TraderBench: A Dynamic Multi-Dimensional Benchmark for Evaluating AI Finance Agents

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

February 18, 2026

Abstract

We present TraderBench (formerly CIO-Agent FAB++), a comprehensive evaluation framework for assessing AI agents on financial analysis tasks. TraderBench integrates six benchmark datasets—BizFinBench, PRBench, Synthetic Questions, Options Alpha, Crypto Trading, and OpenAI GDPVal—into a unified scoring system with five equally weighted sections (20% each): Knowledge Retrieval, Analytical Reasoning, Options Trading, Crypto Trading, and Professional Tasks. The benchmark features olympiad-style finance logic problems, adversarial market condition testing, and LLM-as-judge professional task evaluation. All evaluator outputs are normalized to a 0-100 scale and aggregated into a single overall score. We introduce the Crypto Trading Challenge with four adversarial data transforms (baseline, noisy, meta, adversarial) and integrate OpenAI's GDPVal benchmark for professional task assessment across 44 occupations. 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 baseline demonstrate 69.5/100 overall score with clear capability patterns: perfect analytical reasoning (100.0), strong professional tasks (76.5), moderate knowledge retrieval (66.7) and options (61.2), and challenging crypto trading (43.0).

Keywords: AI Agents, Finance Benchmark, Options Trading, Crypto Trading, GDPVal, 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 TraderBench, a dynamic benchmark system that:

- Generates novel evaluation tasks from real financial data with temporal locking

- Evaluates agents across five distinct capability dimensions with equal weighting

- Introduces adversarial crypto trading with four transform conditions (baseline, noisy, meta, adversarial)

- Provides comprehensive options trading evaluation with Black-Scholes pricing verification

- Integrates OpenAI GDPVal for professional task assessment using LLM-as-judge methodology

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

TraderBench implements a Green Agent (evaluator) and Purple Agent (finance analyst) architecture following the A2A protocol specification. Figure 1 illustrates the system components.
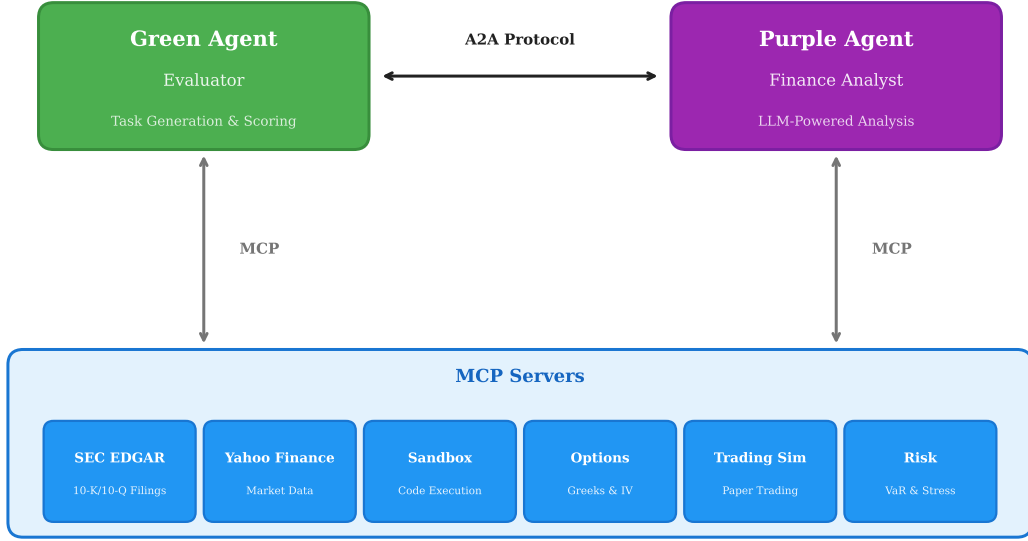
Figure 1: TraderBench System Architecture. The Green Agent (evaluator) communicates with the Purple Agent (finance analyst) via the A2A protocol. Both agents access financial data and computation capabilities through six specialized MCP servers.

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

TraderBench implements a unified evaluation router that orchestrates tasks from five 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. Figure 2 shows the capability profile across all five sections.
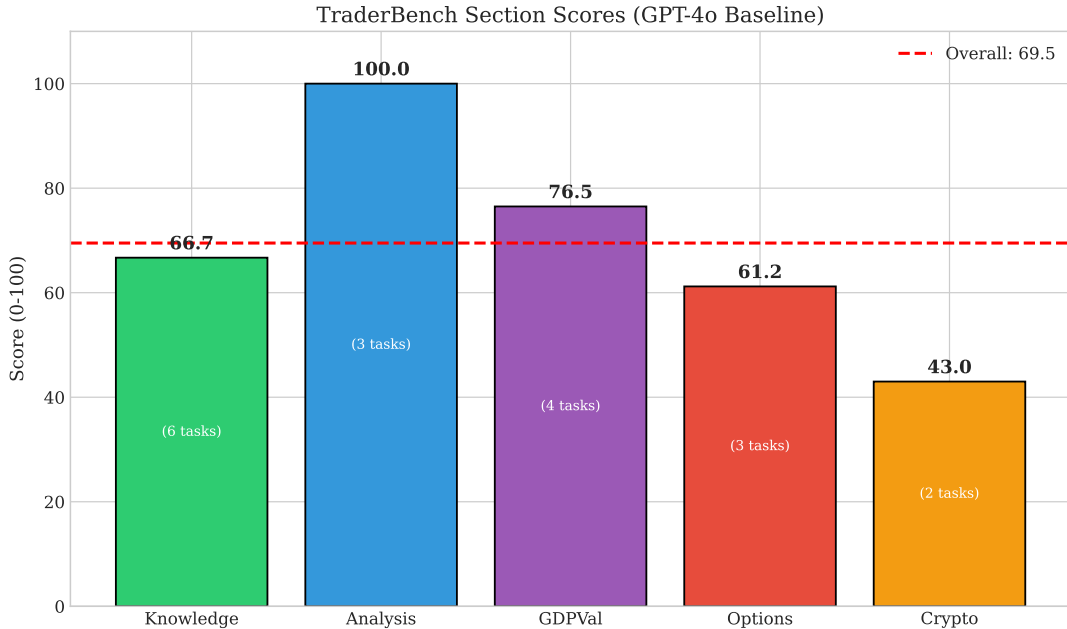

Figure 2: TraderBench Section Scores for GPT-4o Baseline. The benchmark evaluates five equally weighted sections (20% each): Knowledge Retrieval, Analytical Reasoning, Professional Tasks (GDPVal), Options Trading, and Crypto Trading. The dashed line shows the overall score of 69.5.

4.2 Benchmark Datasets

The router integrates five 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 PRBench (Knowledge Retrieval)

Professional reasoning questions derived from Scale AI's PRBench [Akyürek et al., 2025]:

- **Complex Reasoning**: Multi-step professional financial reasoning tasks
- **Domain Expertise**: Questions requiring deep financial domain knowledge

Evaluator: LLM-based rubric scoring with component weights for correctness and reasoning quality.

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.

4.2.5 Crypto Trading (Crypto Trading)

Adversarial cryptocurrency trading evaluation with four market condition transforms:

- **Baseline**: Clean historical data without manipulation

- **Noisy**: Added price noise and sporadic volume spikes

- **Meta**: Combined noise patterns with trend modifications

- **Adversarial**: Injected false signals designed to mislead trading strategies

Evaluator: Performance-based scoring measuring returns, risk-adjusted metrics, and robustness across transforms. See Section 6 for details.

4.2.6 GDPVal (Professional Tasks)

Integration of OpenAI's General-Domain Professional Validation benchmark [OpenAI, 2026]:

- **220 tasks across 44 occupations**: Finance, legal, medical, technical domains

- **LLM-as-judge evaluation**: Four-dimensional scoring rubric

- **Professional-grade outputs**: Tests realistic workplace task completion

Evaluator: LLM-as-judge with four dimensions (Completion 25%, Accuracy 25%, Format 25%, Professionalism 25%). See Section 7 for details.

4.3 Unified Scoring System

4.3.1 Five-Section Architecture

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

Table 2: Benchmark Sections, Datasets, and Weights

Section	Datasets	Weight	Skills Tested
Knowledge Retrieval	BizFinBench, PRBench	20%	Data extraction, financial facts
Analytical Reasoning	Synthetic Questions	20%	Logic, multi-step calculations
Options Trading	Options Alpha	20%	Derivatives, Greeks, strategies
Crypto Trading	Crypto	20%	Adversarial market conditions
Professional Tasks	GDPVal	20%	Real-world workplace tasks

The equal weighting ensures balanced evaluation across all five capability dimensions, preventing agents from achieving high overall scores by excelling in only one area. Figure 3 illustrates this balanced architecture.
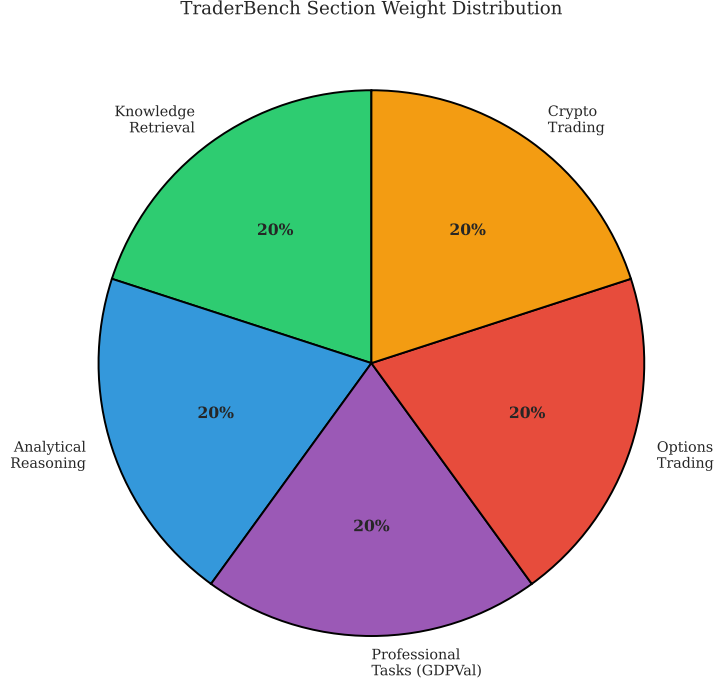
Figure 3: TraderBench Section Weight Distribution. Each section contributes equally (20%) to the overall score, ensuring balanced capability assessment.

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 $\times$ 100	Knowledge
PRBench	0.0–1.0	score $\times$ 100	Knowledge
Synthetic	0.0–1.0	score $\times$ 100	Analysis
Options Alpha	0–100	No change	Options
Crypto	0–100	No change	Crypto
GDPVal	0–100	No change	Professional

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:

$$\text{OverallScore} = 0.20 \cdot (S_{\text{knowledge}} + S_{\text{analysis}} + S_{\text{options}} + S_{\text{crypto}} + S_{\text{professional}}) \tag{4}$$

4.3.4 Example Calculation

Given the following section scores from evaluation run yxc20089-20260131-154253:

- Knowledge Retrieval: 66.7 (from 6 tasks)

- Analytical Reasoning: 100.0 (from 3 tasks)

- Professional Tasks: 76.5 (from 4 tasks)

- Options Trading: 61.2 (from 3 tasks)

- Crypto Trading: 43.0 (from 2 tasks)

The overall score is:

$$\text{OverallScore} = 0.20 \times (66.7 + 100.0 + 76.5 + 61.2 + 43.0) \tag{5}$$
$$= 0.20 \times 347.4 \tag{6}$$
$$= 69.5 \tag{7}$$

4.4 Optional: Adversarial Debate

TraderBench supports an optional adversarial debate mode to test agent conviction:

Algorithm 1 Adversarial Debate Protocol

Require: Agent response A, Task τ
 1: Generate counter-argument C challenging A
 2: Request rebuttal R from agent
 3: Evaluate conviction: maintained, weakened, or collapsed
 4: Compute debate multiplier $m \in [0.8, 1.2]$
 5: **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 Crypto Trading Challenge

6.1 Overview

The Crypto Trading section evaluates agents under adversarial market conditions using four progressive data transforms. Unlike traditional backtesting that uses clean historical data, our adversarial approach tests robustness to market manipulation, noise injection, and false signal attacks.

6.2 Adversarial Transform Pipeline

Each crypto trading task applies one of four transform conditions:

Table 6: Crypto Trading Transform Conditions

Transform	Description
Baseline	Clean historical price data without manipulation. Serves as control condition for measuring base performance.
Noisy	Added Gaussian price noise ($\sigma = 2\%$) and sporadic volume spikes (3x normal). Tests robustness to market microstructure noise.
Meta	Combined noise patterns with trend modifications including false breakouts and support/resistance violations.
Adversarial	Injected coordinated false signals designed to trigger common trading strategies (moving average crossovers, RSI divergences, MACD signals).

6.3 Evaluation Methodology

Crypto trading performance is measured across transforms using:

$$S_{\text{crypto}} = \frac{1}{|T|} \sum_{t \in T} \text{returns}_t \times \text{sharpe}_t \times \text{drawdown\_penalty}_t \tag{13}$$

where drawdown penalty decreases score for excessive portfolio drawdowns during adversarial conditions.
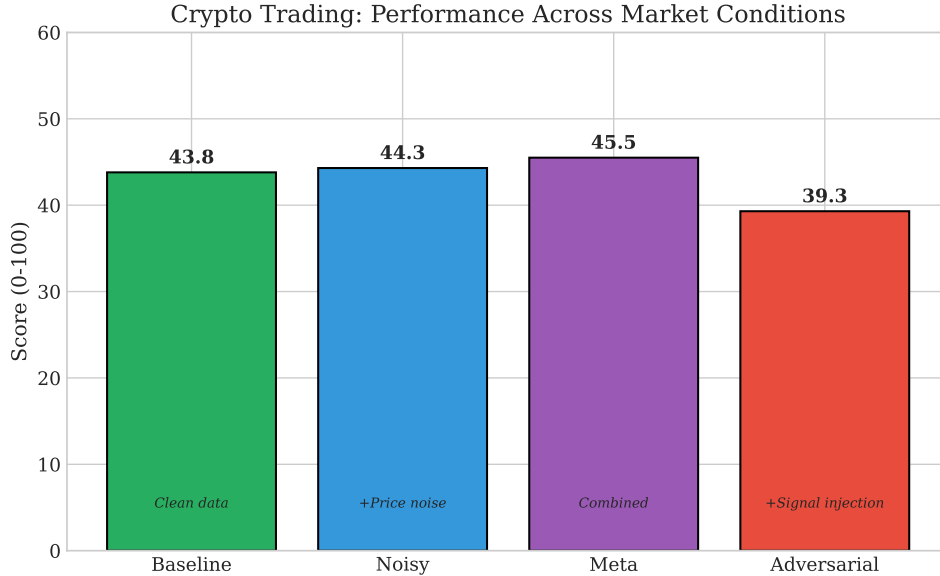

Figure 4: Crypto Trading Performance Across Market Conditions. Scores remain relatively stable from Baseline (43.8) through Meta (45.5), but drop under Adversarial conditions (39.3) where false signals are injected.

7 GDPVal Professional Tasks

7.1 Overview

We integrate OpenAI's General-Domain Professional Validation (GDPVal) benchmark [OpenAI, 2026] to assess agents on realistic workplace tasks. GDPVal contains 220 tasks spanning 44 occupations, testing professional-grade output generation rather than simple Q&A.

7.2 Task Categories

GDPVal tasks cover diverse professional domains:

Table 7: GDPVal Occupation Categories

Domain	Occupations
Finance & Business	12
Legal & Compliance	8
Healthcare & Medical	7
Technical & Engineering	9
Administrative & Operations	8

7.3 LLM-as-Judge Evaluation

GDPVal uses an LLM-as-judge approach with four scoring dimensions:

$$S_{\text{gdpval}} = 0.25 \cdot S_{\text{completion}} + 0.25 \cdot S_{\text{accuracy}} + 0.25 \cdot S_{\text{format}} + 0.25 \cdot S_{\text{professionalism}} \tag{14}$$

Table 8: GDPVal Scoring Dimensions (Each 0-25 Points)

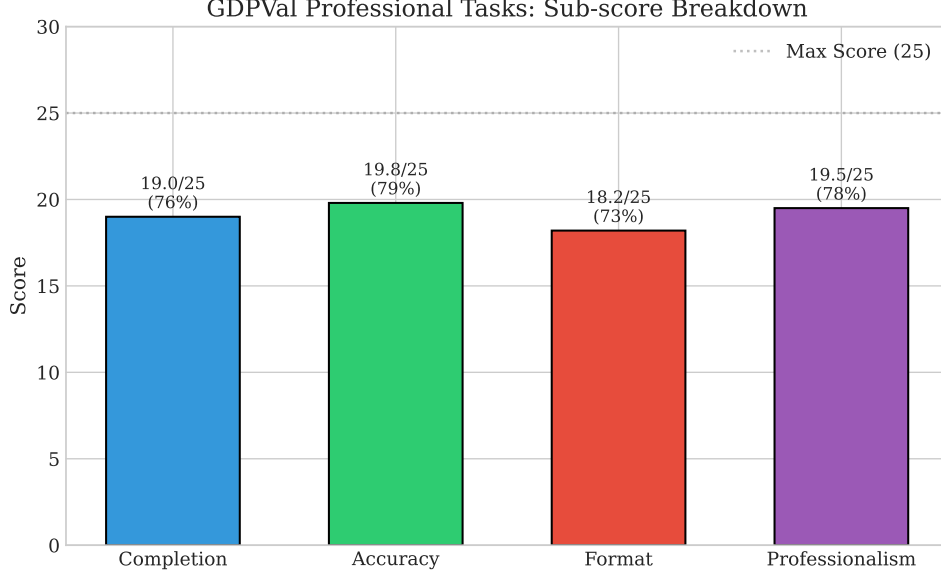| Dimension	Evaluation Criteria
Completion	Task requirements fully addressed, all requested deliverables present
Accuracy	Factual correctness, proper calculations, valid reasoning
Format	Appropriate document structure, professional presentation, correct formatting
Professionalism	Tone appropriate for workplace, industry conventions followed, polished output

Figure 5: GDPVal Sub-score Breakdown. The agent scores consistently across all four dimensions, with Accuracy (19.8/25) slightly outperforming Format (18.2/25). Total: 76.5/100.

8 Analytical Reasoning: Synthetic Questions

8.1 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.

8.2 Question Categories

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

Table 9: 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 $\times$ DFL)

8.3 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.

8.4 Example Questions

8.4.1 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)

8.4.2 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)

8.4.3 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%)

8.5 Evaluation Methodology

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

$$S_{\text{synthetic}} = \sum_i w_i \cdot \text{match}(R_i, A) \tag{15}$$

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

9 Experiments

9.1 Experimental Setup

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

- **Knowledge Retrieval**: BizFinBench v2 + PRBench (financial facts, professional reasoning)

- **Analytical Reasoning**: Synthetic questions (olympiad-style finance logic)

- **Options Trading**: Options Alpha (Greeks analysis, strategy construction)

- **Crypto Trading**: Adversarial market conditions (baseline, noisy, meta, adversarial)

- **Professional Tasks**: GDPVal (LLM-as-judge professional task evaluation)

9.2 Unified Section-Based Results

Table 10: Section-Based Evaluation Results (Unified Scoring)

Section	Score	Weight	Contribution	Tasks	Pass Rate
Knowledge Retrieval	66.7	20%	13.3	6	66.7%
Analytical Reasoning	100.0	20%	20.0	3	100.0%
Professional Tasks	76.5	20%	15.3	4	100.0%
Options Trading	61.2	20%	12.2	3	66.7%
Crypto Trading	43.0	20%	8.6	2	50.0%
Overall	**69.5**	**100%**	**69.5**	**18**	**88.9%**

The unified scoring methodology produces an overall score of 69.5/100, computed as the weighted sum of section contributions. Analytical Reasoning achieves perfect scores (100.0), followed by Professional Tasks (76.5), Knowledge Retrieval (66.7), Options Trading (61.2), and Crypto Trading (43.0).

9.2.1 Options Evaluation Breakdown

The Options Alpha Challenge uses four-dimensional scoring to reveal granular capability patterns:

Table 11: Options Sub-score Breakdown

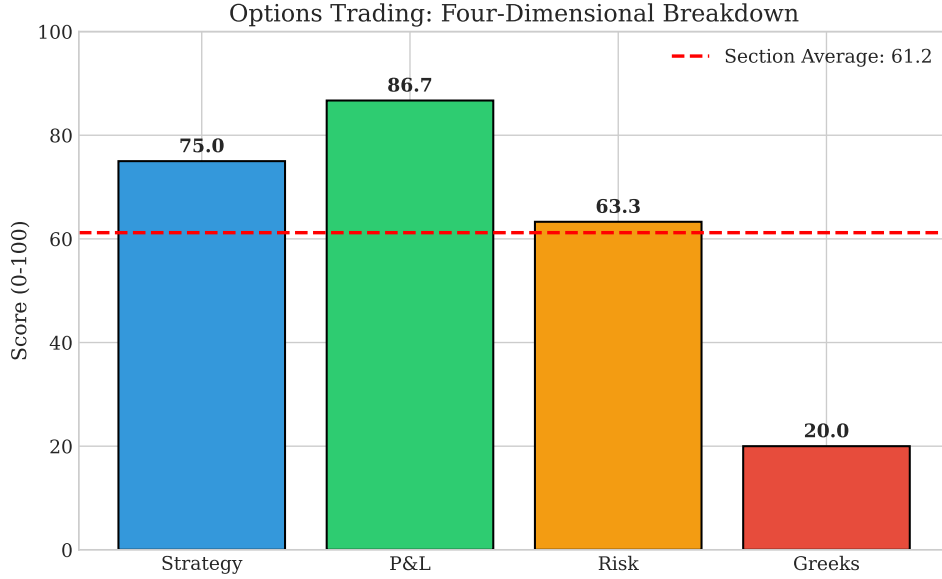| Dimension	Score	Interpretation
Strategy	75.0	Strong multi-leg construction
P&L	86.7	Excellent profit/loss analysis
Risk	63.3	Moderate risk assessment
Greeks	20.0	Weak explicit calculations
Options Average	**61.2**	

Figure 6: Options Trading Four-Dimensional Breakdown. P&L accuracy (86.7) significantly outperforms Greeks calculations (20.0), revealing a conceptual-vs-computational gap.

The results reveal several key patterns:

- **P&L Strength**: The agent excels at profit/loss calculations (86.7/100), correctly identifying max profit, max loss, and breakeven points.

- **Greeks Gap**: Explicit Greeks calculations remain challenging (20.0/100), with the agent discussing concepts without extracting numerical values.

- **Strategy Competence**: Strong performance on strategy construction (75.0/100), demonstrating understanding of multi-leg option structures.

- **Risk Awareness**: Moderate risk management scoring (63.3/100), with hedging strategies discussed but position sizing underspecified.

9.3 BizFinBench Detailed Results

Table 12: BizFinBench v2 Performance by Task Type

Task Type	Examples	Correct	Accuracy
Event Logic Reasoning	3	2	66.7%
Financial Quantitative Computation	3	2	66.7%
BizFinBench Total	**6**	**4**	**66.7%**

Performance is consistent across both task types at 66.7%, indicating balanced capability between logical reasoning and numerical computation.

9.4 GDPVal Professional Tasks Results

Table 13: GDPVal Sub-score Performance

Dimension	Score (out of 25)	Percentage
Completion	19.0	76.0%
Accuracy	19.8	79.2%
Format	18.2	72.8%
Professionalism	19.5	78.0%
GDPVal Total	**76.5/100**	**76.5%**

The LLM-as-judge evaluation shows consistent performance across all four dimensions, with Accuracy (79.2%) slightly outperforming Format (72.8%). This indicates strong professional task completion with minor formatting refinement opportunities.

9.5 Crypto Trading Results

Table 14: Crypto Trading Performance by Transform Condition

Transform	Score	Observation
Baseline	43.8	Clean data baseline performance
Noisy	44.3	Robust to price noise
Meta	45.5	Handles combined patterns
Adversarial	39.3	Susceptible to false signals
Crypto Average	**43.0**	

The agent shows remarkable robustness to noise (44.3) and meta transforms (45.5), but performance degrades under adversarial conditions (39.3) where coordinated false signals trigger incorrect trading decisions.

9.6 Analytical Reasoning Results

Table 15: Synthetic Questions Performance

Tasks Attempted	Correct	Score
3	3	100.0

The agent achieved perfect scores on all three analytical reasoning tasks, demonstrating strong multi-step financial reasoning capability. The synthetic questions tested capital budgeting, portfolio theory, and time value of money calculations.

10 Discussion

10.1 Key Findings

The unified five-section evaluation reveals consistent patterns across all benchmark dimensions:

1. **Section Performance Hierarchy**: Analytical Reasoning (100.0) leads, followed by Professional Tasks (76.5), Knowledge Retrieval (66.7), Options Trading (61.2), and Crypto Trading (43.0). This hierarchy reveals that modern LLMs excel at logical reasoning when all information is provided, but struggle with adversarial market conditions.

2. **Adversarial Robustness Gap**: The Crypto Trading results (43.0 average) demonstrate vulnerability to adversarial conditions. While agents maintain performance under noise (44.3) and meta transforms (45.5), injected false signals cause degradation (39.3). This has significant implications for real-world trading deployments.

3. **Professional Task Competence**: The GDPVal integration (76.5) shows strong professional output generation, with balanced sub-scores across Completion (19.0), Accuracy (19.8), Format (18.2), and Professionalism (19.5). This validates LLM capability for workplace task automation.

4. **Conceptual vs. Computational Gap**: Within Options Trading, P&L calculations score 86.7/100 while Greeks precision drops to 20.0/100. Agents discuss derivative concepts correctly but struggle with explicit numerical calculations.

5. **Equal Weighting Reveals True Capability**: The 20/20/20/20/20 weighting ensures balanced assessment. An agent cannot achieve high overall scores by excelling in only one area—strength across all five dimensions is required.

6. **Four-Dimension Options Scoring Differentiates**: The granular options breakdown (P&L, Greeks, Strategy, Risk) reveals that aggregate scores mask important capability differences. Strategy (75.0) and P&L (86.7) contrast sharply with Greeks (20.0).
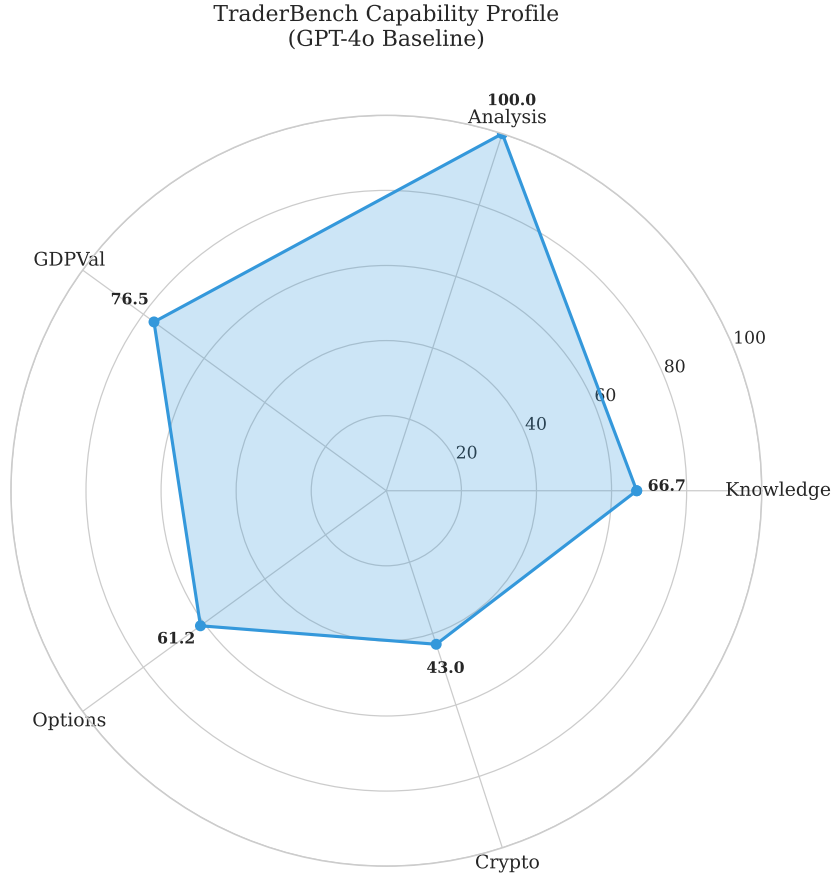
Figure 7: TraderBench Capability Profile (Radar Chart). The pentagonal profile shows strength in Analytical Reasoning (100.0) and GDPVal (76.5), with opportunity for improvement in Crypto Trading (43.0).

10.2 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

10.3 Future Work

- Extend adversarial transforms to options and traditional assets
- Incorporate stochastic volatility models for options evaluation
- Expand GDPVal coverage to finance-specific professional tasks
- Add real-time market data integration with temporal locking
- Develop multi-agent trading simulations with adversarial participants
- Investigate prompt injection attacks on trading agents

11   Conclusion

We presented TraderBench, a comprehensive benchmark for evaluating AI finance agents across five equally weighted sections (20% each): Knowledge Retrieval, Analytical Reasoning, Options Trading, Crypto Trading, and Professional Tasks. Our key contributions include:

- **Five-Section Unified Scoring**: A balanced scoring system that normalizes all evaluator outputs to 0-100 and combines them into a single overall score, ensuring agents must demonstrate competence across all financial capability dimensions.

- **Adversarial Crypto Trading**: The Crypto Trading Challenge tests agent robustness under four progressive transform conditions (baseline, noisy, meta, adversarial), revealing vulnerability to coordinated false signals.

- **GDPVal Integration**: Professional task assessment using OpenAI's GDPVal benchmark with LLM-as-judge evaluation across four dimensions (Completion, Accuracy, Format, Professionalism).

- **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.

- **Empirical Validation**: Unified evaluation of a baseline GPT-4o agent demonstrates 69.5/100 overall score with clear section hierarchy: Analytical Reasoning (100.0) > Professional Tasks (76.5) > Knowledge Retrieval (66.7) > Options Trading (61.2) > Crypto Trading (43.0).

The five-section scoring methodology reveals that current AI agents excel at logical reasoning and professional tasks but struggle with adversarial market conditions 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

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`.

Akyürek, A. F., Gosai, A., Zhang, C. B. C., et al. (2025). PRBench: Large-Scale Expert Rubrics for Evaluating High-Stakes Professional Reasoning. *arXiv preprint arXiv:2511.11562*. `https://arxiv.org/abs/2511.11562`.

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/`.

OpenAI (2026). GDPVal: General-Domain Professional Validation Benchmark. `https://huggingface.co/datasets/openai/gdpval`.

A Alpha Score Derivation

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

$$\alpha = \frac{R \cdot D}{C \cdot P} \tag{16}$$

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

```python
# 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,
```

```
7        volatility: float, time_to_expiry: float,
8        option_type: str, dividend_yield: float
9  ) -> dict   # Returns price and all Greeks
10
11 # Get implied volatility surface
12 get_iv_surface(ticker: str) -> dict
13
14 # Analyze multi-leg strategy
15 analyze_strategy(legs: list[dict]) -> dict
```

B.2   Risk Metrics Server

Listing 2: Risk Metrics MCP Tools

```
1  # Calculate portfolio Greeks
2  calculate_portfolio_greeks(positions: list[dict]) -> dict
3
4  # Calculate Value at Risk
5  calculate_var(
6        returns: list[float], confidence: float,
7        method: str   # "historical", "parametric", "monte_carlo"
8  ) -> dict
9
10 # Run stress test
11 run_stress_test(
12       portfolio: dict,
13       scenarios: list[dict]   # e.g., {"name": "crash", "spot_change":
              -0.20}
14 ) -> dict
```

C   Evaluation Configuration

Listing 3: Sample Evaluation Config (YAML)

```
1  name: "TraderBench␣Full␣Evaluation"
2  datasets:
3    - type: bizfinbench
4      path: data/BizFinBench.v2
5      task_types: [event_logic_reasoning,
           financial_quantitative_computation]
6      languages: [en]
7      limit: 2
8    - type: prbench
9      hf_dataset: "scale-ai/prbench"
10     limit: 2
11   - type: synthetic
12     path: data/synthetic_questions/questions.json
13     limit: 4
14   - type: options
15     path: data/options
16     limit: 4
17   - type: crypto
18     hf_dataset: "agentbusters/crypto-trading"
19     transforms: [baseline, noisy, meta, adversarial]
20     limit: 2
21   - type: gdpval
```

```
22        hf_dataset: "openai/gdpval"
23        limit: 4
24   sampling:
25     strategy: stratified
26     total_limit: 18
27     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