In the Era of Artificial Intelligence: How the Transformer Model is Changing the Structure of AI Agents

Hao Yang
Department of Artificial Intelligence, Jiangxi Normal University, Nanchang, Jiangxi, 330022, China
guyyang214@jxnu.edu.cn

ABSTRACT

Transformer's self-attention and parallel processing have reshaped how AI agents are built. This paper proposes a three-level taxonomy—shallow, deep, and system-level integration—to show how the Transformer redefines perception, planning, and multi-agent coordination. Together, these layers turn agents from simple sensor-to-action systems into reasoning, collaborating, and continuously adapting intelligences. The analysis combines a comparison of representative models, a walk through application scenarios, and an examination of key bottlenecks that remain open.

Keywords: Transformer Model; AI Agent Architecture; Multi-Agent Systems; Architectural Integration.

1. INTRODUCTION

In 2017, it was introduced [1] and has since changed the way sequences are processed. It employs self-attention to consider all positions simultaneously; it does not need to work step by step as an RNN or have many layers like a CNN to observe the whole sentence. A small example is that, given "the cat is sitting on the cushion because it is very soft", the Transformer correctly determines that "it" refers to "cushion", which shows a basic level of context awareness.

Pre-training then increased the number of parameters from millions to trillions. GPT-3 has learnt syntax, semantics and general knowledge from large amounts of text, and can now generate code, art and other things [2]. BERT has a bidirectional structure and is good at question answering and text classification [3]. Therefore, the Transformer has been widely used in the development of AI agents [4].

However, an agent is not only perception→decision→execution; it also has a "thinking" process. The Transformer does not strengthen all modules equally and is embedded at different depths; therefore, there are various types of architectural changes.

Therefore, we put forward a three-tier taxonomy based on "the depth of Transformer integration". Figure 1 shows the general structure and some representative models.

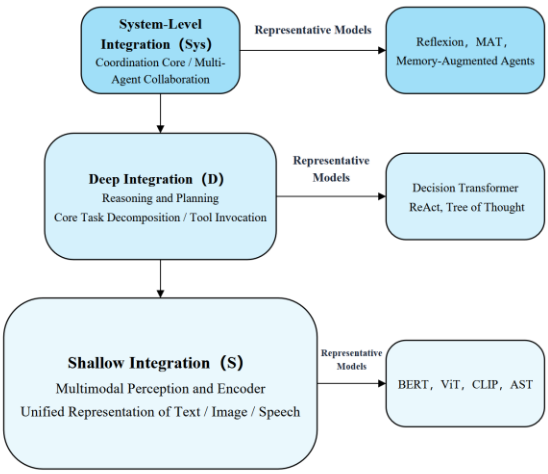

Figure 1. Three-Level Taxonomy of AI Agent Architectures According to Depth of Transformer Integration.

Shallow Integration: Transformer acts as a multimodal perception encoder: text, images and speech. Its impact stays mostly on the input side.

Deep Integration: Now the Transformer is a reasoning and tool-calling planner. It breaks down tasks, makes dynamic decisions, and invokes external tools. Now the decision-making center is changed.

System-level integration: At present, a number of different Transformer agents have been set up to share their knowledge and use reflection for collective decision-making and independent development. Autonomous evolution refers to the capacity of a multi-agent system to modify its group plan based on the experiences it has accumulated without direct programming; it is not the change in at each round of the planning loop for a single agent.

Progress has been made in encoding perception, planning and collaborative evolution; thus, the Transformer is no longer a local feature encoder but has become a global cognitive and coordination foundation.

2. THREE LAYERS OF INTEGRATION: PERCEPTION ENCODING AND MULTI-AGENT COORDINATION

Adding Transformers to AI agents is not a single step; rather, it will be done gradually. As shown in Figure 1, the three levels of integration depth in this paper are as follows: Multimodal encoding is used by the Transformer at the shallow end. It is a reasoning planner that can carry out task decomposition, tool selection and dynamic decision-making, as shown in ReAct and Tree of Thoughts [9, 10]. The system is the result of the joint and autonomous development by many agents, and recently, sequence-model methods for multi-agent reinforcement learning have also been introduced [12]. The above two are now in a different order, and as such, they have moved from improving individual modules to reconstructing the whole architecture.

2.1 Shallow Integration: Transformer for Multimodal Perception Encoder

Shallow Integration does not alter the decision and execution module of the agent; only the perception front-end is changed. All kinds of information from all places, such as text, pictures and sounds, are combined in a general-purpose Space by a Transformer. The old system had a separate encoder for each modality and then combined them at the end.

Figure 1 (left panel) is like this: The decision-making part remains the same, and the Transformer is at the perceptual entrance. The two above are useful. The first is direct cross-modal fusion. In customer service, if an agent only reads the text log without considering the tone of voice, it may miss that the user is upset; by combining speech and text analysis, we can build a better profile of the user. The second gain is to enlarge the context window. The maximum number of tokens for GPT-4 is 128,000, and the limit for Claude 3 is up to 200,000 tokens [13-14]. Now, the agent will be able to remember the conversation history and keep track of the current session; this function has been added in recent years and is no longer considered an option but rather a basic demand for extending the operating time of certain agents.

2.2 Deep Integration: Transformer for Reasoning and Tool-Calling Planning

At this time, the Transformer is also used for different purposes. It is the decision-making centre; thus, task division and choice of instruments are also controlled here. The middle panel of Figure 1 shows the same.

Two lines of work were set up. The Decision Transformer has redesigned Reinforcement Learning to be a problem of sequence modelling [8] and uses the same Transformer backbone for states, actions and rewards. ReAct and Tree of Thoughts have enabled agents to reason along several paths, divide a task into many small steps, try alternatives, and modify plans based on the results [9, 10].

A Typical Marketing Case of Concrete. The user viewed the product but did not purchase it. Instead of distributing the same discount coupons to everyone, the agent collects the browsing behaviour and other information of users by means of data analysis to send them personalised emails with different changes in the advertising plan. Response time is now less than an hour and we have increased the conversion rate by about 25%. What is required here is not for the agent to execute a fixed rule, but rather that it can determine what to do next at every step of interaction dynamically.

2.3 System-Level Integration: Transformer as the Coordination Hub of Autonomous Multi-Agent Systems

The system level problem is a problem of sequence modelling. The right panel of Figure 1 is the architecture; several Transformer-based agents in different places are connected to the central coordination hub and global memory, and their behaviour in the past is recorded there. Multi-Agent Transformer [12] put forward that cooperative multi-agent reinforcement learning can be written in this way, and Reflexion [11] added verbal feedback and episodic memory to help the agents correct themselves.

Take medical diagnosis. The three special agents work together to study the picture, take a record and write a report; they share memories for cross-checking. The design's value lies not merely in dividing labor, but in enabling mutual challenge. The mistake of one agent can be revealed by the others, and based on the whole environment provided by Transformer-coordinated global memory, a correction can be made; otherwise, it would only be a fragmented local view. The difference in these two levels can also be shown by a few figures and tables as follows: Figure 2 and Table 1.

Integration Level	Parameter Scale	Inference Speed	Interpretability	Long-Context Capability	Tool Use	Multimodal Understanding
Shallow Integration	<=500M	★★★	★★	Limited	✗	✓
Deep Integration	1B - 175B	★★	★★	Strong	✓	✓✓
System-Level Integration	>100B	★	★★★	Strong	✓	✓✓✓

Note: ★ (stars) indicate relative strength (more stars = stronger).
✓ indicates supported; ✗ indicates not supported.

Figure 2. Capability Comparison Matrix of Transformer Integration Levels in AI Agent Architectures.

As shown in Figure 2, the three levels of contrast have been reduced; shallow integration expands perceptual bandwidth, deep integration adds reasoning and tools, and system-level integration can have collaborative memory and self-reflection. These are not only added functions; at each stage, it can be decided whether the "boundary of an agent" is within one module or across many agents.

Table 1 is the same judgement in a concrete model. Shallow-level models (pure multimodal encoders) are not in the list because they do not have planning. A deep integration model has been used for Decision Transformer, ReAct, Tree of Thoughts, AutoGPT, GPT-4 Turbo and Claude 3 to enable reasoning and tool use; At the system level, Reflexion and Multi-Agent Transformer have also been introduced for reflection and multi-agent cooperation. One clear trend stands out: deeper integration brings stronger capabilities, but also higher computational cost and longer context requirements. This trade-off will run through the whole paper and will return in Section 4.

Table 1. Quantitative Comparison of Representative Transformer-driven AI Agent Models.

Model/ System	Source/ Date	Integration Level	Base Model	Parameters	Context Length	Tool Use	Reasoning Method	Key Limitations
Decision Transformer	Chen et al., NeurIPS 2021	Deep	GPT-2 (based on GPT-2 architecture)	125M	No built-in limit	✗	Sequential decision-making	No tool use
ReAct	Yao et al., ICLR 2023	Deep	PaLM/GPT	—	—	✓	Reasoning + Acting	Prone to reasoning errors or hallucinations in long tasks
Tree of Thoughts	Yao et al., NeurIPS 2023	Deep	GPT-4	—	—	✗	Systematic exploration	High computational cost, requires multiple LLM calls
Reflexion	Shinn et al., NeurIPS 2023	System-level	GPT-3	—	—	✗	Reflective learning	Strong environmental dependencies, risk of cognitive bias
Multi-Agent Transformer	Wen et al., NeurIPS 2022	System-level	—	—	—	✗	Sequential coordination	Limited to multi-agent scenarios

Table 1. (continued) Quantitative Comparison of Representative Transformer-driven AI Agent Models.

Model/ System	Source/ Date	Integration Level	Base Model	Parameters	Context Length	Tool Use	Reasoning Method	Key Limitations
AutoGPT	Community , 2023	Deep	GPT-4	—	Depends on underlying model (typically 8K–128K)	☑	Goal setting + task planning + tool execution + result verification	Reported in community experiments as costly, prone to loops, unstable with complex tasks
GPT-4	OpenAI, 2023	Deep	Proprietary	Undisclosed	128K	☑	CoT	High cost, API rate limiting
Claude 3	Anthropic, 2024	Deep	Proprietary	Undisclosed	200K	☑	CoT	Relatively weaker on programming tasks

3. REDEFINING THE APPLICATIONS OF AI AGENTS WITH TRANSFORMER.

The three levels of integration from Chapter 2 are not left on paper. Practically speaking, the three problems are: How to link language with physical actions? How to complete the data-to-decision loop? How to have multi-agent systems learn instead of follow rules?

Shallow integration answers the first question. The command “dim the living room lights and play jazz music” can be recognized by speech recognition, semantically parsed, and then the corresponding device APIs are called one after another; a single Transformer backbone does not need modality-specific encoders and is therefore low-latency end-to-end. It is the same; thus, in industrial automation, “start production line 3 and monitor the temperature” will initiate equipment start-up and have sensors observe for any abnormalities. Based on the above reasoning, they are the same; a shallow integration normalizes all types of perception signals into a single kind of semantic format, and then a thin coordination layer maps this format to physical commands.

Marketing and Supply Chain Management will be presented in detail in Section 2.2. A visitor has viewed a product but not purchased; therefore, instead of sending a general coupon, the agent will use the browsing record to send a personalised email and change the advertising plan once. The response time is now a matter of hours, not days; it will increase by about 25%. The supply chain is also the same; based on sales data, weather conditions and fluctuations in demand, a plan will be made to adjust inventory levels and rearrange procurement. Although the area has changed, the pattern is still the same: raw data goes to a reasoning centre and continuously plans how to achieve a goal in that centre.

Medical diagnosis and financial investment are in different environments; therefore, errors here are more severe and the fixed coordination rules are not reliable. System-level integration (2.3) is the main place for all of the agent’s global memory and self-reflection in the Transformer network. The three parts of the Image Analyzer, Record Retriever and Report Generator will be cross-checked to ensure there are no mistakes. The investment manager is observing the market’s mood and reducing the risk of investment due to changes in policy without instructions from management. What is common to all these cases is that they can learn from the previous results, make adjustments spontaneously, and change the form of cooperation flexibly in response to changes.

At the same time, the Arc of smart homes, marketing and collaborative diagnosis are all developing. Shallow Integration Links Language and Action. Finalize the sense-to-act pathway. At the system level, there are now some rules that can learn automatically. Finally, which depth to select will be a compromise between function and speed; this problem will be further explored in the next section.

4. CHALLENGES AND FUTURE DIRECTIONS

Table 2 presents the three main problems at all stages of the integration. Rather than listing problems and their corresponding outlooks separately, the following parts will first delve into the root of each problem, then determine what the current solution has achieved and where it has failed, and finally indicate the direction that may break the deadlock.

Table 2. Comparison of the Main Problems and Countermeasures of Transformer-based AI Agents.

Challenge Category	Core Problem	Existing Solutions and Technical Directions	Future Directions and Research Perspectives
Long-Sequence Processing and Computational Efficiency	$O(n^2)$ complexity bottleneck beyond ~80K tokens	Algorithmic: sparse attention, dynamic routing—partial relief, information loss on very long sequences. Engineering: model compression, hardware acceleration (A100 etc.)—treat symptoms, not root cause.	Linear-complexity variants, SSMs, hybrid architectures (Mamba + Transformer) [19,20]
Interpretability and Responsibility Traceability	Compounded black-box effect; decisions opaque, especially under multi-agent coordination	Visualization: attention heatmaps, path tracing—limited explanatory depth. Normative: mandated decision logs (EU AI Act)—external accountability without true explainability.	Causal inference for decision drivers; neuro-symbolic architectures for auditable reasoning [18]
Multi-Agent Conflict and Efficiency Balance	Misaligned goals, resource contention; “tragedy of the commons” in agent collectives	Coordination: auction, game-theoretic protocols—small-scale, static settings only. Architectural: standardized shared memory—reduces complexity but cannot resolve genuine conflicts.	Adaptive coordination protocols; hierarchical strategies; economics-based mechanism design for incentive alignment

4.1 Long-sequence processing and Computational efficiency

The $O(n^2)$ self-attention cost becomes a hard ceiling when agents must track more than roughly 80K tokens—exactly the regime required by deep planning and multi-turn coordination. Sparse attention [15,16] and content-based routing [17] buy practical speed but sacrifice information fidelity on extremely long sequences; hardware acceleration [22] eases the pressure without touching the architectural root. The more promising turn is toward state-space models and Transformer-SSM hybrids, which already approach Transformer quality while scaling with linear complexity [19,20].

4.2 Interpretability and Traceability of Responsibility

With the introduction of more complex integration and multi-agent cooperation, there has been an increase in compound opacity that cannot be shown by attention heatmaps. Although there is a regulatory requirement in the EU AI Act to keep a log, this log does not explain the reasons for a particular decision. A more reliable way is to find the reason for the choice through causal inference, and neuro-symbolic architectures can directly add auditable rules to the reasoning loop [18].

4.3 Multi-agent cooperation: Conflict and efficiency

The general purpose and the corresponding basic requirements conflict. Auction Mechanisms and game-theoretic Protocols are suitable for small, stable environments, but they are not very flexible in the face of changes in the outside world. Mechanism Design and Hierarchical Coordination in Economics are still developing; however, they provide a sound basis for reforming the incentives of agents so that self-interest promotes the interests of society rather than leading to a deadlock.

5. CONCLUSION

Three levels of taxonomy for shallow, deep and system-level integration were proposed in this paper to classify the reorganisation of AI agent architecture by the Transformer. Shallow integration will be a multimodal encoder; deep integration is a reasoning planner that can break down tasks and call tools; system-level integration is a coordination centre that can share memory and multi-agent learning. Work on Decision Transformer, ReAct, Tree of Thoughts, Reflexion and Multi-Agent Transformer [8-12] has converged to the conclusion that the Transformer is not only a feature extractor but also a general sequence-modeling substrate that can absorb reasoning, planning and coordination in a single framework.

There is a disadvantage to this. The following levels are more difficult to calculate, less intuitive and have more coordination problems. The purpose of the taxonomy is to present these choices clearly. The three directions we need to go in the future are: linear-complexity architecture for arbitrarily long context, interpretability framework with auditable reasoning path, and incentive-aware coordination to shift competition towards collective benefit. With the development of technology, Transformer-based agents will move from scripted tools toward autonomous, reasoning partners.

REFERENCES

[1] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[2] T. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 1877–1901.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[4] r. Bommasani and others, “On the Opportunities and Risks of Foundation Models”, arXiv preprint arXiv:2108.07258, 2021.

[5] A. Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. arXiv preprint arXiv:2010.11929, 2020.

[6] Y. Gong, Y.-A. Chung and J. Glass, “AST: Audio spectrogram transformer,” arXiv preprint arXiv:2104.01778, 2021.

[7] S. Reed et al. “A generalist agent”, arXiv preprint arXiv:2205.06175, 2022.

[8] L. Chen et al., “Decision transformer: Reinforcement learning via sequence modeling,” in Advances in Neural Information Processing Systems, 2021, vol. 34, pp. 15084–15097.

[9] S. Yao et al., “ReAct: Synergising reasoning and acting in language models”, in Proc. int. Con. Learn. Represent. (ICLR), 2023.

[10] S. Yao et al., “Tree of thoughts: Deliberate problem solving with large language models,” in Advances in Neural Information Processing Systems, 2023, vol. 36.

[11] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” in Advances in Neural Information Processing Systems, 2023, vol. 36.

[12] M. Wen et al., “Multi-agent reinforcement learning is a sequence modeling problem,” in Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 16509–16521.

[13] OpenAI, “GPT-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.

[14] Anthropic, “The Claude 3 model family: Opus, Sonnet, Haiku”, arXiv preprint arXiv:2401.05677, 2024.

[15] M. Zaheer et al., “Big bird: Transformers for longer sequences,” in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 17283–17297.

[16] I. Beltagy, M. E. Peters and A. Cohan, “Longformer: The long-document transformer”, arXiv preprint arXiv:2004.05150, 2020.

[17] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, “Efficient content-based sparse attention with routing transformers,” Trans. Assoc. Comput. Linguistics, vol. 9, pp. 53–68, 2021.

[18] A. Garcez and L. C. Lamb, Neurosymbolic AI: The 3rd wave, arXiv preprint arXiv:2112.05251, 2021.

[19] O.Lieber, “Jamba: A hybrid Transformer-Mamba language model”, arXiv preprint arXiv:2403.19887, 2024.

[20] A. Hatamizadeh and J. Kautz, “MambaVision: A hybrid Mamba-Transformer vision backbone,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025.

[21] A. Dosovitskiy et al., “CARLA: An open urban driving simulator,” arXiv preprint arXiv:1711.03938, 2017.

[22] NVIDIA, “Accelerating TensorFlow on NVIDIA A100 GPUs”, NVIDIA Developer Blog, July 2020. [Online]. Available: https://developer.nvidia.com/blog/accelerating-tensorflow-on-a100-gpus/