Towards Scalable Compound AI Systems: A Vision Paper

Richard Ding *
ChainOpera, Inc.
TensorOpera, Inc.
richard@chainopera.com

Yuhang Yao *
TensorOpera, Inc.
yuhang@tensoropera.com

Dimitris Stripelis
TensorOpera, Inc.
dimitris@tensoropera.com

Min Chang Jordan Ren†
ChainOpera, Inc.
jordan9ren8@gmail.com

Qile Wu
Nanyang Technological University
ChainOpera, Inc.
qile001@e.ntu.edu.sg

Salman Avestimehr
University of South California
TensorOpera, Inc.
ChainOpera, Inc.
avestimehr@tensoropera.com

Chaoyang He
ChainOpera, Inc.
TensorOpera, Inc.
ch@tensoropera.com

Abstract

Compound AI systems represent a promising strategy that combines various AI components to enhance both the quality and reliability of AI-driven tasks, meeting the complex demands of today's applications. In this paper, we examine current trends, methodologies, and technical hurdles linked to these systems, underscoring their role in achieving cutting-edge performance, increasing scalability, and fostering innovation in next-generation AI agent systems. We particularly concentrate on the inference process, which we define as the coordinated operation of multiple AI components to generate dependable, task-specific results. Moreover, compound AI systems present unique challenges—such as balancing latency, accuracy, and overall system coherence—that call for novel frameworks and approaches in design, optimization, and monitoring. We also outline prospective pathways for evolving infrastructure to support the growing complexity of AI workflows and multi-agent systems. By analyzing practical case studies, we demonstrate how this approach is reshaping fields like healthcare, finance, customer service, and autonomous systems, among others. We contend that refining compound AI architectures will not only boost the performance of existing AI models but will also lay the groundwork for the next generation of versatile, high-performance AI applications.

1 Introduction

In recent years, significant strides in large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Team et al., 2023) have fundamentally reshaped artificial intelligence by enabling these models to handle a broad spectrum of tasks through straightforward prompting (Ding et al., 2023). Early investigations and application developments were largely dedicated to exploring the abilities of individual models—enhancing and refining LLMs by scaling their architectures and training on ever-expanding datasets (Kaplan et al., 2020; Anil et al., 2023; Dubey et al., 2024). However, recent evidence points to a transition toward compound AI systems (Zaharia et al., 2024a), which deliberately merge various components—such as LLMs, retrieval systems (Lewis et al., 2020b), and external tools (Schick et al., 2024)—to improve both the precision and reliability of AI-driven tasks. This evolution in system design is particularly important, as it addresses the growing need for higher accuracy, flexibility, and control that surpasses what simple model scaling can achieve (Jain et al., 2024).

In this paper, we examine both the progress made and the unresolved challenges in compound AI systems, with a particular focus on the inference process. We define inference as the coordination of various AI components to yield reliable, task-specific outputs (Han et al., 2024b). In contrast to model training—where optimization primarily enhances the predictive performance of a single

*Equal Contributions
†Work done as research assistant at ChainOpera

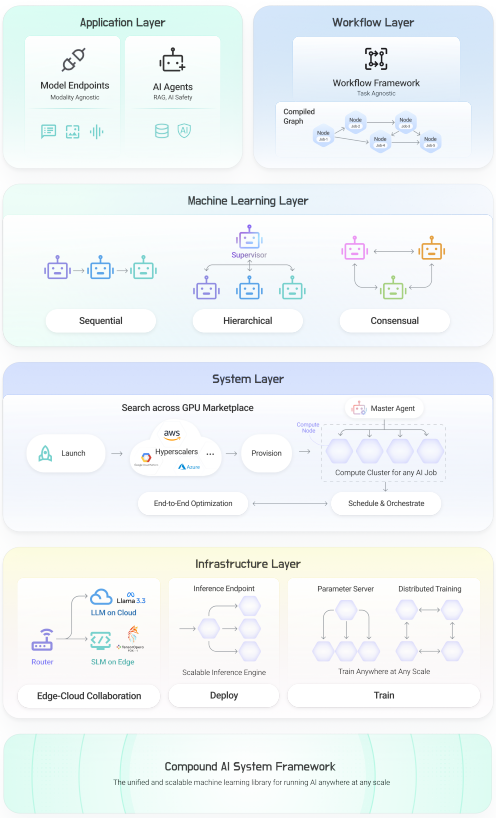

Figure 1: The five-layer view of a Compound AI System can be broken down as follows: The Application Layer delivers user-centric tools, including model endpoints and AI agents. The Workflow Layer offers task-independent orchestration frameworks. The Machine Learning Layer manages AI agents through sequential, hierarchical, and consensus-driven strategies. The System Layer oversees compute resources by leveraging GPU marketplaces and master agents for optimization. Finally, the Infrastructure Layer enables edge-cloud collaboration, scalable inference deployment, and distributed training, supporting AI operations at any scale.

model—the inference phase in compound AI systems requires real-time decision making across multiple elements, such as retrieval and verification modules (Lewis et al., 2020a). Although training these systems presents significant challenges, including aligning diverse objectives, managing interdependencies, and addressing feedback loops, our emphasis on inference underlines its critical role in ensuring the system's operational effectiveness (Santhanam et al., 2024). Strategies like retrieval-augmented generation (RAG), tool-based reasoning, and iterative chain-of-thought sampling highlight the necessity for an adaptive and modular design that boosts accuracy, control, and efficiency (Snell et al., 2024). Furthermore, the inference process in compound AI introduces unique hurdles—such as optimizing latency, accuracy, and overall system coherence—that demand innovative frameworks and methodologies for design, optimization, and monitoring (Davis et al., 2024b). As depicted in Figure 1, the compound AI system is structured into five layers, each addressing specific functionalities and integration requirements to support efficient, modular, and scalable AI deployments (Feng et al., 2024).

1. Application Layer: This highest layer manages user interfaces and tools, including modality-neutral model endpoints that accommodate a range of AI tasks. It also incorporates AI agents designed for functions such as Retrieval-Augmented Generation (RAG) and AI safety, thereby facilitating robust workflows for addressing complex problems.

2. Workflow Layer: This layer establishes orchestration frameworks that are independent of specific tasks. It allows for the creation of compiled graphs for various AI pipelines, modularizing tasks and dependencies to ensure both flexibility and scalability across different applications.

3. Machine Learning Layer: At this level, supervisors are responsible for coordinating AI agents and managing workflows through three distinct operational modes:

   - Sequential: Agents function in a step-by-step, linear process.
   - Hierarchical: Control is organized in layered, task-dependent structures.
   - Consensual: Multiple agents collaborate to make decisions collectively.

4. System Layer: In this layer, AI deployments make use of GPU marketplaces (such as AWS, Azure, and Hyperscalers) to launch, provision, and optimize compute resources. A master agent oversees scheduling, orchestration, and overall optimization, ensuring the smooth operation of large-scale AI workloads.

5. Infrastructure Layer: This foundational layer supports both AI model execution and training across cloud and edge environments. It features:

- Edge-Cloud Collaboration Facilitates integration between cloud-hosted and edge-based large language models.
- Deployment Includes scalable inference engines and endpoints.
- Training Infrastructure Provides support for distributed training, parameter servers, and extensive AI training environments.

Many challenges in compound AI inferencing remain unresolved, with no clear definitions or systematic methods to tackle them (Zaharia et al., 2024b). For example, in retrieval-augmented generation (RAG) pipelines, a core dilemma is whether to devote more computational resources to boosting the retriever's precision, enhancing the generative abilities of the large language model, or increasing the frequency of interactions between the two. Similarly, when merging different types of models—such as pairing a text generation model with text-to-audio or text-to-video models for multimodal outputs—new challenges arise. Developers face decisions about how best to sequence these components; for instance, should the generated text first be semantically refined, or should it be fed directly into audio and video generation pipelines with little preprocessing? Such decisions affect both the latency and the coherence of the final output, yet there is little guidance available on how to integrate these elements effectively. Despite the rising use of compound systems in real-world applications, there is limited agreement on critical design trade-offs, as well as broader issues like control logic design and the integration of various components for overall system optimization. This situation highlights the need for a comprehensive vision paper that clearly defines the inferencing phase in compound AI systems, identifies the main challenges, and sets forth a framework to steer future research. Such a document could provide the essential concepts and methodologies necessary to advance this emerging field and foster a more structured and coherent understanding of the domain.

By investigating the trends, methods, and technical challenges of compound AI inferencing, this paper aims to underscore the growing importance of inferencing strategies in fully unlocking the potential of AI systems. We contend that the design and refinement of compound inference architectures will not only enhance the performance of existing AI models but will also lay the foundation for the next generation of adaptable, high-performance AI applications.

This paper is organized to provide a thorough examination of compound AI systems from several viewpoints, each detailed in its own Section. Section 2 introduces the core concept of Compound AI Systems, outlining its scope and significance in the realm of modern AI. Section 3 explores the infrastructure aspect, focusing on the technical and architectural foundations that enable the integration and operation of various AI components. Section 4 addresses system-level considerations, emphasizing challenges and design principles related to orchestration, modularity, and scalability. Section 5 transitions to machine learning perspectives, discussing how model-specific issues and training versus inference strategies evolve within compound systems. Section 6 delves into workflow dynamics, detailing the sequential and parallel interactions among components to achieve efficient and reliable outcomes. Section 7 shifts the focus to application-specific scenarios, demonstrating how compound AI inference is customized for domains such as medical diagnostics, programming, and multimodal content generation. Section 8 outlines future directions, highlighting unresolved research questions and emerging trends, including strategies for agent training in compound AI systems. Finally, Section 9 reviews the related literature, positioning this paper within the broader research landscape and underscoring its contributions to advancing the development and understanding of compound AI systems.

2 Compound AI Systems

A Compound AI System is a sophisticated framework that brings together multiple interacting components—such as generative language models (LLMs) (Ding et al., 2024), retrieval systems (Cuconasu et al., 2024), symbolic engines (Sprague et al., 2024), and external tools (Qin et al., 2023)—to tackle complex tasks through coordinated functionality rather than relying on a single, monolithic model (Zaharia et al., 2024a). This approach is crucial for achieving state-of-the-art performance in scenarios where standalone models are insufficient, particularly when tasks demand dynamic adaptability, precise control, or enhanced reliability (Shen et al., 2024). By supporting a modular design and tailoring specific components to distinct functions, compound AI systems boost

performance, scalability, and user trust. They represent a paradigm shift in AI development, illustrating how innovations at the system level—such as dynamic orchestration, retrieval-augmented generation, and task-specific chains—can exceed the capabilities of even the most advanced individual models. This trend is increasingly recognized as key to maximizing AI's potential across a wide range of applications, offering a flexible and efficient route for harnessing emerging technologies (Lund and Ting, 2023).

The effectiveness of *Compound AI Systems* hinges on compound AI inference—the process of orchestrating and executing the interactions among these diverse components in real time to produce high-quality outcomes. This form of inference involves not only operating the individual components (e.g., LLMs or retrieval systems) but also managing their communication, exchanging intermediate results, and dynamically adapting to specific task requirements. In doing so, it transforms static architectures into dynamic, context-sensitive applications (Ehsan et al., 2021).

We focus on compound AI inference in this paper because it directly addresses the practical and technical challenges of deploying these systems at scale. Unlike traditional inference for single models, compound inference encompasses multiple layers of complexity, including:

- **Dynamic Workflow Execution**: Components might operate at varying granularities—from token to paragraph outputs—necessitating the coordination of asynchronous or parallel processes (Erbel and Grabowski, 2023).
- **Adaptive Decision-Making**: The system often needs to make runtime choices, such as deciding which tools to activate or how many iterations to run a model, based on both input and intermediate outputs. This requires smart, context-aware orchestration (Liu et al., 2022).
- **Efficiency and Scalability**: It is essential to manage both latency and computational costs, particularly for large-scale applications. Compound inference must allocate resources efficiently among diverse components to ensure real-time performance and reduce overhead (Meng, 2024).
- **Quality Optimization**: By linking multiple models or steps in a sequence, compound inference allows for the iterative refinement of outputs. This process, which can include techniques like self-consistency, ensembling, and contextual verification, improves overall accuracy (Okuyelu and Adaji, 2024).
- **Trust and Transparency**: By structuring the interactions between components explicitly, compound inference facilitates the tracking of result provenance, simplifies debugging, and boosts user confidence through methods such as fact-checking and evidence-based responses (Olateju et al., 2024).

The focus on compound AI inference is driven by its crucial role in unlocking the benefits of compound systems. Although these systems naturally provide flexibility, reliability, and advanced functionality, their full potential is achieved only through smart and efficient inference strategies. This paper explores emerging methodologies, challenges, and solutions for compound AI inference, emphasizing its importance in delivering state-of-the-art performance, boosting scalability, and spurring innovation in next-generation AI systems.

3 Infrastructure Perspectives of Scalable Compound AI Systems

The implementation of composite AI architectures requires strategic planning of computational ecosystems. As processing units rapidly diversify across hardware specifications and network topologies, deployment configuration choices directly impact operational efficiency, response times, and economic viability. This analysis compares three principal implementation frameworks: embedded processing, centralized cloud solutions, and distributed edge-cloud topologies, highlighting their distinct advantages and constraints in facilitating multi-stage AI pipelines. We further explore emerging infrastructure innovations needed to support next-generation intelligent systems.

3.1 On-Device AI Systems

Localized AI implementations utilize integrated computation modules - including mobile processors, neural accelerators, and specialized silicon within smart sensors or edge nodes. These configurations emphasize instantaneous decision-making and information security through localized data handling, circumventing cloud transmission requirements. Embedded solutions prove particu-

larly advantageous for latency-sensitive implementations like industrial robotics, augmented reality navigation, and responsive biometric systems (Ham et al., 2021).

The inherent limitations of memory capacity and parallel processing in embedded systems present unique adaptation challenges for multi-model architectures. Recent approaches like Fox-1's compact language modeling (Hu et al., 2024b) demonstrate techniques for minimizing neural network footprints while preserving functionality. Implementation strategies such as parameter reduction, architectural simplification, and knowledge compression have become essential for deploying sophisticated models on constrained devices. Furthermore, adaptive runtime managers exemplified by platform-specific optimization tools (e.g., Core ML, Qualcomm AI Stack) prove crucial for coordinating interconnected AI components within limited-resource environments.

3.2 Cloud-Based AI Systems

Server-centric intelligent architectures harness enterprise-grade computing clusters to execute demanding computational operations, providing elastic scalability and specialized processing capabilities through tensor processing arrays, quantum annealing units, and high-performance computing grids. This centralized paradigm proves particularly effective for developing and operating sophisticated multi-model architectures, enabling synergistic coordination between knowledge bases, foundation models, and third-party cognitive services (Witanto et al., 2022).

Contemporary server platforms incorporate adaptive coordination tools like Ray and Apache Airflow to govern interconnected AI pipelines. These environments support intelligent workload distribution and resilient error recovery mechanisms, maintaining operational stability during traffic fluctuations. Nevertheless, server-dependent architectures incur temporal delays from network hops and raise governance concerns regarding confidential information stewardship.

3.3 Edge-Cloud Hybrid AI Systems

Distributed cognitive architectures combine localized and centralized computing advantages through intelligent workload segmentation between edge nodes and cloud reservoirs (e.g. TensorOpera Router (Stripelis et al., 2024)). This synergistic model optimally serves composite AI implementations demanding both immediate responsiveness and heavy computational capacity. A representative pattern involves preliminary feature extraction at network periphery, with cognitive-intensive operations like contextual reasoning or cross-domain knowledge synthesis delegated to cloud reservoirs (Williams et al., 2024).

Primary optimization considerations encompass intelligent workload segmentation, network payload overhead reduction, and cross-layer state consistency management. Methodologies including decentralized model refinement, tiered model distribution, and context-aware information distillation prove instrumental in overcoming these constraints. Next-generation coordination platforms such as EdgeX Foundry and Azure IoT Edge offer enhanced capabilities for managing heterogeneous computing strata, facilitating intelligent resource negotiation across infrastructure boundaries.

3.4 Future Directions

The advancement of cognitive computing ecosystems will be shaped by escalating requirements for adaptive intelligence. Prospective innovation trajectories may concentrate on:

- **Context-Aware Resource Negotiation:** Creating cognitive scheduling systems capable of intelligent workload distribution across heterogeneous computing strata, guided by operational telemetry, power consumption optimization, and economic viability thresholds.
- **Federated Cognitive Pipelines:** Advancing decentralized intelligence frameworks that enable secure aggregation of distributed model components through confidential computing frameworks and zero-trust architectural principles.
- **Post-Von Neumann Computing Integration:** Investigating novel processing architectures including quantum-enhanced optimization kernels, neuromorphic sensory processing units, and photonics-based acceleration modules for next-generation cognitive workloads.
- **Self-Optimizing Operational Fabric:** Implementing meta-cognitive controllers for automated infrastructure governance, incorporating predictive anomaly mitigation, autonomous recovery protocols, and elastic ca-

pacity planning across 5G network slicing environments.

- **Universal Interface Specifications:** Developing vendor-agnostic communication standards (e.g., OpenAPI 3.0 extensions, gRPC-Lite interfaces) to enable plug-and-play component ecosystems across multi-vendor infrastructure deployments.

These paradigm shifts will prove essential for maintaining operational scalability, architectural resilience, and total cost of ownership advantages in next-generation cognitive computing environments.

4 System Perspectives of Scalable Composite AI Architectures

4.1 Operational Metrics for Scalable Cognitive Pipelines

The engineering of multi-component intelligent architectures demands rigorous evaluation of operational parameters governing real-world effectiveness, extending beyond conventional single-model frameworks like ScaleLLM (Yao et al., 2024a). Critical evaluation axes encompass transactional capacity, response intervals, memory footprint optimization, and economic sustainability, each requiring specialized optimization strategies to maintain operational integrity across dynamic workload conditions.

Transactional Capacity Enhancement. Cognitive architectures must process high-velocity inference streams through computational pipeline optimization. This demands intelligent workload distribution across decentralized processing units, coupled with concurrency management protocols. Adaptive workload grouping and event-driven processing paradigms can amplify transactional capacity through temporal resource multiplexing and computational phase overlapping (Chen et al., 2022).

Response Interval Minimization. Critical for interactive decision support platforms and real-time operational intelligence systems, latency reduction requires minimizing computational phase transitions. Implementation strategies include hardware-aware acceleration through tensor processing cores and precision scaling strategies. Predictive data staging mechanisms employing computational storage architectures and proximity-aware caching hierarchies prove essential for temporal optimization (Konakanchi, 2024).

Memory Footprint Optimization. Composite architectures integrating knowledge engines, transformer-based models, and cognitive toolchains necessitate advanced memory management. Cross-component parameter multiplexing, dynamic model subspace activation, and adaptive gradient recomputation strategies enable efficient memory utilization. Emerging techniques like holographic parameter representations and selective attention pruning further compress memory requirements for edge deployment scenarios (Kim et al., 2024b).

Economic Sustainability Factors. Lifecycle cost management requires intelligent resource provisioning across infrastructure layers. Multi-fidelity inference hierarchies, where lightweight analyzers filter computation-intensive tasks, combined with cloud-native autoscaling controllers, optimize expenditure. Energy-aware execution models employing transient computing resources and predictive workload shaping algorithms enhance cost-performance ratios (Hill et al., 2024).

4.2 Challenges in Multi-Component Optimization

The convergence of heterogeneous cognitive modules in composite architectures introduces multi-dimensional optimization challenges. Unlike unified model architectures permitting global gradient propagation, composite systems demand specialized co-optimization methodologies due to discrete decision boundaries between components. Advanced approaches including genetic optimization frameworks, automated policy search for pipeline configuration, and differentiable surrogate modeling warrant exploration for system-level tuning. The strategic distribution of computational budgets across functional units—such as optimizing energy allocation between knowledge grounding and generative synthesis phases—necessitates context-aware scheduling engines driven by workload fingerprints and operational constraints (Dessaux et al., 2024).

4.3 Future Directions

Composite cognitive architectures present numerous frontiers for technological advancement, with critical research trajectories including:

Contextual Resource Governance. Creating self-adaptive allocation systems powered by spectral workload analysis and federated performance telemetry. This involves developing meta-optimization layers capable of real-time pipeline morphing through topological recomposition and precision elasticity.

Unified Cognitive Optimization. Pioneering cross-stack optimization toolkits integrating neuro-symbolic optimization surfaces and hybrid convergence algorithms. Emerging paradigms like differentiable constraint satisfaction networks could enable holistic system refinement across continuous-discrete parameter spaces.

Sustainable Cognitive Computing. Addressing ecological impacts through photonic computation fabrics, neuromorphic energy harvesting architectures, and carbon-negative scheduling algorithms. Research into dynamic voltage-frequency scaling for AI accelerators and radiative cooling-enhanced data centers will prove critical.

Convergent System Engineering. Fostering interdisciplinary synergies between distributed systems theory, cognitive science, and cyber-physical security. Developing visual analytics suites for multi-dimensional performance tracing and causal relationship mining will accelerate architectural evolution.

Resilience Engineering. Ensuring operational integrity through adversarial robustness certification frameworks and self-stabilizing consensus protocols. Innovations in homomorphic encryption protocols for pipeline coordination and quantum-resistant authentication mechanisms will strengthen mission-critical deployments.

Addressing these frontiers will enable cognitive architectures to achieve unprecedented levels of operational intelligence, energy proportionality, and fault tolerance, driving their adoption across intelligent infrastructure and autonomous decision systems.

5 Machine Learning Perspectives of Scalable Compound AI Systems

The engineering of scalable composite AI systems presents distinctive machine intelligence challenges, particularly in cross-component optimization, resource governance, and emergent coordination phenomena. Diverging from conventional unified models, composite architectures integrate heterogeneous functional units—including domain-specific processors, knowledge grounding engines, and auxiliary cognitive services—requiring novel optimization paradigms to harmonize subsystem interactions under multidimensional constraints spanning temporal responsiveness, economic viability, and mission-specific success criteria.

5.1 Distributed Cognitive Coordination Mechanisms

Composite architectures necessitate advanced coordination frameworks for autonomous functional units operating in decentralized cognitive ecosystems. Each specialized processor (e.g., neural symbolic reasoners, cross-modal alignment engines, or federated knowledge graphs) must achieve emergent coherence through intelligent coordination protocols. Modern implementations typically employ *Cognitive Orchestrators*—meta-learning controllers that dynamically reconfigure processing workflows using topological awareness derived from component capability registries. These intelligent coordinators enforce constraint satisfaction through real-time performance telemetry analysis and predictive resource negotiation algorithms (Cao et al., 2012).

The coordination challenge intensifies with temporal decision-making requirements. Contemporary solutions employ hypergraph execution blueprints encoding multidimensional dependencies across cognitive units. Adaptive replanning engines utilizing counterfactual reasoning and topological persistence homology analysis enable robust contingency management during runtime anomalies like data stream discontinuities or component degradation. Emerging techniques in multi-agent meta-learning frameworks and hyperdimensional computing paradigms demonstrate potential for autonomous coordination policy evolution through experiential learning (Wang et al., 2022).

5.2 Metacognitive Evaluation Layers

A fundamental innovation in composite architectures is the integration of *Metacognitive Monitors*—specialized evaluation subsystems providing continuous quality assurance across cognitive workflows. These systems implement multidimensional assessment matrices combining learned quality predictors, symbolic rule engines, and anomaly detection networks to audit intermediate cognitive states. Particularly crucial for stochastic generative

components prone to semantic drift or factual inconsistencies, these monitors enable autonomous error correction through iterative refinement cycles (Li et al., 2024a).

Implementation paradigms leverage heterogeneous evaluator ensembles combining neuro-symbolic validators, topological data analysis modules, and quantum-inspired similarity metrics. Advanced architectures employ self-referential learning mechanisms where evaluation criteria dynamically adapt through epistemic uncertainty quantification. This automated refinement capability, reminiscent of conscious metacognition in biological systems, is increasingly realized through techniques like holographic consistency verification and differentiable proof systems (Kim et al., 2024a).

5.3 End-to-End Optimization in Non-Differentiable Systems

While conventional machine learning architectures employ differentiable loss functions for end-to-end training, compound AI systems integrate discrete components like search engines, symbolic reasoning modules, and third-party interfaces that preclude gradient-based optimization. This architectural constraint drives the adoption of specialized coordination mechanisms, including *pipeline-level tuning* for parameter synchronization and *stochastic optimization* techniques for global performance calibration. Innovative frameworks such as DSPy exemplify this paradigm through language model-driven adaptation, systematically refining prompts, contextual exemplars, and component-specific hyperparameters (Cheng et al., 2024a).

The optimization process further requires balancing computational expenditures against functional efficacy. Adaptive selection mechanisms exemplified by FrugalGPT (Chen et al., 2023a) demonstrate cost-aware resource allocation through context-sensitive model routing predicated on task granularity and budgetary parameters. Emerging collaborative optimization frameworks, particularly multi-agent reinforcement learning architectures, provide principled methodologies for coordinating interdependent subtask execution across specialized modules (Mondal et al., 2021).

5.4 Scalability and Resource Allocation

Architectural scalability in compound AI demands strategic distribution of finite computational assets to satisfy operational constraints like latency thresholds and throughput targets (Joshi and Kumar, 2021). Current implementations leverage hierarchical task partitioning coupled with adaptive workload distribution to prioritize critical operations while mitigating resource contention (Marzari et al., 2021). Distributed execution frameworks have become prevalent, employing agent clusters to parallelize task processing through decentralized coordination mechanisms (Olfati-Saber et al., 2007).

Intelligent data orchestration subsystems further enhance resource utilization by segmenting complex queries into atomic operations optimized for heterogeneous storage and compute layers (Cheng et al., 2024b). Advanced metadata management through centralized registries enables dynamic source selection and retrieval protocol optimization using multidimensional metrics encompassing data fidelity, access costs, and infrastructure availability (Peng et al., 2023). These mechanisms collectively establish a robust foundation for elastic system scaling under fluctuating operational demands.

5.5 Future Directions

As compound AI systems continue to evolve, several promising directions for future research and development emerge. These include advancements in multi-agent learning, robust optimization under uncertainty, and the seamless integration of diverse system components.

Adaptive Learning and Transferable Policies. A central challenge in multi-agent coordination lies in developing adaptive mechanisms for dynamic environments and evolving task specifications. Investigations into cross-domain policy transfer and meta-optimization frameworks could empower agents to extrapolate decision-making strategies across heterogeneous operational contexts. Such advancements hold potential to streamline the deployment of compound systems in novel applications by minimizing retraining overhead and computational resource expenditure.

Explainability and Trust. The inherent intricacy of compound AI architectures necessitates breakthroughs in operational transparency and accountability. Emerging systems might incorporate diagnostic XAI modules to generate granular explanations of agent-level reasoning and system-wide workflow patterns. Integrating these interpretability mechanisms with auditing subsystems, such as Judgment Agents, could enable real-time anomaly

diagnosis while fostering user confidence through human-readable behavioral rationales.

Federated and Decentralized Architectures. Architectural innovations emphasizing decentralized governance models could address scalability-privacy trade-offs inherent in centralized systems. By implementing federated coordination protocols, autonomous agents could process tasks through localized computation nodes, thereby reducing network latency and minimizing exposure of sensitive data through on-device retention. This paradigm shift may concurrently enhance fault tolerance by eliminating single points of failure.

Enhanced Optimization Frameworks. Next-generation optimization toolkits must address multidimensional performance criteria as system complexity escalates. Promising avenues include developing pseudo-differentiable computational graphs for hybrid pipelines and creating hybrid optimizers that synergize gradient-informed strategies with evolutionary algorithms. Extensions to frameworks like DSPy could incorporate multi-objective evaluation schemas quantifying ethical dimensions (e.g., algorithmic fairness), operational resilience, and sustainability metrics like carbon-footprint-adjusted efficiency (Cheng et al., 2024a).

Human-AI Collaboration. As these systems become more capable, there is a growing opportunity to redefine how humans and AI collaborate. Future research can explore interactive interfaces and agent architectures that support intuitive user interactions, adaptive feedback loops, and cooperative problem-solving, thus amplifying human creativity and decision-making.

By addressing these areas, compound AI systems can unlock their full potential to deliver reliable, scalable, and context-aware AI applications in enterprise and beyond.

6 Workflow Perspective of Scalable Compound AI Systems

From a workflow perspective, scalable Compound AI systems can be categorized into three distinct architectural paradigms: sequential, parallel, and hybrid workflows. Each paradigm provides unique benefits and caters to different application requirements. Below, we discuss these paradigms with illustrative examples and conclude with future directions.

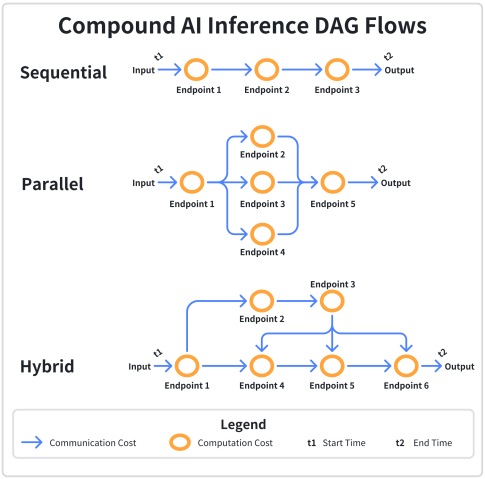

Figure 2: Workflow Architectures of Scalable Compound AI Systems.

6.1 Sequential Inference

Sequential inference embodies workflow architectures where component outputs directly feed successive stages in linear processing chains. This architecture proves particularly effective for applications demanding deterministic, staged data refinement (Fu et al., 2024).

- *Automated Verification Systems*: Frameworks like FacTool (Chern et al., 2023) implement staged validation through sequential phases of claim identification, evidence retrieval, query formulation, and cross-referenced verification (Russo et al., 2024).
- *Context-Enhanced Generation Systems*: Retrieval-augmented architectures prioritize document acquisition through specialized search modules before synthesizing responses via language models (Pouplin et al., 2024).
- *Algorithmic Code Synthesis (e.g., AlphaCode 2)*: These pipelines follow a generate-validate-rank sequence, iteratively producing code candidates, executing test cases, and prioritizing solutions through computational metrics (Zhang et al., 2023b).

6.2 Parallel Inference

Parallel inference orchestrates simultaneous execution of heterogeneous AI modules to address distinct subtasks or explore alternative solution pathways. This architecture proves advantageous for applications requiring distributed computation

or consensus-driven output synthesis (Li et al., 2024c).

Implementation Paradigms:

- *Multi-Modal Retrieval Systems*: Leverage concurrent retrieval architectures combining sparse indexing and semantic embedding techniques to maximize information diversity and precision (Nie et al., 2024).
- *Reasoning Consensus Frameworks*: Systems such as Gemini implement parallelized Chain-of-Thought exploration, sampling divergent reasoning trajectories before synthesizing responses through statistical consensus mechanisms (Rodríguez et al., 2024).
- *Clinical Decision Support (e.g., MedPrompt (Chen et al., 2023b))*: Generate competing diagnostic hypotheses through parallel prompting strategies, subsequently ensembled through probabilistic fusion to enhance clinical reliability (Zavaleta-Monestel et al., 2024).

6.3 Hybrid Inference

Hybrid inference architectures strategically combine linear and concurrent processing chains, merging the deterministic advantages of sequential workflows with the exploratory benefits of parallel computation. This dual-mode operation excels in multi-faceted tasks requiring both interdependent processing stages and independent computational pathways (Hao et al., 2024).

Exemplar Architectures:

- *Multi-Stage Verification Systems*: Building upon FacTool's methodology, these systems perform sequential claim extraction but conduct parallel verification through independent evidence scoring modules (Vural and Kalaman, 2024; Li et al., 2024b).
- *Iterative Reasoning Systems*: Multi-hop QA architectures employ phased context acquisition through sequential retrieval operations while concurrently evaluating multiple reasoning chains per processing stage (Cheng et al., 2024c).
- *Plugin-Enhanced LLM Ecosystems (e.g., ChatGPT [1])*: Feature simultaneous plugin activation (web interfaces, computational engines) managed through sequential orchestration logic to maintain task coherence [2].

6.4 Future Directions

The advancement of workflow architectures in compound AI systems opens several frontiers for innovation:

- **Dynamic Workflow Configuration**: Developing systems capable of real-time architectural adaptation between processing paradigms based on computational demands and task requirements.
- **Holistic System Co-Optimization**: Creating unified optimization frameworks that harmonize component interactions across hybrid workflows while balancing competing performance metrics.
- **Advanced Operational Telemetry**: Designing specialized monitoring interfaces for hybrid architectures to visualize cross-component dependencies and diagnose systemic bottlenecks.
- **Feedback-Informed Architecture Refinement**: Implementing self-improving workflows through continuous performance evaluation and automated configuration adjustments.
- **Cross-Domain Orchestration Mechanisms**: Fostering interdisciplinary solutions integrating distributed systems theory with human-computer interaction principles for robust system governance.

These architectural paradigms – sequential, parallel, and hybrid workflows – establish a modular foundation for addressing diverse computational challenges in compound AI systems. Progress in adaptive workflow engineering will critically influence the development of next-generation intelligent systems prioritizing scalability, operational efficiency, and task-specific reliability.

7 Application Perspectives of Scalable Compound AI Inference

This section examines the transformative impact of compound AI inference across multiple sectors, highlighting its role in advancing healthcare, legal compliance, enterprise operations, and customer

[1] https://chatgpt.com/

[2] https://www.cursor.com/

Type	Workflow Description	Examples
Sequential	Outputs flow linearly between components	Fact-checking (FacTool), RAG pipelines, code generation (AlphaCode 2)
Parallel	Independent components run concurrently	Search engines, self-consistency in CoT, medical diagnosis (MedPrompt)
Hybrid	Combines sequential and parallel workflows	Chatbot verification, multi-hop QA (HotpotQA), tool-augmented LLMs (ChatGPT Plus)

Table 1: Taxonomy of Compound AI Workflow Architectures: System workflows are classified by their component coordination patterns. Sequential architectures enforce strict processing order, parallel designs enable distributed computation, while hybrid systems strategically combine both approaches for balanced performance.

engagement. Through domain-specific implementations, we demonstrate how these architectures address unique challenges while enhancing operational efficiency and decision-making precision.

7.1 Healthcare

- **Diagnostic Decision Support:** Composite AI systems are transforming diagnostic methodologies by integrating generative LLMs with information retrieval architectures and clinical knowledge bases. Systems such as MedPrompt (Chen et al., 2023b) employ similarity-based retrieval of patient histories alongside structured reasoning chains, ensuring diagnostic outputs align with evidence-based guidelines and individualized health profiles.

- **Computational Drug Development:** Platforms like AlphaFold (Abramson et al., 2024) utilize composite inference to synthesize symbolic reasoning engines with machine learning architectures, significantly accelerating molecular dynamics simulations and target interaction predictions.

- **Remote Care Optimization:** Integrated systems combine sensor analytics pipelines with language models to generate interpretable summaries of longitudinal patient data, enabling proactive clinical interventions (Zavaleta-Monestel et al., 2024).

7.2 Legal and Regulatory Compliance

- **Automated Legal Analysis:** Frameworks such as Athena (Xiao and Chen, 2024) facilitate automated processing and condensation of legal texts through retrieval-augmented generation methodologies, cross-referencing statutory databases to ensure jurisprudential accuracy (Guha et al., 2023).

- **Dynamic Compliance Monitoring:** Enterprise solutions dynamically evaluate regulatory documentation through composite systems orchestrating LLMs alongside deterministic rule engines, maintaining real-time adherence to evolving legislative frameworks (Hassani, 2024; Hassani et al., 2024).

7.3 Enterprise and Customer Service

- **Intelligent Conversational Systems:** Contemporary conversational agents, exemplified by ChatGPT Plus, utilize composite inference architectures to interleave language generation with external tool invocation (web interfaces[3], code execution environments[4]), enabling context-aware, multi-modal interactions.

- **Data-Driven Customer Intelligence:** Hybrid AI systems employ retrieval-generation frameworks to analyze extensive customer feedback datasets while establishing cross-domain correlations, driving strategic product optimization (Zhang et al., 2023a).

7.4 Finance

- **Anomaly Detection Architectures:** Modern financial systems synthesize statistical predictors with deterministic rule engines and outlier detection architectures to identify suspicious

[3] https://platform.openai.com/docs/guides/tools-web-search?api-mode=chat

[4] https://microsoft.github.io/autogen/stable/user-guide/core-user-guide/design-patterns/code-execution-groupchat.html

transaction patterns (Hu et al., 2024a; Dubois et al., 2024). This multi-layered approach enhances fraud identification accuracy while minimizing operational disruptions from false alerts.

- **Algorithmic Asset Management:** Systems strategically merge language models for macroeconomic pattern recognition with constraint-based allocation solvers, enabling adaptive portfolio rebalancing through multi-temporal data synthesis (Wang et al., 2024).

7.5 Education and Knowledge Management

- **Personalized Learning Systems:** Adaptive platforms orchestrate generative content personalization with knowledge base query systems, tailoring educational pathways to learner-specific cognitive profiles and competency gaps.

- **Academic Research Automation:** Hybrid AI systems automate literature synthesis pipelines through retrieval-augmented generation frameworks, while tools like AlphaGeometry (Trinh et al., 2024) enhance mathematical discovery by integrating formal logic systems with neural conjecture validation frameworks.[5]

7.6 Technology and Software Development

- **Automated Code Synthesis:** Platforms such as AlphaCode (Siam et al., 2024) synthesize neural code generation with automated validation pipelines, enhancing solution robustness through iterative compilation testing and metric-driven ranking.

- **Intelligent Operations Automation:** Composite DevOps systems converge natural language interfaces with infrastructure telemetry systems, enabling predictive system diagnostics and automated remediation workflows.

7.7 Creative Industries

- **Generative Content Production:** Systems like DALL-E[6] converge generative architectures with editorial constraint modules, enabling streamlined multimodal content production across visual and textual domains.

- **Dynamic Narrative Generation:** Advanced storytelling platforms fuse narrative generation engines with contextual memory architectures, maintaining plot coherence through recursive consistency verification (Kumaran et al., 2024).

7.8 Autonomous Systems and Robotics

- **Autonomous Navigation Systems:** Vehicle control architectures synthesize multimodal sensor fusion systems with probabilistic decision frameworks, ensuring collaborative operational integrity across perception-action cycles.

- **Adaptive Manufacturing Systems:** Industrial automation platforms converge semantic task planning with actuator feedback control systems, achieving sub-millimeter precision in dynamic production environments (Lee and Su, 2023).

7.9 Search and Retrieval

- **Intelligent Search Infrastructure:** Modern engines synthesize neural query interpretation with multi-index retrieval architectures (Gao et al., 2023), enabling context-aware search optimization through real-time relevance feedback loops.[7]

- **Knowledge Curation Systems:** AI-driven pipelines converge semantic parsing engines with structured data extraction toolchains, maintaining enterprise knowledge graphs through continuous verification cycles (Zhang and Soh, 2024).

8 Compound AI Training: Challenges and Future Directions

The advancement of compound AI systems necessitates focused research to address critical challenges in architectural complexity, modular interdependencies, and operational scalability. Below we delineate pivotal research frontiers, particularly emphasizing training methodologies that underpin next-generation system capabilities.

8.1 Enhanced Training Techniques for Compound AI Systems

Training compound architectures presents unique challenges due to their discrete computational elements and heterogeneous component interactions

[5] https://github.com/binary-husky/gpt_academic
[6] https://openai.com/index/dall-e-3/
[7] https://www.bing.com/

(Yao et al., 2024b). Unlike monolithic models benefiting from end-to-end differentiability, these systems demand novel optimization frameworks. Promising research directions include:

- **Gradient Approximation Techniques for Discrete Modules:** Developing differentiable approximation frameworks for non-differentiable operations (e.g., database queries, symbolic solvers) to enable gradient propagation across pipeline stages.
- **Cross-Modular Adaptation Protocols:** Designing joint training regimes that synchronize language model fine-tuning with auxiliary tool calibration (retrieval systems, computational modules) to achieve functional alignment.
- **Adaptive Pipeline Configuration via RL:** Implementing reinforcement learning strategies for dynamic, context-aware module selection and hyperparameter adjustment during inference.
- **Iterative Human Feedback Integration:** Incorporating qualitative human evaluations into training loops to optimize subjective metrics like output interpretability and task-specific usability.
- **Decentralized Parameter Optimization:** Advancing coordinated training methodologies for multi-agent systems through conflict-resistant update synchronization and stability-preserving learning rules. This includes federated optimization approaches where modular components undergo distributed training while exchanging critical performance signals. Key challenges involve mitigating gradient conflicts, establishing robust inter-agent communication protocols, and maintaining convergence stability in heterogeneous parameter spaces.

8.2 Scaling and Resource Optimization

The growing sophistication of compound AI systems demands innovative approaches to computational efficiency and operational scalability. Key research priorities include:

- **Context-Aware Resource Distribution:** Developing intelligent allocation mechanisms that dynamically assign computational assets based on operational demands and latency thresholds, maximizing throughput within cost-performance boundaries.
- **Precision Computational Orchestration:** Advancing pipeline-aware scheduling through proactive workload forecasting and distribution, minimizing resource underutilization via task prioritization and interruptible execution protocols.
- **Budget-Aware Deployment Paradigms:** Adapting resource-efficient frameworks like FrugalGPT for composite architectures, maintaining functional accuracy while optimizing expenditure across interdependent computational modules.

8.3 Architectural Innovations

Exploring novel structural configurations remains crucial for advancing compound AI capabilities:

- **Componentized System Design:** Creating plug-and-play architectures with standardized interfaces, enabling rapid tool-model recombination with reduced integration friction.
- **Input-Sensitive Architectural Reconfiguration:** Engineering context-aware systems capable of real-time module selection and interaction pattern adjustment based on data characteristics.
- **Recursive Refinement Mechanisms:** Investigating bidirectional communication patterns where downstream outputs recursively optimize upstream processing parameters.

8.4 Robustness, Security, and Explainability

Ensuring operational reliability and transparency in complex systems requires:

- **Fault Tolerance Architectures:** Implementing cross-component error diagnostics with automated recovery protocols to maintain service continuity.
- **Adversarial Resilience Frameworks:** Developing security layers that protect tool-model interfaces against prompt injection and data poisoning attacks.
- **Decision Provenance Tracking:** Creating audit trails that visually map information flow across modules, explaining final outputs through component-level contribution analysis.

8.5 Scalable MLOps and Monitoring

Operational sustainability demands advanced infrastructure solutions:

- **Cross-Component Telemetry Systems:** Implementing transactional logging with granular performance metrics across heterogeneous modules.
- **Data Pipeline Synchronization Protocols:** Ensuring consistent data quality standards across retrieval, preprocessing, and model input stages.
- **Iterative Performance Analytics:** Establishing closed-loop systems where operational metrics automatically trigger component retraining or pipeline reconfiguration.

8.6 Human-Centric Design and Applications

Bridging technical capabilities with societal needs necessitates:

- **Low-Code Configuration Interfaces:** Developing visual workflow builders that abstract technical complexity for domain specialists.
- **Human-AI Copilot Paradigms:** Creating interactive systems that adapt reasoning paths based on real-time user feedback and preference signals.
- **Bias Mitigation Protocols:** Implementing fairness-preserving constraints across all pipeline stages, particularly in demographic-sensitive applications.

These research vectors collectively address the dual challenges of technical complexity and real-world applicability in compound AI systems. Advances in coordinated training methodologies and adaptive architectures will prove particularly pivotal, enabling seamless integration of discrete innovations into cohesive intelligent systems. The ultimate goal remains developing AI solutions that transcend individual component capabilities through strategic synergy, while maintaining rigorous standards of efficiency, transparency, and ethical responsibility.

9 Related Work

The evolution of Compound AI systems builds upon progress in modular system design, domain-optimized architectures, and reliability engineering, establishing foundations for adaptive, enterprise-grade AI solutions.

9.1 Modular Frameworks and Architectures

Pioneering frameworks including *LangChain*[8], *AutoGPT*[9], and *DSPy* (Khattab et al., 2023) have redefined modular AI development through automated parameter optimization, resource orchestration, and component interoperability features. Platforms like *Databricks Mosaic AI* [10] exemplify production-grade implementations, merging foundation models with computational toolchains (e.g., vector databases, batch processors) to address industrial-scale challenges.

9.2 Task-Specific and Retrieval-Driven Systems

Compound architectures achieve domain-specific performance enhancements through strategic tool-model integration. *AlphaCode 2* (Siam et al., 2024) demonstrates this via its pipeline combining LLM-generated code candidates with clustering-based ranking, while *MedPrompt* (Nori et al., 2023) enhances clinical reasoning through retrieval-augmented chain-of-thought workflows. Contemporary approaches like Chain-of-Knowledge (Li et al., 2024d) further validate the efficacy of external knowledge integration, outperforming isolated models through structured multi-source reasoning.

9.3 Safety and Reliability

Operational safety remains paramount for high-stakes deployments. Frameworks such as *TorchOpera* (Han et al., 2024a) implement multi-layered safeguards—including hallucination detection and contextual grounding modules—to address operational risks like sensitive data exposure. Modular verification frameworks (Davis et al., 2024a) enhance reasoning reliability by decoupling generation and validation processes, underscoring the critical role of trustworthiness in compound system design.

These developments collectively highlight Compound AI's transformative potential through three core principles: modular extensibility, domain-optimized tool integration, and engineered operational safety.

[8] https://www.langchain.com/

[9] https://github.com/Significant-Gravitas/AutoGPT

[10] https://www.databricks.com/product/machine-learning

10 Conclusion

The advancements and challenges outlined in this paper underscore the transformative potential of compound AI systems as a frontier in artificial intelligence research. By integrating heterogeneous AI components within cohesive architectures, these systems offer superior scalability, adaptability, and reliability across a broad spectrum of applications. Their modular nature mitigates the limitations inherent in monolithic AI models, facilitating dynamic reconfiguration, enhanced computational efficiency, and optimized resource utilization.

This study emphasizes the critical role of inferencing in compound AI systems, wherein the seamless orchestration of diverse components enables real-time decision-making, iterative refinement, and rigorous quality assurance. The stratified architecture of these systems—encompassing application layers, workflow management, machine learning models, system-level optimizations, and infrastructure—provides a structured framework for managing complexity and achieving end-to-end performance optimization.

Despite their advantages, compound AI systems present substantial challenges, including the necessity for novel optimization methodologies, adaptive resource allocation, and improved interoperability. The absence of standardized frameworks for addressing trade-offs in latency, accuracy, and coherence highlights the pressing need for continued research and methodological innovation. Furthermore, ensuring robustness, transparency, and ethical considerations in deployment remains imperative, particularly as these systems are increasingly applied in high-stakes domains such as healthcare, finance, and autonomous systems.

Future research must focus on developing scalable training paradigms, advancing multi-agent coordination mechanisms, and enhancing adaptive and explainable AI workflows within compound AI systems. The integration of emerging technologies such as quantum computing and decentralized architectures has the potential to further expand their capabilities. Moreover, fostering interdisciplinary collaboration and prioritizing user-centric design principles will be crucial to unlocking the full potential of compound AI systems and fostering technological innovation across multiple domains.

In conclusion, compound AI systems represent a paradigm shift in artificial intelligence, offering a pathway to more robust, efficient, and trustworthy AI applications. By addressing the challenges identified in this paper and leveraging the proposed advancements, the research community can drive the evolution of next-generation AI systems capable of meeting the increasing complexity of modern applications and advancing a wide range of scientific and industrial fields through intelligent, scalable solutions.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. 2024. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493—-500.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. 2012. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9:427–438.

Baiqi Chen, Jianpei Dong, Marina Ruelas, Xiangyi Ye, Jinxu He, Ruijie Yao, Yuqiu Fu, Ying Liu, Jingpeng Hu, Tianyu Wu, et al. 2022. Artificial intelligence-assisted high-throughput screening of printing conditions of hydrogel architectures for accelerated dia-

betic wound healing. *Advanced Functional Materials*, 32(38):2201843.

Lingjiao Chen, Matei A. Zaharia, and James Y. Zou. 2023a. Frugalgpt: How to use large language models while reducing cost and improving performance. *ArXiv*, abs/2305.05176.

Xuhang Chen, Chi-Man Pun, and Shuqiang Wang. 2023b. Medprompt: Cross-modal prompting for multi-task medical image translation. *ArXiv*, abs/2310.02663.

Ching-An Cheng, Allen Nie, and Adith Swaminathan. 2024a. Trace is the next autodiff: Generative optimization with rich feedback, execution traces, and llms. In *Neural Information Processing Systems*.

Jie Cheng, Yingbing Chen, Xiaodong Mei, Bowen Yang, Bo Li, and Ming Liu. 2024b. Rethinking imitation-based planners for autonomous driving. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14123–14130.

Keyuan Cheng, Gang Lin, Haoyang Fei, Yuxuan Zhai, Lu Yu, Muhammad Asif Ali, Lijie Hu, and Di Wang. 2024c. Multi-hop question answering under temporal knowledge editing. *ArXiv*, abs/2404.00492.

Ethan Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios. *ArXiv*, abs/2307.13528.

Florin Cuconasu, Giovanni Trappolini, F. Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *ArXiv*, abs/2401.14887.

Jared Quincy Davis, Boris Hanin, Lingjiao Chen, Peter Bailis, Ion Stoica, and Matei Zaharia. 2024a. Networks of networks: Complexity class principles applied to compound ai systems design. *arXiv preprint arXiv:2407.16831*.

Jared Quincy Davis, Boris Hanin, Lingjiao Chen, Peter D. Bailis, Ion Stoica, and Matei Zaharia. 2024b. Networks of networks: Complexity class principles applied to compound ai systems design. *ArXiv*, abs/2407.16831.

Delphine Dessaux, Samuel Buchet, Lucie Barthe, Marianne Defresne, Gianluca Cioci, Simon De Givry, Luis F Garcia-Alles, Thomas Schiex, and Sophie Barbe. 2024. Designing symmetrical multicomponent proteins using a hybrid generative ai approach. *bioRxiv*, pages 2024–06.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Upol Ehsan, Qingzi Vera Liao, Michael J. Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Johannes Erbel and Jens Grabowski. 2023. Scientific workflow execution in the cloud using a dynamic runtime model. *Softw. Syst. Model.*, 23:163–193.

Yanlin Feng, Sajjadur Rahman, Aaron Feng, Vincent Chen, and Eser Kandogan. 2024. Cmdbench: A benchmark for coarse-to-fine multimodal data discovery in compound ai systems. *Proceedings of the Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI*.

Yichao Fu, Peter D. Bailis, Ion Stoica, and Hao Zhang. 2024. Break the sequential dependency of llm inference using lookahead decoding. *ArXiv*, abs/2402.02057.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

MyungJoo Ham, Jijoong Moon, Geunsik Lim, Jaeyun Jung, Hyoungjoo Ahn, Wook Song, Sangjung Woo, Parichay Kapoor, Dongju Chae, Gichan Jang, Yong Min Ahn, and Jihoon Lee. 2021. Nnstreamer: Efficient and agile development of on-device ai systems. *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 198–207.

Shanshan Han, Zijian Hu, Alay Dilipbhai Shah, Han Jin, Yuhang Yao, Dimitris Stripelis, Zhaozhuo Xu, and Chaoyang He. 2024a. Torchopera: A compound ai system for llm safety. *arXiv preprint arXiv:2406.10847*.

Shanshan Han, Yuhang Yao, Zijian Hu, Dimitris Stripelis, Zhaozhuo Xu, and Chaoyang He. 2024b. Torchopera: A compound ai system for llm safety. *ArXiv*, abs/2406.10847.

Zixu Hao, Huiqiang Jiang, Shiqi Jiang, Ju Ren, and Ting Cao. 2024. Hybrid slm and llm for edge-cloud collaborative inference. *Proceedings of the Workshop on Edge and Mobile Foundation Models*.

Shabnam Hassani. 2024. Enhancing legal compliance and regulation analysis with large language models. *2024 IEEE 32nd International Requirements Engineering Conference (RE)*.

Shabnam Hassani, Mehrdad Sabetzadeh, Daniel Amyot, and Jain Liao. 2024. Rethinking legal compliance automation: Opportunities with large language models. *2024 IEEE 32nd International Requirements Engineering Conference (RE)*.

Harry Hill, Cristina Roadevin, Stephen Duffy, Olena Mandrik, and Adam Brentnall. 2024. Cost-effectiveness of ai for risk-stratified breast cancer screening. *JAMA Network Open*, 7(9):e2431715–e2431715.

Sihao Hu, Tiansheng Huang, Ka-Ho Chow, Wenqi Wei, Yanzhao Wu, and Ling Liu. 2024a. Zipzap: Efficient training of language models for large-scale fraud detection on blockchain. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2807–2816, New York, NY, USA. Association for Computing Machinery.

Zijian Hu, Jipeng Zhang, Rui Pan, Zhaozhuo Xu, Shanshan Han, Han Jin, Alay Dilipbhai Shah, Dimitris Stripelis, Yuhang Yao, Salman Avestimehr, Chaoyang He, and Tong Zhang. 2024b. Fox-1 technical report. *arXiv preprint arXiv:2411.05281*.

Swayambhoo Jain, Ravi Raju, Bo Li, Zoltan Csaki, Jonathan Li, Kaizhao Liang, Guoyao Feng, Urmish Thakkar, Anand Sampat, Raghu Prabhakar, and Sumati Jairath. 2024. Composition of experts: A modular compound ai system leveraging large language models. *ArXiv*, abs/2412.01868.

Rajendra P. Joshi and Neeraj Kumar. 2021. Artificial intelligence for autonomous molecular design: A perspective. *Molecules*, 26.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2024a. Spread preference annotation: Direct preference judgment for efficient llm alignment.

Ji Yeon Kim, Seong Hyeon Jo, Ha Sang-hyun, Ki Hwan Kim, Young Jin Kang, and Seok Chan Jeong. 2024b. Comparison of ai model serving efficiency: Response time and memory usage analysis. *AHFE International*.

Sandeep Konakanchi. 2024. Next-generation low-latency architectures for real-time ai-driven cloud services. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*.

Vikram Kumaran, Jonathan Rowe, and James Lester. 2024. Narrativegenie: generating narrative beats and dynamic storytelling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, pages 76–86.

Jay Lee and Hanqi Su. 2023. A unified industrial large knowledge model framework in industry 4.0 and smart manufacturing. *arXiv preprint arXiv:2312.14428*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *ArXiv*, abs/2411.16594.

Dongfang Li, Xinshuo Hu, Zetian Sun, Baotian Hu, Shaolin Ye, Zifei Shan, Qian Chen, and Min Zhang. 2024b. Truthreader: Towards trustworthy document

assistant chatbot with reliable attribution. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Ming-Yu Liu, Kai Li, and Song Han. 2024c. Distrifusion: Distributed parallel inference for high-resolution diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7183–7193.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024d. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*.

Jiaxin Liu, Wenhui Zhou, Hong Wang, Zhong Cao, Wen-Hui Yu, Cheng-Yu Zhao, Ding Zhao, Diange Yang, and Jun Li. 2022. Road traffic law adaptive decision-making for self-driving vehicles. *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2034–2041.

Brady D. Lund and Wang Ting. 2023. Chatting about chatgpt: How may ai and gpt impact academia and libraries? *SSRN Electronic Journal*.

Luca Marzari, Ameya Pore, Diego Dall'Alba, Gerardo Aragon-Camarasa, Alessandro Farinelli, and Paolo Fiorini. 2021. Towards hierarchical task decomposition using deep reinforcement learning for pick and place subtasks. *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 640–645.

Xianghui Meng. 2024. Optimization of algorithmic efficiency in ai: Addressing computational complexity and scalability challenges. *Applied and Computational Engineering*.

Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V. Ukkusuri. 2021. On the approximation of cooperative heterogeneous multi-agent reinforcement learning (marl) using mean field control (mfc). *ArXiv*, abs/2109.04024.

Guangtao Nie, Rong Zhi, Xiaofan Yan, Yufan Du, Xiangyang Zhang, Jianwei Chen, Mi Zhou, Hongshen Chen, Tianhao Li, Ziguang Cheng, Sulong Xu, and Jinghe Hu. 2024. A hybrid multi-agent conversational recommender system with llm and search engine in e-commerce. In *ACM Conference on Recommender Systems*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Olanrewaju M. Okuyelu and Ojima Ojodumi Adaji. 2024. Ai-driven real-time quality monitoring and process optimization for enhanced manufacturing performance. *Journal of Advances in Mathematics and Computer Science*.

Omobolaji Olufunmilayo Olateju, Samuel Ufom Okon, Oluwaseun Oladeji Olaniyi, Amaka Debie Samuel-Okon, and Christopher Uzoma Asonze. 2024. Exploring the concept of explainable ai and developing information governance standards for enhancing trust and transparency in handling customer data. *Journal of Engineering Research and Reports*.

Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. 2007. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95:215–233.

Yuan Peng, Franziska Bathelt, Richard Gebler, Robert Gött, Andreas Heidenreich, Elisa Henke, Dennis Kadioglu, Stephan Lorenz, Abishaa Vengadeswaran, and Martin Sedlmayr. 2023. Use of metadata-driven approaches for data harmonization in the medical domain: Scoping review. *JMIR Medical Informatics*, 12.

Thomas Pouplin, Hao Sun, Samuel Holt, and Mihaela van der Schaar. 2024. Retrieval-augmented thought process as sequential decision making. *ArXiv*, abs/2402.07812.

Yujia Qin, Shi Liang, Yining Ye, Kunlun Zhu, Lan Yan, Ya-Ting Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Marc H. Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *ArXiv*, abs/2307.16789.

María Estrella Vallecillo Rodríguez, Arturo Montejo Ráez, and María Teresa Martín-Valdivia. 2024. Sinai at pan 2024 textdetox: Application of chain of thought with self-consistency strategy in large language models for multilingual text detoxification. In *Conference and Labs of the Evaluation Forum*.

Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. 2024. Face the facts! evaluating rag-based fact-checking pipelines in realistic settings. *ArXiv*, abs/2412.15189.

Keshav Santhanam, Deepti Raghavan, Muhammad Shahir Rahman, Thejas Venkatesh, Neha Kunjal, Pratiksha Thaker, Philip Levis, and Matei Zaharia. 2024. Alto: An efficient network orchestrator for compound ai systems. *Proceedings of the 4th Workshop on Machine Learning and Systems*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. 2024. Small llms are weak tool learners: A multi-llm agent. *ArXiv*, abs/2401.07324.

Md Kamrul Siam, Huanying Gu, and Jerry Q Cheng. 2024. Programming with ai: Evaluating chatgpt, gemini, alphacode, and github copilot for programmers. *arXiv preprint arXiv:2411.09224*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *ArXiv*, abs/2408.03314.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *ArXiv*, abs/2409.12183.

Dimitris Stripelis, Zijian Hu, Jipeng Zhang, Zhaozhuo Xu, Alay Dilipbhai Shah, Han Jin, Yuhang Yao, Salman Avestimehr, and Chaoyang He. 2024. Tensoropera router: A multi-model router for efficient llm inference. *arXiv preprint arXiv:2408.12320*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Trieu Trinh, Yuhuai Wu, Quoc Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*.

Nazmi Ekin Vural and Sefer Kalaman. 2024. Using artificial intelligence systems in news verification: An application on x. *İletişim Kuram ve Araştırma Dergisi*.

Jianrui Wang, Yitian Hong, Jiali Wang, Jiapeng Xu, Yang Tang, Qing-Long Han, and Jürgen Kurths. 2022. Cooperative and competitive multi-agent systems: From optimization to games. *IEEE/CAA Journal of Automatica Sinica*, 9:763–783.

Mengxin Wang, Dennis J. Zhang, and Heng Zhang. 2024. Large language models for market research: A data-augmentation approach. *arXiv preprint arXiv:2412.19363*.

K. Lalith Williams, Y. Durga Prasanth, and M. Jeyaselvi. 2024. Hybrid ai architecture using edge-cloud computing for secure v2x communication. *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, pages 913–920.

Elizabeth Nathania Witanto, Yustus Eko Oktian, and Sang-Gon Lee. 2022. Toward data integrity architecture for cloud-based ai systems. *Symmetry*, 14:273.

Peng Xiao and Liang Chen. 2024. Athena: Retrieval-augmented legal judgment prediction with large language models. *arXiv:2410.11195*.

Yuhang Yao, Han Jin, Alay Dilipbhai Shah, Shanshan Han, Zijian Hu, Yide Ran, Dimitris Stripelis, Zhaozhuo Xu, Salman Avestimehr, and Chaoyang He. 2024a. Scalellm: A resource-frugal llm serving framework by optimizing end-to-end efficiency. *arXiv preprint arXiv:2408.00008*.

Yuhang Yao, Jianyi Zhang, Junda Wu, Chengkai Huang, Yu Xia, Tong Yu, Ruiyi Zhang, Sungchul Kim, Ryan Rossi, Ang Li, et al. 2024b. Federated large language models: Current progress and future directions. *arXiv preprint arXiv:2409.15723*.

Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, and Ali Ghodsi. 2024a. The shift from models to compound ai systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/.

Matei Zaharia et al. 2024b. The shift from models to compound ai systems. *Berkeley Artificial Intelligence Research, February*, 18.

Esteban Zavaleta-Monestel, Ricardo Quesada-Villaseñor, Sebastián Arguedas-Chacón, Jonathan García-Montero, Monserrat Barrantes-López, Juliana Salas-Segura, Adriana Anchía-Alfaro, Daniel Nieto-Bernal, and Daniel E Diaz-Juan. 2024. Revolutionizing healthcare: Qure.ai's innovations in medical diagnosis and treatment. *Cureus*, 16.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. *arXiv preprint arXiv:2404.03868*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023a. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 349–356, New York, NY, USA. Association for Computing Machinery.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B. Tenenbaum, and Chuang Gan. 2023b. Planning with large language models for code generation. *ArXiv*, abs/2303.05510.