

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

Chen Lifan^{1*}, Wei Ding^{1*}, Jingwen Yang¹, Xiuze Zhou²,
Jingan Chen¹, Fan Lin^{1†}

¹Xiamen University

²The Hong Kong University of Science and Technology (Guangzhou)

{chenlifan, dingwei}@stu.xmu.edu.cn, iamafan@xmu.edu.cn

Abstract

Financial report generation is a complex task that requires gathering and reasoning over multi-source information. Recent advances in Large Language Models 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 agent’s 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 prohibitively costly, while existing automated approaches remain insufficient. Fine-tuning-based methods, such as BloombergGPT

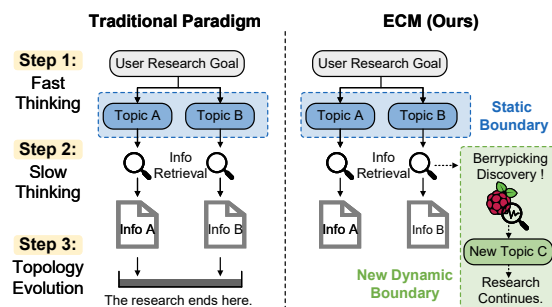


Figure 1: Cogito’s evolutionary reasoning process.

(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., 2024), 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, 2025c; Google, 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.

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 has significantly unlocked the potential of LLMs by generating intermediate reasoning steps. Subsequent approaches like Tree of Thoughts (Yao et al., 2023; Xie et al., 2023) introduced heuristic search and backtracking mechanisms, while Graph of Thoughts (Besta et al., 2024) further broke tree-like constraints, supporting complex logical construction through arbitrary aggregation and looping of thought nodes. However, these

* These authors contributed equally.

† Corresponding author.

foundational architectures inherently rely on predefined static topologies. Even recent dynamic variants (Ning et al., 2024; Zhang et al., 2024, 2025) restrict their adaptability to internal logic correction or search pruning within a closed scope. To handle “unknown unknowns” in complex financial scenarios, reasoning must move beyond optimizing predefined plans. Instead, it requires information-driven topology mutation, which dynamically spawns new exploratory branches in response to unexpected external discoveries.

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, 2013; 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 System (TMS)** (Wegner, 1987) allows 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 within complex research workflows, we incorporate a **Social Collaboration Mechanism (SCM)** based on TMS, which constrains interaction flows through expert role injection and col-

laboration 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.

Our contributions 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 demonstrates that *Cogito* consistently outperforms state-of-the-art (SOTA) open- and closed-source baselines across multiple evaluation criteria. Ablation experiments 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 $R = \{s_1, s_2, \dots, s_N\}$. The objective is to generate the optimal report by maximizing the condi-

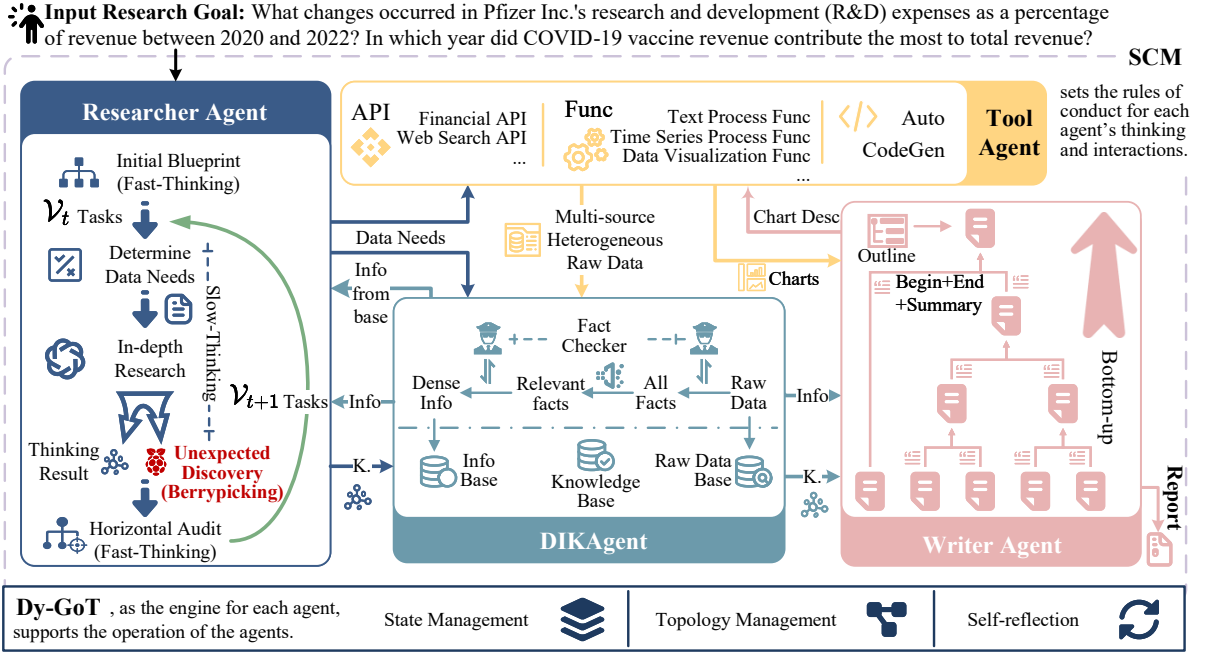


Figure 2: **Overview of the Cogito Framework.** The system is driven by the **Dy-GoT** engine, executing a structured workflow via four agents. The process is initiated and orchestrated by the **ResearcherAgent**, which performs evolutionary reasoning. It dispatches tasks: the **ToolAgent** invokes external 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**.

tional probability:

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

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; Asafelovic and ElishaKay, 2025; Wan et al., 2025; Qiao et al., 2025) define the research boundaries 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 broken down; 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 emerge 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:

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

Here, Γ is the Task Generation Function that outputs a new research plan, while Ω denotes the Evolution Operator, 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 Figure 1 shows, Cogi to 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 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 $\Phi_t : V_t \rightarrow \mathcal{S}$ maps each node to its lifecycle state $s \in \{\text{Ready}, \text{Finished}, \text{Reflecting}, \dots\}$. 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 $\Omega = \{\text{AddNode}, \text{AddEdge}, \dots\}$. When the agent triggers an operator based on intermediate reasoning results (e.g., unexpected discoveries or new sub-task 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 Cogi to 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 collabora-

tion 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 $\rho^+ = \text{ERIM}(\rho_{base}, C_{dom})$, 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 $\mathcal{M} = \{\text{id}_{dst}, \text{Need}_{dst}, \text{Act}_{dst}\}$. Consequently, the upstream agent’s instruction generation policy π is constrained by this perception:

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

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 Cogi to Framework

Supported by the methodology proposed in the previous section, the Cogi to 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 handles 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 spurious facts, distilling this data into high-value information. DIKAgent ensures that subsequent reasoning processes are based on verified, high-density facts not 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. (2025), we construct a dataset of 20 research targets from authoritative financial platforms. The dataset comprises 10 enterprise-level and 10 industry-level tasks to cover diverse research granularities (Appendix B). Given the exceptionally high resource demands of autonomous, multi-source Deep Research, our dataset scale is enough for this specific, open-ended domain.

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, Flesch-Kincaid 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 SOTA general-purpose LLMs with retrieval tools, including DeepSeek-V3.2 (DeepSeek-AI, 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 SOTA in proprietary deep research capabilities, including Gemini DeepResearch (Google, 2025), GPT Deep Research (OpenAI, 2025c), Grok DeepSearch (Grok, 2025), and Perplexity Research (Perplexity, 2025). Please refer to Appendix D for more details.

Implementation Details. We employ Gemini 2.5 Flash and Pro (Gemini Team, 2025) as backbone models, integrated with financial (e.g., Finnhub¹) and general retrieval APIs² (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.

4.2 Main Results

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

Overall, Cogito achieves SOTA performance, attaining total scores of 95.2 and 96.7 at the enterprise- and industry-level respectively. This substantially outperforms the strongest open-source baseline, GPT-Researcher, and exceeds the

¹<https://finnhub.io/>

²<https://tavily.com/>

Domain	Enterprise-Level									Industry-Level									
	Data			Analysis			Pres.			Total	Data			Analysis			Pres.		
Model	Rich.	Auth.	Integ.	Insig.	Logic	Multi.	Prof.	Expr.		Rich.	Auth.	Integ.	Insig.	Logic	Multi.	Prof.	Expr.		
<i>LLMs with Search tools</i>																			
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	
<i>Open Source Agents</i>																			
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	
<i>Closed Source Agents</i>																			
Gemini DeepRes.	14.6	18.7	14.9	9.9	9.7	5.0	9.9	7.2	<u>89.9</u>	<u>14.2</u>	19.0	15.0	9.9	9.7	4.9	10.0	7.4	<u>90.1</u>	
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 DeepRes.	13.5	16.8	13.2	8.1	8.9	5.0	8.8	<u>11.0</u>	85.3	14.1	17.1	13.3	8.1	8.7	4.7	9.1	<u>10.3</u>	85.8	
Cogito (Ours)	14.6	18.7	14.9	<u>9.3</u>	<u>9.4</u>	5.0	<u>9.8</u>	13.5	95.2	14.7	<u>18.8</u>	<u>14.8</u>	<u>9.6</u>	9.7	4.9	<u>9.9</u>	14.3	96.7	

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

Model	Enterprise-Level (Vertical Deep-Dive)				Industry-Level (Horizontal Breadth)			
	Avg. Depth \uparrow	D 1-2 (%) \downarrow	D 3 (%) \uparrow	D \geq 4 (%) \uparrow	Avg. Depth \uparrow	D 1-2 (%) \downarrow	D 3 (%) \uparrow	D \geq 4 (%) \uparrow
GPT DeepRes.	2.13	70.7	25.2	4.1	2.37	66.1	29.6	4.3
Grok DeepSearch	2.05	82.9	17.1	0.0	2.32	64.0	31.0	5.0
Perplexity Res.	1.95	87.6	12.4	0.0	2.07	73.2	25.9	0.9
Gemini DeepRes.	<u>2.72</u>	34.7	57.6	<u>7.6</u>	2.57	47.0	47.8	5.2
Cogito (Ours)	2.79	<u>37.1</u>	<u>46.6</u>	16.4	<u>2.38</u>	<u>61.6</u>	<u>36.0</u>	<u>2.4</u>

Table 2: Quantitative evaluation of Logic Chain Extraction. We categorize the chains into shallow (Depth 1-2), medium (Depth 3), and deep (Depth \geq 4) reasoning.

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

Table 3: Linguistic feature analysis.

closed-source commercial SOTA, Gemini DeepResearch. These results validate the effectiveness of the Dy-GoT engine in handling long-context and multi-source heterogeneous financial report generation tasks. Additionally, Cogito leverages a built-in visualization module to generate professional charts, which provides a notable advantage in presentation quality over models restricted to plain text.

Beyond overall superiority, the analysis of specific dimensions reveals nuanced differences. In terms of data retrieval capability, Cogito performs on par with Gemini DeepResearch in Data Richness and Authority. These metrics are relatively ro-

bust under LLM-based assessments because they primarily reflect relatively objective evidence coverage and source type identification. However, Gemini holds a marginal advantage in subjective analytical metrics such as Insightfulness. To assess whether this difference stems from a genuine reasoning deficiency or a style bias in LLM-as-judge evaluation, we conduct a logical-chain extraction task (Table 2) to cross-validate results via objective quantification of multi-hop exploration depth. The results demonstrate that in Enterprise tasks requiring deep vertical analysis, Cogito exhibits substantial superiority. Specifically, 16.4% of its extracted logic chains reach high-order complexities with a depth of 4 or more, which is more than double the 7.6% achieved by Gemini. This confirms that our evolutionary cognitive mechanism effectively drives deeper exploration. Conversely, in Industry tasks requiring broad horizontal comparisons, Cogito shows bounded depth. We attribute this to an infrastructure gap rather than an algorithmic flaw. Gemini benefits from a proprietary search engine that natively provides su-

perior retrieval coverage and data density, whereas Cogito relies on public search APIs that inherently suffer from lower recall ceilings (He et al., 2024; Coelho et al., 2025).

Finally, we complement the subjective evaluation with objective linguistic metrics to better understand these discrepancies. As shown in Table 3, the text generated by Cogito exhibits a Flesch-Kincaid Grade Level of 14.8, which is considerably higher than Gemini’s 12.8. Furthermore, Cogito demonstrates a higher Max Syntactic Dependency Depth at 16.67 compared to 15.50 for Gemini. 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. This strategic trade-off between casual readability and professional density is well supported by these objective linguistic features. Additionally, Cogito’s high scores in Professionalism (9.8 and 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 of 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 ≥ 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

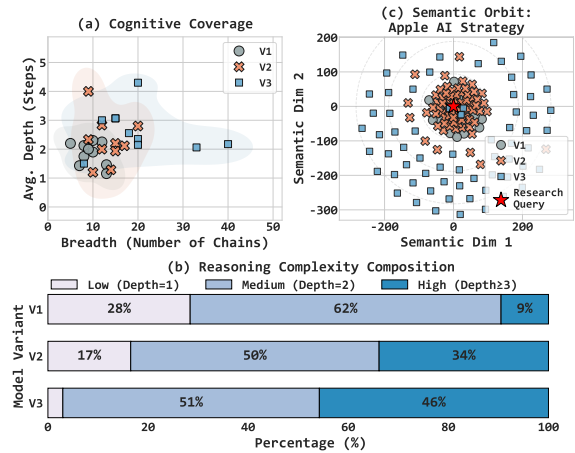


Figure 3: Reasoning analysis. (a) Trade-off between reasoning breadth and depth. (b) Reasoning depth distribution. (c) Semantic space of reasoning chains.

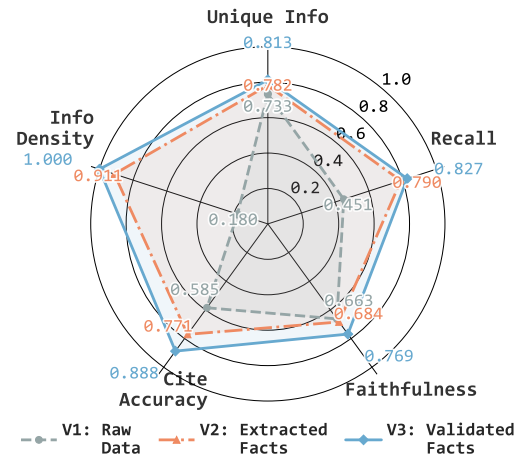


Figure 4: Results of the impact analysis of the DIK.

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, considerably 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) considerably mitigates the “low-density noise” inherent in RAG

Backbone Model	w/o TMS	w/ TMS
Qwen3-32B (Qwen Team, 2025)	6.73	7.27 (+0.54)
GPT-OSS-20B (OpenAI, 2025b)	6.98	7.30 (+0.32)
Llama3.3-70B (Llama Team, 2024)	7.11	7.31 (+0.20)
GPT-OSS-120B (OpenAI, 2025b)	7.52	7.62 (+0.10)

Table 4: 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 5: 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 very high information density.

4.3.3 Analysis of 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 4, the TCPM considerably 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 fewer parameters,

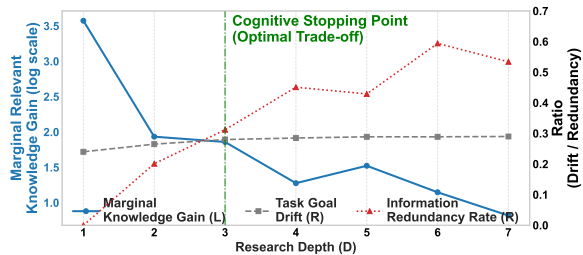


Figure 5: Cognitive Equilibrium Analysis

effectively reducing the system’s reliance on complex prompt engineering. Subsequent ablation experiments on ERIM (Table 5) reveal that removing this mechanism leads to a sharp 10.2% decline in Lexical Diversity (MTLD), a concurrent decrease in the FK-Grade reading level (0.16 ↓), and a regression in the domain-specific depth score achieved by the LLM (4.8 → 4.7). These findings strongly validate that the ERIM successfully activates latent domain-expert representations, thereby considerably enhancing stylistic professionalism while maintaining logical precision, achieving a transformation from “generic exposition” to “professional narrative”.

4.4 Analysis of 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 considerably; 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 pre-training 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.,

2024) 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, 2025c; Google, 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 ToT / GoT variants (Yao et al., 2023; Besta et al., 2024). Recent dynamic frameworks attempt to introduce adaptability but remain confined to closed-world optimization. For instance, DGoT (Ning et al., 2024) utilizes dynamic thresholds primarily for early stopping to reduce computational costs. MCTSr (Zhang et al., 2024) relies on internal self-correction to backtrack logical errors in specific puzzles, while RATT (Zhang et al., 2025) uses retrieval strictly to verify existing nodes. In contrast, Cogito introduces information-driven topology mutation. Rather than merely optimizing a search path within a fixed scope, it dynamically alters the research scope itself, spawning new inquiry branches based on external unexpected discoveries to explicitly address open-world “unknown unknowns”.

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 en-

ables 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. Therefore, we emphasize that Cogito is intended as an augmented intelligence tool to support, not replace, human analysts.

Ethical Considerations

This work adheres to the ACL Code of Ethics. In this study, no human subjects or animal experimentation were involved. The constructed dataset complies with relevant usage guidelines. 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.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61977055 and by the Open Research Projects of Pingtan Research Institute, Xiamen University under Grant No. 202506190004.

References

- 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. <https://github.com/assafelovic/gpt-researcher>.
- 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. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20–27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.
- 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.
- João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and Chenyan Xiong. 2025. Deepresearchgym: A free, transparent, and reproducible evaluation sandbox for deep research. *CoRR*, abs/2505.19253.
- DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556.
- Jonathan St. B. T. Evans and Keith E. Stanovich. 2013. Dual-Process theories of higher cognition. *Perspectives on Psychological Science*, 8(3):223–241.
- Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.
- Google. 2025. Gemini deep research. <https://gemini.google/overview/deep-research>.
- Grok. 2025. Grok ai deepsearch: Real-time research power guide. <https://grokaimodel.com/deepsearch>.
- Guangxin He, Zonghong Dai, Jiangcheng Zhu, Binqiang Zhao, Chenyue Li, You Peng, Chen Wang, and Binhang Yuan. 2024. Zero-indexing internet search augmented generation for large language models. *CoRR*, abs/2411.19478.
- Jiajie Jin, Yuyao Zhang, Yimeng Xu, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. 2025. Finsight: Towards real-world financial deep research. *CoRR*, abs/2510.16844.
- Walter Krämer. 2013. Kahneman, D. (2011): Thinking, Fast and Slow. *Statistical Papers*, 55(3):915.
- 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, and Daochen Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *CoRR*, abs/2307.10485.
- Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.
- Xinyu Ning, Yutong Zhao, Yitong Liu, and Hongwen Yang. 2024. DGoT: Dynamic graph of thoughts for scientific abstract generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4832–4846, Torino, Italia. ELRA and ICCL.
- OpenAI. 2025a. Gpt-5.1: A smarter, more conversational chatgpt.
- OpenAI. 2025b. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925.
- OpenAI. 2025c. 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.
- Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.
- 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. <https://github.com/bytedance/deer-flow>.

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](#). In Brian Mullen and George R. Goethals, editors, *Theories of Group Behavior*, pages 185–208. Springer New York, New York, NY.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. 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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. [Bloomberggpt: A large language model for finance](#). *CoRR*, abs/2303.17564.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024. [Tradingagents: Multi-agents LLM financial trading framework](#). *CoRR*, abs/2412.20138.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Qizhe Xie. 2023. [Self-evaluation guided beam search for reasoning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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. 2023. [Beyond chain-of-thought, effective graph-of-thought reasoning in large language models](#). *CoRR*, abs/2305.16582.

Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. 2024. [Accessing GPT-4](#)

[level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b](#). *CoRR*, abs/2406.07394.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025. [RATT: A thought structure for coherent and correct LLM reasoning](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 26733–26741. AAAI Press.

A Process of ECM

Algorithm 1: Process of ECM

Input : Query q , Environment \mathcal{E} , Max Steps T_{max}

Output : Final Report R^*

- 1 $G_0 \leftarrow \text{InitBlueprint}(q)$;
- 2 $\mathcal{V}_0 \leftarrow \text{GetReadyNodes}(G_0)$;
- 3 $t \leftarrow 0$;
- 4 **while** $\mathcal{V}_t \neq \emptyset$ **and** $t < T_{max}$ **do**
 - // 1. Parallel Execution Phase
 - 5 $\mathcal{O}_t, \mathcal{D}_t \leftarrow \emptyset, \emptyset$;
 - 6 **foreach** $v \in \mathcal{V}_t$ **in parallel do**
 - 7 $o_v, d_v \leftarrow \text{Execute}(v, \mathcal{E})$;
 - 8 $\mathcal{O}_t \leftarrow \mathcal{O}_t \cup \{o_v\}, \mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{d_v\}$;
 - // 2. Horizontal Audit Phase
 - 9 $signal \leftarrow \text{Audit}(\mathcal{O}_t, \mathcal{D}_t, G_t)$;
 - 10 **if** $signal = \text{Stop}$ **then**
 - 11 **break**;
 - // 3. Dynamic Evolution Phase
 - 12 $\Delta_{plan} \leftarrow \Gamma(\mathcal{D}_t, \mathcal{O}_t)$ // Task Generation
 - 13 $G_{t+1} \leftarrow \Omega(G_t, \Delta_{plan})$ // Topology Evol
 - 14 $\mathcal{U}(G_{t+1})$ // State Update
 - 15 $\mathcal{V}_{t+1} \leftarrow \text{GetReadyNodes}(G_{t+1})$;
 - 16 $t \leftarrow t + 1$;
- 17 $R^* \leftarrow \text{SynthesizeReport}(G_t)$;
- 18 **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., Bloomberg³, Reuters⁴) 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 500⁵, NASDAQ-100⁶) 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

³<https://www.bloombergchina.com/>

⁴<https://reutersagency.com/>

⁵<https://www.marketwatch.com/investing/index/spx>

⁶<https://indexes.nasdaqomx.com/Index/Overview/NDX>

scale of 0 to 100 based on three core dimensions: 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:

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

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):

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

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

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
...

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:

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

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 API⁷ 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, FMP⁸ 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.

⁷<https://www.alphavantage.co/>

⁸<https://site.financialmodelingprep.com/>

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.

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

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.