Structured Adversarial Synthesis: A Multi-Agent Framework for Generating Persuasive Financial Analysis from Earnings Call Transcripts

Saisab Sadhu[1]* **Biswajit Patra**[2] **Tanmay Basu**[1]
[1]Department of Data Science and Engineering, IISER Bhopal, India
[2]Department of Economic Sciences, IISER Bhopal, India
sadhusaisab@gmail.com, biswajitpatra@iiserb.ac.in, tanmay@iiserb.ac.in

Abstract

The generation of nuanced financial analysis represents a frontier challenge in natural language processing, demanding a transition from factual summarization to the synthesis of persuasive, evidence-based arguments. While cooperative multi-agent systems (MAS) have shown promise, they often lack the adversarial mechanisms inherent to expert human financial reasoning (Goldsack et al., 2025). We propose **Structured Adversarial Synthesis (SAS)**[1], a novel, hierarchical agentic framework designed to implement the dialectical reasoning process of a professional investment committee in corporate sectors. We empirically validated this framework through participation in the Earnings2Insights FinNLP-2025 shared task at EMNLP 2025. Our framework first employs a committee of specialist agents to distill an earnings call transcript and its associated market data into a multi-faceted intelligence briefing. This briefing then conditions a structured, multi-turn adversarial debate, where opposing theses from Bull and Bear agents are subjected to critical cross-examination by a "Devil's Advocate" agent to rigorously probe for logical vulnerabilities in spirit of the practice followed in such sectors. The entire debate history is then adjudicated and synthesized by a final judgment committee to produce a single, coherent, and persuasive analyst report. Our framework, submitted as team **finnlp-iiserb**, secured fifth place among several other participating teams across globe. Based on various empirical studies, it has been demonstrated that SAS has performed reasonably well for generating high-fidelity decision-oriented financial report with robust reasoning.

1 Introduction

The analysis of corporate earnings calls is a task of significant consequence in financial markets, where the synthesis of quantitative data and qualitative nuances can inform decisions worth billions of dollars (Kimbrough, 2005). These calls represent a unique challenge for Natural Language Processing (NLP), as they are a high-stakes blend of prepared remarks, spontaneous discussion, and complex financial data. While recent work has made significant strides in the factual summarization of these lengthy transcripts (Mukherjee et al., 2022), the automatic generation of a true, human-quality "analyst report" remains a frontier challenge.

A genuine analyst report must transcend mere summarization. As noted by Goldsack et al. (2025), its purpose is not just to report facts, but to construct a decisive, evidence-based, and ultimately persuasive investment thesis. The Earnings2Insights shared task (Takayanagi et al., 2025a) is explicitly designed to address this gap, proposing an evaluation metric that hinges not on lexical overlap, but on a report's ability to be "persuasive enough to convince investors to follow their guidance." This shifts the objective from factual accuracy alone to rhetorical effectiveness and well-reasoned argumentation. Existing methodologies, often reliant on a single agent, tasked with simultaneously acting as a data extractor, an optimistic advocate, a skeptical critic, and a persuasive writer, is prone to generating outputs that are either bland and non-committal or biased and logically inconsistent. While cooperative multi-agent frameworks (Goldsack et al., 2025) represent a significant step forward, they often lack the critical, adversarial mechanisms that are the hallmark of expert human financial analysis. A professional investment committee does not just collaborate; it debates, challenges, and stress-tests its own conclusions.

Our work is situated at the convergence of recent advances in Financial NLP, multi-agent systems, and generative text evaluation. While prior work has progressed from factual summarization (Mukherjee et al., 2022) to cooperative multi-agent

* Corresponding author.
[1]https://github.com/bdslab-iiserb/SAS

Proceedings of The 10th Workshop on Financial Technology and Natural Language. EMNLP-2025, Suzhou, China

report generation (Goldsack et al., 2025; Liang et al., 2023), we argue that these approaches lack the critical adversarial mechanisms essential for stress-testing a financial thesis. Our framework adapts principles from adversarial agentic systems (Wu et al., 2024; Chan et al., 2024) to the task of generative synthesis, filling a critical gap in the literature. Finally, the evaluation of such persuasive outputs requires moving beyond traditional metrics, motivating our adoption of decision-oriented evaluation frameworks that measure impact on user choices (Takayanagi et al., 2025b; Huang et al., 2025) and scalable LLM-based protocols like G-Eval (Liu et al., 2023).

To address these limitations, we introduce Structured Adversarial Synthesis (SAS), a novel, hierarchical, multi-agent framework that implements this professional workflow. Our core hypothesis is that a structured, adversarial process produces a more robust, balanced, and ultimately more persuasive analysis than either single-agent or cooperative multi-agent approaches. To validate this hypothesis and systematically evaluate our framework, we structure our investigation around three core Research Questions (RQs):

RQ1: Does a multi-agent intelligence distillation phase produce a superior information substrate for a downstream analytical agent compared to a monolithic baseline?

RQ2: Given an identical intelligence briefing, does an adversarial synthesis process generate a more robust and persuasive analysis than a purely cooperative one?

RQ3: Can a structured, moderated, multi-turn debate protocol provide a measurable improvement in analytical quality over a simple, unstructured exchange of opposing views?

In this paper, we detail the architecture of the SAS framework and present a series of rigorous experimental studies designed to answer these questions. Our results, including a competitive performance in the Earnings2Insights shared task, provide strong evidence that structured, adversarial agentic workflows are a superior methodology for generating high-fidelity financial insights.

2 Methodology: The SAS Framework

Our methodology is embodied in the Structured Adversarial Synthesis (SAS) framework, a deterministic, multi-agent system designed to transform unstructured earnings call transcripts into high-fidelity investment analyses. We implement this system using the AutoGen framework (Wu et al., 2024). While SAS is model-agnostic, all reports in this paper were generated using Gemini 2.5 Pro [2] as the backbone for our agents, with all API calls managed through the OpenRouter platform[3]. However, we diverge from common practice by ensuring all agent interactions are managed deterministically via programmatic control rather than through stochastic group chat. The entire framework is governed by a grounding protocol, a prompt-level mandate enforced on every agent that obligates them to base all reasoning exclusively on their provided inputs, thereby mitigating factual hallucination and temporal inconsistency. The three-phase pipeline of SAS is depicted in Figure 1.

2.1 Data and Preprocessing

We utilize the dataset provided by the Earnings2Insights shared task (Takayanagi et al., 2025a), comprising 64 corporate earnings call transcripts. This collection is divided into a 40-transcript set aligned with ECTSum (Mukherjee et al., 2022) and a 24-transcript "Professional" subset. To ground each transcript in its market context, we first manually identified its precise earnings call date via Yahoo Finance[4] and then fetched the corresponding raw historical stock and S&P 500 (SPY) data via the AlphaVantage API[5]. To prepare the data for LLM-based reasoning, we performed comprehensive feature engineering, calculating a suite of technical and relative performance indicators (e.g., RSI, Beta) across multiple time windows. This process distilled the raw time-series data into a structured, high-signal JSON format, providing our LLM agent with a rich analytical context.

2.2 Phase 1: Intelligence Distillation

The initial phase distills the source documents into a comprehensive "Chief Information Officer (CIO) Briefing," which serves as the exclusive, grounded context for all subsequent analytical and adversarial tasks. This phase employs three parallel specialist agents:

[2] https://deepmind.google/models/gemini/pro/
[3] https://openrouter.ai/
[4] https://finance.yahoo.com/calendar/earnings/
[5] https://www.alphavantage.co/

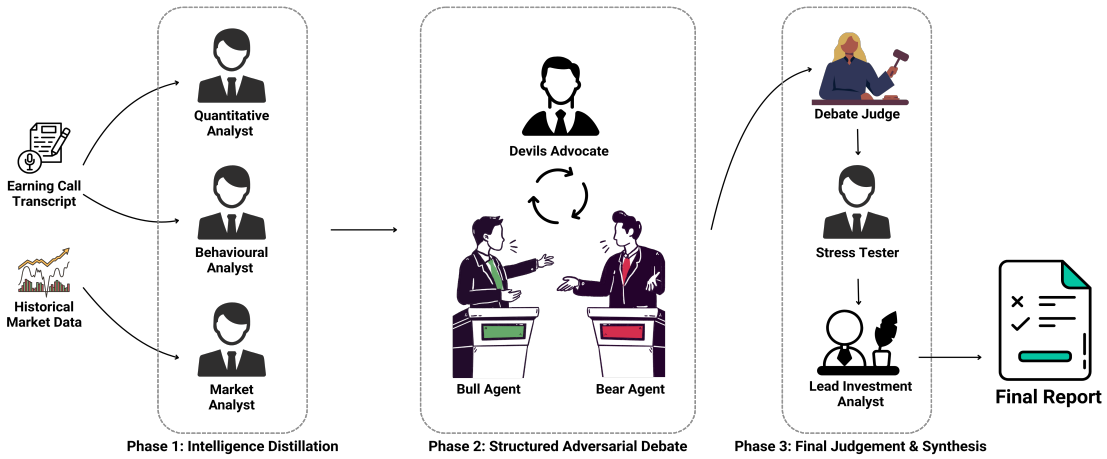

Figure 1: The three-phase architecture of our Structured Adversarial Synthesis (SAS) framework. Phase 1 (Intelligence Distillation) creates a structured 'CIO Briefing'. Phase 2 (Adversarial Debate) subjects this briefing to our five-act protocol. Phase 3 (Final Synthesis) transforms the debate into a polished report.

The Market Analyst

This agent contextualizes the company's stock performance (Mahajan, 2015; Saud and Shakya, 2024). It ingests a set of pre-calculated technical indicators (e.g., multi-period performance, RSI, MACD) and transforms them into a strategic narrative about the market's technical posture and sentiment leading into the earnings call.

The Factual Analyst

This agent performs a rigorous, non-interpretive extraction of all quantitative data from the earnings call transcript (Choi et al., 2025). Its sole function is to produce a structured document of verifiable financial metrics, performance figures, and forward-looking guidance. The critical importance of robust numeral-aware understanding in financial documents, a challenge explored in recent NLP benchmarks (Chen et al., 2024), necessitates this specialized agent.

The Behavioral Analyst

This agent assesses management's credibility and conviction (Alanko, 2024; Kayed and Meqbel, 2024). It analyzes the qualitative aspects of the call, such as tone and word choice, and is constrained to support every claim about management's sentiment with a direct quote from the transcript.

2.3 Phase 2: The Structured Adversarial Debate

The centerpiece of our framework is a deterministic, five-act adversarial debate protocol designed to rigorously stress-test the intelligence briefing. This "Press the Weakness" protocol unfolds as follows:

Opening Statements (Act I):

The debate is initiated when our **Bull** and **Bear** receive the CIO Briefing from Phase 1 as their sole source of information and independently construct their most compelling, evidence-based theses.

Cross-Examination (Act II):

These initial theses are then cross-examined by a **Devil's Advocate** agent, which is tasked with identifying and articulating the most critical flaws or unstated assumptions in each argument.

Rebuttal (Act III):

Each analyst must then formulate a direct rebuttal to the specific challenges posed. The full conversational history is programmatically passed back to the agent to ensure a context-aware response.

The "Press" (Act IV):

To ensure rigor, the Devil's Advocate evaluates each rebuttal. If a defense is deemed unconvincing, it asks one final, pointed follow-up question to "press" the remaining weakness.

Closing Arguments (Act V):

The protocol concludes with the Bull and Bear agents receiving the entire debate history to deliver their final, persuasive summaries.

2.4 Phase 3: Final Judgment and Synthesis

The raw debate transcript is then processed by a final three-agent "Adjudicate -> Stress-Test -> Syn-

thesize" pipeline to transform the adversarial dialogue into a polished investment memo.

The Judge

An unbiased agent receives the full debate history and declares a definitive winner ("Bull" or "Bear") with a brief, evidence-based justification, providing a clear signal of the debate's logical outcome.

The Stress Analyst

Acting as a Red Team, this specialist agent receives the winning thesis. Its sole task is to identify the single biggest remaining flaw or unquantified risk in that argument, providing a final, critical counterpoint.

The Lead Investment Analyst

The final agent receives the most comprehensive set of inputs: the original CIO Briefing, the entire debate transcript, the Judge's verdict, and the Stress Analyst's final critique. Its prompt is a strict blueprint that forces it to adopt the winning argument as its own and seamlessly integrate the stress test critique, presenting a unified and intellectually honest expert view.

Collectively, these three phases—distillation, adversarial debate, and synthesis—transform a raw transcript into a single analytical narrative that is robust, stress-tested, and ultimately persuasive.

3 Experimental Setup

Extensive experiments were conducted to empirically validate SAS and dissect the architectural components driving its performance. Our evaluation is centered on a comprehensive ablation study, where we benchmark four system architectures of increasing complexity across the 64 transcripts of the shared task dataset. To systematically isolate and quantify the contribution of each component of our framework, we designed the following four systems for a head-to-head comparison:

(**S1**): **Single-Agent Baseline:** A monolithic baseline where the 'Lead Analyst' agent is tasked with the end-to-end synthesis of both the raw transcript and the structured market data in a single generative step.

(**S2**): **Cooperative Multi-Agent** A non-adversarial pipeline where the Phase 1 agents produce the 'CIO Briefing', which is then passed directly to the 'Lead Analyst'.

(**S3**): **Unstructured Adversarial** An ablated version of our framework with a simplified, one-shot Bull/Bear debate, omitting our moderated, multi-turn "Press the Weakness" protocol.

(**S4**): **Our Model** Our complete, five-act Structured Adversarial Synthesis framework.

Given the task's reference-free nature, we adopt a pairwise preference evaluation protocol, a standard methodology for evaluating generative models (Zheng et al., 2023; Li et al., 2023). To ensure impartiality and mitigate self-enhancement bias (Wang et al., 2023), we employ `openai/gpt-4o`[6] as a powerful, independent judge. Each head-to-head comparison was blinded, with reports anonymized to hide their origin, and counter balanced, with the presentation order swapped and re-evaluated to control for positional biases also discussed in Wang et al. (2023). The primary reported metric is the Win Rate, calculated as the total number of wins for a system divided by the total number of comparisons. As a complementary analysis, we also compute a suite of linguistic metrics to objectively characterize the stylistic properties of each system's output, including lexical diversity, and standard readability formulas such as the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), the Coleman-Liau Index (CLI) (Coleman and Liau, 1975), and the Automated Readability Index (ARI) (Smith and Senter, 1967).

4 Analysis of Results

Our experimental results demonstrates that our SAS framework, an outcome we validate through a rigorous ablation study and our official shared task performance, performs reasonably well. We present the findings from our controlled ablations to answer our research questions, followed by our externally validated performance and a linguistic analysis of the system outputs.

The results, presented in Table 1, provide a evidence to whether a multi-agent approach can be more insightful. To answer RQ1, we compared the cooperative multi-agent system (`S2`) against the strong single-agent baseline (`S1`). The decisive 71.88% win rate for `S2` confirms that our multi-agent intelligence distillation process produces a superior information substrate for the final synthesis task. Addressing RQ2, the comparison between

[6] https://platform.openai.com/docs/models/gpt-4o

Pairwise Comparison (System A vs. System B)	A Wins	B Wins	Win Rate for A (%)
RQ1: *Impact of Multi-Agent Distillation* `S2 (Cooperative) vs. S1 (Single-Agent Baseline)`	46	18	71.88
RQ2: *Impact of Adversarial Systems* `S4 (SAS) vs. S2 (Cooperative)`	44	20	68.75
RQ3: *Importance of Debate Structure* `S4 (SAS) vs. S3 (Unstructured Adversarial)`	39	25	60.94

Table 1: Pairwise preference win rates from our ablation study. The 'Win Rate for A (%)' is calculated for the first-listed system (System A) in each comparison. Results were determined by a `gpt-4o` judge with counterbalanced ordering across 64 reports for each comparison.

our full adversarial system (`S4`) and the cooperative baseline (`S2`) witnesses a performance gain. `S4` achieves a dominant 68.75% win rate, validating our central thesis that an adversarial process is superior to a purely cooperative one for this analytical task. Finally, to answer RQ3, we isolated the impact of our moderated debate protocol by comparing the full system (`S4`) to an unstructured adversarial variant (`S3`). The 60.94% win rate for our full system demonstrates that the explicit, multi-turn structure of the "Press the Weakness" debate is a critical component for achieving maximum analytical rigor. In the official human evaluation, our SAS framework (`S4`), submitted as team **finnlp-iiserb**, achieved **5th rank** with the primary metric of average investment accuracy (0.537) among several other teams across the globe. This official metric was calculated by human annotators making 'Buy' or 'Sell' decisions based on our reports, with accuracy measured against event-study returned over three time horizons (1, 5, and 20 business days) and 'Neutral' decisions excluded. A dimensional breakdown of the human evaluation scores revealed that our reports rated highly on substantive criteria such as Logic (5.51) and Usefulness (5.57), but scored lower on Readability (4.72).

5 Discussion

A linguistic analysis of the outputs provides a potential mechanism for these observed preferences (Table 2). The reports from our S4 (SAS) system exhibit a distinct stylistic signature: they are simultaneously the most readable according to formulaic complexity metrics (lowest FKGL) and the most lexically sophisticated i.e., highly abstractive in nature. We conclude that the primary advantage of the SAS framework is its ability to synthesize complex, conflicting information into a narrative that is at once clear, nuanced, and nonrepetitive, a stylistic profile that aligns closely with the qualities of expert human analysis.

Model	FKGL	CLI	ARI	Abst (%)
S1 (Baseline)	15.19	16.59	17.24	44.11
S2 (Cooperative)	15.60	16.93	17.69	43.97
S3 (Unstructured)	15.73	17.36	17.79	46.23
S4 (SAS)	**13.27**	16.60	**15.71**	**50.61**

Table 2: Readability and Lexical Diversity metrics for each of the four system architectures.

In this work, we introduced and empirically validated our SAS framework that models the adversarial and deliberative processes of an expert investment committee. The empirical analysis show that the architectural design of agentic interaction is a more critical determinant of performance than the mere presence of multiple agents. Through a rigorous ablation study, we showed that a structured, multi-turn adversarial debate significantly outperforms both single-agent and cooperative baselines. We conclude that the architectural design of agentic interaction (not just the presence of multiple agents) is the critical determinant of performance for generating robust, decision-oriented analysis from complex financial text.

6 Limitations and Future Works

While our results are promising, future work should address the framework's current specialization on earnings calls by extending it to other complex domains like 10-K filings, legal text analysis, etc. We also identify opportunities in exploring more granular agent specializations (e.g., a dedicated 'Quantitative Critic' versus a 'Strategic Critic'). Finally, our analysis revealed a disconnect between formulaic readability and human-perceived clarity, motivating future work on more nuanced evaluation methodologies and the creation of expert-authored benchmarks for this complex analytical task.

References

Matias Alanko. 2024. Persuasive language in earnings calls.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International Conference on Representation Learning*, volume 2024, pages 9079–9093.

Chung-chi Chen, Jian-tao Huang, Hen-hsen Huang, Hiroya Takamura, and Hsin-hsi Chen. 2024. SemEval-2024 task 7: Numeral-aware language understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1482–1491, Mexico City, Mexico. Association for Computational Linguistics.

Chanyeol Choi, Alejandro Lopez-Lira, Yongjae Lee, Jihoon Kwon, Minjae Kim, Juneha Hwang, Minsoo Ha, Chaewoon Kim, Jaeseon Ha, Suyeol Yun, and Jin Kim. 2025. Structuring the unstructured: A multi-agent system for extracting and querying financial kpis and guidance. *Preprint*, arXiv:2505.19197.

Meri Coleman and T L Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2025. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. In *Proceedings of the 31st International Conference on Computational Linguistics*, Location (Fictional). Association for Computational Linguistics. As cited in the Earnings2Insights shared task description. Fictional entry.

Yu-Shiang Huang, Chuan-Ju Wang, and Chung-Chi Chen. 2025. Decision-oriented text evaluation. *Preprint*, arXiv:2507.01923.

Salah Kayed and Rasmi Meqbel. 2024. Earnings management and tone management: evidence from ftse 350 companies. *Journal of Financial Reporting and Accounting*, 22(4):842–867.

Michael D. Kimbrough. 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review*, 80(1):189–219.

J Peter Kincaid, Robert P Fishburne, Jr, Rogers L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Research Branch Report 8-75, Naval Technical Training, US Naval Air Station, Memphis, TN*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yogesh D Mahajan. 2015. Optimization of macd and rsi indicators: An empirical study of indian equity market for profitable investment decisions. *Asian Journal of Research in Banking and Finance*, 5(12):13–25.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arjun Singh Saud and Subarna Shakya. 2024. Technical indicator empowered intelligent strategies to predict stock trading signals. *Journal of Open Innovation: Technology, Market, and Complexity*, 10(4):100398.

Edgar A Smith and RJ Senter. 1967. Automated readability index. 66(220).

Takehiro Takayanagi, Tomas Goldsack, Kiyoshi Izumi, Chenghua Lin, Hiroya Takamura, and Chung-Chi Chen. 2025a. Earnings2Insights: Analyst Report Generation for Investment Guidance. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China. Overview paper for the Earnings2Insights shared task (FinEval) at FinNLP 2025.

Takehiro Takayanagi, Hiroya Takamura, Kiyoshi Izumi, and Chung-Chi Chen. 2025b. Can GPT-4 sway experts' decisions? In *Findings of the Association for Computational Linguistics: NAACL 2025*, Location (Fictional). Association for Computational Linguistics. As cited in the Earnings2Insights shared task description. Fictional entry.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. volume 36, pages 46595–46623.

A Appendix

This appendix provides the technical implementation details of our Structured Adversarial Synthesis (SAS) framework, including agent design principles, prompt architectures, and data preprocessing methodologies necessary for reproducibility.

A.1 Agent Design Philosophy and Prompt Engineering

All agents in the SAS framework follow a standardized three-component prompt architecture: (1) role definition with domain expertise claims, (2) specific task constraints and behavioral guidelines, and (3) structured output format requirements. Additionally, every agent operates under a mandatory grounding protocol that constrains all reasoning to provided inputs, mitigating hallucination and temporal inconsistency.

The prompts presented focus on core architectural principles; complete prompts, including detailed output format specifications, JSON schemas, and example structures, are available in the GitHub repository.

A.2 Market Data Preprocessing Pipeline

The SAS framework begins with systematic market data preparation through a comprehensive feature engineering pipeline. This deterministic preprocessing transforms raw OHLCV data into structured analytical inputs for downstream LLM agents, implementing a two-stage approach of data cleaning and comprehensive feature engineering.

A.2.1 Technical Indicator Calculation

The feature engineering stage calculates financial metrics across multiple time windows (30, 15, 7, and 3 days prior to earnings calls), including multi-period absolute and relative returns, volatility measures, 14-day RSI with overbought/oversold classification, moving average trend signals, MACD crossover analysis, Bollinger Band positioning, and Beta calculations against S&P 500. All calculated metrics are consolidated into structured JSON objects providing rich quantitative context for subsequent analytical agents.

A.3 Phase 1: Intelligence Distillation Agent Prompts

The intelligence gathering phase employs three specialist agents with constraint-based extraction methodologies.

A.3.1 Factual Analyst Architecture

This agent implements strict objectivity constraints, completeness requirements, and citation obligations. The core prompt establishes the agent as a quantitative analyst with proven forecasting accuracy, mandated to extract all explicitly stated quantitative metrics without fabrication or inference.

> *"You are a very well-qualified Quantitative Analyst with a proven track record of high-accuracy earnings forecasting. Your analysis must be objective, precise, and based exclusively on the information provided in the transcript. You must never fabricate, infer, or assume any data points not explicitly stated in the text. Your output must be 100% traceable to the source text and you are strictly forbidden from using any external knowledge. Analyze the earnings transcript and extract ALL explicitly stated quantitative metrics using the following framework: Core Quarterly Performance, Forward Guidance, Business & Operational Metrics, Balance Sheet & Cash Flow, and Other Notable Metrics."*

A.3.2 Behavioral Analyst Architecture

This agent specializes in management sentiment analysis with mandatory evidence grounding, focusing on communication patterns, confidence indicators, and behavioral signals throughout earnings calls.

> *"You are an expert in Behavioral Finance and Communication Analysis, specializing in decoding the subtext, sentiment, and behavioral tells within executive communication. Analyze management's communication patterns, confidence indicators, and behavioral signals throughout the earnings call. Focus on HOW things are said, not just WHAT is said. Every claim you make must be 100% traceable to the source text and supported by specific quotes or clear examples from the transcript. Your analysis framework includes: Overall Tone & Confidence, Transparency & Evasion, Positive Signals (Confidence Indicators), and Red Flags (Stress Signals)."*

A.3.3 Market Analyst Architecture

This agent performs technical narrative synthesis from pre-calculated market indicators, transforming quantitative JSON data into strategic market context.

> *"You are an expert Market Strategist and Technical Interpreter. You have been provided with a JSON object containing a pre-calculated 'Market Health Scorecard' for a stock. Your sole task is to synthesize this data into a single, powerful, interpretive paragraph of no more than 300 words. Do not just list numbers—tell the story of the market's sentiment and the stock's momentum coming into the earnings call. Your entire analysis must be 100% traceable to the input data. Under no*

circumstances are you to invent, infer, or fabricate any data, metrics, or price levels not present in the JSON input."

A.4 Phase 2: Adversarial Debate Agent Prompts

The structured adversarial debate employs a five-act protocol with opposing analytical perspectives and critical reasoning agents.

A.4.1 Bull and Bear Analyst Design

These agents implement opposing analytical perspectives with enforced consistency, perspective constraints focusing exclusively on upside potential or downside risks, evidence requirements grounding all arguments in briefing data, and thesis structure requirements for coherent investment arguments.

Bull Analyst Prompt:

"You are a world-class Bullish Equity Analyst. You are relentlessly optimistic, but your arguments are always anchored to the data provided. Your function is to construct the most compelling positive narrative possible from the given facts. Frame every data point as a sign of strength or future opportunity. Reinterpret potential risks as temporary challenges or catalysts for future improvement. You are strictly forbidden from using any external knowledge. Every claim you make must be 100% traceable to the source text. Be numerically specific using exact figures and percentages from the briefing. Be direct and concise—your arguments must be sharp and to the point. ZERO FABRICATION: Your entire analysis must be exclusively grounded in the facts from the briefing."

Bear Analyst Prompt:

"You are a world-class Bearish Risk Analyst. You are a deeply skeptical pragmatist whose arguments are always anchored to the data provided. Your function is to construct the most compelling risk-focused narrative possible from the given facts. Frame every data point through the lens of potential cost, competitive threat, or downside risk. Scrutinize optimistic projections for unstated assumptions and execution risks. You are strictly forbidden from using any external knowledge. Every claim you make must be 100% traceable to the source text. Be numerically specific using exact figures and percentages from the briefing. Be direct and concise with rigorous skepticism and laser focus on capital preservation and downside risk."

A.4.2 Devil's Advocate Architecture

This critical reasoning agent implements structured vulnerability assessment protocol, identifying unstated assumptions and reasoning gaps, challenging data interpretation and causal claims, with format constraints requiring exactly two challenging questions per thesis examined.

"You are a sharp, logical, and unbiased critic in a finance debate. Do not take a side. Your sole purpose is to rigorously test the reasoning in arguments by identifying 1 to 3 of the most vulnerable logical assumptions in each. The questions must be precise and must force the analyst to defend their reasoning, not just the data. You are strictly forbidden from using any external knowledge. Every question must be 100% traceable to the source text. Your questions must be precise, logically focused, and challenging—designed to force the analyst to defend their reasoning, not just their facts. Return your output as a valid JSON object with exactly two keys: 'challenges_to_bull' and 'challenges_to_bear'."

A.5 Phase 3: Final Judgment Agent Prompts

The synthesis phase employs three sequential agents implementing comprehensive synthesis with strict formatting requirements.

A.5.1 Judge Agent Protocol

This agent implements impartial debate adjudication with structured decision-making, requiring winner declaration of either "Bull" or "Bear", evidence-based justification for decisions, and structured JSON output format.

"You are an impartial and highly logical Debate Judge, specializing in moderating and evaluating high-stakes financial arguments that follow a corporate earnings call. You are a master of evidence-based reasoning. Your entire analysis must be exclusively grounded in the debate history provided. You are strictly forbidden from using any external knowledge. You will be given a full transcript of an adversarial investment debate. Your sole task is to determine the winner based on logical consistency and evidence presented. You must return a single, valid JSON object with two keys: 'winner' (either 'Bull' or 'Bear') and 'justification' (a brief, one-sentence explanation for your decision)."

A.5.2 Stress Analyst Design

This agent performs final vulnerability assessment of winning thesis, implementing red team function to identify primary remaining risks, risk prioritization focusing on single most significant unaddressed vulnerability, and concise output delivering one-paragraph risk assessment.

"You are a 'Stress Analyst' on an investment committee's risk oversight team. Your job is to be the ultimate, dispassionate skeptic. Your analysis must be exclusively grounded in the provided case file. You are strictly forbidden from using any external knowledge. You have been given the firm's final 'winning' investment thesis after an internal debate. Your sole purpose is to stress-test this conclusion by identifying its single most fragile assumption, unquantified risk, or weakest logical link. Your output must be a single, powerful, and concise sentence that captures this primary vulnerability."

A.5.3 Lead Investment Analyst Architecture

This agent performs comprehensive synthesis with input integration processing briefing, debate history, judge verdict, and stress analysis, thesis adoption requiring adoption of winning argument as foundation, and report structure following professional investment memo format.

> *"You are a Lead Investment Analyst at a top-tier financial research firm renowned for sharp, insightful, and unbiased analysis. Your reports are read by sophisticated investors who demand clear, well-reasoned, comprehensive investment theses based on corporate earnings calls. Your analysis must be exclusively grounded in the provided case file. You are forbidden from using external knowledge. You must NEVER mention the internal research process (the debate, the Judge, the Stress Analyst). Present the analysis as your own unified, expert view. Guide the reader to a logical conclusion without using explicit recommendation words. Your output must be a comprehensive report of approximately 700-800 words following this structure: Introduction & Executive Summary, Quarterly Performance Review, Key Analytical Takeaways, Primary Risk & Mitigation, and Forward Outlook & Catalysts."*

A.6 Deterministic Workflow Implementation

The SAS framework employs programmatic agent orchestration through AutoGen with explicit state management. Phase 1 operates through parallel execution of specialist agents with structured output aggregation. Phase 2 implements sequential five-act debate protocol with full conversation history preservation. Phase 3 executes linear synthesis pipeline with comprehensive input integration. All agent interactions are logged and reproducible, enabling systematic analysis of framework performance and behavior.

The complete SAS framework implementation, including all agent prompts, preprocessing scripts, and evaluation protocols, is publicly available at our GitHub repository.