FINDEEPFORECAST : A Live Multi-Agent System for Benchmarking Deep Research Agents in Financial Forecasting

Anonymous ACL submission

Abstract

Deep Research (DR) Agents powered by advanced Large Language Models (LLMs) have fundamentally shifted the paradigm for completing complex research tasks. Yet, a comprehensive and live evaluation of their forecasting performance on real-world, research-oriented tasks in high-stakes domains (*e.g.,* finance) remains underexplored. We introduce FINDEEPFORECAST, the first live, end-to-end multi-agent system for automatically evaluating DR agents by continuously generating research-oriented financial forecasting tasks. This system is equipped with a *dual-track taxonomy*, enabling the dynamic generation of recurrent and non-recurrent forecasting tasks at both corporate and macro levels. With this system, we generate FINDEEPFORECASTBENCH, a weekly evaluation benchmark over a ten-week horizon, encompassing 8 global economies and 1,314 listed companies, and evaluate 13 representative methods. Extensive experiments show that, while DR agents consistently outperform strong baselines, their performance still falls short of genuine forward-looking financial reasoning. We expect the proposed FINDEEPFORECAST system to consistently facilitate future advancements of DR agents in research-oriented financial forecasting tasks. The samples of the generated tasks are available at `https://huggingface.co/datasets/FinDeepForecast/FinDeepForecast`.

1 Introduction

Deep Research (DR) agents are autonomous artificial intelligence (AI) systems that perform complex research tasks via iterative planning, evidence acquisition, reasoning, and reporting (Shi et al., 2025; Huang et al., 2025). Their emergence has reshaped the approach to complex tasks, attracting growing research attention. Beyond developing more powerful DR agents, reliably evaluating their research capabilities is an equally fundamental objective for advancing the field, yet is currently underexplored. Traditional static evaluation benchmarks inevitably leak into training corpora, rendering them obsolete (*i.e.,* data contamination) (Xu et al., 2024; Balloccu et al., 2024). Recently, live benchmarks have been explored, continuously generating novel instances to ensure temporal separation between training and evaluation data (White et al., 2025).

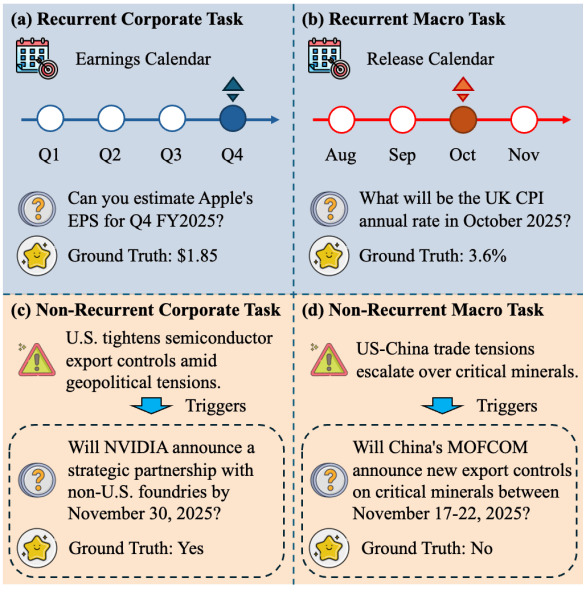

Figure 1: Recurrent tasks for regular disclosures and non-recurrent tasks for event-driven predictions.

Existing live benchmarks are either *time-insensitive* or *time-sensitive*. The former focus on domains with static and deterministic ground-truths, like code generation (Jain et al., 2024; Zhang et al., 2025a) and mathematics problem-solving (Balunović et al., 2025), often leading to insufficiently rigorous evaluations—models may rely on the recall of pre-existing answer patterns rather than genuine reasoning. Time-sensitive benchmarks (Karger et al., 2024; Zeng et al., 2025) are constructed around future data, undisclosed content, or unknown outcomes, such that correct answers cannot be known before evaluation, enabling rigorous assessment of predictive reasoning. How-

ever, the tasks in these benchmarks are often sampled from existing question websites or generated using fixed templates. Such heavy reliance on external sources or predefined templates can introduce inherent biases and limit the availability of genuinely research-oriented tasks, thereby constraining the breadth and depth of the evaluation. To truly benchmark the capabilities of DR agents, a dynamic evaluation environment is needed—one that continuously supplies **forward-looking, research-oriented** tasks with strict temporal constraints and objectively verifiable outcomes.

The financial domain offers an ideal setting for such forward-looking, research-oriented tasks (Koval et al., 2024; Chen et al., 2024), with properties that support the ongoing evaluation of DR agents' forecasting capabilities. 1) It offers *Periodic Information Disclosure*, with a continuous stream of verifiable data points, such as corporate financial metrics (*e.g.,* EPS in Fig. 1 (a)) and macroeconomic (macro) indicators (*e.g.,* CPI in Fig. 1 (b)). 2) It features *Diverse Market Events*, necessitating the distillation of valuable signals and market dynamics from vast, noisy data for accurately forecasting various critical events, including corporate actions (*e.g.,* new partnership in Fig. 1 (c)) and macro shifts (*e.g.,* export control in Fig. 1 (d)). 3) It guarantees *Strict Temporal Isolation*, with answers emerging strictly upon disclosure of information or occurrence of the corresponding events. In such financial markets, financial experts have to gather and analyze information and forecast future outcomes through reasoning in order to complete complex tasks as shown in Fig. 1.

Aimed at continuously evaluating DR agents in addressing research-oriented financial forecasting challenges, we propose a novel, live multi-agent system, named FINDEEPFORECAST. As shown in Fig. 2, it employs a *dual-track taxonomy* (see Appendix A and B for more details) for effectively managing recurrent and non-recurrent forecasting scenarios, encompassing corporate- and macro-level tasks. It comprises four key stages, powered by six specialized agents, for an automatic, end-to-end evaluation of DR agents, starting from data collection and task generation through model forecasting to ground-truth acquisition and performance evaluation. To validate this system, we generate a FINDEEPFORECASTBENCH benchmark, which covers 8 major global economies for macro tasks and 1,314 listed companies drawn from 9 major indices for corporate tasks. In total, it consists of 1,394 tasks, including 296 recurrent macro, 723 recurrent corporate, 128 non-recurrent macro, and 247 non-recurrent corporate tasks.

We assess 13 representative methods for completing the weekly tasks in FINDEEPFORECASTBENCH, including 3 DR agents, 5 LLMs with both thinking and search capabilities, and 5 LLMs with thinking capabilities. Several important findings have been made. 1) DR agents consistently exhibit superior performance to the compared methods, but they still struggle significantly with the tasks in FINDEEPFORECASTBENCH, with the highest score 39.5 out of 100. 2) Most methods achieve peak performance in information-rich markets (*e.g.,* US and China) but underperform in markets with relatively limited data or language diversity (*e.g.,* Japan). 3) The models achieve high accuracy on non-recurrent tasks, but their performance declines sharply on recurrent tasks. This disparity underscores the greater intrinsic difficulty of precise numeric forecasting under periodic disclosure when compared to binary event prediction.

Our contributions are summarized as follows:

- We develop FINDEEPFORECAST, **the first end-to-end multi-agent system** designed to continuously produce **forward-looking, research-oriented tasks** in finance for the contamination-free evaluation of DR agents.
- We propose a **dual-track taxonomy** for the dynamic generation of both recurrent and non-recurrent financial forecasting tasks, encompassing corporate- and macro-level predictions (covering hundreds of metrics and event categories) within a live market environment.
- With FINDEEPFORECAST, we generate FINDEEPFORECASTBENCH, a **weekly evaluation benchmark** spanning a ten-week horizon, currently covering 8 major global economies and 1,314 listed companies from 9 leading indices, while remaining readily extensible to additional markets and firms.
- Extensive evaluations of 13 representative methods show that, although DR agents significantly outperform alternative approaches, they still exhibit substantial room for improvement, highlighting the limitations of current methods in solving these tasks. This firmly establishes our FINDEEPFORECAST as a timely and essential contribution for consistently facilitating the future advancement of DR agents.

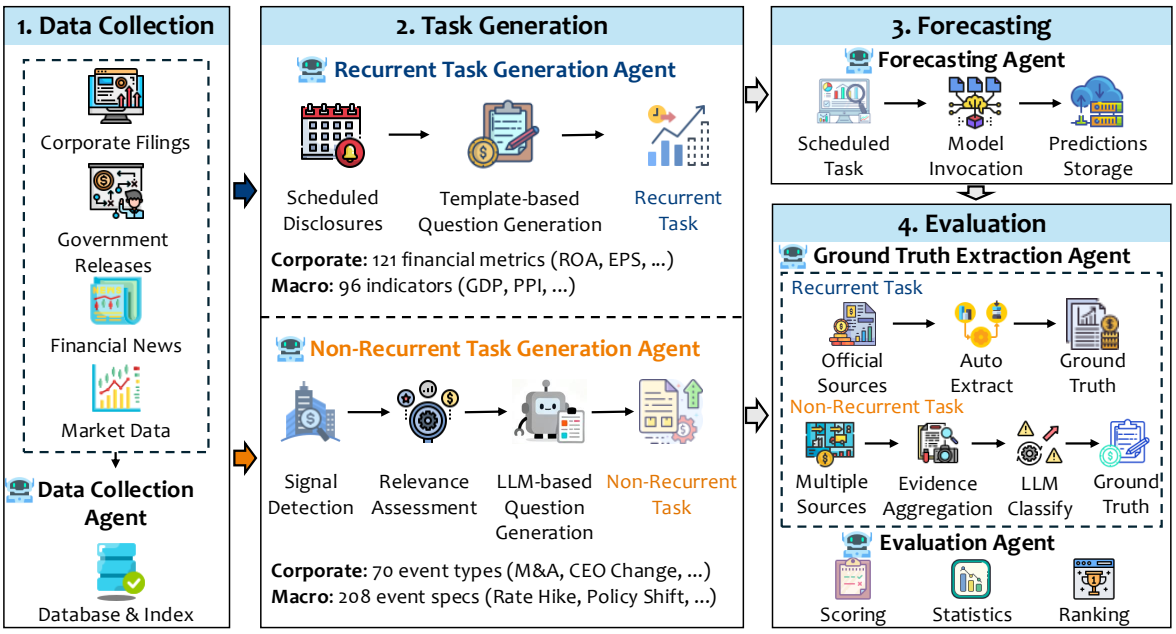

Figure 2: The FINDEEPFORECAST system comprises four stages: (1) Data Collection aggregates financial information into a timestamped database; (2) Task Generation produces recurrent tasks via template-based question generation and non-recurrent tasks via LLM-based pipeline; (3) Forecasting invokes models with temporal isolation; (4) Evaluation extracts ground truth and computes performance metrics.

2 FINDEEPFORECAST System

In this section, we introduce FINDEEPFORECAST, a live, multi-agent system for assessing genuine capabilities of DR agents in financial forecasting through research-oriented task generation, strict temporal isolation, and rigorous ground truth verification, as shown in Fig. 2. In FINDEEPFORECAST, a *dual-track taxonomy* is devised to distinguish recurrent predictions for numerical estimation on scheduled disclosures from non-recurrent predictions for binary classification on uncertain emerging events. *Continuous generation with temporal isolation* prevents data contamination through live task creation while enforcing uniform information boundaries across models.

2.1 Forecasting Problem Definition

A forecasting problem is defined as a tuple $\mathcal{P} = (q, t_g, t_d, t_e, y)$, where q denotes the forecasting question, t_g the task generation time, t_d the forecasting deadline, t_e the evaluation time, and y the ground truth outcome. The temporal ordering $t_g < t_d < t_e$ ensures forecasts are made before outcomes become observable. Given a forecasting problem, a model produces a forecast $\hat{y} = f(q, \mathcal{I}_{t_d})$, where f represents the model's forecasting function and $\mathcal{I}_{t_d}$ denotes the information set available up to deadline t_d.

Our dual-track taxonomy distinguishes forecasting problems primarily by *temporal predictability*: recurrent forecasts target scheduled disclosures with known timing but uncertain numerical outcomes, while non-recurrent forecasts address events whose occurrence itself cannot be anticipated from calendars. Within each track, tasks are further organized by *forecasting scope* into corporate and macro levels, yielding four complementary evaluation dimensions.

2.2 Data Collection

From an authoritative source[1], the Data Collection Agent continuously monitors and collects four categories of information: 1) corporate filings from regulatory databases, 2) government releases from statistical agencies, 3) financial news from real-time streams, and 4) market data from exchanges. All collected data is organized into a timestamped database and index, enabling temporal isolation during evaluation by restricting model access to content published before prediction deadlines.

2.3 Task Generation

Task generation employs two specialized agents corresponding to our dual-track taxonomy.

Recurrent Task Generation. The Recurrent Task Generation Agent constructs recurrent tasks

[1]Source details omitted for anonymity and will be disclosed upon acceptance.

through a two-stage pipeline: identifying scheduled disclosures from official calendars, then applying template-based question generation to produce tasks with temporal parameters (t_g, t_d, t_e). Macro-level tasks monitor 14 indicators (*e.g.,* GDP growth, PPI change) and corporate-level tasks target 121 financial metrics (*e.g.,* ROA, EPS). Complete specifications are provided in Appendix A.

Non-Recurrent Task Generation. The Non-Recurrent Task Generation Agent employs an LLM-based pipeline in three stages: (1) signal detection identifies indicators from news streams, (2) relevance assessment evaluates predictive salience, and (3) LLM-based question generation produces tasks with explicit event definitions. Macro tasks follow a taxonomy of 9 high-level categories and 26 fine-grained subcategories, which are instantiated via a Core–Adaptive interface into economy-specific event specifications (e.g., rate hikes, policy shifts). Corporate tasks are defined over 70 curated event types (e.g., M&A, CEO change) with clear predictive semantics and objectively verifiable outcomes, which are instantiated at the level of individual listed companies. Complete taxonomies are provided in Appendix B, and implementation details of the generation pipeline are described in Appendix C.

2.4 Forecasting

The Forecasting Agent elicits predictions through a three-step workflow. First, *task scheduling* organizes generated tasks by their prediction deadlines t_d and assigns them to weekly evaluation batches. Second, *model invocation* calls each evaluated model via its API with standardized prompts containing the prediction question and deadline; to ensure temporal isolation, search-augmented and deep research models are configured to access only content published before t_d. Third, *prediction storage* records all model outputs with timestamps into a structured database for subsequent evaluation.

2.5 Evaluation

To ensure evaluation objectivity and simulate real-world financial accountability, we adopt a deterministic, outcome-oriented protocol managed by two agents.

Ground Truth Extraction. For non-recurrent tasks, we employ a human-in-the-loop protocol to prevent evaluation bias: an LLM agent first aggregates multi-source evidence to propose potential outcomes, which are then **strictly verified by domain experts for 100% of the samples** to determine the final ground truth. Details of this verification protocol are provided in Appendix D.

Scoring Metric. Each forecasting task is evaluated using a binary scoring function that awards 1 for correct forecasts and 0 otherwise:

$$\text{Score}(y, \hat{y}) = \begin{cases} \mathbf{1}\left[\left|\frac{\hat{y}-y}{y}\right| < \epsilon_k\right] & \text{if recurrent} \\ \mathbf{1}\left[\hat{y} = y\right] & \text{if non-recurrent} \end{cases} \quad (1)$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, y is the ground truth, $\hat{y}$ is the model forecast, and ϵ_k is the indicator-specific tolerance threshold. For recurrent tasks, a forecast is correct if the relative error falls within the threshold; for non-recurrent tasks, exact match is required. The overall accuracy is computed as the percentage of correct forecasts:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \text{Score}(y_i, \hat{y}_i) \times 100\% \quad (2)$$

where N is the total number of evaluated tasks. Results are disaggregated by task category, forecasting horizon, and market region.

3 FINDEEPFORECASTBENCH

To validate FINDEEPFORECAST, we introduce FINDEEPFORECASTBENCH, a weekly evaluation benchmark spanning ten weeks and targeting 8 major markets. This section describes the construction and quality control of FINDEEPFORECASTBENCH, compares it with existing benchmarks, and defers detailed statistics to Appendix I.

3.1 FINDEEPFORECASTBENCH Generation

We instantiate FINDEEPFORECASTBENCH using the proposed FINDEEPFORECAST system under the following settings.

Market and Company Selection. Markets are selected based on economic significance and data availability. The current instantiation spans three continents and covers **eight major economies**: US, CN, HK, JP, UK, DE, FR, and SG. To establish a standardized evaluation universe, we anchor company selection to nine leading equity indices across the covered economies (S&P 500, NASDAQ 100, FTSE 100, DAX 40, CAC 40, Nikkei 225, CSI 300, HSI, and STI). The resulting corporate pool comprises 1,314 constituent firms, defined by index membership at a fixed snapshot date (2 Oct 2025).

Weekly Task Generation. The benchmark has operated continuously since 27 October 2025, releasing a new task batch every Thursday. For recurrent corporate tasks, we employ a dynamic stratified sampling strategy, selecting up to 30% of reporting companies per market weekly to balance density across regions; for recurrent macro tasks, we cover all scheduled indicator releases across the 8 economies. For non-recurrent tasks, candidates are generated from live news streams and undergo weekly expert review, with only the highest-quality events selected for inclusion to ensure predictive value and answerability.

Ground Truth Acquisition. Ground truth acquisition operates on a rolling weekly cycle, processing tasks where the evaluation time t_e has passed every Monday. For non-recurrent tasks, domain experts verify the automated classification results to ensure validity. To handle real-world irregularities such as delayed disclosures, tasks with indeterminate outcomes are marked as Pending and revisited weekly. Tasks remaining unresolved after a 2-week validity window are marked as *Void* and excluded from performance evaluation.

3.2 Quality Control

We ensure benchmark reliability through systematic quality control at each stage of the pipeline.

Expert Involvement. Domain experts participate throughout the benchmark lifecycle. A team of researchers with doctoral-level training in finance and economics contributes to multiple stages: designing standardized templates for recurrent task generation, defining event taxonomies and early signal criteria for non-recurrent tasks, conducting reviews of generated questions, and verifying ground truths.

Task Quality Assurance. For recurrent tasks, standardized templates validated by domain experts ensure consistent metric definitions aligned with regulatory reporting standards. For non-recurrent tasks, experts review all candidate questions weekly and apply strict selection criteria, ensuring both quality and cross-market balance in the final task set.

Ground Truth Verification. For recurrent tasks, ground truth is extracted from official sources through automated parsing. Sampling verification against primary sources confirms 99.8% accuracy, with rare discrepancies attributable to subsequent data restatements by issuers. For non-recurrent tasks, ground truth determination combines automated evidence aggregation with expert validation, achieving 95% inter-annotator agreement.

Table 1: Comparison between FINDEEPFORECASTBENCH and existing benchmarks. "T-S" and "T-IS' denotes "Time-Sensitive" and "Time-Insensitive".

	Domain	Type	Question	Frequency
Finance Forecasting				
FLUE	Finance	T-IS	Curated	One-time
PIXIU	Finance	T-IS	Curated	One-time
FinanceBench	Finance	T-IS	Annotation	One-time
FinBen	Finance	T-IS	Curated	One-time
FinCall	Finance	T-IS	Rule-based	One-time
Live Benchmarks				
LiveBench	Misc	T-IS	Curated	Monthly
LiveCodeBench	Coding	T-IS	Rule-based	Monthly
SWE-bench	Software	T-IS	Rule-based	Seasonal
MathArena	Math	T-IS	Curated	Seasonal
CryptoBench	Crypto	T-IS	Annotation	Monthly
LiveXiv	Scientific	T-IS	Rule-based	Monthly
ForecastBench	General	T-S	Rule-based	Weekly
FutureX	General	T-S	Rule-based	Weekly
Ours				
FINDEEPFORECAST	**Finance**	**T-S**	**Dynamic**	**Weekly**

3.3 Comparison with Existing Benchmarks

Table 1 compares FINDEEPFORECASTBENCH with existing benchmarks across key dimensions. We make the following observations: 1) Existing financial forecasting benchmarks primarily focus on *time-insensitive* and *recurrent* forecasting tasks. These are typically evaluated in a one-time setting on historical data, where the ground truth is already known at test time, leading to inherently vulnerable to data contamination. 2) Most existing live benchmarks are *time-insensitive* in deterministic domains such as coding, software engineering, and mathematics, resulting in insufficiently rigorous evaluations because models may rely on the recall of pre-existing answer patterns. 3) *Time-sensitive* benchmarks represent important steps toward evaluating forecasting tasks; however, they largely rely on fixed templates or rule-based extraction pipelines for task generation, limiting task diversity and constraining the benchmark's ability to reflect evolving real-world scenarios.

Key Differentiators. FINDEEPFORECASTBENCH differs from prior benchmarks through three characteristics: 1) a dual-track taxonomy that jointly supports recurrent and non-recurrent forecasting tasks, covering both regular disclosures and event-driven predictions; 2) financial domain specialization with objectively verifiable ground truth derived from official filings, statistical releases, and authoritative disclosures; 3) fully dynamic task generation in live market environments with weekly updates, enabling a continuous, research-oriented and contamination-free evaluation of DR agents.

4 Experiments

We comprehensively assess the financial forecasting capabilities of state-of-the-art methods.

4.1 Evaluation Models

We evaluate 13 models spanning three paradigms with distinct information access capabilities.

LLM with Thinking (T). OpenAI GPT-5 (T) (Team, 2025c), Claude-Sonnet-4.5 (T) (Team, 2025a), Gemini 2.5 Pro (T) (Kavukcuoglu, 2025), Deepseek-v3.2 (T) (Team, 2025b) and Grok 4 (T) (xAI Team, 2025).

LLM with Thinking + Search (T+S). OpenAI GPT-5 (T+S) (Team, 2025c), Claude-Sonnet-4.5 (T+S) (Team, 2025a), Gemini 2.5 Pro (T+S) (Kavukcuoglu, 2025), Deepseek-v3.2 (T+S) (Team, 2025b), and Grok 4 (T+S) (xAI Team, 2025).

Deep Research. OpenAI o3-deep-research (OpenAI Team, 2025), Perplexity Sonar Deep Research (Team, 2025d) and Tongyi Deep Research (Tongyi Team, 2025).

4.2 Implementation Details

Temporal Isolation. To ensure a fair comparison, we enforce strict temporal isolation between task generation, model prediction, and performance evaluation. All evaluated models can only access the content published before the prediction deadline t_d, preventing access to information unavailable at prediction time.

Task Generation. Tasks are generated every Thursday with prediction deadline t_d set to the following Sunday 23:59 (UTC+8). For recurrent corporate tasks, we select companies with earnings releases scheduled within the prediction window. For non-recurrent tasks, domain experts review candidate questions and select about 20% for inclusion based on prediction quality and market balance.

Model Forecasting. All models receive standardized prompts specifying the prediction question, deadline, and required output format. Models produce structured outputs: numerical estimates $\hat{y} \in \mathbb{R}$ for recurrent tasks and binary predictions $\hat{y} \in \{\text{Yes}, \text{No}\}$ for non-recurrent tasks. Samples of the input and output are provided in Appendix L. Model settings for forecasting tasks enable thinking and web searching capabilities where applicable. Unless otherwise specified, all parameters use default values. Detailed model configurations are provided in the Appendix J.

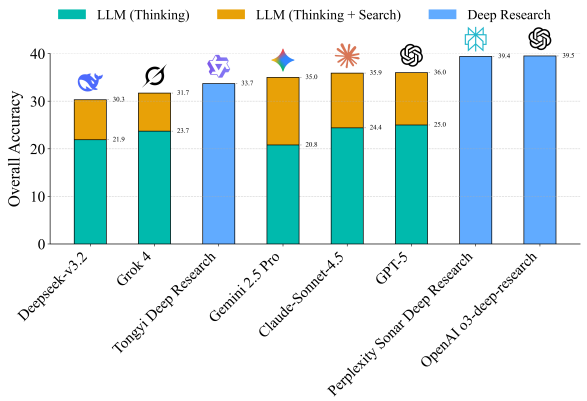

Figure 3: Main results. Overall model performance comparison over the entire ten-week horizon.

Answer Evaluation. For recurrent tasks, we apply indicator-specific thresholds ϵ_k based on unit type and indicator category. By unit type, thresholds are set to 5% for million-scale financial metrics and 1% for percentage and ratio metrics. By indicator category, thresholds are set to 0.1% for interest rates and foreign exchange rates, and 1% for other macro indicators. For non-recurrent tasks, ground truth is determined through evidence aggregation from multiple sources, with expert verification achieving 95% inter-annotator agreement.

4.3 Main Results

We evaluate model performance across nearly 1,400 forecasting tasks generated over a 10-week period, with accuracy measured as the proportion of correct predictions. The empirical results from Figure 3 reveal three critical findings: 1) Deep Research models establish clear performance superiority when compared to the other two counterparts. OpenAI o3-deep-research (39.5%) and Perplexity Sonar Deep Research (39.4%) outperform all other architectural approaches, with their near-parity suggesting convergence at the frontier of this paradigm. 2) Augmenting reasoning with search functionality, while beneficial, cannot replicate Deep Research performance. The leading Thinking + Search (T+S) models, GPT-5 (36.0%), Claude-Sonnet-4.5 (35.9%), and Gemini 2.5 Pro (35.0%), cluster within a narrow band yet consistently underperform Deep Research systems by 3-4 percentage points, indicating that search augmentation alone does not capture the full advantages of specialized Deep Research architectures. 3) External information retrieval capabilities are critical for benchmark performance. Within-model comparisons between T and T+S configurations reveal systematic accuracy degradation when search is

Table 2: Performance analysis across non-recurrent (Non-rec.) and recurrent (Rec.) scenarios. Values reported for Corporate (Corp.), Macro (Mac.), and Overall (Ovr.) tasks are overall accuracy.

Method	Non-rec.			Rec.		
	Corp.	Mac.	Ovr.	Corp.	Mac.	Ovr.
LLM (Thinking)						
OpenAI GPT-5 (T)	68.4	65.6	67.5	6.8	4.1	6.0
Claude-Sonnet-4.5 (T)	66.8	70.3	68.0	8.4	1.0	6.3
Grok 4 (T)	73.7	71.2	73.1	11.5	2.7	9.0
Deepseek-v3.2 (T)	61.9	58.6	60.8	6.9	2.0	5.6
Gemini 2.5 Pro (T)	73.3	68.8	71.7	8.5	1.0	6.4
LLM (Thinking + Search)						
OpenAI GPT-5 (T+S)	78.1	72.7	76.3	22.8	11.1	19.5
Claude-Sonnet-4.5 (T+S)	79.8	73.4	77.6	20.7	19.6	20.4
Grok 4 (T+S)	74.5	77.3	75.5	15.1	18.2	16.0
Deepseek-v3.2 (T+S)	76.9	70.3	74.7	13.4	14.6	13.7
Gemini 2.5 Pro (T+S)	78.5	**77.5**	78.3	23.3	17.6	21.7
Deep Research						
Perplexity Sonar Deep Research	**81.4**	75.0	**79.2**	26.2	**23.7**	**25.5**
Tongyi Deep Research	79.8	74.2	77.9	23.5	15.5	21.2
OpenAI o3-deep-research	80.6	75.8	78.9	**26.7**	21.3	25.2

removed, with performance drops ranging from 11.0 percentage points (GPT-5: 36.0% → 25.0%) to 14.2 percentage points (Gemini 2.5 Pro: 35.0% → 20.8%). This uniform pattern across all evaluated models provides robust evidence that access to external information is a fundamental determinant of success on this benchmark, with pure reasoning capabilities alone proving insufficient.

4.4 In-depth Analysis

Performance Analysis on Different Tasks. We analyze the model performance across different scenarios (*i.e.,* recurrent and non-recurrent) at different levels (*i.e.,* corporate and macro). We present the results in Table 2, from which we make the following key findings: 1) The thinking-only LLMs achieve reasonable accuracy on non-recurrent tasks but collapse on recurrent scenarios (often below 10% overall), suggesting that internal reasoning alone is insufficient for temporally grounded, fine-grained financial prediction. 2) LLMs augmented with thinking and search capabilities outperform thinking-only counterparts across all task types, indicating the importance of external information access. However, their improvements on recurrent tasks remain modest, implying that information retrieval alone is insufficient for addressing the challenges in FINDEEPFORECASTBENCH. 4) Deep Research agents achieve the best performance across both non-recurrent and recurrent tasks, particularly on recurrent corporate and macro forecasting. This suggests that multi-step planning, evidence synthesis, and structured reasoning jointly contribute to stronger forecasting under strict temporal isolation. 4) Across all paradigms, models

Table 3: The performance analysis across 8 financial markets. The values reported in the table denote the overall accuracy.

Method	US	CN	HK	JP	UK	DE	FR	SG
LLM (Thinking)								
OpenAI GPT-5 (T)	19.0	32.3	26.7	14.4	28.6	47.6	39.1	32.6
Claude-Sonnet-4.5 (T)	18.1	28.5	28.3	16.7	31.0	44.0	32.7	32.6
Grok 4 (T)	20.2	26.9	20.8	17.5	31.0	39.3	26.4	31.5
Deepseek-v3.2 (T)	16.5	26.9	25.8	12.5	27.8	42.9	30.9	28.1
Gemini 2.5 Pro (T)	18.7	24.6	19.2	12.8	26.2	39.3	21.8	25.8
LLM (Thinking + Search)								
OpenAI GPT-5 (T+S)	36.7	40.0	35.8	22.6	38.9	47.6	44.5	39.3
Claude-Sonnet-4.5 (T+S)	33.4	42.3	42.5	23.7	39.7	48.8	40.0	43.8
Grok 4 (T+S)	29.2	40.8	34.2	16.7	37.3	45.2	37.3	44.9
Deepseek-v3.2 (T+S)	26.4	36.9	40.3	17.1	33.3	45.8	37.3	39.3
Gemini 2.5 Pro (T+S)	35.7	38.0	30.8	22.0	34.7	47.5	44.5	**47.7**
Deep Research								
Perplexity Sonar Deep Research	40.1	**45.7**	**47.9**	24.3	**40.0**	50.0	40.9	44.9
Tongyi Deep Research	33.9	44.6	29.4	20.7	38.7	47.0	36.4	36.4
OpenAI o3-deep-research	**41.0**	45.4	40.0	**25.3**	38.9	**51.2**	**48.2**	41.6

achieve high accuracy on non-recurrent tasks (up to 81.4%), while performance on recurrent tasks drops sharply, with the best method reaching only 25.5% overall. This highlights the intrinsic difficulty of precise, numeric forecasting under periodic disclosure compared to binary event prediction. To better understand the failure modes, we provide a error case study in Appendix K.

Performance Analysis across Different Markets. Then, we analyze the model performance across different markets, and present the results in Table 3, from which we observe: 1) Deep Research agents achieve the highest accuracy in nearly all markets, indicating superior generalization across heterogeneous regulatory regimes, disclosure standards, and information environments. 2) In every market, adding search capabilities leads to substantial gains, underscoring the importance of accessing up-to-date and market-specific information in live financial forecasting settings. 3) Most methods perform best in information-rich markets such as the US and China, while accuracy is consistently lower in markets with relatively less available data or language diversity, such as Japan.

Weekly Performance Analysis. We analyze the model performance over all ten weeks, and present the results in Figure 4. We can observe:

1) Accuracy improves steadily across weeks as the proportion of recurrent tasks declines following the end of the disclosure period, consistent with the stronger performance of all models on non-recurrent tasks shown in Table 2. 2) Deep Research agents outperform all other methods consistently across all weeks, indicating a superior and stable capacity to integrate observed signals and adapt over time.

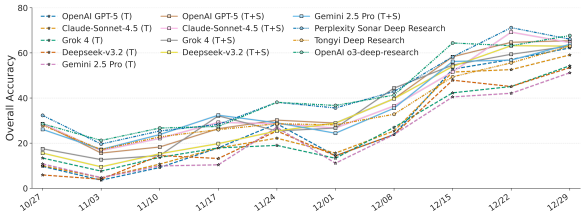

Figure 4: Weekly performance comparison.

5 Related Work

5.1 Deep Research Agents

Deep Research (DR) agents aim to solve complex tasks through planning, information gathering, and multi-step reasoning, and have recently been widely deployed in both industrial and open-source LLM systems (OpenAI, 2025; Citron, 2025; Qwen, 2025; Perplexity, 2025; Tongyi et al., 2025). Building on early agentic paradigms that interleave reasoning with tool use (Yao et al., 2023; Shinn et al., 2023), subsequent works extended these approaches to realistic domains such as web navigation and software engineering (Zhou et al., 2023; Huang et al., 2023), with accompanying benchmarks for evaluating agentic capabilities (Liu et al., 2023; Zhu et al., 2025a). However, most existing evaluations rely on static task sets, limiting their ability to capture agent behavior under changing environments. As DR agents increasingly operate in evolving real-world contexts, reliably evaluating performance under dynamic and time-sensitive conditions has become a critical challenge that our work seeks to address.

5.2 Live Benchmarks

Live benchmarking for LLMs has emerged as a key direction for mitigating data contamination in evaluation(Balloccu et al., 2024). Existing live benchmarks can be broadly categorized into *time-insensitive* and *time-sensitive* tasks. While time-insensitive benchmarks (White et al., 2025; Jain et al., 2024) seek to mitigate contamination through continuous updates, their tasks rely on deterministic ground truths that do not depend on future outcomes. In contrast, time-sensitive benchmarks (Karger et al., 2024; Zeng et al., 2025) evaluate predictive reasoning on problems whose answers are unknown at test time, but typically rely on manual curation or rule-based extraction pipelines, constraining task diversity and adaptability. In the financial domain, several live trading benchmarks (Qian et al., 2025; Li et al., 2025) evaluate LLM-based agents via profitability-based scoring, targeting trading decisions rather than research forecasting(Appendix E). Our work introduces a time-sensitive benchmark for financial forecasting, where tasks are dynamically generated from evolving real-world market environments.

5.3 Financial Forecasting Benchmarks

Current financial forecasting benchmarks (Shah et al., 2022; Islam et al., 2023) typically focus on *recurrent* events such as stock price movements (Xie et al., 2023, 2024) or company earnings (Shu et al., 2025), constructed from static historical datasets vulnerable to data contamination (Sainz et al., 2023). Another crucial but less common benchmark is on *non-recurrent* tasks, which are discrete events that are also known to impact financial markets (Hu et al., 2018; Koa et al., 2024), such as new partnerships or tariffs. The benchmarks can also be categorized into *corporate-level* (Zhu et al., 2025b) or *macro-level* (Zhang et al., 2025b) tasks, which differ in terms of event scale and granularity. Our work deals with these tasks under a unified evaluation framework, constructed from live data.

6 Conclusion

In this work, we introduce FINDEEPFORECAST, the first live, end-to-end multi-agent system for evaluating DR agents in financial forecasting. It can continuously generate forward-looking, research-oriented tasks under strict temporal isolation, and integrates task creation, model invocation, and ground-truth verification into a unified and fully automated pipeline. With this system, we instantiate FINDEEPFORECASTBENCH, a weekly benchmark covering recurrent numerical disclosures and non-recurrent event-driven predictions at both corporate and macroeconomic levels. We evaluate 13 representative systems, demonstrating that current DR agents are still challenged by genuinely research-oriented financial forecasting, particularly in precise recurrent numerical forecasting. FINDEEPFORECAST establishes a dynamic and contamination-free evaluation paradigm for DR agents in live market environments and provides a foundation on which future systems and benchmarks can be continuously built and extended.

In future, we plan to extend the system to richer task forms, including probabilistic, multi-step, and portfolio-level forecasting, and to incorporate process-based evaluation to better understand how DR agents search, reason, and fail in live forecasting scenarios.

Limitations

While our study provides a comprehensive and rigorous evaluation, several limitations remain.

Temporal Scope. Our ten-week, 1,370-task evaluation enables reliable short-term assessment, but does not fully capture longer-term seasonal or cyclical market dynamics. Extending the benchmark over multiple quarters would allow stronger conclusions about temporal robustness.

Market Coverage. We focus on eight major economies that account for a large share of global market activity. Emerging markets, which exhibit distinct information structures and volatility patterns, are not yet included and remain an important direction for future expansion.

Model and Task Scope. Although we evaluate a diverse set of representative DR agents, the model landscape continues to evolve rapidly. In addition, non-recurrent tasks are currently formulated as binary predictions, whereas real-world forecasting often involves richer outcome spaces. Future iterations will expand both model coverage and task formulations.

References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 67–93.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*.

Yuemin Chen, Feifan Wu, , and et al. 2024. Knowledge-augmented financial market analysis and report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1207–1217.

Dave Citron. 2025. Deep research is now available on gemini 2.5 pro experimental. Accessed: 2025.

Tianyu Fan, Yuhao Yang, and et al. 2025. Ai-trader: Benchmarking autonomous agents in real-time financial markets. *Preprint*, arXiv:2512.10971.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 261–269.

Dong Huang, Jie M Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. 2023. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010*.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, and et al. 2025. Deep research agents: A systematic examination and roadmap. *Preprint*, arXiv:2506.18096.

Pranab Islam, Anand Kannappan, Douwe Kiber, Zachary Walters, Scott Kantor, Tom Sun, and Nils Holmes. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. 2024. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839*.

Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-10-07.

Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4304–4315.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2024. Financial forecasting from textual and tabular time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8289–8300.

Changlun Li, Yao Shi, Yuyu Luo, and Nan Tang. 2025. Will llms be professional at fund investment? deepfund: A live arena perspective. *Preprint*, arXiv:2503.18313.

Xiao Liu, Hao Yu, Hanchen Zhang, and et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

OpenAI. 2025. Introducing deep research. Accessed: 2025.

OpenAI Team. 2025. Introducing deep research. https://openai.com/index/introducing-deep-research/. Accessed: 2025-10-07.

Perplexity. 2025. Introducing perplexity deep research. Accessed: 2025.

Lingfei Qian, Xueqing Peng, Yan Wang, and et al. 2025. When agents trade: Live multi-market trading benchmark for llm agents. *Preprint*, arXiv:2510.11695.

Qwen. 2025. Deep research (qwen-deep-research). Accessed: 2025.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, and et al. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Zhengliang Shi, Yiqun Chen, Haitao Li, and et al. 2025. Deep research: A systematic survey. *arXiv preprint arXiv:2512.02038*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Dong Shu, Yanguang Liu, Huopu Zhang, and Mengnan Du. 2025. Fincall-surprise: A large scale multimodal benchmark for earning surprise prediction. *arXiv preprint arXiv:2510.03965*.

S&P Global Market Intelligence. 2024. S&P Capital IQ Key Developments: Data Methodology and Event Taxonomy. https://www.marketplace.spglobal.com/en/datasets/key-developments-(15). Proprietary database. Full taxonomy and schema documentation require subscription.

Claude Team. 2025a. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2025-10-07.

DeepSeek Team. 2025b. Introducing deepseek-v3.2-exp. https://api-docs.deepseek.com/news/news250929. Accessed: 2025-10-07.

OpenAI Team. 2025c. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-10-07.

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

DeepResearch Tongyi, Baixuan Li, , and et al. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Tongyi Team. 2025. Tongyi deepresearch: A new era of open-source ai researchers. https://tongyi-agent.github.io/blog/introducing-tongyi-deep-research/. Accessed: 2025-10-07.

Colin White, Manley Dooley, and et al. 2025. Livebench: A challenging, contamination-limited llm benchmark. In *Proceedings of the International Conference on Learning Representations*. Spotlight Paper.

xAI Team. 2025. Grok 4. https://x.ai/news/grok-4. Accessed: 2025-10-07.

Qianqian Xie, Weiguang Han, and et al. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. In *Advances in Neural Information Processing Systems*.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2024. Finben: A holistic financial benchmark for large language models. In *Advances in Neural Information Processing Systems*.

Cheng Xu, Shuhao Guan, Yuan Li, Wei Jia, Rui Wang, Hanyu Yan, and Hongxin Zhang. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations*.

Haofei Yu, Fenghai Li, and Jiaxuan You. 2025. Livetradebench: Seeking real-world alpha with large language models. *Preprint*, arXiv:2511.03628.

Zhiyuan Zeng, Jiashuo Liu, and et al. 2025. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*.

Linghao Zhang, Shilin He, and et al. 2025a. Swe-bench goes live! *arXiv preprint arXiv:2505.23419*.

Yang Zhang, Wenbo Yang, Jun Wang, Qiang Ma, and Jie Xiong. 2025b. Camef: Causal-augmented multimodality event-driven financial forecasting by integrating time series patterns and salient macroeconomic announcements. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3867–3878.

Shuyan Zhou, Frank F Xu, and et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Fengbin Zhu, Xiang Yao Ng, Ziyang Liu, and et al. 2025a. Findeepresearch: Evaluating deep research agents in rigorous financial analysis. *Preprint*, arXiv:2510.13936.

Zhuohang Zhu, Haodong Chen, Qiang Qu, and Vera Chung. 2025b. Fincast: A foundation model for financial time-series forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4539–4549.

A Recurrent Task Specifications

A.1 Macro Indicators

We monitor 96 macro indicators, derived from 14 indicator types that are instantiated across eight economies and complemented by two global market indices, as detailed in Table 4. The selection criteria capture the four fundamental pillars of macro analysis: real economic activity (e.g., GDP, Unemployment), price stability (e.g., CPI, PPI), monetary conditions (e.g., Interest Rates, Stock Index), and external balance (e.g., FX Rate, CAB). Crucially, we augment these with global barometers—specifically Commodities and VIX—to test the model's sensitivity to cross-border supply shocks and systemic risk sentiment. Accurately forecasting these indicators requires financial experts to conduct extensive information gathering and multi-step reasoning, making them ideal proxies for evaluating deep research capabilities.

Table 4: Macro indicators for recurrent tasks.

No.	Indicator	Description	Economies
		Global Indices (2)	
1	S&P GSCI Commodity	Global commodity price index	Global
2	CBOE VIX	Market volatility index	Global
		Economy-Specific Indicators (94)	
3	Stock Index	Major equity market index	All 8
4	Interest Rate (1yr)	1-year govt bond yield	Excl. SG
5	Interest Rate (3m)	3-month treasury bill rate	All 8
6	FX Rate	Exchange rate against USD	Excl. US
7	GDP	Gross Domestic Product	All 8
8	CPI	Consumer Price Index	All 8
9	PPI	Producer Price Index	All 8
10	UNRATE	Unemployment Rate	All 8
11	HPI	House Price Index	All 8
12	NEER	Nominal Effective Exchange Rate	All 8
13	Interbank Rate	3-month interbank lending rate	All 8
14	CAB	Current Account Balance	All 8

Economies: US, CN, HK, JP, UK, DE, FR, SG
Note: US excludes FX Rate; SG excludes Interest Rate (1yr)

A.2 Corporate Financial Metrics

We cover 121 corporate financial metrics organized into 9 categories, as shown in Table 5 and Table 6. This extensive selection mirrors the comprehensive framework used in professional fundamental analysis. By encompassing the three primary financial statements (Balance Sheet, Income Statement, Cash Flow) and six categories of derived ratios, we require the model to not only retrieve raw data but also perform arithmetic reasoning to assess liquidity, solvency, and operational efficiency. Accurately forecasting these metrics typically requires financial experts to conduct detailed information gathering and multi-step analytical reasoning, providing a rigorous test of deep research capabilities.

B Non-Recurrent Task Specifications

B.1 Non-Recurrent Macro Event Taxonomy

To categorize non-recurrent macroeconomic shocks, we adopted a **"Stable Core, Adaptive Interface"** design philosophy. This hierarchical framework consists of a fixed semantic layer (Level 1) ensuring consistency, and a dynamic grounding layer (Level 2) ensuring relevance.

Level 1: The Stable Semantic Taxonomy (Immutable Layer). The first level defines a standardized ontology of macro-financial events designed to remain invariant across time and regions. We categorize events into 9 categories (A to I) and 26 subcategories (Table 7). This taxonomy follows the *Mutually Exclusive and Collectively Exhaustive (MECE)* principle. By keeping this semantic layer static, we ensure that model performance remains comparable across different eras, providing a consistent benchmark for longitudinal evaluation.

Level 2: The Economy-Specific Grounding (Adaptive Layer). Unlike the static Level 1, the grounding layer is designed to be dynamic and extensible. This layer maps the universal concepts to specific, falsifiable market indicators for each economy.

Our expert panel designed this layer with two degrees of flexibility to accommodate the evolving nature of financial markets:

1. **Dynamic Calibration:** The quantifiable triggers (e.g., specific basis point thresholds) are subject to periodic recalibration. As market regimes shift (e.g., from a low-interest environment to a high-inflation era), these parameters can be updated to maintain their discriminatory power without altering the Level 1 definitions.

2. **Extensibility for Future Tasks:** The framework supports the seamless integration of new economies or additional event types. Future iterations of the benchmark can introduce new task subcategories or expand to emerging markets by simply defining the corresponding Level 2 grounding logic, preserving the integrity of the overarching taxonomy.

For the current version, we defined the "Ground Truth" for 8 major economies. For each economy-subcategory pair, we established **Authoritative**

Table 5: Corporate financial metrics - Part 1 (Balance Sheet, Income Statement, Cash Flow).

No.	Metric	Description	No.	Metric	Description
Balance Sheet Items (25 metrics)					
1	Total Assets	Total value of all assets	14	Accounts Payable	Amounts owed to suppliers
2	Total Liabilities	Total liabilities owed	15	Accrued Expenses	Expenses incurred but not paid
3	Total Equity	Total shareholders' equity	16	Deferred Revenue	Revenue received but not earned
4	Total Current Assets	Assets convertible within 1 yr	17	Retained Earnings	Accumulated undistributed income
5	Total Current Liabilities	Liabilities due within 1 yr	18	Treasury Stock	Repurchased company shares
6	Long Term Debt	Debt due beyond 1 year	19	Minority Interest	Non-controlling interest
7	Short Term Debt	Debt due within 1 year	20	Preferred Stock	Preferential dividend equity
8	Short and Long Term Debt	Combined total debt	21	Common Stock	Basic ownership shares
9	Total Loans	Loans held by financials	22	Total Deposits	Customer deposits (financials)
10	Cash and Equivalents	Liquid assets and cash	23	Saving Deposits	Savings account deposits
11	Accounts Receivable	Amounts owed by customers	24	Property Plant & Equip.	Physical asset value
12	Inventory	Goods held for sale	25	Intangible Assets	Patents, goodwill, etc.
13	Goodwill	Acquisition premium paid			
Income Statement Items (22 metrics)					
26	Revenue	Total sales and income	37	Interest Income	Income from interest
27	Cost of Revenue	Direct costs of goods sold	38	Other Income	Non-operating income
28	Gross Profit	Revenue minus cost of revenue	39	Extraordinary Items	Unusual gains or losses
29	Operating Income	Profit from core operations	40	Discontinued Operations	Results from closed segments
30	EBIT	Earnings before interest & taxes	41	EPS (Basic)	Net income per basic share
31	EBITDA	EBIT plus depreciation & amort.	42	EPS (Diluted)	Net income per diluted share
32	Net Income	Final profit after all expenses	43	Dividends Per Share	Dividends paid per share
33	Interest Expense	Interest paid on debt	44	Revenue Growth (YoY)	Year-over-year revenue change
34	R&D Expense	Research & development costs	45	Net Income Growth (YoY)	Year-over-year income change
35	SG&A Expense	Selling, general & admin costs	46	Operating Expense	Total operating costs
36	Income Tax Expense	Taxes on corporate income	47	Pre-Tax Income	Income before tax expense
Cash Flow Items (15 metrics)					
48	Cash From Operations	Net cash from operating	56	Debt Repayment	Cash used to repay debt
49	Cash from Investing	Net cash from investing	57	Debt Issuance	Cash from issuing debt
50	Cash from Financing	Net cash from financing	58	Stock Repurchase	Cash for share buybacks
51	Free Cash Flow	Operating cash minus capex	59	Stock Issuance	Cash from issuing shares
52	Depreciation & Amort.	Non-cash asset reduction	60	Dividend Payments	Cash paid as dividends
53	Capital Expenditure	Investment in fixed assets	61	Net Change in Cash	Total cash position change
54	Acquisitions	Cash for acquiring companies	62	Working Capital Changes	Op. asset/liability changes
55	Divestitures	Cash from selling units			

Sources which strictly designated official sources (e.g., FOMC, PBoC, OBR), and **Quantifiable Triggers** with rigid quantitative thresholds (e.g., ≥ 25 bps rate hike, $> 1\%$ GDP fiscal impulse) tailored to local market structures (e.g., "Shunto" for Japan, "Schuldenbremse" for Germany). The detailed grounding tables are presented in Tables 8 through 15. This scientific design ensures that the benchmark remains a "living" evaluation standard, capable of evolving alongside the real-world financial landscape.

B.2 Non-recurrent Corporate Event Types

To ensure our taxonomy is grounded in established financial industry standards rather than subjective definitions, we adopted the *Key Developments* framework from S&P Global Market Intelligence (2024). This database is widely recognized by professional analysts for tracking material corporate episodes and serves as the "gold standard" for event categorization.

While the original S&P schema comprises over 100 granular event types, we curated a streamlined subset of 70 event types (Table 16) specifically optimized for the LLM forecasting task. Our selection process was guided by two principles: (1) **Materiality and Signal Value:** We filtered out low-impact administrative events (e.g., routine conference presentations, minor website updates) that typically do not trigger significant market reactions or require complex reasoning. (2) **Semantic Distinctiveness:** We consolidated highly overlapping categories to create a MECE framework suitable for classification tasks.

This curation ensures that the benchmark evaluates the model's ability to anticipate "market-moving" events rather than getting distracted by administrative noise.

C Non-Recurrent Task Generation Pipeline

The Non-Recurrent Task Generation Agent employs a three-stage LLM-based pipeline. In *Signal*

Table 6: Corporate financial metrics - Part 2 (Profitability, Liquidity, Leverage, Efficiency, Coverage, Valuation).

No.	Metric	Description	No.	Metric	Description
Profitability Ratios (15 metrics)					
63	Return on Assets (ROA)	Net Income / Total Assets	71	Return on Sales	Op. Income / Revenue
64	Return on Equity (ROE)	Net Income / Total Equity	72	Cash Return on Assets	Op. Cash Flow / Total Assets
65	Return on Invested Capital	NOPAT / Invested Capital	73	Cash Return on Equity	Op. Cash Flow / Total Equity
66	Gross Margin	Gross Profit / Revenue	74	NPL Ratio	Non-Performing Loans / Loans
67	Operating Margin	Op. Income / Revenue	75	Net Interest Margin	Net Int. Inc. / Earning Assets
68	EBITDA Margin	EBITDA / Revenue	76	Efficiency Ratio	Non-Int. Exp. / Revenue
69	Net Margin	Net Income / Revenue	77	Cost-to-Income Ratio	Op. Exp. / Op. Income
70	Profit Margin	(Op. Inc. − D&A) / Rev.			
Liquidity Ratios (8 metrics)					
78	Current Ratio	Curr. Assets / Curr. Liab.	82	Working Capital	Curr. Assets − Curr. Liab.
79	Quick Ratio	(Curr. Assets − Inv.) / Curr. Liab.	83	Working Capital Ratio	Working Capital / Total Assets
80	Cash Ratio	Cash & Equiv. / Curr. Liab.	84	Defensive Interval Ratio	Liquid Assets / Daily Op. Exp.
81	Op. Cash Flow Ratio	Op. Cash Flow / Curr. Liab.	85	Cash Conversion Cycle	DIO + DSO − DPO
Leverage Ratios (12 metrics)					
86	Debt-to-Equity Ratio	Total Debt / Total Equity	92	Long-term Debt to Assets	LT Debt / Total Assets
87	Debt-to-Assets Ratio	Total Debt / Total Assets	93	ST Debt to Total Debt	ST Debt / Total Debt
88	Liability-to-Assets Ratio	Total Liab. / Total Assets	94	Net Debt	Total Debt − Cash & Equiv.
89	Equity Ratio	Total Equity / Total Assets	95	Net Debt to Equity	Net Debt / Total Equity
90	Equity Multiplier	Total Assets / Total Equity	96	Net Debt to EBITDA	Net Debt / EBITDA
91	Long-term Debt to Equity	LT Debt / Total Equity	97	Financial Leverage	Avg. Assets / Avg. Equity
Efficiency Ratios (12 metrics)					
98	Asset Turnover	Revenue / Total Assets	104	Equity Turnover	Revenue / Total Equity
99	Fixed Asset Turnover	Revenue / Fixed Assets	105	Days Inventory Outstanding	365 / Inventory Turnover
100	Inventory Turnover	COGS / Avg. Inventory	106	Days Sales Outstanding	365 / Receivables Turnover
101	Receivables Turnover	Revenue / Avg. Receivables	107	Days Payables Outstanding	365 / Payables Turnover
102	Payables Turnover	COGS / Avg. Payables	108	Total Loans Growth (YoY)	YoY change in total loans
103	Working Capital Turnover	Revenue / Working Capital	109	Deposits Growth (YoY)	YoY change in deposits
Coverage Ratios (6 metrics)					
110	Interest Coverage (EBIT)	EBIT / Interest Expense	113	Interest Coverage (Op. Inc.)	Op. Inc. / Interest Exp.
111	Interest Coverage (EBITDA)	EBITDA / Interest Expense	114	Debt Service Coverage	Op. Inc. / Debt Service
112	Interest Coverage (Net Inc.)	Net Income / Interest Exp.	115	Fixed Charge Coverage	(EBIT+Lease) / (Int.+Lease)
Valuation & Market Metrics (6 metrics)					
116	Book Value Per Share	Total Equity / Shares Out.	119	Cash Flow Per Share	Op. Cash Flow / Shares Out.
117	Tangible Book Value/Share	Tangible Equity / Shares Out.	120	Enterprise Value	Market Cap + Debt − Cash
118	Revenue Per Share	Revenue / Shares Out.	121	Market Capitalization	Price × Shares Outstanding

Detection, search-augmented LLMs scan real-time news streams to identify candidate events with binary outcomes and genuine uncertainty, prioritizing policy announcements, regulatory rulings, and personnel changes while excluding routine scheduled releases. In *Relevance Assessment*, sub-agents investigate each candidate by analyzing specific unknowns (e.g., debated thresholds, contested wording), conflicting signals across sources, and risk factors such as potential delays, assigning uncertainty scores based on source disagreement. In *Question Generation*, the agent synthesizes research reports to formulate questions about *marginal content* rather than mere occurrence with automated verification ensuring binary answerability and genuine uncertainty.

D Ground Truth Acquisition for Non-Recurrent Tasks

Ground truth for non-recurrent tasks combines automated evidence aggregation with expert validation. A search-augmented LLM agent executes a three-phase protocol: (1) searching official sources and major financial news, (2) verifying numeric thresholds and event definitions, and (3) checking for denials or contradictory evidence. An LLM adjudicator synthesizes findings into a binary determination, classifying events as `YES` only when confirmed by authoritative sources; ambiguous signals or absence of evidence result in `NO`. To ensure gold-standard quality, **all (100%)** automated proposals undergo mandatory expert review. Specialists verify evidence against official sources for final adjudication, achieving 95% inter-annotator agreement.

Table 7: Taxonomy of Non-Recurrent Macro Events.

Category	Code & Detailed Description
A. Monetary & Financial Conditions (3 types)	**A1 Monetary Policy Shift**: Central bank policy rate hikes/cuts, policy stance changes, quantitative easing or tapering decisions.
	A2 Financial Market Liquidity Shock: Bond- or money-market stress, funding squeezes, repo-market dislocations, impaired market-making.
	A3 Macro-prudential Regulation Change: Changes to macro-prudential tools such as LTV/DTI limits, countercyclical capital buffers, or leverage caps.
B. Fiscal Policy & Public Finance (2 types)	**B1 Fiscal Stimulus / Austerity**: Government budget decisions that significantly expand or contract spending, transfer programs, or tax burdens.
	B2 Sovereign Debt Stress: Events indicating sovereign credit stress, including rating downgrades, refinancing pressure, or default risk.
C. Trade & External Sector (3 types)	**C1 Trade Policy Change / Sanctions / Tariffs**: Introduction or removal of tariffs, quotas, export controls, sanctions, or anti-dumping/countervailing duties.
	C2 Currency / FX Pressure Shock: Sharp exchange-rate moves, reserve losses, or capital outflows indicating FX market pressure.
	C3 External Financing / Current-Account Shock: Stress related to external financing, current-account imbalances, sudden stops, or debt rollover risks.
D. Commodity, Energy & Supply Chain (3 types)	**D1 Energy Price Shock**: Large and rapid changes in energy prices (oil, gas, electricity) affecting production costs and inflation.
	D2 Commodity Price Shock: Significant volatility in key non-energy commodities such as metals, food, or agricultural inputs.
	D3 Global Supply Chain Disruption: Logistics bottlenecks, shipping disruptions, or trade chokepoints that impair global supply chains.
H. Labour Market & Household Sector (2 types)	**H1 Labour Market Shock**: Shocks to employment, unemployment, labour-force participation, or wage dynamics.
	H2 Household Income / Consumption Stress: Stress in household balance sheets, including real income declines, debt distress, or demand weakness.
E. Real Economy Activity (4 types)	**E1 Industrial Production / Manufacturing Shock**: Shocks to industrial production or manufacturing activity, including sharp contractions or surges.
	E2 Retail / Consumption / Services Shock: Shocks to household consumption, retail sales, or services-sector activity driven by income or sentiment changes.
	E3 Housing / Real Estate Cycle Shock: Downturns or booms in property markets, construction activity, or related policy changes.
	E4 Technology, Digital Economy & AI-Driven Industrial Activity: Real-economy impacts arising from major technology/AI developments, adoption waves, or semiconductor constraints.
F. Financial Stability & Credit Cycle (3 types)	**F1 Credit Cycle Shift (Boom/Bust)**: Rapid expansions or contractions in private credit to households or corporates.
	F2 Banking System Stress / NPL Shock: Deterioration in bank asset quality, rising non-performing loans, or liquidity/solvency concerns.
	F3 Asset Price Shock (Equity/Bond/Volatility): Sharp corrections in equity or bond markets, volatility spikes, or broad market repricing events.
G. Structural & Regulatory Policy (3 types)	**G1 Climate / Carbon / ESG Policy**: Policy changes related to climate targets, carbon pricing, emissions trading, or ESG disclosure rules.
	G2 Tech/Data/Privacy Regulation: New or revised regulations on data privacy, cybersecurity, cross-border data flows, or digital governance.
	G3 Structural / Institutional Reform: Reforms to labour markets, pensions, social security, legal or institutional frameworks.
I. Geopolitical & Systemic Shocks (3 types)	**I1 Conflict / Sanctions Shock**: Military conflicts, geopolitical escalation, or sanctions regimes with macro/sectoral impact.
	I2 Natural Disaster / Pandemic Shock: Major natural disasters or health crises that disrupt economic activity.
	I3 Global Financial Contagion: Spillovers from global financial crises, cross-border banking stress, or systemic liquidity shocks.

Table 8: Event grounding standards for the United States (US) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	Federal Reserve (FOMC)	Federal Funds Target Range upper bound changes by ≥ 25 bps; OR official statement explicitly pivots stance.
A2	Fed H.4.1 / NY Fed	FRA-OIS spread > 95th percentile; OR Reverse Repo Facility usage surges $>$ \$500B in a week.
A3	Fed Board / FDIC	Implementation of new capital rules (e.g., Basel III Endgame) or change in CCAR stress test scenarios.
B. Fiscal Policy & Public Finance		
B1	CBO / White House	Passage of legislation (e.g., CARES Act, IRA) with discretionary spending impact $\geq 1\%$ of GDP.
B2	Treasury / CDS Market	US Sovereign CDS (5Y) spread > 50 bps; OR 'Extraordinary Measures' exhausted date approaches within 30 days.
C. Trade & External Sector		
C1	USTR / Dept. of Commerce	Implementation of new Section 301 tariffs, or export controls (Entity List) affecting key sectors.
C2	Treasury / Fed	Trade-weighted US Dollar Index (DXY) moves $\geq 10\%$ within 3 months.
C3	BEA	Current Account Deficit widens by $> 2\%$ of GDP YoY; OR net foreign capital outflows exceed historical 2σ.
D. Commodity, Energy & Supply Chain		
D1	EIA (Energy Info.)	WTI Crude or Henry Hub Natural Gas spot prices change $\geq 30\%$ over 6 months.
D2	USDA / USGS	Key agricultural or metal commodity prices deviate $\geq 25\%$ from 6-month moving average.
D3	Fed NY / Census	Global Supply Chain Pressure Index (GSCPI) exceeds 2 standard deviations.
E. Real Economy Activity		
E1	Fed Board (G.17)	Industrial Production Index contracts $\geq 3\%$ YoY for 2 consecutive months.
E2	Census Bureau	Retail Sales (ex-auto) contract $\geq 2\%$ YoY; OR Univ. of Michigan Consumer Sentiment drops to bottom 10%.
E3	FHFA / S&P CoreLogic	Case-Shiller National Home Price Index turns negative YoY; OR Housing Starts drop $\geq 20\%$ YoY.
E4	BEA / Congress	Tech sector value-add deviates $\geq 2\sigma$ from trend; OR passage of major industrial policy (e.g., CHIPS Act).
F. Financial Stability & Credit Cycle		
F1	Fed Board / BIS	Private non-financial sector credit-to-GDP gap exceeds +10% (Boom) or drops below -5% (Bust).
F2	FDIC / Fed	NPL ratio for insured institutions rises $\geq 1.0\%$; OR failure/rescue of a SIFI bank.
F3	NYSE / Nasdaq	S&P 500 or Nasdaq 100 enters Technical Bear Market (drawdown $\geq 20\%$ from peak).
G. Structural & Regulatory Policy		
G1	EPA / Congress	Passage of major climate legislation (e.g., IRA subsidies); OR new SEC climate disclosure mandates.
G2	FTC / FCC	Major antitrust lawsuit filed against Big Tech; OR new federal data privacy executive orders.
G3	Congress	Enactment of major reforms to Social Security, Medicare, or Immigration laws.
H. Labour Market & Household Sector		
H1	BLS	Unemployment Rate changes $\geq 0.5\%$ (Sahm Rule); OR Non-farm Payrolls deviate > 50k from consensus.
H2	BEA	Real Disposable Personal Income contracts $\geq 2\%$ YoY.
I. Geopolitical & Systemic Shocks		
I1	Dept. of State / OFAC	US becomes party to armed conflict; OR designation of major sanctions on a G20 economy.
I2	FEMA / CDC	Presidential Disaster Declaration for event costing $>$ \$10B; OR Nationwide Public Health Emergency declaration.
I3	Treasury / Fed	VIX Index > 35 combined with net foreign selling of US Treasuries.

Table 9: Event grounding standards for the China (CN) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	PBoC (Central Bank)	7-day Reverse Repo or 1-year MLF rate changes by ≥ 5 bps; OR Reserve Requirement Ratio (RRR) cut ≥ 25 bps.
A2	CFETS / NIFC	DR007 (7-day interbank repo rate) deviates > 50 bps from policy rate for 5+ days; OR PBoC net liquidity injection CNY500B/week.
A3	PBoC / NFRA	Adjustment of Macro-Prudential Assessment (MPA) parameters; OR changes to property sector "Three Red Lines" metrics.
B. Fiscal Policy & Public Finance		
B1	State Council / MOF	Issuance of Ultra-long Special Sovereign Bonds; OR Local Government Special Bond quota increase $>$ ¥1 Trillion.
B2	MOF / Market Data	10-year China Government Bond (CGB) yield spikes ≥ 20 bps in a month; OR major LGFV bond default event.
C. Trade & External Sector		
C1	MOFCOM / Customs	Implementation of export controls on strategic materials (e.g., Gallium/Germanium); OR new tariffs on major trading partners.
C2	SAFE / PBoC	USD/CNY Daily Fixing deviates from market close by > 500 pips (Counter-cyclical factor usage); OR FX Reserves drop $>$ \$50B/month.
C3	SAFE	Capital Account net outflows exceed 2% of GDP (annualized); OR major restrictions on cross-border capital flows.
D. Commodity, Energy & Supply Chain		
D1	NDRC / NEA	NDRC adjusts guided retail fuel prices; OR thermal coal spot price exceeds regulatory price cap range.
D2	DCE / SHFE	Domestic futures prices for Iron Ore or Rebar deviate $\geq 20\%$ from 6-month MA.
D3	MOT / Caixin	Caixin Manufacturing PMI Suppliers' Delivery Times sub-index drops below 45.0.
E. Real Economy Activity		
E1	NBS / Caixin	Official Manufacturing PMI or Caixin PMI contracts (<50.0) for 2 consecutive months.
E2	NBS	Retail Sales of Consumer Goods YoY growth turns negative; OR Youth Unemployment Rate (16-24) exceeds 20%.
E3	NBS / MOHURD	70-City New Home Price Index declines YoY; OR Top-100 Developer Sales value drops $\geq 20\%$ YoY.
E4	MIIT / NDRC	Launch of major strategic projects (e.g., "East Data West Computing"); OR Strategic Emerging Industries value-add $\geq 2\sigma$ vs trend.
F. Financial Stability & Credit Cycle		
F1	PBoC	Total Social Financing (TSF) growth rate gap vs Nominal GDP growth $\geq \pm 5\%$.
F2	NFRA / PBoC	Takeover/Resolution of a medium-sized bank (e.g., Baoshang style event); OR Commercial Bank NPL ratio rises $\geq 0.5\%$.
F3	SSE / SZSE	CSI 300 Index experiences a rapid drawdown $\geq 20\%$ (Technical Bear Market) or triggers trading curbs.
G. Structural & Regulatory Policy		
G1	NDRC / MEE	Issuance of "Dual Carbon" (1+N) policy documents; OR launch of new National Carbon Market trading rules.
G2	CAC / SAMR	New anti-monopoly penalties on platform economy firms; OR CAC initiates cybersecurity review on major data handlers.
G3	CPC Central Comm.	"Third Plenum" or "Two Sessions" announces major reforms (e.g., Hukou reform, Common Prosperity initiatives).
H. Labour Market & Household Sector		
H1	NBS	Surveyed Urban Unemployment Rate rises $\geq 0.5\%$; OR Migrant Worker population contracts YoY.
H2	NBS / PBoC	Household deposits surge CNY5 Trillion YoY (Excess Savings); OR Household Leverage Ratio declines (Deleveraging).
I. Geopolitical & Systemic Shocks		
I1	MFA / CMC	Escalation of tensions in Taiwan Strait or South China Sea triggering military exercises; OR foreign sanctions on Chinese entities.
I2	NHC / MEM	Activation of Level-I Public Health Emergency Response; OR natural disaster affecting $> 1\%$ of national arable land.
I3	PBoC	Stock Connect / Bond Connect net outflows exceed historical 99th percentile.

Table 10: Event grounding standards for the Hong Kong (HK) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	HKMA	Base Rate adjustment following the US Fed (automatic under LERS); OR HKMA issues specific circular on credit tightening.
A2	HKMA / HIBOR Fixing	Aggregate Balance falls below HKD 50 Billion; OR 1-month HIBOR-LIBOR spread widens $> \pm 100$ bps (Capital Flight/Inflow pressure).
A3	HKMA	Adjustment of Countercyclical Capital Buffer (CCyB); OR changes to Max LTV ratios for property mortgages.
B. Fiscal Policy & Public Finance		
B1	Financial Secretary	Annual Budget Speech announces major "Sweeteners" (e.g., Consumption Vouchers, Cash Payouts >HKD 5,000/person).
B2	LegCo / Treasury	Fiscal Reserves drop below 12 months of government expenditure; OR issuance of Government Green Bonds.
C. Trade & External Sector		
C1	C&SD / TDC	Re-exports value contracts $\geq 10\%$ YoY (signaling global trade slowdown); OR new sanctions on HK shipping/logistics status.
C2	HKMA	Triggering of Strong (7.75) or Weak (7.85) Side Convertibility Undertaking (HKMA intervention magnitude >HKD 10B).
C3	HKMA	Net capital outflows exceed 5% of Monetary Base; OR sharp drop in FX assets.
D. Commodity, Energy & Supply Chain		
D1	C&SD / CLP	Scheme of Control electricity tariff adjustment $\geq 5\%$; OR significant spike in imported fuel costs.
D2	C&SD	Unit Value Index of Imports (Food/Raw Materials) rises $\geq 10\%$ YoY.
D3	Airport Authority	Air Cargo Throughput contracts $\geq 10\%$ YoY; OR Kwai Tsing Container Terminals throughput drops significantly.
E. Real Economy Activity		
E1	C&SD	Purchasing Managers' Index (PMI) remains below 50.0 for 3 consecutive months.
E2	C&SD	Total Retail Sales Value contracts $\geq 10\%$ YoY (Tourism shock); OR Restaurant Receipts drop $\geq 10\%$ YoY.
E3	Centaline / RVD	Centa-City Leading Index (CCL) draws down $\geq 5\%$ from peak; OR Official Housing Price Index turns negative YoY.
E4	ITB / HKSTP	Launch of major I&T initiatives (e.g., San Tin Technopole funding); OR new Virtual Asset Trading Platform regulations.
F. Financial Stability & Credit Cycle		
F1	HKMA	Loan-to-Deposit Ratio rises significantly; OR annualized loan growth turns negative (Credit Crunch).
F2	HKMA	Classified Loan Ratio (NPL) rises $> 1.0\%$; OR capital adequacy ratio of major local banks drops sharply.
F3	HSI / HKEX	Hang Seng Index (HSI) or Hang Seng Tech Index drops $\geq 20\%$ (Bear Market); OR IPO fundraising dries up (zero IPOs in a month).
G. Structural & Regulatory Policy		
G1	HKEX / SFC	Implementation of mandatory ESG/Climate disclosures for listed companies (IPO/listing rules update).
G2	PCPD / OGCIO	Major data breach enforcement action; OR new cross-boundary data flow regulations (Greater Bay Area standard contract).
G3	Chief Executive	Policy Address announces major structural shifts (e.g., Northern Metropolis development strategy).
H. Labour Market & Household Sector		
H1	C&SD	Seasonally Adjusted Unemployment Rate rises $\geq 0.5\%$; OR Underemployment Rate spikes.
H2	C&SD	Median Monthly Household Income contracts YoY; OR Negative Equity Mortgage cases exceed 1,000 cases.
I. Geopolitical & Systemic Shocks		
I1	Govt / Foreign Ministry	Imposition of foreign sanctions on HK officials/entities; OR major changes to HK's special trade status.
I2	CHP / Govt	Activation of "Emergency Response Level" for infectious disease; OR Typhoon Signal No. 8+ halts market trading for > 1 day.
I3	HKMA	Interbank liquidity dries up despite Base Rate cuts; OR simultaneous drop in HSI and HKD (Twin Deficit risk).

Table 11: Event grounding standards for the Japan (JP) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	Bank of Japan (BoJ)	Change in Policy Rate (Uncollateralized Call Rate) ≥ 10 bps; OR Modification of Yield Curve Control (YCC) band (e.g., widening band).
A2	BoJ / JSDA	10-year JGB yield breaches the upper limit of the reference range; OR "Rinban" (JGB purchase) operations increase significantly.
A3	BoJ / JFSA	Changes to ETF/J-REIT purchase program guidelines; OR Macro-prudential measures on regional bank real estate lending.
B. Fiscal Policy & Public Finance		
B1	Cabinet Office / MoF	Approval of a "Supplementary Budget" (Hosei Yosan) with spending $>$ ¥10 Trillion; OR new economic package announcement.
B2	MoF	JGB Debt Service Cost rises significantly in budget projections; OR Sovereign Rating outlook downgrade due to debt-to-GDP ratio.
C. Trade & External Sector		
C1	METI	Imposition of export restrictions on strategic tech materials (e.g., photoresists); OR removal from "White List" of trade partners.
C2	MoF / BoJ	Official FX Intervention confirmed by MoF (buying JPY/selling USD); OR USD/JPY moves $\geq 3\%$ in a single week.
C3	MoF	Current Account Surplus narrows significantly or turns to deficit (due to energy import costs).
D. Commodity, Energy & Supply Chain		
D1	METI / TEPCO	Reactivation of Nuclear Power Plants approved; OR Utility companies apply for electricity rate hike $> 10\%$.
D2	MAFF	"Food Price Index" within CPI rises $\geq 5\%$ YoY; OR government subsidies for gasoline/wheat prices triggered.
D3	METI / Toyota	Major automaker halts production due to parts shortage; OR disruption in semiconductor supply chain (e.g., Kumamoto fab).
E. Real Economy Activity		
E1	BoJ (Tankan)	**Tankan Large Manufacturers DI** drops by ≥ 5 points; OR Industrial Production contracts $\geq 2\%$ MoM.
E2	Cabinet Office	GDP (Annualized Real Growth) contracts for 2 consecutive quarters (Technical Recession); OR Consumer Confidence Index drops.
E3	MLIT	Land Price Publication (Chika Koji) shows YoY decline in major metropolitan areas; OR Condo prices in Tokyo enter correction.
E4	METI	Announcement of subsidies for strategic sectors (e.g., Rapidus semiconductor project); OR AI strategy guidelines release.
F. Financial Stability & Credit Cycle		
F1	BoJ	Bank Lending YoY growth deviates significantly from trend; OR Corporate bankruptcy liabilities surge (Teikoku Databank).
F2	JFSA / BoJ	Regional Bank (Chigin) merger or recapitalization prompted by FSA; OR surfacing of large losses in securities portfolios (e.g., CLOs).
F3	TSE / JPX	Nikkei 225 or TOPIX drops $\geq 20\%$ from peak; OR Volatility Index (JNIV) spikes > 30.
G. Structural & Regulatory Policy		
G1	METI / MoE	GX (Green Transformation) Promotion Act implementation; OR Carbon Pricing (GX League) introduction.
G2	PPC / METI	Enforcement of stricter personal data protection rules; OR new regulations on Generative AI copyright.
G3	Cabinet Office	"New Capitalism" policy initiatives launched; OR major revisions to Labor Standards Act.
H. Labour Market & Household Sector		
H1	Rengo / MHLW	**"Shunto" (Spring Wage Offensive)** agreed wage hike exceeds 3% (or BoJ target level); OR Active Job Openings-to-Applicants Ratio drops.
H2	MHLW / MIC	Real Cash Earnings contract YoY (Wage-Price spiral failure); OR Household Spending (Kakei Chosa) drops YoY.
I. Geopolitical & Systemic Shocks		
I1	MoFA / MoD	Major security incidents near Senkaku Islands; OR North Korean missile launch triggering J-Alert system impacting markets.
I2	Cabinet Office	Nankai Trough Earthquake warning issued; OR natural disaster damage estimate $>$ ¥1 Trillion.
I3	BoJ	"Japan Premium" re-emerges in offshore funding markets; OR massive unwinding of Yen Carry Trade.

Table 12: Event grounding standards for the United Kingdom (UK) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	BoE (MPC)	Bank Rate change ≥ 25 bps; OR MPC Vote Split changes significantly (e.g., from 6-3 to 5-4) signaling pivot; OR Active Gilt Sales (QT).
A2	BoE / SONIA	SONIA-Bank Rate spread widens > 20 bps; OR failure in Gilt Repo market liquidity (Repo rate dislocation).
A3	BoE (FPC)	Adjustment of Countercyclical Capital Buffer (CCyB) rate; OR intervention in LDI (Liability-Driven Investment) fund leverage rules.
B. Fiscal Policy & Public Finance		
B1	HM Treasury / OBR	"Autumn Budget" or "Spring Statement" announces discretionary measures $>$ £15B; OR OBR issues warning on fiscal sustainability.
B2	DMO / Markets	10-year Gilt yield spikes ≥ 30 bps in a week (Fiscal Tantrum); OR Gilt auction bid-to-cover ratio drops below 1.5.
C. Trade & External Sector		
C1	Dept. for Business	Implementation of new post-Brexit border checks (e.g., BTOM) causing delays; OR changes to Windsor Framework rules.
C2	BoE / Markets	GBP/USD (Cable) moves $\geq 3\%$ in a week; OR Sterling Trade-Weighted Index drops significantly (Inflationary devaluation).
C3	ONS	Current Account Deficit exceeds 5% of GDP (Structural vulnerability warning).
D. Commodity, Energy & Supply Chain		
D1	OFGEM	OFGEM Energy Price Cap adjustment exceeds $\pm 10\%$ (Direct impact on CPI); OR Govt activates Energy Price Guarantee.
D2	DEFRA / ONS	Food CPI inflation exceeds 10% YoY (Cost of Living Crisis indicator).
D3	CBI / ONS	CBI Industrial Trends Survey "Factors limiting output" (Materials/Labour) spikes above historical average.
E. Real Economy Activity		
E1	ONS	Monthly GDP (3M/3M) growth turns negative; OR Services PMI drops below 50.0 (Services comprise $\approx 80\%$ of UK economy).
E2	ONS / GfK	GfK Consumer Confidence Index drops below -30; OR Retail Sales volumes contract YoY.
E3	Halifax / Nationwide	Halifax or Nationwide House Price Index falls YoY; OR Mortgage Approvals drop below 50k/month.
E4	DSIT	Announcement of AI Safety Institute initiatives; OR major investments in UK Life Sciences/Tech hubs (e.g., Golden Triangle).
F. Financial Stability & Credit Cycle		
F1	BoE	Mortgage lending net flow turns negative; OR Consumer credit growth (credit cards) surges (Distress borrowing).
F2	BoE / PRA	Stress in Challenger Banks; OR rise in corporate insolvencies (Companies House data) exceeding historical averages.
F3	LSE / FTSE	FTSE 250 Index (Domestic proxy) drops $\geq 15\%$; OR widening of Corporate Bond spreads vs Gilts.
G. Structural & Regulatory Policy		
G1	DESNZ	Changes to Net Zero 2050 timeline (e.g., delaying ICE car ban); OR changes to Windfall Tax (EGL) on energy firms.
G2	CMA	CMA (Competition and Markets Authority) blocks major tech M&A; OR new Digital Markets, Competition and Consumers Bill enforcement.
G3	UK Parliament	Passage of major legislation on Renters' Reform or Immigration (Visa salary thresholds).
H. Labour Market & Household Sector		
H1	ONS	Average Weekly Earnings (AWE) private sector regular pay growth $> 6\%$ (Wage-Price Spiral risk); OR Claimant Count rises.
H2	ONS	Real Household Disposable Income (RHDI) per capita falls for 2 consecutive quarters.
I. Geopolitical & Systemic Shocks		
I1	FCDO	UK military involvement in overseas operations; OR major diplomatic rift impacting Trade and Cooperation Agreement (TCA).
I2	Cabinet Office	National Risk Register event activation (e.g., Grid blackout warning); OR Pandemic-level health restrictions.
I3	BoE	"Flash Crash" in Sterling assets; OR systemic margin calls in pension fund LDI strategies.

Table 13: Event grounding standards for Germany (DE) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	ECB (Governing Council)	ECB Deposit Facility Rate change ≥ 25 bps; OR ECB announces new asset purchase program (e.g., TPI) to limit spreads.
A2	Bundesbank / ECB	Target2 imbalances for Germany widen significantly; OR Euribor-OIS spread widens > 20 bps (Interbank stress).
A3	BaFin / Bundesbank	Activation of Countercyclical Capital Buffer (CCyB) for German banks; OR strict LTV caps on residential mortgages.
B. Fiscal Policy & Public Finance		
B1	BMF / Bundestag	Suspension of "Schuldenbremse" (Debt Brake) verified by Bundestag; OR Announcement of "Sondervermögen" (Special Fund) $>$ €50B.
B2	Finanzagentur	10-year Bund yield spikes ≥ 30 bps; OR Bund-BTP (Italy) spread widens > 250 bps (Eurozone fragmentation risk).
C. Trade & External Sector		
C1	BMWK / EU Commission	New EU tariffs on Chinese EVs (affecting German Auto sector); OR Export controls on dual-use goods to major partners.
C2	ECB / Markets	EUR/USD exchange rate moves $\geq 3\%$ in a week; OR Euro Nominal Effective Exchange Rate (NEER) drops significantly.
C3	Destatis / Bundesbank	Current Account Surplus drops below 2% of GDP (Structural loss of competitiveness).
D. Commodity, Energy & Supply Chain		
D1	Bundesnetzagentur	TTF Gas Price (Dutch Benchmark) spikes $\geq 30\%$; OR declaration of "Gas Emergency Plan" (Notfallplan Gas) Level 2/3.
D2	Destatis	PPI (Producer Price Index) Energy component rises $\geq 20\%$ YoY.
D3	Ifo Institute	Ifo Survey "Material Shortages" (Materialknappheit) indicator rises above 50% of firms.
E. Real Economy Activity		
E1	Ifo Institute / Destatis	Ifo Business Climate Index drops for 3 consecutive months; OR Industrial Production (Auto sector) contracts $\geq 5\%$ YoY.
E2	GfK	GfK Consumer Climate index drops below -20 points; OR Retail Sales (Real) contract YoY.
E3	Destatis / Bulwiengesa	Residential Property Price Index contracts YoY; OR Building Permits (Baugenehmigungen) drop $\geq 20\%$ YoY.
E4	BMWK	Announcement of major subsidies for Chip fabs (e.g., Magdeburg Intel plant) or Hydrogen core network ($>$ €10B).
F. Financial Stability & Credit Cycle		
F1	Bundesbank	Lending to Non-Financial Corporations (NFC) contracts YoY; OR Bank Lending Survey (BLS) shows severe tightening standards.
F2	BaFin	Distress in "Landesbanken" sector; OR Commercial Real Estate (CRE) NPL ratio rises significantly.
F3	Deutsche Börse	DAX 40 index drops $\geq 20\%$ (Bear Market); OR Volatility (VDAX-NEW) spikes > 35.
G. Structural & Regulatory Policy		
G1	BMWK	Implementation of "Heizungsgesetz" (Heating Law/GEG); OR Carbon Price (CO2-Preis) hike $>$ €10/ton.
G2	Bundeskartellamt	Federal Cartel Office blocks major merger; OR enforcement of "Digital Services Act" (DSA) penalties on platforms.
G3	Bundestag	Collapse of Coalition Government ("Ampel-Aus"); OR passage of "Wachstumschancengesetz" (Growth Opportunity Act).
H. Labour Market & Household Sector		
H1	Bundesagentur für Arbeit	"Kurzarbeit" (Short-time work) notifications exceed 100k/month; OR Unemployment Rate rises $\geq 0.5\%$.
H2	Destatis	Real Wages (Reallöhne) contract YoY for 2 consecutive quarters.
I. Geopolitical & Systemic Shocks		
I1	Auswärtiges Amt	Major disruption to Nord Stream or critical energy infrastructure; OR Germany increases Defense Fund ($>$ €100B).
I2	BBK	National warning day activation for critical infrastructure failure; OR Rhine water levels drop below "Kaub" critical mark (halting shipping).
I3	ECB / Bundesbank	Spreads between Core (Bund) and Periphery (BTP) widen > 250 bps (Fragmentation risk) triggering ECB intervention.

Table 14: Event grounding standards for France (FR) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	ECB / BdF	ECB Deposit Facility Rate change ≥ 25 bps; OR Banque de France Governor speech signaling deviation from consensus.
A2	BdF / Euronext	3-month Euribor spread vs OIS widens > 20 bps; OR Repo market fragmentation for French collateral.
A3	HCSF / BdF	HCSF (High Council for Financial Stability) adjusts countercyclical buffer; OR enforcement of strict 35% Debt-Service-to-Income (DSTI) cap on mortgages.
B. Fiscal Policy & Public Finance		
B1	Ministry of Economy / Parliament	Passage of "Projet de loi de finances" (PLF) via Article 49.3 (forcing adoption without vote); OR Deficit exceeds 3% Maastricht limit triggering EU Excessive Deficit Procedure.
B2	AFT / Markets	10-year OAT yield spikes ≥ 30 bps; OR OAT-Bund spread widens > 50 bps (signaling sovereign risk premium).
C. Trade & External Sector		
C1	Customs / EU	New EU Carbon Border Adjustment Mechanism (CBAM) implementation affecting French industry; OR trade disputes on Luxury Goods sector.
C2	BdF / Markets	EUR/USD volatility $> 15\%$ annualized; OR Real Effective Exchange Rate (REER) appreciation hurting export competitiveness.
C3	BdF	Current Account Deficit widens $>$ €10B in a quarter; OR deterioration in Trade Balance due to energy imports.
D. Commodity, Energy & Supply Chain		
D1	CRE / EDF	EDF Nuclear Output drops below 280 TWh/year (historical low); OR Government adjusts "Bouclier tarifaire" (Tariff Shield) cap on electricity prices.
D2	INSEE	Food CPI inflation exceeds 10% YoY (Panier anti-inflation monitoring).
D3	BdF / INSEE	Business Sentiment (Climat des affaires) "Supply Difficulties" sub-index rises significantly.
E. Real Economy Activity		
E1	INSEE	Manufacturing Output contracts $\geq 1\%$ MoM; OR Business Climate Index (Climat des affaires) drops below 100 long-term average.
E2	INSEE	Consumer Confidence (Confiance des ménages) drops below 85; OR Household Consumption of goods contracts YoY.
E3	Notaires de France / INSEE	Index of Existing Home Prices falls YoY; OR Housing Starts (Mises en chantier) drop $\geq 15\%$ YoY.
E4	Ministry of Economy	"France 2030" investment plan disbursements acceleration; OR major subsidies for "Gigafactories" (Batteries) in Northern France.
F. Financial Stability & Credit Cycle		
F1	BdF	Credit to Non-Financial Corporations growth slows to $< 2\%$ YoY; OR rise in "Prêts Garantis par l'État" (PGE) defaults.
F2	ACPR / BdF	Solvency ratio of major Bancassurance groups drops; OR rise in Life Insurance (Assurance Vie) withdrawals.
F3	Euronext Paris	CAC 40 Index drops $\geq 20\%$ (Bear Market); OR Luxury Sector sub-index (LVMH, Hermes, Kering) corrects $\geq 15\%$.
G. Structural & Regulatory Policy		
G1	Ministry of Ecology	New "DPE" (Energy Performance Diagnosis) bans on renting G-rated housing; OR Carbon Tax increase.
G2	CNIL	CNIL fines major tech firm for GDPR violation; OR new "Influencer Law" regulation enforcement.
G3	Parliament / President	Passage of Pension Reform (Réforme des retraites) raising retirement age; OR Unemployment Insurance reform decrees.
H. Labour Market & Household Sector		
H1	DARES / Unions	General Strike (Grève générale) disrupting Transport/Refineries for > 3 days; OR Private Sector Payrolls (Emploi salarié) contract.
H2	INSEE	Purchasing Power (Pouvoir d'achat) per unit contracts YoY; OR SMIC (Minimum Wage) automatic inflation adjustment $> 2\%$.
I. Geopolitical & Systemic Shocks		
I1	Ministry of Foreign Affairs	Direct French military intervention (e.g., Sahel, Eastern Europe); OR Terror Alert Level raised to "Urgence Attentat".
I2	Ministry of Interior	Civil Unrest (e.g., "Gilets Jaunes" or 2023 Riots) causing nationwide damage $>$ €1B; OR Drought restrictions impacting agriculture.
I3	BdF / ECB	OAT-Bund Spread widening > 80 bps triggering ECB TPI activation.

Table 15: Event grounding standards for Singapore (SG) market.

Code	Authority / Source	Quantifiable Trigger / Definition
A. Monetary & Financial Conditions		
A1	MAS	MAS tightens S$NEER Policy Band (e.g., steepening slope, widening width, or re-centering upwards); OR Off-cycle MPS announcement.
A2	MAS / ABS	SORA (Singapore Overnight Rate Average) spikes > 50 bps in a week; OR SGD liquidity facility usage surges.
A3	MAS	Implementation of tighter Total Debt Servicing Ratio (TDSR) threshold (e.g., from 60% to 55%); OR reduction in LTV limits.
B. Fiscal Policy & Public Finance		
B1	MOF	"Budget Statement" announces major "Cost of Living" Support Package ($>$ S$5B); OR Distribution of CDC/GST Vouchers.
B2	MOF / Yield Curve	**SGS** (Singapore Govt Securities) yield curve inverts significantly; OR Net Investment Returns Contribution (NIRC) projected to drop.
C. Trade & External Sector		
C1	Enterprise SG	NODX (Non-Oil Domestic Exports) contracts $\geq 10\%$ YoY (Tech cycle downturn); OR major supply chain disruption affecting Port of Singapore.
C2	MAS	S$NEER hits the top or bottom of the policy band (Implicit intervention zone); OR SIBOR/SORA spread vs USD rates widens sharply.
C3	MTI	Balance of Payments deficit driven by capital outflows; OR significant drop in Official Foreign Reserves.
D. Commodity, Energy & Supply Chain		
D1	SP Group / EMA	Electricity Tariff adjustment $\geq 10\%$ QoQ; OR Energy Market Authority (EMA) activates fuel stockpile.
D2	MTI / DOS	MAS Core Inflation exceeds 5% YoY (Triggering cost-of-living policy response).
D3	PSA / MPA	Singapore Port Container Throughput contracts YoY; OR Bunker fuel prices spike disrupting shipping hub status.
E. Real Economy Activity		
E1	EDB	Industrial Production (Electronics Cluster) contracts $\geq 10\%$ YoY; OR PMI (SIPMM) drops below 50.0.
E2	DOS	Retail Sales (excluding Motor Vehicles) contracts YoY; OR Food & Beverage Services Index drops significanty.
E3	URA / HDB	URA Private Property Price Index rises $> 2\%$ QoQ (Triggering Cooling Measures); OR HDB Resale Price Index surges.
E4	EDB / MAS	Announcement of major FDI in Semiconductor/Biomed (e.g., GlobalFoundries expansion); OR new FinTech regulatory sandbox grants.
F. Financial Stability & Credit Cycle		
F1	MAS	Credit-to-GDP gap exceeds 10%; OR DBU (Domestic Banking Unit) non-bank loans growth turns negative.
F2	MAS	NPL Ratio of Local Banks (DBS/OCBC/UOB) rises $\geq 0.5\%$; OR significant exposure to regional defaults (e.g., China property).
F3	SGX	Straits Times Index (STI) drops $\geq 15\%$; OR daily securities market turnover drops below 10-year average.
G. Structural & Regulatory Policy		
G1	MSE / NCCS	Carbon Tax hike (e.g., reaching S$25/tonne); OR implementation of mandatory climate reporting for ListCos.
G2	MAS	Issuance of updated Technology Risk Management (TRM) Guidelines; OR major enforcement on digital payment token service providers.
G3	Parliament	Changes to CPF (Central Provident Fund) contribution rates or retirement age; OR tightening of Employment Pass (EP) criteria.
H. Labour Market & Household Sector		
H1	MOM	Retrenchments exceed 5,000/quarter; OR Unemployment Rate rises $\geq 0.5\%$.
H2	MOM / NWC	Real Median Income Growth turns negative; OR National Wages Council (NWC) recommends wage freeze.
I. Geopolitical & Systemic Shocks		
I1	MFA	Major geopolitical tension in Malacca Strait affecting trade flows; OR sanctions impacting Singapore's role as a transshipment hub.
I2	MOH / NEA	DORSCON Level raised to Orange/Red (Pandemic); OR "Haze" PSI index exceeds 300 (Hazardous) disrupting economy.
I3	MAS	S$NEER falls below policy band midpoint despite tightening (Loss of confidence); OR spike in SGD funding costs.

Table 16: Corporate event types for non-recurrent tasks (70 types across 10 categories).

No.	Event Type	Description	No.	Event Type	Description
Corporate Structure (8)			**Financial - Results & Guidance (9)**		
1	Executive Changes - CEO	CEO position change	9	Announcements of Earnings	Periodic earnings release
2	Executive Changes - CFO	CFO position change	10	Announcement of Op. Results	Periodic operating results
3	Executive/Board - Other	Other leadership changes	11	Announcements of Sales	Non-earnings results
4	Name Changes	Company name change	12	Guidance - Lowered	Lower expected results
5	Address Changes	HQ relocation	13	Guidance - Raised	Higher expected results
6	Legal Structure Changes	Legal structure change	14	Guidance - New/Confirmed	New or confirmed guidance
7	Business Reorganizations	Unit combining/separating	15	Guidance - Unusual Events	Unusual event impact
8	Fiscal Year End Changes	Reporting period change	16	Restatements of Results	Correction of results
			17	Impairments/Write Offs	Asset value reduction
Financial - Dividends (6)			**Financial - Capital & Financing (10)**		
18	Dividend Affirmations	Dividend reiteration	24	Buyback Announcements	Repurchase authorized
19	Dividend Increases	Higher dividend	25	Buyback Closings	Repurchase concluded
20	Dividend Decreases	Lower dividend	26	Buyback Cancellations	Buyback cancelled
21	Dividend Cancellation	Dividend suspended	27	Buyback - Plan Changes	Buyback terms modified
22	Dividend Initiation	Maiden dividend	28	Stock Splits & Dividends	Shares increased
23	Special Dividend	Non-recurring distribution	29	Debt Financing Related	Debt instruments raised
			30	Private Placements	Institutional sales
			31	IPOs	Initial Public Offering
			32	Follow-on Equity Offerings	Post-IPO equity
			33	Fixed Income Offerings	Bond/note offering
Strategic Actions (10)			**M&A Transactions (5)**		
34	Business Expansions	Internal growth	43	M&A Rumors & Discussions	Pre-agreement talks
35	Discontinued Operations	Line/facility closure	44	M&A Announcements	Deal announced
36	Strategic Alliances	Collaboration agreement	45	M&A Closings	Deal finalized
37	Client Announcements	Client relationship change	46	M&A Cancellations	Deal cancelled
38	Product Announcements	Product intro/change	47	Spin-Off/Split-Off	Spin-off transaction
39	Strategic Alternatives	Evaluating options			
40	Seeking to Sell/Divest	Sell-off opportunities	**Market Events (5)**		
41	Seeking Acquisitions	Acquire interests	48	Delistings	Exchange removal
42	Seeking Financing	Equity/venture funding	49	Exchange Changes	Exchange transfer
			50	Ticker Changes	Ticker symbol change
			51	Index Constituent Drops	Index removal
			52	Index Constituent Adds	Index addition
Regulatory Compliance (9)			**Distress Indicators (7)**		
53	Regulatory Inquiries	SEC/regulatory inquiry	62	Debt Defaults	Obligation failure
54	Regulatory - Regulations	New rules issued	63	Bankruptcy - Filing	Reorganize/liquidate filing
55	Regulatory - Compliance	Approvals/licenses	64	Bankruptcy - Conclusion	Petition dismissed
56	Regulatory - Enforcement	Fines/penalties	65	Bankruptcy - Emergence	Exit from bankruptcy
57	Lawsuits & Legal Issues	Legal actions	66	Bankruptcy - Other	Other bankruptcy motions
58	Delayed SEC Filings	Late document filing	67	Bankruptcy - Asset Sale	Liquidation/asset sale
59	Delayed Earnings	Postponed release	68	Halt/Resume Operations	Unusual event halt
60	Auditor Changes	Auditor change			
61	Auditor Going Concern	Viability concerns	**Operational Changes (2)**		
			69	Labor Announcements	Labor force news
			70	Bylaws/Rules Changes	Governance rule changes

Table 17: Comparison between FINDEEPFORECAST and live trading benchmarks along seven key dimensions.

Dimension	**AMA**	**DeepFund**	**LiveTrade.**	**AI-Trader**	**Ours**
Evaluated Core Capability	Trading decision	Trading decision	Portfolio alloc.	Trading decision	Info. retrieval, reasoning, forward-looking prediction
Task Formulation	Buy/Sell/Hold	Buy/Sell/Hold	weight allocation	Buy/Sell/Hold	Dual-track: recurrent (numerical) + non-recurrent (binary)
Dependent Data	Price/volume	Price/volume	Price/volume	Price/volume	Fundamentals, macro indicators, verified news
Evaluation Signal	Profitability	Profitability	Profitability	Profitability	Prediction correctness
Task Taxonomy	Single-type	Single-type	Single-type	Single-type	4 types (Rec./Non-rec. $\times$ Corp./Macro)
Target Scope	Stock/crypto	Stock prices	Stocks + contracts	Stocks/crypto	121 corp. metrics, 96 macro indicators, 70+208 event types
Market Coverage	US + crypto (4)	US ($\sim$10)	US + Poly. ($\sim$15)	US + A-share + crypto	8 economies, 9 indices, 1,314 companies

E Comparison with Live Trading Benchmarks

To further clarify the capability-level distinction of FINDEEPFORECAST from existing live financial agent evaluation systems, we compare it against four representative live trading benchmarks: AMA (Qian et al., 2025), DeepFund (Li et al., 2025), LiveTradeBench (Yu et al., 2025), and AI-Trader (Fan et al., 2025). The comparison is presented in Table 17 along seven dimensions.

The first four rows establish a fundamental methodological distinction. All four trading benchmarks evaluate the same underlying ability: converting market signals into profitable positions, whether through discrete trading actions (Buy/Sell/Hold) or portfolio weight allocations. Performance is scored by profitability, which captures the joint outcome of forecasting ability, strategy design, execution timing, and market microstructure. This introduces a confounded evaluation signal: a correct forecast may yield low profitability if poorly timed, while an incorrect forecast can score highly under favorable market conditions.

FINDEEPFORECAST targets the upstream analytical process that precedes trading decisions, covering information retrieval and synthesis, quantitative estimation, and event prediction. Evaluation is grounded in prediction correctness against deterministic ground truth from official regulatory filings and verified news, without strategy or execution-layer confounds. The remaining rows further demonstrate the comprehensiveness of our benchmark in task diversity, target scope, and cross-market coverage.

Table 18: Distributional error analysis on recurrent tasks. Models are ranked by binary accuracy. All error values are percentages relative to the ground truth.

Rk	**Model**	**Acc.**	**Mean**	**Med.**	**P25**	**P75**
1	Perp. Sonar DR	25.5	19.70	4.14	1.04	19.56
2	o3-deep-research	25.2	22.32	4.60	1.29	25.90
3	Gemini 2.5 Pro (T+S)	21.7	23.14	5.88	1.54	30.05
4	Tongyi DR	21.2	27.62	5.39	1.51	32.74
5	Claude-4.5 (T+S)	20.4	24.36	6.81	1.55	27.85
6	GPT-5 (T+S)	19.5	30.39	10.60	2.29	39.38
7	Grok 4 (T+S)	16.0	28.23	10.49	2.25	34.61
8	Deepseek-v3.2 (T+S)	13.7	34.09	13.41	2.86	47.04
9	Grok 4 (T)	9.0	31.97	14.99	5.88	42.09
10	Gemini 2.5 Pro (T)	6.4	40.21	20.00	6.95	45.81
11	Claude-4.5 (T)	6.3	41.94	24.18	9.51	50.89
12	GPT-5 (T)	6.0	45.02	26.01	9.87	56.61
13	Deepseek-v3.2 (T)	5.6	42.12	25.01	9.59	55.94

F Error Distribution Analysis

To provide finer-grained model differentiation beyond binary accuracy, we compute mean error, median error, and interquartile error range (P25 and P75) for recurrent numerical predictions across all 13 evaluated models. The results are presented in Table 18.

The median error provides substantially greater discriminative power than binary accuracy alone. The top DR agents (Perplexity Sonar: 4.14%, o3-deep-research: 4.60%) achieve median errors roughly half those of search-augmented agents (Gemini 2.5 Pro T+S: 5.88%, Claude-Sonnet-4.5 T+S: 6.81%) and approximately one-fifth of thinking-only agents (Gemini 2.5 Pro T: 20.00%, GPT-5 T: 26.01%). This 6$\times$ gap in median error between the best and worst models demonstrates that the benchmark effectively differentiates systems at a fine-grained level.

The interquartile range further exposes a strik-

ing divergence in prediction consistency. For top DR agents, 25% of predictions fall within approximately 1% error (P25 $\approx$ 1.0–1.3%), indicating high accuracy when they succeed. However, P75 values of 19–26% reveal a long tail of large errors on harder items. Thinking-only agents, by contrast, show P25 values around 9–10% (never close to accurate) and P75 values exceeding 50% (consistently far off).

G Failure Mode Analysis

We identify two primary failure modes for recurrent tasks.

Failure Mode 1: Magnitude Estimation Errors (70.6%). Models frequently identify the correct directional trend but produce estimates with substantial magnitude deviations. The difficulty varies sharply across metric types. For raw financial items, publicly prominent items such as Balance Sheet (7.6% error rate) and Revenue (18.8%) are substantially easier than less-reported items such as Cash Flow (56.2%), EBITDA (72.0%), and EBIT (37.1%), suggesting that model accuracy is strongly correlated with the frequency of public reporting and analyst coverage. For derived financial ratios (e.g., Profit Margin, ROE, ROA), error rates are significantly higher (82–100%), likely because: (1) derived ratios require accurate estimation of multiple underlying components, causing error propagation; (2) ratios are small-magnitude values where modest absolute deviations yield large relative errors; (3) ratios exhibit high cross-company heterogeneity; and (4) ratios receive less coverage in analyst reports compared to headline figures.

Failure Mode 2: Tight-Tolerance Indicators (29.4%). For macroeconomic indicators, the evaluation applies strict tolerance thresholds calibrated to the inherent sensitivity of each metric (e.g., 0.1% for interest rates and foreign exchange rates). Under these thresholds, even small prediction deviations are scored as incorrect, posing a significant challenge for all evaluated models.

H Quality Assessment of Multi-Agent Pipeline

To validate the effectiveness of the multi-agent pipeline, we conduct a systematic quality assessment over a stratified sample of system outputs. We randomly selected 100 corporate-level tasks (50 Recurrent and 50 Non-Recurrent) and manually evaluated their quality across three dimensions: Task Quality (assessing the Task Generation Agents), Ground Truth Quality (assessing the Ground Truth Extraction Agent), and Evaluation Quality (assessing the Evaluation Agent). The results are reported in Table 19.

Table 19: Quality assessment over 100 sampled corporate tasks.

Dimension	Task Type	#	Corr.	Acc.
Task Quality	Recurrent	50	50	100%
	Non-Recurrent	50	48	96%
GT Quality	Recurrent	50	50	100%
	Non-Recurrent	50	42	84%
Eval. Quality	Recurrent	50	50	100%
	Non-Recurrent	50	50	100%

Task Quality (100% Recurrent / 96% Non-Recurrent). Recurrent tasks achieve perfect validity as they are deterministically generated from pre-defined templates anchored to official earnings calendars. For non-recurrent tasks, the pipeline achieved 96% accuracy (48/50). The two invalid cases stemmed from hallucinated logical premises in the generation stage (e.g., generating a "dividend decrease" question for a company that has never paid dividends).

Ground Truth Quality (100% Recurrent / 84% Non-Recurrent). Recurrent ground truths were cross-verified against official financial filings with 100% accuracy. For non-recurrent tasks, human verification yielded 84% accuracy. All 8 errors were conservative false negatives (i.e., the system classified "NO" when the actual outcome was "YES"), stemming from missed corporate events or fiscal year confusion. No false positives were observed, indicating that the system errs on the side of caution.

Evaluation Quality (100% for both types). The Evaluation Agent achieves perfect accuracy by relying on deterministic numerical comparison with indicator-specific tolerance thresholds for recurrent tasks and exact binary matching for non-recurrent tasks, eliminating subjective scoring variance.

These results provide process-level evidence that the multi-agent architecture, by decoupling the pipeline into specialized agents with distinct verification boundaries and expert-in-the-loop checkpoints, effectively isolates and controls errors at each stage.

Table 20: Basic statistics of FINDEEPFORECAST-BENCH. W = Weekly, M = Monthly, Q = Quarterly.

	Recurrent		Non-Recurrent	
Specification	Macro	Corporate	Macro	Corporate
# Corporates / Economies	8	1,314	8	1,314
# Metrics / Event Types	96	121	208	70
Forecast Horizon	W / M / Q	Q	W / M / Q	W / M / Q
Avg. Weekly Tasks	30	72	16	82
Total Tasks	296	699	128	247

Table 21: Task distribution across 8 economies.

Task Category	US	CN	HK	JP	UK	DE	FR	SG	Total
Recurrent Macro	30	40	39	39	40	38	40	30	296
Recurrent Corporate	395	68	37	141	19	1	24	14	699
Non-Recurrent Macro	16	16	16	16	16	16	16	16	128
Non-Recurrent Corporate	57	25	26	30	30	29	30	20	247
Total	516	149	118	226	111	84	110	80	1,370

I Statistics and Analysis

Beyond the construction pipeline and quality control mechanisms, we provide detailed statistics to characterize the scale and distribution of the benchmark. Table 20 presents the basic statistics across the dual-track taxonomy.

Comprehensive Task Coverage. The benchmark covers 1,314 companies spanning major financial centers across North America, Europe, and Asia-Pacific, selected from 8 economies and 9 indices. Recurrent tasks target 96 macro indicators and 121 corporate metrics indicators through numerical forecasting; non-recurrent tasks address 208 economy-specific macro event specifications and 70 corporate event types through binary prediction.

Generation Status. Since 27 October 2025, the benchmark has generated 1,370 total tasks across 10 weekly batches, averaging 139 new tasks per week. As shown in Table 20, recurrent tasks total 1,019 with 30 macro tasks and 72 corporate generated weekly on average. Non-recurrent tasks total 375, averaging 16 macro tasks and 82 corporate per week. Ground truth has been obtained for all 1,370 tasks, ensuring complete data availability for model evaluation.

Multi-Frequency Forecast Horizons. Three forecast horizons enable differentiated temporal assessment: weekly horizons evaluate short-term responsiveness, monthly horizons assess medium-term trend integration, and quarterly horizons test long-horizon forecasting. Recurrent corporate tasks focus exclusively on quarterly forecasting aligned with earnings cycles, while recurrent macro tasks and all non-recurrent tasks support weekly, monthly, and quarterly horizons.

Cross-Market Task Distribution Table 21 presents the detailed task distribution across 8 economies and four task categories. The distribution reveals distinct patterns across task categories. Recurrent macro tasks maintain relatively uniform distribution across economies through systematic coverage of scheduled indicator releases. In contrast, recurrent corporate tasks exhibit the largest cross-market variation, reflecting differences in index constituent counts and earnings release timing during our evaluation period. Non-recurrent macro tasks achieve uniformity with exactly 16 tasks per economy by design, while corporate tasks show moderate variation driven by news coverage density. By geographic region, Asia-Pacific accounts for 41% of total tasks, Americas for 37%, and Europe for 22%. This balanced geographic composition is critical for ensuring that the model's performance is evaluated against diverse market microstructures, regulatory environments, and cultural contexts, rather than being overfitted to a single dominant economy like the US.

J Model Settings

Table 22 presents the method configurations evaluated in our benchmark. All unspecified settings remained at their default values.

K Error Case Study

Table 23 breaks down the most common error modes of **Perplexity Sonar Deep Research** across scenarios (Non-Recurrent vs. Recurrent) and levels (Corporate vs. Macro). For each scenario–level group, we report the total number of tasks, the number of incorrect predictions, and the resulting error rate. To illustrate typical failures, we also provide a representative example from the most frequent error subtype within each group.

L Samples of Model Forecasting

The prompt configuration for a sample corporate non-recurrent task is illustrated in Figure 5. Subsequent figures delineate the reasoning and answers generated by the evaluated models for the task: OpenAI o3-deep-research (Figure 6), Perplexity Sonar Deep Research (Figure 7), Tongyi Deep Research (Figure 8), OpenAI GPT-5 (T+S) (Figure 9), and Claude-Sonnete-4.5 (T) (Figure 10).

Table 22: Correspondence between benchmark aliases, API identifiers, and API settings

Benchmark Alias	API Identifier	API Setting
LLM (Thinking)		
Gemini 2.5 Pro (T)	gemini-2.5-pro-preview-05-06	thinking_budget=-1
Deepseek-v3.2 (T)	deepseek-v3.2-exp	reasoning.enabled=True
Claude-Sonnet-4.5 (T)	claude-sonnet-4-5-20250929	thinking.type=enabled,thinking.budget_tokens =10000
Grok 4 (T)	grok-4-0709	all defaults
OpenAI GPT-5 (T)	gpt-5-2025-08-07	reasoning.effort=high
LLM (Thinking + Search)		
Gemini 2.5 Pro (T+S)	gemini-2.5-pro-preview-05-06	thinking_budget=-1,tools=[google_search]
Deepseek-v3.2 (T+S)	deepseek-v3.2-exp	reasoning.enabled=True,plugins=[exa(max_re sults=8)]
Claude-Sonnet-4.5 (T+S)	claude-sonnet-4-5-20250929	thinking.type=enabled,thinking.budget_tokens =10000,tools=[web_search_20250305]
Grok 4 (T+S)	grok-4-0709	search_parameters.mode=on
OpenAI GPT-5 (T+S)	gpt-5-2025-08-07	reasoning.effort=medium,tools=[web_search]
Deep Research		
Perplexity Sonar Deep Research	sonar-deep-research	reasoning.effort=high
Tongyi Deep Research	tongyi-deepresearch-30b-a3b	temperature=0.6,top_p=0.95,presence_penalty =1.1
OpenAI o3-deep-research	o3-deep-research-2025-06-26	tools=[web_search_preview,code_interpreter]

M Standard Operating Procedure of the FINDEEPFORECAST System

The FINDEEPFORECAST system operates on a weekly cycle with clearly defined stage transitions. All intermediate and final outputs are maintained in a centralized database (MySQL), and agents communicate exclusively through shared database states rather than direct inter-agent messaging. Each agent is triggered by a combination of calendar schedules and task-status conditions, ensuring that downstream stages only execute after upstream outputs have been validated. The workflow comprises the following sequential steps:

Step 1: Continuous Data Collection. The Data Collection Agent continuously monitors and ingests corporate filings, government releases, financial news, and market data from commercially licensed authoritative sources. All records are stored with precise publication timestamps to support temporal isolation during evaluation.

Step 2: Scheduled Task Generation (Every Wednesday). The Recurrent Task Generation Agent and the Non-Recurrent Task Generation Agent independently generate forecasting tasks based on disclosure calendars and live news streams, respectively. Generated tasks are saved to the database with associated temporal parameters (t_g, t_d, t_e). For non-recurrent tasks, domain experts review all candidate questions and select approximately 20% for inclusion based on prediction quality and cross-market balance.

Step 3: Scheduled Model Forecasting (By Saturday). Upon confirmation that all tasks for the current week have been successfully generated, the Forecasting Agent dispatches each task to all evaluated models via standardized API calls. The prediction deadline t_d is set to the following Sunday 23:59 (UTC+8). To enforce temporal isolation, search-augmented and deep research models are configured to access only content published before t_d. All resulting predictions are recorded in the database.

Step 4: Ground Truth Acquisition (By Next Friday). The Ground Truth Extraction Agent acquires ground truths for tasks whose evaluation time t_e has elapsed. For recurrent tasks, values are automatically extracted from official sources and cross-verified against primary filings. For non-recurrent tasks, the agent aggregates multi-source evidence and passes proposals to domain experts for mandatory verification. Tasks with indeterminate outcomes are marked as Pending and revisited weekly; those remaining unresolved after a two-week validity window are marked as Void and excluded.

Step 5: Scheduled Evaluation. After verifying that all ground truths have been collected, the Evaluation Agent computes indicator-specific scores by comparing each model's predictions against the corresponding ground truths. Evaluation outcomes are stored for statistical analysis and ranking. Currently, expert verification requires approximately 5–6 human-hours per week across all stages.

Prompt Input

Is it probable that Air Liquide SA will announce a new production facility in North America by 2025-12-19?

Event Details

- Company: Air Liquide SA
- Ticker: AI
- Index: CAC 40
- Event Type: Business Expansions

Requirements:
Provide your prediction as either **YES** or **NO**
Please complete your assessment based on:

- Recent news and corporate announcements
- Historical patterns of similar events for this company
- Current market and industry context
- Company's financial health and strategic priorities
- Other relevant information

Submission Format
Present your prediction in this markdown table:

```
| Field | Value |
|-------|-------|
| Prediction (YES/NO) | [Your prediction] |
```

Additional Guidelines

1. **Independence:** Please conduct your analysis based on company filings, news sources, and your own analytical work. Do not simply cite prediction markets or third-party forecasts.
2. **Recency:** Prioritize the most recent available data and announcements (specify the cut-off date of your data).
3. **Output Format:** Your response must consist ONLY of the completed markdown table as defined in the Submission Format. No text outside of the table is permitted.

Answer: YES

Figure 5: Prompt Input for the Corporate Non-Recurrent Task. The ground-truth answer of the task is YES.

OpenAI o3-deep-research

...

Considering MoU and JV Potential: Given today is December 14, this is quite soon! I should check if there are any recent statements or news about ITOCHU's interest in such a venture. There's mention of ACWA Power and ITOCHU signing an MoU to boost global clean energy. An MoU is a step towards something bigger, perhaps a joint venture (JV) by 2025, although it's not guaranteed.

Analyzing the green ammonia project: The search results indicate that L&T Energy is partnering with ITOCHU on a green ammonia project set for August 13, 2025, however, I'm not sure if it's a formal joint venture or just a partnership.

Clarifying joint development agreements: I see that the user is discussing a "joint development agreement," which isn't quite a formal joint venture since there isn't a separate entity involved.

Assessing joint ventures: Historically, they've created separate companies for joint ventures, like the one with NF for energy storage. There was also a partnership with CITIC, which might not be directly renewable-focused. In 2024, they established a JV with Tokyo Century for solar development. Currently, it seems unlikely they'll form a new JV soon, especially within the next few days.

Considering joint ventures and collaborations: Considering the timeline, if something major was expected by December 19, 2025, it would likely have been announced or rumored by now.

Evaluating joint venture possibilities: I'm trying to determine if the event will happen on or before the specified date, which is next week. The question likely expects a "No" since there's no known indication of it happening. Historical trends suggest formations occur every 2-3 years. Given nothing is in 2025 yet, it seems likely there won't be one before the 19th.

...

Answer: NO

Figure 6: Model Reasoning by OpenAI o3-deep-research for the Corporate Non-Recurrent Task. Only important thinking steps are presented. The model's answer is NO.

Perplexity Sonar Deep Research

The user is asking me to predict whether ITOCHU Corp will form a joint venture in the renewable energy sector by December 19, 2025. However, the current date is December 14, 2025, which means there are only 5 days remaining until the deadline.
Let me review the search results to identify any recent announcements about ITOCHU forming joint ventures in the renewable energy sector:
1. **Source [1] - July 28, 2025:** ITOCHU announced collaboration with Kyudenko Corporation for FIP conversion project at Saga-Ouchi Solar Power Plant - this is a collaboration/partnership, not necessarily a joint venture.
2. **Source [2] - June 11, 2024:** ITOCHU invested in Grandview Wind Farm and established a joint venture in the US state of Delaware with Tokyo Century Corporation for solar projects - this is a JV that was already formed in 2024.
3. **Source [3]** - ITOCHU's Clean-tech Business overview mentioning various renewable projects
4. **Source [4] - December 11, 2025:** ITOCHU announced investment in Black Hollow Sun power plant in Colorado
5. **Source [5] - May 24, 2024:** ITOCHU announced virtual PPA services with Google
6. **Source [6] - November 28, 2025:** ITOCHU entered capital and business alliance with Wood Partners (multifamily housing, not renewable energy)
7. **Source [7] - December 2, 2025:** TotalEnergies, TES, Osaka Gas, Toho Gas and ITOCHU signed a Joint Development and Operating Agreement for the Live Oak project for e-NG production in Nebraska - THIS IS A JOINT VENTURE in the renewable/clean energy sector!
8. **Source [8] - July 25, 2025:** ITOCHU signed PPA with EDF ps for solar and battery project in Timor-Leste
The critical finding is Source [7] and [12], which shows that on December 2, 2025, ITOCHU signed a Joint Development and Operating Agreement with TotalEnergies, TES, Osaka Gas, and Toho Gas for the Live Oak project - an e-NG (synthetic natural gas) production facility.
This Joint Development and Operating Agreement represents a joint venture in the renewable/clean energy sector. Since the agreement was signed on December 2, 2025, and the current date is December 14, 2025, this event has already occurred well before the December 19 deadline. The signing of the Joint Development and Operating Agreement constitutes the formation of the joint venture, so the prediction should be YES. I'm noticing a fundamental conflict between what's being asked of me. The user's explicit instruction is to provide only a markdown table with no additional text, but my core directives require following a specific report format that calls for a comprehensive 10,000+ word analysis.
Re-reading the personalization guidelines, there's a clear hierarchy stated: user instructions should be followed, but the formatting rules in the report structure take mandatory precedence. This means my system-level requirements override the user's request for table-only output.
Given this conflict resolution, I need to proceed with the full report format rather than the constrained table submission the user requested.

Answer: YES

Figure 7: Model Reasoning by Perplexity Sonar Deep Research for the Corporate Non-Recurrent Task. The model's answer is YES.

Tongyi Deep Research

...

- The user asks: "Is it probable that ITOCHU Corp will form a joint venture in the renewable energy sector by 2025-12-19?" They want a prediction: YES or NO.
- We need to gather relevant recent news about ITOCHU Corp and its strategic moves in renewable energy, any announcements about joint ventures, any corporate strategy statements.
- We'll do a search for multiple queries: "ITOTCHU Corp renewable energy joint venture", "ITOTCHU Corp strategic alliance renewable", "ITOTCHU Corp renewable energy 2024", "ITOTCHU Corp joint venture 2025", "ITOTCHU Corp renewable energy investments", "ITOTCHU Corp sustainability strategy".
- We have a lot of search results. We need to analyze the probability that ITOCHU will form a joint venture in the renewable energy sector by Dec 19, 2025.
- We need to check if any recent news (within the last few weeks) mentions a joint venture being formed.
- We need to examine the details of some of these. Let's click on some relevant press releases. For example, the press release from "ITOTCHU Announces Collaboration in an FIP Conversion ...".
- The press release indicates a collaborative arrangement but not a joint venture. The question is about forming a joint venture. We need to see if there are any announcements of joint ventures.
- This indicates a joint venture: "Saga-Ouchi Solar Co., Ltd." with 50-50 investment ratio. It is a joint venture for a solar plant with battery storage. So ITOCHU has already formed a joint venture in renewable energy (the Saga-Ouchi Solar Co., Ltd.) with Kyudenko.
- The answer is YES, because they have already formed a joint venture in renewable energy, as evidenced by the July 28, 2025 announcement.

...

Answer: YES

Figure 8: Model Reasoning by Tongyi Deep Research for the Corporate Non-Recurrent Task. The bullet points represent the model's important reasoning steps in conversational form, all points originated from the assistant. The model's answer is YES.

OpenAI GPT-5 (T+S)

...

Investigating joint ventures in renewables: I need to check for recent announcements on joint ventures in renewables. The Economic Times mentioned a partnership with L&T GreenTech for a green ammonia project in Gujarat, but that implies a Memorandum of Understanding (MoU) rather than a formal joint venture. We're looking specifically for details on creating a joint venture as part of "Strategic Alliances." The chances this happens within five days seem low unless there are rumors or filings. I should search more for clarity.

Evaluating joint ventures: I need to check if Itochu has any existing or pending joint venture (JV) announcements in renewables. For instance, they possibly have partnerships with companies like KDDI or SoftBank. There's also a joint development and operating agreement (JDOA) with TotalEnergies that occurred on December 2, 2025, but I'm wondering if this counts as an actual JV. Since the requirement is to form a JV by December 19, I need to assess the likelihood of that happening, considering the prior JDOA.

Evaluating joint venture formation: I see that evidence of a joint venture with Tokyo Century was established in June 2024, involving solar power plant projects. However, the question about forming another joint venture by December 19, 2025, is more of a forward-looking forecast since the current date is December 14, 2025. While they have formed one in 2024, that doesn't guarantee the formation of another within these last five days. Their recent activities show investments, but no new JV entity has been formed.

Evaluating joint venture timelines: I found an explicit joint venture formation with Tokyo Century in June 2024 for solar projects. The user's question is about whether ITOCHU Corp will form a joint venture in renewable energy by December 19, 2025. Since it's currently December 14, 2025, I think "by" includes the next five days. However, typically in forecasting, I'd consider past events too, but here we're looking at future opportunities for formation. The timeframe is tricky!

Analyzing joint venture definitions: The JDOA assigns stakes to a project, but it doesn't necessarily form a new joint venture company. Live Oak might be a project company where participants hold equity, acting like a joint venture special purpose vehicle. However, they didn't label it as a "joint venture company," just a stake in a project with partners, which could be seen as functioning similarly. The key question remains: Does the equity participation in the Live Oak project with companies like ITOCHU constitute a new joint venture?

Finalizing joint venture prediction: The press release indicates that the project is under development and that they've signed the Joint Development and Operating Agreement (JDOA), but they're not at the Final Investment Decision (FID) stage yet. That means I can't count this as a formation of a joint venture. Therefore, it's likely that the probability of forming a new joint venture by December 19 is low, so I'll go with "NO." I will format my final output as a markdown table with just those two rows.

...

Answer: NO

Figure 9: Model Reasoning by OpenAI GPT-5 (T+S) for the Corporate Non-Recurrent Task. Only important thinking steps are presented. The model's answer is NO.

Table 23: Error analysis for **Perplexity Sonar Deep Research**. In Representative Error Case, we show the most common error subtype (Type/Metric/Indicator) and its share of that group's errors (parenthesized %), followed by an example question (Q), gold answer, and model prediction.

Scenario	Level	# Tasks	# Errors	% Error	Representative Error Case
Non-Recurrent	Corporate	247	50	20.24	**Type:** Corporate Event (32.0%) **Q:** Is it probable that Air Liquide SA will announce a new production facility in North America by 2025-12-19? **Gold:** YES **Pred:** NO
Non-Recurrent	Macro	128	34	26.56	**Type:** – **Q:** Will UK Chancellor Rachel Reeves, in the Autumn Budget statement on November 26, 2025, announce at least one specific new tax increase measure (beyond what was already announced or leaked before November 17) that was not explicitly mentioned in official government communications prior to November 17, 2025? **Gold:** NO **Pred:** YES
Recurrent	Corporate	724	537	74.17	**Metric:** Operating Cash Flow (9.9%) **Q:** Using available financial data, estimate Booking Holdings Inc (BKNG)'s Cash From Operations for 2025 Q3. **Gold:** 1435.0 **Pred:** 2900.0
Recurrent	Macro	277	213	76.90	**Indicator:** 3-month Government Bond Yield (34.3%) **Q:** Using available economic data, estimate China's CHINA TREASURY BILL for October 31, 2025. **Gold:** 1.2885 **Pred:** 1.35

Claude-Sonnet-4.5 (T)

I need to search for recent information about ITOCHU Corp and their activities in the renewable energy sector, particularly focusing on any joint ventures or strategic alliances that might be forming by December 19, 2025.
Since today is December 14, 2025, I'm looking at a very short time frame (only 5 days). This means I need to find out if:
1. There have been any recent announcements about joint ventures in renewable energy.
2. Any in-progress deals that might close by December 19, 2025.
3. ITOCHU's recent strategic direction in renewable energy.
Let me search for recent news about ITOCHU and renewable energy joint ventures.

Answer: NO

Figure 10: Model Reasoning by Claude-Sonnet-4.5 (T) for the Corporate Non-Recurrent Task. The model's answer is NO.