Large Language Models AI Agents and Multi-Agent Systems: A Survey

www.surveyx.cn

Abstract

This survey paper explores the transformative impact of Large Language Models (LLMs), AI agents, and multi-agent systems on AI research and development. It covers architectural innovations, training strategies, and applications of LLMs across diverse domains, emphasizing their role in enhancing natural language processing, conversational AI, and autonomous decision-making. The integration of LLMs into AI agents has advanced their reasoning and adaptability, while multi-agent systems benefit from improved communication and coordination. Despite these advancements, challenges such as biases, ethical considerations, and computational limitations persist, necessitating innovative solutions and robust methodologies. The survey highlights the need for tailored transparency practices and enhanced cybersecurity measures to address these issues. Future research opportunities include refining LLM architectures, improving data augmentation, and exploring dual-modality frameworks to enhance human-machine interaction. By addressing current challenges and leveraging technological advancements, LLMs and related systems are poised to revolutionize various industries, offering significant potential for enhanced problem-solving and decision-making capabilities.

1 Introduction

1.1 Scope and Significance

Large Language Models (LLMs), AI agents, and multi-agent systems are pivotal in advancing AI research and development. This survey covers a broad spectrum of topics, including architectural innovations, training strategies, and fine-tuning methods essential for LLM evolution [1]. Evaluating both quantitative and qualitative intelligence of LLMs is crucial for developing metrics that compare computational methods to human intelligence [2]. The development of Multi-LLM-Agent Systems (MLAS) is examined, focusing on architecture, protocols, and business models that enhance task execution and collaboration among AI agents [3].

LLMs have transformed fields such as dentistry by enabling automated and cross-modal dental diagnosis [4] and enhancing knowledge base question answering (KBQA) for structured data queries [5]. Their integration into AI agents facilitates the automation of Programmable Logic Controller (PLC) code generation and verification, ensuring operational efficiency and safety in industrial control systems [6]. Additionally, enabling LLMs to create their own tools enhances their autonomy and adaptability in problem-solving [7].

In education, LLMs play a significant role in creating environments with multiple conversational agents (CAs), demonstrating their transformative potential [8]. The COVID-19 pandemic has amplified the demand for psychological counseling, highlighting LLMs' importance in scaling global mental health services [9]. Moreover, LLMs and Multilingual LLMs (MLLMs) are crucial for developing robust Machine Translation (MT) systems for low-resource languages in crisis communication [10].

Generated by SurveyX.cn.

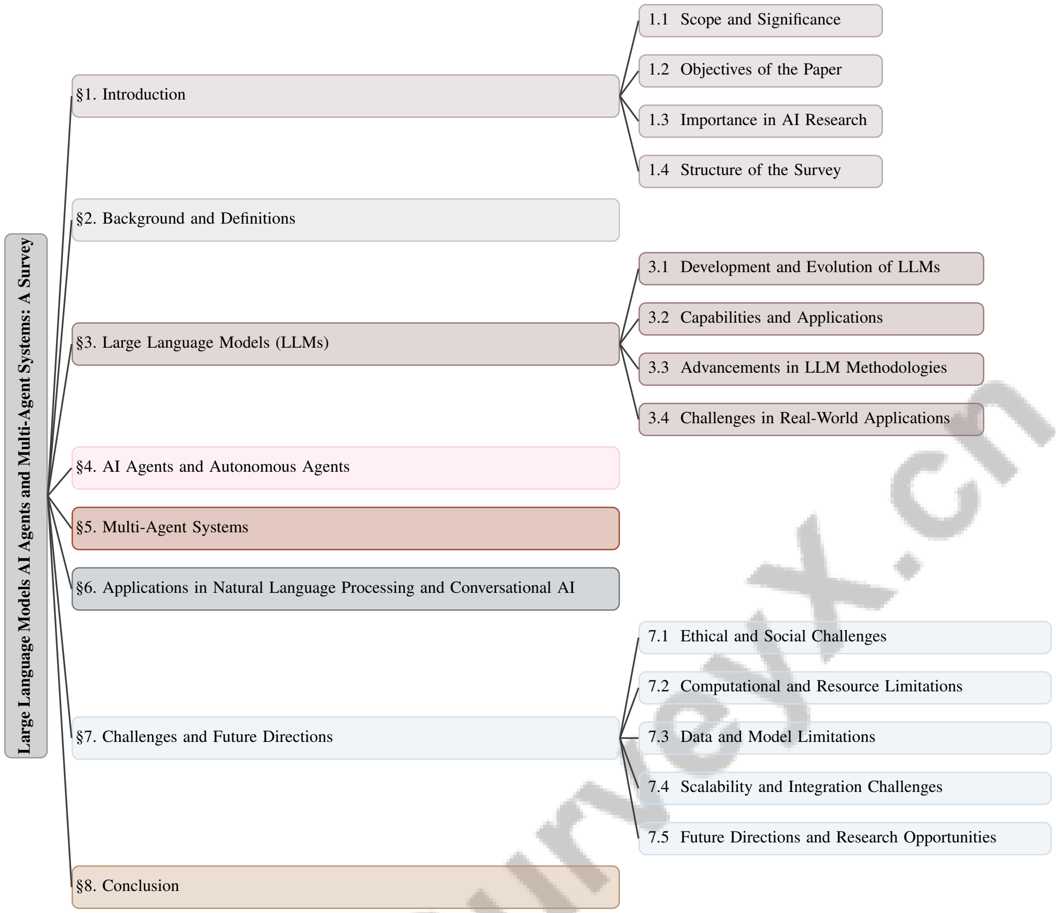

Large Language Models AI Agents and Multi-Agent Systems: A Survey
§1. Introduction
1.1 Scope and Significance
1.2 Objectives of the Paper
1.3 Importance in AI Research
1.4 Structure of the Survey
§2. Background and Definitions
§3. Large Language Models (LLMs)
3.1 Development and Evolution of LLMs
3.2 Capabilities and Applications
3.3 Advancements in LLM Methodologies
3.4 Challenges in Real-World Applications
§4. AI Agents and Autonomous Agents
§5. Multi-Agent Systems
§6. Applications in Natural Language Processing and Conversational AI
§7. Challenges and Future Directions
7.1 Ethical and Social Challenges
7.2 Computational and Resource Limitations
7.3 Data and Model Limitations
7.4 Scalability and Integration Challenges
7.5 Future Directions and Research Opportunities
§8. Conclusion

Figure 1: chapter structure

This survey underscores the substantial impact of LLMs across various fields, particularly in natural language processing (NLP), and emphasizes the need for innovative methodologies to address existing limitations [11]. In cybersecurity, an analysis of over 180 works related to LLMs covers 25 different models and more than 10 downstream scenarios, highlighting their critical role in addressing security challenges [12]. The survey also discusses risks associated with LLM-based agents, including security, privacy, and ethical concerns [13].

Generative AI and LLMs are explored in strategic air traffic flow management to improve efficiency and decision-making processes [14]. The development of semi-autonomous AI-driven agents interacting online on behalf of legal entities is critical, revealing inadequacies in previous methods [15]. The extensive scope and importance of LLMs, AI agents, and multi-agent systems in shaping the future of AI research and development are emphasized, highlighting their role in advancing capabilities and addressing contemporary challenges.

Benchmarks for chat-optimized language models (cLLMs) are essential for evaluating their capabilities in interactive, game-like settings, which aids in understanding situated language comprehension [16]. The integration of LLMs into space applications is pivotal for advancing autonomous spacecraft operations [17]. The rapid growth of data assets has driven advancements across various domains while presenting challenges for human-data interaction [18]. Evaluating LLMs for social intelligence is currently constrained by insufficient scenario diversity and complexity [19]. Traditional information retrieval methods fall short in addressing human curiosity, necessitating a new approach leveraging LLM advancements [20]. The evolving landscape of cyber threats underscores the importance of LLMs in cybersecurity [21], while their integration into process mining addresses the increasing complexity of process data [22].

AI-generated, for reference only.

1.2 Objectives of the Paper

This survey aims to provide a comprehensive examination of recent advancements in Large Language Models (LLMs), AI agents, and multi-agent systems, addressing key gaps in understanding their architectural innovations, training strategies, and diverse applications [1]. A primary objective is to establish benchmarks such as LeSC, which assess the nuanced semantic comprehension capabilities of LLMs, particularly in interpreting common words with uncommon meanings [23]. The survey also focuses on enhancing the performance and applicability of LLMs across domain-specific contexts, including healthcare, law, and finance [11].

In education, the survey investigates the role of LLMs in automating grading and generating feedback, contributing to improved educational outcomes [24]. It introduces frameworks like KoMA, enabling LLM-driven agents to interact, plan, and learn collectively, thereby enhancing decision-making capabilities in multi-agent driving environments [25]. Additionally, it provides a systematic overview of LLM applications in cybersecurity, emphasizing the need for comprehensive literature reviews in this critical area [12].

The performance of conversational LLMs (cLLMs) is evaluated using benchmarks like clembench, which assess instruction-following and engagement in dialogue games [16]. Innovative interaction paradigms such as 'Online Training using External Interactions' are explored, facilitating real-time updates and customization based on user interactions [26]. Moreover, the challenge of transforming raw data into engaging animated data videos is addressed, emphasizing the coordination of diverse components and specialized skills [18].

Evaluating the social intelligence of language agents through diverse and interactive social scenarios using AgentSense is another objective [19]. The survey proposes systems like KwaiAgents, leveraging LLMs as a cognitive core to enhance the efficiency and accuracy of information retrieval [20]. Furthermore, it addresses the shortcomings of traditional honeypots by proposing an LLM-based honeypot for improved data collection and analysis [21]. Lastly, the survey benchmarks LLMs in process mining tasks, focusing on the abstraction and interpretation of process-related data and insight generation [22]. Through these objectives, the survey seeks to advance the field by identifying opportunities and challenges in the integration and application of LLMs, AI agents, and multi-agent systems, ultimately enhancing their interaction capabilities by effectively modeling the behaviors, goals, and beliefs of other agents.

1.3 Importance in AI Research

The significance of Large Language Models (LLMs), AI agents, and multi-agent systems in advancing AI technologies is profound, as they collectively drive innovation across numerous domains. LLMs, with their scalable architectures and emergent abilities, have transformed the landscape of AI research, offering unprecedented capabilities in natural language understanding and generation [1]. These models have been instrumental in enhancing automatic grading systems, addressing resource constraints in educational settings by providing efficient and scalable solutions [24]. Furthermore, LLMs emulate human cognitive processes, significantly improving the effectiveness of autonomous driving systems in complex scenarios [25].

In cybersecurity, the integration of LLMs presents both challenges and opportunities, enhancing existing methodologies and addressing the evolving threat landscape [12]. Tailored approaches to LLM deployment have been shown to significantly improve their performance in specialized applications, underscoring their relevance in advancing AI technologies across various fields [11]. The adaptability of LLMs is further enhanced through methods that allow for real-time personalization, crucial for their continued evolution and applicability [26].

The deployment of LLMs in space operations exemplifies their ability to overcome the limitations of traditional Reinforcement Learning (RL) methods, thereby expanding the potential of AI technologies in novel environments [17]. Additionally, the automation of story generation from raw data highlights efficiency gains in data analysis and communication, demonstrating the broad applicability of LLMs in transforming information processing [18].

Developing comprehensive benchmarks for evaluating social intelligence in AI agents is imperative, as existing frameworks often fail to capture the dynamic nature of social interactions [19]. The integration of LLMs, AI agents, and multi-agent systems into AI research is pivotal, as these technologies continue to push the boundaries of what is possible, addressing contemporary challenges and paving the way for future advancements.

1.4 Structure of the Survey

This survey is meticulously structured to deliver an in-depth examination of Large Language Models (LLMs), their applications, the challenges they face, and their integration into AI agents and multi-agent systems, while also exploring the historical evolution, architectural frameworks, and future directions of this rapidly advancing field [27, 28, 29, 30]. It begins with an introduction that outlines the scope, significance, and objectives of the survey, emphasizing the pivotal role these technologies play in advancing AI research. The introduction also highlights the importance of these systems in various domains and sets the stage for the detailed discussions that follow.

The second section, Background and Definitions, establishes a conceptual framework by defining key terms and exploring the interconnections and distinctions among LLMs, AI agents, autonomous agents, multi-agent systems, natural language processing, conversational AI, and agent-based modeling. This foundational section ensures a clear understanding of the core concepts discussed throughout the survey.

In the third section, Large Language Models (LLMs), the survey delves into the development, capabilities, and applications of LLMs, focusing on their role in natural language processing and conversational AI. This section also addresses the challenges faced by LLMs in real-world applications and highlights recent methodological advancements.

The fourth section, AI Agents and Autonomous Agents, examines the characteristics and functionalities of AI agents, exploring how LLMs enhance their capabilities. It also discusses the human-like reasoning abilities of autonomous agents and their integration with LLMs.

The fifth section, Multi-Agent Systems, explores the significance of multi-agent systems in simulating complex interactions. It discusses cooperative dynamics, agent-based modeling and simulation, and strategies for communication and coordination among agents, emphasizing the role of LLMs in facilitating these processes.

In the sixth section, Applications in Natural Language Processing and Conversational AI, the survey highlights innovative applications and real-world examples of these technologies. It discusses enhancements in language understanding and generation tasks and explores their impact on human-machine interaction.

The seventh section, Challenges and Future Directions, identifies current challenges faced by LLMs, AI agents, and multi-agent systems, including ethical, social, computational, and resource limitations. It also discusses data and model limitations, scalability, and integration challenges, while proposing future research opportunities.

Finally, the conclusion summarizes the key points discussed, reiterating the significance of LLMs, AI agents, and multi-agent systems in advancing AI technologies. Throughout the survey, a framework for categorizing LLMs based on their architectural design is introduced, encompassing task-specific, task-independent, and various transformer-based architectures [31]. This structured approach ensures a coherent narrative, facilitating a deeper understanding of the transformative potential of these AI technologies.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Conceptual Framework and Definitions

Large Language Models (LLMs) are sophisticated AI systems adept at processing, understanding, and generating human language using extensive datasets and complex neural network architectures. These models are defined by scaling laws and emergent capabilities, enabling efficient execution of natural language processing tasks. Domain-Specific Language Model Adaptation (DSLMA) fine-tunes pre-trained models with domain-specific data to enhance adaptability [20]. Benchmarks address challenges in using LLMs for interpreting process mining artifacts [22].

AI agents are computational entities that perceive their environment and act to achieve objectives. Autonomous agents, a subset of AI systems, operate independently in complex environments, utilizing social reasoning and self-evolving mechanisms, often leveraging LLMs for improved decision-making and inter-agent interactions [32, 33]. These agents excel in autonomous decision-making and problem-solving applications.

Multi-agent systems comprise multiple interacting agents that may collaborate or compete, studying emergent behaviors in various contexts. These systems are crucial for AI-generated science (AIGS) and can be optimized for collaborative dynamics. Emphasizing data privacy and monetization, multi-agent systems leverage LLM advancements to enhance capabilities in critical tasks such as information discovery and organization in research [20, 30].

Natural Language Processing (NLP) involves computer-human language interaction, including language understanding and sentiment analysis. Conversational AI, utilizing LLMs, enables natural dialogues, enhancing interaction coherence. LLMs are vital in education, supporting interactive tutoring experiences. Research shows smaller LLMs can be fine-tuned for personalized tutoring, achieving performance similar to larger models at reduced costs, aiding personalized learning and addressing expert-curated dataset challenges [34, 35].

Agent-based modeling (ABM) uses autonomous agents in defined environments to achieve goals, allowing complex system behaviors to emerge. This approach is effective for simulating opinion dynamics and social interactions, utilizing LLMs to reflect human behavior complexities, including polarization and misinformation. Unlike traditional ABMs, LLMs exhibit a bias towards accurate information, fostering consensus aligned with scientific reality but potentially impeding modeling of resistance to accepted views like climate change. Prompt engineering can induce confirmation bias, leading to opinion fragmentation consistent with opinion dynamics research, suggesting real-world discourse refinement could improve LLMs' simulation of human belief evolution [36, 29, 37]. Advanced AI techniques enhance ABM by integrating reasoning capabilities and addressing traditional model limitations.

The symbol grounding problem, enabling artificial agents to derive meaningful interpretations from symbols without external guidance, remains a significant challenge for autonomous AI systems. This issue highlights limitations of approaches relying solely on statistical language patterns, which fail to capture communication's functional aspects essential for effective human interaction. Research into grounded compositional language and neuro-symbolic AI techniques shows promise in developing agents capable of processing symbols while engaging in meaningful reasoning and decision-making, bridging the gap between symbolic representation and practical language use [38, 39, 40]. Transparency in AI models, including LLMs, is crucial for building trust and clarity. Model reporting, evaluation result publication, and stakeholder-targeted explanations are essential. The ongoing debate about LLMs' 'real' understanding and intentionality underscores the need for further exploration of cognitive and ethical implications.

2.2 Interconnections and Distinctions

The interplay between LLMs, AI agents, autonomous agents, multi-agent systems, NLP, conversational AI, and ABM reveals synergies and distinct characteristics. LLMs enhance applications in NLP and conversational AI with their human-like language processing and generation capabilities. However, their probabilistic nature poses challenges such as incomplete responses and misinformation, necessitating robust mechanisms for content accuracy [13].

AI and autonomous agents use LLMs to enhance decision-making and language understanding, benefiting from LLMs' human-like reasoning and adaptability in dynamic environments. Deploying these agents in real-world scenarios, like medical settings, requires ensuring applicability across diverse populations. Designing agent architectures that fully leverage LLM capabilities while developing comprehensive evaluation strategies is critical due to LLMs' unpredictable nature, complicating their deployment as autonomous agents. Essential considerations include Planning, Memory, Tools, and Control Flow, emphasizing the need for robust frameworks to manage stochasticity and resource efficiency. Integrating LLMs into multi-agent systems enhances cognitive abilities in problem-solving, but research gaps remain, necessitating a focused agenda to optimize individual agent capabilities and collaborative synergy. Developing effective evaluation platforms like AgentBench and WebArena is vital for assessing agents' performance in real-world scenarios, facilitating broader adoption across diverse sectors [41, 42, 43].

Multi-agent systems simulate complex interactions among agents, utilizing LLMs to improve communication and coordination. These systems often employ ABM techniques to simulate emergent behaviors in various contexts. Traditional modeling methods inadequately capture complex intelligent systems' dynamic requirements, highlighting the need for innovative approaches. Integrating LLMs within multi-agent environments enhances language processing and interpretation capabilities, enabling agents to autonomously perceive, control, and interact with their surroundings. This advancement fosters more natural and efficient interactions by allowing LLM agents to use external tools, dynamically decompose tasks, and specialize in specific roles, enhancing system flexibility and scalability. The collaborative nature of LLM-based Multi-Agent Systems (LaMAS) addresses complex challenges across domains like customer service and software engineering, supporting the development of robust, resilient agents capable of performing diverse functions in real-world applications [41, 3, 42, 43].

The survey categorizes current methods into centralized, decentralized, and fully decentralized architectures, each with distinct operational mechanisms and control requirements [3]. This categorization highlights diverse strategies in multi-agent systems for managing communication and coordination challenges. The core issue explored is the vulnerability of LLM-based agents to security and privacy threats due to operational complexity and user interaction reliance [13].

While LLMs, AI agents, and multi-agent systems share goals of enhancing language understanding and interaction capabilities, they present unique challenges and opportunities. Distinctions arise from tailored applications and methodologies to navigate complexities in various contexts, such as conducting academic research surveys and automating structured representation extraction from legislative texts for legal expert systems. For example, ResearchArena benchmarks LLMs in information discovery, selection, and organization, highlighting their performance in academic settings, while another study explores LLMs' role in automating structured representation extraction from legislative texts, enhancing legal decision support systems' efficiency [30, 35].

3 Large Language Models (LLMs)

Understanding the foundational aspects of Large Language Models (LLMs) is crucial to appreciating their current capabilities and applications. This section examines the historical context and milestones in the development of LLMs, tracing their evolution from basic systems to advanced architectures. Table 1 offers a detailed comparison of the core features of Large Language Models (LLMs), focusing on their development, capabilities, and methodological advancements. Key advancements and methodologies are highlighted to underscore the transformative impact of LLMs across various domains, setting the stage for a detailed discussion on their development and evolution.

3.1 Development and Evolution of LLMs

The evolution of Large Language Models (LLMs) has been marked by significant advancements in architecture and capabilities, transforming them from basic pattern recognition systems to sophisticated models capable of simulating human-like interactions [7]. This transformation is rooted in the shift from early rule-based systems to modern architectures, highlighting the limitations of prior approaches and the breakthroughs achieved with contemporary models [11]. The transition from statistical and neural language models to large-scale pre-trained models is systematically categorized into stages, emphasizing more efficient training strategies and inference mechanisms [1].

Transformative developments, such as transformer-based architectures like BERT and GPT, have revolutionized the field by enabling LLMs to process extensive datasets and capture complex linguistic patterns [44]. These advancements have facilitated their deployment across diverse domains, including optimization, where they enhance interpretability and decision-making [45]. Improvements in planning capabilities, particularly in task decomposition and multi-plan selection, underscore LLMs' transformative potential in complex decision-making [46].

As illustrated in Figure 2, the key developments in the architecture, application domains, and learning methodologies of LLMs are highlighted, showcasing their evolution into complex systems capable of diverse applications and continuous learning. LLMs have been integrated into specialized frameworks, such as the Conversation as Command Framework for intuitive human-robot interactions and the Richelieu agent for diplomatic tasks. Additionally, the LATM framework exemplifies innovative

AI-generated, for reference only.

approaches to augmenting LLM capabilities, enhancing their adaptability and integration into complex systems [7].

Recent research highlights the need for LLMs to engage in incremental learning through interactions with users and external knowledge sources, addressing historical limitations and facilitating continuous model improvement [26]. This evolution is complemented by the development of benchmarks, such as those for conversational LLMs (cLLMs), which assess performance across interactive scenarios [16]. The introduction of AgentSense, constructing 1,225 diverse scenarios emphasizing goal completion and implicit reasoning, exemplifies innovative benchmarking approaches [19].

The layered architecture for Multi-LLM-Agent Systems (MLAS) emphasizes decentralized control and effective communication protocols, illustrating the integration of LLMs into multi-agent environments [3]. This integration enhances scalability and efficiency, reinforcing LLMs' role in advancing AI technologies across domains. Hybrid frameworks, such as the planning-concluding procedure with a hybrid search-browse toolkit, exemplify innovative strategies for managing and retrieving information [20]. The LLM Honeypot (LLM-H) method showcases fine-tuning pre-trained LLMs on attacker command datasets to create interactive security systems [21]. These advancements highlight LLMs' ongoing evolution and expanding influence.

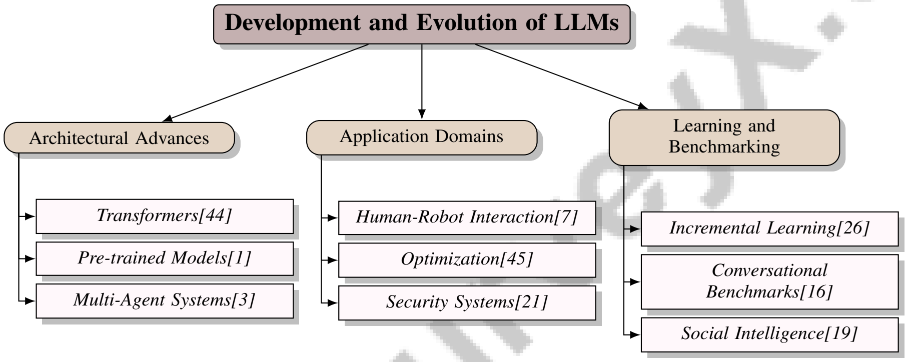


Figure 2: This figure illustrates the key developments in the architecture, application domains, and learning methodologies of Large Language Models (LLMs), highlighting their evolution into complex systems capable of diverse applications and continuous learning.

3.2 Capabilities and Applications

LLMs have demonstrated transformative capabilities across domains, enhancing efficiency, accuracy, and decision-making. In legal text analysis, they convert plain-text legislation into structured representations, offering novel approaches to understanding legal documents [35]. This capability underscores LLMs' multifaceted nature, explained by distinct factors rather than being monolithic [44].

In cybersecurity, LLMs automate threat intelligence and vulnerability detection processes traditionally requiring manual intervention [12]. The LLM Honeypot method exemplifies supervised fine-tuning to mimic Linux server behavior, enabling intelligent responses to attacker-generated data [21].

Healthcare applications include the Multi-Modal LLM for Dentistry (MMLLM), which enhances diagnostic accuracy by integrating visual, textual, and audio data, providing a comprehensive approach to automated dental diagnosis [4]. The Psy-LLM framework leverages LLMs to generate coherent responses to psychological inquiries, enhancing mental health support services [9].

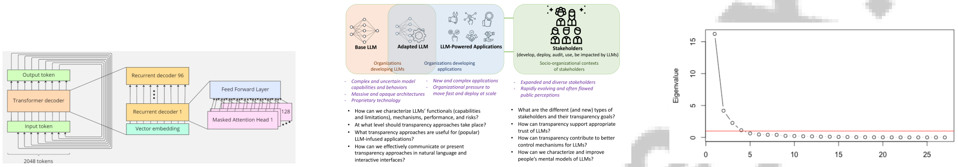
In education, LLMs serve as cost-effective tutors, maintaining performance across key educational metrics. Fine-tuned models, such as the quantized LLaMA-2, demonstrate high accuracy in predicting grades and generating feedback aligned with expert evaluations [24]. Using multiple conversational agents in educational settings increases user engagement and improves learning outcomes, highlighting LLMs' potential in fostering interactive learning experiences [8].

LLMs in autonomous systems enable natural language interaction and informed decision-making in autonomous driving scenarios [47]. Frameworks like CARE, combining chat-based interfaces with multiple LLM-driven agents, showcase LLM capabilities in enhancing user interactions through personalized exploration [48].

In language translation, combining LLMs with fine-tuning techniques and community involvement develops specialized datasets for low-resource languages, improving translation accuracy and responsiveness [10]. This innovation is crucial in crisis communication scenarios requiring accurate translations.

Challenges remain in tasks requiring deep analytical skills, where LLMs underperform compared to keyword-based methods [30]. While LLMs understand formal languages well, their generation capabilities lag, especially for formalized languages [5]. The Conversation as Command Framework demonstrates high command recognition accuracy and effective task execution, with a navigation success rate of 97.96% [49].

Innovative applications include the automatic creation of animated data videos through the Data Director, an LLM-based multi-agent system [18]. In air traffic management, LLMs summarize historical Ground Delay Programs (GDPs), aiding Traffic Managers in efficient decision-making [14]. These examples underscore LLMs' transformative potential across fields, driving innovation and addressing complex challenges.



- Complex and uncertain model capabilities and behaviors
- Massive and opaque architectures
- Proprietary technology

- New and complex applications
- Organizational pressure to move fast and deploy at scale

- Expanded and diverse stakeholders
- Rapidly evolving and often flawed public perceptions

- How can we characterize LLMs' functionals (capabilities and limitations), mechanisms, performance, and risks?
- At what level should transparency approaches take place?
- What transparency approaches are useful for (popular) LLM-infused applications?
- How can we effectively communicate or present transparency approaches in natural language and interactive interfaces?

- What are the different (and new) types of stakeholders and their transparency goals?
- How can transparency support appropriate trust of LLMs?
- How can transparency contribute to better control mechanisms for LLMs?
- How can we characterize and improve people's mental models of LLMs?

(a) Transformer Recurrent Decoder Architecture[50]

(b) Transparency in the Era of Large Language Models: Challenges and Opportunities[51]

(c) Eigenvalue Distribution[44]

Figure 3: Examples of Capabilities and Applications

As shown in Figure 3, Large Language Models (LLMs) have emerged as a transformative force in artificial intelligence, offering unprecedented capabilities and a wide array of applications. The "Transformer Recurrent Decoder Architecture" showcases a sophisticated design integrating a transformer decoder with a recurrent decoder and a feed-forward layer, highlighting the intricate processes in token generation and processing. This architecture exemplifies the complex engineering behind LLMs that enables coherent and contextually relevant outputs. "Transparency in the Era of Large Language Models" addresses the challenges and opportunities presented by LLMs, emphasizing the importance of understanding stakeholder roles in their development and deployment, crucial for ethical and responsible use. The "Eigenvalue Distribution" plot provides insight into LLMs' mathematical underpinnings, illustrating how eigenvalues are distributed within a dataset and offering a glimpse into the structural properties influencing model performance. These examples underscore LLMs' diverse capabilities and potential to revolutionize applications across industries [50, 51, 44].

3.3 Advancements in LLM Methodologies

Recent advancements in Large Language Models (LLMs) are marked by significant methodological innovations, enhancing their natural language understanding and generation capabilities. The Experiential Co-Learning framework integrates past experiences into LLM-powered multi-agent collaboration, improving adaptability and efficiency in dynamic environments [52]. Lifelong learning techniques emphasize model expansion and strategic data selection, allowing LLMs to continuously improve performance across diverse scenarios [53].

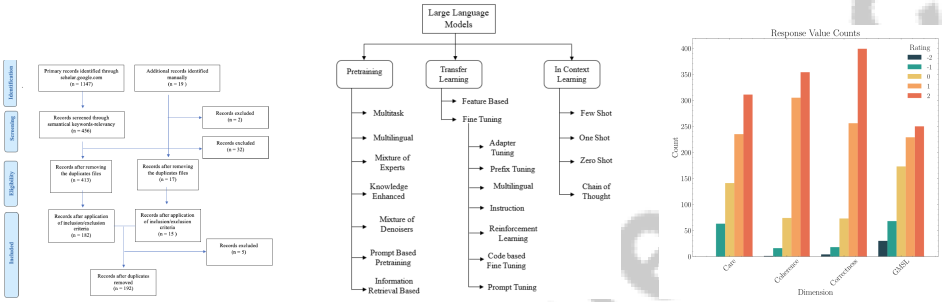
GraphLLM exemplifies the synergy between graph learning and LLMs, significantly boosting accuracy and context efficiency compared to traditional Graph2Text methods [54]. This innovation highlights the potential of integrating different AI methodologies to enhance LLM capabilities. Systematic encoding of qualitative insights into actionable features using LLMs provides a novel approach to translating expert intuition into quantifiable data, a method not leveraged by traditional techniques [55].

AI-generated, for reference only.

In decision-making, integrating LLMs with evolutionary algorithms (EAs) allows for automated, nuanced explanations of complex multi-objective optimization results, enhancing interpretability and transparency [45]. An LLM-agnostic memory architecture integrating database selection and data value memories enables seamless information retrieval from multiple databases, facilitating more informed and contextually relevant outputs [56].

MLPrompt, utilizing multilingual prompts to enhance LLM reasoning, contrasts with existing methods relying solely on dominant language processing, broadening LLM applicability across linguistic contexts [57]. Integrating regression and generative approaches within a unified system allows LLMs to perform effectively in scoring and feedback generation tasks, demonstrating versatility and adaptability in educational settings [24].

These methodological advancements underscore LLMs' ongoing evolution, enhancing capabilities and expanding applicability across diverse fields. The focus on improving efficiency, aligning models with human preferences, and reducing the environmental impact of LLM training highlights the potential for future research to refine and optimize these models [58].



(a) Identification, Screening, Eligibility, and Included Records[59]

(b) Large Language Models[31]

(c) Response Value Counts[34]

Figure 4: Examples of Advancements in LLM Methodologies

As shown in Figure 4, the exploration of advancements in Large Language Models (LLMs) methodologies is multifaceted, as illustrated by the diverse examples. "Identification, Screening, Eligibility, and Included Records" provides a comprehensive overview of the systematic research process, delineating stages from initial identification of records to inclusion. This structured approach underscores the meticulous nature of LLM research. "Large Language Models" offers a detailed flowchart highlighting methods employed in training, including Pretraining, Transfer Learning, and In Context Learning, emphasizing the complexity involved in developing LLMs. The "Response Value Counts" bar chart presents a quantitative analysis of response values across dimensions like Care, Coherence, Correctness, and GMSL, showcasing evaluative metrics used to assess LLM performance. These examples provide a comprehensive snapshot of current advancements and methodological approaches in LLMs [59, 31, 34].

3.4 Challenges in Real-World Applications

Large Language Models (LLMs) face numerous challenges in real-world applications, impacting performance and scalability. A significant issue is their limited understanding of nuanced lexical semantics, essential for tasks requiring deep semantic comprehension [23]. The foundational abilities enabling LLMs to excel across tasks are not well-defined, complicating benchmark development to accurately reflect capabilities and predict behavior [44].

The computational demands of training LLMs are substantial, requiring vast data and resources, posing barriers to widespread deployment. Aligning LLMs with human values and ethical standards remains challenging, often resulting in models failing to meet societal expectations [1]. In domains like Mixed Integer Programming (MIP), natural language ambiguity and complex rule sets challenge LLMs' ability to generate precise structured data, complicating structured problem-solving [57].

Processing large-scale graphs presents challenges, as LLMs exhibit poor performance due to the absence of explicit reasoning paths, limiting effectiveness in tasks requiring detailed graph reasoning

[60]. The generic nature of LLMs poses difficulties in domain-specific applications, affecting accuracy and relevance in specialized fields [11]. Existing benchmarks often fail to adequately challenge LLMs in interactive settings, leading to an incomplete understanding of capabilities in real-world scenarios [16].

LLMs lack adaptive learning and dynamic decision-making capabilities characteristic of human traders, crucial for achieving market equilibrium in competitive environments [61]. The rigidity of current methods, not supporting real-time updates or accommodating individual user preferences, restricts adaptability and responsiveness [26]. Ensuring accuracy of generated content and managing task decomposition complexity remain significant obstacles in applications like automated data storytelling [18].

A critical challenge is suboptimal exploitation of LLM capabilities in planning, memory utilization, and tool integration, affecting performance in information retrieval tasks [20]. Addressing these challenges requires ongoing research and development to enhance LLM robustness and adaptability, ensuring successful integration into diverse real-world applications.

In recent years, the evolution of artificial intelligence (AI) has led to a growing interest in the categorization and functionality of AI agents, particularly in relation to autonomous agents. Understanding the distinctions and connections between these entities is crucial for both theoretical exploration and practical application. As illustrated in Figure 6, this figure presents a comprehensive hierarchical structure of AI agents and autonomous agents, detailing their characteristics and functionalities.

This figure illustrates the key challenges faced by Large Language Models (LLMs) in real-world applications, categorized into semantic comprehension, structured data generation, and adaptive learning limitations. It highlights the need for improved semantic understanding, better alignment with human values, and enhanced capabilities in domain-specific tasks and adaptive learning. The diagram effectively categorizes the core features and operational aspects of AI agents, while also addressing their limitations and advancements. Notably, it highlights the capabilities of autonomous agents, which exhibit human-like reasoning and are enhanced through interactive experiences. Furthermore, the integration of large language models (LLMs) is thoroughly detailed within this framework, emphasizing significant advancements in capabilities, applications, and future developments. This visual representation not only aids in the clarification of complex concepts but also enriches the narrative by providing a structured overview of the current landscape of AI technologies.

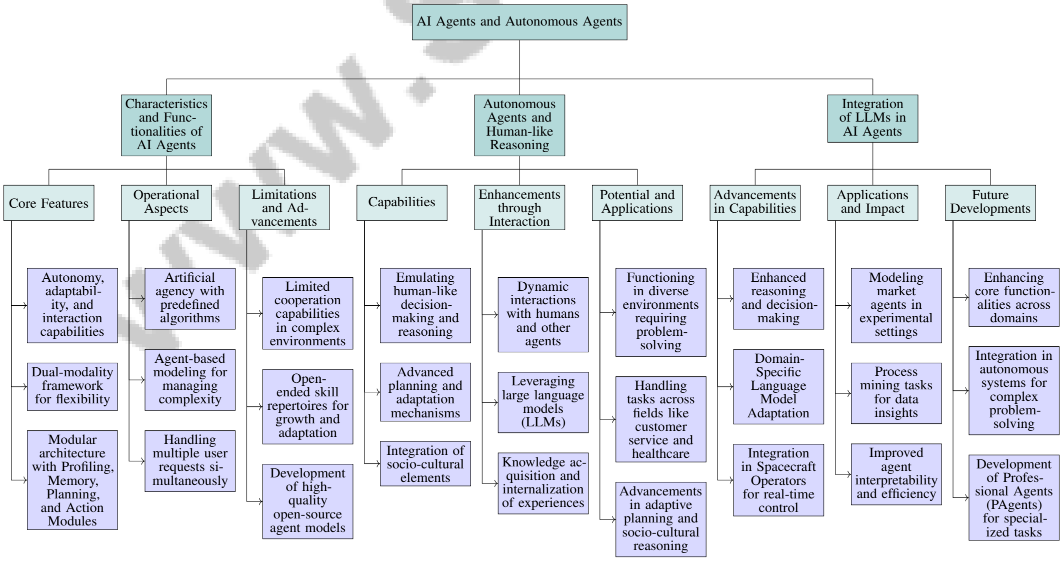

AI Agents and Autonomous Agents
Characteristics and Functionalities of AI Agents
Autonomous Agents and Human-like Reasoning
Integration of LLMs in AI Agents
Core Features
Operational Aspects
Limitations and Advancements
Capabilities
Enhancements through Interaction
Potential and Applications
Advancements in Capabilities
Applications and Impact
Future Developments
Autonomy, adaptability, and interaction capabilities
Dual-modality framework for flexibility
Modular architecture with Profiling, Memory, Planning, and Action Modules
Artificial agency with predefined algorithms
Agent-based modeling for managing complexity
Handling multiple user requests simultaneously
Limited cooperation capabilities in complex environments
Open-ended skill repertoires for growth and adaptation
Development of high-quality open-source agent models
Emulating human-like decision-making and reasoning
Advanced planning and adaptation mechanisms
Integration of socio-cultural elements
Dynamic interactions with humans and other agents
Leveraging large language models (LLMs)
Knowledge acquisition and internalization of experiences
Functioning in diverse environments requiring problem-solving
Handling tasks across fields like customer service and healthcare
Advancements in adaptive planning and socio-cultural reasoning
Enhanced reasoning and decision-making
Domain-Specific Language Model Adaptation
Integration in Spacecraft Operators for real-time control
Modeling market agents in experimental settings
Process mining tasks for data insights
Improved agent interpretability and efficiency
Enhancing core functionalities across domains
Integration in autonomous systems for complex problem-solving
Development of Professional Agents (PAgents) for specialized tasks

Figure 5: This figure illustrates the hierarchical structure of AI agents and autonomous agents, categorizing their characteristics, functionalities, and the integration of large language models (LLMs). It showcases the core features and operational aspects of AI agents, their limitations, and advancements. Furthermore, it highlights the capabilities of autonomous agents in human-like reasoning, their enhancements through interaction, and potential applications. The integration of LLMs is detailed, emphasizing advancements in capabilities, applications, and future developments.

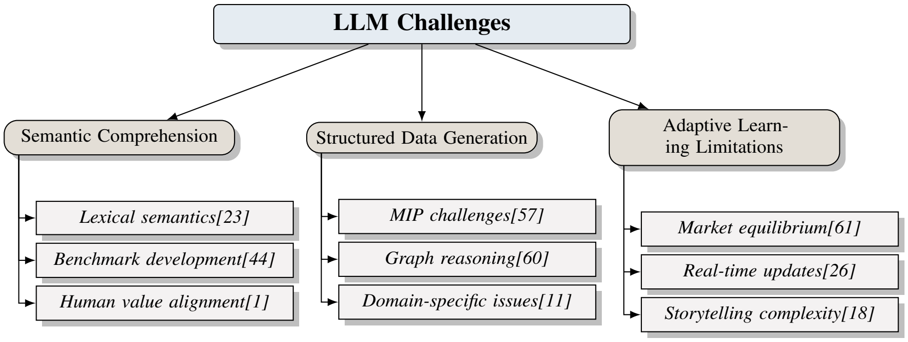


Figure 6: This figure illustrates the key challenges faced by Large Language Models (LLMs) in real-world applications, categorized into semantic comprehension, structured data generation, and adaptive learning limitations. It highlights the need for improved semantic understanding, better alignment with human values, and enhanced capabilities in domain-specific tasks and adaptive learning.

Feature	Development and Evolution of LLMs	Capabilities and Applications	Advancements in LLM Methodologies
Architecture	Transformer-based	Transformer Recurrent	Experiential Co-Learning
Primary Application	Optimization	Legal Analysis	Multi-agent Collaboration
Key Challenge	Incremental Learning	Analytical Skills	Environmental Impact

Table 1: This table provides a comparative analysis of the key features of Large Language Models (LLMs) across three main dimensions: their development and evolution, capabilities and applications, and advancements in methodologies. It highlights the architectural designs, primary applications, and key challenges associated with each dimension, offering insights into the ongoing evolution and diverse applications of LLMs.

4 AI Agents and Autonomous Agents

4.1 Characteristics and Functionalities of AI Agents

AI agents are computational entities designed to perceive environments, process information, and execute actions to achieve specific goals. They are characterized by autonomy, adaptability, and the ability to interact with other agents and humans through various modalities. A dual-modality framework enhances the flexibility of autonomous robots, accommodating diverse user needs and environmental contexts [62].

The architecture of AI agents typically includes modules such as a Profiling Module for defining roles, a Memory Module for information management, a Planning Module for decision-making, and an Action Module for executing decisions [27]. This modular design allows AI agents to perform complex tasks using machine learning techniques for enhanced predictions and decision-making [32].

AI agents operate under a notion of artificial agency, functioning without true intelligence, relying on predefined algorithms and data-driven learning processes [63]. This understanding is crucial for recognizing the limitations and potential of AI agents in real-world applications.

Their flexibility and adaptability are particularly evident in agent-based modeling, allowing them to manage complexity more effectively than traditional methods [64]. In cooperative multi-robot systems, each robot's heterogeneous capabilities necessitate coordinated actions to achieve shared goals [65].

AI agents can handle multiple user requests simultaneously, reducing wait times and improving user satisfaction [66]. However, the cooperation capabilities of LLM-augmented autonomous agents (LAAs) are limited in complex environments, hindering effective collaboration with both human and AI counterparts [67].

These agents are distinguished by their open-ended skill repertoires, enabling growth and adaptation over time, enhancing their effectiveness in dynamic environments [68]. The development of high-quality open-source agent models advances AI agents, providing a structured approach to improving functionalities and applications [69].

AI-generated, for reference only.

4.2 Autonomous Agents and Human-like Reasoning

Autonomous agents exhibit significant capabilities in emulating human-like decision-making and reasoning, essential for effective operation in dynamic environments. They utilize advanced planning and adaptation mechanisms to navigate their surroundings, often outperforming static planning methods. The AdaPlanner framework illustrates this adaptability by adjusting plans in real-time, enhancing performance in complex tasks [70].

The concept of human-like reasoning in autonomous agents is enriched by integrating socio-cultural elements, as seen in Vygotskian autotelic agents designed to reflect human socio-cultural situatedness in their decision-making processes [68]. By internalizing cultural norms and utilizing language as a cognitive tool, these agents achieve reasoning levels that closely mirror human cognition.

The development of autonomous agents capable of human-like reasoning is further enhanced through dynamic interactions with humans and other agents, leveraging large language models (LLMs) to improve cognitive functions and problem-solving skills. These interactions facilitate knowledge acquisition and enable agents to internalize experiences, fostering professional-level competencies and enhancing effectiveness in real-world applications [71, 43, 72, 20, 68]. This learning process involves refining decision-making strategies and incorporating environmental feedback, enhancing agents' adaptability to evolving contexts.

Investigations into autonomous agents powered by LLMs highlight their potential for human-like decision-making and complex reasoning, indicating significant capabilities for functioning in diverse environments that require advanced problem-solving. As LLMs evolve, they increasingly handle tasks across various fields, such as customer service and healthcare, while addressing challenges like multimodality and value alignment. This progress suggests that autonomous agents may soon play crucial roles in daily life, assisting in routine communications and intricate diagnostics [72, 33, 43]. Advancements in adaptive planning and socio-cultural reasoning frameworks represent significant strides in developing more sophisticated and human-like AI systems.

4.3 Integration of LLMs in AI Agents

The integration of Large Language Models (LLMs) into AI agents has significantly advanced their reasoning, decision-making, and interaction capabilities, resulting in more sophisticated systems. Domain-Specific Language Model Adaptation enhances this integration by customizing LLMs for specialized tasks, improving their applicability and effectiveness across various domains [11].

In autonomous systems, LLMs have been integrated into Spacecraft Operators (LLM-SO), enhancing autonomous agents' capabilities by interpreting mission telemetry and generating real-time control commands [17]. This integration exemplifies how LLMs provide real-time insights and decision-making support, improving operational efficiency in complex environments.

LLMs also model market agents in controlled experimental settings, offering AI-driven perspectives that simulate competitive market dynamics and explore strategic interactions among agents [61]. This application highlights LLMs' potential to model complex economic behaviors and enhance AI agents' strategic decision-making processes.

Furthermore, LLMs are leveraged in process mining tasks, generating insights from complex data and applying various prompting strategies to enhance AI agents' interpretability and efficiency in data-driven environments [22]. This versatility underscores LLMs' ability to handle large datasets and extract meaningful patterns that inform agent actions.

Through these integrations, LLMs enhance AI agents' core functionalities and expand their applicability across various domains, driving innovation and improving AI-driven systems' efficiency. The ongoing advancement of LLM methodologies is set to significantly enhance AI agents' capabilities, empowering them to tackle increasingly intricate challenges and engage in more intuitive interactions. As LLMs evolve, they are being integrated into autonomous systems that leverage their cognitive abilities for complex problem-solving across sectors, including software engineering and professional services. This integration fosters improved planning, resource management, and collaboration among multiple agents, enhancing overall robustness and scalability. The development of frameworks like Professional Agents (PAgents) aims to create specialized, interactive agents capable of performing at a professional level, thereby pushing the boundaries of AI's role in real-world applications [72, 41, 42, 43].

5 Multi-Agent Systems

5.1 Cooperative Multi-Agent Systems

Cooperative multi-agent systems represent an advanced paradigm where agents collaboratively work towards shared goals, leveraging individual strengths to enhance collective efficiency and performance. The CoPlanner framework exemplifies this by allowing agents to concentrate on specific problem aspects, thereby optimizing resource allocation and task execution [73]. Similarly, the MACNET framework underscores the importance of high-quality collaboration, yielding superior problem-solving and resource utilization [74]. These frameworks facilitate the development of functional communication systems that enable agents to autonomously evolve flexible language use, essential for seamless interactions and adaptation to dynamic conditions [38].

The BOLAA framework further enhances multi-agent capabilities by structuring LLM-augmented autonomous agents (LAAs) development in stages, ensuring effective contributions towards collective goals [75]. In multi-robot task planning, benchmarks assess communication frameworks' effectiveness, focusing on task success rates and token efficiency [65]. The KoMA framework demonstrates LLM-driven agents collaborating in complex driving environments, employing knowledge-driven strategies to navigate dynamic scenarios [25].

These systems have vast potential across applications, from educational technology to simulating opinion dynamics. They automate tasks and enhance user interactions, showcasing LLMs' transformative potential in multi-agent environments [76]. Innovative applications include LLM-powered agents automating data analysis and design, such as creating animated data videos from raw data [18].

5.2 Agent-Based Modeling and Simulation

Agent-based modeling (ABM) and simulation are crucial for exploring complex interactions among autonomous agents, providing insights into emergent phenomena. AgentTorch represents a significant advancement, enhancing scalability and flexibility in simulations [77]. It supports flexible agent architecture specifications and fully differentiable environments, enabling precise calibration and analysis of behaviors.

ABM effectively simulates social systems, organizational dynamics, and market dynamics, offering a platform for testing hypotheses about interactions and decision-making processes. Frameworks like AgentTorch model millions of agents and interactions within realistic environments, leveraging LLMs to capture adaptive behaviors and conduct comprehensive analyses, such as public health during the COVID-19 pandemic [43, 36, 18, 77, 30]. This capability is vital in economics, allowing simulation of market dynamics and assessment of policy impacts on behavior.

ABM also explores decentralized decision-making, where agents use local information to yield complex global patterns. This enhances understanding of adaptive behaviors in dynamic environments, integrating LLMs to improve simulation realism and effectiveness. For instance, AgentTorch integrates LLMs to model intricate scenarios like the COVID-19 pandemic, providing insights into isolation and employment behaviors' interplay on health and economic outcomes [43, 77]. This integration is crucial for understanding individual decisions' contributions to system-level dynamics.

The synergy between ABM and frameworks like AgentTorch enhances large-scale simulation capabilities, allowing exploration of diverse scenarios, including behavioral dynamics' impacts on health and economic outcomes. By leveraging LLMs, AgentTorch captures intricate adaptive behaviors, improving complex system modeling realism and enhancing ABM's utility in policy-making and scientific discovery [43, 77, 20, 69, 30]. This scalability is crucial for studying complex systems with numerous interacting agents.

As illustrated in Figure 7, the hierarchical structure of agent-based modeling highlights key frameworks, applications, and challenges. This figure emphasizes the role of frameworks like AgentTorch and AgentBench in advancing scalability and realism, while also addressing applications in social systems, economic dynamics, and public health. Challenges such as scalability, realism, and complex interactions are also depicted, underscoring the multifaceted nature of ABM.

Agent-based modeling and simulation provide a robust framework for analyzing multi-agent systems' dynamics, offering insights into collective behaviors' mechanisms and factors influencing system

AI-generated, for reference only.

stability and change. Advancements like AgentTorch enable simulation of millions of agents with high-resolution adaptive behaviors, particularly through LLM integration. This approach enhances interaction realism and facilitates exploration of complex phenomena, such as isolation and employment impact during the COVID-19 pandemic. Moreover, LLM-based agents simulating opinion dynamics reveal potential and limitations, highlighting biases affecting consensus-building on societal issues. These developments underscore ABM's transformative potential in policy-making and scientific discovery, emphasizing the need for continuous refinement to accurately reflect human behavior [37, 77]. Ongoing development of frameworks like AgentTorch promises to expand ABM's capabilities, enabling more detailed simulations of multi-agent environments.

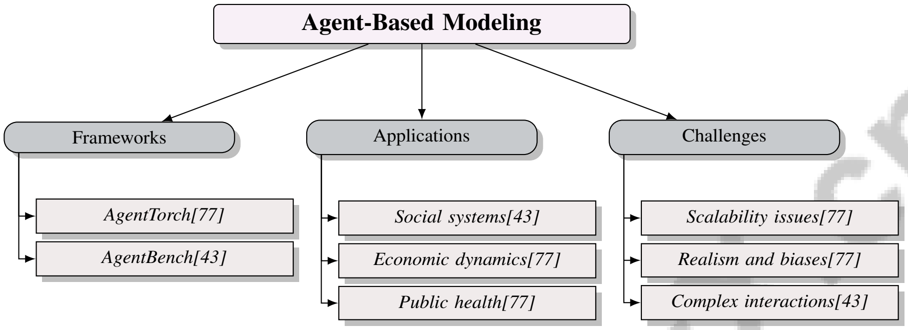


Figure 7: This figure illustrates the hierarchical structure of agent-based modeling, highlighting key frameworks, applications, and challenges. It emphasizes the role of frameworks like AgentTorch and AgentBench in advancing scalability and realism, while also addressing applications in social systems, economic dynamics, and public health. Challenges such as scalability, realism, and complex interactions are also depicted.

5.3 Communication and Coordination Strategies

Effective communication and coordination in multi-agent systems are crucial for collaboration and task execution. Developing communication strategies enables agents to share information, negotiate roles, and synchronize actions. Establishing shared language or protocols enhances interactions between human and AI agents, fostering dialogue that allows AI to adapt and seek clarification when necessary. This structured framework improves coordination and collaboration, leading to more efficient human-AI partnerships [38, 71, 72].

Agents' ability to autonomously develop communication methods highlights their potential for adaptive and flexible strategies. This capability allows agents to align interactions with contextual nuances and task-specific demands, enhancing effectiveness in information retrieval and decision-making [72, 20, 78, 30]. By fostering environments where agents autonomously refine protocols, multi-agent systems achieve higher efficiency and adaptability.

Effective coordination involves strategic resource allocation and task scheduling, crucial for system performance. Advanced techniques, such as those employed by LLMs, optimize decision-making in complex environments. For example, LLMs adjust pricing and resource distribution in markets, and in multi-robot scenarios, they manage coordination challenges across agents. Communication frameworks, including centralized, decentralized, and hybrid models, improve task success rates and scalability, leading to effective multi-agent operations [46, 65, 79]. Techniques like task decomposition and role assignment ensure effective labor division and dependency management. Coordination mechanisms, such as auctions or negotiation protocols, facilitate conflict resolution and task allocation, maximizing collective utility.

Integrating LLMs into multi-agent systems enhances communication by equipping agents with advanced natural language processing capabilities, enabling comprehension and generation of human-like text. This advancement improves interactions across sectors like customer service and healthcare, addressing challenges like multimodality, value alignment, and biases. Techniques like prompting and in-context learning refine these capabilities, leading to resilient autonomous agents assisting in complex tasks from email responses to medical diagnoses [29, 43]. This integration enhances agents' ability to interact with users and other agents, particularly valuable in scenarios requiring collaboration or sophisticated communication skills.

5.4 Challenges and Innovations in Multi-Agent Systems

Multi-agent systems face challenges and innovations critical for advancing capabilities and applications. A significant challenge is reliance on specific environmental configurations, which may not generalize, potentially undermining communication strategies' robustness [38]. This limitation highlights the need for adaptable communication frameworks functioning in diverse environments.

LLMs' integration presents opportunities and challenges. While LLMs enhance decision-making by simulating outcomes and mitigating irreversible mistakes, as shown by the WMA approach in web navigation [80], they also introduce privacy and interaction protocol design challenges [3]. Innovative solutions are needed to ensure LLM-augmented agents operate securely and efficiently.

Agent-based modeling frameworks, such as AgentTorch, offer innovations by providing scalability to millions of agents and capturing adaptive behaviors using LLMs [77]. This scalability is essential for modeling complex dynamics, enabling exploration of scenarios and interactions. However, AI agents' dependency on human data and lack of adaptability remain critical challenges [81].

In interactive environments, architectures like AriGraph outperform existing memory architectures, supporting long-term memory and decision-making in LLM agents [82]. This innovation enhances cognitive capabilities, allowing informed decisions based on historical data and interactions.

Despite advancements, current architectures struggle to enable agents to recognize and respond to non-cooperative behaviors, limiting effectiveness in dynamic environments [67]. Additionally, algorithm rigidity and absence of psychological factors influencing human decision-making pose limitations [61]. Ongoing research is required to create more flexible and adaptive agent architectures.

LLMs' impact on labor markets, particularly in China, illustrates broader socio-economic implications, with potential to reverse labor demand trends and exacerbate structural unemployment [83]. This underscores the need for understanding multi-agent systems' societal impacts and developing strategies to mitigate negative effects.

Multi-agent systems face significant challenges and innovative solutions. Addressing LLMs and advanced modeling frameworks' challenges can enhance capabilities, allowing systems to autonomously tackle complex problems across domains like software engineering and healthcare. Leveraging collaborative strengths enables robust problem-solving, dynamic task decomposition, and external tool utilization, paving the way for fully autonomous, scalable, and trustworthy LLM-based multi-agent systems (LaMAS) capable of efficient real-world operation [41, 3, 42, 43].

6 Applications in Natural Language Processing and Conversational AI

6.1 Innovative Applications and Real-World Examples

Large Language Models (LLMs) are revolutionizing diverse fields through their versatile applications. The KoMA framework illustrates their utility in multi-agent systems, enhancing decision-making in highway traffic scenarios [25]. Similarly, the KSPDG challenge showcases LLMs' capacity to improve autonomous spacecraft operations [17]. In cybersecurity, LLM-based honeypots engage attackers in realistic interactions, strengthening security measures [21]. The KAgentSys framework integrates LLMs with relational databases, improving factual and time-sensitive query accuracy [20]. LLMs also excel in process mining, analyzing event logs from sectors like traffic management and healthcare [22].

These applications demonstrate LLMs' transformative potential, addressing unique challenges and enhancing efficiency across sectors [59, 55]. Figure 8 provides a visual representation of the diverse applications of LLMs across various fields, highlighting key frameworks and challenges in autonomous systems, cybersecurity, and information retrieval. This figure serves to reinforce the narrative by illustrating the interconnectedness of these innovative applications and their practical implications.

6.2 Enhancements in Language Understanding and Generation

Advancements in LLMs have significantly improved language understanding and generation. The FATE-LLM framework enhances fine-tuning in federated learning, maintaining data privacy and

AI-generated, for reference only.

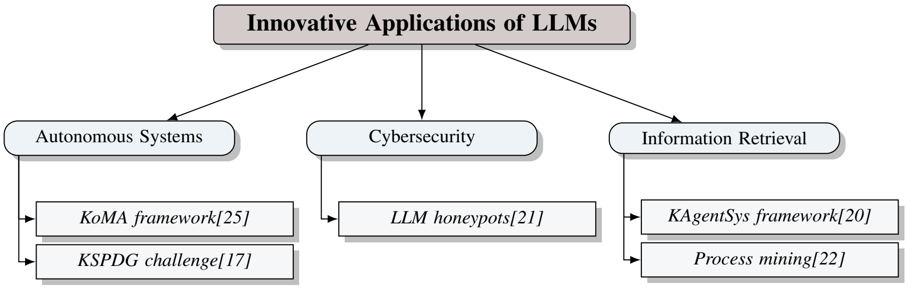


Figure 8: This figure illustrates the diverse applications of Large Language Models (LLMs) across various fields, highlighting key frameworks and challenges in autonomous systems, cybersecurity, and information retrieval.

adaptability [84]. LLM agents exhibit human-like polarization and social networking, simulating complex social dynamics [85]. Despite these advancements, challenges persist, such as ensuring reliability in critical areas like climate action [86]. Methodologies like AdaInfer improve inference efficiency and accuracy [87]. Incremental learning from user interactions allows for personalized models, enhancing language tasks [26]. Collaborative LLM systems, as seen in Corex, improve reasoning accuracy in complex tasks [88]. However, LLMs lack genuine understanding, necessitating improved communication frameworks [89, 90]. Detection methodologies using auxiliary classifiers enhance the identification of LLM-related harms [91]. These advancements underscore LLMs' potential in automating complex tasks and addressing ethical concerns [59, 36, 29, 31, 10].

6.3 Impact on Human-Machine Interaction

The integration of LLMs, AI agents, and multi-agent systems is transforming human-machine interaction, enabling sophisticated engagement across sectors like customer service and healthcare [72, 28, 51, 43]. LLMs' ability to process human-like language facilitates natural interactions, bridging computational processes with user-friendly interfaces. Effective interaction requires grounding symbols through embodied experiences and social interaction [92]. In smart homes and autonomous vehicles, AI agents equipped with language processing capabilities enhance collaboration and task execution [27, 51, 42, 43]. In education, LLMs act as virtual tutors, providing personalized learning experiences [34, 8]. Despite advancements, challenges in ensuring safety, reliability, and alignment with human values persist. Transparency and ethical implications require ongoing research and development [93, 51]. Developing intelligible communication frameworks is essential for enhancing AI systems' transparency and trustworthiness. The integration of LLMs into human-machine interaction signifies a pivotal advancement, facilitating autonomous task execution and transforming operational processes [51, 3]. By fostering natural communication, these technologies have the potential to revolutionize human-machine interactions, driving innovation and improving quality of life across various domains.

7 Challenges and Future Directions

The deployment of Large Language Models (LLMs), AI agents, and multi-agent systems presents a multifaceted array of challenges that require careful consideration. Ethical and social implications, particularly regarding bias, decision-making, and unintended consequences, demand thorough scrutiny to ensure responsible societal integration. The following subsections explore these challenges, emphasizing the need for a comprehensive understanding to facilitate equitable technology deployment.

7.1 Ethical and Social Challenges

The integration of LLMs, AI agents, and multi-agent systems introduces significant ethical and social challenges that necessitate proactive management. A critical concern is bias in training data, which can result in skewed outputs in sensitive areas like education and healthcare, exacerbating existing inequalities [11]. In dynamic environments such as autonomous driving, LLMs must address complex ethical considerations, balancing safety, privacy, and efficiency [25]. The cybersecurity domain also poses risks, as inadequately secured LLMs may inadvertently aid malicious activities, with LLM-based honeypots potentially being identified by attackers if they exhibit predictable behaviors [21].

LLMs' vulnerabilities, including susceptibility to attacks and reproducibility challenges, highlight the need for robust methodologies to enhance security and reliability [12]. The ethical landscape is further complicated by reliance on limited case studies, necessitating standardized approaches for reliable and ethical outputs.

Maintaining knowledge persistency across sessions and integrating new information without overfitting are ethical responsibilities to ensure LLMs' accuracy over time [26]. Limitations in numerical reasoning and prompt quality raise further ethical concerns about output reliability [17]. Ensuring scalability and accuracy in large-scale systems, such as graph reasoning, poses ethical challenges in managing communication and coordination among agents [60]. Benchmarking social intelligence among language agents emphasizes the need for effective and ethical interactions [19].

Addressing these challenges requires a balanced regulatory approach that fosters collaboration among ethicists, technologists, and policymakers. Prioritizing the development of transparent and interpretable AI systems can mitigate ethical challenges, enhance user understanding, and ensure technologies yield sustainable societal benefits. Implementing tailored ethical frameworks and dynamic auditing systems can further align AI systems with human values [94, 59, 51, 95, 55].

7.2 Computational and Resource Limitations

The deployment of LLMs, AI agents, and multi-agent systems is significantly hindered by computational and resource constraints, affecting scalability and accessibility. A core challenge is the extensive computing resources required for training and fine-tuning LLMs, which can be prohibitive for smaller organizations, limiting applications like automatic grading and feedback generation [24]. This issue is compounded by the distribution of high-quality data across isolated entities, restricting collaborative training efforts [84].

Traditional methods face computational limitations, necessitating innovative approaches to enhance efficiency [96]. The infeasibility of simulating adaptive agent behaviors at scale poses barriers to developing comprehensive agent-based models, as current frameworks struggle with efficiency in large-scale simulations [77]. Additionally, computational costs associated with extensive search processes can lead to slower execution times for tasks requiring thorough exploration [97].

In autonomous vehicles, the reliance on processed sensory data complicates real-time decision-making, as integrating LLMs may not accommodate the immediate visual perception required in dynamic environments [47]. Existing studies often lack a unified approach to multimodal inputs and interactions, further complicating technology deployment [13].

Despite the potential for reduced costs in data generation and improved performance on out-of-distribution tasks, computational demands remain a significant barrier. Techniques relying on a single forward pass struggle to extend to sequential generative tasks, limiting their applicability in complex scenarios [87]. Moreover, maintaining memory graphs in frameworks like AriGraph incurs substantial computational costs, impacting scalability and effectiveness in real-world applications [82].

The reliance on a limited sample of LLMs in studies can lead to subjective interpretations of their capabilities, highlighting the need for broader evaluations to ensure comprehensive insights into computational requirements [44]. Addressing these computational and resource limitations is crucial for advancing the capabilities and accessibility of LLMs, AI agents, and multi-agent systems, thereby facilitating their integration into diverse applications.

7.3 Data and Model Limitations

The effectiveness of LLMs, AI agents, and multi-agent systems is significantly constrained by inherent data and model limitations. A primary challenge is the reliance on extensive volumes of high-quality training data, often inaccessible to many researchers, hindering model diversity and robustness. Recent advancements in data augmentation techniques show promise in generating diverse training examples, yet challenges related to controllable and multimodal data augmentation persist [98, 99, 29]. The reliance on static datasets can lead to outdated models that fail to reflect evolving linguistic patterns and user interactions.

LLMs often overfit to specific tasks or domains due to narrow training data, complicating generalization across diverse contexts. This lack of adaptability is exacerbated by the absence of reliable techniques for steering their behavior and the difficulty experts face in interpreting their inner workings [99, 29]. The probabilistic nature of LLMs can result in incomplete or inaccurate outputs, especially in tasks requiring deep semantic comprehension, and their inability to effectively integrate multimodal data further limits applicability in complex real-world scenarios.

Biases in training data pose another critical concern, propagating through models and leading to skewed outputs that compromise the fairness and reliability of AI systems. Identifying and mitigating these biases is essential for ethical AI deployment. Furthermore, the opacity of LLM-based agents presents significant risks, complicating the assessment of their reliability and trustworthiness [13].

Scalability remains a significant limitation, as the computational resources required for training and deploying LLMs are substantial, posing barriers to widespread adoption and innovation. The constraints imposed by limited computational resources highlight the urgent need for developing more efficient model architectures and innovative training methodologies capable of functioning within existing infrastructure, especially considering ethical considerations, model biases, and interpretability challenges [29, 98, 30, 35, 99].

Addressing data and model limitations is crucial for advancing the capabilities of LLMs, AI agents, and multi-agent systems. By creating detailed risk assessment frameworks and pioneering solutions to emerging challenges, researchers can enhance the effectiveness, equity, and adaptability of these technologies, transforming expert domain knowledge into quantifiable features for improved predictive analytics and addressing complex cybersecurity issues [12, 55].

7.4 Scalability and Integration Challenges

Scalability and integration of LLMs, AI agents, and multi-agent systems present significant challenges that must be addressed for enhanced effectiveness in real-world applications. A primary issue is the complexity of managing state transitions and event handling within these systems, which can introduce substantial processing overhead [66]. This complexity often results in inefficiencies that hinder seamless AI system operations, particularly as they scale to handle larger datasets and more intricate interactions.

Current studies frequently struggle with scalability and adaptability in dynamic environments, underscoring the need for improved integration of feedback mechanisms within agent planning processes. This adaptability is crucial for maintaining the relevance and effectiveness of AI systems operating in diverse and changing contexts [33].

The challenge of scalability is further complicated by limitations in LLM-based conversational recommender systems (LLMCRS), where introducing new expert models can strain existing frameworks and impede adaptability [100]. As these systems evolve, developing strategies for integrating new models without compromising performance or scalability becomes increasingly important.

Trust in AI systems is another critical factor influencing their integration and scalability. Developing a comprehensive metric framework for assessing trust in AI agents is essential to maintain user confidence as these systems evolve and become more complex [101]. Maintaining user trust is vital for the successful deployment and adoption of AI technologies in real-world applications.

Moreover, existing benchmarks often fail to measure the nuanced and context-dependent nature of harmful outputs, leading to gaps in detection capabilities [91]. This limitation underscores the necessity for sophisticated evaluation frameworks that can accurately assess the performance and safety of AI systems as they scale and integrate into various domains.

Addressing the scalability and integration challenges of LLMs and multi-agent systems is essential for enhancing their capabilities as autonomous agents. Despite impressive performance across various natural language processing tasks, LLMs face significant hurdles in real-world applications due to unpredictability and limitations in domain-specific analytical tasks. By focusing on critical aspects such as planning, memory management, and resource efficiency, researchers can bridge the gap between theoretical advancements and practical implementations, fostering the development of robust LLM agents capable of navigating complex environments [30, 42]. Through innovative solutions and comprehensive evaluation frameworks, researchers can enhance the scalability, adaptability, and trustworthiness of these technologies, ensuring successful integration into diverse real-world environments.

7.5 Future Directions and Research Opportunities

The future development of LLMs, AI agents, and multi-agent systems presents numerous research opportunities aimed at overcoming current limitations and enhancing capabilities. A significant focus should be on refining LLM architectures and quantization strategies to improve performance and accessibility across diverse domains [24]. In autonomous driving, advancing frameworks like KoMA and exploring more complex scenarios could enhance training efficiency and decision-making capabilities [25].

In cybersecurity, there is a pressing need to develop robust LLMs tailored for cybersecurity applications by improving training methodologies and integrating them with traditional tools to bolster system defenses [12]. Enhancing the adaptability of models like the LLM-based honeypot to emerging threats and improving engagement with various attack patterns are crucial areas for future research [21].

Advancements in data augmentation methods and unsupervised learning techniques are essential to address challenges faced by LLMs, thereby enhancing generalization capabilities [11]. Improving user interaction and model adaptability, with potential enhancements in balancing model efficiency and user engagement, is another critical area of focus [26].

For space operations, refining fine-tuning processes using diverse datasets and exploring multimodal inputs are necessary to support complex decision-making tasks [17]. In storytelling and data visualization, integrating global optimization frameworks and human-in-the-loop systems with multimodal LLMs could significantly enhance narrative generation capabilities [18].

Expanding scenarios and refining evaluation processes in benchmarks such as AgentSense are essential for addressing current limitations and improving the assessment of social intelligence in AI agents [19]. Additionally, future research could focus on enhancing the integration of diverse knowledge sources and improving system adaptability to various user contexts, thereby elevating the utility of LLMs in information retrieval and knowledge management tasks [20].

In process mining, future research should refine benchmarks to include more diverse datasets and tasks, exploring additional LLMs and their applications to enhance process analysis and optimization [22]. By addressing these research opportunities, the future development of LLMs, AI agents, and multi-agent systems is poised to be transformative, with the potential to revolutionize various industries and enhance human-machine interaction.

8 Conclusion

Large Language Models (LLMs), AI agents, and multi-agent systems are pivotal in pushing the boundaries of AI technologies, demonstrating significant advancements across diverse fields. The profound impact of LLMs on natural language processing and conversational AI is evident, as they offer enhanced capabilities in language comprehension and generation. Their integration into AI agents has led to improved reasoning and decision-making, resulting in more adaptive and intelligent systems. Multi-agent systems, leveraging LLMs, have shown potential in simulating intricate interactions and enhancing communication and coordination among agents.

Despite these advancements, challenges remain, particularly concerning biases, ethical issues, and the inherent limitations of LLM-based systems. Addressing transparency in LLMs requires a nuanced approach that considers their unique characteristics and the varied requirements of stakeholders. In the realm of cybersecurity, LLMs hold considerable promise for advancing practices, though further research is needed to address their current limitations and enhance their applicability.

Looking ahead, there are abundant research opportunities. LLMs are well-positioned to refine data augmentation processes and model contexts more effectively through natural language. The development of dual-modality frameworks is anticipated to enhance the naturalness of human-robot

AI-generated, for reference only.

interactions, marking a significant stride in AI technology. Furthermore, the integration of LLMs with optimization algorithms has shown to significantly enhance problem-solving efficiency and effectiveness, indicating promising directions for future research.

References

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[2] Nils Körber, Silvan Wehrli, and Christopher Irrgang. How to measure the intelligence of large language models?, 2024.

[3] Yingxuan Yang, Qiuying Peng, Jun Wang, Ying Wen, and Weinan Zhang. Llm-based multi-agent systems: Techniques and business perspectives, 2024.

[4] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and Bing Shi. Chatgpt for shaping the future of dentistry: The potential of multi-modal large language model, 2023.

[5] Jinxin Liu, Shulin Cao, Jiaxin Shi, Tingjian Zhang, Lunyiu Nie, Linmei Hu, Lei Hou, and Juanzi Li. How proficient are large language models in formal languages? an in-depth insight for knowledge base question answering, 2024.

[6] Zihan Liu, Ruinan Zeng, Dongxia Wang, Gengyun Peng, Jingyi Wang, Qiang Liu, Peiyu Liu, and Wenhai Wang. Agents4plc: Automating closed-loop plc code generation and verification in industrial control systems using llm-based agents, 2024.

[7] Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

[8] Samuel Rhys Cox. The use of multiple conversational agent interlocutors in learning, 2023.

[9] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models, 2023.

[10] Séamus Lankford and Andy Way. Leveraging llms for mt in crisis scenarios: a blueprint for low-resource languages, 2024.

[11] Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260, 2024.

[12] Jie Zhang, Haoyu Bu, Hui Wen, Yongji Liu, Haiqiang Fei, Rongrong Xi, Lun Li, Yun Yang, Hongsong Zhu, and Dan Meng. When llms meet cybersecurity: A systematic literature review, 2024.

[13] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, Yunjun Gao, Yingcai Wu, and Shouling Ji. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents, 2024.

[14] Sinan Abdulhak, Wayne Hubbard, Karthik Gopalakrishnan, and Max Z. Li. Chatatc: Large language model-driven conversational agents for supporting strategic air traffic flow management, 2024.

[15] Jesse Wright. Here's charlie! realising the semantic web vision of agents in the age of llms, 2024.

[16] Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. Clembench: Using game play to evaluate chat-optimized language models as conversational agents, 2023.

[17] Victor Rodriguez-Fernandez, Alejandro Carrasco, Jason Cheng, Eli Scharf, Peng Mun Siew, and Richard Linares. Language models are spacecraft operators, 2024.

[18] Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. From data to story: Towards automatic animated data video creation with llm-based multi-agent systems, 2024.

[19] Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios, 2024.

[20] Haojie Pan, Zepeng Zhai, Hao Yuan, Yaojia Lv, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. Kwaiagents: Generalized information-seeking agent system with large language models, 2024.

[21] Hakan T. Otal and M. Abdullah Canbaz. Llm honeypot: Leveraging large language models as advanced interactive honeypot systems, 2024.

[22] Alessandro Berti and Mahnaz Sadat Qafari. Leveraging large language models (llms) for process mining (technical report), 2023.

[23] Jinyang Wu, Feihu Che, Xinxin Zheng, Shuai Zhang, Ruihan Jin, Shuai Nie, Pengpeng Shao, and Jianhua Tao. Can large language models understand uncommon meanings of common words?, 2024.

[24] Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. Investigating automatic scoring and feedback using large language models, 2024.

[25] Kemou Jiang, Xuan Cai, Zhiyong Cui, Aoyong Li, Yilong Ren, Haiyang Yu, Hao Yang, Daocheng Fu, Licheng Wen, and Pinlong Cai. Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models, 2024.

[26] Juhao Liang, Ziwei Wang, Zhuoheng Ma, Jianquan Li, Zhiyi Zhang, Xiangbo Wu, and Benyou Wang. Online training of large language models: Learn while chatting, 2024.

[27] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents, 2025.

[28] Saurabh Pahune and Manoj Chandrasekharan. Several categories of large language models (llms): A short survey. *arXiv preprint arXiv:2307.10188*, 2023.

[29] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3, 2023.

[30] Hao Kang and Chenyan Xiong. Researcharena: Benchmarking large language models' ability to collect and organize information as research agents, 2025.

[31] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024.

[32] Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems, 2018.

[33] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[34] Menna Fateen and Tsunenori Mine. Developing a tutoring dialog dataset to optimize llms for educational use, 2024.

[35] Samyar Janatian, Hannes Westermann, Jinzhe Tan, Jaromir Savelka, and Karim Benyekhlef. From text to structure: Using large language models to support the development of legal expert systems, 2023.

[36] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review, 2024.

[37] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of llm-based agents, 2024.

[38] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[39] Thilini Wijesiriwardene, Amit Sheth, Valerie L. Shalin, and Amitava Das. Why do we need neuro-symbolic ai to model pragmatic analogies?, 2023.

[40] Haoyi Xiong, Zhiyuan Wang, Xuhong Li, Jiang Bian, Zeke Xie, Shahid Mumtaz, Anwer Al-Dulaimi, and Laura E. Barnes. Converging paradigms: The synergy of symbolic and connectionist ai in llm-empowered autonomous agents, 2024.

[41] Junda He, Christoph Treude, and David Lo. Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead, 2024.

[42] Chris Sypherd and Vaishak Belle. Practical considerations for agentic llm systems, 2024.

[43] Saikat Barua. Exploring autonomous agents through the lens of large language models: A review, 2024.

[44] Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities, 2023.

[45] Gaurav Singh and Kavitesh Kumar Bali. Enhancing decision-making in optimization through llm-assisted inference: A neural networks perspective, 2024.

[46] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey, 2024.

[47] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles, 2023.

[48] Yingzhe Peng, Xiaoting Qin, Zhiyang Zhang, Jue Zhang, Qingwei Lin, Xu Yang, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Navigating the unknown: A chat-based collaborative interface for personalized exploratory tasks, 2024.

[49] Linus Nwankwo and Elmar Rueckert. The conversation is the command: Interacting with real-world autonomous robot through natural language, 2024.

[50] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, and Julian Eggert. A glimpse in chatgpt capabilities and its impact for ai research, 2023.

[51] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[52] Chen Qian, Yufan Dang, Jiahao Li, Wei Liu, Zihao Xie, Yifei Wang, Weize Chen, Cheng Yang, Xin Cong, Xiaoyin Che, Zhiyuan Liu, and Maosong Sun. Experiential co-learning of software-developing agents, 2024.

[53] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey, 2024.

[54] Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. Graphllm: Boosting graph reasoning ability of large language model, 2023.

[55] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[56] Zongyue Qin, Chen Luo, Zhengyang Wang, Haoming Jiang, and Yizhou Sun. Relational database augmented large language model, 2024.

AI-generated, for reference only.

[57] Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. Large language models are good multi-lingual learners : When llms meet cross-lingual prompts, 2024.

[58] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[59] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[60] Yuwei Hu, Runlin Lei, Xinyi Huang, Zhewei Wei, and Yongchao Liu. Scalable and accurate graph reasoning with llm-based multi-agents, 2024.

[61] Jingru Jia and Zehua Yuan. An experimental study of competitive market behavior through llms, 2024.

[62] Linus Nwankwo and Elmar Rueckert. Multimodal human-autonomous agents interaction using pre-trained language and visual foundation models, 2024.

[63] Luciano Floridi. Ai as agency without intelligence: On chatgpt, large language models, and other generative models. *Philosophy & technology*, 36(1):15, 2023.

[64] Ankit Chaudhary and Jagdish L. Raheja. A formal approach for agent based large concurrent intelligent systems, 2011.

[65] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems?, 2024.

[66] Antonio A. Ginart, Naveen Kodali, Jason Lee, Caiming Xiong, Silvio Savarese, and John Emmons. Asynchronous tool usage for real-time agents, 2024.

[67] Manuel Mosquera, Juan Sebastian Pinzon, Manuel Rios, Yesid Fonseca, Luis Felipe Giraldo, Nicanor Quijano, and Ruben Manrique. Can llm-augmented autonomous agents cooperate?, an evaluation of their cooperative capabilities through melting pot, 2024.

[68] Cédric Colas, Tristan Karch, Clément Moulin-Frier, and Pierre-Yves Oudeyer. Language and culture internalisation for human-like autotelic ai, 2022.

[69] Jianguo Zhang, Tian Lan, Ming Zhu, Zuxin Liu, Thai Hoang, Shirley Kokane, Weiran Yao, Juntao Tan, Akshara Prabhakar, Haolin Chen, Zhiwei Liu, Yihao Feng, Tulika Awalgaonkar, Rithesh Murthy, Eric Hu, Zeyuan Chen, Ran Xu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Silvio Savarese, and Caiming Xiong. xlam: A family of large action models to empower ai agent systems, 2024.

[70] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplanner: Adaptive planning from feedback with language models, 2023.

[71] Nikhil Mehta, Milagro Teruel, Patricio Figueroa Sanz, Xin Deng, Ahmed Hassan Awadallah, and Julia Kiseleva. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback, 2024.

[72] Zhixuan Chu, Yan Wang, Feng Zhu, Lu Yu, Longfei Li, and Jinjie Gu. Professional agents – evolving large language models into autonomous experts with human-level competencies, 2024.

[73] Danqing Wang, Zhuorui Ye, Fei Fang, and Lei Li. Cooperative strategic planning enhances reasoning capabilities in large language models, 2024.

[74] Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large-language-model-based multi-agent collaboration, 2025.

[75] Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents, 2023.

[76] Irene Weber. Large language models as software components: A taxonomy for llm-integrated applications, 2024.

[77] Ayush Chopra, Shashank Kumar, Nurullah Giray-Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. On the limits of agency in agent-based models, 2024.

[78] Xinzhe Li and Ming Liu. Rethinking chatgpt's success: Usability and cognitive behaviors enabled by auto-regressive llms' prompting, 2024.

[79] Ryan Y. Lin, Siddhartha Ojha, Kevin Cai, and Maxwell F. Chen. Strategic collusion of llm agents: Market division in multi-commodity competitions, 2024.

[80] Hyungjoo Chae, Namyoung Kim, Kai Tzu iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation, 2024.

[81] Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. Richelieu: Self-evolving llm-based agents for ai diplomacy, 2024.

[82] Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Mikhail Burtsev, and Evgeny Burnaev. Arigraph: Learning knowledge graph world models with episodic memory for llm agents, 2024.

[83] Qin Chen, Jinfeng Ge, Huaqing Xie, Xingcheng Xu, and Yanqing Yang. Large language models at work in china's labor market, 2023.

[84] Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. Fate-llm: A industrial grade federated learning framework for large language models, 2023.

[85] Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P. Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents, 2025.

[86] Denis Havlik and Marcelo Pias. Common errors in generative ai systems used for knowledge extraction in the climate action domain, 2024.

[87] Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. Not all layers of llms are necessary during inference, 2024.

[88] Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration, 2024.

[89] Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024.

[90] A. Baskar, Ashwin Srinivasan, Michael Bain, and Enrico Coiera. A model for intelligible interaction between agents that predict and explain, 2024.

[91] Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E. Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M. Daly, Kirushikesh DB, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Nishtha Madaan, Sameep Mehta, Erik Miehling, Keerthiram Murugesan, Manish Nagireddy, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Tillmann, Aashka Trivedi, Kush R. Varshney, Dennis Wei, Shalisha Witherspooon, and Marcel Zalmanovici. Detectors for safe and reliable llms: Implementations, uses, and limitations, 2024.

[92] Sara Incao, Carlo Mazzola, Giulia Belgiovine, and Alessandra Sciutti. A roadmap for embodied and social grounding in llms, 2024.

[93] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

[94] Jon Chun and Katherine Elkins. Informed ai regulation: Comparing the ethical frameworks of leading llm chatbots using an ethics-based audit to assess moral reasoning and normative values, 2024.

[95] Zijun Liu, Kaiming Liu, Yiqi Zhu, Xuanyu Lei, Zonghan Yang, Zhenhe Zhang, Peng Li, and Yang Liu. Aigs: Generating science from ai-powered automated falsification, 2024.

[96] Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms, 2025.

[97] Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents, 2024.

[98] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024.

[99] Samuel R Bowman. Eight things to know about large language models. *arXiv preprint arXiv:2304.00612*, 2023.

[100] Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. A large language model enhanced conversational recommender system, 2023.

[101] Sivan Schwartz, Avi Yaeli, and Segev Shlomov. Enhancing trust in llm-based ai automation agents: New considerations and future challenges, 2023.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.