
Evaluating Strategic Reasoning in Forecasting Agents

Tom Liptay
FutureSearch
tom@futuresearch.ai

Dan Schwarz
FutureSearch
dan@futuresearch.ai

Rafael Poyiadzi
FutureSearch
rafael@futuresearch.ai

Jack Wildman
FutureSearch
jack@futuresearch.ai

Abstract

Forecasting benchmarks produce accuracy leaderboards but little insight into why some forecasters are more accurate than others. We introduce Bench to the Future 2 (BTF-2), 1,417 pastcasting questions with a frozen 15M-document research corpus in which agents reproducibly research and forecast offline, producing full reasoning traces. BTF-2 detects accuracy differences of 0.004 Brier score, and can distinguish differential agent strengths in research vs. judgment. We build a reference forecaster significantly more accurate than any single frontier agent, and use it to evaluate agent strategic reasoning without hindsight bias. We find the better forecaster differs primarily in its pre-mortem analysis of its blind spots, and consideration of black swans. Expert human forecasters found the dominant strategic reasoning failures of frontier agents are in assessing political and business leaders’ incentives, and judging their likelihood to follow through on stated plans.

1 Introduction

Forecasting real-world outcomes is extremely difficult, with no clear upper bound on accuracy. AI forecasting systems have shown continuous improvement since Autocast [Zou et al., 2022] in 2022, but existing benchmarks produce leaderboards without explaining them.

Frontier AI forecasters are agents that search and read web pages to research each question. Most efforts to evaluate forecasting agents have used live forecasting questions, but since the ground truth state of the world changes continuously, forecasts on the same question are not reproducible, making failure analysis subject to hindsight bias. Approaches like ForecastBench [Karger et al., 2025], the Metaculus AI Benchmarking Series [Metaculus, 2024], and Prophet Arena [Yang et al., 2026] use live forecasting questions.

But reproducible approaches risk data leakage. Alur et al. [2025], for instance, relies on an LLM-judge to detect potential information leakage of the future. Leakage has been documented in the approach used by Phan et al. [2024].

The original Bench to the Future (BTF) [Wildman et al., 2025] introduced a hermetic offline corpus, a snapshot of tens of thousands of web pages for each forecasting question, captured at question creation time, ensuring that models cannot access information after the question is generated. However, it had only 299 easy, correlated questions from early 2025, predating most frontier models’ training cutoffs.

2 The BTF-2 Benchmark

Bench to the Future 2 (BTF-2) is based on a dataset of 1499 questions discussed in [Bosse et al. \[2026\]](#), originally asked in October 2025 and resolved in December 2025. Each question has an average of $\approx 10,100$ web pages scraped and stored offline (range: 6,000–19,000), for a total of ≈ 16.2 million documents, of which ≈ 8.7 million pages are unique.

We reduced this question set to 1417 questions, removing questions where timing mismatches between generation, scraping, and resolution made them misleading. We applied conservative filters, removing any question with ambiguous resolution criteria, risk of incorrect resolution, or outcomes knowable before the question was posed. The questions span regulatory and policy actions ($\approx 23\%$), US government and public policy ($\approx 13\%$), macroeconomics and markets ($\approx 12\%$), international security and diplomacy ($\approx 11\%$), court cases and investigations ($\approx 10\%$), and smaller clusters covering Gaza war diplomacy, extreme weather, Iran’s nuclear program, commercial space launches, COP30 climate negotiations, and elections.

LLMs with training cutoffs before Oct 2025 can pastcast these questions without contamination risk. (We verified that Opus 4.7, with a January 2026 cutoff, showed leakage and cannot be evaluated on BTF-2.) We use the RetroSearch system described in [\[Wildman et al., 2025\]](#) to provide Search and Page Read tools to the forecasting agents, approximating live internet search.

2.1 Validating the statistical power

The best forecast from [Bosse et al. \[2026\]](#) used live-web research summaries with Gemini 3.0 Pro Preview and a structured prompt with forecasting tips (the “Bosse et al. prompt”). We re-ran this exact setup—same research summaries, same prompt—with the newer Gemini 3.1 Pro Preview.

Table 1: Validating detection of model improvements: Gemini 3.1 Pro vs. 3.0 Pro.

Forecaster	N	Brier	Δ	95% CI
Gemini 3.1 Pro (re-run)	1417	0.129	—	—
Gemini 3.0 Pro (original)	1417	0.138	0.009	[0.002, 0.016]

*Both forecasters used the same pre-gathered live-web research summaries and the Bosse et al. prompt; they differ only in model version. Δ is the Brier difference relative to the top row. Best Brier is **bolded**.*

The newer model achieves significantly better accuracy on the same questions with the same approach (Table 1). This is consistent with Gemini 3.1 Pro being an overall more intelligent model.

2.2 Validating the question difficulty

To validate that BTF-2 questions are genuinely difficult, but tractable, we measured divergence across repeated independent rollouts of a strong forecasting agent.

We selected 200 questions that live (prospective) Gemini 3 Pro runs from [Bosse et al. \[2026\]](#) had identified as difficult, based on high Brier scores and inter-run variance. We then evaluated their difficulty in BTF-2 using an Opus 4.6 agent, the strongest single forecasting agent as shown in Table 3. We ran 8 independent Opus 4.6 agent rollouts on each question. Table 2 shows how many of the 8 rollouts landed on the correct side of 50%.

Table 2: Opus 4.6 Agent rollout consistency on 200 difficult questions (8 independent rollouts each).

Correct rollouts (of 8)	0	1–3	4	5–7	8
Questions	89 (45%)	36 (18%)	5 (3%)	34 (17%)	36 (18%)

On 38% of questions, Opus produced rollouts on both sides of the 50% threshold (mean per-question $\sigma = 0.08$). Of the remainder, we cannot distinguish between those that are inherently surprising, e.g. a forecast of 20% would be excellent given the information available, but the event in fact occurs, as it should one in five times, vs. those that exceed the forecasting agent’s ability.

3 Methodology

Forecasting attempts in BTF-2 produce a full sequence of searches, page reads, thoughts, and a final rationale, and can be re-run many times. This enables a series of experiments on the research process and strategic reasoning of the forecasters, in addition to experiments on accuracy.

Agent evaluation. We address three questions: (1) Which frontier LLM makes the forecasting agent? (2) Does the best agent’s advantage come from research strategy or from judgment? (3) How do frontier agents compare to a state-of-the-art forecaster?

In all experiments with forecasting agents, we evaluate agents run by frontier models at high effort with training window cutoffs before Oct 2025, from their commercial APIs. All agents use a ReAct-style architecture [Yao et al., 2023]. Specifically, we use FutureSearch’s ReAct implementation [FutureSearch, 2026], whose system prompt, toolkit, and time-management mechanism were tuned against Deep Research Bench [FutureSearch, 2025], a benchmark of challenging web-research tasks. We gave the ReAct agents an iteration budget of 10, so up to 10 times, agents can call a combination of parallel Search and Read tools. A typical agent makes ≈ 10 – 20 tool calls, takes about 5–6 minutes total, processes ≈ 80 – 200 k input and output tokens, looks at 50–300 search snippets, reads 5–20 pages in full, and costs \$0.15–\$0.50. For Q(2), we compare agents to models given pre-gathered research.

Reasoning patterns across different forecasting agents. To understand what distinguishes better forecasting agents, we score each forecast rationale against Tetlock’s CHAMPS KNOW framework [Tetlock and Gardner, 2015, Chang et al., 2016], which identifies 10 dimensions associated with better human forecasting accuracy. For each rationale, a Gemini 3.1 Pro agent ranks the 10 dimensions from 1 (most prominent) to 10 (least prominent). We compare dimension profiles across models and between individual forecasting agents and the refinement ensemble (Section 4.4).

Strategic reasoning mistakes of the best forecasting agents. To understand the biggest weaknesses of the best forecasting agents, we had expert human forecasters examine a sample of forecasts to judge whether the failures were due to a significant error in judgment—one that an expert human forecaster would be unlikely to make.

To isolate LLM judgment from scaffold effects, we used the best single agent from Section 4.1 (no ensembling, no forecasting guidance).

To ensure the failures were not simply human hindsight bias from knowing the outcome, in addition to looking at the worst $\approx 5\%$ by absolute accuracy score, we also looked at the worst $\approx 5\%$ as measured by the accuracy difference between that no-guidance forecasting agent and the SOTA forecasting agent from Section 4.3, for a total of 130 questions.

To verify reproducibility, we re-ran the agent and retained only cases where both runs made the same error.

4 Results

4.1 Comparing Frontier Agents

We first evaluate agents with no guidance on forecasting practices, approximating the experience of asking ChatGPT, Claude, or Gemini on a standard “pro” subscription tier.

We evaluated Claude Opus 4.6 (thinking: “adaptive”, effort: “high”), Google Gemini 3.1 Pro Preview (thinking_level: “high”), OpenAI GPT-5.4 (reasoning_effort: “high”), and Grok 4.20 Beta. Every agent was given the following straightforward prompt:

“You have been given a prediction question with its resolution criteria. Your task is to research this question and produce the most accurate probabilistic forecast you can. Write a brief rationale summarizing your research and reasoning, then provide your final forecast as a probability between 0 and 100.”

Table 3: Frontier agent accuracy on BTF-2.

Agent	N	Brier	Δ	95% CI	Calibration	Refinement
Opus 4.6 Agent	1417	0.131	—	—	0.005	0.073
Gemini 3.1 Pro Agent	1417	0.143	0.012	[0.002, 0.021]	0.012	0.067
GPT-5.4 Agent	1417	0.151	0.020	[0.012, 0.029]	0.010	0.057
Grok 4.20 Beta Agent	1300	0.165	0.033	[0.022, 0.044]	0.003	0.039

All agents use the simple prompt (Section 4.1) with no forecasting guidance. Δ is the Brier difference relative to Opus 4.6 Agent. Best value in each column is **bolded**. Grok ran on $N = 1300$ due to technical failures on 117 questions.

Table 4: Judgment over fixed evidence.

Model	N	Brier	Δ	95% CI	Calibration	Refinement
Gemini 3.1 Pro	1417	0.141	0.012	[0.006, 0.018]	0.012	0.069
Opus 4.6	1417	0.153	0.024	[0.015, 0.033]	0.015	0.061
GPT-5.4	1417	0.156	0.027	[0.019, 0.034]	0.026	0.068
Grok 4.20 Beta	1417	0.163	0.033	[0.025, 0.042]	0.020	0.056

Models were given pre-gathered research summaries with no ability to search further. Δ is the Brier difference relative to the best single-model forecast (0.129; Table 1). Best value in each column is **bolded**.

Table 3 shows the resulting Brier scores. Δ corresponds to the difference in Brier score with Opus 4.6 on the same set of questions. 95% CI is calculated with bootstrapping throughout the paper. Calibration and refinement are calculated as in Bosse et al. [2026].

The Opus 4.6 agent’s superior accuracy over the other 3 leading agents is statistically significant at the 95% confidence level. The other three agents have a clear ordering in accuracy, with Gemini 3.1 Pro second, followed by GPT-5.4 and then Grok 4.20 Beta.

Brier scores decompose into calibration and refinement. Calibration measures alignment between forecast probabilities and actual frequencies (lower is better), reflecting how well a forecaster knows the limits of its knowledge. Refinement measures ability to differentiate between outcomes (higher is better), reflecting how much useful information a forecaster brings to bear.

Opus 4.6 is better calibrated than Gemini 3.1 Pro and GPT-5.4. Surprisingly, we find that the Grok 4.20 agent is the best calibrated among the agents, despite having the worst accuracy. This is the equivalent of a meteorologist who always forecasts the base rate for rain in London, never looking at a radar. They are perfectly calibrated, but have zero refinement.

Refinement rankings match the accuracy ordering.

4.2 Isolating Judgment from Research

What explains Opus 4.6’s lead from Section 4.1? Is it superior research strategy in choosing which searches to run and pages to read, or superior judgment over the evidence it collects? To isolate judgment, we gave all four models the pre-gathered research summaries from Bosse et al. [2026], with no ability to search further, and this simple prompt:

“Based on the information provided, forecast the probability of a ‘YES’ outcome as a percentage between 0 and 100. Write a brief rationale, then provide your final answer.”

Forecasting agents with this pre-compiled research did significantly worse than when conducting their own research. Opus 4.6 declined from 0.131 to 0.153, with worse calibration and refinement, indicating its agent research finds information the pre-gathered summaries miss.

Refinement improved for Gemini 3.1 Pro, GPT-5.4, and Grok 4.20, suggesting the Bosse et al. research exceeded what those agents found independently. Calibration declined for all models, for unclear reasons.

Table 5: Best Brier score by research source and prompt.

	Simple Prompt	Bosse et al. Prompt
Agent does research	Opus 4.6 (0.131)	Opus 4.6 (0.131)
Bosse et al. research provided	Gemini 3.1 Pro (0.141)	Gemini 3.1 Pro (0.129)

Table 6: SOTA and component forecasting agent performance.

Forecasting Agent	N	Brier	Δ	95% CI	Calibration	Refinement
SOTA Forecasting Agent	1367	0.119	—	—	0.002	0.081
Opus 4.6 Agent	1367	0.130	0.011	[0.006, 0.017]	0.005	0.075
Opus 4.6 Agent run 2	1367	0.130	0.011	[0.006, 0.017]	0.005	0.074
Gemini 3.1 Pro Agent	1367	0.141	0.023	[0.015, 0.031]	0.012	0.069
GPT-5.4 Agent	1367	0.152	0.034	[0.025, 0.042]	0.010	0.056
<i>Mean 4 agents</i>	<i>1367</i>	<i>0.125</i>	<i>0.007</i>	<i>[0.002, 0.011]</i>	<i>0.007</i>	<i>0.081</i>

Δ is the Brier difference relative to the SOTA Forecasting Agent. Best value in each column is **bolded**. $N = 1367$ (50 questions excluded due to technical issues in multi-agent runs).

We also tested Opus 4.6 as an agentic researcher with the elaborate Bosse et al. prompt. Table 5 summarizes the best Brier scores under each combination of agent does research vs. research provided, and simple prompt vs. prompt with forecasting best practices.

The Gemini 3.1 Pro re-run of the Bosse et al. method (Table 1) remains the most accurate single-model forecast. But BTF-2 evaluates agents conducting their research, the first row above, under which Opus 4.6 is best under both simple and elaborate agent prompts. Two examples of the full trace of the Opus 4.6 agent researching and forecasting are in the Appendix A.1.

4.3 State-of-the-art forecasting on BTF-2

BTF-2’s reproducibility enabled us to build a state-of-the-art (SOTA) forecasting agent substantially more accurate than any frontier agent. Its rationales provide a hindsight-bias-free standard for analyzing where frontier agents differ (Section 4.4) and fail (Section 4.5).

Accuracy-improving approaches include: (1) Taking the mean of multiple runs of the same agent (“wisdom of the crowd”); (2) Providing the agents with forecasts to related questions; (3) Re-calibrating forecasts based on historic calibration curves, (4) Asking a longer-scope version of the question first, and (5) Detecting and correcting for biases in forecast rationales.

The mean of four agents outperforms any individual (Table 6), showing a wisdom-of-crowds effect. The SOTA forecasting agent did significantly better than this mean.

The SOTA forecasts are also remarkably well calibrated, despite no explicit calibration step in the approach. Enhanced calibration appears to be an *emergent property* of the accuracy improvements. The calibration curves of the individual agents all have a distinct sag to them, similar to what is found on Metaculus or prediction markets with human forecasters, that the SOTA forecasting agent doesn’t show.

4.4 What do better forecasting agents focus on?

The SOTA forecasting agent from Section 4.3 makes this comparison possible. Because it is significantly more accurate than any single frontier agent, differences in what it emphasizes reflect what better forecasters focus on, without the risk of hindsight bias that would arise from knowing the outcomes.

To characterize these differences, we use Tetlock’s CHAMPS KNOW framework [Tetlock and Gardner, 2015], a set of 10 dimensions developed in the Good Judgment Project. Training human forecasters on these principles improved accuracy by 6–12% in a randomized trial [Chang et al., 2016]. We used Gemini 3.1 Pro to rank-order dimensions by prominence in each rationale. Table 7 shows top-3 frequency.

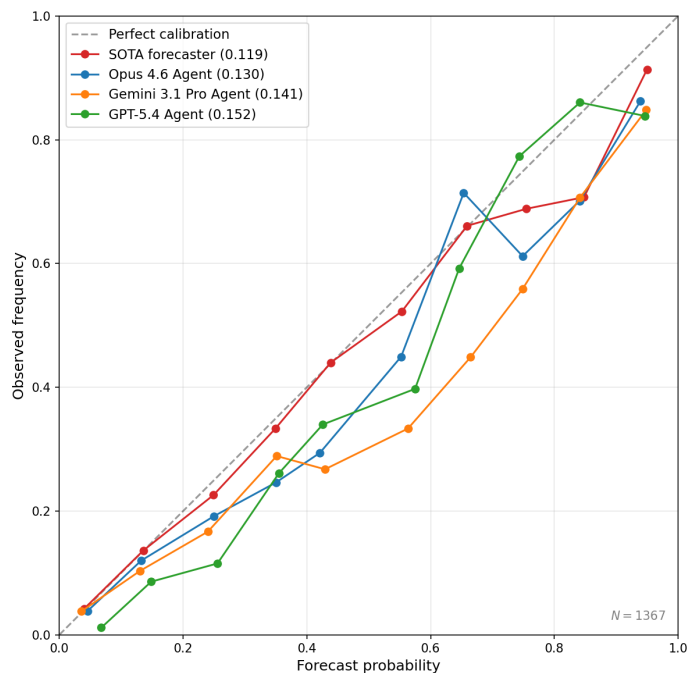


Figure 1: Calibration curves for the top three frontier agents and the SOTA forecaster ($N = 1367$). Dashed diagonal indicates perfect calibration. The SOTA forecaster exhibits markedly better calibration, closely tracking the diagonal. Brier scores in parentheses.

Table 7: CHAMPS KNOW top-3 frequency by forecasting agent.

Dim	Description	SOTA Agent	Opus 4.6 Agent	Gemini 3.1 Pro Agent	GPT-5.4 Agent
N	Norms & Protocols	63.1%	50.7%	57.6%	55.7%
K	Know power players	43.2%	33.6%	37.4%	16.4%
C	Comparison Classes	38.9%	34.1%	36.5%	24.4%
H	Hunt for info	38.0%	94.3%	84.4%	97.4%
P	Pre/Post-mortem	37.8%	9.5%	4.3%	6.8%
S	Select right Qs	28.1%	39.6%	40.3%	67.5%
M	Math/stats models	22.6%	26.5%	25.0%	18.9%
O	Other perspectives	20.3%	5.1%	1.7%	1.6%
A	Adjust & update	5.1%	5.9%	12.1%	10.8%
W	Wildcards	2.9%	0.7%	0.7%	0.3%

Values show the percentage of rationales in which each dimension ranked in the top 3. Agent labels correspond to the agents from Table 3.

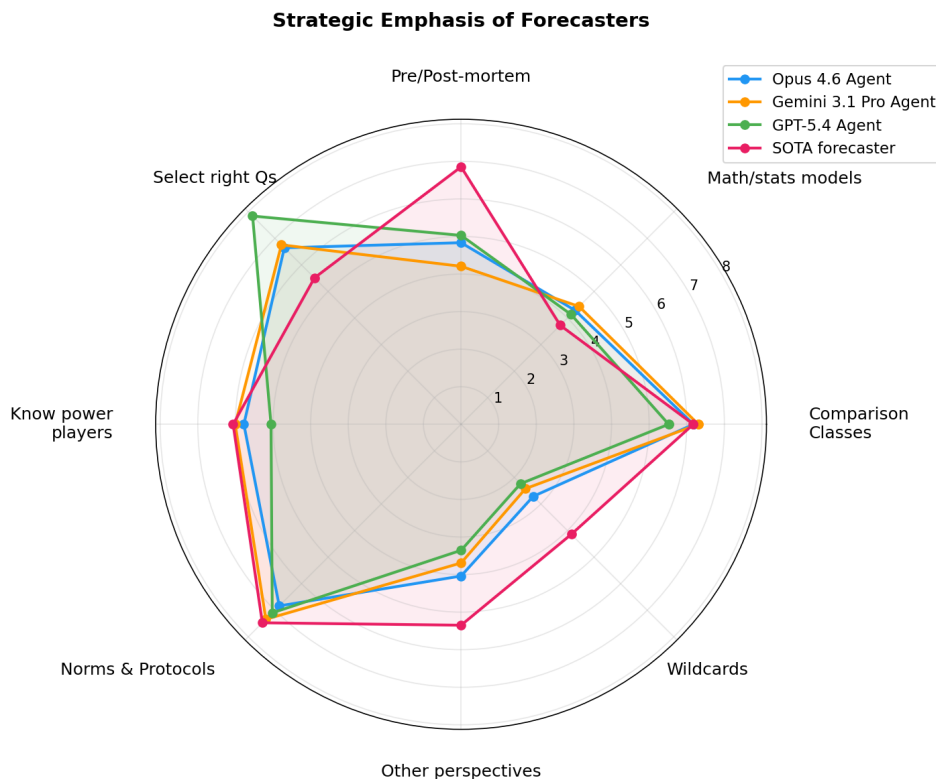


Figure 2: CHAMPS KNOW radar plot showing mean Borda score per dimension (rank 1 = 10 points, rank 10 = 1 point), excluding Hunt for Info and Adjust & Update which are uninformative for comparison. The SOTA forecasting agent emphasizes Pre/Post-mortem analysis, Other perspectives, and Wildcards far more than any individual frontier model agent.

Since rationales cannot include all considerations, these rankings reflect both strategy and prioritization. The SOTA forecasting agent also has structure in its rationale that is not present in the other agents. Human expert forecasters concluded this led to unrealistic comparisons in two of the factors: Hunt for Info, and Adjust and Update, so removed them from the direct comparison radar chart in Figure 2.

The biggest differences in the remaining 8 factors between the SOTA forecasting agent and the less accurate forecasting agents were Pre/Post-mortem, Other perspectives, and Wildcards. Pre/Post-mortem analysis refers to whether the forecasting agent considers, if its forecast ends up being wrong, the most likely reasons why this would occur. Other perspectives and Wildcards, taken together, refer to black swans, unknown unknowns, and blind spots.

All three are epistemic. The primary gap is awareness of uncertainty and the limits of one’s knowledge.

4.5 Strategic reasoning mistakes of the best forecasting agents

Expert human forecasters examined runs from the Opus 4.6 forecasting agent with the simple prompt from Section 4.1, reflecting the best accuracy without any special guidance on forecasting best practices. We selected the least accurate $\approx 5\%$ of questions overall, and the least accurate $\approx 5\%$ compared to the SOTA forecasting agent from Section 4.3, for a total of 130 questions. This ensures examined failures reflect achievable improvements, not merely surprising outcomes.

Even on these 130 questions, experts judged most forecasts high quality. On 24 of these 130 questions, they judged there were significant strategic reasoning errors. These were broadly in two categories: reasoning errors specific to the forecasting task, and reasoning errors in strategic foresight and world modeling.

Errors specific to the forecasting tasks can be broadly classified as one of these three categories: (1) Considering the most extreme version of an outcome, showing why that extreme outcome was unlikely, and downweighting the likelihood of a milder outcome; (2) Extrapolating from base rates too strictly even when the process leading to that base rate has been disrupted; (3) Producing a great forecast but then falling back to a more conservative prior.

Errors in strategic foresight and world modeling generally fell into two categories: (1) Failing to judge in which cases political or business leaders will follow through on their stated position; and (2) failing to understand the primary incentives faced by political or business leaders.

We include one case study of each of these two types of strategic errors. Full details are provided in Appendix A.1 and Appendix A.2.

4.5.1 Case Study: Judging when a political leader will follow through on their stated position

Question: “Will ASUU (Academic Staff Union of Universities, Nigeria) declare a nationwide university strike lasting at least 7 consecutive days?” (Oct 15–Dec 31, 2025)

Opus 4.6 Forecasting Agent: 72% / 75% SOTA Forecasting Agent: 30% Resolution: No

ASUU had begun a two-week warning strike on October 13, two days before the window opened, as part of a publicly stated escalation sequence: 14-day ultimatum, then two-week warning strike, then indefinite strike. The Opus 4.6 agent gave 72%, anchoring on this sequence and on a quote from ASUU National President Chris Piwuna that the next phase “will be total and there will be no going back,” which it positioned as its primary evidence. In fact, ASUU’s National Executive Council suspended the warning strike on October 22, giving the government a one-month window. When rumors of a November 21 indefinite strike circulated, ASUU’s national office explicitly denied them. On December 23, ASUU and the Federal Government signed a landmark renegotiated agreement, the first since 2009, including a 40% salary uplift. No qualifying strike occurred.

The Opus 4.6 agent’s error was treating Piwuna’s maximalist rhetoric as a commitment rather than as bargaining leverage. The same press conference contained an explicit hedge, “we will meet after the expiration to decide when to begin an indefinite and comprehensive strike action,” reserving the very decision the forecasting agent treated as already made. The SOTA forecasting agent at 30% prospectively identified structural signals pointing toward de-escalation: active negotiations had produced progress on five of seven demands; ASUU typically affords the government a grace period of several weeks after suspending a warning strike; the Nigeria Labour Congress’s parallel four-week ultimatum provided a face-saving clock; and Nigerian universities traditionally close for Christmas. The Opus 4.6 agent cited the negotiation progress but discounted it under the escalation-rhetoric frame, and never considered the grace-period pattern or academic-calendar seasonality.

Agent trace analysis confirmed that the Opus forecasting agent encountered and paraphrased Piwuna’s hedge (“the union will meet again after the two-week period”), but in its final analysis, it dropped this conditionality.

4.5.2 Case Study: Modeling an actor’s incentives

Question: “Will the Câmara dos Deputados (Brazilian House of Representatives) Plenary approve the lead proposition containing PL 1.874/2022 (National Circular Economy Policy) between 2025-10-15 and 2025-12-31?”

Opus 4.6 Forecasting Agent: 30% / 35% SOTA Forecasting Agent: 70% Resolution: Yes

Brazil’s circular economy bill (PL 1.874/2022) had been repeatedly scheduled for plenary votes throughout 2025 but never voted on. The forecasting agent gave 30%, anchoring on this pattern of scheduling without action, pending procedural motions, and industry opposition calling the substitute text a “legislative monster.” It cited Finance Minister Haddad’s statement that the bill was a government priority, but treated that as one factor among many. On October 29, the Câmara approved the bill, 12 days before COP30 opened in Belém. Brazil’s Frente Parlamentar Ambientalista had explicitly delivered a document to the Chamber presidency titled “The Urgent Legacy That

the Legislative Branch Needs to Leave for Brazil at COP30,” listing the circular economy bill as a priority.

The Opus 4.6 agent never asked why the Lula government would want this chronically stalled bill passed now. The answer was that Brazil was hosting COP30, the UN climate summit, in November. This transformed the bill from a slow procedural grind into a must-ship showcase item for the host country. The word “COP30” does not appear anywhere in the Opus 4.6 agent’s rationale. The SOTA forecasting agent at 70% prospectively identified COP30 as “an exceptionally strong political catalyst”. The Opus 4.6 agent weighted the September 11 industry complaints heavily despite its own evidence of a September 24 breakthrough agreement meant to resolve those very complaints.

Agent trace analysis of the first Opus 4.6 agent run showed 17 web searches focused on bill numbers and procedural status, but no query contained “COP30,” “climate,” or related. The second Opus 4.6 agent run happened to find COP30 information through a broader query, yet the final synthesis still gave only 35%, heavily discounting these signals in favor of the procedural base rate.

5 Discussion

BTF-2 is the first forecasting benchmark with a large number of difficult, diverse questions, and a hermetic corpus for reproducibility. It can distinguish small accuracy differences, capturing differential strengths between strong frontier agents at statistical significance. The performance of frontier agents on BTF-2 also validate the pastcasting methodology and RetroSearch agent toolkit, achieving higher accuracy than agents that had access to the live web. This reproducibility, along with the detailed agent traces, enable experimentation to develop stronger forecasters, without the hindsight bias that comes from using the question resolutions as feedback.

The stronger forecaster used in this paper has a 0.011 Brier score improvement over the Opus 4.6 agent, the most accurate frontier agent not given forecast-specific guidance. To put this in context, a 0.010 Brier improvement is equivalent to a 50 percentage point advantage on 1 out of every 25 questions, or a 10 percentage point advantage on every question. Though expert human forecasters judge that the Opus 4.6 agent is a great forecaster, this Brier difference points to significant headroom.

Investigations into agent traces and rationales offer insights into this headroom. The CHAMPS KNOW analysis finds that better forecasters are distinguished by epistemics, particularly pre-mortem analysis, correcting for blind spots, and considering black swans. The case studies show that forecasting agents sometimes fail to model the incentives of political and business leaders, treating stated positions as commitments rather than strategic moves. This suggests paths forward to build better agentic forecasters, as well as the methodology to evaluate that progress.

Limitations BTF-2 covers one time period (Oct–Dec 2025), so cannot evaluate models with a more recent training window cutoff. The RetroSearch methodology requires periodically producing new benchmarks.

BTF-2 is weighted toward geopolitics, policy, and macroeconomics. The finding that the most serious failures are in incentive-modeling may partly reflect question composition.

The CHAMPS KNOW analysis is done by a single LLM call on final rationales, not full traces. As the expert human forecast reviews uncovered, evaluating the full series of searches, page reads, and thoughts of each agent can add information that rationales alone lack.

While BTF-2 does enable controlled experiments in reasoning strategies for agents, causally isolating the role of strategic factors is challenging. Prompting agents to pay more or less attention to individual factors does not ensure they actually do this.

We applied expert human review on only 130 questions of 1417, due to costs and time. This covered all the questions with low accuracy scores for Opus 4.6 agents, as well as divergence between the Opus 4.6 agent and the SOTA forecaster. But it neglects questions where the agent was “right for the wrong reason”. A more complete analysis would likely find strategic reasoning failures in questions with high accuracy scores.

Future Work A natural followup to the next version of BTF is to include conditional forecasting questions. These may be more useful in strategic decision making, and also are significantly harder, allowing more exploration of the room between the best agents.

Acknowledgments and Disclosure of Funding

We thank the FutureSearch team for infrastructure support, and the expert forecasters who reviewed case studies.

References

- Rohan Alur, Olivier Boulais, Guillaume Dufort, Tao Gao, Alexander Kuzmin, Marcos Lopez de Prado, and Patrice Simard. AIA forecaster: Technical report. *arXiv preprint arXiv:2511.07678*, 2025.
- Nikos I. Bosse, Peter Mühlbacher, Jack Wildman, Lawrence Phillips, and Dan Schwarz. Automating forecasting question generation and resolution for AI evaluation. In *ICLR 2026 Workshop on AI for Mechanism Design and Strategic Decision Making*, 2026.
- Welton Chang, Eva Chen, Barbara Mellers, and Philip Tetlock. Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11(5):509–526, 2016.
- FutureSearch. Deep research bench. <https://evals.futuresearch.ai>, 2025.
- FutureSearch. FutureSearch agents: ReAct-based web research. <https://futuresearch.ai>, 2026.
- Ezra Karger, Horace Bastani, Yutong Chen, Zachary Jacobs, Thomas Malone, and Philip Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. *arXiv preprint arXiv:2409.09718*, 2025.
- Metaculus. AI forecasting benchmark series. <https://www.metaculus.com/project/ai-benchmarking/>, 2024.
- Long Phan, Adam Khoja, Mantas Mazeika, and Dan Hendrycks. LLMs are superhuman forecasters. 2024. Preprint, September 2024.
- Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.
- Jack Wildman, Nikos I. Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, and Lawrence Phillips. Bench to the future: A pastcasting benchmark for forecasting agents. *arXiv preprint arXiv:2506.21558*, 2025.
- Zelong Yang, Hanchen Liu, Jiaxin Wu, and Haifeng Xu. LLM-as-a-prophet: Understanding predictive intelligence with Prophet Arena. In *International Conference on Learning Representations*, 2026.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*, 2023.
- Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, and Dan Hendrycks. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*, 2022.

A Full Case Study Details

This appendix provides the complete forecasting question, resolution criteria, and a verbatim reconstruction of the Opus 4.6 agent’s research process for each case study discussed in Section 4.5. The traces are reconstructed from observation logs and show every web search query, page read, and intermediate reasoning turn exactly as produced by the agent. The agent’s final rationale appears inline as the last turn of each trace.

A.1 ASUU Strike

Forecasting Question

Between 00:00 WAT Oct 15 and 23:59 WAT Dec 31, 2025, will ASUU declare a nationwide university strike lasting at least 7 consecutive days?

Resolution Criteria

Forecasting target: Resolve YES if there is any single episode, initiated during the window from 00:00 WAT on 2025-10-15 through 23:59 WAT on 2025-12-31 (inclusive), in which ASUU’s national leadership declares a nationwide strike that remains in effect for at least 168 uninterrupted hours (7 consecutive days). Resolve NO otherwise.

Definitions and scope:

- *ASUU:* Academic Staff Union of Universities, Nigeria. “ASUU’s national leadership” means the National Executive Council (NEC) or national officers acting on behalf of the union (e.g., the National President), consistent with reporting on the September 2025 ultimatum.
- *Strike/industrial action:* A formally declared work stoppage by academic staff represented by ASUU. A “strike” includes a “warning strike” or an “indefinite strike” if declared by ASUU’s national leadership.
- *Nationwide:* The declaration is not limited to a specific institution, zone, or subset; it is announced to apply across universities nationally (e.g., described as “nationwide,” “national,” “across universities nationwide”). It must at least cover ASUU chapters in Nigeria’s federal universities. It is not necessary to verify compliance by every campus; the declaration’s stated scope controls.
- *Consecutive days:* Counted as 168 continuous hours from the declared/reported strike start time until the strike’s suspension or expiration is officially announced. Weekends and public holidays count toward the total. If the strike is suspended or called off before 168 hours elapse, the episode does not qualify.
- *Time window:* Only strike episodes that are initiated (i.e., declared to begin) between 00:00 WAT on 2025-10-15 and 23:59 WAT on 2025-12-31 are eligible. A strike that starts before 2025-10-15 but continues into the window does not qualify; a strike that starts within the window and continues beyond 2025-12-31 can still qualify so long as it reaches 168 hours within the window.
- *Exclusions:* Actions limited to a single university/zone; actions solely by other unions (e.g., SSANU/NASU) without an ASUU national strike declaration; isolated boycotts, protests, or sit-ins absent a strike declaration by ASUU national leadership; or “no work” days without an overarching national strike declaration.

Evidence required to resolve:

- *Primary evidence:* An official ASUU national announcement (e.g., press release on its official channels such as website or verified social media) declaring a nationwide strike and stating a start time/date.
- *Corroboration:* Reports by at least two reputable Nigeria-based outlets confirming the nationwide strike and its start date (e.g., Premium Times, Punch, Vanguard, The Nation, The Guardian Nigeria, Channels TV).

Adjudication notes:

- *Start time:* If ASUU’s announcement provides a start time/date, use that; otherwise, use the earliest start time/date reported by at least two independent outlets.
- *End time/duration:* If ASUU announces a suspension/call-off, use that time; otherwise, if at least two independent outlets report the strike remained in effect without suspension for 168 hours from the start, that satisfies duration. Partial exemptions (e.g., essential services) do not break continuity if the strike remains officially in effect.
- *Multiple episodes:* If multiple strike episodes occur, a single qualifying episode suffices for YES.
- *Time zone:* All times are WAT (UTC+01:00).

Agent Research Trace

Turn 1

THOUGHT:

- The forecast must evaluate whether the strike is initiated (begins) specifically between 00:00 WAT on 2025-10-15 and 23:59 WAT on 2025-12-31.
- The strike must be declared by ASUU's national leadership (NEC or national officers), not local chapters or other unions.
- The strike must be nationwide in scope, specifically including Nigeria's federal universities.
- The strike must last for at least 168 consecutive hours (7 full days) to meet the resolution criteria.
- The forecast must account for the exclusion of any strike that starts before 2025-10-15, even if it continues into the window.
- The final forecast must be provided as a probability between 0 and 100.
- The response must include a rationale summarizing research and reasoning for the forecast.

Turn 2

SEARCH: ASUU nationwide strike October 2025 Nigeria

SEARCH: ASUU ultimatum September 2025 warning strike

SEARCH: ASUU strike 2025 latest news Nigeria

Turn 3

SEARCH: ASUU warning strike October 13 2025 end date indefinite strike

SEARCH: ASUU indefinite strike after warning strike October 2025

READ: channelstv.com — *ASUU Declares Two-Week Warning Strike*

READ: guardian.ng — *ASUU confirms talks with FG's team as strike enters seventh day*

READ: vanguardngr.com — *ASUU Strike: Union reports progress in talks with FG*

Turn 4

THOUGHT: As of October 19, 2025, ASUU has reported progress in its negotiations with the Federal Government. The union's National Strike Coordinating Committee (NSCC) confirmed that the government's negotiation team, led by Alhaji Yayale Ahmed, has presented documented responses and offers regarding key issues. Progress has been noted in: release of third-party deductions; payment of promotion arrears; mainstreaming of Earned Academic Allowances (EAA); resolution of issues concerning the confiscation of University of Abuja land; addressing the victimization of members at KSU, LASU, and FUTO.

The document does not explicitly state what will happen immediately after the current warning strike ends, nor does it confirm that an indefinite strike will follow. Engagements are continuing on the draft renegotiated 2009 Agreement. The union urged its members to remain "united and resolute in the few days ahead."

Turn 5

THOUGHT: As of October 19, 2025, ASUU is in the seventh day of a two-week warning strike. Negotiations are ongoing with the Federal Government's Renegotiation Team, led by Alhaji Yayale Ahmed. ASUU reports that "progress is being recorded" on five of its seven-point demands, including the release of third-party deductions, payment of promotion arrears, and the mainstreaming of Earned Academic Allowances (EAA).

The strike is currently in its seventh day. Since the strike began on or around October 13, 2025, it has already met the 168-hour threshold required by the resolution criteria. Because the strike was initiated within the eligible window (October 15–December 31, 2025) and has already reached the 7-day duration requirement, the condition for a "YES" resolution has been met.

Turn 6

THOUGHT: ASUU directed all branches across the country to commence a full withdrawal of services beginning at midnight on Monday, October 13, 2025. The document describes the current action as a "two-week warning strike." While the article notes that the union has previously discussed an "indefinite strike" as a potential escalation, it does not explicitly confirm that an indefinite strike will automatically follow the two-week warning period. The union's decision-making process relies on National Executive Council (NEC) meetings and the government's response to their demands.

Turn 7

THOUGHT: The critical detail here is that the two-week warning strike started on **October 13, 2025**, which is **before** the resolution window of October 15, 2025. The criteria explicitly excludes strikes that start before October 15 even if they continue into the window. So the current warning strike doesn't qualify.

The question becomes: Will ASUU declare a **new** strike (e.g., an indefinite strike) initiated on or after October 15?

SEARCH: ASUU indefinite strike after warning strike ends October 27 2025
SEARCH: ASUU warning strike ends October 26 what next indefinite
SEARCH: ASUU NEC meeting after warning strike escalation indefinite 2025
READ: channelstv.com — *NLC Asks FG To Meet ASUU's Demands Or Risk Strike*
READ: punchng.com — *ASUU rejects FG's proposal, begins two-week strike today*
READ: thenationonlineng.net — *JUST IN: ASUU declares two-week warning strike*

Turn 8

THOUGHT: ASUU commenced a two-week warning strike effective from midnight on Monday, October 13, 2025, following the expiration of a 14-day ultimatum issued on September 28, 2025. The document does not explicitly state that an indefinite strike is guaranteed to follow. However, the union's demands—which include the conclusion of the renegotiated 2009 FGN-ASUU agreement, release of withheld salaries, and sustainable funding—remain unresolved. The union leadership rejected the government's recent proposal, describing it as a “total departure” from previous agreements. The history of ASUU strikes often involves a progression from warning strikes to indefinite industrial action when negotiations fail.

Turn 9

THOUGHT: ASUU has declared a two-week warning strike, which commenced on Monday, October 13, 2025. Regarding the escalation sequence, the ASUU National President, Prof. Chris Piwuna, stated that the union will meet again after the two-week period. If the government fails to address their demands during this window, the union plans to take “action that will certainly be different from this one.” Piwuna explicitly noted that if they embark on the next phase of action, “it will be total and there will be no going back,” implying an escalation to an indefinite strike.

Turn 10

THOUGHT: The Nigeria Labour Congress (NLC) has issued a four-week ultimatum to the Federal Government to meet ASUU's demands. NLC President Joe Ajaero stated that if the government fails to address these issues within this timeframe, the NLC will initiate an indefinite strike in solidarity with educational unions. ASUU has already declared a two-week “total and comprehensive” warning strike, which followed the expiration of a 14-day ultimatum on September 28. The document does not explicitly state what happens immediately after this specific two-week warning strike ends.

Turn 11

READ: dailypost.ng — *Strike: ASUU to hold strategic NEC meeting as action enters week two*
SEARCH: ASUU NEC meeting October 21 2025 indefinite strike decision
SEARCH: ASUU warning strike ends October 27 next steps indefinite action

Turn 12

THOUGHT: The NEC meeting is scheduled to evaluate the union's position on the ongoing strike. ASUU branches held congresses on Monday, October 20, to vote on whether to suspend or continue the current labour action. While the official decision rests with the NEC, reports indicate that many branches have voted in support of continuing the strike. Officials from various universities (University of Abuja, Nnamdi Azikiwe University, and the University of Nigeria, Nsukka) indicated that the matter is “far from being over” and that the congresses were largely a formality to confirm support for the strike's continuation. The union is not “ready to throw in the towel yet.” The NLC has issued a separate four-week ultimatum to the government.

Turn 13

FINAL ANSWER (72%): The key issue is that ASUU declared a two-week warning strike starting October 13, 2025, which is BEFORE the resolution window begins on October 15. The resolution criteria explicitly state: “A strike that starts before 2025-10-15 but continues into the window does not qualify.” Therefore, the current warning strike (Oct 13–27) does NOT count toward resolution. For YES resolution, ASUU must declare a NEW nationwide strike episode that begins on or after October 15, 2025, and lasts at least 168 consecutive hours (7 days). The most likely qualifying scenario is that after the two-week warning strike ends (≈October 27), ASUU's NEC declares an indefinite strike as the next escalation step. Key evidence supporting this:

1. ASUU's planned escalation sequence was explicitly: 14-day ultimatum → 2-week warning strike → indefinite strike. ASUU National President Prof. Chris Piwuna stated the next phase of action “will be total and there will be no going back.”
2. As of October 21, ASUU is holding a strategic NEC meeting to evaluate the strike, with many branches voting to continue/escalate. Officials from several universities indicated the matter is “far from being over.”

3. The NLC has issued a 4-week ultimatum threatening solidarity action if ASUU's demands aren't met.
4. The FG has implemented a "no work, no pay" policy, which could harden both sides' positions.

Evidence that could prevent escalation:

1. ASUU reports progress in negotiations with FG's Renegotiation Team on 5 of 7 demands.
2. The Senate has intervened to mediate.
3. Historical precedent shows some warning strikes end with partial concessions.

However, ASUU's demands are structural and longstanding (dating to 2009), making full resolution within weeks unlikely. The resolution window extends through December 31, 2025, a very long period that provides multiple opportunities for escalation. Even if the current round of negotiations yields temporary calm, the underlying issues could trigger a new strike at any point during this ≈ 10 -week window.

ASUU's historical track record strongly favors escalation: in most recent cycles, warning strikes have been followed by indefinite strikes. The 2022 strike lasted approximately 8 months after similar escalation patterns.

Balancing the strong signals of escalation against the reported negotiation progress and government/Senate intervention, I estimate a probability of approximately 72%.

A.2 Brazilian Circular Economy Bill

Forecasting Question

Will the Câmara dos Deputados Plenary approve the lead proposition containing PL 1.874/2022 (National Circular Economy Policy) between 2025-10-15 and 2025-12-31?

Resolution Criteria

Binary resolution (Yes/No).

Observation window: 2025-10-15 00:00:00 UTC through 2025-12-31 23:59:59 BRT. Brasília Time (BRT) is UTC−03:00; Brazil has no DST in 2025.

What counts as YES: During the observation window, the official Câmara dos Deputados *ficha de tramitação* shows that the Plenário approved the proposition that is the *proposição principal* (lead) for the package that includes PL 1.874/2022 at the time of the vote (currently PL 3.899/2012), with an event explicitly indicating "Aprovado" for the proposition in Plenary. If PL 1.874/2022 is *desmembrada* (de-appended) and itself becomes the item voted in Plenary, a Plenary approval of PL 1.874/2022 within the window also counts as YES.

What does NOT count: Approval of urgency (e.g., REQ 1495/2024) or scheduling alone; approval of a committee opinion without a Plenary approval of the proposition; approval of the *texto-base* alone without subsequent Plenary approval of the proposition.

How to resolve:

1. Identify the then-current *proposição principal* for the package including PL 1.874/2022 by checking its *ficha de tramitação*.
2. Open the *ficha* of that lead proposition (as of 2025-10-01, PL 3.899/2012) and check the 2025 events for a Plenary voting entry indicating "Aprovado" during the observation window.
3. If PL 1.874/2022 itself is voted as principal in the Plenary during the window, check its *ficha* for the Plenary approval event.
4. If the *ficha(s)* explicitly show such a Plenary approval event within the observation window, resolve YES. Otherwise, resolve NO.

Key context: PL 1.874/2022 was presented in the Câmara on 2024-03-21 and is currently *apensada* under PL 3.899/2012. The lead proposition's *ficha* shows a relator designation (Dep. Luciano Vieira, 2025-03-20), urgency regime metadata, and multiple 2025 plenary entries/schedules. As of late September 2025, there is no recorded Plenary approval.

Agent Research Trace

Turn 1

THOUGHT:

- Verify the *proposição principal* (lead proposition) currently associated with PL 1.874/2022 on the official Câmara dos Deputados *ficha de tramitação*.

- Ensure the forecast specifically targets the final Plenary approval (“Aprovado”) rather than the approval of urgency (REQ 1495/2024), scheduling, or the *texto-base*.
- Confirm that the predicted approval event falls within the specific window of October 15, 2025, to December 31, 2025.
- The final response must contain both a reasoning rationale and a numerical probability score between 0 and 100.

Turn 2

READ: camara.leg.br (PL 1.874/2022 *ficha*) — *Portal da Câmara dos Deputados*

READ: camara.leg.br (PL 3.899/2012 *ficha*) — *Portal da Câmara dos Deputados*

SEARCH: PL 1874/2022 ‘economia circular’ Câmara plenário aprovado 2025

SEARCH: PL 3899/2012 plenário aprovado 2025

Turn 3

THOUGHT: The “Situação” field lists PL 1.874/2022 as *apensado* to PL 1755/2022. However, the most recent legislative action (20/03/2025) identifies the bill as *apensado* to PL 3.899/2012. Given that the latter is the most recent update, PL 3.899/2012 is the current lead proposition.

Status: “Aguardando Criação de Comissão Temporária pela MESA; Pronta para Pauta no Plenário (PLEN).”

The bill has not been approved in the Plenary. Active *desapensação* requests (REQ 3873/2025, REQ 4116/2025) are pending. No scheduled Plenary vote recorded.

Turn 4

THOUGHT: PL 3.899/2012 remains active and is currently subject to Plenary appreciation. It has **not** been approved in the Plenary in 2025.

Most recent events: 2025-10-15 PRLP n. 8 PLEN by Dep. Luciano Vieira; 2025-10-08 PRLP n. 7 PLEN and Plenary session where matter not appreciated; 2025-10-02 REQ 4116/2025 (*desapensação*); 2025-09-26 PRLP n. 6 PLEN; 2025-09-24 REQ 3873/2025 (*desapensação*).

Throughout 2025, the proposition has been scheduled for multiple Plenary sessions but in every instance: “Matéria não apreciada em face do encerramento da Sessão” or “Matéria não apreciada em face do encerramento da Ordem do Dia.”

Turn 5

SEARCH: PL 3899/2012 economia circular plenário votação outubro 2025

SEARCH: economia circular câmara plenário votação 2025 aprovado

READ: politicabrasileira.com.br — *Projeto sobre economia circular deve ser pauta de reunião de líderes da Câmara*

READ: blogs.correiobraziliense.com.br — *PL da Economia Circular virou ‘monstro legislativo’, dizem representantes do setor produtivo*

Turn 6

THOUGHT: As of September 11, 2025, PL 3899/2012 is facing significant opposition from the productive sector, which characterizes the current substitute text as a “legislative monster.” Obstacles: the bill has been distorted by the attachment (*apensação*) of dozens of other projects, resulting in a text with over 100 articles; industry argues the text imposes unrealistic obligations (mandatory Circularity Plans, changes to energy matrices) and disproportionate sanctions (fines up to 4% of gross revenue, capped at R\$ 100 million, product seizures, license suspensions); the sector actively supports PL 1874/2022 instead of the current substitute for PL 3899/2012.

Turn 7

THOUGHT: As of September 13, 2025, the project was expected to be brought to a meeting of party leaders by the President of the Chamber. The Chamber had previously scheduled the project for a Plenary vote in the week prior to September 13, 2025, but parliamentarians did not have sufficient time to deliberate on it. There is no specific date provided for a future vote, only that it is considered a priority. Finance Minister Fernando Haddad cited it as a government priority.

Turn 8

SEARCH: PL 3899 economia circular plenário outubro 2025 votação

SEARCH: economia circular câmara deputados votação plenário outubro novembro 2025

SEARCH: câmara deputados pauta plenário semana outubro 2025

Turn 9

SEARCH: PL 3899 economia circular votação plenário câmara outubro novembro 2025

SEARCH: economia circular câmara aprovado plenário site:camara.leg.br 2025

READ: camara.leg.br (PL 3.899/2012 *ficha*, re-read) — *Portal da Câmara dos Deputados*

Turn 10

THOUGHT: October 2025 events on the PL 3.899/2012 *ficha*: 02/10/2025 REQ 4116/2025 (*desapensação*) by Dep. Elcione Barbalho; 07/10/2025 “Matéria não apreciada em face do encerramento da Ordem do Dia”; 08/10/2025 same, plus PRLP n. 7 by Dep. Luciano Vieira; 15/10/2025 PRLP n. 8 by Dep. Luciano Vieira. The matter has **not** been voted or approved in Plenary in October 2025.

Turn 11

SEARCH: PL 3899 economia circular aprovado câmara outubro 2025

SEARCH: câmara deputados pauta plenário economia circular semana

Turn 12

READ: ndmais.com.br — *Sustentabilidade: governo, oposição e setor produtivo fecham acordo ‘histórico’*

SEARCH: economia circular acordo câmara votação plenário site:camara.leg.br OR site:agenciacamara.leg.br 2025

SEARCH: economia circular acordo governo oposição câmara votação outubro 2025

Turn 13

THOUGHT: On Wednesday, September 24, 2025, a “historic” consensus was reached between the government, the opposition, and various sectors (including agribusiness, industry, and environmentalists) to de-append (*desapensar*) PL 1.874/2022 from PL 3.899/2012. The agreement means PL 1.874/2022—which has already been approved by the Senate—will be separated from the older PL 3.899/2012, allowing it to proceed as the primary text for a vote in the Plenary. The decision to de-append rests with the President of the Chamber, Hugo Motta. The expectation is for the matter to be voted under an urgency regime. No specific date for the Plenary vote was provided.

Turn 14

SEARCH: PL 1874 OR PL 3899 economia circular desapensação câmara outubro 2025

SEARCH: economia circular câmara deputados plenário votação semana outubro 2025

READ: camara.leg.br (PL 1.874/2022 *ficha*, re-read) — *Portal da Câmara dos Deputados*

Turn 15

THOUGHT: October 2025: one entry recorded—02/10/2025 REQ 4116/2025 by Dep. Elcione Barbalho requesting *desapensação*. The bill has **not** been de-appended. Status remains “Apensado ao PL 1755/2022.” Two requests pending (REQ 3873/2025 on 24/09/2025 and REQ 4116/2025 on 02/10/2025). No record of Plenary vote. The bill is marked “Pronta para Pauta no Plenário” but the ongoing procedural debate regarding its attachment status makes a Plenary approval uncertain.

Turn 16

SEARCH: economia circular OR PL 3899 OR PL 1874 câmara votação plenário 2025 after:2025-10-15

SEARCH: câmara deputados plenário pauta votação semana 20 outubro 2025

Turn 17

FINAL ANSWER (30%): PL 1.874/2022 (National Circular Economy Policy) is currently *apensada* to PL 3.899/2012, which serves as the lead proposition. As of October 15, 2025, PL 3.899/2012 has NOT been approved in Plenary. Throughout 2025, the bill has been repeatedly scheduled for Plenary sessions but consistently not voted, with entries showing “Matéria não apreciada em face do encerramento da Ordem do Dia.” The rapporteur (Dep. Luciano Vieira) has been actively filing preliminary opinions, up to PRLP 8 as of October 15, 2025.

Several factors support potential approval:

1. The bill has urgency status since November 2024.
2. Finance Minister Haddad cited it as a government priority.
3. A “historic” agreement was reached on September 24, 2025, between government, opposition, and productive sectors to de-append PL 1.874/2022 from PL 3.899/2012, potentially simplifying its path to a vote.
4. End-of-year legislative sessions typically see more bills cleared.

However, significant obstacles remain:

1. The de-appending has NOT yet been effectuated—requests REQ 3873/2025 and REQ 4116/2025 are still pending.
2. The productive sector called the current substitute text a “legislative monster” with over 100 articles and unrealistic obligations.
3. The bill has failed to reach a vote in every Plenary session it was scheduled for in 2025.
4. The procedural path forward remains uncertain due to pending de-appending requests.

While there is political will and approximately 2.5 months remain in the observation window, the repeated failure to vote despite being on the agenda, ongoing procedural uncertainty about de-appending, and substantial industry opposition make approval uncertain but possible given end-of-year legislative dynamics.