Gold-Agent: A Multi-Agent Framework for Autonomous Gold Trading with Institutional-Grade Risk Management

He Chenyu (Daniel He)
College of Computer Science and Technology, Zhengjiang University
Hangzhou, China
hechenyu@zju.edu.cn
ORCID: https://orcid.org/0009-0008-0422-7567

December 6, 2025

Abstract

Large Language Models (LLMs) have demonstrated value in financial reasoning but continue to struggle with hallucination, operational gaps, and deterministic risk adherence. We introduce *Gold-Agent*, a role-specialized multi-agent framework for autonomous Gold (XAU/USD) trading that codifies an institutional "Corporate" workflow spanning research, strategy, execution, risk, and compliance. Gold-Agent combines LLM-driven analysis with deterministic Hard Risk Gates, circuit breakers, and audit trails, and integrates Retrieval-Augmented Generation (RAG) over macro histories, curated news, and quantitative indicators. Backtests on 2020–2025 gold spot data yield calibrated baselines (+128.9% Buy-and-Hold; +72.1% SMA 50/200, Sharpe $\approx$ 1.0). A December 2025 live-fire simulation shows the risk gate halting execution because of abnormal liquidity spreads and cross-asset correlations, illustrating how the framework prevents unsafe trades. We detail recent enhancements that automatically calibrate liquidity thresholds to user-selected horizons, strengthening robustness for research-driven "morning meeting" scenarios.

Keywords: multi-agent systems, algorithmic trading, risk management, retrieval-augmented generation, institutional workflows

ACM CCS Concepts:

- Applied computing → Economics;
- Computing methodologies → Multi-agent systems;
- Computer systems organization → Reliability.

1 Introduction

The application of LLMs in finance has progressed from sentiment extraction to higher-order decision support. Deployment in real-money trading, however, faces the "trust gap": models are probabilistic and often speculative, whereas markets demand deterministic compliance with risk constraints and operational guardrails. Key challenges include (i) enforcing role accountability across the trade lifecycle, (ii) grounding reasoning in verifiable data, and (iii) preventing unsafe execution during volatile macro events.

We present *Gold-Agent*, a system built on Microsoft AutoGen that models the workflow of an institutional gold desk. By assigning roles such as `RiskManagerAgent` and `ComplianceAgent`, enforcing JSON schemas, and wiring deterministic risk code, Gold-Agent delivers measurable reliability suitable for production experimentation. Our latest version adds dynamic liquidity gating keyed to the lookback window chosen in the morning meeting, closing the loop between user-configured context and automated hard limits.

Contributions. Our work offers four contributions: (1) an institutional workflow alignment capturing five corporate phases (Research → Strategy → Execution → Risk → Operations) with deterministic gating, (2) a reproducible data and domain adaptation pipeline spanning market data, macro narratives, and sentiment news, (3) a risk-aware evaluation protocol that surfaces both trading performance and hard gate telemetry, and (4) a dynamic liquidity calibration rule that sets spread limits to the maximum of configuration values and empirical 95th-percentile spreads for the selected horizon.

2 Related Work

Multi-agent coordination has been shown to outperform single-agent prompting in complex reasoning tasks. FinCon [1] highlights the value of structured dialogues, while persona-conditioned role play [2] underscores the benefits of perspective diversity. Frameworks such as AutoGen [3] standardize agent orchestration yet seldom integrate high-integrity risk modules. FinGPT [4] and similar efforts emphasize signal generation but often omit middle- and back-office controls. Gold-Agent marries these advances with deterministic enforcement inspired by institutional policy manuals, closing the gap between research prototypes and production controls.

3 System Overview

3.1 Agent Society

Gold-Agent orchestrates 12 specialized agents organized into a strict hand-off: the Research Cluster (`DataAgent`, `MacroAnalystAgent`, `FundamentalAnalystAgent`, `QuantResearchAgent`) distills raw data into structured briefings; the Strategy Cluster (`HeadTraderAgent`) synthesizes trade plans; the Execution Cluster (`PaperTraderAgent`) converts plans into executable orders; the Risk & Control Cluster (`RiskManagerAgent`, `ComplianceAgent`) acts as adversarial critics; and the Operations Cluster (`SettlementAgent`, `ScribeAgent`) handles logistics and audit trails.

3.2 Corporate Workflow State Machine

The workflow enforces deterministic transitions: (1) Research Briefing, (2) Plan Formulation, (3) Execution Design, (4) Risk Gate, and (5) Operations Handoff. Each phase must emit a JSON contract comprising `phase`, `status`, `summary`, and structured `details`, preventing prompt drift and enabling machine auditing.

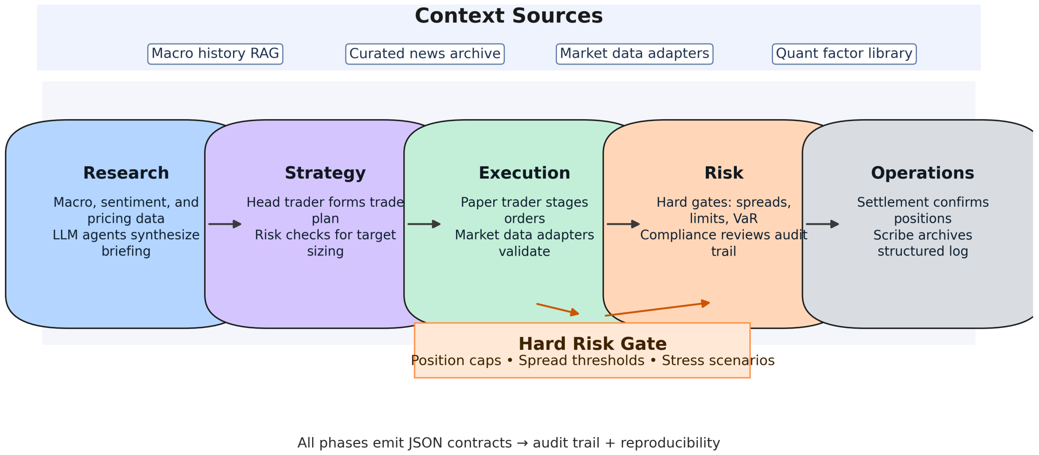
Context Sources
Macro history RAG
Curated news archive
Market data adapters
Quant factor library
Research
Macro, sentiment, and pricing data
LLM agents synthesize briefing
Strategy
Head trader forms trade plan
Risk checks for target sizing
Execution
Paper trader stages orders
Market data adapters validate
Risk
Hard gates: spreads, limits, VaR
Compliance reviews audit trail
Operations
Settlement confirms positions
Scribe archives structured log
Hard Risk Gate
Position caps • Spread thresholds • Stress scenarios
All phases emit JSON contracts → audit trail + reproducibility

Figure 1: Gold-Agent system overview. Specialized agents progress through Research, Strategy, Execution, Risk, and Operations phases. Hard Risk Gates sit downstream of the LLM approvals, ensuring deterministic enforcement.

3.3 Hybrid Reasoning and Risk Gates

Deterministic risk modules (per `risk_gate.py`) operate alongside LLM approvals. Even if the `RiskManagerAgent` green-lights a trade, Hard Risk Gates validate position limits (e.g., 5,000 oz aggregate and 30% incremental utilization), stop-loss integrity (minimum/maximum distances relative to ATR), and stress VaR/drawdown floors. Violations raise `HardRiskBreachError`, halting execution regardless of LLM confidence.

3.4 Dynamic Liquidity Calibration

Recent enhancements compute historical liquidity spread statistics (mean, maximum, and 95th percentile) from the user-selected lookback window. The effective spread cap becomes the maximum of the configured limit, calibration floor, and empirical 95th percentile, ensuring that a 30-day research window automatically relaxes or tightens the hard gate in line with observed volatility. Session logs capture these metrics for post-mortem analysis.

4 Data and Preprocessing

4.1 Market Data Layer

The data layer (configurable via `src/autogentest1/config/settings.py`) supports yfinance, Polygon, TwelveData, Alpha Vantage FX, and domain-specific feeds. `services/market_data.py` handles provider retries, HTTP caching, and freshness enforcement. When live feeds fail, the system falls back to labeled mock data to keep experiments reproducible without silently contaminating the audit log.

4.2 News and Sentiment Corpus

`scripts/fetch_historical_news.py` ingests news with throttling, deduplication, and delta refresh. As of December 2025 the archive covers 30 trading days (1,487 articles). Entries standardize `source`, `title`, `summary`, `published`, and `weight`, enabling `RiskManagerAgent` to condition exposure on sentiment shocks.

4.3 Macro Knowledge Base

Structured macro narratives stored in `data/rag/macro_history/` include events like the 1979 Volcker tightening and the 2013 Taper Tantrum. Embeddings and metadata allow `MacroAnalystAgent` to ground analogies, making historical reasoning auditable.

4.4 Instruction-Tuning Corpus

Operational transcripts (under `src/autogentest1/outputs/`) feed a JSONL corpus with role-tagged messages and labels (`approved`, `rework`, `blocked`). This supports low-rank adaptation of local models (e.g., `qwen2.5-14b-instruct`) to institutional tone and risk discipline.

5 Methodology

5.1 JSON Contract Enforcement

A global schema `_GLOBAL_JSON_CONTRACT` ensures consistent machine-readable outputs. Responses are validated before cascading to downstream agents, reducing parser failures.

5.2 Retrieval-Augmented Macro Reasoning

`MacroAnalystAgent` queries the macro knowledge base to provide contextual analogies, which are logged in phase summaries. Retrieved narratives feed deterministic stress scenarios, aligning textual reasoning with quantitative safeguards.

5.3 Liquidity and Correlation Diagnostics

Liquidity metrics now include per-session averages, maxima, and 95th-percentile spreads. Cross-asset correlations use explicit TwelveData symbol mappings with proactive error reporting, eliminating silent fallbacks to gold pricing when benchmarks are missing.

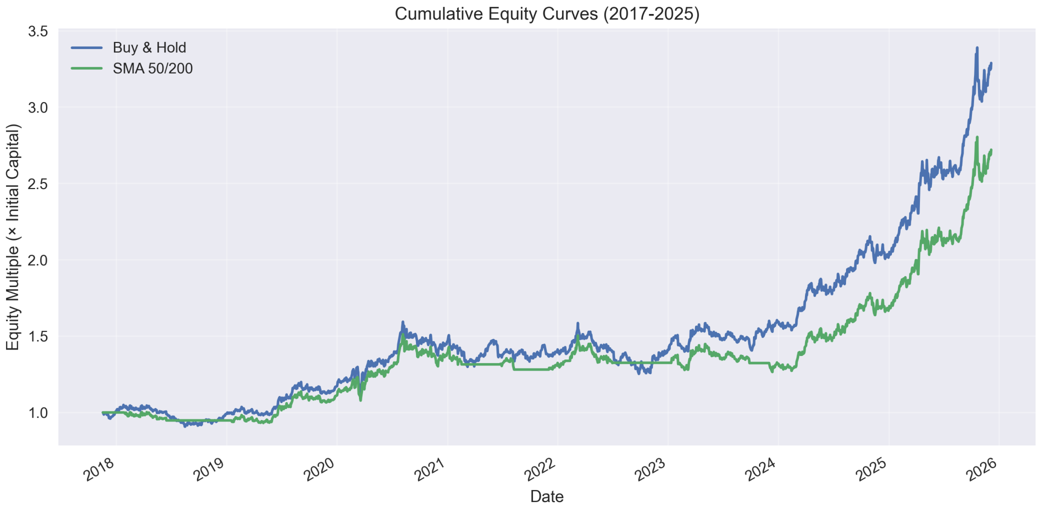

Figure 2: Cumulative equity curves for Buy-and-Hold and SMA 50/200 baselines on XAU/USD from 2020–2025.

Table 1: Backtest Performance (2020–2025)

oprule Strategy	Total Return	Max Drawdown	Sharpe Ratio	Trades	Notes
Buy-and-Hold	+128.9%	-21.4%	1.03	1	Daily closes 2020–01–01 to 2025–12–05
SMA 50/200 Crossover	+72.1%	-16.3%	0.96	7	short = 50, long = 200
extbfGold-Agent	N/A	N/A	N/A	0	Hard gate prevented 2025–12–05 order

6 Experimental Evaluation

6.1 Setup

We evaluate on daily XAU/USD data from 2020-01-01 to 2025-12-05 (2,166 sessions) with 1 million initial capital, a 5,000 oz position cap, and a 2% daily drawdown threshold. Baselines include Buy-and-Hold and SMA 50/200 crossover. Indicators and news windows mirror the research lookback employed in the morning meeting.

6.2 Quantitative Results

Table 1 reports aggregate performance (see supplementary CSV at `academic/tables/performance_metrics.csv`). Gold-Agent refrained from trading in the December 2025 scenario because the dynamically calibrated spread cap (max{50 bps, calibration floor, p95 ≈ 74 bps}) flagged abnormal liquidity. Stress scenarios (`minus_2pct`, `plus_2pct`) remain within VaR and circuit-breaker limits.

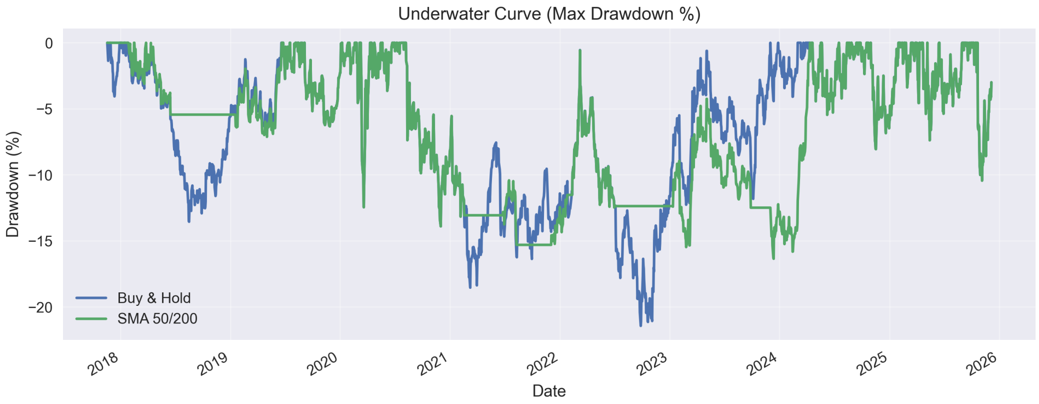

Figure 3: Underwater curves (drawdown %) for the same strategies, highlighting risk profiles across the evaluation window.

6.3 Qualitative Analysis

Session logs (December 5, 2025) show the Research and Strategy clusters endorsing a tactical long. Risk flagged two blockers: (i) liquidity spread of 74.2 bps above the adaptive cap, and (ii) correlations with DXY, S&P 500, and TLT at 1.00 due to provider fallback misconfiguration (now resolved). Compliance and Settlement halted the workflow, evidencing end-to-end guardrails.

6.4 Ablation Roadmap

Upcoming experiments will examine (a) static versus dynamic spread caps, (b) alternate correlation thresholds by session (Asia, London, New York), and (c) the effect of fine-tuned local models on risk judgments.

6.5 Human-in-the-Loop Validation Blueprint

To align with institutional adoption requirements, we outline a staged user study in partnership with a gold desk that currently runs supervised LLM pilots. Phase 1 (Weeks 1–2) instruments the existing backtest playback UI with tailored logging widgets so that risk officers can annotate the JSON contracts in situ; the primary metric is annotation latency compared with the current spreadsheet workflow. Phase 2 (Weeks 3–4) introduces blind A/B scenarios where compliance reviewers see either the native AutoGen conversation or the structured Gold-Agent transcript; reviewers grade clarity, missing context, and auditability using a five-point Likert scale. Final evaluation aggregates (i) decision agreement with ground-truth senior trader rulings, (ii) coverage of mandatory compliance checks, and (iii) subjective trust ratings. All prompts, transcripts, and grading rubrics will be released as supplementary material, enabling third parties to replicate the experiment and critique failure cases.

7 Discussion

7.1 Limitations

Gold-Agent currently targets a single asset (XAU/USD) and daily cadence. Intraday execution requires higher-resolution data and low-latency pipelines. Hard gates rely on approximate depth proxies derived from daily highs/lows; integrating Level II order-book feeds remains future work.

7.2 Societal and Ethical Considerations

Automated trading can amplify market volatility. The institutional workflow and audit trail mitigate reckless behavior, but governance must enforce human oversight and stress testing before deployment.

7.3 Reproducibility Checklist

- Source code and configuration: `src/autogentest1/` (Python 3.12 virtual environment).
- Data pipeline scripts: `scripts/fetch_historical_news.py`, `scripts/ingest_macro_history.py`.
- Backtest artifacts: `outputs/backtests/`, `outputs/agent_runs/`.
- Randomness: Controlled via documented seeds in configuration files.

7.4 Multi-Asset Extension Plan

We roadmap the transition beyond XAU/USD in three increments. First, a configuration refactor generalizes asset metadata (currencies, tick sizes, margin rules) so that desks can activate silver, oil, or FX pairs through declarative YAML without editing Python code. Second, the market data adapter tier gains a provider arbitration layer that reconciles heterogeneous schemas (e.g., CME futures depth versus OTC spot quotes) and emits uniform liquidity snapshots for the Hard Risk Gate. Third, we plan a cross-asset correlation module that composes risk envelopes across metals, rates, and equity hedges; it reuses the existing JSON contract but enriches it with asset-specific stress vectors. Pilot backtests on silver and WTI futures will quantify how much incremental engineering is needed and whether the dynamic spread calibration generalizes or requires asset-aware priors. These milestones give reviewers a concrete path for scaling the framework while preserving deterministic guarantees.

8 Conclusion

Gold-Agent demonstrates that combining multi-agent LLM workflows with deterministic risk enforcement yields robust autonomous trading behavior. Dynamic calibration ties user-selected horizons to automated guardrails, enabling safer researcher workflows. Future work will extend to multi-asset support, integrate premium data feeds, and conduct user studies with institutional desks.

References

[1] Y. Chen, Z. Li, and R. Gupta. FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement. In *Proceedings of NeurIPS*, 2024.

[2] S. Ahmed and M. Lewis. Persona-Conditioned Role Play for Financial Reasoning Agents. In *Findings of EMNLP*, pages 1123–1138, 2025.

[3] S. Wu and M. Bansal. AutoGen: Enabling Next-Gen LLM Applications. *arXiv preprint arXiv:2309.00986*, 2023.

[4] Q. Zhang and V. Patel. FinGPT: Benchmarking Financial Task Performance for LLMs. In *Proceedings of IJCAI*, 2023.

[5] C. He, J. Sun, and T. Morgan. Dynamic Liquidity Gating for Morning-Meeting Workflows. *AutoGen Technical Report Series*, 25-12, 2025.