we introduce the current challenges presented by AI in the areas of knowledge empowerment, collaboration, and co-evolution. Section 3 gives an overview of knowledge-empowered LLMs. In Section 4, we present cutting-edge methods in model collaboration. Section 5 delves into recent techniques for model co-evolution. In Section 6, we explore how knowledge-empowered, collaborative, and co-evolved AI advances science, engineering, and society. Section 7 showcases potential future advancements and applications of knowledge-empowered, collaborative, and co-evolved AI. Finally, in Section 8, we summarize the key insights from this survey.

2. Challenges

In this section, we identify four major types of challenges for current AI models: task heterogeneity, model heterogeneity, data heterogeneity, and security and privacy concerns.

2.1. Task heterogeneity

Existing AI models are primarily developed for distinct tasks, scenarios, and applications with differing or even conflicting optimization objectives and evaluation metrics, which results in theoretical and practical challenges regarding collaboration and co-evolution among these task-specific models. We identify three types of research challenges in task heterogeneity. First, the disparities in training objectives may hinder the model's evolutionary process, which particularly occurs in the model training phase. A notable example is optimizing generative adversarial networks, in which a generator and a discriminator are jointly optimized in an adversarial manner, making it extremely difficult to reach an equilibrium. Thus, it is a challenging problem to balance the divergent objectives and stabilize the training dynamics. Second, the lack of shared knowledge can prevent collaboration and co-evolution. This is because, while models made for completely different tasks may develop unique expertise, such knowledge cannot be easily leveraged across tasks without a common framework. Third, it is difficult for models to reach consensus due to communication barriers between different models; thus, interpreting and acting on outputs from other models becomes challenging.

2.2. Model heterogeneity

Model heterogeneity mainly refers to drastic architecture discrepancies between different AI models that lead to crucial challenges hindering model synergism. Typical examples range from collaboration between models with different architectures and levels of complexity to models with divergent learning paradigms. First, differing input and output representations may make it difficult to align the features well. For example, two convolutional backbones may differ significantly in model depth and width (e.g., hybrid model collaboration), resulting in a varying number of neurons and feature maps and eventually leading to inflexibility for model collaboration. Besides, due to incompatibility in intermediate representations, two fundamentally different learning paradigms may pose a major challenge to model collaboration. An example is the question of how to enable collaboration between symbolic AI and connectionistic AI, which requires the translation of logical rules into numerical formats or vice versa. Therefore, effectively extracting, transferring, and aligning shared knowledge between heterogeneous models will promote model utilization.

2.3. Data heterogeneity

In real-world scenarios, data from different devices or sources are often not independent and identically distributed (non-IID), resulting in significant variations in data distributions. For example, data collected from different end users, functionally different sensors in embodied AI, different patients, or different enterprises may differ in features, labels, and domains, leading to phenomena including class imbalance, covariate shift, and concept drift that remarkably affect the generalization performance of model coordination.

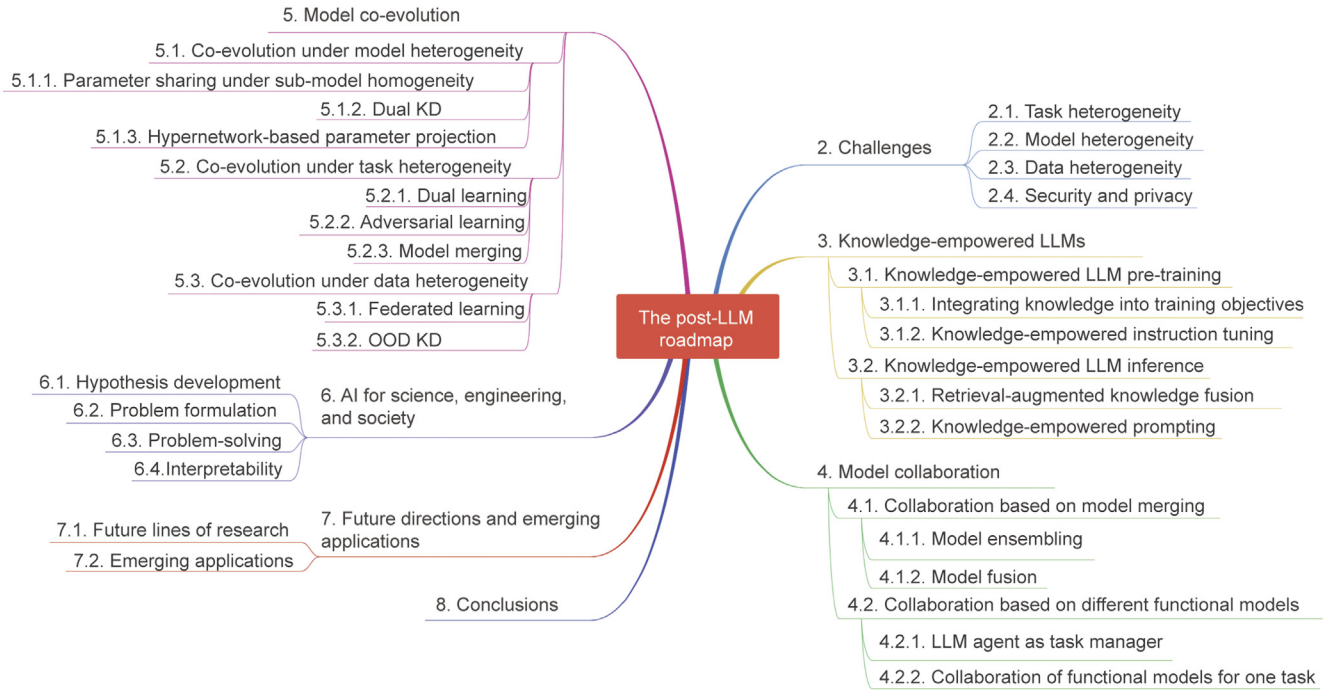

Fig. 2. Outline of the survey. OOD: out-of-distribution; KD: knowledge distillation.

Moreover, data from multiple sources may be inconsistently labeled or may exhibit varying levels of annotation quality. For example, some data might have been mislabeled or contain noise, which is likely to introduce performance degradation during model collaboration or co-evolution. Data modality differences introduce additional challenges, such as inconsistent data representations (e.g., spatial images versus temporal audio) and imbalanced data modalities (e.g., missing or sparse modality). Data heterogeneity introduces particular challenges due to varying data distributions, modality-specific challenges, fusion difficulties, and training complexities. It is essential to effectively address these challenges in order to build collaborative and robust systems that can handle complex, real-world tasks involving diverse data sources.

2.4. Security and privacy

As protected by laws and regulations (e.g., General Data Protection Regulation (GDPR)), data security and privacy are critical concerns in model collaboration, especially in distributed and decentralized machine learning systems, where multiple entities (e.g., devices and organizations) contribute to training a global model without sharing raw data. Even though raw data is not directly shared, model updates through gradients or features can still inadvertently expose sensitive information about the underlying data. For example, certain patterns in gradients can be reverse-engineered to reconstruct the original data. Also, collaborated models are vulnerable to inference attacks that exploit the learned model or its outputs to deduce information about the training data, with typical threats such as model inversion attacks and membership inference attacks. Although mechanisms such as differential privacy offer a potential solution to guarantee privacy, the excessive noise injected by the framework can reduce the overall model performance. Thus, protecting against such privacy attacks while maintaining the utility of the model is a significant challenge in collaborative environments. Moreover, models are vulnerable to poisoning attacks, in which adversaries attempt to corrupt the global model by injecting malicious updates. For example, adversaries can send malicious model updates that deliberately degrade the performance of the global model, often targeting specific sub-tasks or objectives. While robust aggregation mechanisms (e.g., Byzantine-resilient algorithms) can detect malicious clients, designing such a mechanism is complex, especially in environments where clients' contributions are diverse and their trustworthiness cannot be assumed. As collaborative AI systems continue to grow, security and privacy concerns will remain central to the development of secure and privacy-preserving model-collaboration paradigms.

3. Knowledge-empowered LLMs

Their reliance on unsupervised training on large-scale corpora often leaves LLMs devoid of practical, real-world knowledge, limiting their current applicability in knowledge-intensive tasks. To bridge this gap, researchers have explored various strategies to empower LLMs with external knowledge sources. These approaches involve integrating knowledge during pre-training through specialized training objectives, augmenting model inputs with relevant information, and leveraging knowledge during instruction-tuning and inference. This section delves into these methodologies, outlining how they enhance LLMs' capabilities by making them more knowledgeable and effective.

3.1. Knowledge-empowered LLM pre-training

Existing LLMs mostly rely on unsupervised training on a large-scale corpus and thus lack practical real-world knowledge. Previous works that integrate knowledge into LLMs can be categorized into two parts: ① integrating knowledge into training objectives, and ② knowledge-empowered instruction tuning.

3.1.1. Integrating knowledge into training objectives

Zhou et al. [13] constructed a minimal, high-quality dataset and fine-tuning protocol to align pre-trained models with user interaction style, leveraging stylistically coherent yet topically diverse prompts and responses. Akyürek et al. [14] employed the technique of integrating domain knowledge into training objectives by comparing and contextualizing two types of training data attribution (TDA) methods (gradient-based and embedding-based methods), which analyze model behavior at different stages of the training process to assess influence on predictions, alongside a baseline information retrieval method (BM25) that uses lexical similarity for fact tracing without model dependency. The research efforts in this category focus on designing knowledge-aware training objectives. For example, Shen et al. [15] leveraged a knowledge graph (KG) structure to assign a masking probability. Entities that can be reached within a certain number of hops are considered important and are given a higher masking probability during pre-training. Zhang et al. [16] further controlled the balance between token-level and entity-level training losses. Tian et al. [17] followed a similar fusion approach to inject sentiment knowledge during LLM pre-training by determining words with positive and negative sentiment and assigning a higher masking probability to those identified as sentiment words. It feeds both sentences and corresponding entities into LLMs and trains them to predict alignment links between textual tokens and entities in KGs. Gao [18] enhanced input tokens by incorporating entity embeddings and includes an entity prediction pre-training task. Wang et al. [19] directly employed both a KG embedding training objective and a masked token pre-training objective into a shared transformer-based encoder. The deterministic LLM [20] focused on pre-training language models to capture deterministic factual knowledge. It only masks the span that has a deterministic entity as the question and introduces additional clue contrast learning and clue classification objective. Xiong et al. [21] first replaced entities in the text with other same-type entities and then feeds them into LLMs and pre-trains the model to distinguish whether the entities have been replaced or not.

3.1.2. Knowledge-empowered instruction tuning

Ji et al. [22] demonstrated elasticity within LLMs, where the model's alignment can be inversely adjusted through a compression-based protocol, revealing a resistance to alignment that favors the retention of broader pre-training distributions over fine-tuning adjustments. Zhang et al. [23] shed some light on various kinds of instruction-tuning techniques. Gekhman et al. [24] examined how the inclusion of "Unknown" examples within the fine-tuning dataset affects a model's performance; the researchers find that an increased proportion of these examples not only risks overfitting but also hampers the model's generalization, while "MaybeKnown" examples prove most beneficial for balanced performance across knowledge types. KG instruction tuning utilizes facts and the structure of KGs to create instruction-tuning datasets. LLMs finetuned on these datasets can extract both factual and structural knowledge from KGs, enhancing their reasoning ability. Wang et al. [25] first designed several prompt templates to transfer structural graphs into natural language text and then proposes two self-supervised tasks to finetune LLMs. OntoPrompt [26] proposed an ontology-enhanced prompt tuning that can place knowledge of entities into the context of LLMs and fine-tune them on several downstream tasks. Luo et al. [27] fine-tuned LLMs on a KG structure to generate logical queries. Luo et al. [28] presented a planning-retrieval-reasoning framework, fine-tunes on a KG

structure to generate relation paths, and uses these paths to retrieve valid reasoning paths from the KGs for LLMs to conduct faithful reasoning and generate interpretable results.

3.2. Knowledge-empowered LLM inference

While the methods described in Section 3.1 could effectively fuse knowledge into LLMs, they are limited because real-world knowledge changes and they do not permit updates without retraining. Thus, recent research has focused on keeping knowledge and text spaces separate during inference, particularly for question answering (QA) tasks.

3.2.1. Retrieval-augmented knowledge fusion

Ovadia et al. [29] evaluated the leveraging of an auxiliary knowledge base to retrieve relevant information for a given query, combining it with pre-existing model context to enhance a language model's responses to knowledge-intensive tasks. This approach outperforms traditional fine-tuning by offering dynamic, contextually enriched knowledge integration. Retrieval augmented generation (RAG) combines nonparametric and parametric modules. Yang et al. [30] involved an iterative multi-stage process, IM-RAG, in which a reasoner, retriever, refiner, and progress tracker collaborate through reinforcement learning and supervised fine-tuning; this combination enables an LLM to construct, refine, and finalize answers by progressively retrieving, refining, and synthesizing relevant information in a structured, retrieval-augmented reasoning loop. Given input text, we can retrieve relevant documents via maximum inner product search [31], treat them as hidden variables, and feed them into the output generator as additional context. The model presented in Lewis et al. [32] outperforms other baseline models in open-domain QA and can generate more specific and factual text. Story-fragments improves the architecture by adding a module to determine salient knowledge entities. Wu et al. [33] improved efficiency by encoding external knowledge into memory and using fast search. Guu et al. [34] proposed a knowledge retriever for the pre-training stage to improve open-domain QA. Logan et al. [35] selected facts from a KG using the current context to generate sentences. Zhang et al. [36] leveraged multimodal large language models (MLLMs) in conjunction with a neural combinatorial optimization solver to address the combinatorial explosion challenge of ancient manuscript restoration, implementing a two-stage pipeline in which MLLMs perform initial fragment matching while a neural solver optimizes candidate fragment selection, particularly in open-world settings with outliers. Sun et al. [37] represented a KG triple as a sequence of tokens and concatenates them with the sentences, randomly masking either the relation token in the triple or tokens in the sentences. However, this approach may cause knowledge noise. Sun et al. [38] used unified word-knowledge graph to further reduce knowledge noise. Zhang et al. [39] intended to improve LLMs' representations toward those entities by determining long-tail entities and replacing them with pseudo token embedding. Yu et al. [40] leveraged external dictionaries to improve the representation quality of rare words by appending their definitions from the dictionary at the end of input text and training the language model to align rare word representations and discriminate whether the input text and definition are correctly mapped.

3.2.2. Knowledge-empowered prompting

Knowledge-empowered prompting designs a prompt to convert structured knowledge into text sequences for LLMs during inference. Li et al. [41] used a predefined template to convert KG triples into short sentences. Luo et al. [42] sampled relation paths from KGs, verbalizes them, and feeds them into LLMs to generate logical rules. Chain-of-knowledge (CoK) [43] uses a sequence of triples for prompting to elicit LLMs' reasoning ability. KG prompting is a simple way to combine LLMs and KGs without retraining, but the prompt is usually manually designed and thus requires a great deal of effort.

4. Model collaboration

Research on collaboration between AI models is an increasingly prominent field, centered on the cooperation of models with different sizes, structures, or functions. The goal is to leverage the models' respective strengths to achieve performance or efficiency superior to that of a single model. This collaborative approach not only focuses on the complementarity between large and small models but also involves the integration of different types of models, such as deep learning models and traditional machine learning models, to harness the powerful capabilities of large models alongside the efficiency and interpretability of small models. With the rapid advancement of deep learning technology, large models have gained significant attention due to their outstanding performance; however, they often require substantial computational resources, which limits their application in resource-constrained environments and increases their opacity, making their decision-making process difficult to understand. Therefore, exploring the collaborative modes of models to enhance performance and usability has become a research hotspot.

Model collaboration can be categorized into two types based on the collaboration strategy. The first type is model merging, exemplified by the mixture of experts (MoEs) [44], which combines several relatively small expert models to achieve or even surpass the performance of a large model. The second type involves the collaboration of different functional models, such as using a large model agent to coordinate specialized small models to complete specific tasks [45].

4.1. Collaboration based on model merging

In the field of machine learning, a single model often struggles to achieve optimal performance. Model merging is an effective strategy to improve prediction accuracy and robustness; it enhances performance by combining the prediction results, structures, or parameters of multiple models to mitigate the shortcomings of individual models.

4.1.1. Model ensembling

One type of model merging, known as model ensembling, is performed by aggregating the predictions of individual models [46]. The most straightforward model ensemble approach is the simple averaging method, where the final prediction is obtained by averaging the prediction results of all models. However, this method is only reasonable when the performance of each classifier is similar. If one classifier performs significantly worse than the others, the final prediction may not be as good as that of the best classifier in the group. A better approach to ensemble classifiers is to use weighted averaging, where the weights are learned from the validation set. For classification problems, voting [47] is a commonly used model ensemble strategy. The final prediction is selected by having multiple models vote on the predicted classes, with the class receiving the most votes being chosen. Voting can involve either running different models or running a single model multiple times. Stacking [48] is a more complex model ensemble method that uses the prediction results of multiple different models as inputs to train a new model, which then produces the final prediction. This approach effectively leverages the predictive capabilities of different models.

4.1.2. Model fusion

MoE [49] is a sparsely gated deep learning model consisting of two key components: a gate network (GateNet) and expert

networks (Experts). The gate network is responsible for dynamically deciding which expert model should be activated based on the input data's characteristics in order to generate the best prediction. Experts are a group of independent models, each specialized in handling a specific sub-task. Through the gate network, the input data are allocated to the most suitable expert model for processing, and the outputs of different models are weighted and fused to obtain the final prediction result. For example, Mixtral 8×7B [50], a modification of the Mistral 7B model, is a sparse MoE models that includes eight experts per layer. This results in a 47B parameter model that—against several benchmarks—can rival or outperform larger models such as Llama2 70B [51].

Model collaborative computing based on model merging can integrate the strengths and expertise of various models and reduce the bias and errors that may arise from a single model, thereby improving the accuracy and reliability of decisions. Moreover, model fusion can enhance models' interpretability and transparency. For example, in an MoE systems, each expert model's role in and contribution to specific tasks can be clearly identified, providing clearer explanations for the final decision.

4.2. Collaboration based on different functional models

Another typical model collaboration approach is an intelligent agent system composed of multiple functional models. While large models provide broad knowledge and advanced reasoning capabilities, such as mathematical reasoning, programming, and task planning [52], they may be less accurate in handling domain-specific tasks compared with smaller, specialized models. Thus, an effective mechanism is needed to integrate the general capabilities of large models with the specialized expertise of small models, ensuring that the agent system can flexibly handle different tasks and environments.

Collaboration based on different functional models can be divided into two types. In one type of collaboration, LLMs act as intelligent agents, serving as task managers that call upon various specialized models to accomplish different tasks. In the other type, LLMs work together with other specialized models, such as diffusion models, to complete a specific task. With the support of LLMs, the task can be executed more effectively.

4.2.1. LLM agent as task manager

Researchers have begun building intelligent agent systems based on the collaboration between LLMs and small specific models [53]. Specifically, they use LLMs as the brains or controllers of these agents, extending the perception and action space by scheduling SMs. Early works were aimed at enhancing the tool-learning capabilities of LLMs. For example, both tool augmented language models (TALMs) [54] and Toolformer [55] fine-tune language models to learn to use external tool application programming interface (API). HuggingGPT [56] further utilizes LLMs as the brain and SMs as tools, solving complex problems through collaboration between LLMs and SMs.

Chain of thought (CoT) [45], tree of thoughts (ToT) [57], and graph of thoughts (GoT) [58] techniques enable LLM-based agents to demonstrate reasoning and planning capabilities comparable to those of symbolic and reinforcement learning-based agents [59]. These systems can also learn from feedback and execute new actions, gaining the ability to interact with their environment [60]. LLM-based agents can interact seamlessly, forming multi-agent systems that promote collaboration and competition between multiple agents [61].

4.2.2. Collaboration of functional models for one task

LLMs can help specialized models perform specific tasks more effectively. For example, in image-generation tasks, while Stable Diffusion [62] can generate high-quality images, it struggles to control the output strictly based on the prompts. LLMs can better understand prompts and guide the behavior of the generation model, leading to improved controllability in the image-generation process. Wu et al. [63] proposed a framework that generates an image from the input prompt, assesses its alignment with the prompt, and performs self-corrections on the inaccuracies in the generated image. Steered by an LLM controller, this framework turns text-to-image generation into an iterative closed-loop process, ensuring correctness in the resulting image. Wang et al. [64] proposed a training-free method for text-to-image generation and editing. It utilizes the reasoning ability of MLLMs to improve compositionality in diffusion models. This method breaks down complex image generation into simpler tasks for different subregions using regional diffusion. It integrates text-guided generation and editing in a closed-loop system, improving its generalization capabilities.

Some specialized SMs can also enhance the capabilities of MLLMs. For example, Sachin et al. [65] used visual models such as semantic segmentation and instance segmentation to improve MLLMs' performance in object-counting tasks.

5. Model co-evolution

Model co-evolution refers to a dynamic process in which multiple models evolve together to solve complex, heterogeneous tasks and share insights across diverse environments. In this context, models not only adapt and improve based on their individual learning paths but also influence each other's development, ensuring efficient cross-task generalization, parameter sharing, and knowledge transfer. This process becomes essential in scenarios characterized by varied architectures, task requirements, or data distributions, as co-evolution enables models to collaboratively address the heterogeneity by balancing specialization and generalization. The resulting co-adaptation yields models that are more robust, efficient, and capable of solving a wider array of tasks, especially under the constraints of resource limitations and privacy concerns, which are typical of decentralized and federated environments.

This section is organized into three subsections that explore the co-evolution of models under different types of heterogeneity—namely, model, task, and data heterogeneity. Section 5.1 focuses on co-evolution under model heterogeneity, discussing techniques such as parameter sharing, dual knowledge distillation (KD), and hypernetwork-based parameter projection. Section 5.2 addresses co-evolution under task heterogeneity, examining methods such as dual learning, adversarial learning, and model merging. Lastly, Section 5.3 explores co-evolution under data heterogeneity, with a focus on federated learning and out-of-distribution (OOD) KD. Each section examines specialized strategies for optimizing model collaboration and efficiency in diverse environments.

5.1. Co-evolution under model heterogeneity

5.1.1. Parameter sharing under sub-model homogeneity

In the context of parameter sharing under sub-model homogeneity, recent works have significantly advanced the balance between model-specific learning and shared parameter efficiency. Haller et al. [66] introduced "sparse sharing," which utilizes overlapping subnetworks within a larger model, improving parameter efficiency through iterative magnitude pruning (IMP), based on the Lottery Ticket Hypothesis. Ding et al. [67] extended this idea by proposing the multiple-level sparse sharing model (MSSM), which enables more granular control through task-specific and shared features at different network levels. Wang et al. [68]

introduced multitask prompt tuning (MPT), which distills shared knowledge into transferable prompts for efficient adaptation across LLMs. Zhang et al. [69] employed a shared encoder across tasks in their contrastive learning model for blind image-quality assessment, dynamically adjusting shared parameters to boost performance. In a different domain, Chen et al. [70] introduced group detection transformer, which applies a group-wise parameter-sharing mechanism across object queries, significantly improving the efficiency of detection transformers. Ghosh et al. [71] proposed iterative federated clustering algorithm (IFCA), in which shared representation layers are employed across user clusters in federated learning, enabling parameter sharing across distributed environments while preserving cluster-specific learning. Lastly, Ye et al. [72] presented OpenFedLLM, a federated learning framework for LLMs, in which parameter sharing across decentralized systems is achieved through federated instruction tuning and value alignment, allowing collaborative learning without exposing raw data. These works collectively highlight the power of parameter sharing to enhance model efficiency, reduce redundancy, and enable robust performance across heterogeneous task settings and domains.

5.1.2. Dual KD

Dual KD has emerged as a pivotal strategy under the paradigm of model co-evolution, particularly addressing the challenges of model heterogeneity. In this approach, models simultaneously assume the dual roles of both student and teacher, fostering bidirectional knowledge transfer and enhancing learning efficacy across diverse architectures. Unlike traditional unidirectional distillation, dual KD leverages mutual learning, as demonstrated in frameworks such as mutual contrastive learning (MCL) [73], adaptive cross-architecture mutual knowledge distillation (ACMKD) [74], and all-in-one knowledge distillation (AIO-KD) [75]. For example, AIO-KD enables the simultaneous optimization of multiple student models through dynamic gradient detaching and mutual learning strategies, optimizing knowledge exchange without sacrificing the teacher model's performance. Similarly, in the context of semi-supervised learning, multistage collaborative knowledge distillation (MCKD) [76] refines pseudo labels iteratively across multiple student models, preventing overfitting and fostering generalization in sequence-generation tasks. This duality is also critical in tasks such as text-to-image synthesis, in which an adaptive teacher–student collaboration [77] refines student outputs through iterative guidance by means of an oracle mechanism. Additionally, frameworks such as Selective-FD [78] ensure that knowledge sharing is efficient and accurate, selectively filtering ambiguous or OOD predictions in federated learning environments. Collectively, these methods demonstrate the power of dual KD for addressing both architectural and domain-specific discrepancies and thus enhancing model performance and generalization through iterative and collaborative learning processes.

5.1.3. Hypernetwork-based parameter projection

The concept of hypernetwork-based parameter projection has emerged as a robust strategy for addressing model heterogeneity in co-evolutionary systems, particularly when dealing with large-scale models such as pre-trained language models. Hypernetworks, originally introduced to generate weights for target networks, can facilitate the transfer of information across heterogeneous models by learning a mapping from a shared latent space to the diverse parameter spaces of different models. This projection technique is especially beneficial in scenarios where models have been fine-tuned on distinct tasks or domains and a unified mechanism is required to harmonize the varied representations. By utilizing hypernetworks, it becomes feasible to dynamically generate task-specific parameters for a target model, effectively adapting the model to different inputs or tasks without the need for exhaustive retraining. In the context of knowledge fusion, hypernetworks allow for the seamless integration of heterogeneous model outputs, as demonstrated by approaches such as knowledge fusion for large language model (FUSELLM) [79] and mixture-of-adaptations (AdaMix) [80], in which the alignment of tokenizations or adaptation modules is a critical factor. This method aligns well with other model averaging techniques, such as model soups [81] and ensemble strategies [82], by enhancing the parameter space exploration while preserving the unique characteristics of each model through modularity. Additionally, methods such as regression mean (RegMean) [83] and ranking-based merging (RankMean) [84], which focus on parameter fusion without requiring downstream data, highlight the flexibility of hypernetwork-based projection in optimizing the fusion of diverse model parameters. By effectively navigating the parameter projection space, hypernetworks can create a more coherent and efficient model co-evolution process in heterogeneous environments.

5.2. Co-evolution under task heterogeneity

5.2.1. Dual learning

Dual learning has emerged as a powerful paradigm for tackling task heterogeneity in model co-evolution by leveraging the intrinsic duality between paired tasks to enhance learning efficiency and performance across diverse domains. For unbiased learning to rank (ULTR), Yu et al. [85] proposed the contextual dual learning algorithm with listwise distillation (CDLA-LD), which combines a listwise-input ranking model employing self-attention to capture local context with a pointwise-input model for distilling relevance judgments, outperforming existing methods on the Baidu-ULTR dataset by mitigating position and contextual biases. For constrained optimization, Park and Van Hentenryck [86] introduced self-supervised primal-dual learning (PDL), a method that jointly trains primal and dual networks without pre-solved instances by mimicking the augmented Lagrangian method to balance optimality and feasibility, achieving negligible constraint violations and minor optimality gaps. Fei et al. [87] enhanced dual learning by aligning structural information between tasks, introducing syntactic structure co-echoing and cross-reconstruction in text-to-text generation, and using syntactic–semantic alignment in text-to-non-text scenarios, thus significantly improving performance across tasks such as machine translation and image captioning. For video captioning, Ji et al. [88] developed an attention-based dual learning (ADL) approach that establishes a bidirectional flow between videos and captions using a multi-head attention mechanism to focus on effective information, resulting in more accurate and coherent captions. Li et al. [89] presented a multi-pass dual learning (MPDL) framework for stylized dialogue generation, leveraging mappings among the context and responses of different styles and incorporating discriminators to ensure stylistic consistency, and achieved state-of-the-art results. Additionally, frameworks such as dual learning enhanced auto-reflective translation (DUAL-REFLECT) [90] enhance LLMs for reflective translation through dual learning feedback mechanisms, while the dual learning with dynamic KD (DL-DKD) framework [91] integrates contrastive language–image pre-training (CLIP) models into partially relevant video-retrieval tasks by employing a teacher–student network with dynamic KD, further demonstrating the merits of dual learning for addressing task heterogeneity under model collaboration.

5.2.2. Adversarial learning

Adversarial learning is pivotal in model co-evolution under task heterogeneity. It frames objectives as adversarial games between competing models or components, thereby enhancing robustness,

alignment, and performance across diverse tasks. An LLM-enhanced adversarial editing system for lexical simplification employs confusion and invariance losses to predict lexical edits, effectively distinguishing complex words from simple ones while preserving semantics [92]. Latent adversarial training removes undesirable behaviors in LLMs by using targeted adversaries to elicit and mitigate harmful outputs [93]. In AI-text detection, adversarial learning between a paraphraser and a detector enhances robustness against paraphrasing attacks [94]. Worst-class adversarial training addresses class imbalance in adversarial robustness by focusing on improving the worst-performing classes using no-regret dynamics [95]. In weakly supervised semantic segmentation, adversarial learning between a classifier and a reconstructor improves segmentation precision by encouraging the classifier to produce more accurate class activation maps [96]. By fostering adversarial interactions, these methods effectively tackle the challenges of task heterogeneity in collaborative models.

5.2.3. Model merging

Model merging under task heterogeneity is a technique for creating a unified model that can handle heterogeneous tasks while minimizing interference. Basic methods such as parameter averaging [97], despite being straightforward, often result in suboptimal performance due to task conflicts. To address this, weighted-based approaches, such as spherical linear interpolation [98], optimize merging coefficients by evaluating the importance of each model or task vector, with some techniques extending this to layer-wise or parameter-specific weighting using methods such as layer-wise adaptive model merging [99] or merging models with fisher-weighted averaging [100]. Subspace-based methods, including trim, elect sign and merge (TIES-MERGING) [101] and drop and rescale (DARE) [102], focus on pruning unimportant parameters and leveraging the over-parameterized nature of neural networks to merge sparse subspaces, thereby reducing task interference. Routing-based strategies dynamically adjust merging during inference, thus adapting to input-specific variations. Examples include twin-merging [103] and weight-ensembling MoE [104], which use routing networks to guide the merging process. Finally, post-calibration techniques, such as representation surgery [105], address representation bias in merged models by aligning the representations of merged and independent models to enhance performance. Together, these methods provide a sophisticated toolkit for merging models in multi-task learning environments in order to optimize performance while addressing the complexities introduced by task heterogeneity.

5.3. Co-evolution under data heterogeneity

5.3.1. Federated learning

Federated learning addresses the challenges of data heterogeneity in model co-evolution by leveraging the powerful multimodal capabilities of LLMs and the low computational requirements and swift response times of SMs. The essence of federated learning lies in enabling LLMs to enhance the performance of SMs in domain-specific tasks while rigorously protecting data privacy. To augment LLMs' performance through training, OpenFedLLM [72] involves a comprehensive pipeline that includes federated instruction tuning (FedIT) and federated value alignment (FedVA), which optimize instruction adherence and model alignment while safeguarding data privacy. Zhang et al. [106] introduced multimodal large language model assisted federated learning (MLLM-FL) to bolster federated learning performance on heterogeneous and long-tailed data distributions through global multimodal pretraining, federated finetuning, and global alignment, effectively mitigating data heterogeneity while minimizing privacy risks and computational burdens on client devices. Bai et al. [107] developed a federated learning scheme for fine-tuning LLMs that dynamically adjusts the low-rank adaptation (LoRA) ranks based on individual client resources, thus enhancing the effective use of diverse client capabilities and improving generalization across heterogeneous tasks and resources. For collaborative model performance enhancement, FedMKT [108] involves a framework for federated selective mutual knowledge transfer and token alignment using a minimum edit distance, which enhances the performance of both LLMs and SMs. To improve the effectiveness of SMs through LLMs, Li et al. [109] extracted generalized and domain-specific knowledge from LLMs via synthetic data generation and then transfers this knowledge to local SMs while preserving privacy. Fan et al. [110] developed PDSS, a framework that employs the step-by-step distillation of LLMs to augment the capabilities of SMs, utilizing advanced strategies for prompt and rationale encoding to maintain information integrity during the perturbation and subsequent distillation of domain-specific knowledge.

5.3.2. OOD KD

KD involves training computationally efficient specialized models as student models to replicate the performance of more powerful LLMs as teacher models. This process reduces resource demands without significantly impacting performance, thereby facilitating broader deployment of LLMs. Traditional distillation techniques using synthetic or data-free approaches often suffer performance declines in OOD scenarios. To address these challenges, Gholami et al. [111] used a task-agnostic framework for OOD KD, which iteratively leverages feedback from LLMs to refine the specialized models, thus enhancing their generalizability. Li et al. [112] targeted OOD distillation challenges in vision language models (VLMs) by improving prompt coherence and enriching language representations in teacher models in order to better align vision-language tasks between teacher and student models. Agarwal et al. [113] developed generalized knowledge distillation (GKD), which uses reinforcement-learning-based fine-tuning to align the training and inference distributions, informed by the teacher model's feedback on the student models' outputs. Chen et al. [114] used a perturbation distillation approach that integrates modifications in score, class, and instance levels to distill knowledge to SMs, specifically addressing domain generalization challenges.

6. AI for science, engineering, and society

The post-LLM era marks a significant shift in the role of AI across multiple domains—particularly in science, engineering, and society. These domains share common challenges and unique characteristics that necessitate the tailored application of AI methodologies. Fig. 3 depicts the outline of this section, which elaborates on hypothesis development, problem formulation, problem-solving, and the interpretability of AI applications in science, engineering, and society, exploring how knowledge, collaboration, and co-evolution underpin these advances.

6.1. Hypothesis development

The development of hypotheses is a foundational challenge shared across science, engineering, and society domains. Hypotheses can take on various forms, depending on the domain. In science, hypotheses often serve as theoretical propositions aimed at explaining natural phenomena and are typically crafted to be empirically tested [29]. For example, hypotheses in scientific research might predict the effects of a specific variable on a biological process or forecast the outcome of a chemical reaction under certain conditions. In engineering, hypotheses often manifest as

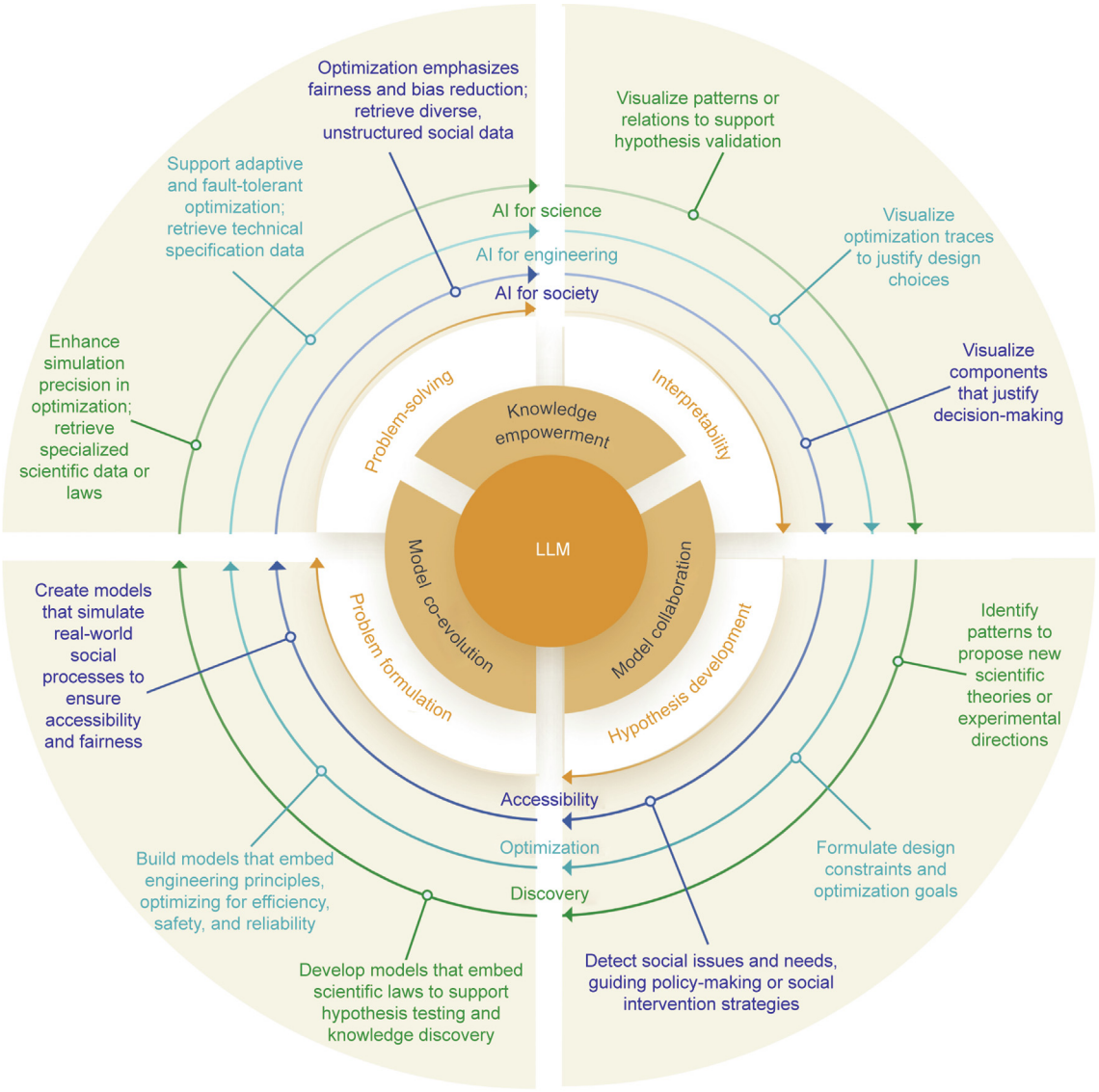

Fig. 3. Outline of AI for science, engineering, and society.

objectives designed to achieve specific goals or meet operational constraints [115]. For example, the operation of complex systems such as power grids, space stations, or autonomous vehicles often requires setting hypotheses related to efficiency goals and safety constraints. These hypotheses are more practical and serve as a basis for system design and decision-making, helping engineers determine the optimal settings and controls for achieving the desired performance under the given limitations. In societal contexts, hypotheses are often related to behavioral or policy outcomes [22]. For example, an AI model might hypothesize that specific interventions (e.g., public awareness campaigns or infrastructure adjustments) could lead to better outcomes in areas such as healthcare accessibility or traffic management. These hypotheses are typically tested in simulations or pilot programs prior to broader implementation. Despite the diversity in hypothesis types across these domains, there are shared categories of hypotheses, such as predictions about system behavior under various scenarios or predictions that serve as the basis for simulation models in order to validate different configurations before real-world implementation. These shared and unique hypotheses guide subsequent formulation and problem-solving processes in all three domains.

In the post-LLM era, knowledge-empowered AI models are instrumental in crafting these hypotheses, as they incorporate domain-specific expertise and thus enhance both accuracy and reliability. For example, advanced meteorological AI models such as Pangu [116], FengWu [117], and FuXi [118] could be integrated with domain-specific knowledge to improve renewable energy (e.g., wind and solar) forecasting, which is crucial for the integration of renewable energy sources into power systems. Collaboration among multiple smaller AI models also plays a critical role in validating hypotheses by cross-verifying outcomes from diverse perspectives, thereby enhancing the robustness of the hypotheses. This collaborative approach helps mitigate biases and provides a more holistic understanding of the problem space. Moreover, co-evolution fosters the iterative refinement of hypotheses. Through ongoing learning from both successes and failures, models can evolve to develop more nuanced and effective hypotheses. In this way, the post-LLM advancements contribute significantly to transforming hypothesis development, enabling deeper theoretical reasoning and more extensive data-driven exploration across science, engineering, and societal applications. The iterative process of co-evolution leads to hypotheses that are more adaptive to changing

environments, better aligned with domain-specific challenges, and ultimately more capable of driving meaningful advancements in each respective field.

6.2. Problem formulation

The application of large models for modeling the real world is currently a focal point of science, engineering, and society research. There are three types of modeling in this research domain: the modeling of objective entities, the modeling of objective environments, and the modeling of objective laws.

For entity modeling, multi-agent systems are introduced to effectively simulate personalized roles such as students and teachers in educational scenarios [119]. The strategic integration of simulated rules based on large model agents has propelled the educational field forward by uncovering the principles of education and teaching.

For environment modeling, the key challenge is how to realize the organized interaction between multiple agents. A proposal has emerged for a virtual classroom platform that leverages the power of multi-agent systems. The virtual platform applies large model agents to simulate multiple students and explore the cultivation of their academic abilities. Yue et al. [119] integrated domain knowledge in the teaching process into the classroom simulation process. Utilizing well-crafted role simulators, the exploration of classroom teaching processes is meticulously conducted through the orchestration of meaningful role interactions.

Exploring objective laws is an important goal for the development of AI [120]. To this end, PINNs [11] have been proposed that utilize physical laws to improve the model's predictive accuracy and generalization ability for the physical world. Compared with traditional neural networks, PINNs can achieve predictive outcomes that adhere to physical laws using a more modest amount of training data. Moreover, they exhibit enhanced resilience to noise and other interference. PINNs have been widely applied in many fields of physical research, such as fluid mechanics and heat conduction research [12]. In the study of heat conduction, they can help analyze physical-world objective phenomena such as heat diffusion [121]. Although research on PINNs has made great progress, problems such as slow training and difficulty in convergence still remain. Furthermore, PINNs perform poorly when processing high-dimensional data and solving high-dimensional equations.

6.3. Problem-solving

The application of AI has undergone extensive development for problem-solving in the domains of science, engineering, and society [122]. When symbolicism prevailed in the development of AI, many studies [123] designed various logical automatic reasoners using first-order logic and higher-order logic for scientific research, such as automatic mathematical provers [124] and automatic physical reasoners [125]. However, the amount of knowledge (i.e., logical rules) stored in these manually designed reasoners is often limited, and they may perform unsatisfactorily on more complex science problems. With the rise of deep learning, researchers [123] turned their attention to large-scale neural networks with greater knowledge retention and utilization capabilities. Such studies are roughly divided into two categories based on the function of neural networks. ① One way is to design deep learning models as retrievers [122]. The deep learning models are responsible for retrieving the knowledge needed for each reasoning step from knowledge databases, thus assisting in the step-by-step solution of a science problem. ② Another way is to regard deep learning models as pure memorizers [126]. During the training process, deep learning models fully memorize knowledge. In the subsequent inference process, the deep learning models directly produce the sufficient and complete solution without needing to retrieve knowledge databases.

LLMs designed for real-world problem-solving, such as the mathematical model DeepSeek_prover_v1.5 [126], take the ability to induce and store domain knowledge of neural networks to the extreme. The collaboration of multiple agents has also been gradually applied to the field of science research with the development of large models. For example, similar to the separation of computation and verification in the real world, separate mathematical problem-solving agents and mathematical conclusion verification agents are set up [127]. The effective collaboration of these two types of agents has achieved more accurate solutions to mathematical problems.

6.4. Interpretability

In AI-driven research, aside from reaching conclusions, explaining the reasoning process is an important issue. To this end, a meta-programming framework, MetaGPT [128], has been proposed, which integrates SOPs into the workflows of multi-agent systems. This framework is designed to enhance task decomposition and coordination, which are critical for managing complexity in software engineering projects. By encoding SOPs into prompt sequences, MetaGPT allows agents to operate with human-like domain expertise, verifying intermediate results and reducing errors. By mimicking the behavior of human experts, this approach of integrating SOPs increases the interpretability of model operations. Improving a model's capability for human–computer interaction is another way to improve interpretability. Building on this concept, Qian et al. [129] introduced as a framework for software development driven by multiple LLM-based agents. These agents collaborate through natural language and programming language exchanges, guided by chat chains and a dehallucination mechanism to improve software completeness, executability, and consistency.

Supporting hypotheses is another important goal in realizing model interpretability. Fang et al. [130] presented KANO, a KG-enhanced molecular contrastive learning method that integrates chemical domain knowledge to provide interpretable molecular representations and superior prediction performance. KANO generates functional prompts that evoke downstream task-related knowledge, thus enhancing the interpretability of the model's predictions. Li et al. [131] introduced modSAR, an optimization-based quantitative structure–activity relationship (QSAR) modeling technique that offers transparent and explainable predictions by pinpointing key breakpoint features and crafting piecewise linear regression equations. The model's ability to generate clear rules and assign Shapley additive explanations (SHAP) values to molecular fragments enhances its justification of its predictions, making it a valuable tool for drug discovery.

7. Future directions and emerging applications

7.1. Future lines of research

Beyond the topics covered above, several important and relevant areas warrant further exploration.

(1) **Embodied AI.** Embodied AI is a promising post-LLM direction. Collecting high-quality robotic datasets is labor intensive, and over-reliance on simulation data intensifies the sim-to-real gap, requiring collaborative dataset creation and improved simulators. Efficiently integrating human demonstration data and advancing cognition in complex environments are also critical in building adaptive models. Additionally, enabling causal reasoning, continual learning, and unified evaluation benchmarks will be essential for robust, scalable, and generalizable embodied AI systems.

(2) **Brain-like AI.** AI systems and algorithms inspired by the structure and functions of the human brain seek to emulate the brain's parallel processing, adaptability, and efficiency to enhance computational models. Interdisciplinary integration with neuroscience could yield AI models that closely mirror human cognitive functions by adopting insights into brain-based learning, memory, and decision-making processes. Advances in neuroscience could also inspire robust brain-like AI models capable of naturalistic emotional and contextual responses, enhancing the potential for human–computer empathy and adaptability. Moreover, significant opportunities lie in developing scalable, efficient, and responsible AI frameworks that can operate reliably in real-world applications, especially in resource-constrained or sensitive domains. By integrating insights from the structure and adaptability of neural circuits, researchers can enhance the resilience, efficiency, and transparency of brain-like AI models, ultimately moving closer to designing AI that is responsible, adaptable, and responsive to complex human-centered needs.

(3) **Non-transformer foundation models.** Despite the prominence of transformer architectures in large foundational models, several alternative architectures show promise as potential replacements. Hyena [132] introduces an efficient structure by integrating data-controlled gating with implicitly parametrized long convolutions, providing a subquadratic solution to large-scale sequence processing. Other models leverage state space models (SSMs) [133] to achieve linear scaling and improved efficiency over traditional transformers. RetNet [134], which replaces multi-head attention with a multi-scale retention mechanism, captures sequence information effectively while reducing memory usage and significantly accelerating training. Thus, these models can be seen as viable and efficient transformer alternatives.

(4) **LLM-involved model generation.** Leveraging LLMs to generate small, task-specific models by summarizing user requirements and a few in-domain data into latent variables, which are then decoded to produce tailored AI models directly usable for prediction [135], can be a promising post-LLM direction.

7.2. Emerging applications

In the post-LLM landscape, the next generation of AI, characterized by knowledge empowerment, model collaboration, and co-evolution, will definitely redefine the capabilities of AI and reshape our perceptions of these new AI systems. Its continually evolving nature will bring new possibilities to our real-world society, meeting the highly complex demands of more specialized, adaptative, and human-aligned applications.

The characteristic of knowledge empowerment suggests that the post-LLM AI systems will increasingly emphasize the fusion of more specialized, factual, and structured information, significantly enhancing their expertise in specific fields with precision and logical reasoning and eventually surpassing today's general-purpose AI models. In particular, with the integration of rich knowledge sources accumulated from science, engineering, and human society, next-generation AI is expected to delve deeper into exploring scientific laws, generating new hypotheses for scientific research and discoveries, and predicting the trajectory of events. For example, in the field of mathematics, the integration of AI will become more widespread, with large-scale neural networks being utilized to store mathematical knowledge and to conduct reasoning, and with the accuracy of problem-solving being improved through multi-agent collaboration. This will benefit other emerging AI interdisciplinary fields as well, such as online education, physics, and more. For example, the personalized application of AI will appear in the field of education, enriching teaching interactions and experiences by simulating and integrating interactive insights between students and teachers. In the realm of physics, technologies such as PINNs will leverage the laws of physics to enhance models' predictive accuracy and generalization capabilities.

Model collaboration in the post-LLM era will involve deeper collaboration among both heterogeneous data and heterogeneous models. By fusing data from multiple sources (e.g., text, images, audio, and sensory signals), omni-modal AI systems will gain a more holistic understanding of the physical world, which will be particularly useful in fields such as autonomous vehicles, cross-media content generation, and digital twins. Collaboration between large (general-purpose) and small (specialized) models is another emerging trend in collaborative AI. Large models exhibit strong capabilities in generation, reasoning, and knowledge integration, while SMs display merits such as efficiency, low latency, security, and privacy. Achieving deeper collaboration between large and small models is a future development trend, involving not only effective data exchange but also knowledge sharing and task decomposition to address complex task scenarios, particularly in areas such as embodied intelligence. As application scenarios expand, personalized and adaptive collaborative systems incorporating large and small models will become a significant development direction in areas including intelligent assistants and service robots.

Model co-evolution, inspired by biological ecosystems, is another key ingredient in establishing the next generation of AI, in which AI models evolve collectively, learning and adapting in an interdependent process. This dynamic and continually evolving relationship between collective models is expected to remarkably advance the intelligence level and adaptation capability of AI systems, enhancing their robustness to dynamic and unknown physical-world changes. Merging diverse functional models may be a potential approach to co-evolving AI systems, as this can synthesize information from multiple models into a cohesive, unified framework. Nevertheless, several crucial challenges still remain to be resolved, including the lack of a deeper theoretical understanding of the merging mechanism—especially in scenarios involving models trained on different datasets or for different tasks—and the high computational and memory costs of merging schemes, among other issues. Co-evolved AI models have broad potential applications, such as autonomous driving, mining robotics, and industrial manufacturing.

Knowledge-empowered, collaborative, and co-evolving AI will likely bring AI systems to a new level with higher intelligence, resilience, and autonomy, expanding AI's capability to handle complex real-world applications such as scientific discoveries, engineering design, personalized education, manufacturing, and more. Meanwhile, the increasing autonomy and interconnectivity of AI models may also present challenges in terms of safety and societal impacts, requiring mechanisms to monitor and control AI systems in order to prevent unforeseen behaviors.

8. Conclusions

In this survey, we explored the evolving landscape of AI beyond LLMs, with a focus on the paradigms of knowledge-empowered AI, model collaboration, and the co-evolution of AI systems. While LLMs have significantly advanced AI capabilities, they present inherent challenges in scalability and adaptability. To address these limitations, we discussed a range of post-LLM techniques and applications aimed at building more robust, scalable, and adaptable AI models. This survey sheds light on potential roadmap points in the post-LLM era for researchers and practitioners. The ongoing evolution of AI requires continued innovation and interdisciplinary collaboration to build systems that are not only powerful but also adaptable and aligned with human values.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (62441605).

Compliance with ethics guidelines

Fei Wu, Tao Shen, Thomas Bäck, Jingyuan Chen, Gang Huang, Yaochu Jin, Kun Kuang, Mengze Li, Cewu Lu, Jiaxu Miao, Yongwei Wang, Ying Wei, Fan Wu, Junchi Yan, Hongxia Yang, Yi Yang, Shengyu Zhang, Zhou Zhao, Yueting Zhuang, and Yunhe Pan declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Pan Y. 2018 special issue on artificial intelligence 2.0: theories and applications. Front Inform Technol Electron Eng 2018;19(1):1–2.
[2] Lyu YG. Artificial intelligence: enabling technology to empower society. Engineering 2020;6(3):205–6.
[3] Lyu YG, Wu F. Toward a more general empowering artificial intelligence. Engineering 2023;25:1–2.
[4] Lyu YG, Wu F. Further empowering humans in specific fields and rethinking AGI testing. Engineering 2024;34:1–2.
[5] Li DF, Xu F. Synergizing knowledge graphs with large language models: a comprehensive review and future prospects. 2024. arXiv:2407.18470.
[6] Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. IEEE Trans Knowl Data Eng 2024;36(7):3580–99.
[7] Kau A, He X, Nambissan A, Astudillo A, Yin H, Aryani A. Combining knowledge graphs and large language models. 2024. arXiv:2407.06564.
[8] Yuan B, Chen Y, Zhang Y, Jiang W. Hide and seek in noise labels: noise-robust collaborative active learning with LLMs-powered assistance. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 10977–1011.
[9] Hao Z, Jiang H, Jiang S, Ren J, Cao T. Hybrid SLM and LLM for edge-cloud collaborative inference. In: Proceedings of the Workshop on Edge and Mobile Foundation Models; 2024 Jun 3–7; Tokyo, Japan. New York City: Association for Computing Machinery (ACM); 2024. p. 36–41.
[10] Zhang K, Wang J, Ding N, Qi B, Hua E, Lv X, et al. Fast and slow generating: an empirical study on large and small language models collaborative decoding. 2024. arXiv:2406.12295.
[11] McClenny LD, Braga-Neto UM. Self-adaptive physics-informed neural networks. J Comput Phys 2023;474:111722.
[12] Sharma P, Chung WT, Akoush B, Ihme M. A review of physics-informed machine learning in fluid mechanics. Energies 2023;16(5):2343.
[13] Zhou C, Liu P, Xu P, Iyer S, Sun J, Mao Y, et al. LIMA: less is more for alignment. In: Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023); 2023 Dec 10–16; New Orleans, LA, USA. Trier: NeurIPS Proceedings; 2024.
[14] Akyürek E, Bolukbasi T, Liu F, Xiong B, Tenney I, Andreas J, et al. Towards tracing factual knowledge in language models back to the training data. 2022. arXiv:2205.11482.
[15] Shen T, Mao Y, He P, Long G, Trischler A, Chen W. Exploiting structured knowledge in text via graph-guided representation learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; online. Stroudsburg: Association for Computational Linguistics (ACL); 2020. p. 8980–94.
[16] Zhang D, Yuan Z, Liu Y, Zhuang F, Chen H, Xiong H. E-BERT: a phrase and product knowledge enhanced language model for E-commerce. 2020. arXiv:2009.02835.
[17] Tian H, Gao C, Xiao X, Liu H, He B, Wu H, et al. SKEP: sentiment knowledge enhanced pre-training for sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; online. Stroudsburg: Association for Computational Linguistics (ACL); 2020. p. 4067–76.
[18] Gao T. Knowledge authoring and question answering with KALM. 2019. arXiv:1905.00840.
[19] Wang X, Gao T, Zhu Z, Zhang Z, Liu Z, Li J, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation. Trans Assoc Comput Linguist 2021;9:176–94.
[20] Li S, Li X, Shang L, Sun C, Liu B, Ji Z, et al. Pre-training language models with deterministic factual knowledge. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 7–11; Abu Dhabi, UAE. Stroudsburg: Association for Computational Linguistics (ACL); 2022. p. 11118–31.
[21] Xiong W, Du J, Wang WY, Stoyanov V. Pretrained encyclopedia: weakly supervised knowledge-pretrained language model. 2019. arXiv:1912.09637.
[22] Ji J, Wang K, Qiu T, Chen B, Zhou J, Li C, et al. Language models resist alignment. 2024. arXiv:2406.06144.
[23] Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, et al. Instruction tuning for large language models: a survey. 2023. arXiv:2308.10792.
[24] Gekhman Z, Yona G, Aharoni R, Eyal M, Feder A, Reichart R, et al. Does fine-tuning LLMs on new knowledge encourage hallucinations? 2024. arXiv:2405.05904.
[25] Wang J, Huang W, Qiu M, Shi Q, Wang H, Li X, et al. Knowledge prompting in pre-trained language model for natural language understanding. 2022. arXiv:2210.08536.
[26] Ye H, Zhang N, Deng S, Chen X, Xiong F, Chen X, et al. Ontology-enhanced prompt-tuning for few-shot learning. In: Proceedings of the ACM Web Conference 2022; 2022 Apr 25–29; online. New York City: Association for Computing Machinery (ACM); 2022. p. 778–87.
[27] Luo H, Tang Z, Peng S, Guo Y, Zhang W, Ma C, et al. ChatKBQA: a generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. 2023. arXiv:2310.08975.
[28] Luo L, Li YF, Haffari G, Pan S. Reasoning on graphs: faithful and interpretable large language model reasoning. 2023. arxiv:2310.01061.
[29] Ovadia O, Brief M, Mishaeli M, Elisha O. Fine-tuning or retrieval? Comparing knowledge injection in LLMs. 2023. arXiv:2312.05934.
[30] Yang D, Rao J, Chen K, Guo X, Zhang Y, Yang J, et al. IM-RAG: multi-round retrieval-augmented generation through learning inner monologues. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024 Jul 14–18; Washington, DC, USA. New York City: Association for Computing Machinery (ACM); 2024. p. 730–40.
[31] Mussmann S, Ermon S. Learning and inference via maximum inner product search. In: Proceedings of the International Conference on Machine Learning; 2016 Jun 20–22; New York City, NY, USA. Birmingham: Proceedings of Machine Learning Research; 2016. p. 2587–96.
[32] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. New York City: Association for Computing Machinery (ACM); 2020. p. 9459–74.
[33] Wu Y, Zhao Y, Hu B, Minervini P, Stenetorp P, Riedel S. An efficient memory-augmented transformer for knowledge-intensive NLP tasks. 2022. arXiv:2210.16773.
[34] Guu K, Lee K, Tung Z, Pasupat P, Chang MW. REALM: retrieval augmented language model pre-training. In: Proceedings of the International Conference on Machine Learning; 2020 Jul 13–18; online. Birmingham: Proceedings of Machine Learning Research; 2020. p. 3929–38.
[35] Logan R, Liu NF, Peters ME, Gardner M, Singh S. Barack's wife Hillary: using knowledge graphs for fact–aware language modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. Stroudsburg: Association for Computational Linguistics (ACL); 2019. p. 5962–71.
[36] Zhang Y, Li H, Zhang S, Wang R, He B, Dou H, et al. LLMCO4MR: LLMs-aided neural combinatorial optimization for ancient manuscript restoration from fragments with case studies on Dunhuang. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision—ECCV 2024. Cham: Springer; 2024.
[37] Sun Y, Wang S, Feng S, Ding S, Pang C, Shang J, et al. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. 2021. arXiv:2107.02137.
[38] Sun T, Shao Y, Qiu X, Guo Q, Hu Y, Huang X, et al. CoLAKE: contextualized language and knowledge embedding. In: Proceedings of the 28th International Conference on Computational Linguistics; 2023 Dec 8–13; Barcelona, Spain. Stroudsburg: Association for Computational Linguistics (ACL); 2020. p. 3660–70.
[39] Zhang T, Wang C, Hu N, Qiu M, Tang C, He X, et al. DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding. Proc Conf AAAI Artif Intell 2022;36(10):11703–11.
[40] Yu W, Zhu C, Fang Y, Yu D, Wang S, Xu Y, et al. Dict-BERT: enhancing language model pre-training with dictionary. 2021. arXiv:2110.06490.
[41] Li S, Gao Y, Jiang H, Yin Q, Li Z, Yan X, et al. Graph reasoning for question answering with triplet retrieval. 2023. arXiv:2305.18742.
[42] Luo L, Ju J, Xiong B, Li YF, Haffari G, Pan S. ChatRule: mining logical rules with large language models for knowledge graph reasoning. 2023. arXiv:2309.01538.
[43] Wang J, Sun Q, Chen N, Li X, Gao M. Boosting language models reasoning with chain-of-knowledge prompting. 2023. arXiv:2306.06427.
[44] Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. 2017. arXiv:1701.06538.
[45] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2022.
[46] Kraaijenbrink J, Wijnhoven F. Managing heterogeneous knowledge: a theory of external knowledge integration. Knowl Manag Res Pract 2008;6(4):274–86.
[47] Dogan A, Birant D. A weighted majority voting ensemble approach for classification. In: Proceedings of the 2019 4th International Conference on Computer Science and Engineering (UBMK); 2019 4th International Conference on Computer Science and Engineering (UBMK 2019); 2019 Sep 11–15; Samsun, Türkiye. New York City: IEEE; 2019. p. 1–6.

[48] Kwon H, Park J, Lee Y. Stacking ensemble technique for classifying breast cancer. Healthc Inform Res 2019;25(4):283–8.
[49] Du N, Huang Y, Dai AM, Tong S, Lepikhin D, Xu Y, et al. GLaM: efficient scaling of language models with mixture-of-experts. In: Proceedings of the 39th International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. Birmingham: Proceedings of Machine Learning Research; 2022. p. 5547–69.
[50] Wang K, Xu Y, Wu Z, Luo S. LLM as prompter: low-resource inductive reasoning on arbitrary knowledge graphs. 2024. arXiv:2402.11804.
[51] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. 2023. arXiv:2307.09288.
[52] Nayak A, Timmapathini HP. LLM2KB: constructing knowledge bases using instruction tuned context aware large language models. 2023. arXiv:2308.13207.
[53] Wang H, Li R, Jiang H, Tian J, Wang Z, Luo C, et al. BlendFilter: advancing retrieval-augmented large language models via query generation blending and knowledge filtering. 2024. arXiv:2402.11129.
[54] Parisi A, Zhao Y, Fiedel N. TALM: tool augmented language models. 2022. arXiv:2205.12255.
[55] Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Hambro E, et al. Toolformer: language models can teach themselves to use tools. In: Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems; 2023 Dec 10; New Orleans, LU, USA. New York City: Association for Computing Machinery (ACM); 2023.
[56] Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face. 2023. arXiv:2303.17580v4.
[57] Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. In: Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems; 2023 Dec 10; New Orleans, LU, USA. New York City: Association for Computing Machinery (ACM); 2023.
[58] Besta M, Blach N, Kubicek A, Gerstenberger R, Podstawski M, Gianinazzi L, et al. Graph of thoughts: solving elaborate problems with large language models. 2024. arXiv:2308.09687v4.
[59] Rabby G, Auer S, D'Souza J, Oelen A. Fine-tuning and prompt engineering with cognitive knowledge graphs for scholarly knowledge organization. 2024. arXiv:2409.06433.
[60] Ein-Dor L, Toledo-Ronen O, Spector A, Greta S, Dankin L, Halfon A, et al. Conversational prompt engineering. 2024. arXiv:2408.04560.
[61] Yu Z, Ouyang X, Shao Z, Wang M, Yu J. Prophet: prompting large language models with complementary answer heuristics for knowledge-based visual question answering. 2023. arXiv:2303.01903.
[62] Lu X, Liao Y, Liu C, Lio P, Hui P. Heterogeneous model fusion federated learning mechanism based on model mapping. IEEE Internet Things J 2022;9(8):6058–68.
[63] Wu TH, Lian L, Gonzalez JE, Li B, Darrell T. Self-correcting LLM-controlled diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. New York City: IEEE; 2024. p. 6327–36.
[64] Wang Y, Zhu S, Fu F, Miao X, Zhang J, Zhu J, et al. Efficient multi-task large model training via data heterogeneity-aware model management. 2024. arXiv:2409.03365.
[65] Sachin DN, Annappa B, Hegde S, Abhijit CS, Ambesange S. FedCure: a heterogeneity-aware personalized federated learning framework for intelligent healthcare applications in IoMT environments. IEEE Access 2024;12:15867–83.
[66] Haller M, Lenz C, Nachtigall R, Awayshehl FM, Alawadi S. Handling non-IID data in federated learning: an experimental evaluation towards unified metrics. In: Proceedings of the 2023 IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC); 2023 Nov 14–17; Abu Dhabi, UAE. New York City: IEEE; 2023. p. 0762–70.
[67] Ding K, Dong X, He Y, Cheng L, Fu C, Huan Z, et al. MSSM: a multiple-level sparse sharing model for efficient multi-task learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021 Jul 11–15; online. New York City: Association for Computing Machinery (ACM); 2021. p. 2237–41.
[68] Wang Z, Panda R, Karlinsky L, Feris R, Sun H, Kim Y. Multitask prompt tuning enables parameter-efficient transfer learning. 2023. arXiv:2303.02861.
[69] Zhang W, Zhai G, Wei Y, Yang X, Ma K. Blind image quality assessment via vision-language correspondence: a multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 18–22; Vancouver, BC, Canada. New York City: IEEE; 2023. p. 14071–81.
[70] Chen Q, Chen X, Wang J, Zhang S, Yao K, Feng H, et al. Group DETR: fast DETR training with group-wise one-to-many assignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2–6; Paris, France. New York City: IEEE; 2023. p. 6633–42.
[71] Ghosh A, Chung J, Yin D, Ramchandran K. An efficient framework for clustered federated learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–10; Vancouver, BC, Canada. Red Hook: Curran Associates Inc.; 2020.
[72] Ye R, Wang W, Chai J, Li D, Li Z, Xu Y, et al. OpenFedLLM: training large language models on decentralized private data via federated learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2024 Aug 25–29; Barcelona, Spain. New York City: Association for Computing Machinery (ACM); 2024. p. 6137–47.
[73] Yang C, An Z, Zhou H, Zhuang F, Xu Y, Zhang Q. Online knowledge distillation via mutual contrastive learning for visual recognition. IEEE Trans Pattern Anal Mach Intell 2023;45(8):10212–27.
[74] Ni J, Tang H, Shang Y, Duan B, Yan Y. Adaptive cross-architecture mutual knowledge distillation. In: Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG); 2024 May 27–31; Istanbul, Türkiye. New York City: IEEE; 2024. p. 1–5.
[75] Miao Z, Zhang W, Su J, Li X, Luan J, Chen Y, et al. Exploring all-in-one knowledge distillation framework for neural machine translation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. Stroudsburg: Association for Computational Linguistics (ACL); 2023. p. 2929–40.
[76] Zhao J, Zhao W, Drozdov A, Rozonoyer B, Sultan MA, Lee JY, et al. Multistage collaborative knowledge distillation from a large language model for semi-supervised sequence generation. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 14201–14.
[77] Starodubcev N, Fedorov A, Babenko A, Baranchuk D. Your student is better than expected: adaptive teacher–student collaboration for text-conditional diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. New York City: IEEE; 2024. p. 9275–85.
[78] Shao J, Wu F, Zhang J. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. Nat Commun 2024;15:349.
[79] Wan F, Huang X, Cai D, Quan X, Bi W, Shi S. Knowledge fusion of large language models. 2024. arXiv:2401.10491.
[80] Wang Y, Agarwal S, Mukherjee S, Liu X, Gao J, Awadallah AH, et al. AdaMix: mixture-of-adaptations for parameter-efficient model tuning. 2022. arXiv:2205.12410.
[81] Wortsman M, Ilharco G, Gadre SY, Roelofs R, Gontijo-Lopes R, Morcos AS, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: Proceedings of the International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. Seattle: Proceedings of Machine Learning Research; 2022. p. 23965–98.
[82] Arpit D, Wang H, Zhou Y, Xiong C. Ensemble of averages: improving model selection and boosting performance in domain generalization. 2022. arXiv:2110.10832.
[83] Jin X, Ren X, Preotiuc-Pietro D, Cheng P. Dataless knowledge fusion by merging weights of language models. 2022. arXiv:2212.09849.
[84] Perin G, Chen X, Liu S, Kailkhura B, Wang Z, Gallagher B. RankMean: module-level importance score for merging fine-tuned LLM models. In: Proceedings of the Findings of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 1776–82.
[85] Yu L, Bi K, Ni S, Guo J. Contextual dual learning algorithm with listwise distillation for unbiased learning to rank. 2024. arXiv:2408.09817.
[86] Park S, Van Hentenryck P. Self-supervised primal-dual learning for constrained optimization. Proc Conf AAAI Artif Intell 2023;37(4):4052–60.
[87] Fei H, Wu S, Ren Y, Zhang M. Matching structure for dual learning. In: Proceedings of the International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. Seattle: Proceedings of Machine Learning Research; 2022. p. 6373–91.
[88] Ji W, Wang R, Tian Y, Wang X. An attention based dual learning approach for video captioning. Appl Soft Comput 2022;117:108332.
[89] Li J, Xia Y, Yan R, Sun H, Zhao D, Liu T, et al. Stylized dialogue generation with multi-pass dual learning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Sep 6–14; online. Red Hook: Curran Associates Inc.; 2021.
[90] Chen A, Lou L, Chen K, Bai X, Xiang Y, Yang M, et al. DUAL-REFLECT: enhancing large language models for reflective translation through dual learning feedback mechanisms. 2024. arXiv:2406.07232.
[91] Dong J, Zhang M, Zhang Z, Chen X, Liu D, Qu X, et al. Dual learning with dynamic knowledge distillation for partially relevant video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 4–6; Paris, France. New York City: IEEE; 2023. p. 11302–12.
[92] Wang Y, Sun T, Li S, Yuan X, Ni W, Hossain E, et al. Adversarial attacks and defenses in machine learning-empowered communication systems and networks: a contemporary survey. IEEE Comm Surv and Tutor 2023;25(4):2245–98.
[93] Cheng P, Yang Y, Li J, Dai Y, Hu T, Cao P, et al. Adversarial preference optimization: enhancing your alignment via RM-LLM game. In: Proceedings of the Findings of the Association for Computational Linguistics; 2024 Aug 11–16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 3705–16.
[94] Tan K, Luo K, Lan Y, Yuan Z, Shu J. An LLM-enhanced adversarial editing system for lexical simplification. 2024. arXiv:2402.14704.
[95] Sheshadri A, Ewart A, Guo P, Lynch A, Wu C, Hebbar V, et al. Targeted latent adversarial training improves robustness to persistent harmful behaviors in LLMs. 2024. arXiv:2407.15549.
[96] Hu X, Chen PY, Ho TY. RADAR: robust AI-text detection via adversarial learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2023. p. 15077–95.
[97] Thota S, Vangoor VKR, Reddy AK, Ravi CS. Federated learning: privacy-preserving collaborative machine learning. DLBSAR 2019;5:168–90.

[98] Goddard C, Siriwardhana S, Ehghaghi M, Meyers L, Karpukhin V, Benedict B, et al. Arcee's mergekit: a toolkit for merging large language models. 2024. arXiv:2403.13257.
[99] Yang E, Wang Z, Shen L, Liu S, Guo G, Wang X, et al. AdaMerging: adaptive model merging for multi-task learning. 2024. arXiv:2310.02575.
[100] Matena M, Raffel C. Merging models with fisher-weighted averaging. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2022. p. 17703–16.
[101] Yadav P, Tam D, Choshen L, Raffel C, Bansal M. TIES-MERGING: resolving interference when merging models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2023.
[102] Yu L, Yu B, Yu H, Huang F, Li Y. Language models are super Mario: absorbing abilities from homologous models as a free lunch. 2023. arXiv:2311.03099.
[103] Lu Z, Fan C, Wei W, Qu X, Chen D, Cheng Y. Twin-merging: dynamic integration of modular expertise in model merging. 2024. arXiv:2406.15479.
[104] Tang A, Shen L, Luo Y, Yin N, Zhang L, Tao D. Merging multi-task models via weight-ensembling mixture of experts. 2024. arXiv:2402.00433.
[105] Yang E, Shen L, Wang Z, Guo G, Chen X, Wang X, et al. Representation surgery for multi-task model merging. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria. Seattle: Proceedings of Machine Learning Research; 2024.
[106] Zhang J, Yang HF, Li A, Guo X, Wang P, Wang H, et al. MLLM-FL: multimodal large language model assisted federated learning on heterogeneous and long-tailed data. 2024. arXiv:2409.06067.
[107] Bai J, Chen D, Qian B, Yao L, Li J. Federated fine-tuning of large language models under heterogeneous language tasks and client resources. 2024. arXiv:2402.11505.
[108] Fan T, Ma G, Kang Y, Gu H, Song Y, Fan L, et al. FedMKT: federated mutual knowledge transfer for large and small language models. 2024. arXiv:2406.02224.
[109] Li H, Zhao X, Guo D, Gu H, Zeng Z, Han Y, et al. Federated domain-specific knowledge transfer on large language models using synthetic data. 2024. arXiv:2405.14212.
[110] Fan T, Kang Y, Chen W, Gu H, Song Y, Fan L, et al. PDSS: a privacy-preserving framework for step-by-step distillation of large language models. 2024. arXiv:2406.12403.
[111] Gholami M, Akbari M, Hu C, Masrani V, Wang J, Zhang Y. GOLD: generalized knowledge distillation via out-of-distribution-guided language data generation. 2024. arXiv:2403.19754.
[112] Li X, Fang Y, Liu M, Ling Z, Tu Z, Su H. Distilling large vision-language model with out-of-distribution generalizability. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 4–6; Paris, France. New York City: IEEE; 2023. p. 2492–503.
[113] Agarwal R, Vieillard N, Zhou Y, Stanczyk P, Ramos S, Geist M, et al. On-policy distillation of language models: learning from self-generated mistakes. In: Proceedings of the Twelfth International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria. London: ICLR; 2024.
[114] Chen Z, Wang W, Zhao Z, Su F, Men A, Meng H. PracticalDG: perturbation distillation on vision-language models for hybrid domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jul 16–22; Seattle, WA, USA. New York City: IEEE; 2024. p. 23501–11.
[115] Feng S, Sun H, Yan X, Zhu H, Zou Z, Shen S, et al. Dense reinforcement learning for safety validation of autonomous vehicles. Nature 2023;615(7953):620–7.
[116] Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q. Accurate medium-range global weather forecasting with 3D neural networks. Nature 2023;619(7970):533–8.
[117] Chen K, Han T, Gong J, Bai L, Ling F, Luo JJ, et al. FengWu: pushing the skillful global medium-range weather forecast beyond 10 days lead. 2023. arXiv:2304.02948.
[118] Zhong X, Chen L, Liu J, Lin C, Qi Y, Li H. FuXi-extreme: improving extreme rainfall and wind forecasts with diffusion model. 2023. arXiv:2310.19822.
[119] Yue M, Mifdal W, Zhang Y, Suh J, Yao Z. MathVC: an LLM-simulated multi-character virtual classroom for mathematics education. 2024. arXiv:2404.06711.
[120] Müller J, Zeinhofer M. Achieving high accuracy with PINNs via energy natural gradient descent. In: Proceedings of the International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. New York City: IEEE; 2023.
[121] Aymerich E, Pisano F, Cannas B, Sias G, Fanni A, Gao Y, et al. Physics informed neural networks towards the real-time calculation of heat fluxes at W7-X. Nucl Mater Energy 2023;34:101401.
[122] Yang K, Swope A, Gu A, Chalamala R, Song P, Yu S, et al. LeanDojo: theorem proving with retrieval-augmented language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2024.
[123] Zhan B. AUTO2, a saturation-based heuristic prover for higher-order logic. In: Proceedings of Interactive Theorem Proving; 2016 Aug 22–25; Nancy, France; 2016.
[124] Steen A, Sutcliffe G, Scholl T, Benzmüller C. Solving modal logic problems by translation to higher-order logic. In: Proceedings of the International Conference on Logic and Argumentation; 2023 Sep 10–12; Hangzhou, China. Cham: Springer Nature Switzerland; 2023. p. 25–43.
[125] Foulis DJ, Randall CH. The empirical logic approach to the physical sciences. In: Hartkämper A, Neumann H, editors. Foundations of quantum mechanics and ordered linear spaces. Marburg: Advanced Study Institute; 1973. p. 230–49.
[126] Xin H, Ren ZZ, Song J, Shao Z, Zhao W, Wang H, et al. DeepSeek-prover-V1. 5: harnessing proof assistant feedback for reinforcement learning and Monte-Carlo tree search. 2024. arXiv:2408.08152.
[127] Zhou JP, Staats C, Li W, Szegedy C, Weinberger KQ, Wu Y. Don't trust: verify-grounding LLM quantitative reasoning with autoformalization. 2024. arXiv:2403.18120.
[128] Hong S, Zheng X, Chen J, Cheng Y, Wang J, Zhang C, et al. MetaGPT: meta programming for multi-agent collaborative framework. 2023. arXiv:2308.00352.
[129] Qian C, Liu W, Liu H, Chen N, Dang Y, Li J, et al. ChatDev: communicative agents for software development. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistic; 2024 Aug 11–16; Bangkok, Thailand. Stroudsburg: Association for Computational Linguistics (ACL); 2024. p. 15174–86.
[130] Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nat Mach Intell 2023;5(5):542–53.
[131] Li Y, Cardoso-Silva J, Kelly JM, Delves MJ, Furnham N, Papageorgiou LG, et al. Optimisation-based modelling for explainable lead discovery in malaria. Artif Intell Med 2024;147:102700.
[132] Poli M, Massaroli S, Nguyen E, Fu DY, Dao T, Baccus S, et al. Hyena hierarchy: towards larger convolutional language models. In: Proceedings of the International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. Seattle: Proceedings of Machine Learning Research; 2023. p. 28043–78.
[133] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. 2023. arXiv:2312.00752.
[134] Sun Y, Dong L, Huang S, Ma S, Xia Y, Xue J, et al. Retentive network: a successor to transformer for large language models. 2023. arXiv:2307.08621.
[135] Tang Z, Lv Z, Zhang S, Wu F, Kuang K. ModelGPT: unleashing LLM's capabilities for tailored model generation. 2024. arXiv:2402.12408.