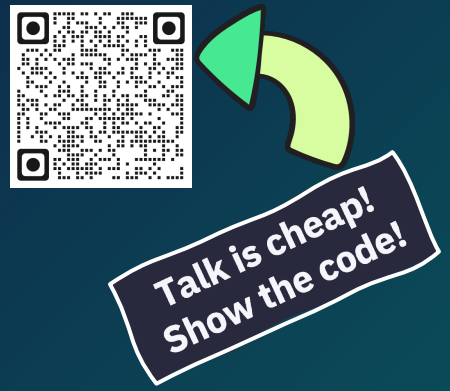
NeurIPS 2025 Datasets and Benchmarks Track

Time Travel is Cheating:

Going Live with DeepFund for Real-Time Fund Investment Benchmarking

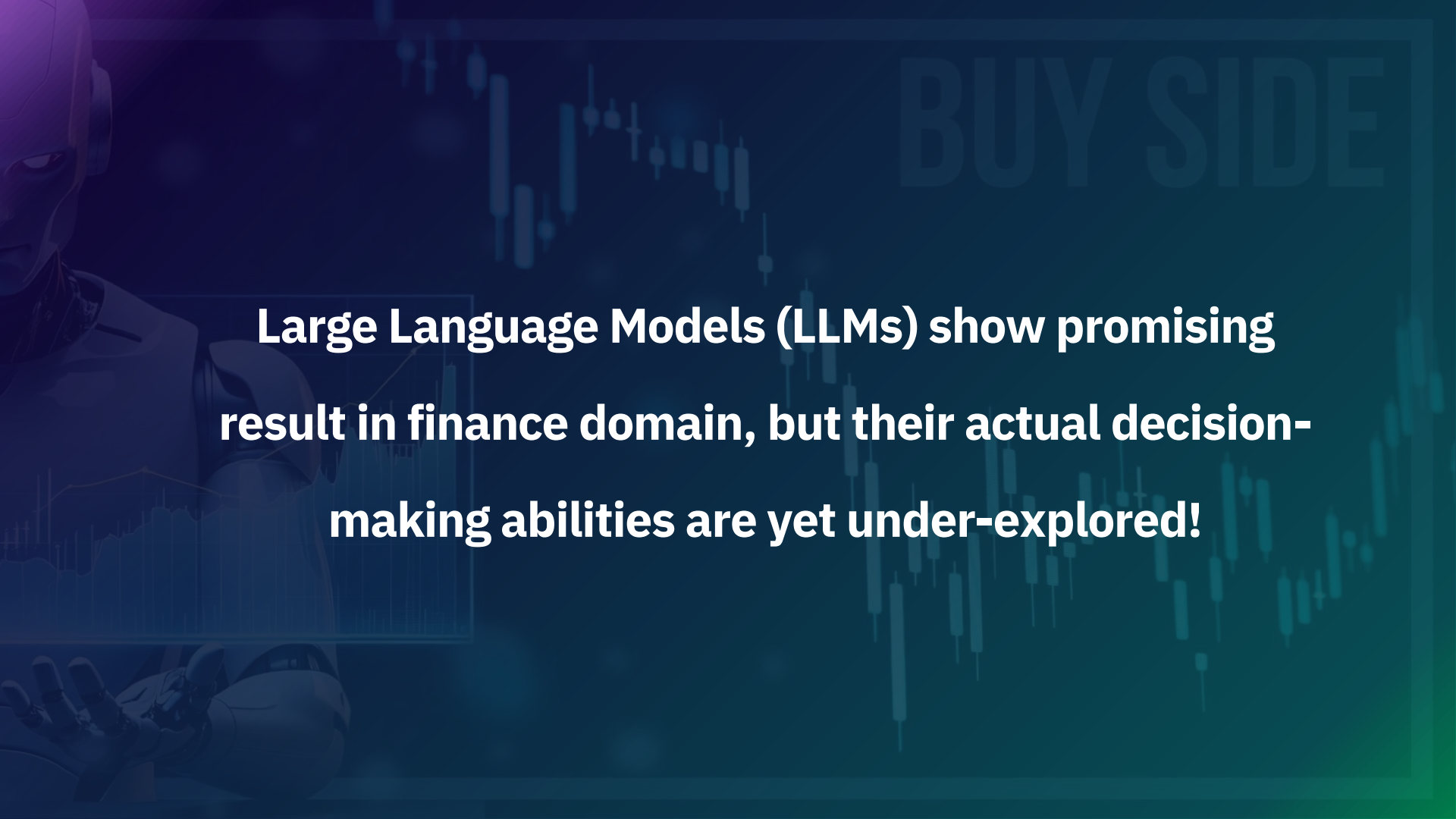


Changlun Li[1,2], Yao Shi[1,2], Chen Wang[1], Qiqi Duan[1], Runke Ruan[1], Weijie Huang[1], Haonan Long[1], Lijun Huang[1], Nan Tang[1,2], Yuyu Luo[1,2]

HKUST(GZ) [1], Paradoox AI Research[2]

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

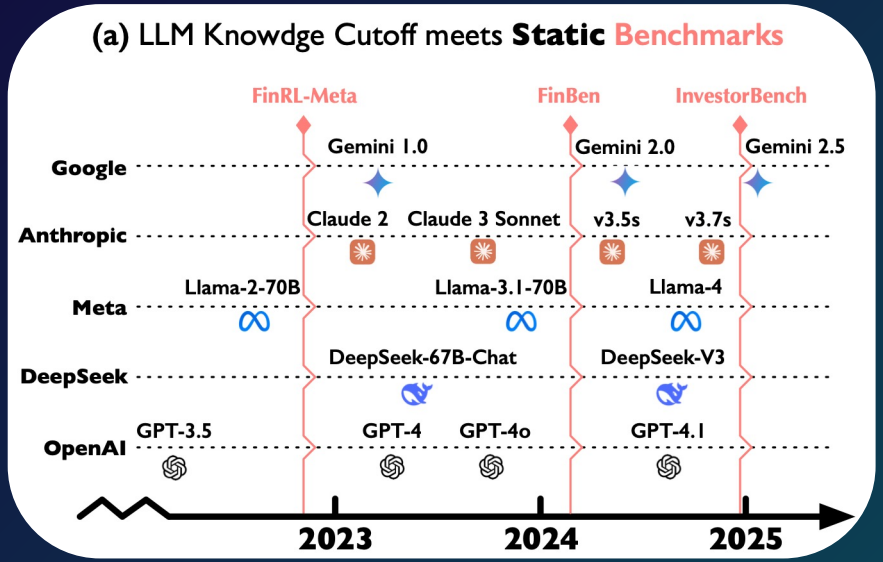
PARADOX.AI



Large Language Models (LLMs) show promising result in finance domain, but their actual decision-making abilities are yet under-explored!

The Time Travel Problem

Back-testing is problematic for LLM-driven approaches as they have been pre-trained on the very historical data used for testing, leading to severe **information leakage**.



Large language models (LLMs) **cannot** be trusted for economic forecasts during periods covered by their training data.[1]

[1] Lopez-Lira, Alejandro et al. “The Memorization Problem: Can We Trust LLMs' Economic Forecasts?” (2025). arXiv/2504.14765

Benchmark Goes Live

Live Forward Testing

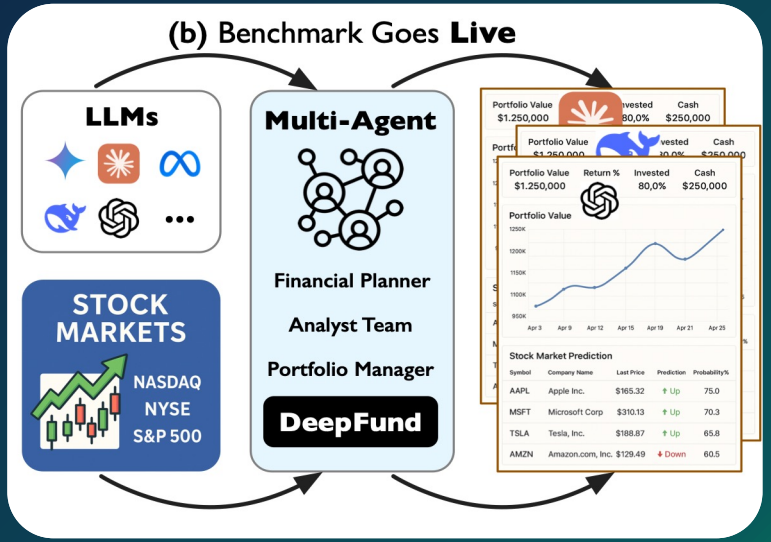
- A novel benchmarking tool that supports world market to mitigate information leakage.

Multi-Agent Workflow

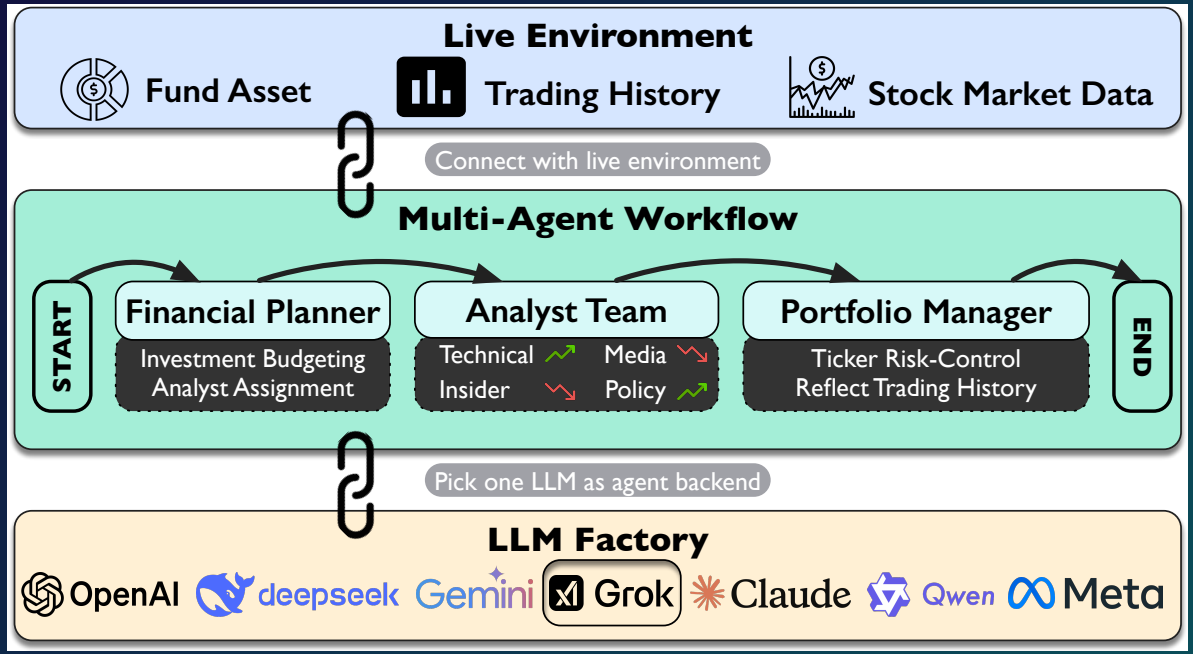
- Mimics real-world fund management: Financial Planner, Analyst Team, and Portfolio Manager.

Empirical Findings

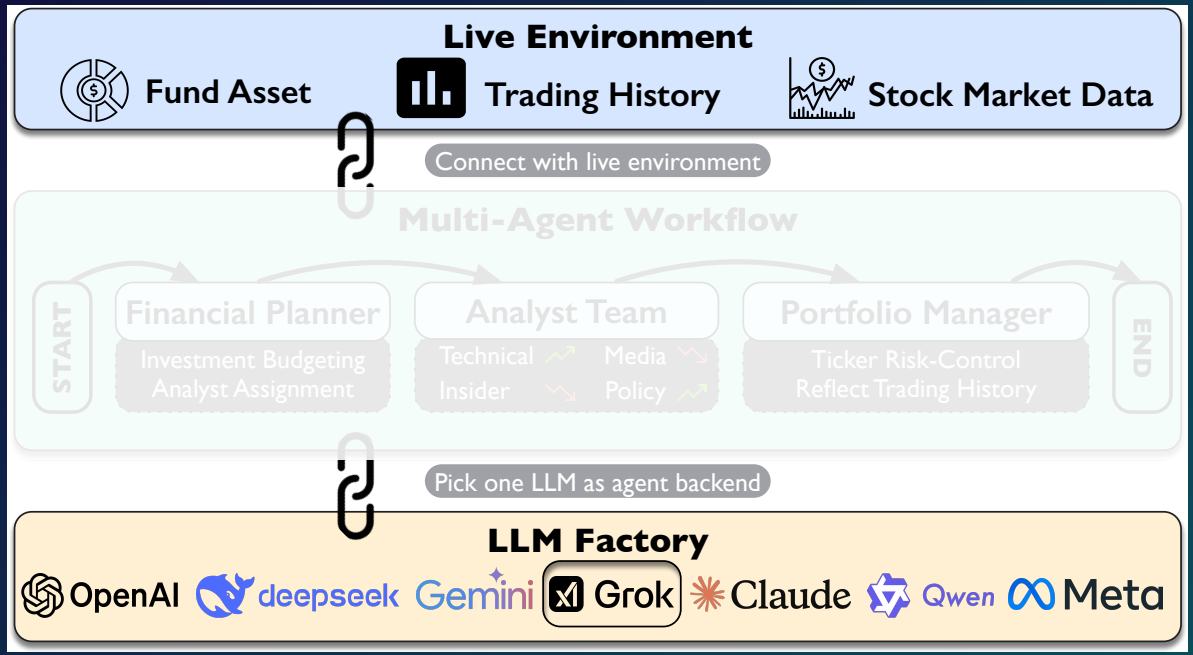
- Live environment interaction with various LLMs reveals significant performance disparities and distinct trading behaviors.



DeepFund Architecture



Connect to Live Env. + LLM Plug-in



Database Storage

Provider Endpoint

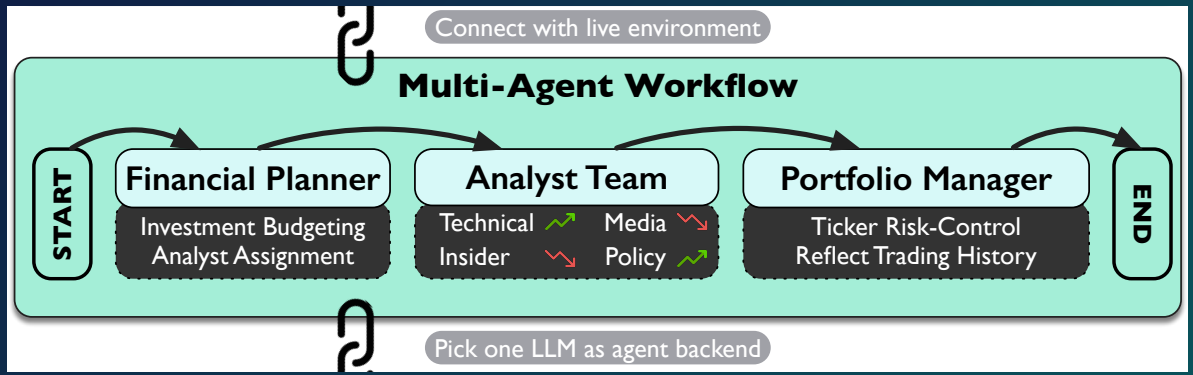
API

Multi-Agent Workflow BY LangChain

Financial Planner: Strategically determines analytical priorities and assigns tasks to analysts (deterministic or dynamic modes).

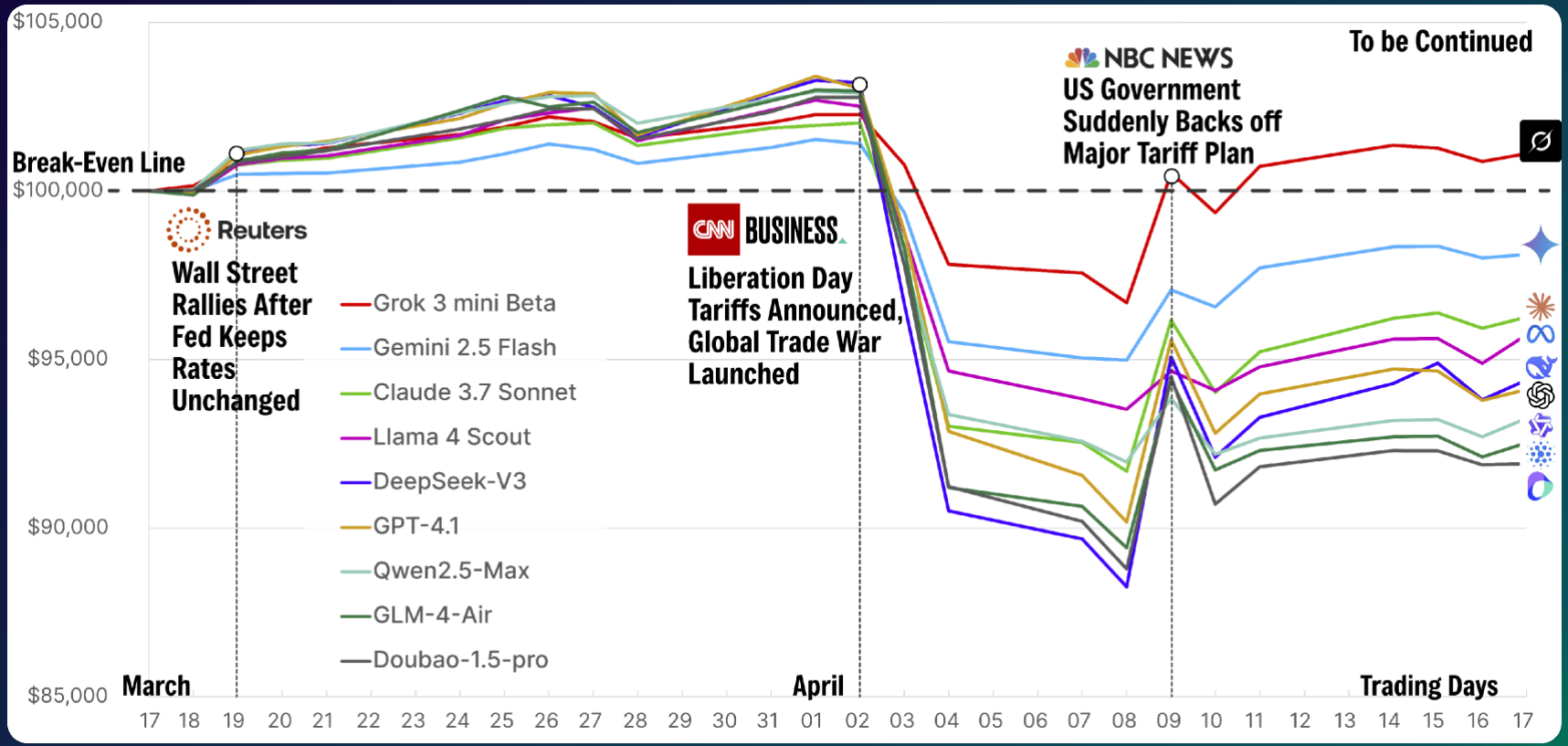
Analyst Team: Specialized agents (Fundamental, Technical, Insider, Company News, Macro Economic, Policy) analyze domain-specific data, generating Bullish, Bearish, and Neutral signals with justifications.

Portfolio Manager: Integrates analyst signals, makes Buy/Sell/Hold decisions, manages risk, reflects on history.



Arena at a glance

Setting: One-month trading period, 100K initial cash
Portfolio: Buffett's Berkshire Hathaway Top 5 holdings



Most LLMs Lost Money!

Only Grok 3 managed to secure a positive CR (+1.1%).
Half LLMs beat S&P 500.

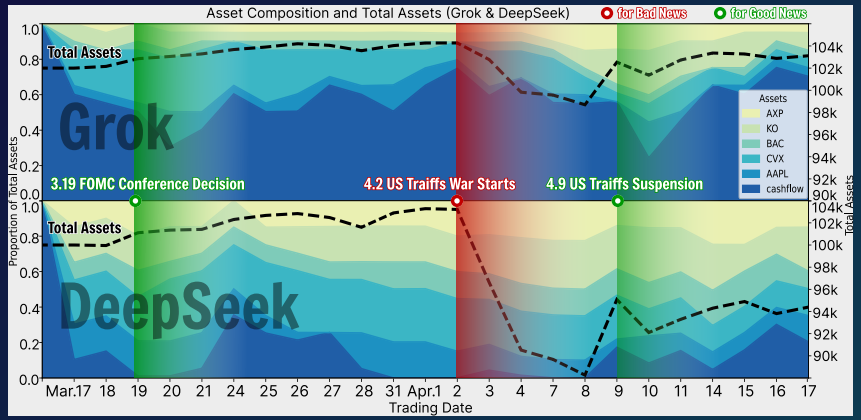
Table 3: Overall trading performance of LLMs in DeepFund, sorted by **CR** (↓).

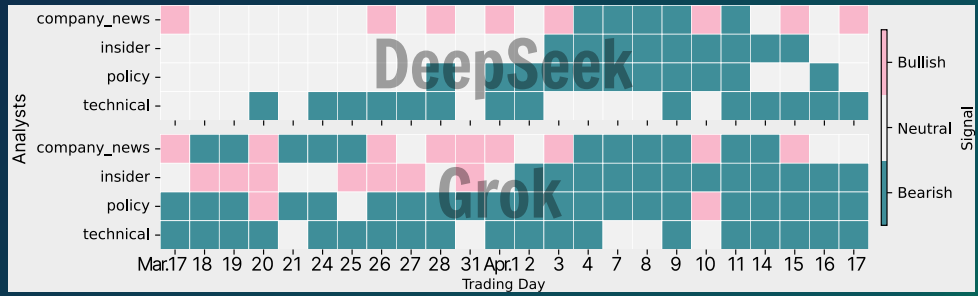
Model Version	**CR(%)**	**CR$_{bnh}$ (%)**	**SR**	**MDD (%)**	**WR (%)**	β	α
Grok 3 mini Beta	**+1.1**	-3.09	0.51	5.5	61	0.42	0.2
Gemini 2.5 Flash	-1.9	-1.58	-1.37	6.4	61	0.35	0.0
Claude 3.7 Sonnet	-3.7	-2.94	-1.45	10.1	70	0.64	0.0
Llama 4 Scout	-4.3	-3.62	-2.42	8.9	61	0.36	-0.1
DeepSeek-V3	-5.7	-5.6	-1.39	14.5	57	0.94	0.0
GPT-4.1	-5.9	-4.41	-1.87	12.8	52	0.77	0.0
Qwen2.5-Max	-6.7	-4.86	-3.12	10.7	65	0.48	-0.2
GLM-4-Air	-7.5	-3.90	-2.31	13.2	57	0.78	-0.1
Doubao-1.5-pro	-8.1	-5.37	-2.35	13.6	65	0.84	-0.1
S&P 500	-6.91	NA	0.3	13.7	NA	1.00	0.0

Distinct Trading Personalities

The Aggressive Speculator: Immediately invested ~90% of its cash. This high concentration and low cash reserve made it extremely vulnerable to the tariff-driven crash. Eventually, led to the largest dropdown (-14.5%).

The Prudent Manager: Was more cautious. It held a large cash reserve (~60%) , which allowed it to diversify risk and, crucially, seize opportunities by buying during the market dip and profiting from the rebound.





Read more experimental analysis from our paper!

From "Time Travel" to Real-World Value

Timely contribution: We introduced DeepFund, the first live, multi-agent benchmark for a fair, leakage-free evaluation.

The Sobering Result: In a real-time, volatile market, most cutting-edge LLMs (like Claude 3.7 and DeepSeek-V3) **are not** profitable fund managers. They struggle with risk control and adapting to live events.

Future Direction: assess across regime change, consider realistic execution frictions, and expand to new market with longer timeframe.

Thanks for listening!

Paper with Code

Read paper &
Star us at GitHub!

Contact Us: Prof. Yuyu Luo, yuyuluo@hkust-gz.edu.cn

Data Intelligence and Analytics Lab

Data Science and Analytics Thrust, Information Hub

The Hong Kong University of Science and Technology (Guangzhou)