Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA37319

The Three-Stage Trajectory of Large Language Models: 2022 to 2026

Kunyu Bian[1], Ruikai Mao[2], Xiaofan Xu[3,*]

[1]No.9 High School, Harbin, China
[2]Middlefield Collegiate Institute, Markham, Canada
[3]Yew Chung International School, Hong Kong, China
**Corresponding Author. Email: 200081925@hkstudent.ycis.com*

Abstract. This literature review puts forward a three-stage model of Large Language Models (LLMs), delineates the research goals of LLM in three stages and predicts future trends. Against the rapid, fragmented advancement of large-language-model technologies, systematic sorting of their evolutionary trajectories helps researchers grasp core contradictions and shifting research priorities. It also offers actionable reference for subsequent technical exploration, practical deployment and policy formulation in the LLM domain. In Phase 1 (2022 to 2023), LLMs have improved reasoning and general capabilities, while facing conflicts between capability and reliability. In Phase 2 (2023 to 2025), LLMs enhance evaluation and robustness, while filling the gap between benchmarks and real-world performance. In Phase 3 (2025 to 2026), LLMs improve autonomous action and safe deployment, while balancing accessibility and security. This dilemma catalyzed Phase 4 - "World Model Era", emphasizing physical simulation and causal reasoning. This work synthesizes 12 milestone papers and sorts out the development process and future trend of LLM.

Keywords: Large Language Models, three-stage evolution, three-phase, model evaluation

1. Introduction

Large Language Models (LLMs) change from simple word processors into advanced artificial intelligence systems, in which they develop multiple abilities. Particularly, they can now perform logical reasoning well, communicate in a number of ways, and also use tools to perform complex tasks [1-3].

This rapid development of LLMs produces a large amount of research. However, the research in that field is divided into many areas, such as reasoning and coding, which seems unrelated, making it difficult to see a clear development path.

This paper divides LLM development into three-phases. The phases are based on research goals, key conflicts, and evaluation methods. Phase 1 covers from 2022 to 2023. Phase 2 covers from 2023 to 2025. Phase 3 covers from 2025 to 2026. These phases do not have strict time boundaries: they overlap with each other. In addition, problems in one phase can lead to new research in the next phase. For example, Phase 1 focused on finding new abilities through scaling laws [4]. At the same

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA37319

time, researchers were already concerned about hallucination and evaluation problems [5]. Later, dynamic evaluation showed the limits of passive LLMs [6, 7]. Researchers then started to study autonomous agents and safer agent skills [3, 8]. Therefore, the three-phases should not be seen as completely separate periods. They are connected by the problems found in each phase.

2. Phase 1: what can LLMs do

2.1. Core papers and key findings

The main goal of Phase 1 was to discover the abilities of LLMs. GPT-4 is an important example of this development: according to the technical evaluation, GPT-4 performed better than GPT-3.5 on several benchmarks in three aspects [2].

Firstly, GPT-4 performed well on the MMLU benchmark (MMLU tests model understanding in 57 fields). GPT-4 achieved 86.4%, while GPT-3.5 achieved 70.0%, and Flan-PaLM model achieved 75.2% [2]. GPT-4 was therefore 16.4 percentage points higher than GPT-3.5 and 10.8 percentage points higher than Flan-PaLM. This result shows that GPT-4 had better knowledge and understanding. On one hand, changes in model scale helped. On the other hand, training data and optimization methods played a role as well [4]. Secondly, GPT-4 performed well on reasoning tasks [2]. It achieved 95.3% on HellaSwag, while GPT-3.5 achieved 85.5%. GPT-4 also reached 96.3% on the AI2 Reasoning Challenge (ARC), which was 10.8 percentage points higher than GPT-3.5 [2]. These results show that GPT-4 had stronger reasoning abilities: it could understand information better as well as solve more complex tasks.

Thirdly, GPT-4 also improved code generation. It achieved 67.0% on HumanEval, and GPT-3.5 achieved 48.1% [2]. The result shows that GPT-4 could understand programming syntax better than the previous model because it can generate code more effectively. Therefore, GPT-4 is more useful for software development. Altogether, findings in those three aspects show that LLM development depended on more than model size: scaling laws were important, while training strategies were also important [4].

2.2. AGI potential conflicts with fundamental limitations

The progress of the LLM, GPT-4, led to discussions about Artificial General Intelligence (AGI). Bubeck et al. argued that GPT-4 showed early characteristics of AGI because of its strong performance across different tasks [9]. Other studies also showed the growing abilities of LLMs. Wei et al. showed that Chain-of-Thought prompting could improve reasoning. Zhao et al. discussed both the progress and problems of LLMs [1, 10].

However, these abilities did not mean that LLMs were always reliable. LLMs could produce false information. Specifically, GPT-4 did not achieve perfect scores on every benchmark, which shows that strong performance did not mean that it could solve every task correctly [2]. In addition, CoT prompting did not always produce reliable results. Its performance could also depend on model scale. Therefore, reliability rose to become a problem of LLMs. These problems created a conflict in Phase 1: LLMs were becoming more capable, but their performance was still not fully reliable. Researchers thus needed to look for a way to balance capability and reliability of LLMs.

DOI: 10.54254/2755-2721/2026.BA37319

2.3. Scaling laws and reasoning emergence

The main research focus of this phase is to improve the reasoning capability. Representative research shows that under the guidance of computing optimal scaling laws, through larger models and more effective prompting strategies, reasoning performance has been continuously improved. As reasoning ability becomes increasingly complex, researchers need a more systematic way to evaluate these new abilities [1, 4].

3. Phase 2: how well do LLMs perform

3.1. Core papers and evaluation shift: moving from static benchmarks to dynamic assessment

Zhao et al. proposed all-encompassing evaluation dimensions of reasoning, robustness and downstream applications [10]. GPT-4 Technical Report assessed the models on many benchmarks. Bubeck et al. have extended the scope of assessment to complex, real-life situations, and Wei et al. have also included the reasoning process in their evaluations [1, 2, 9]. Kaddour et al., Naveed et al. and Matarazzo and Torlone have also systematically integrated the above evaluation dimensions in an all-encompassing review [6, 7, 11]. According to the above research, benchmark accuracy alone cannot reflect the ability of new LLMs. Evaluation gradually developed into multidimensional assessment and capability analysis.

3.2. Synthesis of methodological conflicts: standardized scores vs. real-world failure modes

The GPT-4 Technical Report mainly used the standardised benchmark, while Bubeck et al. focused on real-world problem-solving ability [2, 9]. Wei et al. focused on reasoning ability; Zhao et al. proposed an all-encompassing evaluation system for many kinds of functions [1, 10]. At the same time, the essential studies have shown that high benchmark scores often hide serious production failure modes, such as hallucination and logical inconsistency [5, 6]. As shown by the above figures, evaluating high-end LLMs has become increasingly difficult.

3.3. Systematization, critical assessment, and the emergence of agentic demands

In addition to benchmark performance, others have also studied reasoning ability, reliability and real-world applicability more often. At the same time, research has been conducted to explore the shortcomings of models and hallucination mechanisms, as well as safety problems [5]. Exposing these static performance deficiencies naturally drove the demand for dynamic, autonomous agent execution in Phase 3 [3]. 4. Phase 3: Agentic Expansion From Passive Evaluation to Safe Governance in Dynamic Environments

4. Phase 3: how can LLMs act safely in dynamic environments

4.1. Core papers and capability shift: from text processing to autonomous agent execution

Building upon the dynamic evaluation frameworks established in Phase 2, Phase 3 is a fundamental paradigm shift from "passive LLM capability assessment" to "autonomous execution and safe deployment in dynamic environments" [3, 6, 10].

DOI: 10.54254/2755-2721/2026.BA37319

4.2. Ideological & technical conflicts: inclusive accessibility vs. sandboxed security governance

The dialogue of these foundational works indicates a deep-seated ideological and technical conflict in the Phase 3 research. Ahuja et al. point out the systemic multilingual deficiencies and advocate for universal accessibility to expand the benefits of AI inclusively [12]. Xu and Yan also put forward the problem of operating security and believe that an unchecked increase in agent capabilities creates a dangerous execution loophole [8]. Therefore, a serious dialectical contradiction has arisen: the pursuit of expansion in general-purpose capabilities directly conflicts with the requirement for strict security governance and sandbox restrictions.

4.3. Phase focus & transition: bridging the safety-fairness trade-off and the catalyst for world models

Together, these opposing positions define the core paradox of this period: fairness calls for wide accessibility, while safety demands restricted boundary enforcement. It has not been resolved, and thus the main research topic for Phase 3 is "accessibility versus security". Therefore, the current problem is expected to motivate a new approach in the future; that is, the difficulty of ensuring safety solely by adding rules after the fact requires moving away from static language representations, laying the theoretical foundation for physical grounding and internal causality in the emerging World Model Era.

5. Synthesis and prediction: a three-stage evolution model and future directions

5.1. The three-stage evolution model

All nine milestone papers point to the three-stage evolution model. Each phase could be defined with a central question with a distinct research focus:

Phase 1. The foundational question was "What can LLMs do?". Research was also based around the observation of LLMs' abilities unlocked through scaling and innovative prompting strategies, namely the Chain-of-Thought prompting demonstrated by Wei et al. [1].

Phase 2. As capabilities expanded further, the driving question changed to "How well do LLMs perform?". There are comprehensive surveys like Zhao et al. that proposed a structured evaluation framework covering reasoning, real-world applicability [10]. While Naveed et al. and Matarazzo and Torlone further unified it by providing holistic overviews with their inherent limitations, such as hallucination and logical inconsistency [7, 11].

Phase 3. The current stage of LLMs. With a fully mature understanding of capabilities and limitations, what else could be explored? The question evolved to "How can LLMs act safely in dynamic environments?" For example, the Multilingual Survey stressed an essential factor of fairness by showcasing the poor performance of the technology in non-English languages, requiring wide and inclusive distribution [12]. Meanwhile, Xu and Yan formulated the concept of "Agent Skills," noting that community-created agent skills have 26.1% of vulnerabilities and require a security-first governance model [8].

5.2. How conflicts drove the evolution

The evolution from one phase to another was not exactly smooth and linear.

During phase 1, the models were argued to be fabricating facts and lacked true understanding of the topic. Bubeck et al.'s qualitative evidence for AGI sparks contrasts with findings that CoT

DOI: 10.54254/2755-2721/2026.BA37319

prompting was actually brittle and dependent on model scale [9]. The conflict between "general capability" and "unreliable execution" makes it clear that personal performance was insufficient. This links to phase 2, where the entire research agenda started to build robust, multidimensional evaluation frameworks.

There is conflict in phase 2 as well, due to the evaluation itself. The initial standardized static benchmarks were challenged by survey papers, recommending dynamic, real-world-based assessments [6, 10]. The field of work became aware that the model may be able to ace a reasoning test but is never able to do so in a production setting. This "benchmark" and "deployment reliability" are the conflict of phase 2.

In phase 3, the conflict is the pressing unaddressed conflict of our time, the conflict between extensive accessibility and extensive security. The Multilingual Survey argues that linguistic borders must be torn down to provide services to all of humanity, which is an expanding idea [12]. Xu and Yan prove that free expansion of agents' abilities through community efforts creates security gaps [8]. The "accessibility vs. security" conflict powers the current research agenda and drives the future paradigm shift.

5.3. Predicting phase 4: the world model era

This three-phase contradiction leads to the new era, which is known as the "World Model Era" since this paradox of fairness and safety continues to exist due to the lack of knowledge about physical consequences in the real-world by an LLM model. Nevertheless, safe actions require prediction of potential outcomes by agents, which cannot be resolved by language comprehension alone. This means that there will be a need for world simulation models capable of comprehending cause-and-effect relations. It can be seen that the future research on LLMs will take this new path by shifting the focus onto world simulations. An environment that allows them to predict future states and make conclusions. This integrates sensory inputs like vision or audio with reinforcement learning in simulated environments, allowing LLMs to understand the real-world better than its language functions. Some examples of these are video generation models that could potentially work as world simulators and the focus of agent skills research on sandboxed executions [8]. These will provide some support for the prediction that LLMs need stronger abilities to understand changes in their environments. In this new Age of World Models era, LLMs will be judged based on how colloquial their reasoning is internally.

6. Conclusion

This literature review explored nine important works from an evolutionary conflict-driven perspective, thus showing the existence of a consistent three-stage development trajectory in LLM research. Phase 1 was marked by the discovery of reasoning skills driven by scaling and Chain-of-Thought prompting. Phase 2 was seen for the trend towards systematic evaluation, as comprehensive surveys replaced personal observations and showed significant gaps in the ability to be robust and truthful. Phase 3 was associated with the agentic trend, where the emphasis was on secure deployment in dynamic environments, which created a deep conflict between linguistic fairness and security. This shows that the progress of the field is not only a result of technological advances but of the tension created by opposing views.

The key value added by this paper is the development of an integrated understanding of a disjointed literature by formulating a model for the evolution of the literature through three distinct stages, wherein internal conflicts consistently advance the research agenda. This gives the researcher

DOI: 10.54254/2755-2721/2026.BA37319

a conceptual map not only of where previous studies stand but also where the literature is heading, specifically in this paper's prediction of a future World Model Era. However, the major drawback of this paper lies in the selective inclusion of only nine milestones, providing in-depth analysis of the selected papers but possibly missing out on other equally significant perspectives. Furthermore, the quick rate of change in LLMs makes it difficult to draw rigid lines between the phases and the prediction is still purely speculative.

Author contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 35*, 24824–24837.
[2] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.
[3] Wang, L. et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345.
[4] Hoffmann, J. et al. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
[5] Ji, G. et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.
[6] Kaddour, J. et al. (2023). Challenges and applications of large language models. arXiv preprint arXiv:2307.10169.
[7] Matarazzo, A., & Torlone, R. (2025). A survey on large language models with some insights on their capabilities and limitations. arXiv preprint arXiv:2501.04040.
[8] Xu, R., & Yan, Y. (2026). Agent skills for large language models: Architecture, acquisition, security, and the path forward. arXiv preprint arXiv:2602.12430.
[9] Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.
[10] Zhao, W. X., et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
[11] Naveed, H., et al. (2025). A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol., 16*(5), Art. no. 106.
[12] Ahuja, K., Diddee, H., Oregi, R., Sitaram, S., & Choudhury, M. (2024). A survey of multilingual large language models. *Patterns, 5*(12), p. 101111.