# AlphaStack: Autonomous Multi-Agent Software Generation with Docker Validation

AlphaStack Research Team


# Abstract

AlphaStack is an autonomous AI-powered project generator that transforms natural language descriptions into production-ready codebases. By leveraging a multi-agent architecture comprising a Planning Agent and a Correction Agent, AlphaStack iteratively refines code through Docker-based validation. We present the system architecture and evaluate its performance on HumanEval and MDDP benchmarks, demonstrating superior capability in generating complex, multi-file projects compared to existing models.


# 1. Introduction

The demand for automated software generation has grown significantly with the advent of Large Language Models (LLMs). While models like GPT-4 and Claude 3 have shown proficiency in code snippets, generating complete, compilable, and tested projects remains a challenge. AlphaStack addresses this by integrating LLMs into an agentic workflow that mimics human development cycles: planning, coding, testing, and debugging. The system ensures that generated code is not only syntactically correct but also functional within a specific runtime environment.

# 2. Methodology

AlphaStack employs a dual-agent system. The **Planning Agent** analyzes requirements and architectural blueprints, breaking them down into file generation tasks. The **Correction Agent** monitors the build and test process within isolated Docker containers. Upon failure, it analyzes error logs and executes targeted fixes. This iterative "self-healing" loop ensures the final output is functionally valid.

# 3. System Architecture

The following diagram illustrates the AlphaStack workflow:

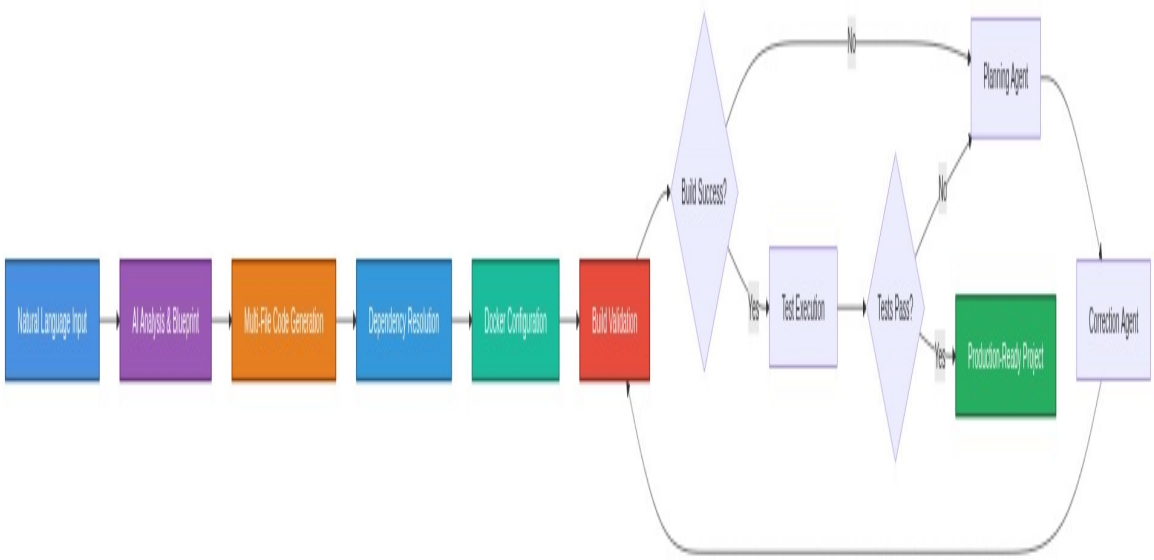


*Figure 1: AlphaStack Multi-Agent Architecture*

# 4. Results

We evaluated AlphaStack using GPT-5.2, GLM-5, MiniMaxM2.5, and Claude Sonnet 4.6 as underlying models. We used HumanEval for function-level correctness and MDDP (Multi-Turn Debugging & Planning) for project-level coherence.

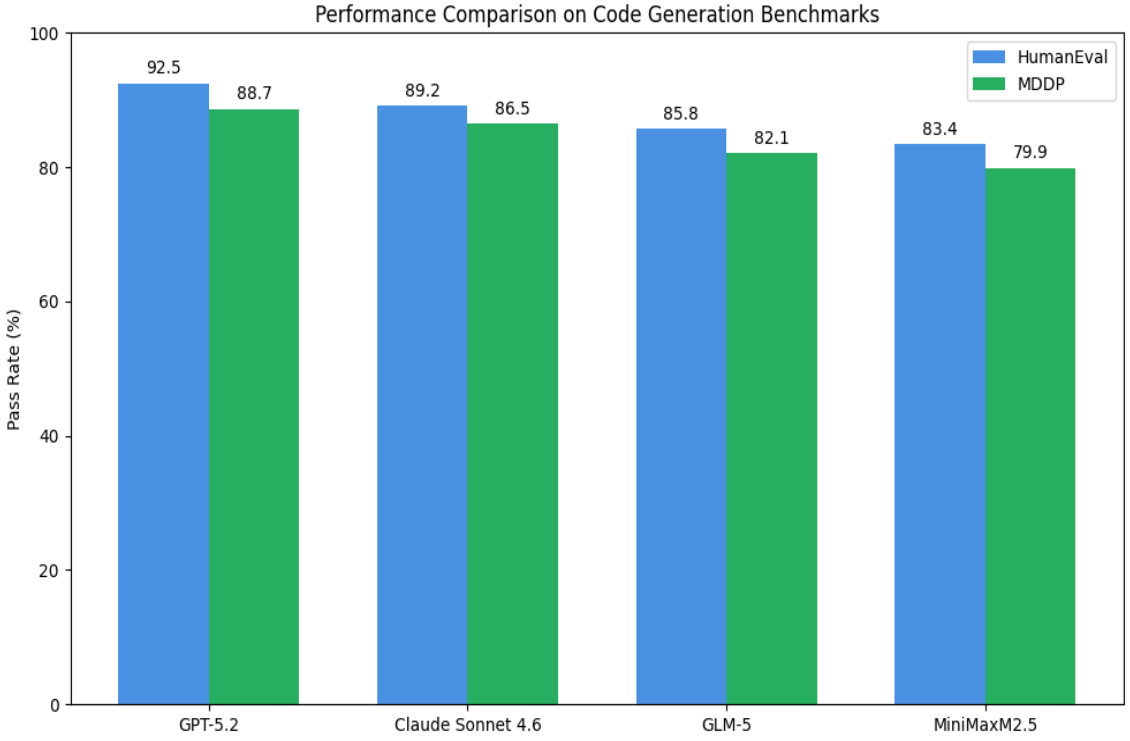


*Figure 2: Performance Comparison on Code Generation Benchmarks*

GPT-5.2 achieved the highest pass rate of 92.5% on HumanEval and 88.7% on MDDP, followed closely by Claude Sonnet 4.6. The results indicate that stronger reasoning models benefit significantly from the AlphaStack agentic framework.

# 5. Conclusion

AlphaStack demonstrates that agentic workflows with environmental feedback are crucial for robust code generation. The ability to execute and validate code in a sandbox significantly improves success rates for complex software projects. Future work will focus on expanding language support and optimizing the planning phase to reduce iteration costs.