Cogito: A Cognitive Agentic Framework Driven by Dynamic Graph of Thoughts for Financial Report Generation

Anonymous ACL submission

Abstract

Financial report generation is a complex task that requires gathering and reasoning over multi-source information. Recent advances in Large Language Models (LLMs) have made them a promising solution for automating this process. However, the reasoning paths in traditional Chain-of-Thought paradigms are inherently constrained by predefined, static computational topologies, rendering them ill-equipped to handle the dynamic uncertainties of real-world financial environments. To tackle this challenge, we propose **Cogito**, a cognitively grounded agentic framework for professional financial report generation. At its core, Cogito is driven by Dynamic Graph of Thoughts, a novel reasoning mechanism that models the agents reasoning process as an evolving topology for adaptive exploration. We further introduce a Social Collaboration Mechanism to facilitate coordinated agent interaction. Finally, Cogito is instantiated as a multi-agent system, where four specialized agents collaboratively execute the end-to-end report generation task. Extensive experiments on enterprise- and industry-level financial report generation benchmarks demonstrate the superiority of Cogito in data quality, analytical validity, and presentation quality.

1 Introduction

Financial Report Generation is a knowledge-intensive and labor-intensive complex task. Experts not only perform multi-turn, iterative retrieval in open and noisy information environments, but also leverage deep domain expertise to reconstruct fragmented market data into logically coherent professional narratives. Manual execution is impractical due to its prohibitive cost, and existing automated approaches remain insufficient. Fine-tuning-based methods, such as BloombergGPT (Wu et al., 2023), inject static knowledge but struggle to adapt to real-time market dynamics. Multi-agent frameworks, such as FinDebate (Cai et al., 2025) and TradingAgents (Xiao et al., 2025), enhance professionalism through role-playing, but are constrained by predefined static workflows and can only conduct research on preset topics. Meanwhile, general methods such as Deep Research (OpenAI, 2025b; Gemini, 2025; Perplexity, 2025), though excelling at retrieval, lack critical domain expertise. Consequently, constructing an agent system that simultaneously exhibits domain depth and dynamic adaptability to handle the high uncertainty and open-ended exploration requirements of financial scenarios, remains a critical research challenge.

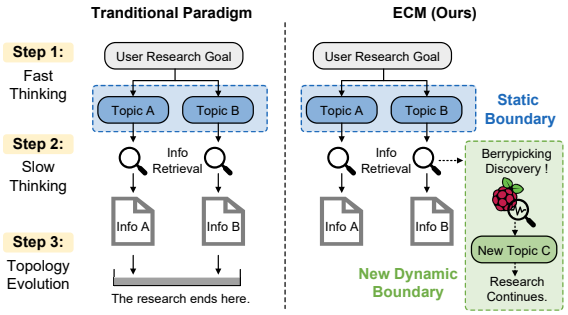

Figure 1: Cogito's evolutionary reasoning process.

To address this issue, we turn our attention to the reasoning mechanisms of Large Language Models (LLMs). Beginning with "Let's think step by step," Wei et al.'s (2022) Chain-of-Thought (CoT) has significantly unlocked the potential of LLMs by generating intermediate reasoning steps. Subsequent approaches like Tree of Thoughts (ToT) (Yao et al., 2024; Xie et al., 2023) introduced heuristic search and backtracking mechanisms, while Graph of Thoughts (GoT) (Besta et al., 2024) further broke tree-like constraints, supporting complex logical construction through arbitrary aggregation and looping of thought nodes. However, these mainstream architectures are constrained by predefined static topologies, which limit their ability to handle "unknown unknowns."

As Descartes stated, "*Cogito, ergo sum*." Humans can establish cognitive anchors (*Sum*) in research situations full of uncertainty, precisely because they possess the ability to think dynamically and evolutionarily (*Cogito*). This cognitive flexibility is well supported by established theories in cognitive science, which together characterize human reasoning as dynamic, adaptive, and socially coordinated. Specifically, **Dual Process Theory (DPT)** (Krämer, 2014; Evans and Stanovich, 2013) reveals how humans flexibly switch between "intuitive planning (System 1 / Fast)" and "deep research (System 2 / Slow)." **Bates' Berrypicking Model** (Bates, 1989) explains how human exploration paths evolve with new clues, while **Transactive Memory Systems (TMS)** (Wegner, 1987) allow humans not only to know "what" but also to perceive "who needs to know what and who is responsible for what" within a team. This naturally raises the question of whether Agentic AI can be endowed with similar dynamic thinking capabilities, evolving from a static executor into an active subject of cognitive learning.

Inspired by the aforementioned theories, we first introduce an **Evolutionary Cognitive Mechanism (ECM)** to overcome the limitations of traditional static reasoning. Grounded in DPT and Bates Berrypicking Model, the ECM employs a three-step cognitive process to capture unexpected discoveries and drive the continuous evolution of the reasoning topology, as shown in Figure 1. To operationalize this evolving cognitive process, we design **Dynamic Graph of Thoughts (Dy-GoT)** as its execution engine. Dy-GoT models reasoning as a dynamically evolving graph, supporting runtime node generation and reflective backtracking. In addition, to support effective collaboration in complex research workflows, we incorporate a **Social Collaboration Mechanism (SCM)** based on TMS, which constrains the interaction flow through expert role injection and collaboration perception. Built upon the ECM, Dy-GoT, and SCM, we propose **`Cogito`**, a professional research report generation framework designed for open-ended and uncertain financial scenarios. Within `Cogito`, four specialized agents collaborate to execute the end-to-end workflow: the *ResearcherAgent* orchestrates evolutionary reasoning, the *ToolAgent* and *DIKAgent* compress heterogeneous retrieved data into high-value knowledge, and the *WriterAgent* synthesizes the final report.

The main contributions of this paper are as follows:

- We introduce Dy-GoT, a novel dynamic reasoning mechanism that enables runtime evolution of reasoning topologies, addressing the limitations of static architectures.
- We propose `Cogito`, a unified agentic framework driven by Dy-GoT for adaptive and collaborative financial research in open-ended settings.
- Extensive experiments on enterprise- and industry-level financial report generation benchmarks demonstrate that `Cogito` consistently outperforms state-of-the-art open- and closed-source baselines across multiple evaluation criteria. Ablation study further validate the effectiveness of each component.

2 Methodology

This section begins with the task formalization (§2.1), followed by the Evolutionary Cognitive Mechanism (§2.2) designed for adaptive planning. To operationalize this logic, Dy-GoT (§2.3) is introduced as the underlying state-driven engine. Finally, the Social Collaboration Mechanism (§2.4) is detailed, describing how specialized agents are coordinated within this dynamic environment.

2.1 Problem Formulation

Financial report generation task aims to generate a comprehensive, logically structured, and factually accurate long-form research report that incorporates external knowledge. Formally, given a user query q representing the research intent, the model interacts with an open-domain environment $\mathcal{E}$ to iteratively acquire relevant data. We define the target report R as a hierarchical sequence of sections $R = \{s_1, s_2, \ldots, s_N\}$. The objective is to generate the optimal report by maximizing the conditional probability:

$$R^* = \arg\max_R P(R \mid q, \mathcal{E}) \tag{1}$$

where the generation of each section s_i depends on the query q, the data retrieved from $\mathcal{E}$, and the coherence with preceding sections.

2.2 Evolutionary Cognitive Mechanism

Traditional methods (Wang et al., 2023; Assafelovic and ElishaKay, 2025; Wan et al., 2025; Qiao et al., 2025) define the research boundaries

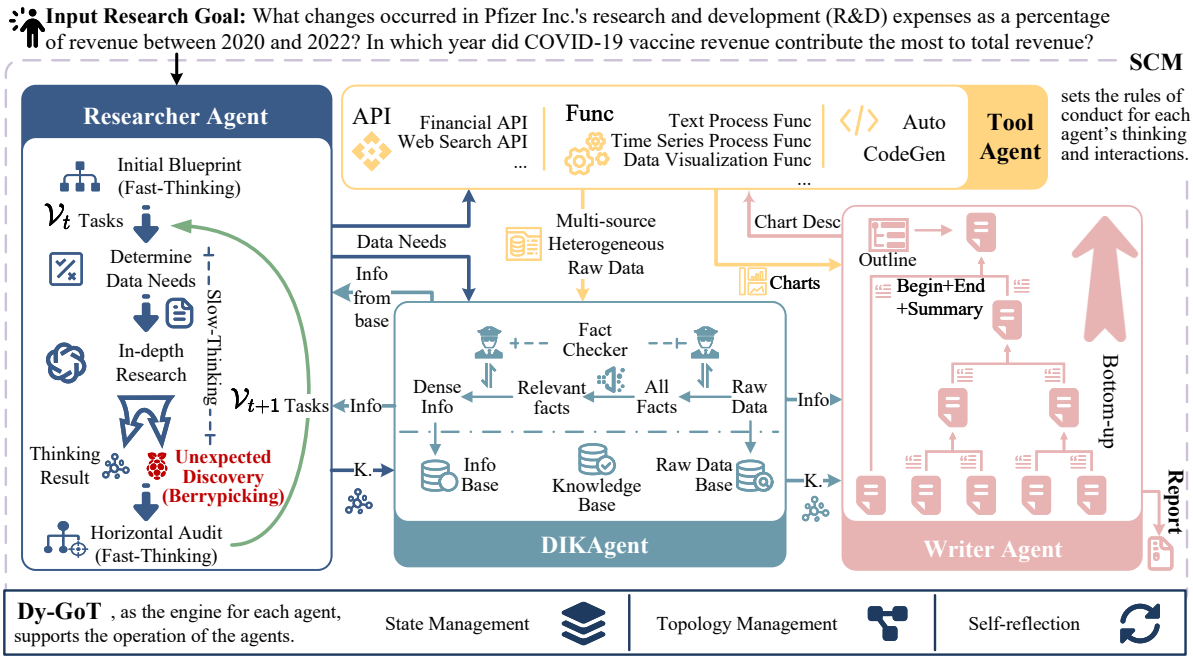

Figure 2: **Overview of the Cogito Framework.** The system is driven by the **Dy-GoT** engine, executing a structured workflow via four agents. The process is initiated and orchestrated by the **ResearcherAgent**, which performs evolutionary reasoning. It dispatches tasks: the **ToolAgent** invokes external tools, and the **DIKAgent** is responsible for knowledge curation. This knowledge is synthesized by the **WriterAgent** into the final report. **Arrow** colors denote data output sources, and all inter-agent interactions are governed by the **SCM**.

through problem decomposition, and then perform retrieval operations to supplement the details. This paradigm relies on a strong implicit assumption that all necessary information in the solution space is "known information." This is a static research paradigm where the boundaries of the research are defined early on when the subproblems are breakdown; consequently unexpected findings during the search do not lead to new research plans. In contrast, in real-time and complex financial research tasks, critical information is often hidden within the "unknown unknowns" due to the temporal limitations of the LLM's internal knowledge. Such pivotal directions only upon after executing a search step and obtaining intermediate feedback.

To capture this dynamic value, Cogito introduces an Evolutionary Cognitive Mechanism (ECM) that combines DPT and Bates' Berrypicking model. After utilizing fast thinking to complete the blueprint planning, the system initiates slow thinking by entering a batch-based vertical research mode. At time step t, a batch of active nodes $\mathcal{V}_t$ is executed in parallel. The resulting deep reasoning outputs $\mathcal{O}_t$ and unexpected discoveries $\mathcal{D}_t$ are fed collectively into a horizontal auditing mechanism. According to the berrypicking model, the research path gradually becomes clearer through multiple retrievals and reflections. Therefore, the auditing function $\mathcal{A}$ uses fast thinking to evaluate the current cognitive state and dynamically decides whether to construct the next batch of research tasks (e.g., fine-tuning the research direction or delving deeper) or to terminate the research:

$$G_{t+1} = \begin{cases} \Omega(G_t, \Gamma(\mathcal{D}_t, \mathcal{O}_t)), & \text{if } \mathcal{A}(\mathcal{V}_t) \to \text{Cont.} \\ G_t, & \text{if } \mathcal{A}(\mathcal{V}_t) \to \text{Stop} \end{cases} \quad (2)$$

Here, Γ represents the Task Generation Function that outputs a new research plan, while Ω denotes the Evolution Operator, which is responsible for topologically integrating the generated plan into the current graph G_t. This mechanism ensures that Cogito does not merely append tasks, but restructures its cognitive trajectory based on intermediate findings. As shown in Figure 1, Cogito simulates the adaptive flow of human information retrieval; every act of "berrypicking" provides feedback that reshapes the subsequent search space, allowing the agent to transition from initial uncertainty to structured expertise. The detailed procedure of this cognitive architecture is formally summarized in Algorithm 1.

2.3 Dynamic Graph of Thoughts

To support the physical implementation of ECM, we design the Dynamic Graph of Thoughts (Dy-

GoT) as the underlying reasoning engine. Unlike standard GoT approaches, which primarily seek an optimal path within a predefined static graph. Dy-GoT models cognition as a Finite State Machine. Formally, we define the reasoning graph as a time-variant tuple $G_t = (V_t, E_t, \Phi_t)$, where V_t represents the set of cognitive units (thoughts or actions) at time t, and $\Phi_t : V_t \to \mathcal{S}$ maps each node to its lifecycle state $\mathcal{S} \in \{\mathsf{Ready}, \mathsf{Finished}, \mathsf{Reflecting}, \dots\}$. State transitions follow the working logic of each Agent. This state-driven mechanism provides a unified protocol for diverse complex workflows, enabling diverse reasoning logics to be abstracted into state transitions and dependency management on a graph.

To facilitate dynamic structure construction, we define a set of evolution operators $\Omega = \{\mathsf{AddNode}, \mathsf{AddEdge}, \dots\}$. When the agent triggers an operator based on intermediate reasoning results (e.g., unexpected discoveries or new subtask requirements) during execution, the system performs a graph update $G_{t+1} = \Omega(G_t)$. The engine then immediately recalculates the readiness of affected subgraphs via the state update function $\mathcal{U}(G_{t+1})$. This achieves automated reasoning orchestration without human intervention, constituting the universal execution engine shared by all agents within the `Cogito` framework.

2.4 Social Collaboration Mechanism

In financial report generation, complex tasks such as processing multi-source heterogeneous data and authoring long-form reports far exceed the capability boundaries of a single agent. Therefore, we construct a collaborative team following a multi-agent system paradigm and design a Social Collaboration Mechanism (SCM) based on TMS. It consists of two parts: an Expert Role Injection Mechanism (ERIM) and a Team Collaboration Perception Mechanism (TCPM), to improve collaboration efficiency and quality.

To ensure professional depth, the ERIM first generates specific professional identities based on the solution space. These identities are injected into the base prompt ρ_{base} alongside domain context C_{dom} via $\rho^+ = \mathsf{ERIM}(\rho_{base}, C_{dom})$, forcing the agent to load domain-specific thinking paradigms. To improve interaction efficiency, when an upstream agent (e.g., ResearcherAgent) calls a downstream agent, it uses TCPM to explicitly perceive the downstream agent's profile $\mathcal{M} = \{\mathrm{id}_{dst}, \mathrm{Need}_{dst}, \mathrm{Act}_{dst}\}$. Consequently, the upstream agent's instruction generation policy π is constrained by this perception:

$$I_c = \pi(\mathcal{M}, \rho^+) \tag{3}$$

This mechanism ensures that inter-agent communication is constrained at the source. For instance, knowing the downstream agent is a ToolAgent responsible for fact-checking (perceived via $\mathcal{M}$), the ResearcherAgent will automatically filter out subjective commentary-oriented search intents. The downstream agent then efficiently executes tasks based on these precise instructions.

3 The `Cogito` Framework

Supported by the methodology proposed in the previous section, the `Cogito` Framework (Figure 2) decouples the complex report generation task and delegates it to four specialized agents with distinct functions for collaborative execution.

ResearcherAgent serves as the cognitive hub and orchestrator. Driven by the ECM, it decomposes macro-goals into executable blueprints and conducts in-depth analysis. Crucially, through a continuous Horizontal Audit, it evaluates the strategic value of intermediate findings to dynamically determine whether to spawn new exploration branches or deepen existing ones, thereby driving the adaptive evolution of the research topology.

ToolAgent functions as the system's perceptual interface and execution unit, bridging the closed-world limitations of the LLM. It encapsulates a diverse toolkit, ranging from APIs to code generators, to handle data retrieval, preprocessing, and visualization. Operating under SCM constraints, it transcends passive execution by autonomously selecting tools aligned with the specific intent instructions received from upstream agents.

DIKAgent is responsible for knowledge curation, equipped with a fact checker, and manages the system's data flow and persistent memory. It receives Raw Data acquired by the ToolAgent and executes a rigorous "Extraction-Verification-Fusion" process. Through fact extraction and self-reflective verification, it eliminates irrelevant and false facts, distilling this data into high-value information. The existence of the DIKAgent ensures that subsequent reasoning processes are based on verified, high-density facts rather than massive raw

noise, mitigating the interference of low-quality context on cognition.

WriterAgent is the final presenter of the research results, focusing on organizing knowledge (thought results) into a logically rigorous and coherent narrative. Upon receiving inputs from the ResearcherAgent, it executes a dependency-based recursive bottom-up writing strategy. Adhering to the outline's topological structure, the agent prioritizes synthesizing leaf-node micro-sections by aggregating relevant knowledge within each scope. As sub-node content (text, tables, charts) solidifies, it progressively aggregates upwards to formulate macro-arguments. This ensures that every high-level conclusion in the final report is strictly supported by granular, traceable evidence.

4 Experiments

We conduct extensive experiments to answer the following four core Research Questions (RQs):

RQ 1: How does `Cogito` perform against SOTA baselines in the financial report generation task?

RQ 2: How does the ECM compare to static planning in adapting to "unknown unknowns"?

RQ 3: Compared to standard retrieval methods, what exactly do DIK strategy change, and how do they affect reasoning quality?

RQ 4: How does SCM affect professionalism and interaction efficiency in the multi-agent team?

4.1 Experimental Settings

Dataset. Following Jin et al.'s (2025), we construct a dataset of 20 research targets from authoritative financial platforms. The set comprises 10 enterprise-level and 10 industry-level tasks to cover diverse research granularities (Appendix B).

Evaluation Metrics. We evaluate report quality using a quantitative framework (Total Score: 100) across Data, Analytical, and Presentation dimensions. Additionally, we report objective linguistic metrics including Fog Index, FK Grade, Max Dependency Depth, and Distinct-3 to quantify writing sophistication and complexity (Appendix C). To mitigate randomness, each evaluation is run 5 times, and the average value is reported.

Baselines. We compare `Cogito` against three distinct categories of baselines: (1) LLMs with Search Tools. Representing the direct integration of state-of-the-art general-purpose LLMs with retrieval tools, including DeepSeek-V3.2 (DeepSeek-AI et al., 2025), GPT-5.1 (OpenAI, 2025a), and Claude Sonnet 4.5 (Anthropic, 2025). (2) Open-Source Agentic Frameworks. We select frameworks representing the highest standard for "Deep Research" tasks within the open-source community, including GPT-Researcher (Assafelovic and ElishaKay, 2025), OpenManus (Liang et al., 2025), and DeerFlow (Walnut and Li, 2025). (3) Closed-Source Commercial Deep Research Systems. This category represents the current State-of-the-Art (SOTA) in proprietary deep research capabilities, including Gemini DeepResearch (Gemini, 2025), GPT Deep Research (OpenAI, 2025b), Grok DeepSearch (Grok, 2025), and Perplexity Research (Perplexity, 2025) (Appendix D).

Implementation Details. We employ Gemini 2.5 Flash and Pro (Comanici et al., 2025) as backbone models, integrated with financial (e.g., Finnhub[1]) and general retrieval APIs [2] (Appendix E). We use Gemini 3 Pro as the evaluator. A single research run takes 12+ minutes, increasing with the complexity of the research until it reaches the maximum limit. To facilitate reproducibility, the code and example reports are available.[3]

4.2 Main Results

To answer RQ1, we compare `Cogito` against 10 baselines across three distinct categories at two granularities of research: Enterprise-level and Industry-level. The main experimental results are presented in Table 1.

Overall, `Cogito` achieves SOTA performance in both granularities surpassing all baselines by a significant margin. Specifically, `Cogito` attains total scores of 95.2 and 96.7 at the Enterprise and Industry levels, respectively. This performance not only substantially outperforms the strongest open-source baseline, GPT-Researcher (78.4 and 79.2), but also exceeds the closed-source commercial SOTA, Gemini DeepResearch (89.9 and 90.1). These results strongly validate the effectiveness of the Dy-GoT-based agentic framework in handling long-context, multi-source heterogeneous financial report generation tasks, demonstrating its superior capability in managing complex financial information flows. Notably, `Cogito` exhibits a substantial advantage in Presentation Quality, a key

[1] `https://finnhub.io/`
[2] `https://tavily.com/`
[3] They are provided in the accompanying software archive.

Domain	Enterprise-Level									Industry-Level								
Model	Data		Analysis			Pres.			Total	Data		Analysis			Pres.			Total
	Rich.	Auth.	Integ.	Insig.	Logic	Multi.	Prof.	Expr.		Rich.	Auth.	Integ.	Insig.	Logic	Multi.	Prof.	Expr.	
LLMs with Search tools																		
DeepSeek-V3.2	8.0	13.5	11.6	6.6	7.7	3.9	8.3	6.4	66.0	6.3	8.6	9.5	5.6	6.7	4.6	7.3	5.0	53.6
GPT-5.1	12.5	15.6	11.7	6.4	8.0	5.0	8.0	5.6	72.8	10.6	13.4	11.3	6.5	7.8	**4.9**	7.8	5.3	67.6
Claude 4.5	9.6	15.7	13.7	8.3	8.8	3.4	8.9	5.8	74.2	10.5	15.8	14.3	9.1	9.3	3.6	9.6	8.0	80.2
Open Source Agents																		
GPT-Researcher	13.4	16.6	13.4	8.3	8.4	3.4	8.8	6.1	78.4	12.7	16.4	13.5	8.2	8.8	3.3	9.1	7.2	79.2
OpenManus	3.6	7.2	7.5	3.5	5.9	4.6	6.2	2.4	40.9	3.9	6.5	7.8	4.4	6.2	4.3	6.5	2.8	42.4
Deer-Flow	12.4	16.1	13.1	7.5	8.5	3.1	8.5	9.8	79.0	11.9	15.4	12.5	7.4	8.4	3.3	8.8	7.6	75.3
Closed Source Agents																		
Gemini DeepRes.	**14.6**	**18.7**	**14.9**	**9.9**	**9.7**	**5.0**	**9.9**	7.2	89.9	14.2	**19.0**	**15.0**	**9.9**	**9.7**	**4.9**	**10.0**	7.4	90.1
GPT DeepRes.	12.9	15.6	11.9	6.6	7.4	5.0	7.9	5.2	72.5	13.3	15.9	12.8	6.2	7.0	4.5	8.2	6.5	74.4
Grok DeepSearch	10.7	15.0	11.8	6.6	7.8	4.6	8.1	5.6	70.2	10.8	14.0	11.7	6.7	7.3	4.3	8.0	6.4	69.2
Perplexity	13.5	16.8	13.2	8.1	8.9	5.0	8.8	11.0	85.3	14.1	17.1	13.3	8.1	8.7	4.7	9.1	10.3	85.8
Cogito (Ours)	**14.6**	**18.7**	**14.9**	9.3	9.4	**5.0**	9.8	**13.5**	**95.2**	**14.7**	18.8	14.8	9.6	**9.7**	**4.9**	9.9	**14.3**	**96.7**

Table 1: Main results on the Enterprise and Industry benchmarks. The best and second-best results are highlighted in **bold** and underlined, respectively.

Metric	Gemini (SOTA)	Cogito (Ours)
Professionalism		
Fog Index (↑)	14.75	**18.39**
FK Grade (↑)	12.81	**14.82**
Complexity		
Max Dep. Depth (↑)	15.50	**16.67**
Distinct-3 (↑)	0.834	**0.839**

Table 2: Linguistic feature analysis.

driver of its overall lead. While existing models score poorly on Expressiveness due to their restriction to plain text generation, Cogito leverages a built-in visualization module to bridge this gap. Instead of solely presenting text or tables, Cogito automatically generates informative, professional-grade charts grounded in narrative logic, offering intuitive support for complex arguments.

We further conduct a detailed analysis based on specific evaluation dimensions. In terms of the breadth and authority of information retrieval, Cogito performs on par with the industrial leader, Gemini DeepResearch, indicating SOTA-level capability in financial fact capture. Notably, while Gemini holds a slight advantage in qualitative metrics (e.g., Insightfulness and Structural Logic), a cross-reference with objective linguistic metrics reveals that this discrepancy reflects divergent stylistic orientations. As shown in Table 2, the text generated by Cogito exhibits a Flesch-Kincaid Grade Level (FK-Grade) of 14.8 (corresponding to a professional level), which is significantly higher than Gemini's 12.8 (undergraduate level). Furthermore, Cogito demonstrates a higher Max Syntactic Dependency Depth (16.7 vs. 15.5). This suggests that while current SOTA models tend to generate "fluent and accessible" content to align with the preferences of LLM judges, Cogito prioritizes the structural rigor and syntactic sophistication essential for professional financial reporting, reflecting a strategic trade-off between casual readability and professional density that is well-supported by objective metrics. Additionally, Cogito's high scores in Professionalism (9.8/9.9) correlate with the aforementioned high FK-Grade, further corroborating its ability to precisely leverage industry terminology and avoid the colloquial tendencies often observed in general-purpose models.

4.3 Ablation Study

4.3.1 Analysis in Reasoning Capability

To address RQ2, we conduct a comparative ablation study on the full evaluation dataset by configuring Cogito's reasoning engine into two ablated variants: CoT (V1) and ToT (V2). As the analysis of the reasoning chain structure extracted from the research results shows (Figure 3a and 3b), Dy-GoT (V3) demonstrates a significant dual advantage in "Breadth-Depth" coverage, overcoming the shallow linear constraints of V1 and the limited horizontal breadth of V2. Notably, high-order reasoning chains (Depth $\geq$ 3) account for 46% of the total in V3. This provides strong empirical validation for DPT, and specifically confirms how the integration of "Fast and Slow" thinking facilitates both broad and deep reasoning. Furthermore, we conduct a semantic analysis of the reasoning chains, visualized using t-SNE dimensionality reduction on nomic-embed-text embeddings. Using the "Apple AI Strategy" task as a representative case (Figure 3c), the results reveal that V3 not only precisely covers the "known unknowns" (Inner Ring) clustered around the research core

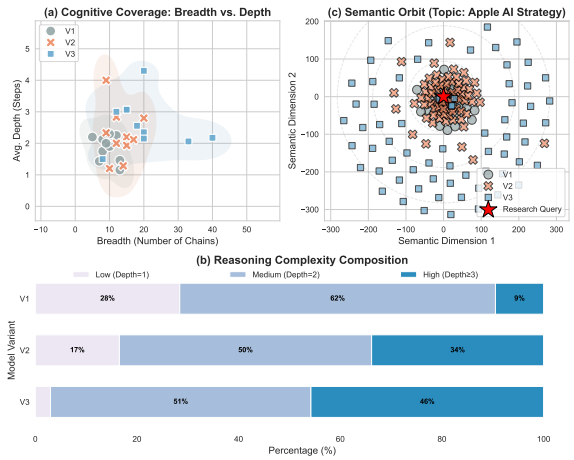

Figure 3: Results of the reasoning capability analysis. (a) The cognitive coverage scatter plot with KDE illustrates the trade-off between reasoning breadth and average depth. (b) A stacked bar chart showing the distribution of reasoning depths. (c) Visualization of the semantic space of reasoning chains.

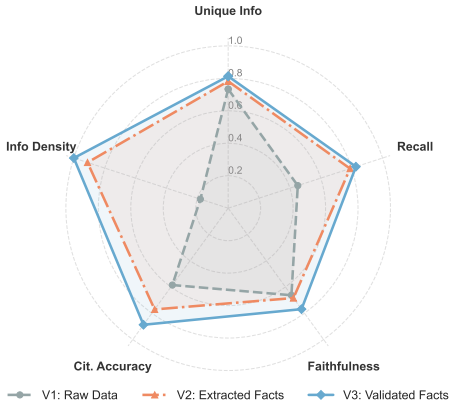

Figure 4: Results of the impact analysis of the DIK.

but also successfully extends its exploratory reach into the semantically sparse "unknown unknowns" (Outer Ring) under the guidance of the Berrypicking Model's ECM. This capability to dynamically reconstruct search paths in response to acquired information enables `Cogito` to transcend predefined retrieval boundaries, significantly elevating both robustness and the upper bound of exploration in highly uncertain environments.

4.3.2 Analysis in DIK strategy

To address RQ3, we focus on the DIKAgent to deconstruct the independent contributions of "Information Distillation" and "Fact Checking" to reasoning quality via the ablation study. The experimental data is derived from the intermediate processes recorded during report generation. As visualized in Figure 4, the distillation from Raw Data (V1) to Fact Extraction (V2) significantly mitigates the "low-density noise" inherent in RAG

Backbone Model	w/o TMS	w/ TMS
Qwen3-32B (Yang et al., 2025)	6.73	**7.27** (**0.54**↑)
GPT-OSS-20B (Agarwal et al., 2025)	6.98	**7.30** (0.32↑)
Llama3.3-70B (Grattafiori et al., 2024)	7.11	**7.31** (0.20↑)
GPT-OSS-120B (Agarwal et al., 2025)	7.52	**7.62** (0.10↑)

Table 3: Impact of TCPM on feasibility and granularity of retrieval commands.

Variant	NPL Metrics				LLM Eval
	Distinct-2	FK-Grade	MTLD	Max-Dep	Avg. Score
w/o PRI	0.569	14.94	96.68	15.1	4.7
w/ PRI	**0.583**	**15.10**	**106.56**	**15.7**	**4.8**

Table 4: Impact of ERIM on Generation Quality.

systems. This is evidenced by a dramatic expansion along the info density axis, where Information Density surges from 0.180 to 0.911, proving that the distillation mechanism effectively filters corpus noise. However, subsequent ablation of the verification mechanism reveals that this compression comes at a cost. Relying solely on extraction (V2) induces compressive hallucinations, resulting in a noticeable retraction in the Faithfulness dimension, which drops to 0.684. Consequently, the subsequent incorporation of Fact Checking (V3) drives a robust rebound in Faithfulness to 0.769, while pushing Citation Accuracy to its peak at 0.888. This comparison provides compelling evidence that the fact-checking mechanism serves as a critical safeguard against hallucinations in long-text generation. Such a recovery in fidelity is particularly vital in financial contexts, as it ensures that the distilled, high-density insights remain strictly grounded in verifiable evidence, thereby mitigating the risk of downstream misjudgment in professional auditing. By working in synergy with the distillation mechanism, this strategy enables the system to ensure high fidelity in both logic and evidence while maintaining extreme information density.

4.3.3 Analysis in SCM

To address RQ4, we isolate the two components of SCM, namely ERIM and TCPM, and evaluate each independently using data derived from the intermediate processes. As shown in Table 3, the TCPM significantly optimizes the quality of tool-oriented instruction construction by the ResearcherAgent; this improvement demonstrates high robustness across different backbone models and exhibits a notable "performance compensation effect" for models with smaller parameters, effectively reducing the system's reliance on complex prompt engineering. Subsequent ablation experiments on ERIM (Table 4) reveal that removing this mechanism leads to a sharp 10.2% decline in Lex-

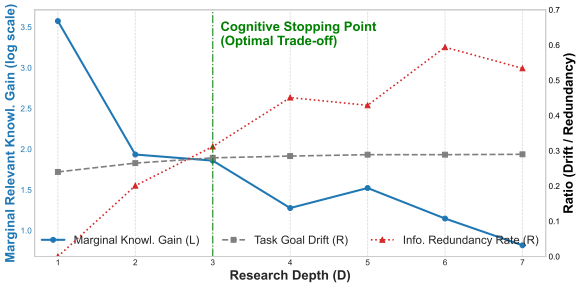

Figure 5: Cognitive Equilibrium Analysis

ical Diversity (MTLD), a concurrent decrease in the FK-Grade reading level ($0.16 \downarrow$), and a regression in the domain-specific depth score achieved by the LLM ($4.8 \rightarrow 4.7$). These findings strongly validate that the ERIM successfully activates latent domain-expert representations, thereby significantly enhancing stylistic professionalism while maintaining logical precision, achieving a qualitative leap from "generic exposition" to "professional narrative".

4.4 Analysis in Depth-Utility Equilibrium

To determine the optimal Cognitive Stopping Point, we examine the trade-off between Marginal Relevant Knowledge Gain and Information Redundancy Rate. As visualized in Figure 5, the system maximizes the acquisition of unique mission-critical entities at a depth of 3. Beyond this threshold, the influx of new insights diminishes significantly; meanwhile, redundancy rises due to the accumulation of repetitive facts. This confirms $d = 3$ as the cost-effective equilibrium. These results validate that although the Horizontal Audit effectively prevents semantic drift, a predefined depth limit remains vital to prevent excessive computational overhead on saturated information.

5 Related Work

Financial Agents. Early financial LLMs such as BloombergGPT (Wu et al., 2023) and FinGPT (Liu et al., 2023) primarily focused on constructing domain-general models through pretraining or supervised fine-tuning on financial corpora. However, the high latency of parameter updates has shifted research toward RAG and Multi-Agent Systems. For instance, frameworks like FinRobot (Yang et al., 2024), FinDebate (Cai et al., 2025), and TradingAgents (Xiao et al., 2025) introduced role-playing mechanisms, enhancing decision-making professionalism by simulating debates between analysts and traders. Nevertheless, they lack the autonomous planning capabilities required for open-ended financial problems, rendering them insufficient for generating high-fidelity reports that require adaptive long-horizon reasoning.

Deep Research Systems. Recent "Deep Research" systems address long-horizon information aggregation. Commercial solutions (OpenAI, 2025b; Gemini, 2025; Perplexity, 2025; Grok, 2025) excel in general domains but suffer from closed ecosystems and a lack of domain-specific depth or multimodal synthesis. Conversely, open-source initiatives (Assafelovic and ElishaKay, 2025; Liang et al., 2025; Walnut and Li, 2025) attempt to replicate these capabilities but often struggle with information noise and hallucination in long contexts due to simplistic underlying reasoning logic. A gap remains for an open-source framework that combines deep reasoning with the rigor required by financial standards.

Thought Reasoning. To handle complex tasks, reasoning topologies have evolved from linear CoT (Wei et al., 2022) to tree-based (ToT) (Yao et al., 2024) and graph-based (GoT) (Besta et al., 2024) structures. Many agent frameworks (Shao et al., 2024; Li et al., 2025; Wan et al., 2025) have also begun to combine fast and slow thinking to improve overall inference capabilities, facilitate multi-perspective reasoning, and mitigate illusions through retrieval. However, their workflows are based solely on predefined static topologies and cannot handle emerging "unknown unknowns" encountered in real-world financial research.

6 Conclusion

We propose `Cogito`, a cognitively grounded agentic framework for open-ended financial report generation. By integrating an Evolutionary Cognitive Mechanism with Dynamic Graph of Thoughts and a Social Collaboration Mechanism, `Cogito` enables adaptive reasoning, evolutionary exploration, and coordinated multi-agent collaboration in complex financial scenarios. Through extensive benchmark evaluations and fine-grained ablation analyses, we show that `Cogito` surpasses strong open- and closed-source baselines across multiple evaluation criteria for professional financial research generation. In future work, we will employ ERIM to extend the applicability of `Cogito` to open-ended reporting tasks across diverse domains.

Limitations

The complexity of the evolutionary cognitive architecture and the rigorous fact-verification process incur significant token consumption, limiting the feasibility of evaluation on larger-scale datasets. Additionally, constraints on retrieval API quality and rate limits hinder the potential for deeper evolution in certain research inquiries. Consequently, this study focuses on validating the framework's effectiveness under current conditions. Future work will focus on optimizing resource allocation and enhancing the efficiency of the evolutionary architecture.

Potential Risks

While `Cogito` is designed to assist professionals in producing high-quality financial research, potential risks remain regarding reliability and misuse. Despite the integration of the DIKAgent for rigorous fact-checking, the inherent probabilistic nature of LLMs means that subtle hallucinations or omission of critical risk factors cannot be entirely eliminated, which could lead to financial losses if used as the sole basis for investment decisions. Furthermore, there is a dual-use risk where the system's capability to generate authoritative-sounding reports could be exploited by malicious actors to mass-produce fabricated narratives for market manipulation. Therefore, we emphasize that `Cogito` is intended as an augmented intelligence tool to support, not replace, human analysts, and a "human-in-the-loop" verification mechanism remains essential for final decision-making.

Ethics Statement

This work adheres to the ACL Code of Ethics. In this study, no human subjects or animal experimentation was involved. The constructed dataset complies with relevant usage guidelines, ensuring that no privacy is violated. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, and 106 others. 2025. gpt-oss-120b & gpt-oss-20b model card. Preprint, arXiv:2508.10925.

Anthropic. 2025. Claude sonnet 4.5. `https://www.anthropic.com/claude/sonnet`.

Assafelovic and ElishaKay. 2025. Gpt-researcher: An llm agent that conducts deep research (local and web) on any given topic and generates a long report with citations. [Online; accessed 2025-12-27].

Marcia J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. Online Review, 13(5):407–424.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. Proceedings of the AAAI Conference on Artificial Intelligence, 38(16):17682–17690.

Tianshi Cai, Guanxu Li, Nijia Han, Ce Huang, Zimu Wang, Changyu Zeng, Yuqi Wang, Jingshi Zhou, Haiyang Zhang, Qi Chen, Yushan Pan, Shuihua Wang, and Wei Wang. 2025. FinDebate: Multi-agent collaborative intelligence for financial analysis. In Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing, pages 268–282, Suzhou, China. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint, arXiv:2507.06261.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. Preprint, arXiv:2512.02556.

Jonathan St. B. T. Evans and Keith E. Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science, 8(3):223–241. PMID: 26172965.

Gemini. 2025. Gemini deep research. https://gemini.google/overview/deep-research/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

Grok. 2025. Grok ai deepsearch: Real-time research power guide. https://grokaimodel.com/deepsearch/.

Jiajie Jin, Yuyao Zhang, Yimeng Xu, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. 2025. Finsight: Towards real-world financial deep research. Preprint, arXiv:2510.16844.

Walter Krämer. 2014. Kahneman, d. (2011): Thinking, fast and slow. Statistical Papers, 55.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025. Webthinker: Empowering large reasoning models with deep research capability. Preprint, arXiv:2504.21776.

Xinbin Liang, Jinyu Xiang, Zhaoyang Yu, Jiayi Zhang, Sirui Hong, Sheng Fan, and Xiao Tang. 2025. Openmanus: An open-source framework for building general ai agents.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. Preprint, arXiv:2307.10485.

OpenAI. 2025a. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/.

OpenAI. 2025b. Introducing deep research. https://openai.com/index/introducing-deep-research/.

Perplexity. 2025. Introducing perplexity deep research. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research.

Shuofei Qiao, Zhisong Qiu, Baochang Ren, Xiaobin Wang, Xiangyuan Ru, Ningyu Zhang, Xiang Chen, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2025. Agentic knowledgeable self-awareness. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12601–12625, Vienna, Austria. Association for Computational Linguistics.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Walnut and Henry Li. 2025. Deerflow is a community-driven deep research framework, combining language models with tools like web search, crawling, and python execution, while contributing back to the open-source community. [Online; accessed 2025-12-27].

Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. 2025. A cognitive writing perspective for constrained long-form text generation. In Findings of the Association for Computational Linguistics: ACL 2025, pages 9832–9844, Vienna, Austria. Association for Computational Linguistics.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Daniel M. Wegner. 1987. Transactive Memory: A Contemporary Analysis of the Group Mind, pages 185–208. Springer New York, New York, NY.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. Preprint, arXiv:2303.17564.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2025. Tradingagents: Multi-agents LLM financial trading framework. In The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. 2023. Self-evaluation guided beam search for reasoning. In Advances in Neural Information Processing Systems, volume 36, pages 41618–41650. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388.

Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and Christina Dan Wang. 2024. Finrobot: An open-source ai agent platform for financial applications using large language models. CoRR, abs/2405.14767.

Yao Yao, Zuchao Li, and Hai Zhao. 2024. Beyond chain-of-thought, effective graph-of-thought reasoning in language models. Preprint, arXiv:2305.16582.

A Process of ECM

Algorithm 1: Process of ECM

Input : Query q, Environment $\mathcal{E}$, Max Steps T_{max}
Output : Final Report R^*

$G_0 \leftarrow \text{InitBlueprint}(q)$;
$\mathcal{V}_0 \leftarrow \text{GetReadyNodes}(G_0)$;
$t \leftarrow 0$;
while $\mathcal{V}_t \neq \emptyset$ **and** $t < T_{max}$ **do**
  `// 1. Parallel Execution Phase`
  $\mathcal{O}_t, \mathcal{D}_t \leftarrow \emptyset, \emptyset$;
  foreach $v \in \mathcal{V}_t$ **in parallel do**
    $o_v, d_v \leftarrow \texttt{Execute}(v, \mathcal{E})$;
    $\mathcal{O}_t \leftarrow \mathcal{O}_t \cup \{o_v\}, \mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{d_v\}$;
  `// 2. Horizontal Audit Phase`
  $signal \leftarrow \texttt{Audit}(\mathcal{O}_t, \mathcal{D}_t, G_t)$;
  if $signal = \text{Stop}$ **then**
    break;
  `// 3. Dynamic Evolution Phase`
  $\Delta_{plan} \leftarrow \Gamma(\mathcal{D}_t, \mathcal{O}_t)$ `// Task Generation`
  $G_{t+1} \leftarrow \Omega(G_t, \Delta_{plan})$ `// Topology Evol`
  $\mathcal{U}(G_{t+1})$ `// State Update`
  $\mathcal{V}_{t+1} \leftarrow \text{GetReadyNodes}(G_{t+1})$;
  $t \leftarrow t + 1$;
$R^* \leftarrow \text{SynthesizeReport}(G_t)$;
return R^*;

B Dataset Details

To rigorously evaluate the capability of agents in generating professional financial reports, we construct a dataset covering 20 distinct research topics, divided equally into enterprise-level and industry-level tasks. The selection of these targets is guided by three criteria: high market attention, data availability for the 2015-2024 period, and the complexity of the required reasoning (e.g., linking qualitative strategies to quantitative financial outcomes). The prompts encompass both English and Chinese queries to test the system's cross-lingual retrieval and generation capabilities.

Data Construction and Selection Criteria. The research targets are identified through a rigorous screening of top-tier financial news aggregators (e.g., Bloomberg[4], Reuters[5]) and equity research platforms. We adhere to three core selection criteria: (1) Information Density & Availability: We select entities from major indices (e.g., S&P 500[6], NASDAQ-100[7]) to ensure sufficient public disclosures, earnings transcripts, and news coverage for retrieval. (2) Reasoning Complexity: Simple data retrieval tasks (e.g., "What is Apple's revenue?") are excluded. Instead, we design "multi-hop" reasoning tasks that require linking strategic decisions to financial outcomes (e.g., "Impact of Tesla's price cuts on Gross Margin" or "Meta's efficiency strategy vs. Net Profit"). (3) Sector Diversity: The dataset spans Technology, Healthcare, Consumer Discretionary, Aerospace, Finance, and Energy, testing the agent's generalization ability across different business models.

Task Granularity. The dataset is structured into two granularities to test different cognitive capabilities. The 10 Enterprise-level tasks focus on deep-dive analysis of single entities, requiring the agent to perform longitudinal comparisons (e.g., Starbucks Pre- vs. Post-COVID) and causal inference (e.g., Apple's AI strategy impact). The 10 Industry-level tasks require horizontal comparative analysis, testing the agent's ability to synthesize information across multiple competitors (e.g., The Cloud Computing Capex war among AWS, Azure, and Google) and assess macroeconomic impacts (e.g., High-interest rates on the Solar or Banking sectors).

C Evaluation Metrics Details

C.1 Subjective Evaluation (LLM-as-a-Judge)

To ensure a fair and comprehensive assessment of the generated financial reports, we employ an advanced LLM (Gemini3 Pro) as the evaluator. The model is instructed to score each report on a scale of 0 to 100 based on three core dimensions:

[4] https://www.bloombergchina.com/
[5] https://reutersagency.com/
[6] https://www.marketwatch.com/investing/index/spx
[7] https://indexes.nasdaqomx.com/Index/Overview/NDX

Prompt 1: Evaluation Prompt for Financial Reports

The research reports above are generated by Agents using the request {research goal}. Please score the reports from the following 3 dimensions (using integers only) and output the result in the specified table format.

Data Metrics (35): Measure whether the data sources used in the final report are rich and authoritative.

- Richness (15): Are the data sources in the report rich?
- Authority (20): Are the data sources in the report authoritative and professional?

Analytical Effectiveness (35): Measure whether the analysis results in the report provide sufficient information and insights to the user.

- Entity Completeness (15): The completeness of entities relevant to the research requirements included in the report (company names, person names, specific events, core data metrics, etc.).
- Insightfulness (10): Does the report provide critical analysis, original insights, and forward-looking suggestions?
- Structural Logic (10): Does the report expression follow a certain narrative logic and possess readability?

Presentation Quality (30): Measure the presentation quality of the final report.

- Multilingual Adaptability (5): Does the language used in the report comply with the language used or required by the user's question?
- Language Professionalism (10): Does the language conform to the professional terminology of the analysis domain?
- Expressiveness of Presentation Form (15): Does the report contain multimodal expression methods such as charts, and do they support the arguments in terms of narrative logic, including informativeness and aesthetics?

Output Format:

	Data (35)	Analysis (35)	Presentation (30)	Total (100)
Report A	Score + Explanation	...	...	...
Report B	...	...	...	...
...	...	...	...	...

Data Metrics, Analytical Validity, and Presentation Quality. The exact prompt used for evaluation is presented in Prompt 1.

C.2 Objective Linguistic Metrics

In addition to semantic evaluation, we utilize five objective metrics to quantify the linguistic complexity, structural depth, and lexical richness of the generated text. All metrics were computed using Python 3.8+ environment.

Flesch-Kincaid Grade Level (FK-Grade). This metric indicates the U.S. school grade level required to understand the text. We implemented the calculation using standard syllabification rules provided by the `textstat` library. The score is calculated as:

$$0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59 \tag{4}$$

A higher FK-Grade indicates a more complex text, which is often expected in professional financial reporting.

Gunning Fog Index (Fog Index). Similar to FK-Grade, the Fog Index estimates the years of formal education needed to understand the text on the first reading. It specifically penalizes long sentences and complex words (words with three or more syllables):

$$0.4 \left[\left(\frac{\text{total words}}{\text{total sentences}} \right) + 100 \left(\frac{\text{complex words}}{\text{total words}} \right) \right] \tag{5}$$

Financial reports typically aim for a higher Fog Index compared to casual conversation, reflecting professional depth.

Max Syntactic Dependency Depth. This metric measures the syntactic complexity of sentences by calculating the maximum depth of the dependency tree (distance from the root to the furthest leaf node). We utilized the `spacy` library (v3.8.11) with the `en_core_web_sm` model for dependency parsing. Unlike surface-level length metrics, dependency depth captures the extent of nested clauses and logical subordination. A greater depth implies a more structurally complex sentence structure, characteristic of formal analytical writing.

Measure of Textual Lexical Diversity (MTLD). To robustly evaluate the richness of vocabulary independent of text length, we employ the Measure of Textual Lexical Diversity (MTLD). Unlike simple Type-Token Ratio (TTR), which decreases as text length increases, MTLD calculates the average length of word strings that maintain a TTR above a threshold (default 0.72). We used the `lexical-diversity` Python package (v0.1.1). The input text was tokenized using the library's

built-in word-level tokenizer (`ld.tokenize`) to ensure accurate word boundary detection, excluding punctuation and sub-word tokens. A higher MTLD score indicates a broader range of terminology and reduced lexical repetition.

Distinct-3 Score. To further evaluate phrase-level diversity, we calculate the Distinct-3 score. It is the ratio of unique trigrams (sequences of three consecutive words) to the total number of trigrams in the generated text:

$$\text{Distinct-3} = \frac{\text{Count(unique trigrams)}}{\text{Count(total trigrams)}} \quad (6)$$

We implemented this using the `nltk` (v3.9.2) library's N-gram generation utilities. A higher Distinct-3 score indicates a richer vocabulary and more varied sentence construction.

D Baselines Details

We compare `Cogito` against three distinct categories of baselines to ensure a comprehensive evaluation:

(1) LLMs with Search Tools. This category represents the direct integration of state-of-the-art general-purpose LLMs with retrieval capabilities.

- **DeepSeek-V3.2 w/ Search:** The latest iteration of the DeepSeek-V3 series, featuring "Sparse Attention" architecture. We utilize its official API with the online parameter enabled for real-time information retrieval.
- **GPT-5.1 w/ Search:** OpenAI's newest flagship model featuring "Instant" and "Thinking" modes. We evaluate the model's native browsing capability via the Assistant API to answer complex research queries.
- **Claude Sonnet 4.5 w/ Search:** Anthropic's most advanced reasoning model, known for its superior coding and agentic capabilities. We connect it with a standard Google Search tool to evaluate its synthesis performance.

(2) Open-Source Agentic Frameworks. We selected frameworks representing the highest standard for "Deep Research" tasks within the open-source community.

- **GPT-Researcher:** A widely adopted autonomous agent that streamlines the research process by generating research questions, triggering parallel crawlers, and aggregating information into a final report.
- **OpenManus:** An open-source reproduction of the commercial Manus AI, developed by the MetaGPT community. It features a planner-executor architecture designed for long-horizon task handling and tool invocation.
- **DeerFlow:** ByteDance's community-driven Deep Research framework. It orchestrates specialized agents (Planner, Searcher, Coder) to conduct end-to-end research, supporting multi-modal output and complex reasoning chains.

(3) Closed-Source Commercial Deep Research Systems. This category represents the current State-of-the-Art (SOTA) in proprietary deep research capabilities.

- **Gemini DeepResearch:** Google's specialized agent powered by Gemini 3 Pro. It features multi-turn planning, deep navigation into sub-pages, and integration with Google's extensive index for high-fidelity financial data sourcing.
- **GPT Deep Research:** OpenAI's autonomous research agent built on the GPT 5.1 reasoning models. It performs iterative search, analyzes multiple sources simultaneously, and synthesizes findings into a document with granular citations.
- **Grok DeepSearch:** xAI's real-time research assistant leveraging the Grok-4 model. It is optimized for integrating social sentiment (via X data) with traditional financial news for rapid market signal analysis.
- **Perplexity Research:** A commercial research engine that combines iterative search logic with the Sonar Pro model. It focuses on reducing hallucinations through strict citation grounding and multi-step query refinement.

E API Configuration Details

To support the multi-dimensional data requirements of `Cogito`, we integrated a suite of specialized APIs. These interfaces provide the agent with necessary structured financial data and unstructured market information. We detail the specific utility of each API below:

Finnhub. We use Finnhub as our primary data source to retrieve insider trading records, institutional ownership changes, and company ESG scores, among other things. Its concise JSON response format allows our system to quickly parse high-frequency trading signals.

Alpha Vantage. This API[8] is employed to construct long-term historical time series, which is crucial for the agent to calculate accurate technical indicators (e.g., Moving Averages, RSI) over the 20232024 analysis period.

Financial Modeling Prep (FMP). Although not explicitly listed in the main text, FMP[9] acts as the backbone for fundamental analysis. We query FMP for standardized financial statements (Income Statement, Balance Sheet, Cash Flow) and key valuation ratios (PE, PB, Debt-to-Equity). Its ability to provide quarterly (10-Q) and annual (10-K) filings data ensures the agent bases its fundamental reasoning on audited figures.

Tavily Search API. Unlike the structured financial APIs above, Tavily is used as a search engine optimized for LLMs. It handles open-ended queries (e.g., "Starbucks restructuring plan details 2024") by aggregating content from news outlets, earning call transcripts, and industry blogs.

We confirm that all data acquisition and usage strictly adhere to the Terms of Service of the respective API providers listed above. The data retrieved is utilized solely for non-commercial academic research purposes. We do not redistribute the raw proprietary datasets; only the derived analysis and generated reports are presented as research artifacts.

F Additional Evaluation Prompts

In our ablation studies and fine-grained analysis, we employed specific prompts to evaluate the reasoning depth, professional quality, and entity coverage of the generated reports. The specific prompts used for these metrics are detailed below.

Logic Chain Extraction. To quantify the depth of reasoning, we tasked the LLM with extracting explicit causal chains from the reports using prompt 2. We filter out simple factual statements to focus on analytical depth.

Professional Depth Scoring. We assessed the domain expertise of the reports using a tiered scoring rubric that differentiates between superficial terminology usage and data-backed, forward-looking analysis. (Prompt 3)

Entity Coverage and Importance Analysis. To evaluate the recall of key information, we extracted Named Entities (NER) and asked the evaluator to judge their critical relevance to the specific research mission using prompt 4.

[8] https://www.alphavantage.co/
[9] https://site.financialmodelingprep.com/

Prompt 2: Logic Chain Extraction Prompt

The above research results were generated by the {variation}-Agent based on the research goal "{research goal}". Please extract all logical reasoning chains related to the research goal and output them in a table format.
A logical chain consists of two parts: Evidence and n-level Inferences:

- Evidence (Premise): xxx
- Inference x (Potential inference, where x is a number from 1 to n)

Please interpret the logical chains as exhaustively as possible; do not omit any. Note: Logical chains with the same premise but different inferences are not considered the same chain. All components of the logical chain must be strictly extracted directly from the document's research results; do not fabricate any part.
Ignore simple factual statements (e.g., "Apple's revenue is 100 billion USD").
Focus on analysis (e.g., "Revenue declined -> due to supply chain issues -> which were caused by...").
I need to use your output structure to evaluate the Agent's research capabilities, so please ensure you output all compliant logical chains. If the output cannot be completed in one response, please continue in the next response.
Table Format:
Serial No.	Logic Chain	Logic Chain Depth
:—	:—	:—
1	Evidence -> Inference 1 -> ...	

Prompt 3: Professional Depth Scoring

Both reports above are research reports addressing the question "{research goal}". Please score the **professional domain depth** of each report (1-5, integers only).
Scoring Criteria:

- **1-2 points:** Professional terminology usage.
- **2-4 points:** Professional terminology + Data support (Evidence).
- **5 points:** Professional terminology + Data support + Forward-looking insights (Note: Empty talk about the future does not count as forward-looking).

Prompt 4: NER and Importance Prompt

The above content is the research result generated by Agents based on the Project Mission: "{research goal}". Please perform the following operations:
1. Extract all Named Entities (NER) from the reports (Company names, person names, specific events, core data metrics, etc.).
2. Judge whether the entity is critical to answering the Project Mission? (Yes/No).
Finally, output the results in a table format:
Serial No.	Entity Name	Is Critical (Yes/No)
:—	:—	:—
1		
2		