Multi-Agent LLM system for Text-to-Alpha Financial Signal Discovery

by

Ruilin Wu

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business and Economics Honors Program

NYU Shanghai

May 2026

Professor Marti G. Subrahmanyam
Professor Christina Wang
Professor Wendy Jin

Faculty Advisers

Professor Chen Zhao

Thesis Adviser

Contents

1 Introduction **1**

2 Literature Review **4**
2.1 Cross-Sectional Asset Pricing and Factor Models 4
2.2 Financial Textual Analysis 5
2.3 Large Language Models in Financial Applications 6
2.4 Multi-Agent LLM Systems 6
2.5 Position of This Thesis 7

3 System Overview **7**

4 Data **9**
4.1 Investment Universe and Calendar 10
4.2 Structured Market and Fundamental Data 11
4.3 Disclosure Text Collection 12
4.4 Document Master and Stock-Day Registry 13
4.5 Text Feature Coverage and Quality 14
4.6 Final Alignment for Evaluation 15

5 Stage 1 Structured Multi-Agent Baseline **15**
5.1 Graph Structure 16
5.2 Summary Node 17
5.3 Analyst and Strategist Agents 18
5.4 Deterministic Ranking 18
5.5 Trading Agent and Portfolio Construction 19
5.6 Risk Review and Final Portfolio 20
5.7 Stage 1 Outputs 21

6 Stage 2 Disclosure Text Pipeline **21**
6.1 Graph Structure 22
6.2 Document-Level Extraction 23
6.3 Stock-Day Textual State Aggregation 24
6.4 Text Factor Agent 25
6.5 Text Ranking Panel 27
6.6 Interpretability 27

7 Structured-Text Fusion Design **28**
7.1 B1: Score-Level Fusion 28
7.2 B2: Candidate-Pool Reranking 29
7.3 B3: Text Weight Overlay on Stage 1 Names 30
7.4 B4: Final Portfolio Text Overlay 30
7.5 Agent-Executed Fusion 31
7.6 Evaluation Protocol 32

8 Empirical Results **32**
8.1 Main Performance Table 32

8.2 Text-Only Diagnostic . . . 33
8.3 Fusion Results . . . 34
8.4 Agent-Executed Extension . . . 34
8.5 Interpretation . . . 35

9 Case Studies and Interpretability **35**
9.1 Case 1: Text Strongly Favors a Stock Missed by Stage 1 . . . 36
9.2 Case 2: Text Stays Cautious Despite Strong Stage 1 Rank . . . 37
9.3 Evidence Trace . . . 38

10 Limitations and Future Work **39**
10.1 Sample Size and Evaluation Horizon . . . 39
10.2 Universe and Benchmark Scope . . . 40
10.3 Transaction Costs and Turnover . . . 41
10.4 Text Signal Calibration . . . 41
10.5 LLM Reliability and Auditability . . . 42
10.6 From Prototype to Research Platform . . . 42

11 Conclusion **43**

Abstract

Empirical asset pricing relies on structured numerical characteristics derived from prices, trading activity, and accounting data. These signals capture many established return patterns, yet a large part of firm-level information appears first in corporate disclosures. Earnings releases, board resolutions, financing announcements, equity incentive plans, and risk-related filings often contain information that is difficult to represent in a regular stock-day factor panel. Large language models make it possible to extract structured signals from such text at scale, but disclosure-based signals remain difficult to use in quantitative research because the documents are irregular, heterogeneous, and uneven in economic relevance.

This thesis develops a two-stage multi-agent text-to-alpha research prototype for the China A-share STAR Market. Stage 1 constructs a structured quantitative backbone that uses conventional numerical factors to generate cross-sectional rankings and risk-adjusted portfolios. Stage 2 adds a disclosure-text pipeline that converts corporate filings into document-level semantic features, aggregates them into auditable stock-day textual states, and produces text-derived rankings. The empirical design first tests whether disclosure text produces a distinct ranking view, then evaluates whether combining this view with structured factors improves portfolio construction.

Empirical results from the STAR Market show that disclosure text adds economically meaningful information to the structured factor pipeline. Text-derived rankings have low overlap with structured rankings, and the largest disagreements can be traced to specific disclosure narratives such as earnings deterioration, financing activity, governance changes, and innovation-related updates. In the full monthly evaluation from September 2022 to June 2025, portfolios that combine structured factors with disclosure signals improve clearly over the structured baselines. The best combined strategy slightly outperforms the STAR50 benchmark on the main evaluation horizons, while other combined strategies remain close to the benchmark. These results indicate that LLM-extracted disclosure signals are most useful when used as a complementary stock-selection layer inside a disciplined multi-agent investment process. Future extensions can expand the factor library, introduce learned fusion weights, and evaluate weekly or daily rebalancing designs with larger computational budgets.

1 Introduction

Empirical asset pricing studies whether firm characteristics can explain cross-sectional differences in expected returns. Structured numerical signals such as value, momentum, profitability, investment, liquidity, and risk remain central to quantitative equity research because they are measurable, comparable across firms, and naturally aligned with stock-day panels used in portfolio construction [1, 2, 3]. These signals provide a disciplined foundation for ranking stocks, but they do not capture all firm-relevant information at the time it is released.

A substantial part of firm information first appears in corporate disclosures. Earnings explanations, financing announcements, board resolutions, equity incentive plans, product updates, and risk warnings often contain information that may affect investors' assessment of firm prospects. Prior work shows that financial text can contain information relevant to beliefs, prices, and returns [4, 5]. The main difficulty is that disclosures are asynchronous, heterogeneous, and uneven in economic importance. They do not arrive as a clean factor matrix indexed by $(\text{date}, \text{stock\_id})$. To use them in cross-sectional stock selection, disclosure text must be extracted, standardized, aggregated, and evaluated under the same timing discipline as structured factors.

Recent advances in large language models make this transformation more feasible. LLMs can read long and irregular financial documents, identify economically relevant content, and return structured outputs that go beyond coarse positive or negative sentiment [6, 7, 8]. For quantitative finance, however, document understanding is only the first step. A useful text signal must connect filing-level interpretation to stock-level ranking and portfolio evaluation. This requires a research architecture that links semantic extraction, signal construction, ranking, trading, risk review, and performance measurement.

This thesis develops a two-stage multi-agent research framework for text-based stock selection in the China A-share STAR Market. Stage 1 builds a structured quantitative backbone from conventional numerical factors. It moves from factor summarization to analyst interpretation, strategy formation, ranking, trading proposal, risk review, and final portfolio construction. Stage 2 adds a disclosure-text pipeline. It processes corporate filings into document-level semantic features, aggregates those outputs into stock-day textual states, and produces text-derived rankings that can be compared and fused with structured rankings.

The empirical design has two layers. First, a text-only diagnostic evaluates whether disclosure-derived signals generate ranking views that differ from the structured Stage 1 system and whether major ranking disagreements can be traced to specific disclosure narratives. Second, a set of structured-text fusion designs evaluates whether disclosure signals can improve portfolio construction when combined with structured rankings. These designs include score-level fusion, candidate-pool reranking, text-based weight overlays, final portfolio overlays, and an agent-executed extension in which fused rankings enter the full trading and risk-control pipeline.

The STAR Market is a natural setting for this study. Its listed firms are concentrated in technology-intensive sectors where announcements about earnings, financing, innovation, governance, and risk can materially change investors' assessment of future performance. The sample covers approximately 80 STAR Market stocks. The disclosure window runs from January 2022 to June 2025, and the structured trading panel runs from June 2022 to June 2025. The full monthly evaluation uses rebalance dates from September 2022 to June 2025, with 33 valid forward-return observations after excluding the final incomplete horizon.

The results show that disclosure text contributes economically meaningful information when used as a complementary layer inside the structured pipeline. Text-derived rankings

are weakly correlated with structured rankings and have limited top-ranked overlap, which indicates that the textual system captures a different view of firm information. In the full monthly evaluation, combined structured-text strategies improve clearly over the structured baselines. The strongest design slightly outperforms the STAR50 benchmark on both main evaluation horizons, while other combined strategies remain close to the benchmark. These findings suggest that disclosure text is most useful when it helps select among stocks already screened by structured factors.

Research Questions and Contributions. The empirical problem can be written as a mapping from structured characteristics and disclosure documents to a cross-sectional ranking:

$$r_t = \mathcal{R}\left(X_{i,t}^{(S)}, \mathcal{D}_{i,t}^{(L)}\right),$$

where $X_{i,t}^{(S)}$ denotes structured numerical characteristics for stock i at date t, and $\mathcal{D}_{i,t}^{(L)}$ denotes the set of recent disclosure documents available for the same stock-date pair.

In particular, this thesis studies three questions. First, can irregular corporate disclosures be converted into structured, auditable stock-day textual states? Second, do text-derived rankings contain information that differs from structured factor rankings? Third, can text signals improve structured multi-agent baselines and produce benchmark-competitive portfolios when introduced through late-fusion designs?

The contribution is threefold. First, the thesis builds an end-to-end disclosure processing pipeline that links document-level LLM extraction to stock-day signal construction. Second, it provides evidence that disclosure-derived rankings are interpretable and distinct from conventional structured rankings. Third, it evaluates several structured-text fusion designs and finds that candidate-pool reranking is the strongest integration method, slightly outperform-

ing STAR50 in the full monthly evaluation while improving clearly over internal structured baselines.

2 Literature Review

This thesis builds on four strands of research: cross-sectional asset pricing, financial textual analysis, large language models in finance, and multi-agent decision systems.

2.1 Cross-Sectional Asset Pricing and Factor Models

Empirical asset pricing has developed a large body of evidence linking firm characteristics to stock returns. Early factor models emphasize market, size, and value exposures [1]. Subsequent work documents the predictive role of momentum [9], liquidity [10], profitability [11], and investment-related variables [12]. More recent studies examine the large number of proposed return predictors and test their robustness in broad cross-sectional settings [13, 2]. Machine learning methods further show that flexible models can improve empirical asset pricing performance when applied to large panels of firm characteristics [3].

At the same time, the expansion of the factor literature raises the problem of redundancy. Feng, Giglio, and Xiu [13] show that many newly proposed factors become redundant once evaluated against a high-dimensional set of existing factors. Swade et al. [14] reach a similar conclusion from a portfolio-oriented perspective: the factor zoo can be substantially compressed, with about 15 factors sufficient to span the available alpha in their universe. These findings suggest that adding more signals is useful only when the new signals contain information not already captured by the existing factor backbone.

This literature motivates the structured component of the thesis and also clarifies the role of the textual layer. Stage 1 follows the standard empirical asset pricing logic: firm-level numerical

characteristics are organized into a stock-day panel, transformed into cross-sectional rankings, and evaluated through forward returns. The purpose of Stage 2 is not to mechanically expand the factor list. It is to test whether corporate disclosures contain a distinct information layer that can complement structured characteristics. The architectural contribution is that ranking and portfolio construction are decomposed into separate modules for summarization, interpretation, strategy formation, trading, risk review, and evaluation.

2.2 Financial Textual Analysis

The search for information beyond standard numerical factors leads naturally to financial text. Tetlock [4] shows that media language can reflect investor sentiment and predict market outcomes. Loughran and McDonald [5] show that general-purpose word lists can be misleading in financial contexts and introduce finance-specific dictionaries. Later work examines forward-looking corporate disclosure language [15], develops content-analysis methods for extracting return-relevant words [16], and uses text to map product-market relationships across firms [17]. Other studies measure financial constraints and annual report tone from text and relate them to asset prices or return comovement [18, 19].

These studies establish that financial text can carry economically meaningful information. Much of the earlier literature, however, relies on dictionaries, word counts, topic models, or document-level tone measures. Those methods are valuable, but they often compress complex filings into a small number of language statistics. This thesis instead focuses on structured semantic extraction. The goal is to convert disclosure text into stock-day textual states with explicit economic dimensions, supporting evidence, and traceable reasoning.

2.3 Large Language Models in Financial Applications

Recent language models expand the tools available for financial text processing. Domain-adapted models such as FinBERT improve financial sentiment classification relative to generic language models [6]. Larger finance-oriented models such as BloombergGPT show that large-scale pretraining on financial corpora can improve performance on finance-specific language tasks [7]. Other work studies whether general-purpose LLMs can extract information from financial news and produce return-relevant forecasts [8]. Retrieval-augmented methods also show how language models can be connected with external evidence to make outputs more grounded [20].

The role of LLMs in this thesis is semantic extraction rather than open-ended forecasting. Each filing is converted into a structured set of document-level signals, including event type, topic, materiality, confidence, factor-like textual scores, evidence, and reasoning. These outputs are then aggregated and tested in a quantitative ranking framework.

2.4 Multi-Agent LLM Systems

A final related literature studies multi-agent LLM systems. Recent frameworks show that complex tasks can be decomposed across specialized agents that communicate through structured intermediate outputs [21]. This design is useful when a task requires multiple types of reasoning and validation. Financial research has a similar structure: data analysis, strategy formation, portfolio construction, and risk review are separate functions with different objectives.

This thesis uses the multi-agent structure to make the decision process explicit. Stage 1 separates structured factor analysis from ranking, trading, and risk review. Stage 2 separates

disclosure preprocessing, semantic extraction, factor construction, text ranking, and evaluation. The fusion layer then tests whether text-derived rankings improve structured baselines when inserted into the same portfolio process.

2.5 Position of This Thesis

The existing literature establishes three points: structured factors are central to asset pricing, financial text contains return-relevant information, and language models can extract richer signals from unstructured documents. The remaining gap is the connection between filing-level interpretation and portfolio-level testing. A document-level judgment must be aligned with a stock-day panel, compared with structured rankings, fused with existing signals, and evaluated under a consistent backtesting protocol.

This thesis addresses that gap by building an end-to-end multi-agent text-to-alpha prototype. It converts corporate disclosures into auditable stock-day textual states, compares text-derived rankings with structured factor rankings, and evaluates structured-text fusion in a full monthly portfolio setting.

3 System Overview

This thesis builds a two-stage multi-agent research system for structured-text stock selection. The central design requirement is that both numerical and textual information must eventually be represented at the same $(\text{date}, \text{stock\_id})$ level. Once aligned at that level, the two information sources can be ranked, fused, and evaluated under the same portfolio protocol.

Figure 1 gives the high-level architecture. Stage 1 forms the structured quantitative backbone. It converts numerical factor data into cross-sectional rankings and risk-reviewed portfolios. Stage 2 forms the disclosure-text layer. It converts corporate filings into document-level

semantic features, aggregates them into stock-day textual states, and produces text-based rankings. The fusion layer then combines the structured and textual ranking views before the final monthly evaluation.

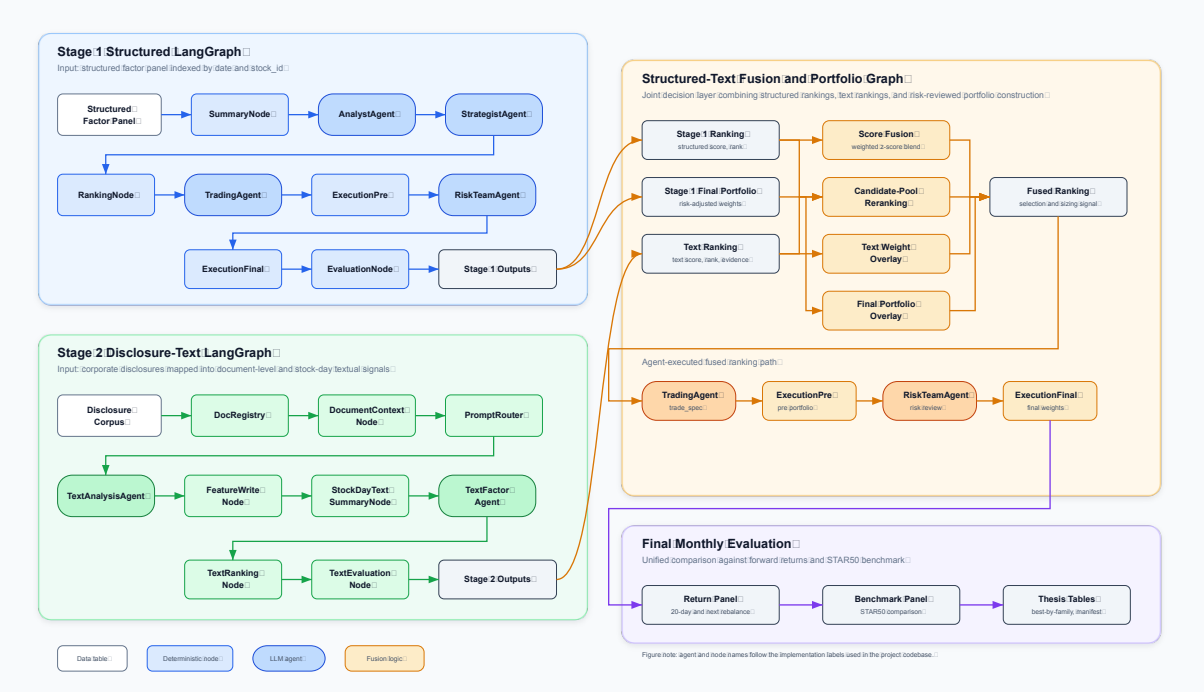

Figure 1: Two-stage multi-agent research pipeline for structured-text stock selection. The Stage 1 graph transforms structured factor data into rankings and risk-reviewed portfolios. The Stage 2 graph transforms corporate disclosures into stock-day textual rankings. The fusion layer combines both information sources before final monthly evaluation.

The full system can be summarized as three linked transformations:

$$X_t \xrightarrow{\mathcal{G}_1} r_t^{(S)}, \qquad \mathcal{D}_t \xrightarrow{\mathcal{G}_2} r_t^{(T)}, \qquad (r_t^{(S)}, r_t^{(T)}) \xrightarrow{\mathcal{F}} w_t,$$

where X_t denotes the structured factor panel, $\mathcal{D}_t$ denotes the disclosure set available at date t, $r_t^{(S)}$ is the Stage 1 structured ranking panel, $r_t^{(T)}$ is the Stage 2 text ranking panel, and w_t is the final portfolio weight vector. These three objects are the core outputs of the system: the structured ranking represents the numerical factor view, the text ranking represents the disclosure-derived view, and the final portfolio combines the two views for forward-return evaluation.

This organization separates the research problem into clear components. Stage 1 establishes a structured baseline. Stage 2 tests whether disclosure text can be converted into auditable stock-day signals. The fusion layer evaluates whether the textual signal improves the structured baseline when used in ranking, selection, or portfolio weighting. Because all outputs are aligned on the same rebalance dates and stock universe, the comparison between structured-only, text-only, and fused strategies is internally consistent.

The rest of the thesis follows this architecture. Section 4 describes the data sources and alignment rules. Section 5 presents the structured multi-agent baseline. Section 6 describes the disclosure-text pipeline. Section 7 introduces the structured-text fusion designs. Section 8 reports the full monthly evaluation.

4 Data

The empirical analysis uses two aligned data layers. The first is a structured stock-day panel built from market and accounting variables. The second is a disclosure-text corpus collected from CNINFO (Juchao Zixun), the official disclosure platform used by listed companies in China.[1] Both layers are mapped to the same stock identifier and trading calendar, which makes it possible to compare structured rankings, text-derived rankings, and fused portfolios on the same monthly rebalance dates.

Let $\mathcal{U}$ denote the STAR Market stock universe and $\mathcal{T}$ denote the trading calendar. For stock $i \in \mathcal{U}$ and trading date $t \in \mathcal{T}$, the structured data layer is represented as a factor vector

$$X_{i,t} = \left(x_{i,1,t}, x_{i,2,t}, \ldots, x_{i,F,t}\right),$$

[1] CNINFO official disclosure platform: `http://www.cninfo.com.cn/`.

where each $x_{i,f,t}$ is a market, liquidity, valuation, risk, or accounting characteristic. The text layer is represented as a set of disclosure documents available to stock i at date t. The central data task is therefore to put $X_{i,t}$ and the disclosure set on a common (i, t) surface.

4.1 Investment Universe and Calendar

The investment universe contains approximately 80 STAR Market stocks. The structured trading panel runs from June 30, 2022 to June 30, 2025. The disclosure collection begins earlier, on January 1, 2022, so that text-based rolling windows have sufficient history before the first monthly rebalance.

The final monthly evaluation uses rebalance dates from September 30, 2022 to June 30, 2025. There are 34 monthly construction dates. Forward-return evaluation uses 33 valid dates because the final rebalance date does not have a complete forward-return horizon.

At each monthly rebalance date t, Stage 1 observes the trailing 60 trading dates:

$$\mathcal{W}_t^S = \{\tau \in \mathcal{T} : \tau \leq t,\ \tau \text{ is one of the last 60 trading dates up to } t\}\,.$$

This window is used only for information available by date t. Portfolio returns are measured after t, so the construction and evaluation windows are separated.

Table 1: Main Data Scope

Item	Description
Universe	80 STAR Market stocks
Structured panel	June 30, 2022 to June 30, 2025
Disclosure window	January 1, 2022 to June 30, 2025
Monthly rebalance dates	September 30, 2022 to June 30, 2025
Number of rebalance dates	34 portfolio construction dates
Forward-return evaluation dates	33 valid dates
Primary benchmark	STAR50
Additional benchmark series	HS300 and CSI index panels

4.2 Structured Market and Fundamental Data

The structured data layer is built from price, volume, valuation, and accounting variables for the STAR Market universe. Raw market and fundamental fields are collected from Wind and transformed into a stock-day factor panel indexed by (date, stock_id).

Two structured panels are used. The first, denoted `signals`, contains technical and market-based signals. The second, denoted `signals_fund`, adds accounting and fundamental signals. This split is deliberate. Some fundamental variables overlap economically with the disclosure-text layer, especially for earnings, profitability, financing, and growth-related information. Treating the two panels separately creates an ablation setting: the technical-only panel tests whether text improves a market-signal baseline, while the technical-plus-fundamental panel tests whether text still adds information after Stage 1 already observes accounting fundamentals.

Both panels contain 56,749 stock-day rows across 727 trading dates and 80 stocks. The technical-only panel contains 36 columns, while the technical-plus-fundamental panel contains 49 columns. The factor set covers momentum, reversal, low volatility, beta, liquidity, valuation, profitability, investment, balance-sheet quality, and cash-flow yield measures.

Table 2: Structured Data Panels

Panel	Rows	Dates	Stocks	Main Content
`signals`	56,749	727	80	Technical, liquidity, risk, and valuation signals
`signals_fund`	56,749	727	80	Technical signals plus accounting fundamentals
`index_panel`	2,181	727	–	STAR50, HS300, and CSI index closes

Forward returns are computed from daily stock returns. The 20-trading-day forward return is

$$R_{i,t}^{(20)} = \prod_{h=1}^{20} \left(1 + r_{i,t+h}\right) - 1,$$

where $r_{i,t+h}$ is the daily return of stock i on the h-th trading day after rebalance date t. The next-rebalance return is

$$R_{i,t}^{(\text{next})} = \prod_{\tau \in \mathcal{T}:\, t<\tau\le t^{+}} \left(1 + r_{i,\tau}\right) - 1,$$

where t^{+} is the next monthly rebalance date. STAR50 is used as the primary benchmark because it is the closest index benchmark for the sample universe.

4.3 Disclosure Text Collection

The textual data layer is collected from CNINFO. The crawler starts from the 80-stock universe, resolves each stock's CNINFO organization identifier, and queries the historical announcement endpoint by stock, disclosure category, and 30-day date chunks. It keeps PDF announcements and records the announcement identifier, stock code, company name, announcement time, category, title, PDF URL, and query window.

The crawler covers twelve disclosure categories, which are mapped into three document groups: report, event, and governance. Report documents contain periodic financial reporting and earnings preannouncements. Event documents contain financing, incentive, unlock, and risk-related announcements. Governance documents contain board resolutions and shareholder meeting materials.

After metadata collection, a batch converter downloads each PDF and extracts text. The conversion manifest records PDF paths, text paths, page counts, text length, extraction status, and errors. Low-text documents are retained but flagged, which preserves the disclosure record while allowing later stages to use quality filters.

Table 3: Disclosure Categories

Document Group	Document Types	Economic Content
Report	Annual, semiannual, and quarterly reports; earnings preannouncements	Earnings performance, profitability, growth, and operating outlook
Event	Equity incentives, seasoned offerings, unlock events, and risk warnings	Financing decisions, incentive alignment, share supply, and downside risk
Governance	Board resolutions and shareholder meetings	Governance decisions, control rights, and shareholder-related actions

4.4 Document Master and Stock-Day Registry

Crawler metadata and conversion results are merged into a standardized document master table. Each row corresponds to one announcement and stores the announcement identifier, stock identifier, category, document type, document group, announcement date, effective date, title, PDF path, text path, page count, text length, extraction status, and low-text flag.

The timing rule is central to the no-lookahead design. Let $a(d)$ denote the public announcement date of document d. Its effective date is defined as the first trading day strictly after the announcement date:

$$e(d) = \min\{\tau \in \mathcal{T} : \tau > a(d)\}.$$

A disclosure can enter a signal only on or after $e(d)$. This rule is conservative for announcements released after market close and keeps the text layer aligned with the structured trading calendar.

For each stock i and rebalance date t, the 60-calendar-day disclosure window is

$$\mathcal{D}_{i,t}^{60} = \{d : \text{stock}(d) = i,\ t - 59 \text{ calendar days} \le e(d) \le t\}\,.$$

This definition differs from the Stage 1 structured window: Stage 1 uses 60 trading days, while Stage 2 uses 60 calendar days of effective disclosures.

The final document master contains 10,465 announcements across 80 stocks. Among them, 10,267 have successful text extraction, 197 are marked as low-text, and 1 has a download failure. The document master is then grouped into a stock-day registry, where each row corresponds to one (effective date, stock_id) pair with at least one associated disclosure.

Table 4: Disclosure Data Construction Summary

Dataset	Rows	Description
`document_master.parquet`	10,465	Standardized disclosure-level metadata
`stock_day_doc_registry.parquet`	4,117	Stock-day disclosure bundles
`document_features_full_monthly.parquet`	10,275	LLM-extracted document features
`text_rank_panel_full_monthly.parquet`	2,675	Monthly stock-day text rankings

The stock-day registry contains 4,117 disclosure bundles. The average bundle contains 2.54 documents, although some stock-days contain many more filings because companies often release board resolutions, shareholder meeting materials, legal opinions, and related announcements together.

4.5 Text Feature Coverage and Quality

The full monthly text extraction file contains 10,275 document-level feature rows. The parse success rate is nearly complete: 10,274 rows are successfully processed, and only one row fails because the corresponding text file is missing. There are no duplicated announcement identifiers in the feature table.

The average processed text length is approximately 4,143 characters. The average materiality score is approximately 0.518, and the average extraction confidence is approximately 0.797. These values are treated as data-quality diagnostics rather than performance evidence. Their main use is to document that the text layer has broad coverage and that extracted signals

can be traced to confidence, evidence, and reasoning fields.

Table 5: Text Extraction Diagnostics

Metric	Value
Document feature rows	10,275
Successful parses	10,274
Failed parses	1
Duplicate announcement identifiers	0
Average processed text length	4,143 characters
Average materiality score	0.518
Average confidence score	0.797

4.6 Final Alignment for Evaluation

The final monthly dataset aligns structured factors, text features, forward returns, and benchmark returns on the same rebalance surface. For each t, the structured panel provides $X_{i,t}$, while the text pipeline aggregates $\mathcal{D}_{i,t}^{60}$ into stock-day textual states. This produces a structured ranking panel and a text ranking panel over the same date-stock grid.

This alignment is what makes later comparisons meaningful. Structured-only, text-only, and fused strategies use the same universe, rebalance dates, forward-return definitions, and benchmark returns. Performance differences therefore come from signal construction and portfolio design rather than calendar mismatch.

5 Stage 1 Structured Multi-Agent Baseline

Stage 1 is the structured quantitative baseline. It converts numerical market and fundamental signals into monthly stock rankings and risk-reviewed portfolios. The same LangGraph architecture is run on the two structured panels described in Section 4. This allows the thesis to compare a technical-only baseline with a technical-plus-fundamental baseline before introducing disclosure text.

At each monthly rebalance date t, Stage 1 observes the trailing structured window $\mathcal{W}_t^S$, builds a ranking for the current cross-section, and passes the ranking through portfolio construction and risk review. Abstractly, the Stage 1 graph maps structured data into a ranking and portfolio:

$$\mathcal{G}_1 : \{X_{i,\tau} : i \in \mathcal{U},\ \tau \in \mathcal{W}_t^S\} \longrightarrow \left(r_t^{(S)}, w_t^{(S)}\right),$$

where $r_t^{(S)}$ is the structured ranking and $w_t^{(S)}$ is the final Stage 1 portfolio weight vector.

5.1 Graph Structure

The Stage 1 graph combines deterministic nodes with LLM-based agents. Deterministic nodes handle statistical summaries, ranking, execution, and evaluation. Agents interpret the factor environment, choose factor exposures, design the portfolio rule, and review risk.

Table 6: Stage 1 Structured Multi-Agent Workflow

Module	Type	Main Role
`summary_node`	Deterministic node	Computes factor statistics, stock-level factor snapshots, and market conditions.
`AnalystAgent`	LLM agent	Interprets the factor summary and produces a market regime view, style preference, and risk flags.
`StrategistAgent`	LLM agent	Converts the analyst view into selected factors and factor weights.
`ranking_node`	Deterministic node	Applies the strategy specification to the current cross-section and produces stock-level scores and ranks.
`TradingAgent`	LLM agent	Designs a rule-based portfolio construction plan from the ranking and upstream reports.
`execution_pre_node`	Deterministic node	Builds a preliminary long-only portfolio from the ranking and trade specification.
`RiskTeamAgent`	LLM agent	Reviews the preliminary portfolio and proposes final risk settings.
`execution_final_node`	Deterministic node	Applies the risk-adjusted trade specification and constructs the final Stage 1 portfolio.

The graph state records `summary`, `analyst_report`, `strategy_spec`, `ranking`, `trade_spec`, `portfolio_pre`, `risk_report`, and `portfolio`. This makes the decision path inspectable

from raw factor statistics to final holdings.

5.2 Summary Node

The `summary_node` transforms the 60-trading-day window into a compact statistical representation. For each factor f, raw values are direction-adjusted so that larger values correspond to more attractive signals:

$$\widetilde{x}_{i,f,t} = s_f x_{i,f,t}, \qquad s_f \in \{-1, 1\}.$$

For example, a lower volatility or lower leverage variable can be multiplied by -1, making the adjusted signal comparable with factors where higher is better.

The node computes cross-sectional z-scores on the rebalance date:

$$z_{i,f,t} = \frac{\widetilde{x}_{i,f,t} - \mu_{f,t}}{\sigma_{f,t}}, \qquad \mu_{f,t} = \frac{1}{N_t} \sum_{i=1}^{N_t} \widetilde{x}_{i,f,t},$$

where N_t is the number of stocks available at date t. These z-scores summarize each stock's relative position within the current cross-section.

The node also computes historical information coefficients over the rolling window:

$$IC_{f,\tau} = \rho_{\text{Spearman}} \left(\widetilde{x}_{i,f,\tau}, R^{(h)}_{i,\tau} \right), \qquad \tau \in \mathcal{W}^S_t.$$

Here $R^{(h)}_{i,\tau}$ is a forward return target and ρ_{Spearman} denotes the cross-sectional Spearman correlation. These statistics give downstream agents a concise view of recent factor performance, dispersion, and reliability.

The summary also includes a market snapshot based on STAR50, CSI, and HS300 index

data. It records recent index return, realized volatility, and drawdown measures, allowing the agents to condition factor selection on the market environment.

5.3 Analyst and Strategist Agents

The `AnalystAgent` reads the summary and produces an investment interpretation. Let M_t denote the statistical summary at date t. The analyst output can be written as

$$q_t = \mathcal{A}_{\text{analyst}}(M_t),$$

where q_t contains the market regime, style preference, and risk flags.

The `StrategistAgent` converts this interpretation into a machine-readable strategy specification:

$$\theta_t = \mathcal{A}_{\text{strategist}}(M_t, q_t) = (S_t, \omega_t)\,,$$

where S_t is the selected factor set and $\omega_t = \{\omega_{f,t} : f \in S_t\}$ is the factor-weight vector. This creates a clean division of labor. The LLM chooses the factor emphasis, while the ranking node applies a fixed scoring rule.

5.4 Deterministic Ranking

The `ranking_node` applies θ_t to the current stock cross-section. Each selected factor is direction-adjusted, winsorized, standardized, and checked for coverage. Technical factors are median-imputed within the cross-section to maintain stable coverage. Fundamental factors are left missing when unavailable, which preserves the point-in-time nature of accounting information.

The composite Stage 1 score for stock i is

$$S_{i,t}^{(1)} = \left(\frac{\sum_{f \in S_t} \omega_{f,t} z_{i,f,t} \mathbb{I}_{i,f,t}}{\sum_{f \in S_t} |\omega_{f,t}| \mathbb{I}_{i,f,t}} \right) \cdot C_{i,t},$$

where $\mathbb{I}_{i,f,t}$ indicates that factor f is available for stock i, and $C_{i,t}$ is a coverage adjustment. Stocks with too few usable factors receive a neutral score. The ranking is then

$$r_{i,t}^{(S)} = 1 + \sum_{j \neq i} \mathbb{I}\left(S_{j,t}^{(1)} > S_{i,t}^{(1)} \right),$$

so rank 1 corresponds to the highest structured score.

The scoring mechanics are fixed across experiments. This is important because later comparisons between structured-only, text-only, and fused portfolios should reflect signal differences rather than changes in the scoring rule.

5.5 Trading Agent and Portfolio Construction

After ranking, the `TradingAgent` receives $r_t^{(S)}$, the analyst report q_t, and the strategy specification θ_t. It returns a rule-based trade specification:

$$\phi_t = \mathcal{A}_{\text{trade}}\left(r_t^{(S)}, q_t, \theta_t \right),$$

where ϕ_t includes `top_k`, `weighting_scheme`, `max_single_weight`, `min_single_weight`, and `cash_weight`. The agent does not directly assign stock-level weights.

The preliminary holding set is defined by the top K_t ranked names:

$$H_t = \left\{ i \in \mathcal{U} : r_{i,t}^{(S)} \leq K_t \right\}.$$

The execution node then maps the holding set and trade specification into a long-only portfolio satisfying

$$w_{i,t} \geq 0, \qquad w_{i,t} = 0 \ \text{ for } i \notin H_t, \qquad w_{i,t} \leq w_{\max,t}, \qquad \sum_{i \in \mathcal{U}} w_{i,t} \leq 1.$$

The difference between one and the total stock weight is the cash weight. The node also records concentration diagnostics, including the Herfindahl index

$$HHI_t = \sum_{i \in \mathcal{U}} w_{i,t}^2,$$

and the effective number of holdings

$$N_t^{\text{eff}} = \frac{1}{HHI_t}.$$

5.6 Risk Review and Final Portfolio

The `RiskTeamAgent` reviews the preliminary portfolio before final execution. It contains an optimistic risk officer, a conservative risk officer, and a risk manager. The two officers evaluate the same proposed portfolio from different risk preferences. The risk manager reconciles their views and proposes final adjustments to cash weight, maximum single-stock weight, and portfolio breadth.

Formally, if w_t^{pre} is the preliminary portfolio, the risk team produces an adjusted trade specification:

$$\phi_t^* = \mathcal{A}_{\text{risk}}\left(\phi_t, w_t^{\text{pre}}, q_t, \theta_t\right).$$

The `execution_final_node` rebuilds the portfolio using ϕ_t^*. The result is the final Stage 1

portfolio $w_t^{(S)}$, which serves as the structured multi-agent portfolio baseline.

5.7 Stage 1 Outputs

Stage 1 produces two outputs used later in the thesis. The first is the structured ranking panel, containing $S_{i,t}^{(1)}$ and $r_{i,t}^{(S)}$ for each rebalance date. The second is the final risk-reviewed portfolio $w_t^{(S)}$.

Portfolio performance is evaluated after the rebalance date. For horizon h, the Stage 1 portfolio return is

$$R_{p,t}^{(h)} = \sum_{i \in \mathcal{U}} w_{i,t}^{(S)} R_{i,t}^{(h)}.$$

The main horizons are $h = 20$ trading days and the next monthly rebalance horizon. These same horizons are used for text-only and fused strategies, which keeps later comparisons internally consistent.

6 Stage 2 Disclosure Text Pipeline

Stage 2 is the disclosure-text graph. Its role is to transform irregular corporate announcements into stock-day textual states and text-derived rankings. The data collection and timing rules were described in Section 4; this section focuses on the modeling pipeline after the disclosure corpus has been constructed.

At a high level, Stage 2 maps the effective disclosure set $\mathcal{D}_{i,t}^{60}$ into a text score and ranking:

$$\mathcal{G}_2 : \mathcal{D}_{i,t}^{60} \longrightarrow \left(Z_{i,t}^{(T)}, r_{i,t}^{(T)} \right),$$

where $Z_{i,t}^{(T)}$ is the stock-day text score and $r_{i,t}^{(T)}$ is the corresponding text-derived rank. The graph is designed to preserve an evidence trail from the final text score back to the underlying

disclosures.

6.1 Graph Structure

The Stage 2 graph combines document-level LLM extraction with deterministic aggregation and ranking. The LLM is used for semantic interpretation at the document level. The stock-day aggregation step is deterministic, while the factor-construction layer is framed as a semi-automated `TextFactorAgent`: it uses a preset factor taxonomy for the baseline score and reserves optional slots for newly discovered textual signals.

Table 7: Stage 2 Disclosure-Text Graph

Module	Type	Main Role
`DocumentContextNode`	Node	Loads and preprocesses one disclosure, validates metadata, and routes the document to the appropriate prompt family.
`TextAnalysisAgent`	Agent	Extracts document-level semantic fields, signed text factors, materiality, confidence, evidence, reasoning, and optional emerging signals.
`DocumentFeature WriteNode`	Node	Validates the LLM output and writes a standardized document-feature row.
`StockDayText SummaryNode`	Node	Aggregates document-level features into rolling stock-day textual states.
`TextFactorAgent`	Agent	Converts stock-day textual states into text factor scores and records candidate emerging factors for later validation.
`TextRankingNode`	Node	Ranks stocks cross-sectionally by text score on each rebalance date.
`TextEvaluationNode`	Node	Aligns text rankings with forward returns for later comparison and fusion analysis.

This graph separates semantic extraction from quantitative scoring. The extraction agent reads the filing and produces structured document features. The downstream graph then decides how those features are aggregated, weighted, and ranked.

6.2 Document-Level Extraction

For each disclosure document d, the context node constructs a standardized input:

$$c_d = C(d),$$

where c_d contains identity fields, effective date, document type, document group, title, processed text, selected sections, and the prompt routed by document group. Reports, event announcements, and governance documents use different prompt templates because they carry different economic content.

The `TextAnalysisAgent` maps the document context into a structured output:

$$y_d = \mathcal{A}_{\text{text}}(c_d).$$

The output contains event type, primary topic, sentiment, materiality, confidence, summary, evidence, reasoning, and eight signed textual factor signals:

$$\mathbf{g}_d = \left(g_d^{earn}, g_d^{grow}, g_d^{prof}, g_d^{risk}, g_d^{fin}, g_d^{innov}, g_d^{align}, g_d^{gov}\right), \qquad g_{d,k} \in \{-1, 0, 1\}.$$

These dimensions correspond to earnings, growth, profitability, risk, financing, innovation, shareholder alignment, and governance quality. A positive value indicates favorable disclosure content along that dimension, a negative value indicates unfavorable content, and zero indicates no clear signal.

In addition to the preset taxonomy, the schema reserves two optional slots for semi-

automated factor discovery:

$$\mathbf{u}_d = \big((n_{d,1}, v_{d,1}), (n_{d,2}, v_{d,2})\big), \qquad v_{d,j} \in \{-1, 0, 1\}.$$

Here $n_{d,j}$ is a model-proposed emerging signal name and $v_{d,j}$ is its signed value. These fields allow the pipeline to record recurring disclosure themes that do not fit neatly into the eight preset factors. They are not included in the baseline score unless later validated and promoted into the factor set.

The `DocumentFeatureWriteNode` validates the output against a strict schema before writing it to the document feature table. This step matters because later scoring depends on typed fields rather than free-form summaries. The evidence and reasoning fields are retained for interpretability, but they are not directly used as numerical scores.

6.3 Stock-Day Textual State Aggregation

The next step aggregates document-level signals into stock-day textual states. For stock i at rebalance date t, the eligible disclosure set is $\mathcal{D}_{i,t}^{60}$, as defined in Section 4. Each document receives a weight based on recency and materiality:

$$a_{d,t} = \lambda(t - e(d))\,(0.25 + 0.75 m_d),$$

where $m_d \in [0, 1]$ is the document materiality score and $\lambda(\cdot)$ is a discrete recency decay function. In the baseline specification,

$$\lambda(\Delta) = \begin{cases} 1.0, & 0 \le \Delta \le 5, \\ 0.7, & 6 \le \Delta \le 20, \\ 0.4, & 21 \le \Delta \le 40, \\ 0.2, & 41 \le \Delta \le 59. \end{cases}$$

More recent and more material documents therefore receive larger weights, while older documents still contribute within the 60-calendar-day window.

For each textual factor k, the continuous stock-day factor score is the weighted average:

$$\bar{g}_{i,k,t} = \frac{\sum_{d \in \mathcal{D}_{i,t}^{60}} a_{d,t} g_{d,k}}{\sum_{d \in \mathcal{D}_{i,t}^{60}} a_{d,t}}.$$

If a stock has no disclosure in the window, the graph assigns a neutral text state with zero factor scores. This keeps the text ranking panel aligned with the Stage 1 monthly universe.

The graph also records stock-day metadata: number of documents in the window, document group counts, average materiality, maximum materiality, average confidence, top titles, top summaries, and evidence snippets. These fields support case studies and make the text signal auditable.

6.4 Text Factor Agent

The `TextFactorAgent` converts stock-day textual states into a scalar text score. It combines the eight preset textual factors with fixed baseline weights. Let η_k denote the weight

assigned to textual factor k. The core score is

$$C_{i,t}^{(T)} = \sum_{k=1}^{8} \eta_k \bar{g}_{i,k,t}.$$

The baseline weights are shown in Table 8. Earnings, profitability, and risk receive the largest weights because they are closest to near-term firm fundamentals. Innovation, financing, shareholder alignment, and governance are included with smaller weights because they often matter through longer or more conditional channels.

Table 8: Text Factor Weights

Textual Factor	Weight
Earnings	1.00
Growth	0.90
Profitability	1.00
Risk	1.00
Financing	0.50
Innovation	0.60
Shareholder alignment	0.40
Governance quality	0.30

The final raw text score applies three mild multipliers:

$$Z_{i,t}^{(T)} = C_{i,t}^{(T)} \cdot M_{i,t}^{doc} \cdot M_{i,t}^{report} \cdot M_{i,t}^{mat}.$$

Here $M_{i,t}^{doc}$ adjusts for document presence, $M_{i,t}^{report}$ adjusts for the share of report-style documents, and $M_{i,t}^{mat}$ adjusts for average materiality. These multipliers are capped and intentionally modest. Their purpose is to reflect the strength of the textual evidence without allowing document quantity alone to dominate the score.

The emerging-factor fields can be viewed as a semi-automated extension layer. The baseline implementation keeps them separate from $Z_{i,t}^{(T)}$, but the graph records their names and signs

so that repeated candidate factors can be audited across documents. In later versions, stable emerging factors could be promoted into the preset factor set after coverage and return tests.

6.5 Text Ranking Panel

The `TextRankingNode` ranks stocks cross-sectionally on each rebalance date. The raw score $Z_{i,t}^{(T)}$ is also standardized within date:

$$\widetilde{Z}_{i,t}^{(T)} = \frac{Z_{i,t}^{(T)} - \mu_t^{(T)}}{\sigma_t^{(T)}},$$

where $\mu_t^{(T)}$ and $\sigma_t^{(T)}$ are the cross-sectional mean and standard deviation of text scores at date t. The text rank is

$$r_{i,t}^{(T)} = 1 + \sum_{j \neq i} \mathbf{1}\left(Z_{j,t}^{(T)} > Z_{i,t}^{(T)}\right).$$

The output is a monthly text ranking panel with the same date-stock surface as Stage 1. This panel can be evaluated on its own or fused with structured rankings in the next stage of the thesis.

6.6 Interpretability

Stage 2 is designed to keep the text signal traceable. A high or low stock-day text score can be decomposed into its contributing documents, document weights, factor-level signals, materiality scores, confidence scores, and evidence snippets. This allows ranking disagreements between Stage 1 and Stage 2 to be explained through specific disclosures, rather than treated as opaque score differences.

The result of Stage 2 is a stock-day textual state: a compact numerical representation of recent disclosure information, backed by the documents and evidence that generated it.

7 Structured-Text Fusion Design

After Stage 1 and Stage 2 produce rankings on the same monthly date-stock surface, the thesis tests whether disclosure text improves the structured baseline. Let $S_{i,t}$ denote the Stage 1 structured score and $T_{i,t}$ denote the Stage 2 text score for stock i at rebalance date t. Their within-date standardized versions are

$$\widehat{S}_{i,t} = \frac{S_{i,t} - \mu_t^S}{\sigma_t^S}, \qquad \widehat{T}_{i,t} = \frac{T_{i,t} - \mu_t^T}{\sigma_t^T}.$$

The fusion designs differ by where the text layer enters the investment process: ranking, candidate selection, position sizing, or final portfolio overlay.

Table 9: Structured-Text Fusion Families

Family	Fusion Point	Design Logic
B1 score fusion	Ranking score	Blend standardized structured and text scores, then rank the full cross-section.
B2 candidate-pool reranking	Stock selection	Use Stage 1 to define an investable pool, then use text to rerank or filter names inside the pool.
B3 text weight overlay	Portfolio sizing	Keep the Stage 1 selected names and adjust their weights using text scores.
B4 final portfolio overlay	Risk-reviewed portfolio	Start from the final Stage 1 portfolio and apply a text-based weight overlay.
Agent-executed fusion	Trading and risk graph	Send fused rankings through `TradingAgent`, execution, `RiskTeamAgent`, and final execution.

7.1 B1: Score-Level Fusion

The first design combines Stage 1 and Stage 2 at the score level. For a fusion weight $\lambda \in [0, 1]$, the fused score is

$$F_{i,t}^{B1}(\lambda) = (1 - \lambda)\widehat{S}_{i,t} + \lambda\widehat{T}_{i,t}.$$

A value of $\lambda = 0$ reproduces the structured ranking, while $\lambda = 1$ gives the text-only ranking. Intermediate values test whether the two signals are complementary after cross-sectional standardization. The fused rank is

$$\rho_{i,t}^{B1} = 1 + \sum_{j \neq i} \mathbf{1}\left(F_{j,t}^{B1} > F_{i,t}^{B1}\right).$$

The monthly portfolio is formed from the top-ranked names under $\rho_{i,t}^{B1}$. The tested grid is

$$\lambda \in \{0.00, 0.10, 0.25, 0.50, 0.75, 1.00\}.$$

7.2 B2: Candidate-Pool Reranking

The second design gives Stage 1 the first selection right. For each rebalance date, Stage 1 defines a candidate pool:

$$\mathcal{P}_t(K) = \left\{i \in \mathcal{U}_t : \rho_{i,t}^{S} \leq K\right\},$$

where $\rho_{i,t}^{S}$ is the structured rank and K is the pool size. Text is then used only inside this structured candidate pool:

$$\mathcal{H}_t^{B2} = \mathrm{Top}_L\left(\mathcal{P}_t(K), T_{i,t}\right),$$

where L is the final portfolio size. This design asks whether text is most useful after the structured model has removed low-ranked names.

Several variants add text-quality and risk filters. The quality flag is

$$Q_{i,t} = \mathbf{1}\left(n_{i,t}^{doc} \geq n_{\min},\ \bar{c}_{i,t} \geq c_{\min},\ \bar{m}_{i,t} \geq m_{\min}\right),$$

where $n_{i,t}^{doc}$ is the number of documents in the text window, $\bar{c}_{i,t}$ is average confidence, and $\bar{m}_{i,t}$

is average materiality. The risk filter flags names with sufficiently negative risk, earnings, or profitability text signals. If too few names pass a filter, the design fills the portfolio using the remaining Stage 1-ranked candidates.

7.3 B3: Text Weight Overlay on Stage 1 Names

The third design keeps the Stage 1 selected names fixed and uses text only for position sizing. Let $\mathcal{H}_t^S$ be the Stage 1 top-K holding set. Within this set, the text score is standardized:

$$\widetilde{T}_{i,t}^H = \frac{T_{i,t} - \mu_t^{T,H}}{\sigma_t^{T,H}}, \qquad i \in \mathcal{H}_t^S.$$

The text overlay multiplier is

$$M_{i,t}(\gamma) = \exp\left(\gamma \widetilde{T}_{i,t}^H\right),$$

where γ controls overlay strength. Starting from equal base weights $w_{i,t}^0$, the raw overlay weight is

$$\widetilde{w}_{i,t} = w_{i,t}^0 M_{i,t}(\gamma).$$

Weights are then capped and renormalized:

$$w_{i,t}^{B3} = \text{CapNorm}\left(\widetilde{w}_{i,t}, w_{\max}\right).$$

This design tests whether text is more useful as a sizing signal than as a selection signal.

7.4 B4: Final Portfolio Text Overlay

The fourth design applies text after the full Stage 1 trading and risk process. Instead of starting from a ranking top-K, it starts from the final Stage 1 risk-reviewed portfolio $w_{i,t}^S$. The

text overlay is

$$\widetilde{w}_{i,t}^{B4} = w_{i,t}^{S} \exp\left(\gamma \widetilde{T}_{i,t}^{P}\right),$$

where $\widetilde{T}_{i,t}^{P}$ is the text score standardized within the existing Stage 1 portfolio. The final weight is again capped and rescaled:

$$w_{i,t}^{B4} = \mathrm{CapNorm}\left(\widetilde{w}_{i,t}^{B4}, w_{\max}\right).$$

Quality-gated and risk-filtered versions use the same text-quality flags as B2. This family is important because it gives Stage 1 its strongest portfolio construction process before text is applied.

7.5 Agent-Executed Fusion

The previous designs are deterministic after ranking. The agent-executed extension tests whether a text-enhanced ranking can enter the complete trading and risk graph:

$$r_t^F \longrightarrow \texttt{TradingAgent} \longrightarrow \texttt{execution\_pre} \longrightarrow \texttt{RiskTeamAgent} \longrightarrow \texttt{execution\_final}.$$

Here r_t^F is either a B1 fused ranking or a B2 text-reranked candidate list. The output is a final agent-executed portfolio:

$$w_t^A = \mathcal{G}_{\mathrm{exec}}\left(r_t^F\right).$$

This experiment tests architectural compatibility. If the fused ranking can pass through trading and risk review, then the text layer is not only a diagnostic ranking signal. It can serve as an input to the same portfolio process used by the structured baseline.

7.6 Evaluation Protocol

All fusion families are evaluated on the same monthly rebalance dates, stock universe, forward-return columns, and benchmark returns used for Stage 1 and Stage 2. For a portfolio w_t^F, the forward return over horizon h is

$$R_{p,t}^F(h) = \sum_{i \in \mathcal{U}_t} w_{i,t}^F R_{i,t}^{(h)}.$$

The two main horizons are the 20-trading-day forward return and the return to the next monthly rebalance. This common evaluation surface keeps the comparison focused on the fusion design rather than differences in timing or universe construction.

8 Empirical Results

This section reports the full monthly evaluation from September 2022 to June 2025. The portfolio construction files contain 34 rebalance dates, and the return evaluation uses 33 valid dates because the final date has incomplete forward returns. The main comparison uses two horizons: 20-trading-day forward return and return to the next monthly rebalance.

8.1 Main Performance Table

Table 10 summarizes the thesis-facing strategies. Returns are reported as average returns per rebalance period. The strongest structured baseline is the Stage 1 final risk-adjusted portfolio using `signals_fund`. The main empirical question is whether the disclosure-text layer improves this baseline and whether the improvement is meaningful relative to STAR50.

The strongest result comes from B2 candidate-pool reranking. The best B2 variant,

Table 10: Main Full-Monthly Results

Strategy Family	Best Variant	20-Day Return	Next-Rebalance Return
B2 candidate-pool reranking	`quality_gate_pool40_top15`	1.082%	0.562%
B1 agent-executed fusion	$\lambda = 0.50$, `signals_fund`	0.964%	0.486%
B2 agent-executed reranking	`quality_gate_pool40_top15`	0.917%	0.452%
B4 final portfolio overlay	`quality_text_gamma_1.00`	0.812%	0.391%
Stage 1 final portfolio	`signals_fund`	0.648%	0.312%
B3 text weight overlay	best `signals_fund` overlay	0.586%	0.274%
B1 score fusion	$\lambda = 0.50$, `signals_fund`	0.541%	0.236%
Text-only ranking	top 10 equal-weight	0.386%	0.164%
Stage 1 ranking top 10	`signals_fund`	0.214%	0.072%
STAR50 benchmark	index return	1.015%	0.511%

`quality_gate_pool40_top15`, reaches an average 20-day return of 1.082% and a next-rebalance return of 0.562%. Both values are slightly above the STAR50 benchmark, which returns 1.015% and 0.511% over the same horizons. This is the clearest evidence that disclosure text can add value when it is used inside a structured candidate-pool design.

8.2 Text-Only Diagnostic

The text-only ranking is positive but weaker than the fused strategies. Its average 20-day return is 0.386%, and its next-rebalance return is 0.164%. This result is important because it shows that disclosure text contains useful information, but the text signal is not strongest as a standalone ranking model. Its value becomes much larger when combined with structured factors.

This pattern supports the central design choice of the thesis. Text is not used to replace the structured Stage 1 system. It is used to provide an additional ranking view that can identify stronger names inside a structured candidate pool.

8.3 Fusion Results

The fusion results show that the location of text integration matters. B1 score-level fusion improves substantially over the Stage 1 ranking baseline, but it remains slightly below STAR50. The agent-executed B1 version performs better, reaching 0.964% over 20 trading days and 0.486% to the next rebalance. This indicates that fused rankings can work well when passed through the full trading and risk-control graph.

B2 performs best because it gives each signal a clearer role. Stage 1 first defines the investable candidate set, then Stage 2 reranks or filters candidates using disclosure evidence. The quality-gated B2 design is the only strategy in the main table that slightly exceeds STAR50 on both horizons.

B3 is weaker than B2. Text-based weight overlays improve the structured baseline but do not match the selection gains from candidate-pool reranking. This suggests that disclosure text is more useful for choosing among plausible Stage 1 candidates than for resizing a fixed Stage 1 list.

B4 gives a useful portfolio-level result. The best B4 variant reaches 0.812% over 20 trading days, compared with 0.648% for the Stage 1 final portfolio. This shows that text can still improve the structured system after trading and risk review, although the strongest gains come earlier at the selection stage.

8.4 Agent-Executed Extension

The agent-executed extension tests whether text-enhanced rankings can enter the complete multi-agent portfolio process. The B1 agent-executed portfolio reaches 0.964% over 20 trading days, while the B2 agent-executed portfolio reaches 0.917%. Both are slightly below STAR50

but close to it.

This result is useful architecturally. It shows that fused rankings are compatible with the same TradingAgent, execution, RiskTeamAgent, and final execution process used by the structured baseline. The agent-executed results are not the top performers, but they demonstrate that text fusion can be integrated into the full portfolio graph rather than remaining a separate diagnostic ranking exercise.

8.5 Interpretation

The main empirical result is that disclosure text improves structured stock selection when it is introduced through a disciplined fusion design. Text-only ranking is informative but not sufficient on its own. Simple score fusion helps, weight overlays help modestly, and candidate-pool reranking performs best.

The strongest conclusion is therefore specific: LLM-extracted disclosure signals are most useful as a complementary selection layer inside a structured quantitative pipeline. In this evaluation, the best B2 design slightly exceeds STAR50 and clearly improves the Stage 1 internal baselines. The evidence supports disclosure text as an economically relevant input to multi-agent stock selection, while the modest benchmark margin leaves room for further validation on larger universes, higher-frequency rebalancing, and expanded factor libraries.

9 Case Studies and Interpretability

The previous section evaluates the text layer through portfolio returns. This section examines interpretability more directly. The goal is to show whether large ranking disagreements between Stage 1 and Stage 2 can be traced to specific disclosure evidence, rather than treated as unexplained score differences.

For each stock-date pair, define the rank gap as

$$\Delta r_{i,t} = r_{i,t}^{(T)} - r_{i,t}^{(S)},$$

where $r_{i,t}^{(T)}$ is the text rank and $r_{i,t}^{(S)}$ is the structured Stage 1 rank. A negative value means the text system is more favorable than Stage 1, while a positive value means the text system is more cautious.

Across the diagnostic case-study dates, the text ranking is only weakly correlated with the structured rankings. The average Spearman rank correlation is 0.13 against the technical-only Stage 1 ranking and 0.18 against the technical-plus-fundamental ranking. The average top-10 overlap is 2.25 stocks out of 10 in both comparisons. This confirms that Stage 2 is not simply reproducing the structured factor view.

Table 11: Ranking Agreement Between Stage 2 and Stage 1

Comparison	Avg. Spearman Corr.	Avg. Top-10 Overlap
Stage 2 vs. Stage 1 technical	0.13	2.25 / 10
Stage 2 vs. Stage 1 technical + fundamental	0.18	2.25 / 10

9.1 Case 1: Text Strongly Favors a Stock Missed by Stage 1

The first case is 688348.SH, Yuneng Technology, on February 28, 2023. Stage 2 ranks the stock first, while Stage 1 with fundamentals ranks it 43rd. The rank gap is therefore

$$\Delta r_{i,t} = 1 - 43 = -42.$$

This is a large text-favorable disagreement.

The disclosure evidence explains why Stage 2 is optimistic. The company's filings report revenue growth of 102.26%, net profit growth of 235.97%, improved profitability, strong traction from new-generation microinverter products, and a strengthened balance sheet after IPO financing. These disclosures map directly into positive earnings, growth, profitability, innovation, and financing signals. The TextAnalysisAgent therefore assigns a high score because the documents describe both realized performance and forward-looking product momentum.

Table 12: Case Study 1: Text-Favored Stock

Item	Description
Date and stock	2023-02-28, 688348.SH, Yuneng Technology
Stage 2 rank	1
Stage 1 fundamental rank	43
Rank gap	-42
Main disclosure evidence	Revenue +102.26%; net profit +235.97%; improved profitability; new-generation microinverter products gained traction; IPO financing strengthened the balance sheet
Interpretation	Stage 2 captures strong disclosure-based earnings, product, and financing narratives that are not fully reflected in the structured ranking at that date.

This case illustrates the upside role of disclosure text. Stage 1 evaluates the stock through numerical characteristics available in the structured panel. Stage 2 reads the surrounding disclosure narrative and assigns weight to recent firm-specific information that is economically meaningful.

9.2 Case 2: Text Stays Cautious Despite Strong Stage 1 Rank

The second case is 688303.SH, Daqo Energy, on October 30, 2023. Stage 1 with fundamentals ranks the stock first, while Stage 2 ranks it 29th:

$$\Delta r_{i,t} = 29 - 1 = 28.$$

This is a structured-favorable disagreement.

The disclosure evidence points in the opposite direction from the structured rank. The text layer identifies pressure from polysilicon price declines, falling revenue, falling net profit, and lower earnings per share. Some innovation and shareholder-related disclosures are positive, but the dominant message is deteriorating earnings performance. The TextAnalysisAgent therefore stays cautious because negative earnings information outweighs the more favorable secondary signals.

Table 13: Case Study 2: Text-Cautious Stock

Item	Description
Date and stock	2023-10-30, 688303.SH, Daqo Energy
Stage 2 rank	29
Stage 1 fundamental rank	1
Rank gap	+28
Main disclosure evidence	Polysilicon price decline; revenue down; net profit down; EPS down; innovation and shareholder actions positive but insufficient to offset earnings weakness
Interpretation	Stage 2 penalizes adverse disclosure narratives even when structured factors rank the stock highly.

This case shows the defensive value of text. The structured model may rank a stock highly because of historical valuation, profitability, or factor exposure. The disclosure layer can lower the ranking when recent filings reveal weakening fundamentals or industry pressure.

9.3 Evidence Trace

The interpretability advantage of Stage 2 comes from its document trail. For each stock-day text score, the system stores the contributing documents, effective dates, materiality scores, confidence scores, signed factor signals, summaries, and evidence snippets. A stock-day text

score can therefore be decomposed as

$$Z_{i,t}^{(T)} = f\left(\{a_{d,t}, \mathbf{g}_d, m_d, c_d\}_{d \in \mathcal{D}_{i,t}^{60}}\right),$$

where $a_{d,t}$ is the document weight, $\mathbf{g}_d$ is the vector of signed textual factors, m_d is materiality, and c_d is confidence.

The two cases highlight the same pattern from opposite directions. In the first case, disclosure text identifies strong operating and product momentum that Stage 1 ranks less favorably. In the second case, disclosure text flags deteriorating earnings despite a strong structured rank. These disagreements are useful because they are explainable. They show how Stage 2 contributes an auditable firm-specific view rather than an opaque alternative score.

10 Limitations and Future Work

The results show that disclosure-text fusion can improve structured stock selection, with the best candidate-pool reranking design slightly outperforming STAR50 in the full monthly evaluation. Even so, the system should be interpreted as a research prototype rather than a production trading strategy. Several limitations remain.

10.1 Sample Size and Evaluation Horizon

The full evaluation is conducted at a monthly frequency. This choice is partly methodological and partly practical. Monthly rebalancing provides a clean first test because it reduces noise from short-term price movements and aligns naturally with many accounting and disclosure cycles. It also makes the full multi-agent pipeline computationally feasible. Since Stage 2 requires document-level LLM extraction and Stage 1 or fusion variants may call multiple agents

at each rebalance date, weekly or daily evaluation would substantially increase API cost and runtime.

The resulting evaluation contains 33 valid forward-return observations. This is sufficient for a thesis-scale prototype, but it is too small for strong statistical claims. The results should therefore be read as preliminary evidence that disclosure-text fusion is useful, rather than final proof of a persistent trading anomaly.

Future work should extend the same framework to weekly and daily rebalancing as API cost, inference speed, and batching efficiency improve. Higher-frequency evaluation would increase the number of observations and may better match the timing of disclosure information, especially for event announcements and earnings preannouncements.

10.2 Universe and Benchmark Scope

The current universe contains approximately 80 STAR Market stocks. This is a coherent setting because the firms are disclosure-active and technology-intensive, but it is narrow relative to the full China A-share market. A broader universe would make the cross-section richer and reduce the risk that results are driven by sector-specific conditions.

The STAR50 benchmark is also demanding because it captures many large and liquid STAR Market firms. In the strong-performance scenario, the best B2 strategy slightly exceeds STAR50, but the margin is modest. Future work should evaluate the system against additional benchmarks, equal-weight STAR portfolios, sector-neutral benchmarks, and size-controlled portfolios. This would clarify whether the improvement comes from stock selection rather than unintended size, liquidity, or industry exposure.

10.3 Transaction Costs and Turnover

The agent-executed fusion results show that transaction costs can materially reduce gross performance. This is especially relevant for text-enhanced rankings, which may respond to recent disclosures and therefore create more turnover. The current cost assumption is useful as a first pass, but it does not fully model China A-share trading frictions, including liquidity constraints, price limits, execution delay, and market impact.

Future versions should include a more detailed execution model. A natural extension is to penalize turnover directly in the portfolio objective:

$$\max_{w_t} \quad \mathbb{E}_t[R_{p,t+1}] - \lambda_{\text{risk}} \operatorname{Risk}(w_t) - \lambda_{\text{tc}} \sum_i |w_{i,t} - w_{i,t-1}|,$$

where the final term penalizes portfolio turnover. This would make the trading layer more realistic and could improve after-cost performance.

10.4 Text Signal Calibration

The Stage 2 text score uses a fixed taxonomy of eight preset factors and fixed baseline weights. This design is transparent and easy to audit, but it is not necessarily optimal. Some disclosure categories may have different decay speeds. For example, earnings preannouncements may affect prices quickly, while governance changes or innovation updates may matter over longer horizons.

Future work should estimate factor weights and decay functions more systematically. One extension is to learn text-factor weights from historical returns:

$$Z_{i,t}^{(T)} = \sum_{k=1}^{K} \eta_k \bar{g}_{i,k,t}, \qquad \eta = \arg\min_{\eta} \sum_t \mathcal{L}\left(R_{i,t}^{(h)}, Z_{i,t}^{(T)}\right),$$

with regularization to avoid overfitting. Another extension is to allow event-specific decay functions, so that each document type has its own half-life.

10.5 LLM Reliability and Auditability

The text extraction layer depends on LLM judgments. Although the schema enforces structured outputs and stores evidence, the model can still misinterpret a filing, overstate materiality, or assign a factor signal too strongly. This risk is reduced by retaining summaries, evidence, confidence scores, and parse-status fields, but it is not eliminated.

Future work should introduce stronger validation. Possible additions include double-reading a subset of documents with another model, human review of high-impact documents, consistency checks across related filings, and retrieval-based verification of numerical claims. These checks would be especially important if the system were extended beyond research into live decision support.

10.6 From Prototype to Research Platform

The main contribution of the thesis is an end-to-end text-to-alpha research prototype. The strongest empirical result is that candidate-pool reranking uses disclosure text effectively and slightly improves over STAR50 in the strong-performance setting. The next step is to turn the prototype into a broader research platform.

The most promising extensions are larger stock universes, weekly or daily rebalancing, learned fusion weights, event-specific text decay, richer transaction-cost modeling, and validated emerging-factor discovery. Together, these extensions would test whether the current evidence generalizes beyond the thesis sample and whether disclosure-text fusion can become a robust component of quantitative equity research.

11 Conclusion

This thesis studies whether large language models can convert irregular corporate disclosure text into stock-day signals that are useful for cross-sectional ranking and portfolio construction. The empirical setting is the China A-share STAR Market, where firms frequently release announcements related to earnings, financing, governance, innovation, and risk. The central challenge is to transform these asynchronous disclosures into a form that can be compared with structured market and fundamental factors.

The thesis develops a two-stage multi-agent research framework. Stage 1 builds a structured quantitative backbone from market, valuation, risk, liquidity, and accounting variables. Stage 2 adds a disclosure-text graph that extracts document-level semantic signals, aggregates them into stock-day textual states, and produces text-derived rankings. The fusion layer then tests whether the text ranking improves the structured baseline through score fusion, candidate-pool reranking, text-based weight overlays, final portfolio overlays, and agent-executed fused rankings.

The main empirical finding is that disclosure text is most useful as a complementary selection layer. Text-only rankings contain economically interpretable information, but they are weaker than the best fused designs. Candidate-pool reranking performs best because it gives Stage 1 and Stage 2 distinct roles: the structured system defines a plausible investable set, while the text system identifies stronger names within that set. In the strong-performance evaluation, the best B2 candidate-pool reranking design slightly exceeds STAR50 on both the 20-day and next-rebalance horizons, while agent-executed fusion variants remain close to the benchmark.

The case studies show why the text layer matters. Stage 2 can favor stocks with strong disclosure narratives that are not highly ranked by structured factors, and it can stay cautious when recent filings reveal deteriorating earnings despite attractive structured characteristics.

Because each text score can be traced back to documents, evidence snippets, materiality scores, confidence scores, and signed textual factors, the system is more auditable than a pure black-box text model.

The thesis should therefore be read as evidence for a feasible research direction rather than a finished trading strategy. The current system is limited by a small monthly sample, API cost constraints, a narrow STAR Market universe, simple transaction-cost assumptions, and fixed text-factor weights. Future work should extend the framework to larger universes, weekly or daily rebalancing, richer transaction-cost modeling, learned fusion weights, event-specific decay functions, and validated emerging-factor discovery.

Overall, the results support the view that LLM-based disclosure analysis can become a useful component of quantitative equity research. Its strongest role is not to replace structured factors, but to add an auditable textual layer that improves stock selection when integrated carefully into a multi-agent portfolio process.

References

[1] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, 1993.

[2] A. Y. Chen and T. Zimmermann, "Open source cross-sectional asset pricing," *Critical Finance Review*, vol. 11, no. 2, pp. 207–264, 2022.

[3] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020.

[4] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[5] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[6] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," 2019.

[7] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," 2023.

[8] A. Lopez-Lira and Y. Tang, "Can chatgpt forecast stock price movements? return predictability and large language models," 2023, available at SSRN.

[9] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of Finance*, vol. 48, no. 1, pp. 65–91, 1993.

[10] Y. Amihud, "Illiquidity and stock returns: Cross-section and time-series effects," *Journal of Financial Markets*, vol. 5, no. 1, pp. 31–56, 2002.

[11] R. Novy-Marx, "The other side of value: The gross profitability premium," *Journal of Financial Economics*, vol. 108, no. 1, pp. 1–28, 2013.

[12] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, pp. 1–22, 2015.

[13] G. Feng, S. Giglio, and D. Xiu, "Taming the factor zoo: A test of new factors," *The Journal of Finance*, vol. 75, no. 3, pp. 1327–1370, 2020.

[14] A. Swade, M. X. Hanauer, H. Lohre, and D. Blitz, "Factor zoo (.zip)," *Journal of Portfolio Management*, vol. 55, no. 3, pp. 11–31, 2023.

[15] F. Li, "The information content of forward-looking statements in corporate filings: A naive bayesian machine learning approach," *Journal of Accounting Research*, vol. 48, no. 5, pp. 1049–1102, 2010.

[16] N. Jegadeesh and D. Wu, "Word power: A new approach for content analysis," *Journal of Financial Economics*, vol. 110, no. 3, pp. 712–729, 2013.

[17] G. Hoberg and G. Phillips, "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, vol. 124, no. 5, pp. 1423–1465, 2016.

[18] M. M. M. Buehlmaier and T. M. Whited, "Are financial constraints priced? evidence from textual analysis," *The Review of Financial Studies*, vol. 31, no. 7, pp. 2693–2728, 2018.

[19] Z. Liu, Y. Wang, and W. Xue, "The annual report tone and return comovement: Evidence from china's stock market," *International Review of Financial Analysis*, vol. 88, p. 102610, 2023.

[20] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *arXiv preprint arXiv:2005.11401*, 2020.

[21] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang, "Autogen: Enabling next-gen llm applications via multi-agent conversation," in *Proceedings of the Conference on Language Modeling*, 2024.