CIO-Agent FAB++: A Dynamic Multi-Dimensional Benchmark for Evaluating AI Finance Agents

Team AgentBusters
AgentBeats Competition 2026
https://github.com/yxc20089/AgentBusters

January 16, 2026

Abstract

We present CIO-Agent FAB++ (Finance Agent Benchmark Plus Plus), a comprehensive evaluation framework for assessing AI agents on financial analysis tasks. FAB++ integrates four benchmark datasets—BizFinBench, Public CSV, Synthetic Questions, and Options Alpha—into a unified scoring system with three weighted sections: Knowledge Retrieval (30%), Analytical Reasoning (35%), and Options Trading (35%). The Analytical Reasoning section features 20 olympiad-style finance logic problems spanning capital budgeting, portfolio theory, fixed income, corporate finance, and derivatives. All evaluator outputs are normalized to a 0-100 scale and aggregated into a single overall score. We introduce the Options Alpha Challenge, a specialized track testing Black-Scholes pricing, Greeks analysis, and multi-leg strategy construction with four-dimensional scoring. Our framework leverages the Agent-to-Agent (A2A) protocol for standardized communication and Model Context Protocol (MCP) servers for real-time financial data access. Experimental results on a GPT-4o mini baseline demonstrate 60.44/100 overall score with clear capability patterns: strong knowledge retrieval (83.33) versus weaker analytical reasoning (50.00) and options trading (51.25).

Keywords: AI Agents, Finance Benchmark, Options Trading, Agent Evaluation, A2A Protocol, MCP

1 Introduction

The rapid advancement of large language models (LLMs) has enabled the development of sophisticated AI agents capable of performing complex financial analysis tasks [Brown et al., 2020]. However, evaluating these agents presents significant challenges: financial reasoning requires numerical precision, temporal awareness, and domain expertise that traditional NLP benchmarks fail to capture adequately.

Existing finance benchmarks suffer from several limitations:

1. **Static evaluation**: Fixed question sets become memorized by models during training, leading to inflated performance metrics.
2. **Single-dimensional scoring**: Most benchmarks evaluate only answer correctness, ignoring reasoning quality and methodology.
3. **Lack of temporal constraints**: Agents may inadvertently access future information, violating realistic trading scenarios.

4. **Limited options coverage**: Few benchmarks evaluate quantitative finance skills like derivatives pricing and risk management.

We address these limitations with CIO-Agent FAB++, a dynamic benchmark system that:

- Generates novel evaluation tasks from real financial data with temporal locking
- Evaluates agents across multiple dimensions including macro reasoning, fundamental accuracy, and execution quality
- Introduces adversarial debate to test conviction and robustness
- Provides comprehensive options trading evaluation with Black-Scholes pricing verification

2 Related Work

2.1 Financial Benchmarks

The Finance Agent Benchmark (FAB) [Bigeard et al., 2025] introduced structured evaluation of AI agents on earnings analysis tasks. BizFinBench [Lu et al., 2025] expanded coverage to include Chinese financial markets and multi-turn reasoning. However, these benchmarks use static question sets vulnerable to data contamination.

2.2 Agent Communication Protocols

The Agent-to-Agent (A2A) protocol [A2A Protocol, 2025] standardizes communication between AI agents, enabling interoperability across different implementations. The Model Context Protocol (MCP) [MCP, 2024] provides a unified interface for agents to access external tools and data sources.

2.3 Options Pricing Models

The Black-Scholes-Merton model [Black and Scholes, 1973, Merton, 1973] remains the foundation for options pricing. Extensions include stochastic volatility models [Heston, 1993] and jump-diffusion processes [Merton, 1976].

3 System Architecture

3.1 Overview

FAB++ implements a Green Agent (evaluator) and Purple Agent (finance analyst) architecture following the A2A protocol specification. Figure 1 illustrates the system components.

```
+------------------------------------------------------------------+
|                    AgentBusters System                             |
+------------------------------------------------------------------+
|  +---------------+      A2A Protocol      +---------------+       |
|  | Green Agent   |<--------------------->| Purple Agent  |       |
|  | (Evaluator)   |                       | (Analyst)     |       |
|  | Port: 9109    |                       | Port: 9110    |       |
|  +-------+-------+                       +-------+-------+       |
|          |                                       |                |
|          |     +-----------------------------+   |                |
|          |     |       6 MCP Servers         |   |                |
|          |     | SEC EDGAR  | Yahoo Finance  |   |                |
|          |     | Sandbox    | Options Chain  |   |                |
|          |     | Trading Sim| Risk Metrics   |   |                |
|          |     +-----------------------------+   |                |
+------------------------------------------------------------------+
```

Figure 1: FAB++ System Architecture

3.2 Green Agent (Evaluator)

The Green Agent serves as the benchmark orchestrator, responsible for:

- Dynamic task generation from financial data templates
- Multi-dimensional response evaluation
- Adversarial counter-argument generation
- Alpha Score computation

3.3 Purple Agent (Finance Analyst)

The Purple Agent represents the system under test, implementing:

- Financial data retrieval via MCP servers
- LLM-powered analysis generation
- Options strategy construction
- Risk assessment and position sizing

3.4 MCP Server Infrastructure

We deploy six MCP servers providing specialized financial capabilities:

Table 1: MCP Server Specifications

Server	**Port**	**Capabilities**
SEC EDGAR	8101	10-K/10-Q filings, XBRL parsing, temporal locking
Yahoo Finance	8102	Real-time quotes, historical data, lookahead detection
Python Sandbox	8103	Secure code execution for numerical computations
Options Chain	8104	Black-Scholes pricing, Greeks calculation, IV surface
Trading Simulator	8105	Paper trading, slippage modeling, P&L tracking
Risk Metrics	8106	VaR computation, Sharpe/Sortino ratios, stress testing

4 Evaluation Methodology

4.1 Overview: The Benchmark Router

FAB++ implements a unified evaluation router that orchestrates tasks from four distinct benchmark datasets, each targeting different financial reasoning capabilities. The router samples questions from each dataset according to a configurable strategy (stratified, random, or sequential) and routes responses to dataset-specific evaluators. All evaluator outputs are then normalized and aggregated into a single overall score.

```
+------------------------------------------------------------------+
|                     FAB++ Benchmark Router                        |
+------------------------------------------------------------------+
|                                                                   |
|  +------------------+    +------------------+    +---------------+|
|  | BizFinBench v2   |    | Public CSV       |    | Synthetic     ||
|  | - Event Logic    |    | - Beat/Miss      |    | - NPV/IRR     ||
|  | - Quantitative   |    | - Market Analysis|    | - Portfolio   ||
|  +--------+---------+    +--------+---------+    | - Fixed Income||
|           |                       |              | - Options     ||
|           v                       v              | - Corporate   ||
|  +------------------+    +------------------+    +-------+-------+|
|  | Exact Match      |    | Rubric-Based     |            |        |
|  | Evaluator        |    | LLM Evaluator    |            v        |
|  +--------+---------+    +--------+---------+    +---------------+|
|           |                       |              | Rubric-Based  ||
|           |                       |              | LLM Evaluator ||
|           |                       |              +-------+-------+|
|           |                       |                      |        |
|           +----------+------------+----------------------+        |
|                      |                                            |
|                      v                                            |
|           +---------------------+         +------------------+    |
|           | Score Normalizer    |         | Options Alpha    |    |
|           | (0-1 -> 0-100)      |         | - Greeks         |    |
|           +----------+----------+         | - Strategies     |    |
|                      |                    | - P&L/Risk       |    |
|                      |                    +--------+---------+    |
|                      |                             |              |
|                      |     +-----------------------+              |
|                      |     |                                      |
|                      v     v                                      |
|           +----------------------------------+                    |
|           |     Unified Scoring Engine       |                    |
|           | Knowledge(30%) + Analysis(35%)   |                    |
|           |      + Options(35%) = Overall    |                    |
|           +----------------------------------+                    |
+------------------------------------------------------------------+
```

Figure 2: FAB++ Benchmark Router Architecture

4.2 Benchmark Datasets

The router integrates four complementary benchmark datasets:

4.2.1 BizFinBench v2 (Knowledge Retrieval)

A bilingual benchmark from Lu et al. [2025] testing financial fact retrieval and quantitative computation:

- **Event Logic Reasoning**: Temporal ordering of financial events

- **Financial Quantitative Computation**: Precise numerical calculations (e.g., EPS, margins)

Evaluator: Exact match with 1% tolerance for numerical answers.

4.2.2 Public CSV (Knowledge Retrieval)

Questions derived from the Finance Agent Benchmark [Bigeard et al., 2025] public dataset:

- **Beat or Miss**: Earnings surprise detection against analyst consensus
- **Market Analysis**: Qualitative interpretation of market events

Evaluator: LLM-based rubric scoring with component weights.

4.2.3 Synthetic Questions (Analytical Reasoning)

Twenty olympiad-style finance logic problems requiring multi-step reasoning without external data retrieval. Topics include capital budgeting, portfolio theory, fixed income, corporate finance, options, and forex. See Appendix D for the complete question bank.

- **Self-contained**: All information provided in the question
- **CFA-level difficulty**: Undergraduate to professional curriculum
- **Definitive answers**: Unambiguous correct solutions for objective scoring

Evaluator: LLM-based rubric scoring (methodology 30%, calculation 30%, answer 40%).

4.2.4 Options Alpha (Options Trading)

Specialized evaluation track for derivatives knowledge:

- **Greeks Analysis**: Delta, gamma, theta, vega calculations
- **Strategy Construction**: Multi-leg options strategies (spreads, condors, straddles)
- **P&L Analysis**: Max profit/loss, breakeven calculations
- **Risk Management**: Position sizing, hedging strategies

Evaluator: Four-dimensional scoring (P&L 25%, Greeks 25%, Strategy 25%, Risk 25%). See Section 5 for details.

4.3 Unified Scoring System

4.3.1 Three-Section Architecture

Tasks are grouped into three weighted sections based on the skills they test:

Table 2: Benchmark Sections, Datasets, and Weights

Section	Datasets	Weight	Skills Tested
Knowledge Retrieval	BizFinBench, Public CSV	30%	Data extraction, financial facts
Analytical Reasoning	Synthetic Questions	35%	Logic, multi-step calculations
Options Trading	Options Alpha	35%	Derivatives, Greeks, strategies

4.3.2 Score Normalization

Different evaluators produce scores on different scales. All scores are normalized to 0-100 before aggregation:

Table 3: Score Normalization by Dataset

Dataset	**Raw Range**	**Normalization**	**Section**
BizFinBench	0.0–1.0	score × 100	Knowledge
Public CSV	0.0–1.0	score × 100	Knowledge
Synthetic	0.0–1.0	score × 100	Analysis
Options Alpha	0–100	No change	Options

4.3.3 Final Score Calculation

The overall score is computed in three steps:

Step 1: Section Scores. For each section s, compute the mean normalized score across all tasks in that section:

$$S_s = \frac{1}{|T_s|} \sum_{t \in T_s} \text{normalize}(\text{score}_t) \tag{1}$$

Step 2: Weight Redistribution. If any section has no tasks, redistribute weights proportionally:

$$w'_s = \frac{w_s}{\sum_{j \in \text{active}} w_j} \tag{2}$$

Step 3: Weighted Aggregation. Compute the final overall score:

$$\text{OverallScore} = \sum_{s \in \text{active}} w'_s \cdot S_s \tag{3}$$

With all sections active and default weights:

$$\text{OverallScore} = 0.30 \cdot S_{\text{knowledge}} + 0.35 \cdot S_{\text{analysis}} + 0.35 \cdot S_{\text{options}} \tag{4}$$

4.3.4 Example Calculation

Given the following section scores from an evaluation run:

- Knowledge Retrieval: 83.33 (from 6 tasks)
- Analytical Reasoning: 50.00 (from 2 tasks)
- Options Trading: 51.25 (from 2 tasks)

The overall score is:

$$\begin{align}
\text{OverallScore} &= 0.30 \times 83.33 + 0.35 \times 50.00 + 0.35 \times 51.25 \tag{5}\\
&= 25.00 + 17.50 + 17.94 \tag{6}\\
&= 60.44 \tag{7}
\end{align}$$

4.4 Optional: Adversarial Debate

FAB++ supports an optional adversarial debate mode to test agent conviction:

Algorithm 1 Adversarial Debate Protocol

```
Require: Agent response $A$, Task $\tau$
Generate counter-argument $C$ challenging $A$
Request rebuttal $R$ from agent
Evaluate conviction: maintained, weakened, or collapsed
Compute debate multiplier $m \in [0.8, 1.2]$
return Multiplier $m$
```

When debate is enabled, the section score is adjusted by the debate multiplier before aggregation.

5 Options Alpha Challenge

5.1 Black-Scholes Implementation

The Options Chain MCP server implements the Black-Scholes-Merton model with dividend yield:

$$d_1 = \frac{\ln(S/K) + (r - q + \sigma^2/2)T}{\sigma\sqrt{T}} \tag{8}$$

$$d_2 = d_1 - \sigma\sqrt{T} \tag{9}$$

Call and put prices:

$$C = Se^{-qT}N(d_1) - Ke^{-rT}N(d_2) \tag{10}$$

$$P = Ke^{-rT}N(-d_2) - Se^{-qT}N(-d_1) \tag{11}$$

where S is spot price, K is strike, r is risk-free rate, q is dividend yield, σ is volatility, and T is time to expiration.

5.2 Greeks Calculation

We compute the standard Greeks for evaluation:

Table 4: Options Greeks Formulas

Greek	**Call**	**Put**
Delta (Δ)	$e^{-qT}N(d_1)$	$-e^{-qT}N(-d_1)$
Gamma (Γ)	$\frac{e^{-qT}n(d_1)}{S\sigma\sqrt{T}}$	Same as call
Theta (Θ)	$-\frac{Se^{-qT}n(d_1)\sigma}{2\sqrt{T}} - rKe^{-rT}N(d_2)$	Complex
Vega (ν)	$Se^{-qT}\sqrt{T}n(d_1)$	Same as call
Rho (ρ)	$KTe^{-rT}N(d_2)$	$-KTe^{-rT}N(-d_2)$

5.3 Options Evaluation Scoring

The Options Evaluator uses a four-dimensional scoring rubric:

$$S_{\text{options}} = 0.25 \cdot S_{\text{P\&L}} + 0.25 \cdot S_{\text{Greeks}} + 0.25 \cdot S_{\text{Strategy}} + 0.25 \cdot S_{\text{Risk}} \tag{12}$$

Table 5: Options Scoring Dimensions

Dimension	**Evaluation Criteria**
P&L Accuracy	Max profit/loss calculations, breakeven points, probability of profit
Greeks Accuracy	Delta, gamma, theta, vega values within 5% tolerance
Strategy Quality	Correct leg identification, strike selection rationale, structure validity
Risk Management	Position sizing, hedging strategy, exit criteria definition

6 Analytical Reasoning: Synthetic Questions

6.1 Overview

The Analytical Reasoning section evaluates agents on self-contained finance logic problems that require multi-step reasoning without external data retrieval. Unlike BizFinBench or Options Alpha tasks that test data extraction and domain-specific calculations, synthetic questions assess fundamental financial reasoning ability.

6.2 Question Categories

We curate 22 olympiad-style finance questions across 10 topic areas:

Table 6: Synthetic Question Topics

Topic	**Count**	**Example Concept**
Capital Budgeting	2	NPV crossover rate
Portfolio Theory	3	Beta adjustment, leverage
Fixed Income	4	Bond pricing, duration immunization
Corporate Finance	3	FCFF, Modigliani-Miller
Options & Derivatives	4	Put-call parity, risk-neutral valuation
Time Value of Money	2	Present value comparisons
Valuation	2	Gordon Growth Model, DCF
Forex	1	Covered interest arbitrage
Corporate Actions	1	Stock splits
Leverage	1	Combined leverage (DOL × DFL)

6.3 Question Design Principles

Synthetic questions are designed to:

1. **Be self-contained**: All necessary information is provided in the question; no external data retrieval required.

2. **Test logical reasoning**: Questions require multi-step deduction, not memorized formulas.
3. **Have definitive answers**: Each question has an unambiguous correct answer for objective scoring.
4. **Cover CFA-level finance**: Topics span undergraduate to professional finance curriculum.

6.4 Example Questions

6.4.1 Capital Budgeting (NPV Crossover)

> *"A company has two mutually exclusive projects. Project A requires $100,000 investment and returns $150,000 in Year 1. Project B requires $100,000 investment and returns $180,000 in Year 2. At what discount rate are the two projects equally attractive (i.e., have equal NPV)?"*

Answer: 20% (derived by setting $\text{NPV}_A = \text{NPV}_B$ and solving for r)

6.4.2 Duration Immunization

> *"A pension fund has liabilities with duration of 15 years. It holds two bonds: Bond A with duration 5 years and Bond B with duration 20 years. What percentage of the portfolio should be invested in Bond B to immunize against interest rate changes?"*

Answer: 66.67% (weighted average duration must equal liability duration)

6.4.3 Interest Rate Swap

> *"Company X can borrow at fixed 8% or floating LIBOR+1%. Company Y can borrow at fixed 10% or floating LIBOR+2%. If they enter a swap where X borrows floating and Y borrows fixed, splitting gains equally, what fixed rate does X effectively pay?"*

Answer: 7.50% (comparative advantage analysis: total gain = 1%, each party gains 0.5%)

6.5 Evaluation Methodology

Synthetic questions use LLM-based semantic evaluation with structured rubrics:

$$S_{\text{synthetic}} = \sum_i w_i \cdot \text{match}(R_i, A) \tag{13}$$

where R_i are rubric components (methodology, calculation, final answer) and A is the agent's response. The evaluator checks:

- **Methodology** (30%): Correct problem setup and formula selection
- **Calculation** (30%): Accurate intermediate computations
- **Final Answer** (40%): Correct numerical result within tolerance

7 Experiments

7.1 Experimental Setup

We evaluated a baseline Purple Agent using GPT-4o mini as the underlying LLM. The evaluation was conducted through the Green Agent A2A server using a unified multi-dataset configuration that tests across all four dataset types mapped to three benchmark sections:

- **Knowledge Retrieval**: BizFinBench v2 (financial facts) + Public CSV (market analysis)
- **Analytical Reasoning**: Synthetic questions (olympiad-style finance logic)
- **Options Trading**: Options Alpha (Greeks analysis, strategy construction)

7.2 Unified Section-Based Results

Table 7: Section-Based Evaluation Results (Unified Scoring)

Section	Score	Weight	Contribution	Tasks	Accuracy
Knowledge Retrieval	83.33	30%	25.00	6	83.3%
Analytical Reasoning	50.00	35%	17.50	2	50.0%
Options Trading	51.25	35%	17.94	2	0.0%
Overall	**60.44**	**100%**	**60.44**	**10**	**80.0%**

The unified scoring methodology produces an overall score of 60.44/100, computed as the weighted sum of section contributions. Knowledge Retrieval (data extraction tasks) scores highest at 83.33, while Analytical Reasoning (logic puzzles) and Options Trading (derivatives) both score around 50.

7.2.1 Options Evaluation Breakdown

Table 8: Options Task Performance by Category

Task	Category	P&L	Greeks	Strategy	Risk
strategy_001	Strategy Construction	100	30	85	70
greeks_002	Greeks Analysis	80	0	60	60
Average		**90**	**15**	**72.5**	**65**

Table 9: Options Final Scores (Weighted Average)

Task ID	Raw Score	Normalized
strategy_001 (Iron Condor SPX)	71.25/100	0.7125
greeks_002 (Portfolio Delta)	50.0/100	0.500
Options Average	**60.62/100**	**0.606**

The results reveal several key patterns:

- **P&L Strength**: The agent excels at profit/loss calculations (90/100 average), correctly identifying max profit, max loss, and breakeven points.

- **Greeks Gap**: Explicit Greeks calculations remain challenging (15/100), with the agent discussing concepts without extracting numerical values.
- **Strategy Competence**: Strong performance on strategy construction (72.5/100), demonstrating understanding of multi-leg option structures.
- **Risk Awareness**: Moderate risk management scoring (65/100), with hedging strategies discussed but position sizing underspecified.

7.3 BizFinBench Detailed Results

Table 10: BizFinBench v2 Performance by Task Type

Task Type	Examples	Correct	Accuracy
Event Logic Reasoning	3	3	100%
Financial Quantitative Computation	3	1	33.3%
BizFinBench Total	**6**	**4**	**66.67%**

The agent demonstrates strong logical reasoning (100% on event ordering) but struggles with precise numerical calculations (33.3% on quantitative tasks), where small deviations exceed the 1% tolerance threshold.

7.4 Public CSV Detailed Results

Table 11: Public CSV Dataset Performance

Question Category	Correctness	Score	Result
Market Analysis (US Steel)	4/4	1.0	Correct
Beat or Miss (TJX Margin)	0/2	0.0	Incorrect
Public CSV Total		**0.50**	**50%**

The rubric-based evaluation reveals that qualitative analysis questions (market context) score higher than quantitative beat/miss questions requiring specific BPS calculations.

7.5 Analytical Reasoning Results

Table 12: Synthetic Questions Performance by Topic

Topic	Question	Score	Result
Time Value of Money	Present value comparison	100.0	Correct
Fixed Income	Perpetuity price change	0.0	Incorrect
Average		**50.0**	**50%**

The agent correctly solved the present value comparison problem (choosing between $10,000 today vs. $12,500 in 3 years at 8% discount rate) but failed the perpetuity pricing question requiring calculation of percentage price change when interest rates move from 5% to 6%. This pattern—success on simpler TVM problems, failure on more abstract bond math—is consistent with findings in other sections.

8 Discussion

8.1 Key Findings

The unified section-based evaluation reveals consistent patterns across all three benchmark sections:

1. **Section Performance Hierarchy**: Knowledge Retrieval (83.33) significantly outperforms both Analytical Reasoning (50.00) and Options Trading (51.25). This suggests agents are better at extracting and reporting financial data than performing multi-step logical reasoning or complex derivative calculations.

2. **Conceptual vs. Computational Gap**: Within each section, agents demonstrate stronger conceptual understanding than precise numerical execution. In Options, P&L calculations score 80/100 while Greeks precision drops to 15/100. In Analytical Reasoning, simple TVM problems score 100% while perpetuity math fails completely.

3. **Weighted Scoring Reveals True Capability**: The 30/35/35 weighting ensures that overall scores reflect balanced capability. An agent excelling only at data retrieval (Knowledge) cannot achieve high overall scores without competence in reasoning (Analysis) and derivatives (Options).

4. **Synthetic Questions Fill a Gap**: The Analytical Reasoning section tests skills not covered by BizFinBench or Options Alpha—self-contained logic puzzles requiring financial domain knowledge but no data retrieval. The 50% accuracy suggests significant room for improvement.

5. **Options 4-Dimension Scoring Differentiates**: The granular options breakdown (P&L, Greeks, Strategy, Risk) reveals that aggregate scores mask important capability differences. An agent scoring 51.25/100 on Options may excel at P&L (80) while failing Greeks (15).

8.2 Limitations

- Ground truth for subjective tasks (macro analysis) relies on reference summaries
- Options pricing assumes Black-Scholes model validity
- Adversarial debate quality depends on counter-argument generation

8.3 Future Work

- Extend to multi-agent trading simulations
- Incorporate stochastic volatility models
- Add real-time market data integration
- Develop specialized evaluators for emerging asset classes

9 Conclusion

We presented CIO-Agent FAB++, a comprehensive benchmark for evaluating AI finance agents across three weighted sections: Knowledge Retrieval (30%), Analytical Reasoning (35%), and Options Trading (35%). Our key contributions include:

- **Unified Section-Based Scoring**: A weighted scoring system that normalizes all evaluator outputs to 0-100 and combines them into a single overall score, enabling meaningful comparison across diverse financial tasks.

- **Olympiad-Style Synthetic Questions**: 22 curated finance logic problems spanning capital budgeting, portfolio theory, fixed income, corporate finance, and derivatives—testing reasoning ability without data retrieval dependency.

- **4-Dimension Options Scoring**: The Options Alpha Challenge provides granular assessment across P&L accuracy, Greeks precision, strategy quality, and risk management—revealing capability patterns masked by aggregate scores.

- **Dynamic Weight Redistribution**: When sections are disabled, weights redistribute proportionally, ensuring meaningful scores regardless of evaluation configuration.

- **Empirical Validation**: Unified evaluation of a baseline GPT-4o mini agent demonstrates 60.44/100 overall score with clear section hierarchy: Knowledge Retrieval (83.33) > Analytical Reasoning (50.00) ≈ Options Trading (51.25).

The section-based scoring methodology reveals that current AI agents excel at financial data extraction but struggle with multi-step logical reasoning and precise derivative calculations—key areas for future improvement. The system is publicly available at `https://github.com/yxc20089/AgentBusters` with Docker images for immediate deployment:

```
ghcr.io/yxc20089/agentbusters-green:latest
ghcr.io/yxc20089/agentbusters-purple:latest
```

Acknowledgments

We thank the AgentBeats Competition organizers at Berkeley RDI for inspiring this work. We acknowledge the contributions of the A2A Protocol and MCP communities for enabling standardized agent communication.

References

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1-2):125–144.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Bigeard, A., Nashold, L., Krishnan, R., and Wu, S. (2025). Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks. *arXiv preprint arXiv:2508.00828*. `https://arxiv.org/abs/2508.00828`.

Lu, G., Guo, X., Zhang, R., Zhu, W., and Liu, J. (2025). BizFinBench.v2: A Unified Dual-Mode Bilingual Benchmark for Expert-Level Financial Capability Alignment. *arXiv preprint arXiv:2601.06401*. https://arxiv.org/abs/2601.06401.

Google and Linux Foundation (2025). Agent-to-Agent Protocol: An open protocol enabling communication and interoperability between opaque agentic applications. https://github.com/a2aproject/A2A.

Anthropic (2024). Model Context Protocol. https://modelcontextprotocol.io/.

A Alpha Score Derivation

The Alpha Score is designed to reward accurate, robust, and efficient agent responses:

$$\alpha = \frac{R \cdot D}{C \cdot P} \tag{14}$$

where:

- $R = \text{RoleScore} \in [0, 100]$
- $D = \text{DebateMultiplier} \in [0.8, 1.2]$
- $C = \ln(1 + \text{Cost})$ (logarithmic cost penalty)
- $P = 1 + \text{LookaheadPenalty}$ (temporal violation penalty)

The logarithmic cost penalty ensures diminishing returns for expensive computations, while the lookahead penalty harshly penalizes agents that access future information.

B MCP Server API Reference

B.1 Options Chain Server

Listing 1: Options Chain MCP Tools

```
# Get options chain for a ticker
get_options_chain(ticker: str, expiration: str) -> dict

# Calculate Black-Scholes price
calculate_option_price(
    spot: float, strike: float, rate: float,
    volatility: float, time_to_expiry: float,
    option_type: str, dividend_yield: float
) -> dict  # Returns price and all Greeks

# Get implied volatility surface
get_iv_surface(ticker: str) -> dict

# Analyze multi-leg strategy
analyze_strategy(legs: list[dict]) -> dict
```

B.2 Risk Metrics Server

Listing 2: Risk Metrics MCP Tools

```
# Calculate portfolio Greeks
calculate_portfolio_greeks(positions: list[dict]) -> dict

# Calculate Value at Risk
calculate_var(
    returns: list[float], confidence: float,
    method: str  # "historical", "parametric", "monte_carlo"
) -> dict

# Run stress test
run_stress_test(
    portfolio: dict,
    scenarios: list[dict]  # e.g., {"name": "crash", "spot_change": -0.20}
) -> dict
```

C Evaluation Configuration

Listing 3: Sample Evaluation Config (YAML)

```
name: "FAB++␣Full␣Evaluation"
datasets:
  - type: synthetic
    path: data/synthetic_questions/questions.json
    limit: 50
  - type: bizfinbench
    path: data/BizFinBench.v2
    task_types: [event_logic_reasoning, financial_quantitative_computation]
    languages: [en]
    limit_per_task: 20
  - type: public_csv
    path: finance-agent/data/public.csv
    limit: 100
sampling:
  strategy: stratified
  total_limit: 100
  seed: 42
```

D Complete Synthetic Question Bank

The following 20 questions comprise the Analytical Reasoning section. Questions are organized by topic with difficulty ratings (E=Easy, H=Hard, X=Expert).

D.1 Data Retrieval Questions (2)

1. **[E] Quantitative Retrieval**: What was AAPL's EBITDA in fiscal year 2024?
   Answer: $134.66B

2. **[E] Qualitative Retrieval**: Describe AAPL's main business and products.
   Answer: Apple Inc. designs, manufactures, and markets smartphones (iPhone), tablets (iPad), computers (Mac), wearables (Apple Watch), and provides digital services.

D.2 Capital Budgeting (2)

3. **[H] NPV Crossover Rate**: A company has two mutually exclusive projects. Project A requires \$100,000 investment and returns \$150,000 in Year 1. Project B requires \$100,000 investment and returns \$180,000 in Year 2. At what discount rate are the two projects equally attractive?
   Answer: 20%

4. **[H] FCFF Calculation**: Company ABC has EBIT of \$10 million, depreciation of \$2 million, capital expenditures of \$3 million, and working capital increase of \$1 million. The tax rate is 25%. What is the Free Cash Flow to Firm?
   Answer: \$5.5 million

D.3 Portfolio Theory (3)

5. **[H] Beta Adjustment**: An investor holds Stock X (60% weight, β=1.2) and Stock Y (40% weight, β=0.8). To reduce portfolio beta to 1.0 by adjusting only Stock X weight (remainder in risk-free), what should be the new weight of Stock X?
   Answer: 50% (or 83.3% depending on interpretation)

6. **[X] Leverage & Standard Deviation**: Stock XYZ has expected return 12% and σ=25%. Risk-free rate is 4%. To achieve 16% expected return using XYZ and risk-free, what is the portfolio's standard deviation?
   Answer: 37.5%, Leverage: 1.5x

7. **[H] Combined Leverage**: A company has DOL=2.5 and DFL=1.6. If sales increase by 10%, by what percentage will EPS change?
   Answer: 40%

D.4 Fixed Income (4)

8. **[X] Bond Pricing Arbitrage**: A zero-coupon bond (\$1,000 face, 5-year) trades at \$680. A 6% coupon bond trades at par. What is the arbitrage-free price of a 5-year 8% coupon bond?
   Answer: \$1,085.35

9. **[X] Duration Immunization**: A pension fund has 15-year liability duration. Bond A has 5-year duration, Bond B has 20-year duration. What percentage in Bond B to immunize?
   Answer: 66.67%

10. **[H] Perpetuity Price Change**: A perpetual bond pays \$50 annually. If rates rise from 5% to 6%, what is the percentage price change?
    Answer: -16.67%

11. **[H] Gordon Growth Model**: Stock just paid \$2.00 dividend, growth rate 6%, required return 10%. What is intrinsic value?
    Answer: \$53.00

D.5 Corporate Finance (3)

12. **[H] Leverage Effect on ROE**: Firm has $500M assets, D/E=1.5, 6% interest, 30% tax, ROA=10%. What is ROE?
    Answer: 11.2% (or 18.1% depending on formula used)

13. **[X] Modigliani-Miller Homemade Leverage**: Firm A is all-equity (1M shares at $50). Firm B has $20M debt, $30M equity. How can an investor owning 10% of A replicate 10% of B's equity?
    Answer: Borrow $2M, invest $5M in A, net investment $3M

14. **[H] Stock Split**: Company has 100,000 shares at $25. After 3-for-2 split, what are new price and shares outstanding?
    Answer: $16.67/share, 150,000 shares

D.6 Options & Derivatives (4)

15. **[X] Bull Call Spread**: Buy $100 call for $8, sell $110 call for $3. What are max profit, max loss, and breakeven?
    Answer: Max Profit $5, Max Loss $5, Breakeven $105

16. **[H] Risk-Neutral Probability**: Stock at $100 can go up 20% or down 15%. Risk-free rate 5%. What is risk-neutral probability of up move?
    Answer: 57.14%

17. **[X] Put-Call Parity**: Call priced at $8, stock at $100, strike $95, risk-free 5%, 1-year expiry. What should put price be?
    Answer: $1.37 (or negative indicating arbitrage)

18. **[X] Interest Rate Swap**: X borrows at 8% fixed or LIBOR+1%. Y borrows at 10% fixed or LIBOR+2%. If they swap (X floating, Y fixed) and split gains equally, what fixed rate does X pay?
    Answer: 7.50%

D.7 Time Value of Money (2)

19. **[H] Present Value Comparison**: Choose between $10,000 today or $12,500 in 3 years at 8% discount rate. Which is better and by how much?
    Answer: Option A ($10,000 today) better by $75.15

D.8 Forex (1)

20. **[X] Covered Interest Arbitrage**: Spot EUR/USD=1.10, 1-year forward=1.12, USD rate=5%, EUR rate=3%. Is there arbitrage? Calculate profit per $1M.
    Answer: Yes, approximately $7,273 profit per $1M