On the Role of Tools in Agentic LLM-Based Trading Systems

Notebook for the FinMMeval Lab at CLEF 2026

Maximilian Heil[1,*,†], Aleksandar Pramov[1] and Boyang Gong[1]

[1]Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

Abstract

Large language model (LLM)-based agentic systems are increasingly applied to financial decision-making tasks, where they must process heterogeneous information sources and act under uncertainty. While prior work has demonstrated the viability of multi-agent financial architectures, it remains unclear which tools and reasoning components materially improve performance. In particular, little empirical evidence exists on whether increasing tool sophistication and information-processing quality leads to better downstream investment decisions. In this paper, we present DS@GT StockTron, a multi-agent trading system developed for FinMMEval 2026 Task 3. The system combines specialized analyst agents for technical analysis, sentiment analysis, section-aware news processing, and machine learning-based return forecasting, coordinated by an LLM-based portfolio manager. We conduct a controlled ablation study on TSLA and BTC over the FinMMEval development period from 2024-08-02 to 2026-01-10 in order to isolate the contribution of individual tools and architectural components. Our results show that tool quality materially affects agentic trading performance. More sophisticated and context-aware sentiment-processing pipelines consistently outperform naive sentiment aggregation, while augmenting the system with a predictive machine learning forecasting tool further improves performance, particularly for BTC. In contrast, introducing short-term reflective memory over past decisions consistently deteriorates results, suggesting that naive memory mechanisms may amplify recency bias and market noise rather than improve decision quality. The strongest-performing configuration achieves a cumulative return of +20.1% and Sharpe ratio of +0.54 on BTC during the ablation studies. Live competition results provide a more mixed out-of-sample picture: the final system remains competitive for TSLA relative to buy-and-hold, but performs poorly for BTC. Overall, the findings suggest that carefully designed tool augmentation and higher-quality information processing constitute important levers for improving agentic financial systems, potentially more so than prompt engineering or architectural complexity alone.

Keywords

LLM agents, Financial Trading, Sentiment Analysis, Machine Learning,

1. Introduction

Autonomous LLM-based agents are increasingly applied to financial decision-making, where they must interpret heterogeneous information, such as price history, news, and company filings, and act under time constraints. FinMMEval 2026 Task 3 [1] operationalizes this setting as a daily trading competition: given yesterday's market context, an agent must emit a BUY, HOLD, or SELL decision within three minutes.

Researchers have already proposed a variety of agentic systems for financial decision-making, such as InvestorBench [2], TradingAgents [3], and DeepFund [4]. These frameworks transfer core agentic AI building blocks, including agent specialization, memory, tool usage, and multi-agent orchestration, to financial investment settings. Typically, such systems equip agents with elementary financial tools, for example, the computation of technical indicators from historical price series or sentiment analysis of company-related news. In addition, memory and reflection mechanisms allow agents to incorporate past decisions and portfolio states into subsequent actions. Depending on their designated role, LLM-

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

† Main contributor.

mheil7@gatech.edu (M. Heil); apramov3@gatech.edu (A. Pramov); bgong36@gatech.edu (B. Gong)

0009-0002-6459-6459 (M. Heil); 0009-0005-9049-1337 (A. Pramov); 0009-0004-9915-9050 (B. Gong)

based agents may act as market analysts generating BUY, HOLD, or SELL recommendations, stylized investors imitating the philosophies of influential market participants, risk managers evaluating portfolio constraints, or portfolio managers making the final investment decision.

While prior work demonstrates the viability of these architectures, the specific set of tools and reasoning mechanisms available to the agents is likely to materially influence their financial performance. This motivates the question of which additional components may improve the effectiveness of agentic trading systems. First, traditional systematic investment strategies frequently rely on predictive time-series models, ranging from simple momentum approaches to more sophisticated machine learning and ensemble methods ([5], [6]). Yet, despite their importance in quantitative finance, it remains unclear whether providing such predictive models as explicit tools to LLM-based agents improves downstream investment decisions. We investigate this question empirically.

Second, both qualitative and quantitative investors often incorporate asset-specific sentiment information, such as company news, earnings commentary, or macroeconomic developments, into their decision-making process. Since LLM-based agents are particularly well suited for processing and reasoning over textual information, enhanced sentiment-analysis techniques represent a promising avenue for improving financial performance. In particular, we investigate whether more targeted and context-aware sentiment incorporation can improve agentic investment decisions.

Third, reflection constitutes a common design pattern in agentic AI systems, enabling agents to evaluate the quality of their previous decisions and adapt future behavior accordingly. We therefore examine whether incorporating reflective reasoning over past portfolio actions leads to measurable performance improvements.

Putting these considerations together, we address the following three research questions:

RQ1: *Does incorporating a predictive time-series machine-learning model as an agent tool improve investment performance relative to an otherwise identical system without it?*
We hypothesize that providing the portfolio manager with a structured quantitative forecast from a machine-learning model improves investment performance.

RQ2: *Does sophisticated sentiment analysis improve performance over naive sentiment computation?*
We hypothesize that sentiment signals derived from selectively filtered, forward-looking, and asset-relevant news provide more informative inputs than sentiment measures obtained by naively aggregating all available text. In particular, incorporating not only the sentiment level itself, but also sentiment dynamics such as changes and temporal trends, may allow the agent to better capture shifts in market expectations and investor positioning. We therefore investigate whether a more targeted and context-aware sentiment pipeline can reduce informational noise and improve the portfolio manager's directional accuracy.

RQ3: *Does reflecting on the performance of past decisions improve investment performance relative to an otherwise identical system without such reflection?*
We hypothesize that equipping the portfolio manager with a memory of recent hit rates of its own decisions will improve performance by allowing it to account for recent signal quality and avoid repeating systematic errors.

To answer these questions, we build DS@GT StockTron, a multi-agent trading workflow. The system is based on DeepFund and includes various analysts that analyze newly arrived information and propose a BUY, HOLD, or SELL recommendation. In its simplest form, the analysts use naïve technical indicators or a trivial sentiment analysis to form recommendations, which are then aggregated by a portfolio manager into a final investment decision. All recommendations and decisions are stored in an SQLite database which can also be accessed as decision memory and supports the advanced tools evaluated in this study.

We evaluate the system through ablation studies on the FinMMEval development dataset, covering TSLA and BTC from 2024-08-02 to 2026-01-10. We do not draw a separate train/test split from this development period for the agentic system itself; instead, the period is used as a chronological

walk-forward backtest to compare system variants under the same period. The live competition period from 2026-05-11 to 2026-07-04 serves as the true out-of-sample test, for which we additionally report leaderboard results. Separately, the predictive machine-learning tool is trained and validated on an independent S&P 500 dataset.

The remainder of this paper is organized as follows: Section 2 reviews related work on LLM-based financial agents, tool augmentation, sentiment analysis, and memory; Section 3 describes the FinMMEval development dataset and the auxiliary S&P 500 data used for the machine-learning tool; Section 4 presents the DS@GT StockTron architecture and its analyst, portfolio-manager, and memory components; Section 5 reports the ablation results; Section 6 discusses the main findings and limitations; and Sections 7 and 8 outline future work and conclude.

2. Related Work

The landscape of LLM-based financial agents has grown rapidly. InvestorBench [2] provides a standardized benchmark spanning stocks, cryptocurrencies, and ETFs, and demonstrates that architectural design choices and memory mechanisms can materially affect performance. TradingAgents [3] proposes a multi-agent architecture designed to mirror a real trading desk, separating the roles of analysts, researchers, traders, and risk managers. HedgeFundAgent [7] demonstrates an open-source implementation combining fundamental, sentiment, and technical signals. DeepFundAgent [4] further emphasizes the importance of live benchmarking and leakage-free evaluation, motivating our strictly out-of-sample validation protocol for the machine-learning tool.

A central design question in these systems is how agents should use tools. In many existing financial agents, tools are relatively simple: a Python function computing RSI or MACD, a retrieval step over news, or a sentiment classifier applied to raw text. Tool use and augmentation have received broader attention in the LLM literature [8], but the isolated effect of tool augmentation in financial agents remains under-studied. This is particularly relevant for systematic investment settings, where quantitative tools such as return forecasts, technical indicators, or risk models have long been central to investment decisions. In equities, Gu, Kelly, and Xiu [9] show that machine-learning models can extract statistically meaningful return predictability from large cross-sections of firm-level data. Similarly, Krauss et al. [6] find that tree-based and neural network models can achieve modest but economically relevant directional predictability for S&P 500 stocks. These findings motivate our use of a random forest model as a structured forecasting tool available to the agent, rather than as a standalone trading strategy.

Sentiment analysis is another important information source for financial decision-making. Tetlock [10] shows that negative media sentiment contains information about stock market movements. However, applying sentiment analysis in an agentic trading system raises an additional challenge: not all available text is equally informative for investment decisions. Raw news summaries, filings, and market commentary may contain backward-looking descriptions, repeated information, generic market context, or conflicting bullish and bearish statements. Naively aggregating all available text can therefore dilute the signal and increase noise. Prior work on financial text analysis suggests that the source, context, and economic relevance of textual information matter substantially [10, 11]. This motivates our enhanced sentiment pipeline, which filters for asset-relevant and forward-looking information, and additionally considers sentiment change and short-term sentiment trends rather than only the current sentiment level.

Finally, reflection and memory are common design patterns in agentic AI systems. Existing financial agent frameworks often allow the agent to condition on past actions or portfolio states, with the intuition that this may help correct repeated mistakes. InvestorBench [2], for example, highlights the importance of memory design in financial agents. At the same time, financial markets are noisy and non-stationary, so short-term memory may also amplify recent random outcomes rather than improve decision quality. This motivates our empirical evaluation of whether a simple hit-rate-based reflection mechanism improves performance relative to an otherwise identical system without reflection.

3. Data

The competition dataset includes two assets: Tesla (TSLA) and Bitcoin (BTC). A true out-of-sample evaluation of the system is performed via a live test from 2026-05-11 to 2026-07-04. DS@GT StockTron receives structured data (closing price history) and unstructured data (a news summary and, where applicable, 10-K and 10-Q filings) each day via HTTPS requests from the organizers. The news summary may include an extensive LLM-based summarization of a) comprehensive summary of events and themes, b) key themes and developments, c) overall market sentiment based on these articles, d) trading activity and price action, e) trading activity themes, f) sentiment signals for BTC or TSLA. The data used for system development and the ablation studies are sourced from the FinMMEval development dataset [12], spanning a period from 2024-08-02 to 2026-01-10.

Table 1 shows the daily return and word count distributions for TSLA and BTC in the FinMMEval development dataset. The return distributions are approximately centered on zero: the median return is 0.10 % for TSLA and 0.03 % for BTC, yet the tails are markedly asymmetric: TSLA spans $[-15.4\,\%, +22.7\,\%]$ and BTC $[-8.6\,\%, +12.1\,\%]$, reflecting the occasional large events that tend to dominate cumulative performance over this window. TSLA is also considerably more volatile (daily std. 4.10 % vs. 2.35 % for BTC), and its interquartile range $(-1.98\,\%, +2.82\,\%)$ is substantially wider than BTC's $(-1.18\,\%, +1.23\,\%)$, indicating meaningfully larger day-to-day dispersion even outside extreme tail events. This is consequential for market timing these assets: on the majority of days BTC moves less than approximately $\pm 1.2\,\%$, while TSLA typically remains within roughly $\pm 2.8\,\%$, so performance will still be disproportionately driven by a relatively small number of large-move days regardless of average signal quality. Moreover, TSLA's word-count distribution is on average smaller (TSLA: mean 995 words, BTC: mean 1190 words) and more variable (TSLA: std. 293 words, BTC: std. 165 words) than BTC's. This is driven by weekends or public holidays with less market news about the stock. A qualitative inspection of the daily news summaries also reveals that they exhibit low actionable information content. They consist of pervasive hedging language that potentially produces near-neutral sentiment regardless of actual market conditions. Early TSLA entries in particular are near-empty, while BTC summaries, though longer and more consistent, systematically present both bullish and bearish angles, further dampening any directional signal available to a sentiment model. Finally, the 528-day window covers a predominantly bullish period for both assets, so a naive long-biased strategy would have performed well regardless of signal quality; with only 527 BTC return observations and 361 TSLA trading-day observations, combined with daily volatility of 2–4 %, the statistical power to detect a genuine edge is low, and any reported Sharpe ratio or drawdown figure should be treated as indicative rather than conclusive.

Table 1
Summary statistics for daily returns and news word counts of the FinMMEval development dataset.

Variable	Asset	N	Mean	Std	Min	Q25	Median	Q75	Max
Daily Return (%)	TSLA	361	0.28	4.10	-15.43	-1.98	0.10	2.82	22.69
	BTC	527	0.09	2.35	-8.56	-1.18	0.03	1.23	12.05
Word Count	TSLA	528	994.83	292.86	297.00	822.75	1016.50	1201.25	1815.00
	BTC	528	1189.56	165.21	3.00	1085.00	1183.50	1292.50	1799.00

Note: TSLA returns cover 361 genuine trading days (weekends and U.S. market holidays excluded). BTC returns (prices) cover 527 (528) calendar days reflecting its 24/7 trading schedule. Word counts cover all 528 calendar days for both assets since news summaries are generated daily including weekends. Returns are percentage price changes; word counts proxy for LLM input token length.

Although BTC trades continuously, all subsequent backtests use a common U.S. equity trading-day calendar for both assets to keep the decision frequency comparable across TSLA and BTC.

Our agentic system is complemented with a predictive time-series machine-learning model. This model is trained on an independent dataset of all S&P 500 constituents from 2001-01-01 to the beginning

of the FinMMEval development dataset (2024-07-31), downloaded via `yfinance`. For the live competition system, we use data available up to 2026-05-08, i.e., the last trading day before the live evaluation starts.[1] This yields approximately 2.8 million stock-day observations across 500 stocks. During the competition, our agentic system autonomously downloads new daily observations for online learning.

4. Methodology

4.1. System Architecture

DS@GT StockTron follows a two-stage pipeline in the spirit of DeepFund (Figure 1):

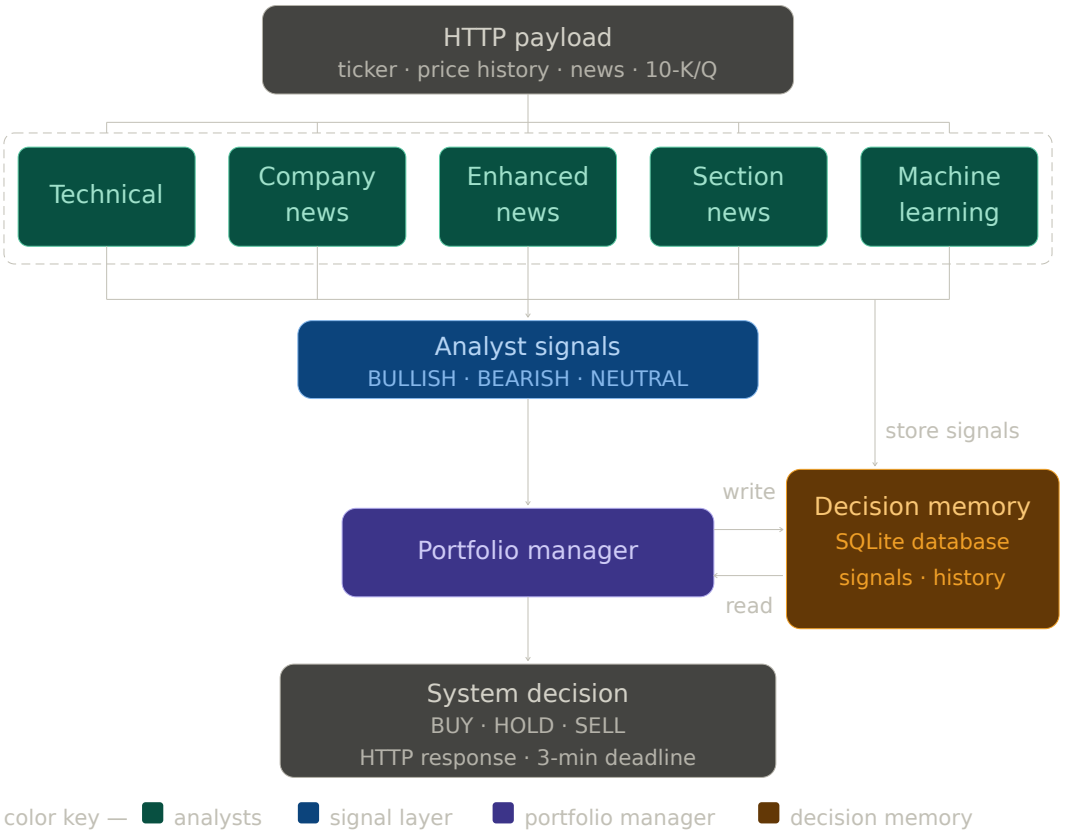

Figure 1: DS@GT StockTron system architecture

1. **Data request via HTTP API.**
2. **Parallel analyst agents.** Up to five LLM-backed specialist agents run concurrently and each emits a BULLISH / NEUTRAL / BEARISH signal with a confidence score and a short rationale. They retrieve the necessary price data for the analysts themselves and use a single specified tool. The agents are described in detail in Section 4.2.
3. **Portfolio manager.** An LLM-backed agent that receives all analyst signals and produces a final BUY / HOLD / SELL decision with a rationale taking into account the memory.
4. **Decision memory.** A persistent SQLite-based memory layer records the system's daily information state and decisions. After each trading day, the database stores the analyst signals, confidence scores, rationales, final portfolio-manager action, and portfolio state. If configured, the portfolio

[1] We acknowledge that the index constituents in the predictive time-series machine-learning model are not point-in-time, because such data are very costly to obtain. Since the predictive model is one of several tools to which the agent has access, we believe that this limitation does not substantially impair the research system.

manager queries the most recent k days of decisions and their realized outcomes to compute a rolling hit rate before reconciling the current day's analyst signal.

5. **Data response via HTTP API.**

The workflow is implemented in LangGraph. For all development-period ablation experiments reported in Section 5, we use `ChatGPT-4.1-nano` as the LLM backbone for all LLM-based analyst agents and the portfolio manager. For the final live competition system, we keep the same architecture but upgrade the LLM backbone to `ChatGPT-5.5` to improve higher-level reasoning and multi-signal reconciliation. The ablation study therefore identifies the strongest architecture under a fixed lightweight backbone, rather than proving that the same component ranking necessarily holds across larger LLM backbones. This design choice is consistent with prior evidence that architectural choices and tool design can materially affect the performance of financial agents [12].

4.2. Analyst Agents

Technical agent This analyst agent uses Python functions to compute specialized technical indicators: exponential weighted moving averages (trend), mean reversion including Bollinger Bands, relative strength index, volatility and support and resistance price levels. All but the last indicator are rule-based and transformed to qualitative signals: BULLISH / NEUTRAL / BEARISH. Subsequently, the LLM backbone is prompted with the results of the technical analysis and infers a directional signal (again: BULLISH / NEUTRAL / BEARISH). This agent and the company news agent are from DeepFund and represent the baseline tools used in our ablation study.

Company news agent. Passes the full news summary of TSLA or BTC to an LLM prompt and asks for an aggregate sentiment direction. Returns BULLISH / NEUTRAL / BEARISH.

Enhanced news agent. Implements a three-stage filtering pipeline before sentiment aggregation: (1) *splitting*: splits the news summaries given their headlines into subsets (a) comprehensive summary of events and themes, b) key themes and developments, c) overall market sentiment based on these articles, d) trading activity and price action, e) trading activity themes, f) sentiment signals; (2) *relevance*: scans each single subset of the news summary and discards items unrelated to the company's or coin's core business; (3) *sentiment analysis*: computes the sentiment on each filtered subset with a forward-looking emphasis, yielding a more targeted signal. The sentiment subsets are finally aggregated to a single TSLA or BTC news sentiment. In addition, the agent also computes the sentiment change relative to the prior day and the overall sentiment trend over the last 10 trading days using the signal memory. In summary, this single agent returns three BULLISH / NEUTRAL / BEARISH signals for the aggregated sentiment, sentiment change, and sentiment trend.

Section news agent. This agent complements the Enhanced news agent by providing a more structured view of the news flow, but without explicitly modeling sentiment change and trend over time. The agent first ingests the daily news summary and, importantly, available 10-K or 10-Q filings. It then classifies information into several financially relevant sections such as company fundamentals, product demand, macroeconomic developments, regulation, competition, and filings using a lightweight hybrid classification pipeline. Each section is scored independently by the LLM as BULLISH, NEUTRAL, or BEARISH together with a confidence estimate. Finally, the section-level signals are aggregated into a single directional `AnalystSignal` representing the overall news environment for the trading day.

Machine-learning agent. The machine-learning agent wraps a random forest classifier as a callable forecasting tool. The target variable is next-day direction (up/down). Features are entirely price- and volume-derived, computed over the 1-day, 5-day, and 20-day horizons. Hyperparameters have been derived via a validation split that covers the last 6 years before the FinMMEval development dataset with an embargo of 1 month. As a result, the best parameters are number of trees: 500, maximum depth:

5, minimum samples N required to split an internal node: $N = 100$, number of features F considered at each split: $\sqrt{F}$). The agent updates the model daily using an online learning protocol that incorporates new observations without revisiting the full training set. There is just a single model for both TSLA and BTC due to the lack of cross-sectional observations for exchange-traded coins.

4.3. Portfolio Manager and Decision Memory

The portfolio manager is an LLM-backed orchestration agent that receives the full set of BULLISH/NEUTRAL/BEARISH analyst signals together with their confidence scores and rationales, and produces a single BUY / HOLD / SELL action with a natural-language justification. Its central capability is therefore to aggregate the information from the analysts and take a final investment decision. Furthermore, when reflection is enabled, the portfolio manager accesses the decision memory and retrieves the recent decision history of k days, including the signals it previously received, the actions it took, and their outcomes - and is explicitly prompted to reconcile conflicting analyst views against its recent track record. This enables a self-improvement loop in which the agent can, for instance, discount a signal type it has repeatedly acted on unsuccessfully. We test in ablation studies whether such reflection can be helpful for market-timing assets.

Finally, the decision memory stores all intermediate and final outputs (signals, rationales, actions, and full prompt contexts), providing both the runtime memory that drives reflection and a complete audit trail for post hoc analysis and the ablation studies in Section 5.

5. Results

5.1. Ablation Study

We backtest the proposed system and evaluate our research questions in a market timing setting, in which the system emits a pure signal; cumulative return is computed by following each BUY/HOLD/SELL decision without position sizing. All market-timing experiments were backtested on the FinMMEval development dataset from 2024-08-02 to 2026-01-10 using the common trading-day calendar described in Section 3. Table 2 reports the results of a series of ablation studies designed to isolate the contribution of individual architectural components to the overall performance of the agentic trading system. We evaluate each configuration using cumulative return (CR), Sharpe ratio (SR), and maximum drawdown (MD).

The buy-and-hold benchmarks over this period are +114.3% for TSLA and +47.1% for BTC. Such exceptionally strong performance over the backtesting period highlights the challenging nature of the market-timing task. In strongly upward-trending markets frequent trading decisions can easily underperform passive exposure due to mistimed entries and exits. Consequently, the objective of the agentic system is not merely to generate positive returns, but to produce risk-adjusted performance improvements and reduce downside risk relative to simpler active approaches under realistic trading conditions. Importantly, none of the evaluated agentic configurations should be interpreted as outperforming a passive buy-and-hold strategy over this development period. We therefore interpret the ablation results primarily as evidence about the relative contribution of individual system components, rather than as evidence of absolute superiority over passive exposure.

The ablation framework starts from a baseline configuration and incrementally modifies individual components while holding the remaining system architecture fixed. This allows us to assess the marginal contribution of specific design choices, including enhanced sentiment processing, section-aware news analysis, predictive machine-learning signals, additional specialized agents, and memory-based reflection mechanisms. In this setting, the baseline serves as the reference architecture against which all subsequent experiments are compared. The baseline configuration consists of a portfolio manager (no reflection) agent supported by two core information sources: (i) a technical agent that analyzes historical price dynamics using standard technical indicators, and (ii) a news agent that performs naive sentiment aggregation over all available textual information. The subsequent experiments

then evaluate targeted modifications to this baseline. Experiments 1 and 2 investigate increasingly sophisticated forms of news and sentiment processing replacing the naive sentiment aggregation. Experiments 3 and 4 analyze the effect of replacing the technical-analysis component with a predictive machine-learning model and its interaction with enhanced sentiment signals. Experiment 5 extends the architecture to the full five-agent configuration. Lastly, Experiments 6 and 7 evaluate whether introducing memory over previous decisions and portfolio states improves performance. Here, we enable the reflection mechanism of the portfolio manager while keeping the baseline agents.

Table 2
Ablation study results for the FinMMEval development dataset from 2024-08-02 to 2026-01-10 using the common trading-day calendar. CR = Cumulative Return, SR = Sharpe Ratio, MD = Max Drawdown.

Experiment	TSLA			BTC		
	CR	SR	MD	CR	SR	MD
Buy-and-Hold	+114.3%	—	—	+47.1%	—	—
Baseline (`technical+news`)	-54.5%	-0.76	-65.8%	-18.0%	-0.29	-32.5%
Exp. 1: Enhanced news	-23.5%	$\mathbf{-0.13}$	-53.7%	-18.5%	-0.30	-40.6%
Exp. 2: Section news	-41.6%	-0.47	-52.3%	$+2.7\%$	$+0.21$	-26.8%
Exp. 3: ML model	-23.8%	-0.45	-41.5%	$+0.8\%$	$+0.15$	-26.7%
Exp. 4: Enhanced news + ML	-31.9%	-0.37	-74.2%	$+11.2\%$	$+0.40$	-18.4%
Exp. 5: All five agents	-39.5%	-0.42	-61.5%	$\mathbf{+20.1\%}$	$\mathbf{+0.54}$	-25.7%
Exp. 6: Memory 5d	-61.8%	-1.14	-88.2%	-15.9%	-0.27	-32.5%
Exp. 7: Memory 10d	-50.5%	-0.85	-63.2%	-7.0%	-0.04	-25.6%

Based on the results, we can establish the following as answers to the postulated research questions:

Baseline: The baseline configuration performs poorly across both assets, yielding negative cumulative returns and Sharpe ratios for both TSLA and BTC. Relative to the strong buy-and-hold benchmark, this suggests that a simple combination of technical indicators and naive sentiment aggregation is insufficient to consistently capture profitable market-timing signals in highly volatile and trend-driven markets. At the same time, the baseline establishes an important reference point for the ablation studies, as it demonstrates that merely introducing an agentic architecture does not automatically lead to strong financial performance. Instead, the quality of the underlying tools, information processing, and reasoning mechanisms appears to be critical for achieving meaningful improvements.

RQ1 (Machine-learning model as tool): In Experiment 3, we replace the technical agent with the random forest machine-learning model, leaving all other baseline components unchanged. This consistently improves BTC performance relative to the baseline (SR: from -0.29 to $+0.15$). When paired with the enhanced news agent (Exp. 4), BTC performance improves further (SR: from -0.30 to $+0.40$, CR: -18.0% to $+11.2\%$). For TSLA, results are more mixed: the predictive time-series machine-learning model signal improves performance alongside naive news ($\Delta\text{SR} = +0.31$) but slightly worsens performance when combined with enhanced news ($\Delta\text{SR} = -0.24$). Overall, the usefulness of the predictive time-series machine-learning tool appears to be asset-dependent and interacts with the quality of the accompanying sentiment signal. The BTC results should be interpreted cautiously, because the forecasting model is trained on S&P 500 equities and therefore operates under substantial cross-market distribution shift when applied to a cryptocurrency.

To assess the standalone predictive quality of the random forest tool, we evaluate it on S&P 500 stocks over the held-out period from 2024-08-02 to 2026-01-10, after training on data available through 2024-07-31. Predictions are generated in chronological order before each new observation is incorporated through the online update protocol. Table 3 reports a directional accuracy of 51.5%, which is consistent with the financial machine-learning literature: Gu et al. [9] report similar figures for tree-based models

on U.S. equities, and Krauss et al. [6] find 52–53% next-day accuracy for random forests on S&P 500 stocks. The high recall (75.2%) indicates that the model skews toward calling BULLISH, a bias that proves consequential in the backtest (Section 6).

Table 3
Random Forest out-of-sample test metrics on S&P 500 stocks over the held-out period from 2024-08-02 to 2026-01-10.

Metric	Value
Accuracy	51.5%
Precision	52.3%
Recall	75.2%
F1-Score	61.7%
AUC-ROC	50.8%

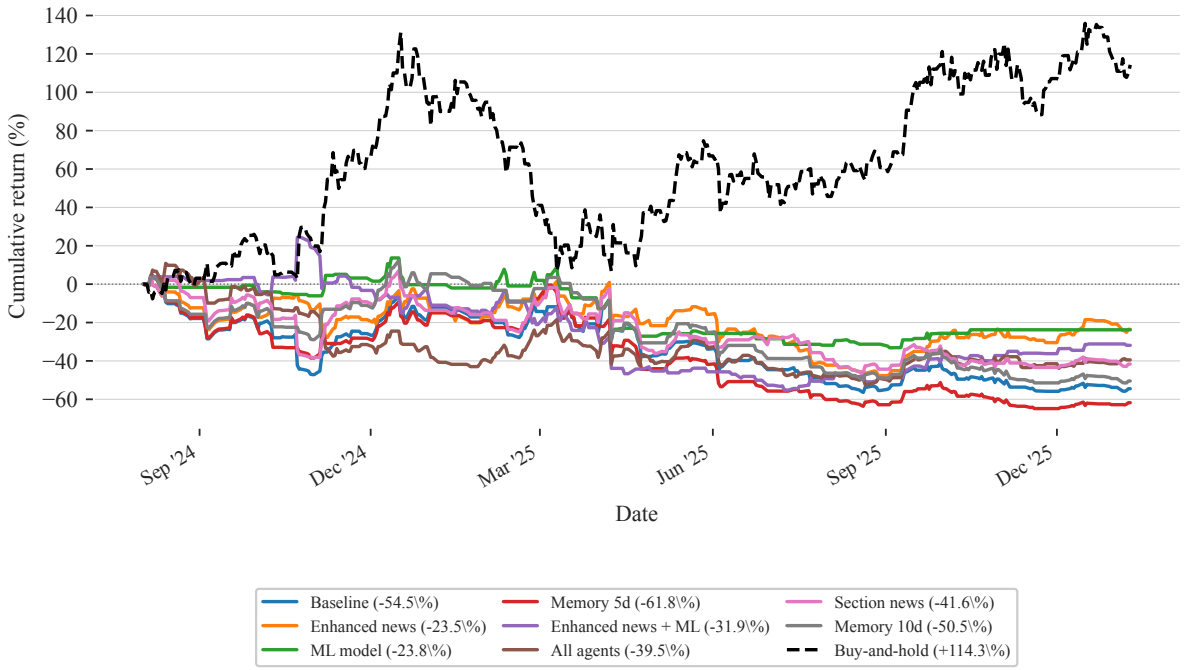

Figure 2: Cumulative Return of Tesla

RQ2 (Enhanced sentiment analysis): Both enhanced sentiment-analysis variants outperform the naive baseline, suggesting that the quality of textual information processing materially affects agent performance. Experiment 1 (`Enhanced news`) improves TSLA performance substantially (SR: from -0.76 to -0.13) while also reducing drawdowns, although the effect on BTC remains limited. In contrast, Experiment 2 (`Section news`) produces the strongest improvement for BTC, turning cumulative returns positive (from -18.0% to $+2.7\%$) and increasing the Sharpe ratio from -0.29 to $+0.21$, while also moderately improving TSLA.

Overall, the results indicate that more targeted and context-aware sentiment processing provides more informative signals than naive sentiment aggregation. Importantly, neither enhanced variant simultaneously degrades both assets, supporting the hypothesis that sophisticated sentiment analysis is additive within agentic trading systems.

RQ3 (Reflection): Experiments 6 and 7 investigate whether augmenting the portfolio manager with a short-term memory of past decision performance improves investment outcomes relative to the

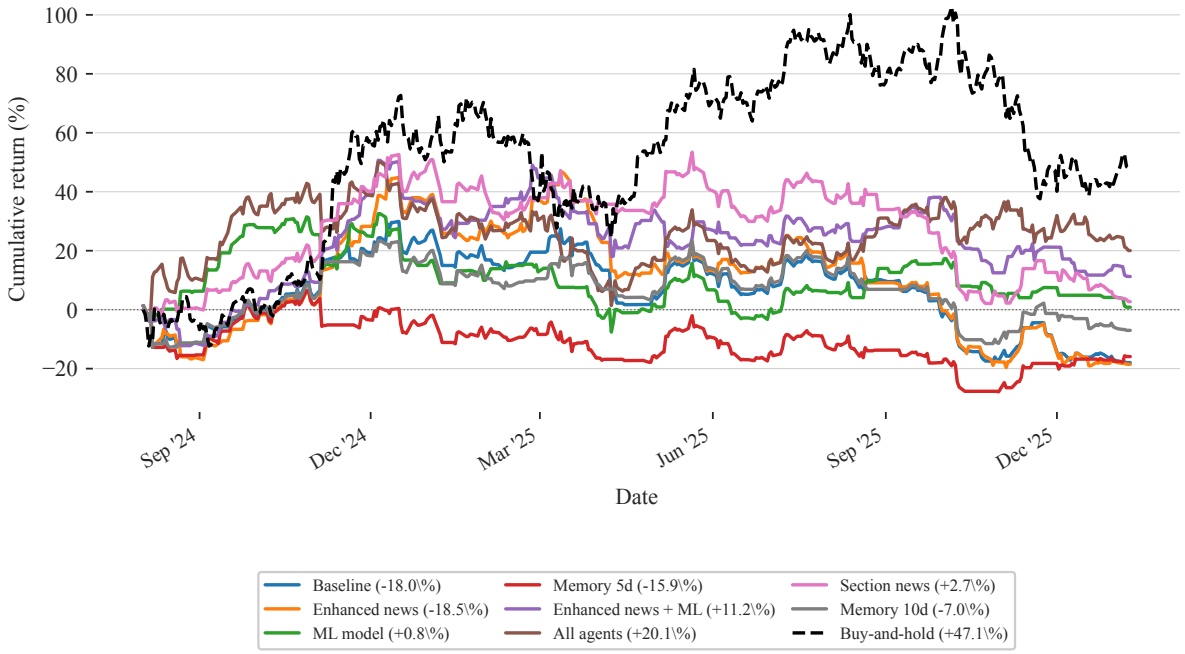

Figure 3: Cumulative Return of BTC

Table 4
Live competition results for DS@GT StockTron from 2026-05-11 to 2026-07-04. CR = Cumulative Return, SR = Sharpe Ratio, B&H = Buy-and-Hold.

Asset	CR	SR	Excess vs. B&H
TSLA	-2.41%	-0.63	+12.41%
BTC	-14.20%	-6.81	+13.05%

baseline system without reflection. More specifically, the agent is provided with a rolling memory of its recent hit rate and past decisions over 5-day and 10-day windows, respectively, with the goal of enabling adaptation to recent signal quality and reducing repeated systematic mistakes.

Contrary to our hypothesis, the introduction of reflective memory consistently deteriorates performance across both assets. For TSLA, the 5-day memory configuration produces the weakest overall result of all experiments (SR: -1.14, MD: -88.2%), while the 10-day memory setup also underperforms the baseline. Similarly, for BTC, neither memory configuration improves risk-adjusted returns, with both variants maintaining negative Sharpe ratios. These findings suggest that the simple reflection mechanism employed in this study may have introduced excessive sensitivity to recent market noise and short-term fluctuations, leading the agent to overreact to recent successes or failures. Overall, the results provide little evidence that naive short-term reflective memory alone improves financial decision-making in this setting.

5.2. Live Competition Results

In addition to the ablation study, we report the live competition performance of the final DS@GT StockTron system. Unlike the ablation experiments, where we isolate tool- and architecture-level effects under a fixed lightweight backbone, the live system uses the complete five-agent architecture with `ChatGPT-5.5`. The live results should therefore be interpreted as an external out-of-sample evaluation of the final submitted system.

Table 4 reports the final leaderboard results as of 2026-07-04. Due to an organizer-side issue, system

decisions for TSLA have been ignored since 2026-06-26, with positions treated uniformly as flat.

The live results show mixed out-of-sample performance. For TSLA, DS@GT StockTron records a cumulative return of -2.41% and a Sharpe ratio of -0.63, but still achieves a positive excess return of +12.41% relative to buy-and-hold over the same period. During the live competition, the system ranked among the top five TSLA submissions until 2026-06-22, when an incorrect directional trade substantially reduced its relative standing. Due to an organizer payload issue, the system had no opportunity to recover its performance thereafter: from 2026-06-26 onward, all TSLA decisions were ignored and TSLA positions remained flat across competitors. As a result, the system finished in the middle of the TSLA leaderboard. For BTC, the system performed poorly in absolute and relative competition terms, with a cumulative return of -14.20%, a Sharpe ratio of -6.81, and a ranking close to the bottom of the leaderboard. Although the excess return versus buy-and-hold is positive because BTC buy-and-hold performance was also weak over the same window, the negative absolute return, low Sharpe ratio, and near-bottom leaderboard ranking indicate that the BTC component did not generalize successfully out of sample.

6. Discussion

The most striking finding in the ablation study is the strong asymmetry between assets. While the best-performing configuration achieves a positive cumulative return of +20.1% and Sharpe ratio of +0.54 for BTC, none of the evaluated configurations generate positive returns for TSLA. This suggests that the usefulness of individual tools and agent interactions is highly regime- and asset-dependent. In particular, the Random Forest model exhibits a pronounced BULLISH recall bias (75.2%), which aligns naturally with BTC's persistent upward trend during the evaluation period. In contrast, TSLA exhibited substantially higher idiosyncratic volatility and more abrupt reversals, reducing the effectiveness of directional signals. The findings therefore indicate that agentic trading performance depends not only on the sophistication of the architecture itself, but also on the compatibility between tool characteristics and the statistical properties of the underlying asset.

Among all evaluated components, enhanced sentiment processing emerges as the most consistently beneficial modification. Both enhanced sentiment variants outperform the naive baseline and improve either risk-adjusted returns, cumulative returns, or drawdown characteristics across the two assets. This supports the hypothesis that signal quality is more important than raw information quantity. Naive aggregation of all available news likely introduces substantial noise, causing the portfolio manager to overreact to irrelevant or weakly informative textual signals. In contrast, more targeted and context-aware sentiment extraction appears to provide more stable and economically meaningful information for downstream decision-making.

The machine-learning forecasting tool provides mixed but informative results. Replacing the technical-analysis agent with the Random Forest model consistently improves BTC performance and, in combination with enhanced sentiment processing, produces some of the strongest overall results in the study. At the same time, the improvements for TSLA remain limited and inconsistent. This highlights an important interaction effect between predictive models and accompanying information sources: the usefulness of a forecasting tool depends not only on its standalone predictive accuracy, but also on how its signal interacts with sentiment and market context within the broader agentic system.

The reflection experiments provide particularly interesting evidence. Contrary to our initial hypothesis, introducing short-term memory over past decisions consistently deteriorates performance across both assets. The 5-day memory configuration produces the weakest TSLA result of all experiments, suggesting that the portfolio manager may overreact to recent successes or failures. This behavior resembles recency-driven extrapolation documented in behavioral finance, where investors place excessive weight on recent outcomes when forming expectations about future returns [13]. The results therefore indicate that naive episodic memory mechanisms are not automatically beneficial in financial agentic systems and may require substantially more sophisticated designs to provide useful context without destabilizing decision-making.

All evaluated active configurations underperform the passive buy-and-hold benchmark, particularly for TSLA. This is an important limitation of the present system. The result suggests that, in a strongly upward-trending development period, the proposed agentic market-timing strategies do not provide sufficient directional accuracy to compensate for missed upside participation. In other words, the system can improve relative to weaker active baselines through better tools and information processing, but it does not yet constitute a competitive replacement for passive exposure in this setting. Consequently, the primary contribution of this study is not the demonstration of absolute outperformance, but the controlled evaluation of how different tools and architectural components affect relative agentic performance. In this regard, the ablation studies still provide useful evidence on the marginal value of enhanced sentiment processing, predictive time-series tools, and memory mechanisms.

The live competition results qualify the ablation findings in two important ways. First, they show how strongly the evaluation regime affects the interpretation of market-timing performance. During the development-period backtest, buy-and-hold achieved very large positive returns for both TSLA and BTC, making it difficult for any active timing strategy to keep pace with passive exposure. In the live period, by contrast, buy-and-hold was negative for both assets, and DS@GT StockTron generated positive excess returns relative to buy-and-hold despite negative absolute returns. This suggests that the system may add value by reducing downside participation in weaker regimes, even when it does not generate positive standalone performance. Second, the live results substantially weaken the favorable BTC conclusion from the ablation study. In the development-period ablations, BTC was the asset for which the full five-agent architecture performed best, achieving a cumulative return of +20.1% and Sharpe ratio of +0.54. In the live competition, however, the BTC component produced a cumulative return of -14.20% and Sharpe ratio of -6.81, finishing close to the bottom of the leaderboard. This backtest–live reversal suggests that the apparent BTC edge was not robust out of sample and was likely sensitive to the development-period regime, the equity-trained forecasting model, the use of an equity-style trading calendar, and the common architecture applied across assets with very different market structures. Consequently, the BTC live result should be interpreted as evidence that the current simplifications are insufficient for cryptocurrency trading.

7. Future Work

The current evaluation in the ablation study is limited to two assets over approximately 18 months. This is insufficient to distinguish skill from luck and to capture full market cycles. Extending to a larger cross-section of U.S. equities would enable stock selection rather than market timing, an approach with stronger empirical support for long-horizon excess returns [14, 15]. Future versions of DS@GT StockTron should also account for the fact that BTC can be traded every day. Alternative machine-learning architectures and training strategies for the time-series-based tool should also be considered. The random forest achieves 51.5% accuracy with a BULLISH-skewed recall profile. Neural network ensembles or gradient-boosted trees may achieve higher precision. Training on cryptocurrency-specific features rather than S&P 500 cross-sections could produce a better signal for BTC, addressing the asset-dependent performance noted above. Furthermore, the failure of hit-rate episodic memory motivates alternative designs. Attention-weighted decision history, embedding-based retrieval of analogous historical episodes, or explicit regime detection could provide useful context without the recency amplification effect observed in our ablations. Lastly, the competition provides 10-K and 10-Q filings alongside news. A dedicated SEC filing agent with structured extraction of key risk factors and guidance language could provide signals orthogonal to the current news pipeline.

8. Conclusions

In this work, we introduced **DS@GT StockTron**, a multi-agent LLM-based trading system for the FinMMEval 2026 benchmark. The system combines specialized analyst agents for technical analysis, sentiment processing, section-aware news analysis, and machine-learning forecasting, coordinated

through an LLM-based portfolio manager. Through a controlled ablation study on TSLA and BTC, we investigated how different tools and reasoning mechanisms affect agentic financial decision-making.

Our results show that the quality of tools materially influences performance. In particular, enhanced and structured sentiment analysis consistently improves results relative to naive sentiment aggregation, supporting the view that signal quality is more important than raw information volume in LLM-based trading systems. Furthermore, incorporating a predictive machine-learning model as an explicit forecasting tool improves performance for BTC and, in combination with enhanced sentiment processing, yields the strongest overall configuration in the development-period ablation study. In contrast, naive reflection via short-term memory of past decisions consistently deteriorates performance, suggesting that simple episodic memory mechanisms may amplify recency bias and market noise rather than improve decision quality.

Overall, our study suggests that carefully designed tool augmentation can improve the relative performance of LLM-based financial agents beyond prompt engineering alone. For the final competition system, we therefore adopt the complete five-agent architecture identified in the ablation studies while upgrading the reasoning backbone to `ChatGPT-5.5`. Live competition results provide a mixed out-of-sample picture: the system remains competitive for TSLA relative to buy-and-hold, but performs poorly for BTC, highlighting the limitations of applying equity-trained tools and equity-style trading assumptions to cryptocurrency markets. These findings suggest that future agentic trading systems may require more asset-specific forecasting tools, trading calendars, and market-structure-aware decision rules.

Acknowledgements

The authors thank the DS@GT ARC community, the Georgia Tech PACE team for providing computational resources [16], and the FinMMEval 2026 [17] organizers for providing the competition infrastructure and dataset. This work was completed as part of the Georgia Tech OMSCS program.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT 5.5 and Claude Sonnet 4.6 in order to: drafting content, generating literature review, paraphrasing and rewording, improving writing style, grammar and spelling check, citation management, and generating Figure 1. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Z. Xie, L. Qian, G. Georgiev, D. Dimitrov, R. Elbadry, F. Zhang, X. Peng, J. Huang, V. Jani, Y. Dai, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, X. Liu, P. Nakov, Overview of the FinMMEval 2026 task 3: Financial decision making, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[2] H. Li, Y. Cao, Y. Yu, S. R. Javaji, Z. Deng, Y. He, Y. Jiang, Z. Zhu, K. Subbalakshmi, G. Xiong, et al., InvestorBench: A benchmark for financial decision-making tasks with LLM-based agents, https://arxiv.org/abs/2412.18174, 2024. ArXiv:2412.18174.

[3] Y. Xiao, E. Sun, D. Luo, W. Wang, TradingAgents: Multi-agents LLM financial trading framework, https://arxiv.org/abs/2412.20138, 2024. ArXiv:2412.20138.

[4] C. Li, Y. Shi, C. Wang, Q. Duan, R. Ruan, W. Huang, H. Long, L. Huang, N. Tang, Y. Luo, Time travel is cheating: Going live with DeepFund for real-time fund investment benchmarking, https://arxiv.org/abs/2505.11065, 2025. ArXiv:2505.11065.

[5] T. J. Moskowitz, Y. H. Ooi, L. H. Pedersen, Time series momentum, Journal of financial economics 104 (2012) 228–250.

[6] C. Krauss, X. A. Do, N. Huck, Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500, European Journal of Operational Research 259 (2017) 689–702. doi:10.1016/j.ejor.2016.10.031.
[7] Virattt, AI hedge fund team, https://github.com/virattt/ai-hedge-fund, 2025. Accessed: 2025-11-04.
[8] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, in: Advances in Neural Information Processing Systems, volume 36, 2023.
[9] S. Gu, B. Kelly, D. Xiu, Empirical asset pricing via machine learning, Review of Financial Studies 33 (2020) 2223–2273. doi:10.1093/rfs/hhaa009.
[10] P. C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, Journal of Finance 62 (2007) 1139–1168. doi:10.1111/j.1540-6261.2007.01232.x.
[11] T. Loughran, B. McDonald, When is a liability not a liability? textual analysis, dictionaries, and 10-Ks, Journal of Finance 66 (2011) 35–65. doi:10.1111/j.1540-6261.2010.01625.x.
[12] L. Qian, X. Peng, Y. Wang, V. J. Zhang, H. He, H. Smith, Y. Han, Y. He, H. Li, Y. Cao, Y. Yu, A. Lopez-Lira, P. Lu, J.-Y. Nie, G. Xiong, J. Huang, S. Ananiadou, When agents trade: Live multi-market trading benchmark for LLM agents, 2025. URL: https://arxiv.org/abs/2510.11695. doi:10.48550/arXiv.2510.11695. arXiv:2510.11695.
[13] R. Greenwood, A. Shleifer, Expectations of returns and expected returns, The Review of Financial Studies 27 (2014) 714–746.
[14] E. F. Fama, K. R. French, Common risk factors in the returns on stocks and bonds, Journal of Financial Economics 33 (1993) 3–56. doi:10.1016/0304-405X(93)90023-5.
[15] E. F. Fama, K. R. French, A five-factor asset pricing model, Journal of Financial Economics 116 (2015) 1–22. doi:10.1016/j.jfineco.2014.10.010.
[16] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: http://www.pace.gatech.edu.
[17] Z. Xie, Y. Dai, R. Elbadry, V. Jani, X. Peng, L. Qian, G. Georgiev, D. Dimitrov, F. Zhang, J. Huang, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, X. Liu, P. Nakov, Overview of FinMMEval 2026: Multilingual and multimodal financial evaluation, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.