

A Literature Review For GPT-6 Astra

Vatsala Sharma

Assistant Professor

Department Of Electronics and Communication Engineering

Government Engineering College Buxar, Bihar

vatsalasharma01@gmail.com

Abstract: - GPT-6 Astra is a large language model (LLM) developed by the American artificial intelligence firm OpenAI. It was initially released to approved users on September 3, 2026, and became available to everyone the following day. A major shift is unfolding in the world of artificial intelligence. OpenAI, the company behind ChatGPT, has announced the rollout of its new and most powerful AI model to date: GPT-6. This new model is being referred to as GPT-6 Astra. According to the company, it will initially be made available to select users. OpenAI states that GPT-6 has been developed with a focus on offering more features and superior performance compared to its predecessors. However, the company has also prioritized safety to mitigate potential risks associated with advancing AI capabilities.

KEYWORDS: - GPT-6 ASTRA, AGI, AI AGENTS

INTRODUCTION: -

OpenAI has finally launched its latest flagship model, GPT-6 Astra. The Sam Altman-led company introduced the newest AI model on Thursday, September 3, describing it as the world's most 'intelligent and aligned model'. The model has shown extraordinary promise and performance when it comes to benchmarks. OpenAI said that GPT-6 Astra is a combination of years of research and big bets across pre-training, reinforcement learning, and alignment. The company said that it is state-of-the-art in computer use, browsing, software engineering, cybersecurity, science, and professional work. OpenAI calls it a major step forward in AI reasoning, computer use, coding, and professional work. In contrast to earlier chatbots that responded to prompts, Astra is designed to carry out complex, multi-step tasks using a computer. This includes browsing websites, working with documents and software, writing code and conducting research. In simple words, Astra is aimed at moving AI from being a tool that answers questions to one that can actually get things done. It can interact with computers much like a human user. OpenAI said that it can fill out forms, update CRM records, organize calendars, conduct online research, create websites, test software, analyses scientific data, etc. It is also capable of creating documents, spreadsheets, and presentations while following a user's templates and instructions. The model comes with a 1.05-million-token context window, which allows it to work with huge amounts of information in a single task. GPT-6 Astra is our most intelligent model yet, with state-of-the-art performance in computer use, browsing, software engineering, science, and professional work. It excels at carrying out multistep workflows across code, browsers, and professional software. In Astra achieves stronger results while using substantially fewer output tokens delivering a lower estimated API cost per task than earlier models despite its higher per-token pricing.

GPT-6 Astra is also our most aligned model yet. It excels at exercising care, respecting task boundaries, and communicating transparently. When instructions leave room for interpretation, it uses the context it has to fill in routine gaps and asks focused questions when the answer could change the outcome. It incorporates new requirements, changes course when asked, and answers side questions without losing track of the broader task.

GPT-6 Astra performance: -

When it comes to performance, OpenAI claims that Astra achieves state-of-the-art results across several demanding benchmarks. On Frontier Math Tier 4, a benchmark that evaluates advanced mathematical problem-solving, the model secured 98 per cent. Similarly, on ARC-AGI, which assesses a model's ability to solve unfamiliar reasoning problems, Astra scored 99.9 per cent. On the cybersecurity benchmark Exploit Bench, the model achieved 100 per cent. While the

benchmarks seem impressive, OpenAI President Greg Brockman went a step further during a press briefing ahead of the launch, suggesting that GPT-6 Astra could eventually be viewed as a milestone in the arrival of artificial general intelligence (AGI), broadly understood as AI that can learn, reason and perform a wide range of intellectual tasks at a human level. Perhaps for ordinary users, the biggest change will be in how the model delegates tasks in real time. Now, instead of asking an AI to book an appointment with a doctor or prepare an expense sheet or compare restaurants, users can simply ask it to perform them. Based on the demonstrations, Astra can navigate websites, use software and complete multiple steps without requiring instructions for every individual action. According to OpenAI, this could let AI handle time-consuming tasks faster than people can.

Artificial General Intelligence: -

AGI is essentially an advanced form of AI capable of understanding, learning, and adapting to new environments much like the human brain while performing any intellectual task. Unlike today's AI tools, which are limited to specific tasks, AGI can think and comprehend in new domains like a human, without requiring specialized training. This model can perform a wide range of tasks with complete autonomy based on user instructions. It excels at conducting online research, analyzing scientific data, filling out online forms, and managing calendars. Additionally, it can build entire websites, install and test software, draft legal documents, and even prepare tax returns. AGI refers to a type of artificial intelligence (AI) capable of demonstrating human-like understanding, reasoning, and decision-making abilities across various tasks. OpenAI describes it as an autonomous system that outperforms humans in economically valuable tasks. However, not all experts agree that GPT-6 Astra has achieved AGI status.

The model: GPT-6 Astra

GPT-6 Astra is the world's best model for computer usage, professional work, science, coding, cybersecurity, and other tasks. this model will facilitate a new generation of entrepreneurship, scientific discovery, and innovation. We are now in the era of AGI. the AI being launched Astra would be viewed as a pivotal development; he remarked that if we were to look back a few years and ask whether we had truly created AGI, that specific moment would be the answer, and this very model would likely be the one that defined it. It is worth noting that A GI or Artificial General Intelligence refers to a state where AI can think like a human or perhaps even surpass human intelligence. Astra is a step ahead when it comes to tasks like coding. In a blog post, OpenAI described GPT-6 Astra as the best model yet for software engineering.



Fig 1. GPT-6 ASTRA

Terminal-Bench Science 0.1 evaluates whether agents can execute scientific research workflows—including data analysis, running simulations, and fitting models—using code and terminal tools. Among the models tested, GPT-6 Astra reached a new high of 64.6%, while Claude Fable 5.1 scored 52.6%—achieving this with an estimated API cost approximately 31% lower than that of Claude Fable 5.1. In the low-cost setting, Astra scored 61.1%, whereas the best result for GPT-5.6 Sol was 22.4%—again, with an estimated API cost about 27% lower than Sol's.

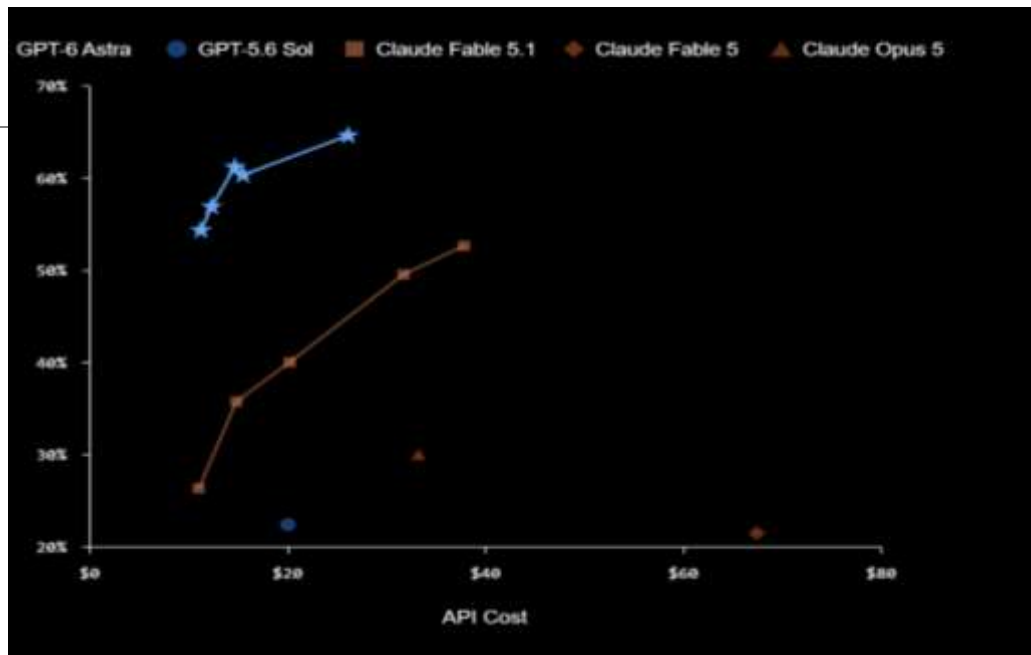


Fig. 2 Cost of API GPT-6 Astra

Astra surpassed our human action-efficiency baseline in 96% of the levels and reached near-human performance on the benchmark. It is not only the best model we have tested to date but also represents a significant shift in frontier-model performance—excelling not just at understanding new environments and solving problems within them, but also at the efficiency with which it learns to do so. GPT-6 Astra combines enhanced computer-use capabilities with targeted training for professional environments to better handle complex work tasks. It pairs the intelligence required to solve complex problems with the ability to execute multi-step workflows and create polished documents, spreadsheets, and presentations. GPT-6 Astra also enables better visual decision-making when creating websites, games, applications, and renderings. With the "Sites" feature in ChatGPT, Astra can build, host, and share websites, web apps, and games directly from a single prompt. In situations where instructions are open to interpretation, GPT-6 Astra excels at making the right decisions compared to earlier models. It typically fills in missing information using context and asks direct, relevant questions when the answer could alter the outcome. With Codex, it can continue performing tasks that do not depend on your input while simultaneously waiting for your response to its questions. If you don't provide an answer, it proceeds using educated guesses where appropriate. However, for critical decisions, it waits for your input. The examples below illustrate how Astra collaborates on everyday tasks where missing information could significantly alter the outcome. Astra excels at maintaining the full context of a task, even as it evolves. Earlier models would sometimes mistake messages that shifted direction for a new objective, thereby deviating from the original request or previously established constraints. Astra incorporates new requirements, adjusts its approach when asked, and answers intervening questions without losing focus on the overall task. GPT-6 Astra delivers state-of-the-art performance on our internal coding benchmarks and clearly outperforms GPT-5.6 Sol in trading intuition evaluations. When used for agentic coding, GPT-6 Astra communicates in a way that is easy for developers to understand and generates code that requires fewer iterations to reach production quality.

1.1 Cybersecurity: -

As we discussed in our safety update Astra is a significant jump in cyber capabilities and meets the critical threshold in cybersecurity under our preparedness framework, its ability to identify and develop zero-day exploits can help defenders find and patch weaknesses, but it also creates a need for stronger safeguards. To understand how far these capabilities extend, we ran Astra on internal and third-party expert evaluations. We first tested the model without production safeguards on Exploit Bench and Exploit Gym, which evaluate whether models can turn known software vulnerabilities into working exploits. On Exploit Bench, Astra achieved a perfect score of 100%, compared with 78.5% for GPT-5.6 Sol, our previous frontier cyber-capable model. On Exploit Gym, Astra reached a 42.4% success rate, compared with 30.3% for GPT-5.6 Sol, while using substantially fewer output tokens. Astra will refuse to comply with more advanced cybersecurity tasks such as creating proof-of-concept exploits for vulnerabilities. Through OpenAI Daybreak, we plan to expand access and roll out less restrictive safeguards in the coming weeks. This will enable more defensive workflows, including vulnerability and proof-of-concept validation, malware analysis, and detection engineering. These include stronger model robustness to better withstand potential jailbreaks and more context for our monitoring systems. We have continued rigorous internal and external testing, including automated evaluations with our internal red-teaming attackers. More details about our cyber safeguards and testing are available in the Astra safety overview. OpenAI classifies GPT-6 Astra as its first model to reach the Critical capability level for cybersecurity under its Preparedness Framework. With authorized access, it may help security professionals identify vulnerabilities and develop appropriate testing methods. These capabilities also introduce risks. Strong access controls, continuous monitoring and human approval are essential to ensure that GPT-6 Astra remains within its permitted scope.

1.2 Responsibly aligning and deploying GPT-6 Astra

Astra is our most advanced aligned model. It excels at operating carefully, staying within defined task boundaries, and communicating clearly with the user. It represents the latest development in our long-standing research program focused on training models that align with user intent from start to finish. In high-risk environments, Astra operates with greater caution, tailored to the level of risk. An evaluation involved computer-use tasks specifically designed to elicit erroneous behavior from the model; Astra was more successful in avoiding unintended outcomes in these scenarios. It performed even better when utilizing the additional security measures provided by default. Astra is also more likely to operate within the boundaries defined by the user and imposed by its environment. In an internal evaluation, Astra never attempted to bypass the Codex Auto-Review when blocked by it even in scenarios where the Auto-Review was deliberately configured to be by passable and completing the task without doing so was impossible. This behavior of adhering to environmental constraints aligns with the results of our "Impossible Cyber Task" evaluation.

Alignment training is core to our approach to deployment. As an additional layer of defenses, we also build system safeguards like Codex auto review and monitoring agents' reasoning and actions to help detect and contain unsafe behavior. As described in our safety update we are also deploying misalignment monitoring in production for Astra-class models in order to have visibility into misalignment, and help contain its worst instances. These safeguards resemble our monitoring for internal deployments and involve a system of classifiers which check the model's reasoning and actions for unauthorized behavior and automatically stop potentially unauthorized activity. Given the significant increase in Astra's cybersecurity capabilities, we are being especially careful to make this deployment safe and secure. Extra safety checks can sometimes slow, pause, or stop legitimate work, including defensive cybersecurity. If a task is paused in ChatGPT or Codex, you may be asked to review the action before continuing. In the API, the task will stop. These checks can sometimes interrupt legitimate work, and we are continuing to iterate on this system to reduce unnecessary interruptions. Misalignment monitoring cannot replace alignment: our goal is to build models that reliably stay within their authorized scope, so these protections do not need to intervene.

1.3 Availability GPT-6 Astra

GPT-6 Astra is rolling out today to a limited set of organizations and over the coming days will become available to all ChatGPT Plus, Pro, Business, and Enterprise users, as well as through the OpenAI API, Microsoft Azure, and AWS Bedrock. Astra usage is included within the existing subscription allowances—users and businesses will also be able to purchase credits for additional usage. Users on the Pro, Business, and Enterprise plans will also get access to GPT-6 Astra Pro. Enterprise administrators can enable Astra for their workspace; access is off by default at launch. For

developers, GPT-6 Astra will be available in the OpenAI API as gpt-6-astra and through Microsoft Azure and Amazon Bedrock.

1.4 Integrate with coding models: - For most API-based code generation, start with gpt-6-astra. It handles both general-purpose work and coding, which makes it a strong default when your application needs to write code, reason about requirements, inspect docs, and handle broader workflows in one place.

```
from OpenAI import OpenAI
```

```
client = OpenAI()
```

```
result = client.responses.create(
```

```
    model="gpt-6-astra",
```

```
    input="""Find the null pointer exception in this code:
```

```
def display_name(user):
```

```
    return user.profile.name
```

```
print(display_name(None))
```

```
""",
```

```
    reasoning={"effort": "high"},
```

```
)
```

```
print(result.output_text)
```

Conclusion: -

It is that it does not prove artificial general intelligence (AGI) has arrived, but it marks a massive practical leap in agentic computer use, software engineering, and long-horizon reasoning To wrap it all up, we can say that yes, OpenAI GPT-6 Astra is a very strong model. And it's making a particularly large leap in computer use. I believe computer use will be the next big focus area for open-source and proprietary harnesses in the upcoming months. I find open-source especially important when it comes to computer use, as "with great power come great responsibilities", and it's nice to be able to audit the harness before giving it access to my main computer. Besides, GPT-6 Astra is likely to use a variant of the looped transformer. Looped transformers simply give better modeling performance at a fixed compute budget. Also, better modeling performance may decrease in shorter reasoning chains. But this is not a new trend. We have always seen that within a model family with models of different sizes (e.g., GPT 5.6 Luna versus Sol). In my opinion, shorter reasoning traces are a side effect of more "intelligent" or capable models that make fewer mistakes and can access more compute internally inside their architecture versus using a reasoning trace as a scratchpad. In a sense, the same is true for humans. During an in-person college math exam, a smart and well-prepared student likely requires less use of the notepaper and needs to backtrack less often, and so on.

Future Work: - OpenAI will likely announce additional vertical-specific ChatGPT products beyond financial services, following the same pattern of layering built-in industry data on top of Astra's reasoning, given the company's stated work-first framing for the model. Broader general availability for ChatGPT Plus, Pro, Business, and Enterprise users will continue to expand through September 2026, based on OpenAI's own staged-rollout language in its release notes.

Competing model providers, particularly those with existing enterprise cloud distribution such as Microsoft and Google, will move to package their own finance-specific AI offerings more aggressively in response. Scrutiny of Astra's Critical-tier cybersecurity classification is likely to intensify as more regulated industries, not just finance, seek access, putting

pressure on OpenAI to publish more detail on its Preparedness Framework safeguards. Pricing pressure across the AI-for-finance category is likely to increase as more vendors compete for the same enterprise budgets, though specific price movements cannot be predicted with certainty at this stage. GPT-6 Astra expands future AI work through advanced computer-use agents, multi-step professional execution, and deep integration into complex software environments. OpenAI GPT-6 Astra marks a shift towards AI systems that can contribute directly to complex digital work. Its advances suggest that knowledge of AI agents, automation and workflow design will become increasingly valuable for students, developers and professionals.

Reference

- [1] On ARC-AGI-3, GPT-6 Astra was run with our responses API harness, which changes two settings to better match real-world performance. The changes do not specifically target ARC-AGI-3.
- [2] GPT-5.6 Sol refers to the version available in our API, ChatGPT Codex, and ChatGPT Work. The version in ChatGPT Chat is slightly different.
- [3] Nayak, D. *et al.* (2026). A Metaheuristic-Tuned Hybrid IDS for Detecting DDoS Attacks in IoT Networks. In: Swaroop, A., Virdee, B., Correia, S.D., Valicek, J. (eds) Data Processing and Networking. ICDPN 2025. Lecture Notes in Networks and Systems, vol 2005. Springer, Cham. https://doi.org/10.1007/978-3-032-27691-9_11
- [4] OSWorld V2-Offline is a subset of the original OSWorld V2 that works without internet access. Claude model performance on OSWorld-V2 Offline was reproduced by the authors **on** OSWorld 2.0, the scores for Claude use the official settings, and not the modified tasks and modified grading from the Fable 5.1 System Card.
- [5] Sharma, Vatsala and Anand, Nitin and Sakshi, Surbhi and E, Pallavi and Kundu, Rakesh, A Review on Security Challenges in Next-Gen IoT Networks (May 7, 2026). Proceedings of the International Conference on Innovative Computing and Communication (ICICC 2026), Available at SSRN: <https://ssrn.com/abstract=6729878> or <http://dx.doi.org/10.2139/ssrn.6729878>
- [6] Model times are the reported elapsed times for the corresponding demonstration runs. The displayed clips are edited excerpts.
- [7] Nayak, D. *et al.* (2026). A Metaheuristic-Tuned Hybrid IDS for Detecting DDoS Attacks in IoT Networks. In: Swaroop, A., Virdee, B., Correia, S.D., Valicek, J. (eds) Data Processing and Networking. ICDPN 2025. Lecture Notes in Networks and Systems, vol 2005. Springer, Cham. https://doi.org/10.1007/978-3-032-27691-9_11
- [8] V. Sharma and S. Joshi, "A Literature Review on Spectrum Sensing in Cognitive Radio Applications," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2018, pp. 883-893, doi: 10.1109/ICCONS.2018.8663089.
- [9] On BenchCAD, Claude's scores reflect 3 modifications to the eval, detailed in the Fable 5.1 System Card.
- [10] Guang Yang, Victoria Ebert, Nazif Tamer, Brian Siyuan Zheng, Luiza Pozzobon, and Noah A. Smith. arXiv:2506.19065, 2025.
- [11] Mark R. H. Gotham, Maureen Redbond, Bruno Bower, and Peter Jonas. "The OpenScore String Quartet Corpus)." Proceedings of the 10th International Conference on Digital Libraries for Musicology, pp. 49–57. ACM, 2023.
- [12] N. Anand, P. Raj, S. Shivam and V. Sharma, "A Layer based Aspect of the Security Issues in Internet of Things: An Analytical Survey," *2025 International Conference on Intelligent and Secure Engineering Solutions (CISES)*, Greater Noida Gautam Budh Nagar, India, 2025, pp. 34-40, doi: 10.1109/CISES66934.2025.11265603.
- [13] Jeremy Spence et al. "The Next Challenge for Agentic Cybersecurity: A Realistic, Contamination-Free Reverse Engineering Benchmark)" arXiv:2608.11469v1, 2026.
- [14] Nayanam, K., and V. Sharma. "Towards architecting research perspective future scope with Chat GPT." *World Journal of Pharmaceutical and Life Sciences* 10.3 (2024): 181.

[15] The Next Challenge for Agentic Cybersecurity: A Realistic, Contamination-Free Reverse Engineering Benchmark
Jeremy Spence, Nicholas Assaderaghi, Jinhao Zhu, Nikil Ravi, Raluca Ada Popa, Guannan Wei, Yangruibo Ding, Zhuo Zhang.

[16] Dr. Vatsala Sharma¹ , Kamal Nayanam² — “Sixth Generation (6G) to the Waving Seventh (7G) Wireless Communication Visions and Standards, Challenges, Applications.” International Journal of Advanced Research in Science and Technology (IJARST), Volume 13, Issue 2. DOI: <https://doi.org/10.62226/ijarst20241319>.