

Trade Arena

Multi-Agent Trade System with Game-Theoretic Validation

11

AI AGENTS

4

STRATEGIES

7

LLM MODELS

55K+

OUTPUTS

Priyam Choksi • Vishodhan Krishnan

Do different strategies perform optimally under different market regimes?

Each trading approach has decades of research supporting it. Each has periods where it dramatically outperforms. And each has periods where it completely falls apart.

Our hypothesis: the best approach depends on market conditions. Momentum works in trending markets. Mean reversion works in ranging markets. The real skill is knowing which regime you're in.

HOW WE TESTED IT

Built a system that generates AI-driven market signals, then pits different trading strategies against each other in a competitive tournament.

Not a simulation where strategies run in parallel — an actual competition where capital flows from losers to winners each round.

Stanford Virtual Labs, 2024

Nanobody Design for COVID-19

Stanford used a multi-agent AI system to design nanobodies — tiny antibody fragments. Multiple specialized agents, each with a different perspective, debating and refining molecular structures through structured rounds.

We asked: could this architecture work for financial markets?

THE INSIGHT

Markets are fundamentally about conflicting viewpoints

Bulls versus bears

Momentum versus mean reversion

Technical analysis versus fundamentals

The same adversarial debate structure that worked for molecular biology seemed like a natural fit.

Three components, three questions

01

Analysis Pipeline

11 AI agents process market data through 5 phases: gather → debate → synthesize → evaluate → decide.

Can AI generate meaningful market signals?

73.2%

Directional Accuracy

02

Strategy Arena

4 strategies compete for shared \$1M capital. Winners take from losers via z-score reallocation.

Which approach wins, and does it depend on conditions?

1,704

Tournament Rounds

03

LLM Arena

7 models given identical data and prompts, competing head-to-head over 383 trading days.

Does the AI model itself affect decision quality?

3,707

Total Trades

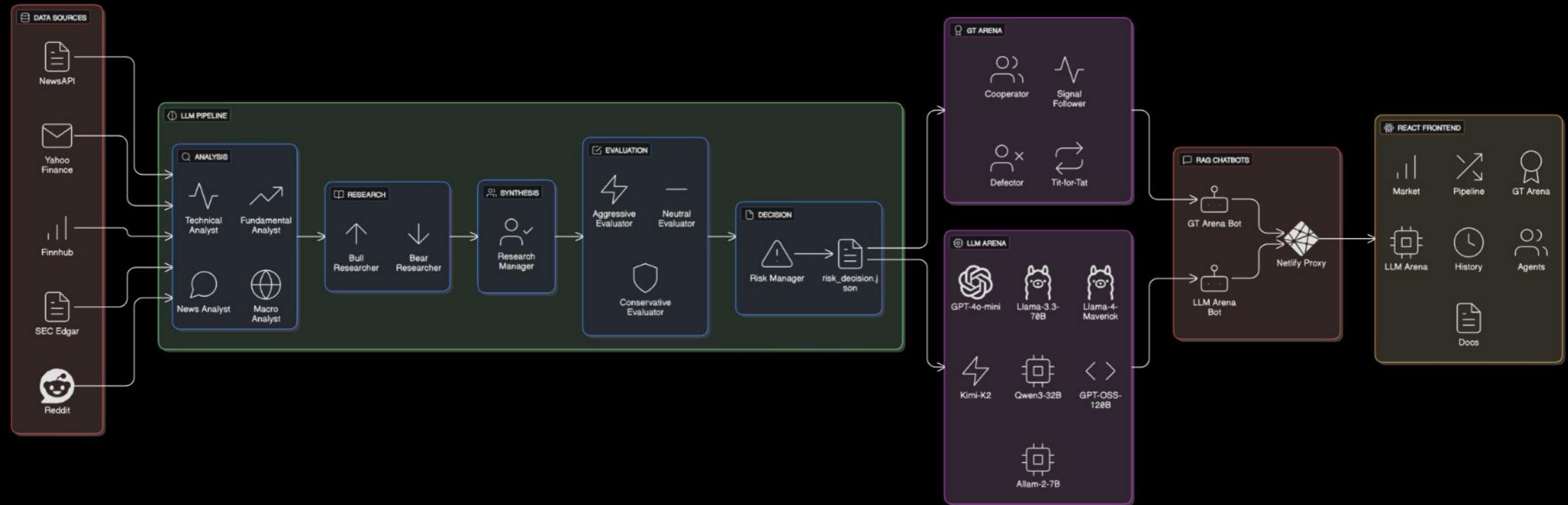
20 tickers (AAPL, NVDA, TSLA, META, GOOGL, MSFT, JPM, GS...) · June 2024 – December 2025 · ~\$17M simulated volume

Like a well-run investment committee

You don't want one analyst making the call. You want multiple perspectives, those perspectives to challenge each other, and risk checks before capital gets deployed.



~3 minutes per stock. Every decision fully traceable — see exactly which signals drove the recommendation and how the debate played out.



Single-model AI can be confidently wrong

When you make models argue against each other, you surface weaknesses before they become expensive mistakes. The 73.2% accuracy doesn't come from any single brilliant agent — it comes from the structure.

Bull Researcher

- Formal thesis with specific catalysts
- Upside price targets
- Must anticipate and rebut bear arguments

Bear Researcher

- Risks and downside scenarios
- What would make this trade fail?
- Counter-arguments to bull case



RESEARCH MANAGER SYNTHESIZES INTO PROBABILITY-WEIGHTED OUTPUT

62%

Bull Case

28%

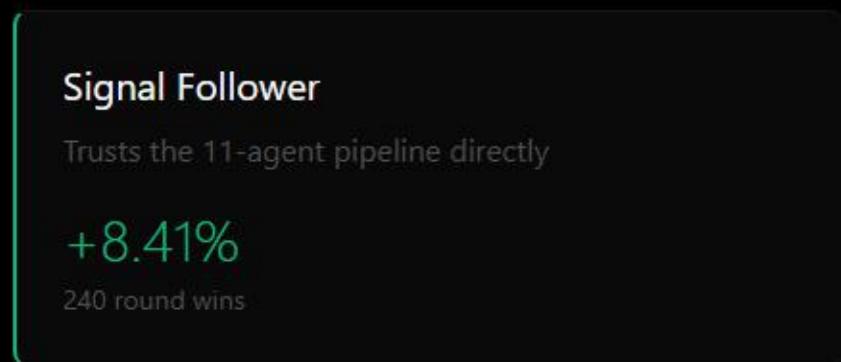
Bear Case

10%

Base Case

Game theory, not backtesting

Backtesting treats strategies in isolation. In real markets, your returns depend on what everyone else is doing. So we built a tournament where strategies compete for a shared capital pool — winners take from losers via z-score reallocation.



\$250K each

Starting Capital

20 tickers

90 rounds each

80.3%

Cooperation Rate

+16.32%

Benchmark (B&H)

No single strategy wins everywhere

RESULTS BY MARKET REGIME (20 TICKERS)

	Signal Follower	Cooperator	Defector	Winner
↗ Bull (13 tickers)	+14.2%	+12.8%	+14.4%	Defector (8)
↘ Bear (5 tickers)	-5.6%	-4.4%	-6.4%	Cooperator (5)
↔ Sideways (2 tickers)	-2.2%	+0.2%	-2.7%	Cooperator (2)

This is exactly what our hypothesis predicted: **the optimal approach depends on conditions.**

Does the AI model itself matter?

7 models. \$250K each. 383 trading days. 3,707 trades. ~\$17M volume. Same data, same prompts — only the architecture differs.

1 Llama-3.3-70B
+32.6%

280 trades · \$31K

2 Qwen3-32B
+32.2%

365 trades · \$330K

3 Kimi-K2
+30.3%

1,068 trades · \$326K

4 GPT-4o-mini
+24.4%

87 trades · \$311K

Less Can Be More

Llama-70B: 280 trades → +32.6%. GPT-OSS: 1,408 trades → +29.9%. Trading frequency doesn't equal returns.

Size ≠ Performance

Llama-70B beat larger models. Architecture and training data matter more than parameter count.

Trading Personalities

Allam-2-7B: 25 trades, 100% buy ratio, never sold. Still returned +21.4%. Pure diamond hands.

What We Learned

01 Market regime matters more than strategy selection

Instead of optimizing strategy parameters, focus on regime detection. If you can accurately identify whether you're in a trending or mean-reverting environment, the right strategy becomes obvious.

02 Adversarial structure improves AI decision-making

73.2% accuracy comes from structured disagreement — bull vs bear, aggressive vs conservative — that forces blind spots to surface. Don't rely on a single model's output for consequential decisions.

03 Model architecture affects decision character

Different LLMs produce systematically different outputs even with identical inputs. Model selection isn't just about capability — it's about behavior. Llama is aggressive, Allam is conservative, some adapt to conditions.

What We Built



Cross-Domain Transfer

Multi-agent architecture from molecular biology → financial markets.
The orchestration pattern is domain-agnostic. Change prompts, swap
data sources.



Novel Validation

Game-theoretic tournament for validating AI signals. Not backtesting
— genuine strategic competition where capital flows from losers to
winners.



Full Transparency

55,000+ JSON files with complete agent reasoning. Every decision
explainable. Every intermediate step logged. Open source on GitHub.

73.2%

Accuracy

+32.6%

Best LLM Return

~\$80

API Cost

Questions?

Happy to discuss methodology, results, or technical implementation.

Priyam Choksi • Vishodhan Krishnan

LIVE DEMO

tradearena.site

REPOSITORY

github.com/priyam-choksi/matsmatsmats

~\$80

API Cost

~3 min

Per Stock

Open Source

55K+ Files

Academic research · Not financial advice