

000 001 002 003 004 005 FUTUREX: AN ADVANCED LIVE BENCHMARK FOR 006 LLM AGENTS IN FUTURE PREDICTION 007 008 009

010 **Anonymous authors**
011 Paper under double-blind review
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029

030 ABSTRACT 031

032 Future prediction is a complex task for LLM agents, requiring a high level of analytical thinking, information gathering, contextual understanding, and decision-making under uncertainty. Agents must not only gather and interpret vast amounts of dynamic information but also integrate diverse data sources, weigh uncertainties, and adapt predictions based on emerging trends, just as human experts do in fields like politics, economics, and finance. Despite its importance, no large-scale benchmark exists for evaluating agents on future prediction, largely due to challenges in handling real-time updates and retrieving timely, accurate answers. To address this, we introduce **FutureX**, a dynamic and live evaluation benchmark specifically designed for LLM agents performing future prediction tasks. FutureX is the largest and most diverse live benchmark for future prediction, supporting real-time daily updates and eliminating data contamination through an automated pipeline for question gathering and answer collection. We evaluate 25 LLM/agent models, including those with reasoning, search capabilities, and integration of external tools such as the open-source Deep Research Agent and closed-source Deep Research models. This comprehensive evaluation assesses agents' adaptive reasoning and performance in dynamic environments. Our goal is to establish a dynamic, contamination-free evaluation standard that drives the development of LLM agents capable of performing at the level of professional human analysts in complex reasoning and predictive thinking.
033

034 1 INTRODUCTION 035

036 The rapid evolution of Large Language Models (LLMs) has catalyzed a fundamental shift in the
037 landscape of artificial intelligence, moving from the generation of coherent text to the creation of
038 autonomous agents capable of complex, goal-oriented behavior (Chase, 2022; Yao et al., 2023; Sig-
039 nificant Gravitas; Park et al., 2023; Roucher et al., 2025b). This transition from passive text gener-
040 ators to active problem-solvers necessitates a corresponding evolution in evaluation methodologies.
041 While foundational benchmarks like MMLU (Hendrycks et al., 2021) and SuperGLUE (Wang et al.,
042 2019) are instrumental in assessing the static knowledge of LLMs, they are insufficient for measur-
043 ing what a model can do when deployed as part of an interactive, goal-seeking system. An agent's
044 performance is defined not just by its underlying model, but by its ability to plan, use external tools,
045 and adapt to a dynamic environment.
046

047 In response, a new generation of agent-centric benchmarks has emerged, primarily focused on eval-
048 uating search (Wei et al., 2025b), tool usage (Yao et al., 2025), and coding skills in controlled or
049 simulated settings (Jimenez et al., 2024). While these benchmarks offer valuable insights into agent
050 capabilities, they largely address static, well-defined problems whose solutions are already known.
051 Further, they fail to address a critical gap: the ability to synthesize dynamic, real-world informa-
052 tion, process it, and perform complex analysis and reasoning, which human experts possess across
053 different domains. Future prediction, in fact, directly addresses these two drawbacks. This task di-
rectly tests an agent's ability to integrate dynamic, real-world information, process it in context, and
generate complex analysis and reasoning about problems whose answers are not yet known to the
world. Such tasks naturally involve a dynamic element, and their primary significance lies in preparing
agents to anticipate and navigate genuinely novel scenarios, mirroring the foresight applied by
human experts across diverse domains.
054

In response to this need, we introduce **FutureX**, a dynamic and live evaluation benchmark specifically designed for LLM agents performing future prediction tasks. FutureX is built upon a semi-automated pipeline that continuously collects future-oriented questions from 195 diverse websites, curated from a pool of 2,008 sites covering areas such as politics, economics, technology, sports, healthcare, and more. This curation process involves both LLM-based agents and human experts, a necessary combination to ensure quality. Each event is associated with a start date (several days prior to the resolution date) and a resolution date. The pipeline automatically collects and stores agents' predictions on the start date. After the resolution date passes, the system dynamically crawls the web to retrieve the ground-truth outcome and scores the agent's prior predictions. FutureX provides four key advantages that directly address the limitations of existing benchmarks:

- **Large-Scale and Broad Domain Coverage:** Using a semi-automated pipeline for question collection and filtering, we currently select 195 websites from a pool of 2,008 as our sources. These selected websites cover a wide range of topics—including politics, economics, finance, sports, and entertainment—making it, to our knowledge, the *largest and most diverse live benchmark* for future prediction.
- **Real-Time Updates:** FutureX continuously collects future-oriented questions from 128 websites, with daily updates to ensure real-time relevance. By dynamically crawling questions and answers, the benchmark maintains both timeliness and diversity in the questions, presenting a unique challenge for LLM agents to process and adapt to constantly evolving information.
- **No Data Contamination:** Focusing exclusively on future events, FutureX inherently eliminates any risk of data contamination, preventing any exploitation of historical information to manipulate the results.
- **Comprehensive & Automated Assessment of LLM Agents:** Building on FutureX, we have developed a fully automated evaluation pipeline that updates future questions daily, runs various LLM agents for each event on its start date, collects event outcomes after the resolution date, and evaluates agents' performance. The models under evaluation include base LLMs, LLMs with reasoning and search capabilities, open-source Deep Research Agents, and closed-source Deep Research Agents, for a total of 25 models.

In addition to the overall results, we conduct an in-depth analysis, including:

- How do LLM agents perform on questions of varying difficulty levels and across different domains? (Section 4.2 and G.1)
- What factors (such as the type of LLM model, agent framework, and question domain) have the most statistically significant impact on performance? (Appendix G.2)
- How do LLM agents perform when making predictions after the resolution date? (Appendix G.3)
- How planning and search capabilities affect the performance? (Appendix G.4 and G.5)

To pave the way of the “Second Half of AI¹”, we firmly believe FutureX has great potential to unlock new research directions for developing LLM agents capable of performing at the level of professional human analysts in real-world, high-stakes domains.

2 RELATED WORK

A significant gap in current LLM agent evaluations is their ability to handle dynamic, real-world information and reason about future events, which are core skills human professional analysts routinely apply across finance, business, politics, and technology trend analysis. Future prediction serves as a critical test for these capabilities, demanding agents to gather up-to-date evidence and anticipate outcomes in an open-ended environment. However, building a reliable benchmark for forecasting presents unique challenges: unlike static question-answering, future events are inherently uncertain and cannot be easily verified in advance. Previous efforts to evaluate LLMs on forecasting, such as backtesting with historical data (Wildman et al., 2025), risk introducing temporal leakage and retrieval contamination due to retrospective information influencing internet searches on past events (Paleka et al., 2025). Furthermore, many existing benchmarks (Guan et al., 2024; Karger et al., 2025; Nako & Jatowt, 2025) focus on evaluating LLMs without search capabilities, which is impractical for real-world future prediction. Benchmarks like those proposed in Guan et al. (2024) and Nako & Jatowt (2025) are also one-time collected and lack the live updates necessary for dynamic events. While ForecastBench (Karger et al., 2025) attempts to address the “future event”

¹<https://ysymyth.github.io/The-Second-Half/>

108 Table 1: Comparison with Previous Benchmarks for Future Prediction. A ✓ in the Live Update
 109 column indicates that a benchmark supports this feature, though may not update regularly. A ✓ in
 110 the LLM Agents column for FutureBench reflects evaluation of only a single open-source agent. In
 111 contrast, ✓✓✓ to denote regular updates and comprehensive coverage of multiple models.

	#Events	#Domain	Live Update	Time	Source	Data			Evaluation		
						LLM Agents	Env.	Frequency	Auto		
ForecastQA (Jin et al., 2021)	10392	-	✗	Past	21 News Websites	✗	Sim.	One-Time	✗		
Autocast (Zou et al., 2022)	6707	5	✗	Past & Future	3 Prediction Markets	✗	Sim.	One-Time	✗		
OpenEPBench (Guan et al., 2024)	983	-	✗	Future	2 News Websites	✗	Sim.	One-Time	✗		
NaviTomorrow (Nako & Jatowt, 2025)	5000	-	✗	Past	4 News APIs	✗	Sim.	One-Time	✗		
ForecastBench (Karger et al., 2025)	6402	8	✓	Future	4 Prediction Markets 5 Databases	✗	Sim.	Monthly	✓		
FutureBench (Together.ai, 2025)	42	-	✓	Future	1 Prediction Markets Several News Websites	✓	Real	Weekly	✓		
FutureX	~500/week	11	✓✓✓	Future	195 Websites	✓✓✓	Real	Daily Weekly	✓		

121 challenge by using only questions about future outcomes, it predominantly evaluates vanilla LLMs,
 122 and relies on prediction market events, dominated by multiple-choice questions. This limits both
 123 the diversity of events included and, critically, the assessment of an agent’s ability to perform open-
 124 ended, real-world information gathering. Similarly, FutureBench (Together.ai, 2025) is restricted to
 125 events from PolyMarket and includes a very small number of events (~30). This limited diversity
 126 in current future prediction benchmarks highlights the inherent difficulty in collecting and evaluating
 127 such events, particularly those with unknown answers or those that have not yet transpired.

128 Although challenging, we adopt this direction because it ensures the absence of information contam-
 129 ination and directly evaluates how effectively an LLM-agent can synthesize real-time data, reason
 130 under uncertainty, and predict future events—capabilities that represent the next frontier for expert-
 131 level AI agents, which is exactly what our proposed FutureX seeks to achieve. More related works
 132 on general LLM agent benchmarks are discussed in Section B.

3 FUTUREX

136 The goal of FutureX is to provide a dynamic, comprehensive, and contamination-free evaluation
 137 of LLM agents’ advanced search and reasoning capabilities, aiming to match or even surpass the
 138 expertise of human professionals. In this section, we will introduce the construction process of
 139 FutureX, as well as its core features.

3.1 DESIGN PRINCIPLES OF FUTUREX

142 To clearly articulate the design philosophy of FutureX and draw a sharp contrast with other bench-
 143 marks, we demonstrate four core design dimensions.

- 144 • **Eliminating Data Contamination.** The design philosophy of FutureX inherently solves the
 145 data contamination problem. By defining its core task as “future prediction”, FutureX guarantees
 146 that the ground-truth answers to all questions have not yet occurred at the time of the agent’s
 147 prediction, making it impossible for them to exist in any model’s training data.
- 148 • **Simulating Real-World Challenges and Evaluating Core Intelligence.** FutureX places the
 149 agent directly into the real world’s information flow (e.g., economic fluctuations, technological
 150 breakthroughs, political elections). This task is holistic, demanding a suite of advanced cognitive
 151 skills, including information gathering, data synthesis, probability weighing, and causal reasoning.
- 152 • **Large-Scale and Cross-Domain Comprehensive Coverage.** FutureX is currently the *largest*
 153 and *most diverse* live benchmark for future prediction. Through a semi-automated data pipeline,
 154 we curate and filter information from 195 high-quality sources, selected from a pool of over 2,000
 155 websites. These sources cover a wide array of domains, including politics, economics, finance,
 156 technology, sports, and entertainment.
- 157 • **Dynamic and Automated Evaluation Process.** One of the core values of FutureX lies in its
 158 highly automated, dynamic, and closed-loop evaluation process. On each event’s designated start
 159 date, it automatically runs the various agent models and stores their predictions. Once the event’s
 160 resolution date has passed, the system again automatically crawls the web to obtain the ground-
 161 truth outcome and scores the agents’ prior predictions. This entire process operates without man-
 162 ual intervention, ensuring the evaluation’s timeliness, objectivity, and scalability.

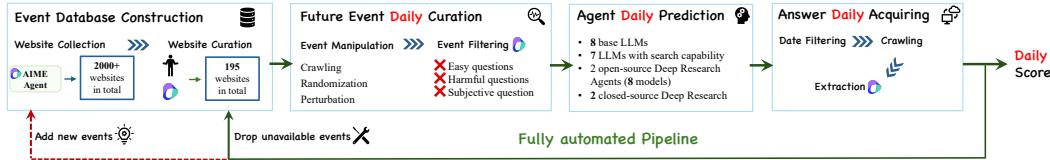


Figure 1: The overall pipeline of FutureX, which consists of event database construction, future event daily curation, answer daily acquisition. The entire pipeline is fully automated and operates on a daily basis.

Overview of FutureX. Following our design principles, FutureX is a live-updating benchmark for future prediction that covers a broad range of source websites and domains. With daily and weekly updates, it features an automated pipeline for event collection, curation, and agent evaluation—all running smoothly and reliably. As shown in Table 1, FutureX demonstrates clear advantages over previous benchmarks from both data and evaluation perspectives, supporting a much broader range of events and a more diverse set of LLM agents for evaluation. Specifically, recent live benchmarks (Karger et al., 2025; Together.ai, 2025) primarily rely on prediction market websites for live updates. However, as demonstrated in Section 3.3, these events tend to be relatively simple, and many involve subjective questions that are not well-suited for rigorous evaluation. In sharp contrast, FutureX collects and curates events from a much broader range of sources to ensure a challenging and high-quality set of evaluation tasks. Moreover, FutureX evaluates 25 models across four different categories, which, to our knowledge, is the first comprehensive benchmark for LLM agents in the domain of future prediction. In comparison, FutureBench (Together.ai, 2025) evaluates only a single open-source agent with a few LLMs.

3.2 CONSTRUCTION OF FUTUREX

As shown in Figure 1, FutureX is an automated, live benchmark that operates on a daily cycle, encompassing four stages: event database construction, future event daily curation, agent daily prediction, and answer daily acquisition. *Each stage is processed on a daily basis.*

3.2.1 EVENT DATABASE CONSTRUCTION

This phase consists of website collection and website curation. During *website collection*, we begin by using the AIME agent (Shi et al., 2025) to gather a large number of website URLs relevant to domains such as politics, economics, finance, technology, and sports, with a total of 2,008 websites. Then for *website curation*, these URLs are then subjected to an initial LLM-based check, utilizing a combination of Seed1.5-Thinking (Seed et al., 2025) and DeepSeek-R1 (Guo et al., 2025). This process performs tasks such as deduplication, assessing the suitability of the websites for question generation, and evaluating their update frequency, reducing the number of websites to 636. The remaining URLs are then manually reviewed, focusing on identifying reliable sources, particularly ranking lists and websites with high update frequency, ensuring that only the most relevant and up-to-date information is included. After this manual review, our initial version has **195** high-quality websites as our current event database, including five types:

- Prediction market websites: Websites that allow users to bet on or predict the outcomes of future events, including political events, sports outcomes, or financial market movements. Examples include gjopen, Polymarket, etc.
- News websites: Websites that provide up-to-date news, analysis, and market-moving events, such as earnings reports, economic policy changes, sports news, technology trends, and geopolitical developments. Examples include Reuters, ESPN, etc.
- Entertainment ranking websites: Websites that provide rankings related to music, movies, TV shows, and other entertainment forms. These rankings are often based on user reviews, sales data, critical acclaim, or popularity metrics. Examples include Billboard, IMDb, etc.
- Government websites: Official websites that provide economic data, regulations, and reports from governments. These include statistical data like GDP, unemployment rates, inflation, fiscal policies, and other public economic indicators. Examples include U.S. Bureau of Economic Analysis, China Agricultural and Rural Information Site, etc.
- Real-time data platforms: Platforms that provide real-time or near-real-time data on various financial markets. These platforms are used to monitor live stock prices, forex rates, cryptocurrency

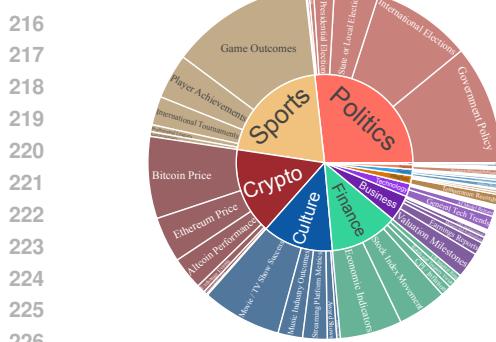
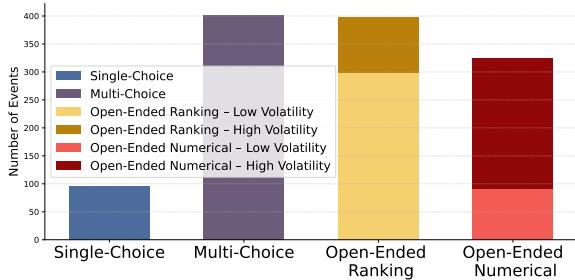
216
217
218
219
220
221
222
223
224
225
226
227
228
Figure 2: Domains of 1,272 events.

Figure 3: Event type distribution.

229 values, and other asset prices, offering instant updates to help with timely trading decisions. Examples include Shenzhen Stock Exchange Site, Sina Finance, etc.

230
231 In addition, as shown in Figure 1, this event database is updated daily to remove events with un-
232 available outcomes and continuously add new events using current high-quality websites as “seeds”.
233 Based on these, we perform daily future event curation to generate prediction questions for each day,
234 including *event manipulation* and *event filtering*, and details are shown in Appendix C.

235 3.2.2 ANSWER DAILY ACQUISITION

236 Answer acquisition is a critical phase in FutureX. While we can pose questions about a wide range
237 of future events, the success of evaluation ultimately depends on whether we can reliably obtain the
238 corresponding answers. To this end, much of our effort in constructing the event database (see Sec-
239 tion 3.2.1) focuses on ensuring answer availability. We carefully select high-quality websites that
240 consistently provide verifiable outcomes on a daily or weekly basis. Building on this, our pipeline
241 automatically retrieves answers each day by following the procedures outlined below.

- 242 1. **Date Filtering.** We filter events whose resolution date aligns with the current day.
- 243 2. **Website Crawling.** With the daily answer acquisition set, we then crawl the corresponding
244 website and extract the core content.
- 245 3. **Answer Extraction.** Based on the core content retrieved from each website, we use the
246 Seed1.5-Thinking model (Seed et al., 2025) to extract the precise answer.

247 In addition, as expected during the initial stages, various types of failure cases arise. To systemati-
248 cally address them, we categorize failures into two types: crawling errors and extraction errors. For
249 crawling errors, such as those caused by anti-bot measures, we update our event database to exclude
250 websites that are no longer accessible. For extraction errors, such as incorrect or missing answers,
251 we manually review the cases and design customized prompts to improve answer extraction accu-
252 racy. With these efforts, the answer acquisition success rate exceeds **97%**, supporting an efficient
253 and fully automated evaluation pipeline that runs on a daily basis.

254 3.3 DATA OF FUTUREX

255 **Comprehensive Domain Coverage.** As introduced in Section 3.2, FutureX achieves comprehen-
256 sive domain coverage through the daily curation of future events from 195 high-quality websites.
257 Between July 20th and August 3rd, our dataset comprises 1,272 events, systematically categorized
258 into 11 main domains—including politics, sports, crypto, culture, finance, business, technology
259 trends, weather, health, and space—each further refined into several sub-categories. As depicted in
260 Figure 10, the distribution across these domains is notably well-balanced, which facilitates a robust
261 and comprehensive evaluation of LLM agents across a multitude of real-world scenarios. We show
262 the examples in major domains in Table 4.

263 **Different Event Types.** FutureX includes four different event types: single-choice, multi-choice,
264 open-ended ranking, and open-ended numerical prediction events. *Single-choice* events require se-
265 lecting one correct answer from options fewer than 4. *multi-choice* events involve identifying multi-
266 ple correct answers, making them inherently more challenging. In addition to events with predefined
267 choices, FutureX features a significant number of *open-ended* events, where no options are provided.
268 There are two types of open-ended events: (1) *ranking* tasks, which typically involve forecasting the
269 order of items on a future leaderboard (e.g., music or movie popularity rankings), and (2) *numerical*

270
271 Table 2: Difficulty tiers and assessed agent’s skills in FutureX.
272
273

Level	Tier	Event Type	Focus	Assessed Agent’s Skills		
				Planning	Reasoning	Searching
1	Basic	Few choices	Choose from fewer than 4 options from a given list.	Weak	Weak	Weak
2	Wide Search	Many Choices	Exhaustive discrimination and Return <i>all</i> correct futures.	Weak	Medium	Medium
3	Deep Search	Open-ended (low volatility)	Interactive search & synthesis Navigate sources (click, scroll, filter) Integrate evidence for an answer.	Medium	Medium	Strong
4	Super Agent	Open-ended (high volatility)	Forecast high-volatility, open-ended events Conduct wide-scope information search Reason and predict under deep uncertainty “Super-agent” tier	Strong	Strong	Strong

284 prediction tasks, which require estimating a specific numeric value (e.g., a stock price or price index). As detailed in Appendix C, we significantly downsampled binary yes-or-no events—primarily
285 affecting single-choice questions—to increase the difficulty of the benchmark. As shown in Figure
286 11, the distribution of the remaining three event types is relatively balanced.

287 **Different Volatility.** While all events in FutureX involve forecasting future outcomes, they vary
288 significantly in how dynamic these outcomes are over time. We analyze the *volatility* of each event
289 based on historical data, measuring how much the target outcome is expected to fluctuate over time.
290 We then tag open-ended events as either “Low Volatility” or “High Volatility”, which are visually
291 indicated by different color depths in Figure 3.

292 **Difficulty Tiers.** Guided by event type and expected volatility (see Figure 11), we partition the
293 benchmark into four progressively harder tiers, basic, wide search, deep search, and super agent,
294 that correspond to the agent capabilities assessed in Levels 1 through 4. An overview of the 4 tiers
295 and the specific agent skills they assess is given in Table 2. The examples of events corresponding
296 to these 4 tiers are shown in Table 3

297

- The **Basic tier** (Level 1) contains single-choice events with options fewer than 4. The predefined
298 options sharply limit the search space, so information retrieval and reasoning are lightweight.
- The **Wide Search tier** (Level 2) comprises multi-choice events with several correct answers. The
300 agent must submit the full set of valid options and nothing more. This tier requires more complex
301 reasoning, and therefore tests whether the agent can perform *exhaustive yet precise discrimination*
302 across multiple plausible options.
- The **Deep Search tier** (Level 3) contains open-ended events whose underlying facts are relatively
303 stable. With no options provided, the agent must propose its own answer, performing multi-
304 step search and reasoning to gather evidence. Because volatility is low, exhaustive information
305 collection should converge on the correct response. This tier thus probes the agent’s ability to
306 *navigate, integrate, and synthesize* reliable information.
- The **Super Agent tier** (Level 4) covers high-volatility, open-ended events. Here the agent must
307 cast a wide net for information and reason probabilistically under shifting signals and deep un-
308 certainty. The task is taxing even for human experts—let alone machines—because the scenarios are
309 complex, ambiguous, and resist simple fact retrieval. This tier therefore probes an agent’s “super”
310 capacity for *nuanced, uncertainty-aware forecasting* in the most demanding real-world settings.

311 Notably, all Level 3 and Level 4 events are generated through our automated pipeline (see Figure 9),
312 which supports scalable event creation while maintaining quality control. This marks a key dis-
313 tinction from prior benchmarks (Karger et al., 2025; Together.ai, 2025), where most events were
314 relatively simple and collected directly from prediction market websites (see Table 1).

317 3.4 EVALUATION PROTOCOL OF FUTUREX

318
319 Unlike traditional static benchmarks, where each query is associated with a known answer, future
320 prediction inherently lacks ground truth at the time of prediction, since the relevant events have not
321 yet occurred. As a result, FutureX introduces an evaluation delay, referring to the time gap between
322 when a prediction is made and when it can be evaluated. For example, suppose an agent makes
323 predictions for several events on July 15th, with resolution dates ranging from July 16th to July
22nd. In this case, performance can only be evaluated after July 22nd, once all outcomes are known,

324 Table 3: Examples of different levels, where the specific date can be replaced with any future date.
325

326 Level	327 Example Events
1. Basic	Ethereum Up or Down on August 20, 2025?
2. Wide Search	Who will win the King of the Mountains / Polka-dot Jersey at the 2025 Tour de France A. Tadej Pogacar', B. Other', C. Jonas Vingegaard, ...
3. Deep Search	Which movies will be in the top 10 of Maoyan Movie Ticketing Rating List as of Beijing Time August 20, 2025.
4. Super Agent	What is the daily purchase transaction amount (in billion - yuan) in the daily transaction information of the Shanghai - Hong Kong Stock Connect on August 20, 2025, Beijing Time?

334 resulting in an evaluation delay of one week. FutureX adopts an *one-week* prediction window, which
335 provides both sufficient event coverage and manageable evaluation latency.
336

337 We use the exact-match and F1-score to evaluate the level-1 and level-2 questions, for the level-3
338 and level-4 questions, we adopt LLM-as-Judge (equation 1) and the clipped MSE score (equation 2)
339 as the evaluation metric. All details can be found in Appendix E.
340

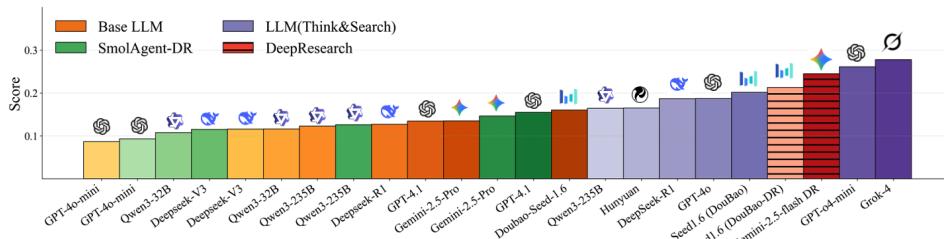
341 4 EXPERIMENTS

342 In this section, we present the main results for FutureX from July 20th to August 3rd². The results
343 in this section cover 25 models, including:

- 344 • **Base LLMs (8 models):** Open-source and closed-source LLMs without tool usage, including *Gemini-2.5-pro*, *DeepSeek V3*, *GPT-4o-mini*, *GPT-4.1*, *DeepSeek R1*, *Qwen3-32B*, *Qwen3-235B*, and *Doubaot-Seed1.6-Thinking*.
- 345 • **SmolAgent for Deep Research (Roucher et al., 2025a) (6 models):** SmolAgent is evaluated
346 with various backbone LLMs, including *Gemini-2.5-pro*, *GPT-4.1*, *GPT-4o-mini*, *Qwen3-235B*,
347 *Qwen3-32B*, and *DeepSeek V3*. Among reasoning models, only *Gemini-2.5-pro* is included here,
348 as others—such as *Doubaot-Seed1.6-Thinking*, *GPT-03/o4-mini*, and *DeepSeek R1*—incur significantly
349 longer runtimes and are therefore currently excluded.
- 350 • **AgentOrchestra (Zhang et al., 2025b) (2 models):** Tested with two backbone LLMs—*Gemini-2.5-pro* and *GPT-4.1*. Due to the complexity of this agent framework and limited compatibility,
351 only these representative models are included for now.
- 352 • **LLMs (Think&Search) (7 models):** Evaluation of advanced commercial LLM services with integrated
353 thinking and searching capabilities, including *Doubaot*, *DeepSeek R1*, *Hunyu*, *Qwen3-235B*, *GPT-4o*,
354 *GPT-04-mini*, and *Grok-4*. *Gemini-2.5-pro* (Think&Search) is on the way.
- 355 • **Deep Research Models (2 models):** Includes top-tier closed-source models tailored for deep
356 research tasks: *Doubaot* and *Gemini Deep Research* (with *Gemini-2.5-flash*).

357 Note that weekly results (till now) of FutureX online competition are shown in Appendix A, with
358 much more advanced models like GPT5-pro, ChatGPT Agent, Claude 4.1 Opus, Gemini Deep
359 Think, etc.

360 4.1 OVERALL RESULTS



374 Figure 4: Overall scores on FutureX between July 20th and August 3rd.

375 As for the overall score, we combine scores from the 4 difficulty tiers (see Table 2) using weights
376 of 10%, 20%, 30%, and 40%, respectively, with heavier weights assigned to the more challeng-
377

378 ²The weekly-updated results after August 3rd are shown in Appendix A.

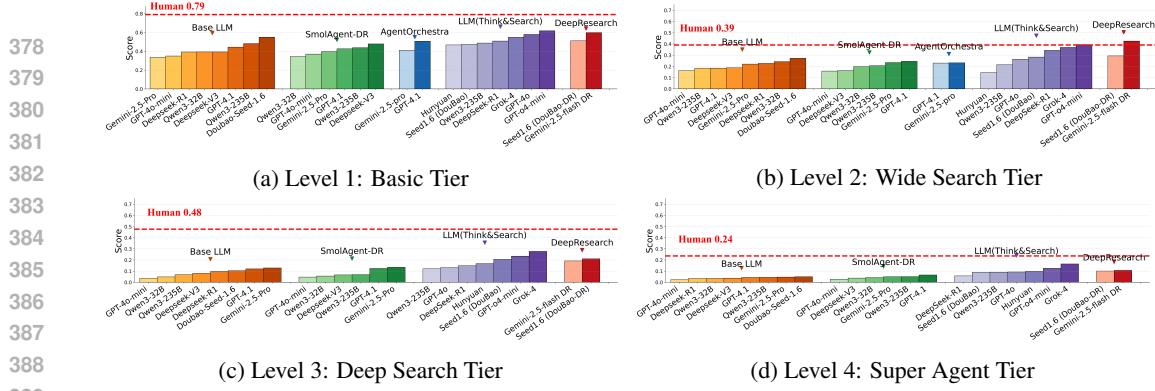


Figure 5: Overall results of different difficulty tiers (between July 20th and August 3rd). Note that since AgentOrchestra is computationally intensive, we evaluate it with only two backbone models for only Level 1 and 2 events.

ing tiers. The overall results are shown in Figure 4, where models of the same type are represented using similar colors for clarity. As shown in Figure 4, across the four model types, **Grok-4** achieves the highest overall performance, followed by **Gemini-2.5-flash Deep Research**, **GPT-o4-mini (Think&Search)**, and **Seed1.6 (DouBao)**. Generally, reasoning models equipped with search capabilities outperform the rest, underscoring the importance of advanced search and reasoning in FutureX. Moreover, SmolAgent-DR (Roucher et al., 2025a) underperforms compared to LLM (Think&Search), likely reflecting differences in their search API capabilities.

4.2 RESULTS ACROSS DIFFICULTY TIERS

We show separate results across 4 difficulty tiers in Figure 5. Our main findings are as follows:

Finding 1. Our difficulty tiers accurately reflect the complexity of the events. We observe a clear, consistent decline in performance across the 4 defined difficulty tiers, which strongly supports the validity of our task stratification. Models achieve high accuracy on Levels 1 and 2, but their performance drops significantly on Level 3 and, in particular, on Level 4. This trend indicates that our difficulty labels effectively correspond to increasing levels of task complexity. Furthermore, by comparing Figure 14 and Figure 15, we find that even within the same domain, model performance declines substantially.

In particular, Level 4 events, which are open-ended and highly volatile, pose significant challenges for current models. These tasks often demand multi-step reasoning, synthesis of ambiguous or incomplete information, long-horizon forecasting, and a deeper understanding of world knowledge and strategic thinking. In our experiments, most models struggle to generate correct responses for these events, and even the strongest models often fail to score at all. In fact, these tasks are so complex that they not only test the limits of current models, but may also serve as a benchmark for measuring superhuman performance in future systems.

Finding 2. Base LLMs perform well on Level 1 and Level 2 events. Level 1 and level 2 events are designed as relatively simple single/multi-choice questions. Our results show that even base LLMs (without tool usage or advanced reasoning capabilities) consistently achieve high accuracy on these tasks. Notably, *DouBao-Seed1.6-Thinking* outperforms several agents equipped with web search tools, including the two Deep Research agents. These events may primarily rely on basic factual recall or straightforward reasoning, which base LLMs are already capable of handling without external tools.

This finding also suggests that Level 1 and Level 2 events are *not sufficiently challenging to distinguish* between models of varying capabilities. While they are useful for establishing a performance baseline, they offer limited insight when evaluating more advanced language models.

Finding 3. Search/tool usage becomes increasingly important for harder events. As the complexity of the events increases, particularly in Level 3, models that incorporate external tools such as web search, calculators, or code execution tend to perform significantly better than those that rely solely on static knowledge. This highlights the critical role of tool-augmented reasoning in handling complex, multi-step problems that cannot be solved through pre-trained information alone.

432 This further demonstrates that FutureX is capable of analyzing advanced search capabilities that are
 433 closely linked to reasoning.
 434

**Finding 4. DouBao-Seed1.6-Thinking excels in knowledge retrieval (Level 1 and Level 2), and
 435 Grok-4 demonstrates exceptional performance on more difficult events (Level 3 and Level 4).**

436 Among base LLMs, we find that *DouBao-Seed1.6-Thinking* performs the best on Level 1 and Level
 437 2 events. Notably, it even outperforms agents equipped with search tools as well as deep research
 438 models. This suggests that, when provided with answer options, *DouBao-Seed1.6-Thinking* is highly
 439 effective at retrieving and applying its internal knowledge to make accurate predictions about future
 440 events, demonstrating strong inherent reasoning capabilities.
 441

442 In contrast, among all evaluated models, *Grok-4* stands out on the most challenging tasks. Re-
 443 markably, it surpasses even premium models such as Gemini Deep Research in both accuracy and
 444 efficiency. Despite operating with fewer searching and faster inference speeds, *Grok-4* and *GPT-
 445 o4-mini* achieve top-tier results, highlighting an impressive balance between reasoning strength and
 446 runtime efficiency.
 447

Finding 5: LLM agents still lag behind humans. In addition to automated model evaluations, we
 448 conducted a human annotation study with 40 industry experts. These participants comprised current
 449 or former employees from the Big Four accounting firms (e.g., KPMG), top consulting firms (e.g.,
 450 McKinsey), and nine leading investment banks (e.g., UBS). We randomly sampled 300 questions
 451 from our test bank and had these experts answer them independently, then computed their average
 452 scores on the same evaluation metrics (red dotted lines).
 453

454 As shown in Figure 5, humans significantly outperform LLM agents on Level 1, Level 3, and Level
 455 4 events, indicating that LLMs still have considerable ground to cover before matching human ex-
 456 pertise. Conversely, for Level 2 events, some models actually surpass human performance; this may
 457 be because these multi-choice questions involve so many options that people often cannot exhaus-
 458 tively compare every possibility. Overall, these results underscore the substantial potential for LLM
 459 agents to aid (and eventually rival) humans in forecasting future events.
 460

4.3 MORE CASE STUDIES

461 In addition to our main results, we design and conduct multiple case studies, including:

1. **Results across Different Domains.** Differnt models have different strengths, for example, GPT models are good at Crypto. Details in Appendix G.1.
2. **Factor Analysis.** We conduct a linear analysis to discover important factors with statistical significance. From it, the “difficulty level” and “event domain” factors are identified as the most critical factor that influence the performance of agents. Details in Appendix G.2
3. **Past vs. Future Prediction.** By comparing historical event searching and future prediction, we could directly characterize the search capability of different models, where Grok-4 exhibits excellent performance. Details in Appendix G.3
4. **Analysis of Agent Trajectories.** The number of tool-call, comprehensiveness of planning and the credibility of source websites all significantly influence the agents’ performance in future prediction. Details in Appendix G.4
5. **Search Analysis.** We count the average number of searches for each model, where we find strong models typically conduct more searches, and Grok-4 searches even more than deep research models. Details in Appendix G.5

477 More recent results of our weekly competition are shown in Appendix A, and a comprehensive
 478 review of related works can be found in Appendix B.
 479

480 5 CONCLUSION

481 FutureX is the first live benchmark that tests LLM agents on real-world future prediction tasks.
 482 Going forward, FutureX offers a flexible platform for improving LLM agents. We are actively
 483 working on adding new domains and data sources to FutureX. By keeping the benchmark live and
 484 diverse, we aim to push agents closer to the level of human experts in making timely, strategic
 485 predictions across a wide range of fields.
 486

486

6 ETHICS STATEMENT

487
488 This work did not involve animal subjects or personally identifiable data. All information used are
489 publicly available and were accessed in accordance with their licenses and terms of use.
490491

7 REPRODUCIBILITY STATEMENT

492
493 All information necessary to reproduce our results is included, covering benchmark construction
494 procedures, the test date, evaluation metrics, and the set of models under evaluation.
495496

REFERENCES

497
498 Harrison Chase. LangChain, October 2022. URL <https://github.com/langchain-ai/langchain>.
499500
501 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li,
502 Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot
503 arena: An open platform for evaluating llms by human preference. In *Forty-first International
504 Conference on Machine Learning*, 2024.505
506 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit
507 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
508 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
509 bilities. *arXiv preprint arXiv:2507.06261*, 2025.510
511 Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. Openep: Open-ended future event
512 prediction. *arXiv preprint arXiv:2408.06578*, 2024.513
514 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
515 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
516 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.517
518 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
519 Steinhardt. Measuring massive multitask language understanding. In *International Conference
520 on Learning Representations*, 2021.521
522 Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando
523 Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free
524 evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.525
526 Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R
527 Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth
528 International Conference on Learning Representations*, 2024.529
530 Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang
531 Ren. Forecastqa: A question answering challenge for event forecasting with temporal text data. In
532 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the
533 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,
534 pp. 4636–4650, 2021.535
536 Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and
537 Philip Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In *The
538 Thirteenth International Conference on Learning Representations*, 2025.539
540 Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gon-
541 zalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and
542 benchbuilder pipeline, 2024.543
544 Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding,
545 Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *The Twelfth Inter-
546 national Conference on Learning Representations*, 2024.

540 Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia:
 541 a benchmark for general ai assistants. In *The Twelfth International Conference on Learning*
 542 *Representations*, 2023.

543

544 Petraq Nako and Adam Jatowt. Navigating tomorrow: Reliably assessing large language models
 545 performance on future event prediction. *arXiv preprint arXiv:2501.05925*, 2025.

546 OpenAI. Introducing SWE-Bench verified, 2024. URL <https://openai.com/index/introducing-swe-bench-verified/>.

547

548 Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language
 549 model forecasters. *arXiv preprint arXiv:2506.00723*, 2025.

550

551 Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang.
 552 Training software engineering agents and verifiers with swe-gym. In *Forty-second International*
 553 *Conference on Machine Learning*, 2025.

554

555 Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and
 556 Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL
 557 <https://arxiv.org/abs/2304.03442>.

558

559 Aymeric Roucher, Albert Villanova del Moral, Merve Noyan, Thomas Wolf, and Clémentine
 560 Fourrier. Open-source deepresearch – freeing our search agents, 2025a. URL <https://huggingface.co/blog/open-deep-research>. Accessed: 2025-07-28.

561

562 Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kau-
 563 nismäki. ‘smolagents’: a smol library to build great agentic systems. <https://github.com/huggingface/smolagents>, 2025b.

564

565 ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang,
 566 Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, et al. Seed1. 5-thinking: Advancing superb rea-
 567 soning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.

568

569 Yexuan Shi, Mingyu Wang, Yunxiang Cao, Hongjie Lai, Junjian Lan, Xin Han, Yu Wang, Jie Geng,
 570 Zhenan Li, Zihao Xia, et al. Aime: Towards fully-autonomous multi-agent framework. *arXiv*
 571 *preprint arXiv:2507.11988*, 2025.

572

573 Significant Gravitas. AutoGPT. URL <https://github.com/Significant-Gravitas/AutoGPT>.

574

575 Together.ai. Futurebench: Evaluating agents’ future prediction capabilities, 2025. URL <https://www.together.ai/blog/futurebench>. Accessed: 2025-07-27.

576

577 Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer
 578 Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language
 579 understanding systems. *Advances in neural information processing systems*, 32, 2019.

580

581 Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won
 582 Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecmp: A simple yet
 583 challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025a.

584

585 Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won
 586 Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecmp: A simple
 587 yet challenging benchmark for browsing agents, 2025b. URL <https://arxiv.org/abs/2504.12516>.

588

589 Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid
 590 Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging,
 591 contamination-limited llm benchmark. In *The Thirteenth International Conference on Learning*
 592 *Representations*, 2025.

593

Jack Wildman, Nikos I Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan
 Schwarz, Lawrence Phillips, et al. Bench to the future: A pastcasting benchmark for forecasting
 594 agents. *arXiv preprint arXiv:2506.21558*, 2025.

594 John Yang, Kilian Leret, Carlos E Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang,
 595 Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. Swe-smith: Scaling data for software
 596 engineering agents. *arXiv preprint arXiv:2504.21798*, 2025.

597

598 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.
 599 React: Synergizing reasoning and acting in language models, 2023. URL <https://arxiv.org/abs/2210.03629>.

600

601 Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. *tau-bench*: A benchmark
 602 for \underline{T}ool-\underline{A}gent-\underline{U}ser interaction in real-world domains.
 603 In *The Thirteenth International Conference on Learning Representations*, 2025.

604

605 Daoguang Zan, Zhirong Huang, Wei Liu, Hanwu Chen, Linhao Zhang, Shulin Xin, Lu Chen, Qi Liu,
 606 Xiaojian Zhong, Aoyan Li, et al. Multi-swe-bench: A multilingual benchmark for issue resolving.
 607 *arXiv preprint arXiv:2504.02605*, 2025.

608

609 Linghao Zhang, Shilin He, Chaoyun Zhang, Yu Kang, Bowen Li, Chengxing Xie, Junhao Wang,
 610 Maoquan Wang, Yufan Huang, Shengyu Fu, et al. Swe-bench goes live! *arXiv preprint arXiv:2505.23419*, 2025a.

611

612 Wentao Zhang, Ce Cui, Yilei Zhao, Yang Liu, and Bo An. Agentorchestra: A hierarchical multi-
 613 agent framework for general-purpose task solving. *arXiv preprint arXiv:2506.12508*, 2025b.

614

615 Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng,
 616 Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building
 617 autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024.

618

619 Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob
 620 Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks.
 621 *Advances in Neural Information Processing Systems*, 35:27293–27305, 2022.

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

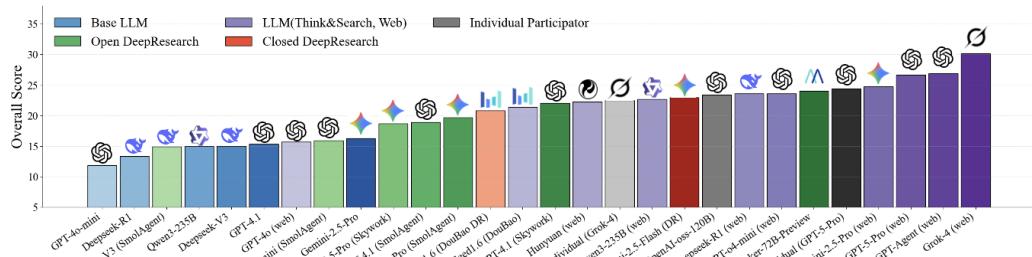
646

647

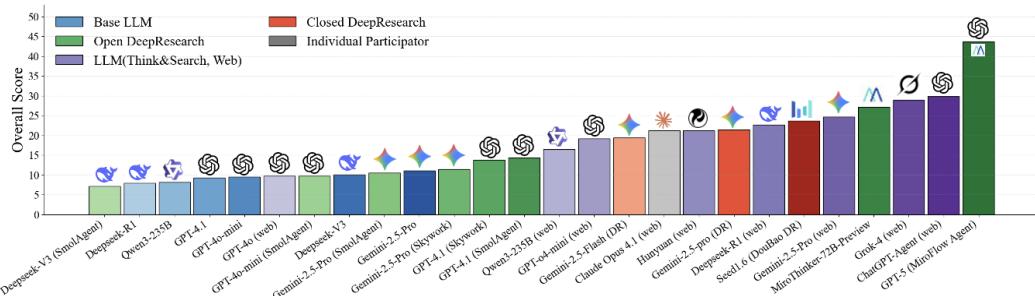
648 A WEEKLY RESULTS OF THE FUTUREX ONLINE COMPETITION

650
651 Following the release of FutureX, we launched a weekly online competition that has attracted par-
652 ticipants from both industry and academia, including MiroMind³, Foresight⁴, Peking University, as
653 well as many independent researchers and practitioners.

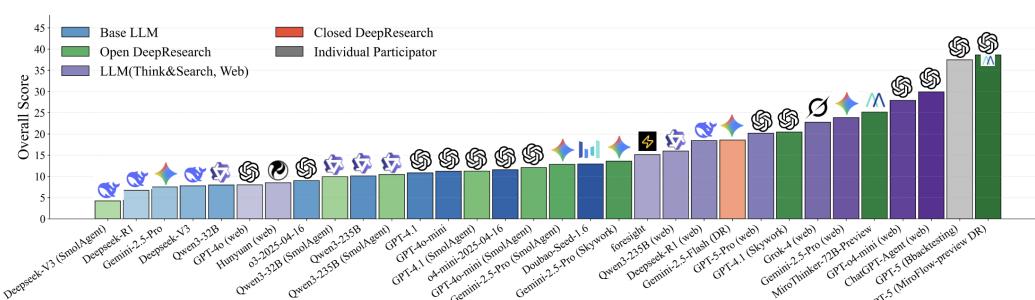
654 In the main body, we report results on 1,272 events conducted between July 20th and August 3rd.
655 In this appendix, we provide complementary results on a weekly basis using more recent data. In
656 order to lower the participation bar, we sample ~100 events every week for the weekly test. For the
657 weekly competition results, in addition to the models evaluated in the main body, we also include
658 GPT-5 Pro (web), ChatGPT Agent (web), Gemini Deep Think (web), and Claude 4.1 Opus.
659



660
661
662
663
664
665
666
667
668
669
670 Figure 6: Results of events between August 27th and September 2nd.
671
672
673



674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701 Figure 7: Results of events between September 3rd and September 9th.
702



702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
22210
22211
22212
22213
22214
22215
22216
22217
22218
22219
22220
22221
22222
22223
22224
22225
22226
22227
22228
22229
22230
22231
22232
22233
22234
22235
22236
22237
22238
22239
22240
22241
22242
22243
22244
22245
22246
22247
22248
22249
22250
22251
22252
22253
22254
22255
22256
22257
22258
22259
22260
22261
22262
22263
22264
22265
22266
22267
22268
22269
22270
22271
22272
22273
22274
22275
22276
22277
22278
22279
22280
22281
22282
22283
22284
22285
22286
22287
22288
22289
22290
22291
22292
22293
22294
22295
22296
22297
22298
22299
22300
22301
22302
22303
22304
22305
22306
22307
22308
22309
22310
22311
22312
22313
22314
22315
22316
22317
22318
22319
22320
22321
22322
22323
22324
22325
22326
22327<br

702 As the competition progresses, we observe that purpose-built agents, such as MiroFlow with GPT-5
 703 and the Bbacktesting Agent with GPT-5⁵, consistently outperform general LLM web products and
 704 deep research agents. This finding indicates that the task of future prediction is, to some extent, both
 705 *learnable* and *optimizable*. Moreover, such results suggest the potential for agents to eventually
 706 surpass human experts, further underscoring the significance of our FutureX benchmark.

712 B COMPREHENSIVE RELATED WORK

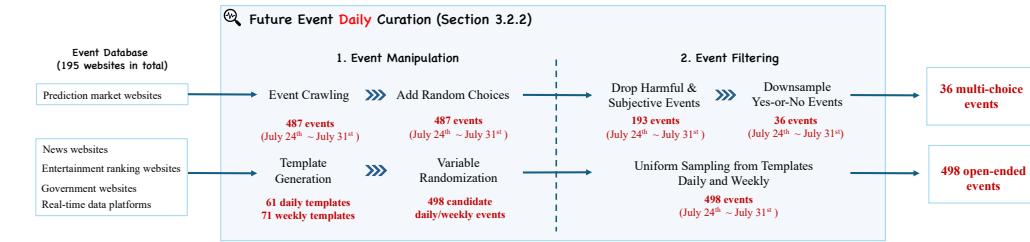
717 In addition to recent benchmarks for future prediction (see Section 2), this section introduce relevant
 718 literature on LLM agent benchmarks, encompassing both established and recent live evaluations.

719 **Agent Benchmark.** A new wave of benchmarks is designed specifically to evaluate LLM-based
 720 agents on complex, interactive tasks rather than isolated questions. For example, AgentBench (Liu
 721 et al., 2024) introduces 8 distinct simulated environments (from operating systems and databases
 722 to web interfaces and games) to assess an agent’s planning, tool use, and decision-making abilities.
 723 WebArena (Zhou et al., 2024) provides high-fidelity simulations of real websites (e-commerce, fo-
 724 rumns, collaborative coding, content management) to test long-horizon web interaction tasks, where
 725 Agents must execute multi-step web browsing actions to accomplish user goals. Another bench-
 726 mark, GAIA (Mialon et al., 2023), focuses on general-purpose assistant capabilities with 466 real-
 727 world questions that require reasoning, multi-modality, web search, and tool use. In the realm of
 728 information retrieval, BrowseComp (Wei et al., 2025a) consists of 1,266 “inverted” questions de-
 729 signed to entangle information and thwart simple lookup. Agents must perform creative, multi-hop
 730 web searches to uncover hidden facts, testing their persistence and strategic search skills beyond
 731 basic fact retrieval. In software engineering, SWE-bench (Jimenez et al., 2024) evaluates agents
 732 on resolving real GitHub issues by generating code patches and verifying them against project test
 733 suites, and based on this, multiple variants are proposed, such as SWE-bench-Verified (OpenAI,
 734 2024), SWE-gym (Pan et al., 2025), Multi-SWE-bench (Zan et al., 2025), and SWE-smith (Yang
 735 et al., 2025). Collectively, these benchmarks offer valuable insights into various aspects of agent
 736 performance—from web navigation and tool use to coding—but they primarily operate in controlled
 737 environments with predefined task scopes and information. Additionally, these benchmarks do not
 738 integrate complex reasoning and information-gathering capabilities, both of which are essential for
 739 our proposed FutureX.

740 **Live Benchmark.** Besides traditional benchmarks with static question sets, live benchmarks have
 741 recently emerged, such as LiveBench (White et al., 2025), LiveCodeBench (Jain et al., 2024) and
 742 SWE-bench-live (Zhang et al., 2025a), which automatically update questions to enable more reli-
 743 able, contamination-free evaluations. Besides, Li et al. (2024) introduce Arena-Hard that is fre-
 744 quently updated from live data in Chatbot Arena (Chiang et al., 2024) to avoid potential over-fitting
 745 or test set leakage. Our proposed FutureX follows this trend, offering a fully automated, live bench-
 746 mark for future prediction—where even the events themselves are live. Compared to the benchmarks
 747 discussed, ours is more aligned with real-world professional analysis scenarios across multiple do-
 748 mains, making it highly practical.

749
 750
 751
 752
 753
 754
 755
⁵These two agents are given by the participants and currently are not open-sourced.

756 **C DETAILS ABOUT DATA CURATION AND FILTERING**



767 Figure 9: The detailed future event daily curation process, which manipulates daily and weekly
 768 events from the event database. As an example, the number of events between July 24th and July
 769 31st remaining after each step is shown in red. Note that the number of events each week varies due
 770 to fluctuations in prediction market events (the first row in the figure).

772 Based on the event templates, we perform daily future event curation to generate prediction questions
 773 for each day. As shown in Figure 9, this process consists of two main steps: event manipulation and
 774 event filtering.

775 **Event Manipulation** This phase involves transforming each website into a concrete future event
 776 format suitable for our pipeline, which varies depending on the type of website.

777 **Prediction market websites** There are already feature future prediction events on prediction market
 778 websites. Therefore, we crawl these events on a daily basis, which are typically *binary or*
 779 *multiple-choice*. For each event, we employ the Seed1.5-Thinking model (Seed et al., 2025) to in-
 780 troduce a set of unrelated (random) choices, thereby increasing the complexity and challenging the
 781 system’s search and reasoning capabilities at a higher level. For instance, if LLM agents were to
 782 query each choice individually, it would significantly reduce efficiency.

783 **Other types of websites** For other websites where future events cannot be directly obtained, we
 784 follow these steps to make more challenging *open-ended questions*:

785 • **Template Generation:** We create an “event template” for each site, which can take variables
 786 (for example, target, date, etc.) as inputs to make the events adaptable over time, even for
 787 the same website. The process is as follows: First, we use an LLM to design candidate
 788 question templates based on the content of each webpage. Next, we specify the input
 789 variables for answer scraping. After the answer is scraped, the LLM checks whether the
 790 answer can be successfully retrieved. To ensure reliability, a human reviewer then verifies
 791 the results and selects the most appropriate question templates. Once an event template
 792 is established, it remains fixed within our pipeline, eliminating the need for recreation in
 793 subsequent iterations. Below are two examples.

794 **Website 1:** <https://www.dongchedi.com/> (China’s largest car review website)
 795 – Variables: rank, date, target
 796 – Event Template: Which car will be ranked {rank} on the {target} board on {date} at Dongchedi?
 797
 798 **Website 2:** <https://www.google.com/finance/>
 799 – Variables: stock, date
 800 – Event Template: What will be the highest point of {stock} on {date}?

801 • **Randomization:** Based on the event templates, in order to prevent asking the same future
 802 event every day, we apply randomization to our templates by varying the input variables
 803 within the same event template (and for the same website). For example, for ranking web-
 804 sites, we may ask the LLM agents to predict different ranks each day or predict ranks within
 805 different sub-ranking lists. For government websites, we may request different statistics or

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
metrics. For real-time data platforms, we may focus on different markets, indexes, or stocks. Below are some examples.

Template 1: Which car will be ranked `{rank}` on the `{target}` board on `{date}` at Dongchedi?

- Q1. Which car will be ranked *1st* on the *SUV Popularity Ranking* board on *September 1st* at Dongchedi?
- Q2. Which car will be ranked *3rd* on the *MPV Sales Ranking* board on *September 15th* at Dongchedi?

Template 2: What will be the highest point of `{stock}` on `{date}`?

- Q1. What will be the highest point of *APPLE* on *September 1st*?
- Q2. What will be the highest point of *NVIDIA* on *September 7th*?

After manipulating the events, we are able to generate ~ 500 daily and weekly future events as candidates out of 195 high-quality websites. Note that each future event is associated with an answer resolution date that will be used in the Answer Daily Acquisition phase (see Section 3.2.2).

Event Filtering For events crawled from the internet (for example, prediction market websites like Polymarket and gjopen), we carefully filter the event set before testing the LLM agents to ensure the validity of the events, where we mainly filter out easy or trivial events, harmful events, and subjective events.

- *Harmful events*: These events include content that may involve discrimination, hate speech, or other harmful factors. Such events can introduce bias or propagate misleading information, undermining the integrity of the predictions. To mitigate this, we use a combination of Seed1.5-Thinking Seed et al. (2025), DeepSeek-R1 Guo et al. (2025), and Gemini-2.5-flash (Comanici et al., 2025) to filter out harmful events from the set before testing to ensure that only appropriate and reliable data is used.
- *Subjective events*: Events that rely on individual opinions or subjective judgment are difficult to predict reliably. These events introduce significant variability in responses, which can disrupt the testing process. To address this, we use LLM-as-a-judge to filter out such events with a combination of Seed1.5-Thinking Seed et al. (2025), DeepSeek-R1 Guo et al. (2025), and Gemini-2.5-flash (Comanici et al., 2025) to ensure reliability. As shown in Figure 9, between July 15th and July 22nd, we drop 294 unsuitable events (both harmful and subjective events). Examples include: “Will we win 100k tomorrow at the mara hackathon?” and “I finish Park’s “Our Nation’s Path” by EOM July?”.
- *Yes-or-No events*: Events with binary choices (such as yes or no, or the outcome of a single match) are relatively easy to predict, with even random guessing achieving an accuracy of 50%. Since we cannot introduce additional choices for these events, we significantly downsample these yes/no binary events. As shown in Figure 9, events collected from prediction market websites are reduced from 193 to 36 for one week.

Through event filtering, we significantly reduce the number of events from prediction market websites to make our benchmark more challenging. This ensures the high quality of FutureX and stands in sharp contrast to previous benchmarks (Together.ai, 2025; Karger et al., 2025), where prediction market events dominate.

Additionally, to maintain event diversity and prevent homogeneity across other types of websites, we randomly select only one question per template per website for inclusion in the daily prediction set, resulting in 61 daily events and 71 weekly events ($61 \times 7 + 71 = 498$ open-ended events every week).

D DETAILS ABOUT FUTUREX DATA

Comprehensive Domain Coverage As introduced in Section 3.2, FutureX achieves comprehensive domain coverage through the daily curation of future events from 195 high-quality websites. Between July 20th and August 3rd, our dataset comprises 1,272 events, systematically categorized into 11 main domains—including politics, sports, crypto, culture, finance, business, technology

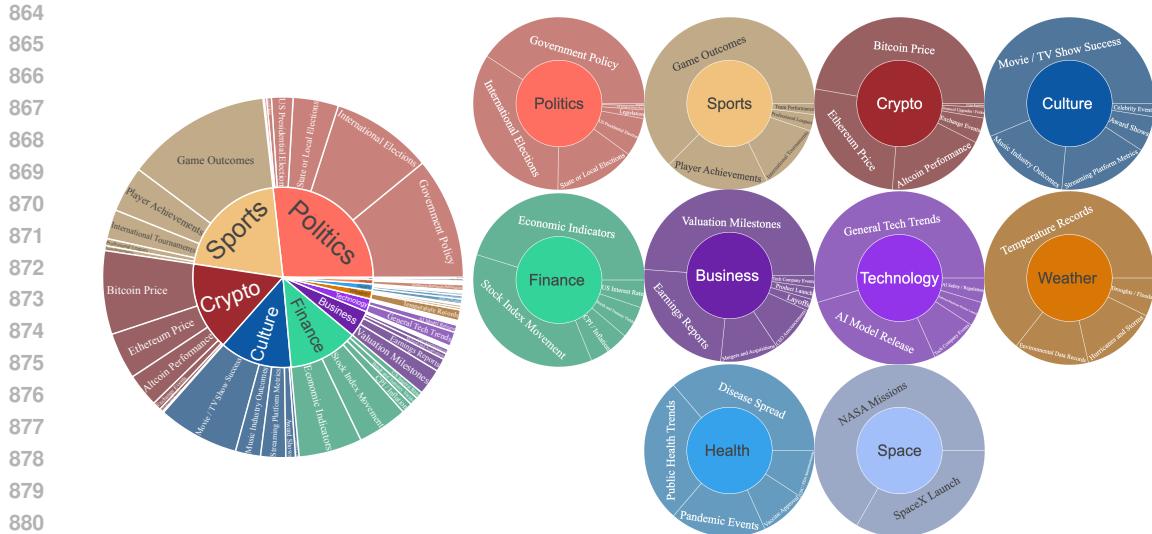


Figure 10: Detailed domains of all *events* included in FutureX, from July 20th to August 3rd, total 1,272 events. These events are organized into 11 main categories—politics, sports, crypto, culture, finance, business, technology, weather, health, and space—with each category containing several sub-categories.

trends, weather, health, and space—each further refined into several sub-categories. As depicted in Figure 10, the distribution across these domains is notably well-balanced, which facilitates a robust and comprehensive evaluation of LLM agents across a multitude of real-world scenarios. This comprehensive domain coverage provides two key advantages:

1. It enables a holistic assessment of LLM agents’ overall future prediction capabilities, as diverse fields often necessitate distinct analytical approaches and specialized reasoning strategies.
2. Coupled with FutureX’s live updating, this breadth makes our benchmark significantly more robust against exploitation or overfitting, and we anticipate this will ensure it remains a challenging and relevant evaluation for the foreseeable future.

We show the examples in major domains in Table 4.

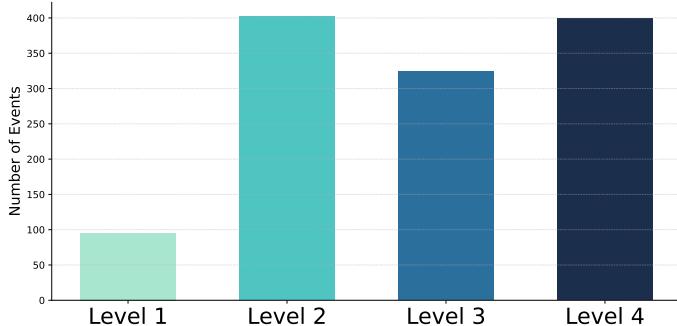


Figure 11: Difficulty level distribution (between July 20th and August 3rd).

918 Table 4: Examples to be Predicted by Domain. We take the date August 20, 2025 as an example,
 919 which can be replaced with any time in the future.

921 Domain	922 Event to be Predicted
923 Culture & Media	Please predict the Top 10 Gross in US dollars on Box Office Mojo's 924 Daily Box Office for August 20, 2025, Beijing Time.
925 Finance & Economy	Please predict the grain price index within the 'Agricultural Product 926 Wholesale Price 200 Index' from the National Agricultural Product 927 Wholesale Market Price Information System for August 20, 2025, 928 Beijing Time.
929 Technology	Please predict the market share percentage of Win10 in the operating 930 system rankings published by 51.LA for August 20, 2025, Beijing 931 Time.
932 Crypto	Please predict what the Fear and Greed Index on CoinMarketCap 933 will be on August 20, 2025, Beijing Time.
934 Business & Companies	Please predict which models will be in the top five of Dongchedi's 935 national popular sedan rankings for August 20, 2025, Beijing Time.

937 E DETAILS ABOUT EVALUATION METRIC

940 As introduced in Section 3.3, we have multiple types of events in the benchmark, each with a differ-
 941 ent evaluation metric. As for single-choice events, the metric is simply the 0-1 error:

$$942 \quad \text{score}(Y, \hat{Y}) = \mathbb{I}(Y = \hat{Y}).$$

944 For multi-choice events, as the answer contains multiple right options (denoted by \mathcal{Y}), the metric is
 945 as follows:

$$946 \quad \text{score}(\mathcal{Y}, \hat{\mathcal{Y}}) = \text{F1-Score}(\mathcal{Y}, \hat{\mathcal{Y}}).$$

947 For open-ended ranking events, such as predicting the top- k ranked items, we treat the ground truth
 948 as an ordered list, denoted by $\{y_1, \dots, y_k\}$. To evaluate predictions $\{\hat{y}_1, \dots, \hat{y}_k\}$, we design the
 949 following metric:

$$951 \quad \text{score}(\{y_1, \dots, y_k\}, \{\hat{y}_1, \dots, \hat{y}_k\}) = \begin{cases} 1, & \text{if } y_i = \hat{y}_i, \text{ for } i = 1, \dots, k \\ 0.8 \times \frac{|\{y_1, \dots, y_k\} \cap \{\hat{y}_1, \dots, \hat{y}_k\}|}{k}, & \text{otherwise,} \end{cases} \quad (1)$$

954 where partial credit (80%) is awarded based on the overlap between the predicted and ground-
 955 truth sets. For open-ended numerical prediction events, where precisely forecasting the outcome is
 956 particularly challenging, we evaluate prediction accuracy relative to the outcome's recent volatility.
 957 Specifically, we define the score as:

$$958 \quad \text{score}(Y, \hat{Y}) = \max \left(0, 1 - \left(\frac{Y - \hat{Y}}{\sigma(Y)} \right)^2 \right), \quad (2)$$

962 where $\sigma(Y)$ denotes the standard deviation of the outcome over the past 7 days. Intuitively, predic-
 963 tions that fall within one standard deviation of the true outcome receive partial credit, while those
 964 beyond one standard deviation receive a score of zero.

965 F DEALING WITH MISSING PREDICTIONS

968 One challenge is handling missing predictions. Since the pipeline runs on a daily basis, if a model
 969 fails to produce a prediction on a given day, it cannot retroactively provide it later. Given the large
 970 number of models evaluated, as well as occasional API instability or cases where a model may
 971 refuse to predict certain future events, missing predictions are inevitable. The ideal approach to
 handling missing predictions is to align the events across all models. However, because the events

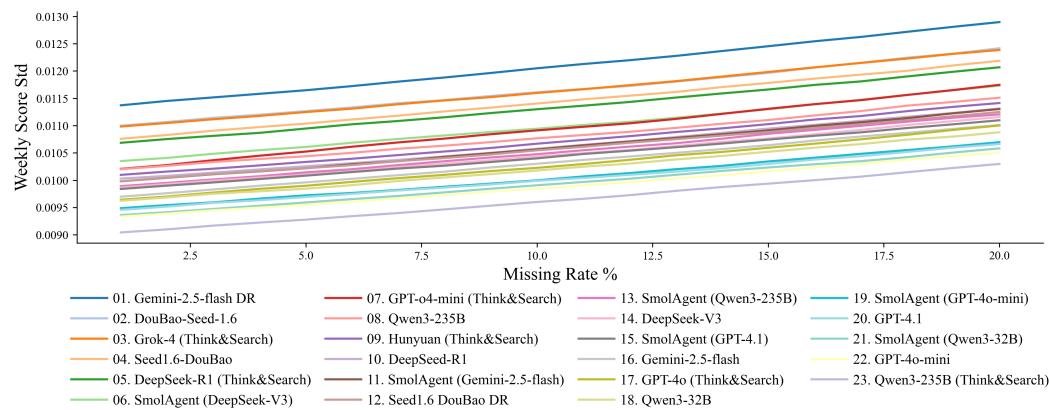


Figure 12: Standard deviation vs. missing rate. The missing rate κ ranges from 1% to 20%; for each model, we plot the standard deviation of its average score.

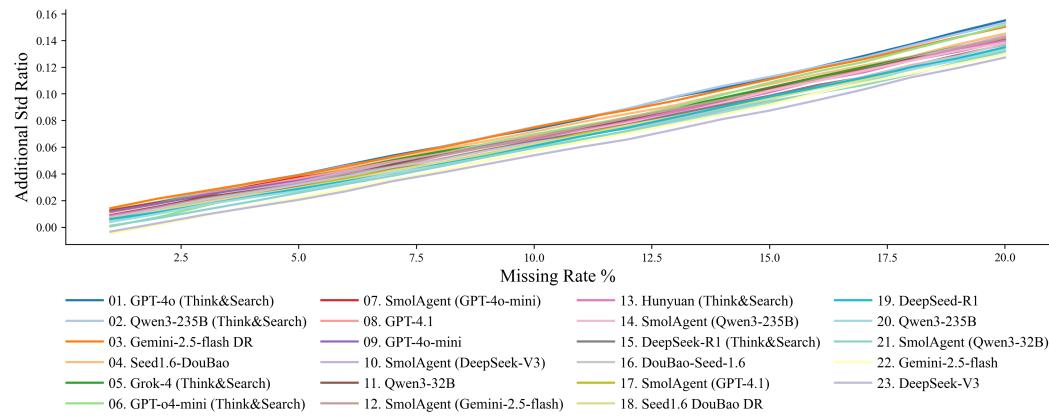


Figure 13: Additional standard deviation ratio vs. missing rate.

with missing predictions typically differ between models, full alignment would substantially reduce the total number of events, which is undesired.

Based on our data, we first analyze the standard deviation (std) introduced by missing predictions using a Monte Carlo simulation. Since we have roughly 500 events each week, in each simulation trial we first sample 500 events (with both results and predictions) and calculate the “true” average score s_i . We then randomly sample $(100 - \kappa)\%$ of these events, where $\kappa\%$ denotes the missing ratio, varying from 1% to 20%, and compute the “pseudo” average score \hat{s}_i . This process is repeated 20,000 times, after which we calculate the “true” standard deviation, $\text{Std}(S)$, for $\{s_1, \dots, s_{20000}\}$, and the “pseudo” standard deviation, $\text{Std}(\hat{S})$, for $\{\hat{s}_1, \dots, \hat{s}_{20000}\}$. The “pseudo” standard deviation is plotted in Figure 12 against the missing rate. In addition, we quantify the relative increase in standard deviation $((\text{Std}(\hat{S}) - \text{Std}(S)) / \text{Std}(S))$ with respect to missing predictions in Figure 13.

The results indicate that the standard deviation remains relatively small. These values are computed from 500 total samples, which roughly corresponds to one week of data. As FutureX progresses and the test size grows, the standard deviation is expected to decrease at a rate proportional to the inverse square root of the sample size. Therefore, given the complexity of the auto-evaluation pipeline and the relatively minor impact of missing predictions, we prioritize increasing the test sample size over strict alignment, allowing for slight misalignments across different models.

We plot the additional standard deviation ratio introduced by missing predictions in Figure 13. Note that the ratio is calculated by $\frac{\text{Std}(\hat{S}) - \text{Std}(S)}{\text{Std}(S)}$.

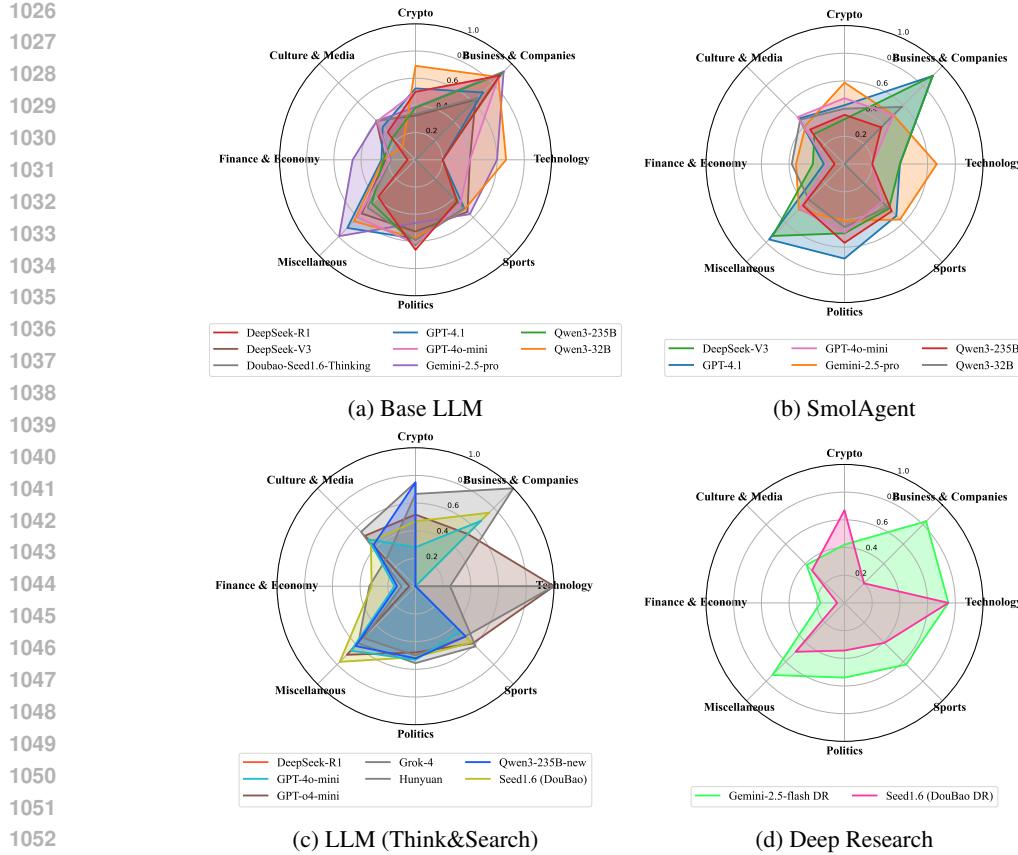


Figure 14: Performance across different domains for Level 1 (Basic Tier) and Level 2 (Wide Search Tier) events.

G MORE EXPERIMENTS AND ANALYSIS

Due to space limit of the main body, we put more experimental results and analyses here.

G.1 RESULTS ACROSS DIFFERENT DOMAINS

In addition to the overall performance, we present domain-specific results in Figure 14 and Figure 15, which highlight the relative strengths of different models across various subject areas. Given the performance gap between Level 1&2 and Level 3&4, we draw two set of figures respectively. Several interesting observations include:

- **Different models have different strengths.** GPT models—including GPT-4.1 (Base LLM), GPT-4.1 (SmolAgent), and GPT-o4-mini (Think&Search)—demonstrate superior performance in *Crypto* and *Technology*. DouBao-Seed1.6-Thinking excels in *Finance&Economy* and *Business&Companies*, while DeepSeek-V3 (SmolAgent) performs exceptionally well in *Politics*, even outperforming closed-source deep-research agents and Think&Search LLMs.
- **Search-enhanced reasoning significantly improves performance in information-driven domains.** For domains like *Culture & Media* and *Technology*, performance increases notably as we move from basic to more advanced reasoning frameworks. This is likely because these domains benefit directly from timely information access and contextual reasoning grounded in real-world updates.
- **Tool using increases the performance differences.** As shown in Figure 14, for Level 1 and Level 2 events, the performance differences among the base LLMs are not very large,

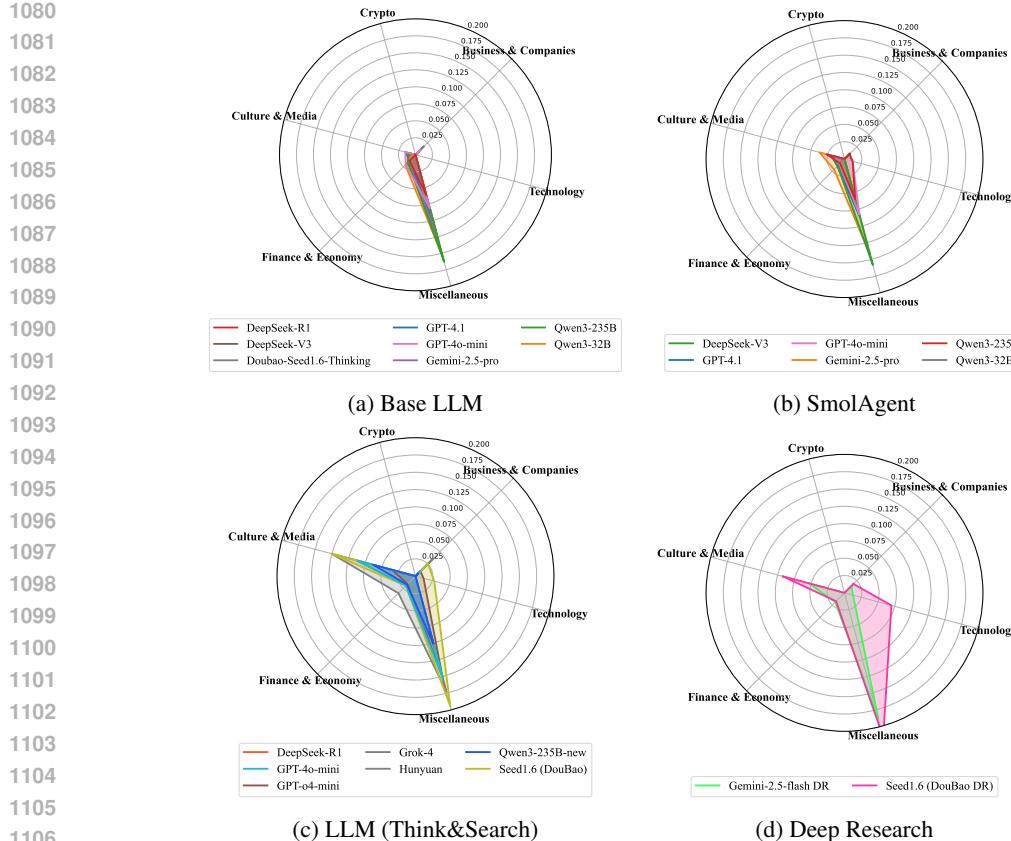


Figure 15: Performance across different domains for Level 3 (Deep Search Tier) and Level 4 (Super Agent Tier) events.

but as tools are added, the gaps in the radar charts widen, likely reflecting each model’s choice of search tools and reasoning style.

- **Advanced searching alone may not be sufficient in complex, abstract domains.** In contrast, for Level 3 and Level 4 events (see Figure 15), even models equipped with strong search capabilities struggle to perform well. These open-ended tasks often demand deeper abstraction, multi-hop reasoning, and a synthesis of heterogeneous information, revealing a potential limitation in the current design of search-augmented agents.

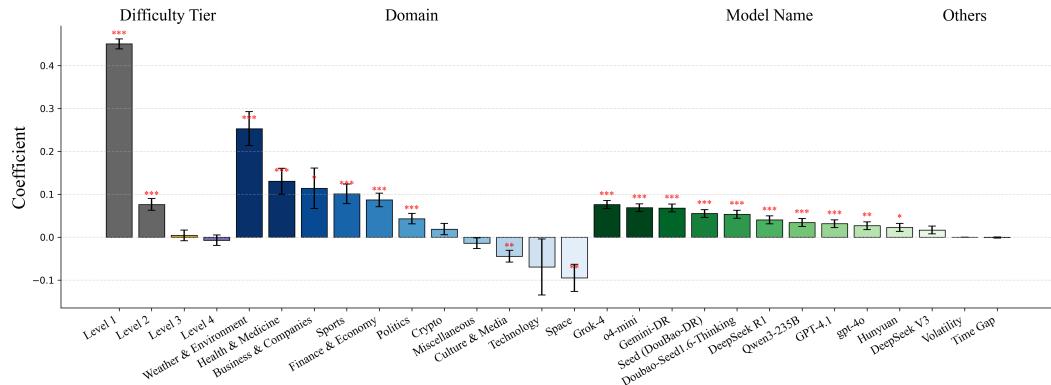
In addition to the overall results, we conduct several in-depth case studies to further understand the current models’ limitations.

G.2 FACTOR ANALYSIS

To systematically investigate the impact of each factor, such as the choice of LLM, event domain, and difficulty tier, we perform a linear regression analysis on each model’s score for each event. Note that the target variable is the score of each model on each event. Figure 16 shows the estimated coefficients for each factor, with *** indicating statistical significance ($p < 0.005$). From the results, we have the following observations:

- **Difficulty level really matters.** Consistent with our earlier findings, difficulty level has a significant impact on model performance. This also validates our overall scoring scheme, in which we assign 10% and 20% weights to Level 1 and Level 2 events, respectively, to place greater emphasis on more challenging cases.

- **Domain also matters.** We observe substantial variation in the coefficients of different domains. This highlights the importance of domain-specific challenges and the need for tailored evaluation.
- **Top models align with the overall leaderboard.** The four highest-performing models (Grok-4, GPT-04-mini, Gemini Deep Research, Seed1.6 (DouBao)) in our per-domain analyses are exactly the same as those in the overall score ranking, confirming the consistency and robustness of our benchmark.

Figure 16: Coefficients of different factors in our linear analysis. The R^2 is 0.418.

G.3 PAST PREDICTION VS. FUTURE PREDICTION

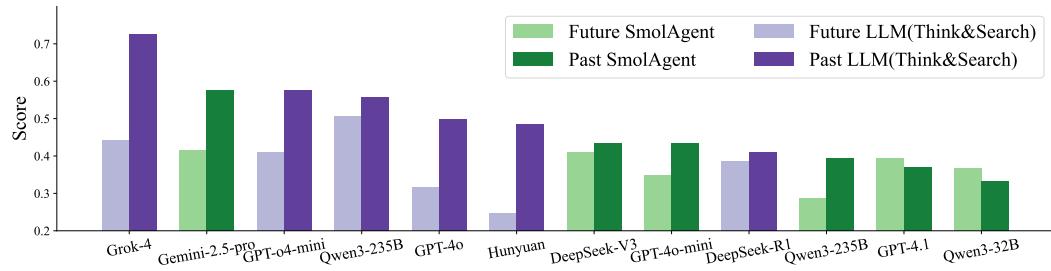


Figure 17: Comparing Past and Future Predictions. We randomly select 30 events from Level 1 and Level 2, then evaluate model performance on two tasks: predicting outcomes before they are known (future prediction) and searching outcomes after they have been resolved (past prediction).

To more precisely assess search capability, we introduce a past-prediction task in which models retrieve each event’s outcome *one week* after its resolution date. Using the *same* set of 30 randomly selected events from Level 1 and Level 2⁶, we report performance scores for both past-prediction and future-prediction in Figure 17. In the figure, dark bars represent past-prediction results, while shallow bars represent future-prediction results. From the results, we find that:

- **Grok-4 leads in search capability, followed by GPT models and Hunyuan.** In the past-prediction task, Grok-4 significantly outperforms all other methods, underscoring its robust and timely information retrieval. GPT-04-mini and GPT-4o also deliver strong past-prediction performance. Interestingly, although Hunyuan achieves impressive past-prediction results, the large gap between its future- and past-prediction scores suggests weaker reasoning ability—meaning that despite effective search, its overall performance suffers.

⁶We focus on Level 1 and Level 2 events because their outcomes are more readily retrievable.

1188

- 1189 • **SmolAgent with Gemini-2.5-pro achieves significant gains, unlike with other base**
 1190 **LLMs.** Within the open-source SmolAgent framework, integrating Gemini-2.5-pro
 1191 yields a marked improvement in past-prediction performance—comparable to most com-
 1192 mercial models (except Grok-4). Crucially, this demonstrates that SmolAgent’s relatively
 1193 modest overall performance cannot be blamed solely on the quality of its search API. By
 1194 contrast, when paired with other base LLMs, SmolAgent shows little to no improvement—
 1195 and in some cases, even a performance decline—suggesting that the underlying search
 1196 capabilities of those models play a significant role in the agent’s effectiveness.

1197 **G.4 PLANNING ANALYSIS OF SMOLAGENT**

1198 To understand the agent’s performance, we examine the agent memory of SmolAgent when paired
 1199 with different LLM backends. As we cannot access the internal memory of closed-source models,
 1200 our analysis focuses on SmolAgent.

1201 SmolAgent’s memory contains a plan to solve the problem, a detailed log of tool usage, and the
 1202 outcome of each tool call in every iteration, providing a rich foundation for our analysis. Examples
 1203 of full memory are shown in Section I. For each study plan generated by an agent, we first conduct
 1204 an evaluation using Gemini-2.5-pro, which assigns a score ranging from 0 to 10 across three key
 1205 dimensions: (1) comprehensiveness (assessing the extent to which the plan covers all necessary
 1206 components and relevant information), (2) source reliability (evaluating the credibility and validity
 1207 of references or data cited in the plan), and (3) plan actionability (measuring how practical and
 1208 executable the proposed steps are in real-world scenarios).

1209 To ensure the fairness and objectivity of the evaluation process, we anonymize the model identity
 1210 by replacing the original model name in the prompt with a generic label “model- i ” (where i is a
 1211 unique numerical identifier). This anonymity mechanism is designed to prevent Gemini-2.5-pro
 1212 from exhibiting potential bias, specifically, avoiding any tendency to inflate scores for study plans
 1213 generated by models within the Gemini series. By eliminating such identity-based influences, we
 1214 aim to obtain more accurate and unbiased assessment results that truly reflect the quality of each
 1215 study plan.

1216 As shown in Table 5, **GPT-4.1** and **Gemini-2.5-pro** obtained significantly higher planning scores
 1217 compared to other models. This result is consistent with their superior overall performance demon-
 1218 strated in Figure 1, which suggests a strong relationship between an agent’s planning capability and
 1219 its future-prediction performance.

1220 A closer look at Table 5 reveals key differences in how models approach planning:

1221

- 1222 • **Comprehensiveness:** Powerful models like GPT-4.1 earn high scores for comprehensiv-
 1223 ness by generating plans that address a wide array of specific and sophisticated risk factors.
 1224 In contrast, weaker models like Qwen3-32B produce plans that are often superficial, cov-
 1225 ering only basic elements and lacking the necessary depth or specific guidance.
- 1226 • **Source Reliability:** We observed that strong models like GPT-4.1 consistently leverage
 1227 authoritative and specialized sources for information. Conversely, weaker models such as
 1228 Qwen3-32B frequently pull information from unvetted sources like Twitter, compromis-
 1229 ing the reliability of their plans.
- 1230 • **Plan Actionability:** Interestingly, certain models, including Deepseek-v3, sometimes reach
 1231 a conclusion within the planning phase itself. This behavior suggests that these models
 1232 may not always rely on search tools to solve problems, instead leveraging their internal
 1233 knowledge base to form a final plan.

1242 Table 5: Analysis of agent planning by scoring the memory in Comprehensiveness, Source Reliability
 1243 and Plan Actionability. The predicted event in the shown example is “What price will Ethereum
 1244 hit July 21-27?”

1245

1246 Model	1247 Criterion	1248 Example	1249 Score	1250 Analysis
1247 GPT-4.1	1248 Comprehensiveness	1249 Upcoming catalysts or risk factors affecting ETH from now until July 2025 (e.g., expected upgrades, ETF approvals/rejections, major regulations, known hack/theft risks, etc.).	1250 9	1251 Covers a wide range of specific and sophisticated risk factors.
		1252 Implied expectations from Ethereum derivatives markets... Source: Deribit, CME, other derivatives market data/analysis.	1253 8	1254 Identifies authoritative, specialized sources for advanced metrics.
		1255 5. Search for information from the derivatives markets (particularly options and futures pricing for ETH with expiry around July 2025) to infer market-implied price expectations and volatility.	1256 9	1257 The step is a clear, specific, and executable instruction.
1258 Gemini-2.5-pro	1259 Comprehensiveness	1260 Ethereum’s Technical Roadmap for 2024–2025: Major upgrades like the “Pectra” fork can act as significant price catalysts.	1261 10	1262 Highly specific and knowledgeable, referencing a key future network upgrade by name.
		1263 Source: Use the search_agent to find reports and articles from sources like Bloomberg, CoinDesk, Messari, Goldman Sachs, JPMorgan, etc.	1264 10	1265 Unmatched in its list of specific, top-tier financial and crypto-native sources.
		1266 6. Analyze all the gathered information (price history, expert forecasts, technical catalysts, macroeconomics, and regulation) to synthesize a coherent outlook.	1267 10	1268 Clearly defines a complex but actionable goal of synthesis.
1269 Qwen3-32b	1270 Comprehensiveness	1271 Any relevant news, events, or macroeconomic indicators that may influence Ethereum’s price during this timeframe.	1272 5	1273 A generic statement that covers the basics but lacks depth or specific direction.
		1274 Source Reliability	1275 News outlets like Reuters, Bloomberg, Coindesk, or crypto-focused forums like Reddit or Twitter/X.	1276 2
		1277 Plan Actionability	1278 Use the search_agent team member to research any upcoming events, news, or macroeconomic factors that could affect Ethereum’s price...	1279 4

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

(continued on next page)

1296 (Table 5 continued)

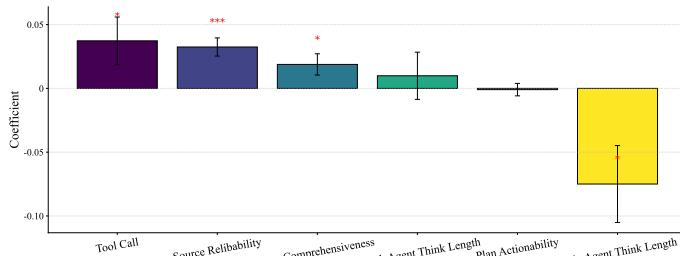
Model	Criterion	Example	Score	Analysis
Deepseek-v3	Comprehensiveness	Facts to derive - Correlation between Bitcoin halving cycles (April 2024) and Ethereum's price 15 months later.	9	Demonstrates a deep, specific, and relevant understanding of crypto market cycles.
	Source Reliability	Sources: Crypto market data platforms (CoinGecko, CoinMarketCap), Ethereum Foundation announcements, financial news (Cointelegraph, Decrypt), and analyst reports (e.g., Ark Invest, Glassnode).	8	Provides a strong list of specific and respected sources across different categories.
	Plan Actionability	Finalize prediction: Select the most plausible options... and format the answer as $\boxed{A, B, \dots}$.	0	The plan's action is to provide a conclusion, which it does immediately, negating the purpose of the plan itself.
GPT-4o-mini	Comprehensiveness	Market conditions or significant events that may affect Ethereum's price around that timeframe (e.g., regulatory changes, technological upgrades, macroeconomic factors).	5	Lists standard categories but remains on a generic, surface-level.
	Source Reliability	This information can be found in articles or publications on cryptocurrency news websites or financial analysis reports.	2	Fails to name any specific sources, making the plan's quality entirely dependent on chance.
	Plan Actionability	4. Review and compile the significant factors that could affect Ethereum's price between now and July 2025, including potential regulatory developments or technological advancements.	4	A vague instruction to "review and compile" without guidance on how to weigh or analyze these factors.
Qwen3-235b	Comprehensiveness	We can calculate potential price ranges using technical analysis tools like moving averages, Fibonacci retracements, etc.	7	Decent scope, and improves its quality by mentioning specific types of analysis tools.
	Source Reliability	Cryptocurrency market forecasts for 2025: To understand expert opinions and analyses regarding the future of Ethereum.	1	A critical failure. It identifies the need for expert analysis but provides zero indication of where to find it.
	Plan Actionability	8. Calculate probabilities for each option based on the analysis and select the most plausible options.	7	A clear, specific, and valuable step that adds a quantitative layer to the plan.

1350 Based on this, we then perform a linear regression analysis to assess the impact of several factors,
 1351 including total tool calls, search text length, overall context length, and the three evaluation scores.
 1352 We restrict our study to Level 1 and Level 2 events due to SmolAgents’ poor performance on Levels
 1353 3 and 4. As shown in Figure 18, we find that:

1354

- 1355 • **Number of tool calls, source reliability, and plan comprehensiveness** exert the strongest
positive effects on the overall score: more frequent tool calling, higher trustworthiness of
 1356 referenced information, and more thorough answer content all drive substantially higher
 1357 user ratings.
- 1358 • In contrast, **main agent think length** carries the most *negative* effect: longer accumulated
 1359 dialogue history introduces noise and redundancy, which hurts the performance.

1360 These results suggest that, for further improvements, SmolAgents could strategically invoke tools,
 1361 rigorously check and cite reliable information, and maintain concise dialogue histories to improve
 1362 the performance.

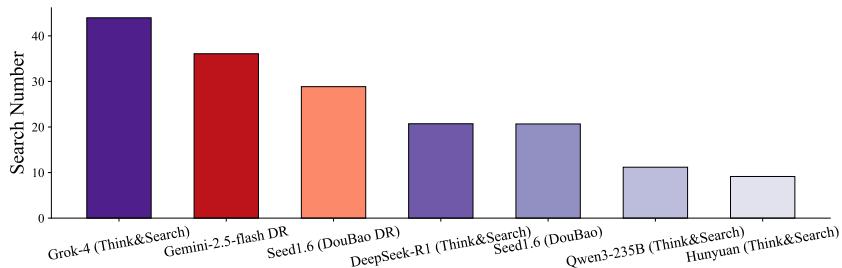


1363 Figure 18: Coefficients of different factors in the linear analysis of SmolAgent’s planning. The R^2
 1364 is 0.518.

1365 G.5 SEARCH ANALYSIS

1366 Figure 19 shows the average number of web queries performed by commercial LLMs with
 1367 Think&Search. Notably, Grok-4 issues the largest number of searches among all evaluated
 1368 models.⁷ This finding has two key implications: first, Grok-4’s high query volume corresponds with its
 1369 superior performance on the FutureX; second, it achieves this result with remarkably low latency—
 1370 completing searches in *less than 5 minutes*, compared to approximately 30 minutes for Gemini deep
 1371 research.

1372 Furthermore, two deep research models conduct substantially more searches than the rest of the
 1373 Think&Search cohort, highlighting their aggressive retrieval strategies. By contrast, Qwen3-235B
 1374 and Hunyuan perform significantly fewer web queries, which may help explain their relatively lower
 1375 performance on the same tasks.



1376 Figure 19: The search number of different models.

1377
 1378
 1379
 1380
 1381
 1382
 1383
 1384
 1385
 1386
 1387
 1388
 1389
 1390
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399
 1400
 1401
 1402
 1403
⁷We are unable to measure the search counts for GPT-o4-mini and GPT-4o (Think&Search), so these models
 are excluded from the comparison.

1404 **H EXAMPLES OF AGENT RESPONSE**
1405
1406
1407

This study analyzes the reasoning processes behind agent's response. To understand why their predictions succeed or fail, we selected several lengthy agent responses as our subject of analysis. We utilized the Gemini-2.5-flash model to analyze the reason that the prediction is correct or incorrect.

We found that the correct predictions (Gemini-2.5-flash-Deep-Research & Grok-4) are based on comprehensive and credible evidence, while the incorrect predictions come from insufficient data (GPT-4o), unreasonable assumption (GPT-o4-mini) or hallucination (Doubao-Seed-1.6-Deep-Research).

Question

How many deaths will there be to claim during the riots in California until the end of July 2025?

- A. the outcome be 0
- B. the outcome be 1 - 5
- C. the outcome be 6 - 10
- D. the outcome be 11 - 15
- E. the outcome be 16 - 20
- F. the outcome be 21 - 25
- G. the outcome be Above 2

The answer is A.

Gemini-2.5-flash-Deep-Research**ANALYSIS**

The prediction is correct due to exhaustive evidence showing no riot-related fatalities, proper distinction of unrelated deaths, and alignment with de-escalation trends.

PREDICTION (SUMMARIZED BY GEMINI-2.5-FLASH)

This report provides an analytical assessment of fatalities directly attributable to the civil unrest and riots in California from June 6 to July 31, 2025. Based on a comprehensive review of the available intelligence and public records for this period, there have been no reported deaths directly resulting from the civil unrest.

Key evidence supporting this conclusion includes:

1. Detailed accounts of the June 2025 Los Angeles riots consistently omit any mention of fatalities.
2. Subsequent protests in July 2025 across the state were predominantly non-lethal in nature.
3. The tragic deaths of three Los Angeles County Sheriff's deputies in an explosion were explicitly identified as an isolated, non-riot-related incident, as they occurred during the handling of explosive devices as an occupational hazard.
4. This event's lack of fatalities stands in stark contrast to historical precedents like the 1965 Watts Riots (34 deaths) and the 1992 Los Angeles Riots (50-63 deaths), highlighting a significant difference in outcomes.

Based on the exhaustive review of the provided data, the estimated number of deaths to claim during the riots in California until the end of July 2025 is 0. Therefore, the correct option is A: the outcome be 0.

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

Grok-4

ANALYSIS

The prediction is correct because it is supported by credible sources, accounts for the contained nature of the protests, and accurately reflects the absence of riot-related fatalities up to the prediction date.

PREDICTION

There have been ongoing protests and occasional riots in California, primarily in Los Angeles, since June 6, 2025, in response to ICE immigration raids. These events have involved clashes with law enforcement, vandalism, arson (e.g., vehicles set on fire), and the use of less-lethal munitions like tear gas, flash-bang grenades, and rubber bullets. A state of emergency remains in effect, with federal forces (including National Guard and Marines) deployed but partially withdrawn by mid-July. Protests continued into late July, including a demonstration on July 23, but have been largely contained to small areas of downtown LA and have not escalated to widespread destruction.

Multiple sources, including Wikipedia, BBC, LA Times, AP News, and NPR, explicitly state no deaths have been directly linked to the protests or riots as of July 25, 2025. Injuries total around 20–40 (affecting protesters, officers, journalists, and others), with over 575 arrests. Two potential incidents were investigated:

- A farmworker died on July 12 after falling during an ICE raid on July 10 (accidental, while fleeing; occurred amid a protest at the site but not from riot violence).
- A body was found near a looting site on June 10, but showed no trauma and the cause remains unknown/unlinked to the unrest.

Neither is attributed as a riot-related fatality in official reports or media summaries. Given the protests' contained nature, reduced troop presence, and lack of fatalities over 1.5 months, it is unlikely significant escalation or deaths will occur in the remaining 6 days through July 31. The total is predicted to be 0.

A

GPT-4o

ANALYSIS

The prediction error arises primarily due to **insufficient information collection and flawed logic**. The prediction cites two deaths but concludes with a range (1–5) that is not well-justified by the provided evidence. Additionally, it fails to account for potential underreporting or delayed confirmation of fatalities, which are common in chaotic events like riots. The prediction also lacks clarity on whether the two deaths are directly caused by the riots or merely coincidental.

Summary: Insufficient data and inconsistent logic in extrapolating the death count range.

PREDICTION

Based on available information up through late July 2025, here is the current situation regarding fatalities connected to the California riots (primarily in Los Angeles):

- According to multiple news reports, **no deaths** had been officially confirmed during the early days of the protests/riots in June — in contrast to the 1992 Rodney King riots, which saw over 60 fatalities (AP News).
- However, media outlets have reported at least **one body was discovered** near a looted area in downtown Los Angeles during the unrest (New York Post).
- Additionally, a **California farmworker named Jaime Alanis died** from wounds sustained during a protest in early to mid-July 2025 (en.wikipedia.org).

As of **July 25, 2025**, that brings the **known death count to two** attributed to events tied to the unrest. While reporting conditions remain complex—and it's possible further deaths could be confirmed later—based on the **best publicly available data** now, the most plausible fatality total in California riots through **July 31, 2025** is in the **1–5 range**.

My prediction: B

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

GPT-o4-mini

ANALYSIS

The prediction is incorrect because it fails to account for the possibility of unforeseen escalations or new incidents that could occur between the time of the prediction (July 28, 2025) and the end of July 2025. The prediction relies on limited data (only one confirmed death) and assumes the situation will remain stable, which is an unreasonable assumption given the volatile nature of riots and protests.

Summary: Unreasonable assumption of stability and insufficient consideration of potential escalations

PREDICTION

Based on official reports and reputable news sources:

- As of June 9, 2025, there were no reported fatalities in the Los Angeles protests and riots (goggrids.com).
- On July 11–12, 2025, Jaime Alanis—a 57-year-old farmworker—died from injuries sustained during a chaotic ICE raid protest, marking the first confirmed riot-related death (people.com, theguardian.com).
- No other deaths directly attributable to these disturbances have been reported through late July 2025.

Given the low fatality rate so far and the nature of these localized events, it is highly likely the total number of riot-related deaths by July 31, 2025 will remain in the 1–5 range.

B

Doubao-Seed-1.6-Deep-Research

ANALYSIS

The model's prediction was completely wrong because its reasoning was based on an entirely fabricated scenario, not a real-world trend analysis. The model invented a narrative that riots in California began in June 2025 due to a massive ICE raid, and it established a false premise that approximately 13 deaths had already occurred by mid-July. Based on this erroneous, self-created “fact,” the model naturally ruled out the possibility of zero deaths and concluded that the final toll would fall within the 11-15 range. The entire analytical process is a classic case of “garbage in, garbage out,” as it was analyzing an event that never happened.

Sumary: The prediction failed not because of flawed analytical logic, but because its analysis was based on a complete “AI hallucination.” The model invented a crisis and its casualties, and then predicted an outcome based on that fiction. In reality, the described events did not occur, making the correct death toll zero.

PREDICTION

Current Situation and Historical Context The 2025 California riots began in early June, triggered by large-scale raids by U.S. Immigration and Customs Enforcement (ICE) in Los Angeles's Latino-majority communities. On June 6, ICE arrested 44 undocumented immigrants at seven locations, including shopping malls and factories. This action quickly sparked intense community protests, which evolved into ongoing social unrest. By June 11, the incidents had spread to major California cities like San Francisco and San Diego, with over 5,000 arrests statewide, 12 deaths, and direct economic losses exceeding \$800 million.

The situation escalated further in July. On July 3, the Trump administration deployed an additional 700 Marine Corps soldiers to California, reinforcing federal control over the state. These troops joined the 4,000 National Guard soldiers already on the ground to maintain order in cities like Los Angeles. However, California Governor Gavin Newsom deemed the move unconstitutional, further intensifying the conflict between the two sides.

1566

1567

1568

1569

1570

1571

1572

1573

1574

On July 10, federal agents clashed with protesters during an immigration raid on a Southern California farm. On-site video footage showed helmeted law enforcement officers throwing tear gas, forcing protesters and workers to retreat as thick smoke engulfed the area, affecting even children. This operation resulted in one death and multiple injuries. On July 18, an explosion at the Biscaluz Training Center in Los Angeles County killed three veteran police officers instantly. The incident was caused by an unexploded device discovered the day before in a Santa Monica apartment garage, which detonated accidentally during handling, brutally exposing a fatal flaw in the law enforcement's high-risk materials handling protocol.

Compared to historical California riots, the 2025 unrest shows new characteristics. The 1992 Los Angeles riots were triggered by the acquittal of white police officers who beat black motorist Rodney King. The riots lasted six days, caused 63 deaths, and resulted in \$1 billion in damages. In contrast, the 2025 riots are centered on immigration policy and have evolved into a power struggle between the federal and state governments. It is noteworthy that although the scale of the 2025 riots is considerable, the current death toll (around 13) remains far below the 63 fatalities of 1992. This suggests that the current situation has not reached the same level of violence as historical highs, but given the ongoing conflict and escalating confrontation, the developments over the next week require close attention.

Key Data and Influencing Factors According to statistics from reliable sources, as of July 24, 2025, the California riots have resulted in approximately 13 deaths:

1. By June 11, 12 deaths had been confirmed.
2. On July 10, the Southern California farm raid resulted in 1 death.
3. On July 18, the explosion at the Los Angeles County training center caused the deaths of three police officers, but whether this incident is directly related to the riots is yet to be confirmed.

It is important to emphasize that the above statistics may have omissions, especially considering other violent conflicts or accidental events that may occur during the riots.

The power struggle between the federal and state governments is a key factor influencing the development of the riots. Since taking office in January 2025, the Trump administration has labeled undocumented immigrants a "national security threat", setting a goal of 3,000 daily arrests and attempting to seal the U.S.-Mexico border. This policy directly impacts California's economy and social structure, as the state's economy is highly dependent on immigrant labor, particularly in the agricultural and construction sectors. California Governor Newsom strongly opposes the federal government's intervention, publicly accusing Trump's actions of being unconstitutional and describing it as a "constitutional war". He believes the White House's goal is to weaken state power and set a precedent for federal intervention in state affairs. This political antagonism not only exacerbates social divisions but also makes it difficult to resolve the conflict through conventional means.

The federal government has deployed a large number of military forces to California. In addition to the 700 Marine Corps soldiers, this includes the 40th Infantry Division of the National Guard and rapid response forces from the Department of Homeland Security. These troops, equipped with riot gear, light armored vehicles, and armed helicopters, have established temporary posts at several strategic locations in Los Angeles. The escalation of law enforcement tactics is also a significant factor in the increase of casualties. On-site videos show law enforcement officers using tear gas, rubber bullets, and flashbang grenades to suppress crowds. This high-intensity enforcement not only directly causes casualties but also easily escalates tensions, leading to further conflict.

Prediction for the Next Week Key factors influencing the death toll include:

1. **Progress of negotiations between the federal and state governments:** If both sides can reach a compromise, it may ease tensions and reduce violent clashes.
2. **Scale and intensity of protest activities:** During the Independence Day period, protests of varying scales were held in many parts of the United States, from Chicago to Los Angeles, from Roanoke, Virginia, to Sioux Falls, South Dakota, and Eugene, Oregon. Such large-scale protests carry a high risk of violence.
3. **Changes in enforcement intensity:** If law enforcement continues to use high-intensity suppression tactics, it may lead to more casualties.
4. **Economic factors:** As the world's fourth-largest economy, California's economic interests may prompt all parties to seek a peaceful resolution.

Expert analysis points out that there are fundamental differences between these riots and the 1992 Los Angeles riots. The 1992 riots stemmed primarily from racial discrimination and judicial injustice,

1620

1621

1622

1623

1624

1625

1626

1627

1628

1629

while the 2025 conflict is more the result of political maneuvering. In the 1992 riots, the jury’s acquittal of the four police officers involved sparked a six-day riot, causing 53 deaths, over two thousand injuries, and property damage estimated at over \$1 billion. Currently, the death toll from the 2025 riots (around 13) is significantly lower than the 1992 level. This is partly due to a faster response from law enforcement and also reflects the differences in the organization and intensity of the current protests. However, with the escalation of the confrontation between the federal and state governments, the situation could still worsen.

Considering the fatality trend over the past month and the current situation, we can make the following predictions:

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1. **Most Likely Scenario:** An increase of 0–2 deaths in the coming week, bringing the total to 13–15. This prediction is based on the current relatively stable death toll trend and the likely scale of conflicts.

2. **Optimistic Scenario:** If the federal and state governments can reach a compromise and protests remain peaceful, there may be no new deaths in the coming week, keeping the total at around 13.

3. **Pessimistic Scenario:** If the conflict escalates further, especially if large-scale violent events similar to 1992 occur, the death toll could increase significantly. However, given the current military deployment and law enforcement capabilities, the possibility of this happening is relatively low.

Final Prediction and Conclusion Taking into account historical data, the current situation, and future trends, my prediction for the total number of fatalities in the California riots by the end of July 2025 is as follows:

B. 1–5 fatalities: This option clearly underestimates the actual situation, as the death toll has already reached around 13 as of July 24.

C. 6–10 fatalities: This option is also below the actual situation, as the current death toll is already over 10.

D. 11–15 fatalities: This is the most likely scenario. Based on the current death toll of around 13 and the prediction of a possible increase of 0–2 deaths in the coming week, the total number of fatalities is most likely to fall within this range.

E. 16–20 fatalities: While this possibility cannot be completely ruled out, it would require multiple severe violent incidents to occur in the coming week to reach this level, making it relatively less probable.

F. 21–25 fatalities and G. Over 25 fatalities: These two scenarios would require a significant escalation in the scale and intensity of the conflict. Given the current military deployment and political environment, these are highly unlikely.

Final Prediction: The most probable death toll range is **D. 11–15 people**.

Risk Assessment and Uncertainties It must be emphasized that this prediction contains a degree of uncertainty, stemming primarily from the following factors:

1656

1657

1658

1659

1660

1661

1662

1663

1. **Changes in political factors:** The relationship between the federal and state governments could change significantly in a short time, affecting the situation’s development.

2. **Possibility of external intervention:** Other political forces or international factors might intervene, altering the nature and intensity of the conflict.

3. **Impact of sudden events:** Unexpected incidents, such as the July 18 explosion, could cause a sudden increase in the death toll.

4. **Unpredictability of protest activities:** The development of large-scale protests is often uncertain and can escalate or subside rapidly.

Despite these uncertainties, based on existing data and analysis, we have reason to believe that the total number of fatalities from the California riots by the end of July 2025 is most likely to fall within the 11–15 range.

1667

1668

1669

1670

1671

I EXAMPLES OF SMOLAGENT’S PLANNING

1672

1673

Smolagent Roucher et al. (2025b) is an open-source agent framework where the planning of agents is visible. The following examples are planning of SmolAgent’s agents, which all follows the same structure:

1674 1. Facts Survey
 1675 (a) Facts to look up
 1676 (b) Facts to derive
 1677 2. Plan

1679 **Question**

1682 What price will Ethereum hit July 21-27?
 1683 A. The outcome be \$4400
 1684 B. The outcome be \$4300
 1685 C. The outcome be \$4200
 1686 D. The outcome be \$4100
 1687 E. The outcome be \$4000
 1688 F. The outcome be \$3700
 1689 G. The outcome be \$3600
 1690 H. The outcome be \$3500
 1691 I. The outcome be \$3400
 1692 J. The outcome be \$3300

1694 **GPT-4.1**

1695 1. FACTS SURVEY

1696 1.1. FACTS GIVEN IN THE TASK

- 1699 • The event to be predicted concerns the price of Ethereum (ETH) during the week of July
 1700 21-27, 2025, with reference time around 2025-07-28T04:00:00Z.
- 1701 • Several options are specified, corresponding to price thresholds ETH might reach or dip to
 1702 during that time:
 - 1703 – A. \$4400
 - 1704 – B. \$4300
 - 1705 – C. \$4200
 - 1706 – D. \$4100
 - 1707 – E. \$4000
 - 1708 – F. \$3700
 - 1709 – G. \$3600
 - 1710 – H. \$3500
 - 1711 – I. \$3400
 - 1712 – J. \$3300
- 1713 • The answer must use the boxed format and list all plausible options.

1714 1.2. FACTS TO LOOK UP

- 1716 • The current (most recent) price of Ethereum (ETH) in USD.
 - 1717 – Source: major crypto price aggregator (CoinGecko, CoinMarketCap, etc.)
- 1718 • Recent trends, volatility, and price movement data for Ethereum in the months leading up to
 1719 the task date (currently June 2024).
 - 1720 – Source: historical charts from crypto price sites.
- 1721 • Macroeconomic factors, crypto market sentiment, and news that could affect ETH price up
 1722 to July 2025.
 - 1723 – Source: news sites, market outlooks, analyst reports.
- 1724 • Historical Ethereum price behavior during similar periods (i.e., how ETH behaved in prior
 1725 years between June/July and the following July).
 - 1726 – Source: price charts or historical returns data.

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

- Upcoming catalysts or risk factors affecting ETH from now until July 2025 (e.g., expected upgrades, ETF approvals/rejections, major regulations, known hack/theft risks, etc.).
 - Source: crypto news, Ethereum Foundation roadmaps, industry reports.
- General market consensus or well-regarded forecasts for ETH price in mid-late 2025.
 - Source: analyst price targets, industry reports, prediction market odds for July 2025.
- Implied expectations from Ethereum derivatives markets (options/futures) expiring around July 2025, if available.
 - Source: Deribit, CME, other derivatives market data/analysis.

1.3. FACTS TO DERIVE

- A range of plausible ETH/USD price targets for the week of July 21-27, 2025, based on available data and trends.
- Assessment of which thresholds (A–J) are more likely to be reached or not based on fundamental, technical, and market sentiment factors.
- Probabilities or relative likelihoods for each option (if possible), to identify which boxed options are justified.
- Reasoning process to rule in/out thresholds based on volatility, historical growth rates, trend continuation or reversal scenarios.
- Synthesis of risk factors or events that could cause extreme moves (upside or downside) by July 2025.

2. PLAN

- Look up the current price of Ethereum (ETH) in USD, as well as its price trend over the past year.
- Gather historical ETH price data from prior years, especially from June/July to the following July, to understand typical annual volatility and trend direction.
- Search for major events, upgrades, crypto-specific and macroeconomic risks, and catalysts expected between now and July 2025 that might affect Ethereum's price.
- Look up analyst price predictions, industry reports, and any well-regarded forecasts for ETH in mid-late 2025.
- Search for information from the derivatives markets (particularly options and futures pricing for ETH with expiry around July 2025) to infer market-implied price expectations and volatility.
- Synthesize all gathered information to construct a reasoned, data-driven price range for ETH during the week of July 21-27, 2025.
- Match the forecasted price range to the specified answer options (A–J) and identify all plausible thresholds ETH could hit during the target week.
- Format the prediction in the required boxed format, including all plausible letter options.

Gemini-2.5-pro

1. FACTS SURVEY

1.1. FACTS GIVEN IN THE TASK

- Asset to Predict:** Ethereum (ETH).
- Prediction Timeframe:** The week of July 21-27, 2025.
- Reference Timestamp:** The prediction is contextualized around 2025-07-28T04:00:00Z.
- Prediction Options:**
 - A. Ethereum reach \$4400

1782

1783

1784

- B. Ethereum reach \$4300
- C. Ethereum reach \$4200
- D. Ethereum reach \$4100
- E. Ethereum reach \$4000
- F. Ethereum dip to \$3700
- G. Ethereum dip to \$3600
- H. Ethereum dip to \$3500
- I. Ethereum dip to \$3400
- J. Ethereum dip to \$3300

1785

1786

1787

1788

1789

1790

1791

1792

1793

1794

1795

1796

1797

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

- **Output Constraint:** The final answer must be a boxed list of plausible options, separated by commas (e.g., \boxed{A, B, C}).

1.2. FACTS TO LOOK UP

- **Current and Historical Price of Ethereum:** We need to know the current price to have a baseline, and the all-time high and recent price action to understand its volatility and historical cycles.
 - **Source:** Use the ‘search-agent’ to query financial data websites or crypto news outlets.
- **Expert Price Predictions for 2025:** We need to gather price targets and forecasts for Ethereum for mid-2025 from reputable financial institutions, crypto analysts, and investment firms.
 - **Source:** Use the ‘search-agent’ to find reports and articles from sources like Bloomberg, CoinDesk, Messari, Goldman Sachs, JPMorgan, etc.
- **Ethereum’s Technical Roadmap for 2024-2025:** We need to identify any major network upgrades, forks, or Ethereum Improvement Proposals (EIPs) scheduled for the period leading up to July 2025. Major upgrades like the “Pectra” fork can act as significant price catalysts.
 - **Source:** Use the ‘search-agent’ to search the official Ethereum Foundation blog, crypto-focused news sites (e.g., The Block, CoinDesk), and developer forums.
- **Macroeconomic Forecasts for mid-2025:** We need to understand the predicted global economic environment for mid-2025, particularly regarding interest rates (e.g., US Federal Reserve’s stance), inflation, and overall market sentiment (risk-on vs. risk-off).
 - **Source:** Use the ‘search-agent’ to find reports from major financial news outlets (Reuters, Wall Street Journal) and international economic bodies (IMF, World Bank).
- **Regulatory Climate for Ethereum:** We need to find information on any anticipated regulatory decisions or changes in legislation concerning Ethereum and crypto assets in major jurisdictions (especially the U.S. SEC) for late 2024 and H1 2025. The status and performance of spot Ether ETFs will be particularly relevant.
 - **Source:** Use the ‘search-agent’ to search for articles on crypto regulation from reputable legal and financial news sources.

1.3. FACTS TO DERIVE

- **Synthesized Market Sentiment:** From the facts we look up, we need to derive a consolidated view of the likely market sentiment for Ethereum in mid-2025. This involves weighing the bullish factors (e.g., successful upgrades, positive expert predictions, favorable macro environment) against bearish factors (e.g., regulatory crackdowns, economic recession, technical delays).
- **A Probable Price Range:** Based on the synthesized sentiment and the specific price targets found, we will derive a most likely price range for Ethereum during the target week. This will not be a single point but a zone of probability.
- **Selection of Plausible Options:** By comparing our derived probable price range with the given options (A-J), we will identify all options that fall within or are very close to our analytical conclusion.

1836

1837

1838

2. PLAN

1839

1840

1841

1842

1843

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855

1856

1857

1858

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

1870

1871

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

1886

1887

1888

1889

Qwen3-32B

1. FACTS SURVEY

1.1. FACTS GIVEN IN THE TASK

- The event to predict is the price of Ethereum (ETH) on July 21–27, around 2025-07-28T04:00:00Z.
- There are multiple potential outcomes listed as options A through J.

1.2. FACTS TO LOOK UP

- Historical and recent price data for Ethereum from reliable financial sources or cryptocurrency market platforms.
 - **Where to find:** Cryptocurrency tracking websites like CoinMarketCap, CoinGecko, or financial data platforms such as Yahoo Finance or Bloomberg.
- Any relevant news, events, or macroeconomic indicators that may influence Ethereum's price during this timeframe.
 - **Where to find:** News outlets like Reuters, Bloomberg, Coindesk, or crypto-focused forums like Reddit or Twitter/X.
- Technical analysis reports or forecasts about Ethereum from trusted analysts or institutions.
 - **Where to find:** Reports from investment firms, analyst blogs, or social media posts by well-known crypto analysts.

1.3. FACTS TO DERIVE

- Based on historical trends and current data, identify if Ethereum is likely to rise or dip in the specified time frame.
- Compare the probability of each option based on available data and expert predictions.
- Determine a set of plausible outcomes (options A–J) that align with the derived probabilities.

1890

1891

1892

2. PLAN

1893

1894

1895

1896

1897

1898

1899

1900

1901

1902

1903

1904

1905

1906

1907

1908

1909

Deepseek-V3

1910

1911

1912

1. FACTS SURVEY

1913

1.1. Facts given in the task

1914

1915

1916

1917

- The event to predict is the price of Ethereum (ETH) between July 21–27, 2025.
- The prediction options range from \$4400 (highest) to \$3300 (lowest).
- The prediction must be formatted as \boxed{A, B, ...} for plausible options.

1918

1.2. Facts to look up

1919

1920

1921

1922

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

- Historical price trends of Ethereum (especially around similar timeframes in previous years).
- Current price of Ethereum and its recent volatility.
- Upcoming Ethereum network upgrades or events (e.g., EIPs, hard forks) that could impact price.
- Macroeconomic factors (e.g., Bitcoin halving cycles, regulatory news) that could influence Ethereum's price.
- Market sentiment indicators (e.g., futures, options, institutional interest).
- Expert predictions or analyst reports for Ethereum's price in mid-2025.
- Sources: Crypto market data platforms (CoinGecko, CoinMarketCap), Ethereum Foundation announcements, financial news (Cointelegraph, Decrypt), and analyst reports (e.g., Ark Invest, Glassnode).

1.3. Facts to derive

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

- Likely price range based on historical growth rates and volatility.
- Correlation between Bitcoin halving cycles (April 2024) and Ethereum's price 15 months later.
- Impact of known future Ethereum upgrades (e.g., Proto-Danksharding) on price.
- Probability distribution of the given options based on historical data and expert consensus.

2. PLAN

1940

1941

1942

1943

1. **Research Ethereum's historical price trends:** Analyze ETH's price movements in similar timeframes (e.g., 15 months post-Bitcoin halving) to identify patterns.
2. **Grab current market data:** Retrieve ETH's current price, volatility, and institutional interest to assess short-term momentum.

1944

1945

1946

1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

3. **Identify upcoming Ethereum events:** Look for confirmed upgrades or events in 2025 that could drive price action.
4. **Analyze macroeconomic factors:** Check for broader market trends (e.g., Fed rate cuts, regulatory clarity) that could impact ETH.
5. **Compile expert predictions:** Aggregate mid-2025 price forecasts from reputable analysts and institutions.
6. **Derive plausible price range:** Combine historical trends, current data, and expert opinions to narrow down the most likely options.
7. **Finalize prediction:** Select the most plausible options (e.g., if \$4000–\$4300 is the consensus range, choose B, C, D, E) and format the answer as $\boxed{A, B, \dots}$.

 $\boxed{B, C, D, E}$

GPT-4o-mini

1. FACTS SURVEY

1.1. FACTS GIVEN IN THE TASK

- The event to predict is the price of Ethereum around July 21-27, 2025.
- The potential price points provided are: \$4400, \$4300, \$4200, \$4100, \$4000, \$3700, \$3600, \$3500, \$3400, and \$3300.

1.2. FACTS TO LOOK UP

- Historical price data of Ethereum, especially around July of previous years, to identify trends or patterns. This data can be found on cryptocurrency data platforms like CoinMarketCap or CoinGecko.
- Predictions or forecasts from cryptocurrency analysts for Ethereum prices in 2025. This information can be found in articles or publications on cryptocurrency news websites or financial analysis reports.
- Market conditions or significant events that may affect Ethereum's price around that time-frame (e.g., regulatory changes, technological upgrades, macroeconomic factors). This can also be sourced from news outlets or industry reports that discuss the future of cryptocurrencies.

1.3. FACTS TO DERIVE

- Trend analysis to determine how historical price movements correlate with specific months or events in the crypto market.
- An aggregate prediction based on gathered forecast data and price trends, leading to a set of plausible options based on market behavior leading up to the specified date.

2. PLAN

1. Gather historical price data for Ethereum around July from previous years using CoinMarketCap or CoinGecko.
2. Analyze the historical price trends to identify patterns or significant events that could provide insight into Ethereum's price movements during July in past years.
3. Search for cryptocurrency market predictions or analysis related to Ethereum for the year 2025 using news websites or financial analysis platforms.
4. Review and compile the significant factors that could affect Ethereum's price between now and July 2025, including potential regulatory developments or technological advancements.
5. Synthesize the information collected, relating historical trends to potential future outcomes.
6. Make a prediction based on the derived data, focusing on the most plausible price point options.

1998

1999

2000

2001

2002

2003

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014

2015

2016

2017

2018

2019

2020

2021

2022

2023

2024

2025

2026

2027

2028

2029

2030

2031

2032

2033

2034

2035

2036

2037

2038

2039

2040

2041

2042

2043

2044

2045

2046

2047

2048

2049

2050

2051

- Format the final answer according to the specified format, ensuring to list all plausible price options identified in the box format.

Qwen3-235B

1. FACTS SURVEY

1.1. FACTS GIVEN IN THE TASK

- The task asks for a prediction about Ethereum's price between July 21-27 around 2025-07-28T04:00:00Z
- The possible options for prediction are:
 - A. \$4400, B. \$4300, C. \$4200, D. \$4100, E. \$4000
 - F. \$3700, G. \$3600, H. \$3500, I. \$3400, J. \$3300

1.2. FACTS TO LOOK UP

- Current Ethereum price and recent price trends: Need to understand the current market situation and trajectory.
- Historical Ethereum price patterns: Useful for identifying seasonal trends or recurring patterns that might help in predicting future prices.
- Cryptocurrency market forecasts for 2025: To understand expert opinions and analyses regarding the future of Ethereum.
- Macroeconomic factors that could influence cryptocurrency prices in 2025: Such as regulatory changes, adoption by major institutions, technological developments, etc.
- Any significant events or developments in the Ethereum ecosystem expected around July 2025: Like protocol upgrades, major partnerships, or ecosystem growth.

1.3. FACTS TO DERIVE

- Based on current trends and historical data, we can derive potential price trajectories for Ethereum leading up to July 2025.
- By analyzing market sentiment and macroeconomic indicators, we can derive likely market conditions for Ethereum in July 2025.
- We can calculate potential price ranges using technical analysis tools like moving averages, Fibonacci retracements, etc.
- Considering all available information, we can derive a probability distribution across the given price levels to identify the most plausible outcomes.

2. PLAN

- Conduct a comprehensive search on the current price of Ethereum, its recent performance, and current market sentiment.
- Investigate historical price patterns of Ethereum to identify any recurring trends or cycles that could help in predicting future prices.
- Search for expert forecasts and market analyses specifically addressing Ethereum's price expectations for 2025.
- Research macroeconomic factors, regulatory developments, and technological advancements that could impact Ethereum's price in 2025.
- Look for any specific information about events or developments in the Ethereum ecosystem expected around July 2025.
- Analyze all gathered information to determine the most likely price range for Ethereum during July 21-27, 2025.
- Map the derived price range to the specific price levels provided in the options (A-J).

2052

2053

2054

2055

2056

2057

2058

2059

2060

2061

2062

2063

2064

2065

2066

2067

2068

2069

2070

2071

2072

2073

2074

2075

2076

2077

2078

2079

2080

2081

2082

2083

2084

2085

2086

2087

2088

2089

2090

2091

2092

2093

2094

2095

2096

2097

2098

2099

2100

2101

2102

2103

2104

2105

8. Calculate probabilities for each option based on the analysis and select the most plausible options.
9. Provide the final answer using the exact box format with the identified options.

J PROMPTS

J.1 PROMPT FOR PROPOSING NEW QUESTION FOR FUTURE PREDICTION

Please help me find some websites of data/lists/rankings/statistical tables.

Requirements:

1. The data/lists/rankings/statistical tables are real and released by authoritative and well-known institutions.
2. The data/lists/rankings/statistical tables have official websites.
3. The data/lists/rankings/statistical tables are **regularly updated**, such as daily, weekly, or monthly. This point is very important.
4. Each data/list/ranking/statistical table is always updated on a **fixed and unchanging** website. This point is very important.
5. The data/lists/rankings/statistical tables are **directly visible**, that is, you can directly see the complete content of the data/list/ranking/statistical table when you open the website, without any need for clicking, drop-down selection, page-turning, downloading, etc. This point is very important.

The following are some examples that meet the requirements and examples that do not meet the requirements. Please visit these websites one by one to understand the above requirements, especially the three points of **regular update**, **fixed and unchanging**, and **directly visible**.

Example 1 that meets the requirements: Top Ten in Terms of Market Capitalization of Stocks on the Shanghai Stock Exchange

Website: <https://www.sse.com.cn/market/stockdata/marketvalue/main/>

The list is updated daily, always on this website, and you can directly see the complete list when you open the website.

Example 2 that meets the requirements: Billboard Hot 100

Website: <https://www.billboard.com/charts/hot-100/>

The ranking is updated weekly, always on this website, and you can directly see the complete ranking when you open the website.

Example 1 that does not meet the requirements: Central Parity Rate of RMB Exchange Rate

Website: <http://www.pbc.gov.cn/zhengcehuobisi/125207/125217/125925/index.html>

Reason for not meeting the requirements: The data of the central parity rate of the RMB exchange rate is updated daily, but it is updated on different websites every day, which does not meet the requirement of “fixed and unchanging”.

Example 2 that does not meet the requirements: Added Value of Industrial Enterprises above Designated Size

Website: <https://www.stats.gov.cn/sj/zxfb/>

Reason for not meeting the requirements: The data of the added value of industrial enterprises above designated size is updated monthly, but it is updated on different websites every month, which does not meet the requirement of “fixed and unchanging”.

2106 J.2 PROMPT FOR FUTURE PREDICTION
21072108 **For the multi-choice questions**
21092110
2111 You are an agent that can predict future events. The event to be predicted: “{title} (around {time}).
2112 {options}”
2113 IMPORTANT: listing all plausible options you have identified, separated by commas, within the box.
2114 For example: \boxed{A} for a single option or \boxed{B, C, D} for multiple options.
2115 Do not use any other format. Do not refuse to make a prediction. Do not say “I cannot predict the
2116 future”. You must make a clear prediction based on the best data currently available, using the box
2117 format specified above.
21182119 **For the other questions**
21202121
2122 You are an agent that can predict future events. The event to be predicted: “Please Predict Beijing
2123 Time {time}, {title}”
2124 IMPORTANT: Your final answer MUST end with this exact format: **PREDICTION**
2125 Do not use any other format. Do not refuse to make a prediction. Do not say “I cannot predict the
2126 future”. You must make a clear prediction based on the best data currently available, using the box
2127 format specified above.
21282129 **K HUMAN ANNOTATION DETAILS**
21302131
2132 In order to compare recent advanced LLMs with human experts, we conduct human annotation
2133 experiments. In this section, we provide more details.
21342135 **K.1 NUMBER OF QUESTIONS PER CATEGORY**
21362137 We involve 31 human experts and collect their predictions prior to the resolution date of each task.
2138 Since FutureX is a live benchmark, it is challenging to gather human predictions on a daily basis,
2139 which makes it nearly impossible to align the tasks for LLMs and human experts consistently. As
2140 a result, we acknowledge that some mismatch between human and LLM tasks is inevitable. The
2141 number of questions in each difficulty tier is as follows:
21422143 Table 6: Scores and Number of Questions by Level
2144

	Level-1	Level-2	Level-3	Level-4
Num Questions	106	72	44	85
Human Score	79%	39%	48%	24%

2149 **K.2 QUALITY CONTROL**
21502151 **Information access for human experts.** As will be demonstrated below, regarding the information
2152 sources for human experts, our annotation guidelines explicitly require experts to use web search
2153 to find relevant information before making their final prediction. This ensures that human experts,
2154 like the agents, are tasked with synthesizing information from external sources rather than relying
2155 solely on their existing knowledge.
21562157 **Consistency of answers.** We acknowledge that having multiple annotators for each question
2158 would enhance the statistical robustness of our results. However, future prediction is a significantly
2159 more labor-intensive task than typical annotation. Our protocol requires each annotator to spend
at least 5 minutes per question, document their thought process and action steps, perform at least
one web search, and record their screen for verification. To ensure high-quality annotations under
2159

2160 these demanding conditions, we conduct a rigorous verification process, including manual checks of
 2161 the recorded logs and screen recordings for each annotation. Given these expensive annotation and
 2162 verification costs, we prioritized in-depth quality control for single, well-vetted annotations over col-
 2163 lecting multiple annotations for the same question. We believe this could maintain the high fidelity
 2164 of our human performance baseline.

2166 K.3 HUMAN ANNOTATION GUIDELINES

2168 We ask human annotators to predict future events by following a specific set of guidelines. The core
 2169 principle of this project is to rely solely on human reasoning and publicly available information. The
 2170 use of any AI tools is strictly forbidden. A critical component of the task is for each annotator to
 2171 provide a detailed and logical thinking process that leads to their final prediction. This ensures that
 2172 every prediction is based on verifiable information and sound human judgment, rather than on an
 2173 AI's output.

Human Annotation Guidelines

ANNOTATION BACKGROUND

We need to collect predictions for future events. You will predict the outcomes of events that will occur within the next seven days, such as the result of a sports match or a company's stock price change. You must gather information from the internet to predict the outcomes of these events that have not yet happened.

ANNOTATION RULES

- **Do not use large language models (LLMs) or AI software for predictions.** If a screencast shows the use of AI software, all tasks for that day will be void.
- **Time Requirements for Solutions** (All expert-level tasks are considered difficult):
 - Each task must take a minimum of **5 minutes** to solve.
 - Each task must include a minimum of **3 steps**.
 - You must consult a minimum of **1 web page**.
- If a screencast shows prolonged pauses or other time-wasting behaviors, the task will be void.
- The entire solution process for each task must be recorded in a screencast.

THOUGHT PROCESS & PREDICTION RATIONALE

You must briefly write down your thought process and the reasoning behind your prediction.

TEMPLATE FOR THOUGHT PROCESS & PREDICTION RATIONALE

1. **Search Keywords:** {keywords}, **Accessed Webpage:** {webpage}
2. **Observations:** {observations}
3. **My Reasoning:** {reasoning} (The reasoning must clearly explain how you reached the prediction from your observations).
4. If a correct prediction can be made, stop. Otherwise, repeat the above steps.

ANNOTATION BONUS

- If the screencast and solution process meet the requirements, you will receive the **basic reward**, even if the prediction is incorrect.
- If the final prediction is correct and the thought process and reasoning are sound, you will receive an **additional bonus** (the bonus for difficult tasks is higher than for simple ones).
- If the recorded thought process and reasoning are unreasonable or perfunctory, you will receive **no reward**, regardless of whether the prediction is correct.

IMPORTANT NOTES

Before submitting the required screencast videos for this project, please ensure they do not contain your private or personal information. If they do, please redact or anonymize the information before submission.

2214

2215

2216

2217

2218

2219

2220

2221

2222

2223

2224

2225

2226

2227

2228

2229

2230

2231

2232

2233

2234

2235

2236

2237

2238

2239

2240

2241

2242

2243

2244

2245

2246

2247

2248

2249

2250

2251

2252

2253

2254

2255

2256

2257

2258

2259

2260

2261

2262

2263

2264

2265

2266

2267

ANNOTATION TASK OUTPUT

- Solution Process
- Screencast
- Prediction Result

L WEBSITES

Regarding the list of websites from which we collect data, we have made sure to only include publicly accessible sites that explicitly allow data scraping, or have no restrictions outlined in their terms of service. To ensure compliance, we have adhered to ethical data collection practices and ensured that our actions align with the site's usage policies.

In this section, we list only some of the publicly accessible websites that do not have restrictions on data access, and each website can generate various questions.

- <https://www.sse.com.cn/market/stockdata/marketvalue/main/market/stockdata/overview/day>
- <https://www.szse.cn/market/stock/active/index.html>
- <http://www.pbc.gov.cn/zhengcehuobisi/125207/125217/125925/index.html>
- <https://piaofang.maoyan.com/dashboard>
- <https://www.billboard.com/charts/hot-100>
- <https://www.stats.gov.cn/sj/zxfb/>
- <https://inside.fifa.com/fifa-world-ranking/men?dateId=id14597>
- <https://www.timeshighereducation.com/cn/world-university-rankings/latest/world-ranking>
- <https://weather.cma.cn/latest>
- <https://indexes.nikkei.co.jp/cz/nkave/index/profile>
- <https://www.dongchedi.com>
- <https://pfsc.agri.cn/>
- <https://xueqiu.com>
- <https://store.steampowered.com>
- <https://www.hulu.com>
- <https://web.51.la>
- <https://www.ufc.com/rankings>
- <https://www.kolrank.com/>
- <https://worldathletics.org>

M THE USE OF LARGE LANGUAGE MODELS

In this work, LLMs were only used as a general-purpose language polishing tool to improve the clarity, fluency, and grammatical accuracy of the manuscript text, without involving any contribution to core research content. All contents were checked and edited by the authors to ensure alignment with the technical content. The authors take full responsibility for the final version of the paper.