BENCHMARKS' FIRST EXAM: 13 LEVELS OF AI BENCHMARKS

Koutian Wu
Benchmark Radar
k@benchmark-radar.org

ABSTRACT

New benchmark papers promise to measure research, self-improvement and discovery, yet no study has placed the stream of new benchmarks on one scale. We place every record that Benchmark Radar collected from July to November 2026 on Benchmarks' First Exam (BFE 1.0), a thirteen-level scale (L0 to L12) ordered by the difficulty of the ability a passing score requires, from retrieval to predicting any state of the world. The census covers 33,735 records: 29,891 papers, repositories and datasets from 41 discovery sources, and 3,844 catalog benchmarks from six leaderboards and registries. A model labeller first decides whether each record introduces an AI benchmark, then assigns a level only when the text supports it. Only 1 of the 4,631 levelled benchmark records (4,525 after exact-identifier deduplication) asks a question whose answer nobody held when it was posed. Three in four levelled new benchmarks (74.0%) still score a single answer to a question, a share that did not fall from August to November. A blind second labeller from a different model family relabelled a random sample of 1,500 records: it agrees on which records are benchmarks ($\kappa = 0.78$) and reproduces the level distribution, although it assigns a level to fewer records. Hard is not the same as unknown: the field keeps building harder exams for answers it already has.

1 INTRODUCTION

A benchmark score means what its tasks require. Morris et al. order AI systems by performance and generality (Morris et al., 2023), and METR measures how long a task an agent can finish (Kwa et al., 2025; METR, 2026). On the benchmark side, the AI4S Benchmark Index Tree orders 361 AI-for-science listings by what was known about the answer when each benchmark was built, and finds that only 32 of them pose a question nobody could answer at the time (AItonomy Foundation, 2026). Because these ladders rank systems or cover one domain, none of them tells a reader where the benchmarks released this month sit across all of AI evaluation.

The gap matters because a benchmark's title can promise more than its tasks test. A benchmark for "AI research" may ask an agent to reproduce a paper whose results the evaluator already holds (Starace et al., 2025; Xu et al., 2026). A benchmark for "self-improvement" may score a fixed engineering task with a known best solution (Wijk et al., 2024; Chan et al., 2024). Such tasks are hard and passing them is informative, but they do not test whether a system can produce an answer that nobody has yet, which is what claims of discovery require.

We separate the two by asking one question of every benchmark: which ability does a passing score require? Benchmarks' First Exam (BFE 1.0) answers it on thirteen levels (Section 2), whose lower levels score a single answer that someone already holds (L0 to L2), then work inside an environment (L3 to L5), then research and evaluation tasks with known outcomes (L6 to L8). Only L9 and above, which cover open problems, facts the world reveals later and chaotic systems, score answers that nobody held when the run started. This thirteen-level test is the first exam a benchmark itself must pass before its scores can support what it claims to measure.

We apply the scale to the full corpus of Benchmark Radar (Wu et al., 2026b) release v0.16.0, which keeps every record its 41 discovery sources and six catalog sources produced between 2026-07-23 and 2026-11-23 (Section 3). The census finds that 74.0% of levelled new benchmarks score one

answer to a question, that this share did not fall over four months, and that 1 of 4,631 levelled benchmark records (4,525 deduplicated by exact identifier) reaches L9 or above (Section 4).

Our contributions are:

- a single scale for AI benchmarks across all domains, with a pass test for each level;
- a census of 33,735 benchmark records, released with every label, its evidence fields and its rationale;
- an agreement check by a blind second labeller, which shows which parts of the census are stable.

2 THE SCALE

BFE 1.0 assigns the hardest ability that a passing score requires (Table 1). A higher level tests a harder ability, with one exception: L9, L10 and L11 test three kinds of unknown answer, and none of them is necessarily harder than another. L12 is the hardest ability; no finite benchmark can test it. A labeller tests the levels from L12 down and assigns the first level whose boundary test passes (Appendix A, Table 3) and whose evidence fields are complete (Table 4). Missing evidence gives `unclassified`, and the record stays in the corpus. A harness, leaderboard or validity study tests no ability of the system under test and receives the type `meta-evaluation` without a level.

The scale merges two earlier orderings. The stage names Retrieval, Replication, Rediscovery and Frontier advancement come from the AI4S Benchmark Index Tree (AItonomy Foundation, 2026). The distinction between rediscovering known theory and discovering new theory follows the Artificial Einstein Test (Keating, 2026). The idea that prediction of unseen data marks the top of the scale follows Pachocki (2026).

3 DATA AND LABELLING

Corpus. Benchmark Radar release v0.16.0 holds two populations with different units, and we keep them apart. The discovery corpus has 29,891 artifact records (papers, repositories, datasets and posts) that 41 discovery sources surfaced between 2026-07-23 and 2026-11-23. The catalog has 3,844 benchmark records from six leaderboards and registries. We label every record and drop none; joins use exact identifiers only. Across the two populations, 770 catalog records share an exact identifier (arXiv ID, DOI, or repository URL) with 769 discovery records; 106 of those pairs receive a level in both populations, agreeing on the exact level in only 21 cases because catalog blurbs are terser than discovery abstracts. We therefore keep the two populations side by side in the main comparison and also report the exact-identifier deduplicated merged total (4,525 levelled records, preferring discovery abstracts on overlap).

Labelling. A model labeller (`gpt-6-luna`, called through the Manus API with one fixed prompt) reads each record's title and abstract or description and answers two questions. First, does the record introduce a benchmark, evaluation suite or test set for AI systems? Second, for records that do, which level does a passing score require? It also fills the scale's evidence fields (Appendix A, Table 4): the scored outcome, the target the agent sees, what the evaluator knew before the run, and the verifier. L9 to L11 additionally require a run-start cutoff and a novelty check or resolution event. Abstracts cover 87% of discovery records; the rest are labelled from titles, and a level then rarely passes its evidence test. After the run we read every record labelled L9 or above and corrected those whose answers the evaluator already held.

Agreement check. A second labeller from a different model family (Gemini, run in Antigravity with the same prompt and rubric) labelled 1,500 records drawn at random with a fixed seed, 1,200 discovery records and 300 catalog records, without access to the first labels. It opened the source of every record it placed at L6 or above. A third model (Claude) relabelled a further 100 records, 40 of them drawn from records the first labeller had levelled. All labellers are models, so agreement shows stability, not correctness.

Table 1: The BFE 1.0 scale (L0 to L12) and the number of levelled benchmarks at each level. Discovery records are new papers, repositories and datasets; catalog records come from leaderboards and registries. Examples are from the scale definition.

Level	Name	A passing score requires	Discovery	Catalog
Band 1: Single known answer (L0–L2)				
L0	Retrieval	returning information that exists in a source	223	116
L1	Quiz questions	a short fixed answer a non-specialist can derive (MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021))	1,373	1,735
L2	Expert-level questions	a fixed answer that needs expertise (HLE (Phan et al., 2025), FrontierMath (Glazer et al., 2024))	295	25
Band 2: Work in an environment (L3–L5)				
L3	Real coding work	code changes in a real project, judged by tests (SWE-bench (Jimenez et al., 2023))	229	51
L4	Long agent tasks	many actions toward one goal over a long horizon (OSWorld (Xie et al., 2024))	152	103
L5	Whole jobs	a complete professional deliverable (GDPval (Patwardhan et al., 2025))	91	7
Band 3: Known-outcome research and evaluation (L6–L8)				
L6	Testing the tests	judging whether a test, grader or result is valid	64	18
L7	Reproducing research	a research result the evaluator holds (PaperBench (Starace et al., 2025))	33	8
L8	Self-improvement	a measured improvement to an AI system (RE-Bench (Wijk et al., 2024))	94	13
Band 4: Unknown answers (L9–L12)				
L9	Open problems	a result nobody had when the benchmark was built	0	0
L10	Hidden facts	a prediction scored when the world reveals it (ForecastBench (Karger et al., 2024))	1	0
L11	Chaotic systems	a prediction of a chaotic system, scored on later observations	0	0
L12	Laplace's demon	a prediction of any state of the world	0	0
Levelled benchmark records			2,555	2,076

4 RESULTS

Most records are not benchmarks. The labeller marks 10,271 of 29,891 discovery records (34%) as new AI benchmarks; 17,995 are method papers, clinical or social-science evaluations, unrelated datasets or personal repositories, and 1,625 are unclear. The second labeller agrees with this decision on 88% of its 1,500 sampled records (Cohen's $\kappa = 0.78$; $\kappa = 0.71$ on discovery records). On catalog records the two agree on 98%, where nearly every record is a benchmark and κ is uninformative.

Few benchmarks disclose enough to be levelled. Only 2,555 of 10,271 new benchmarks (25%) state their task, verifier and answer status clearly enough for a level; 610 are meta-evaluations and 7,106 remain unclassified. The catalog fares better, with 2,076 of 3,838 levelled (54%). The distributions below describe the levelled benchmarks.

Benchmarks still score answers someone holds. Figure 1 shows where the 4,631 levelled benchmark records sit. Among new benchmarks, 74.0% score a single answer to a question (L0 to L2), 18.5% score work in an environment (L3 to L5), and 7.5% fall at L6 to L8. In the catalog these shares are 90.4%, 7.8% and 1.9%. Merging the two populations by exact identifier (4,525 deduplicated levelled records, preferring discovery abstracts on the 106 overlapping levelled pairs) yields

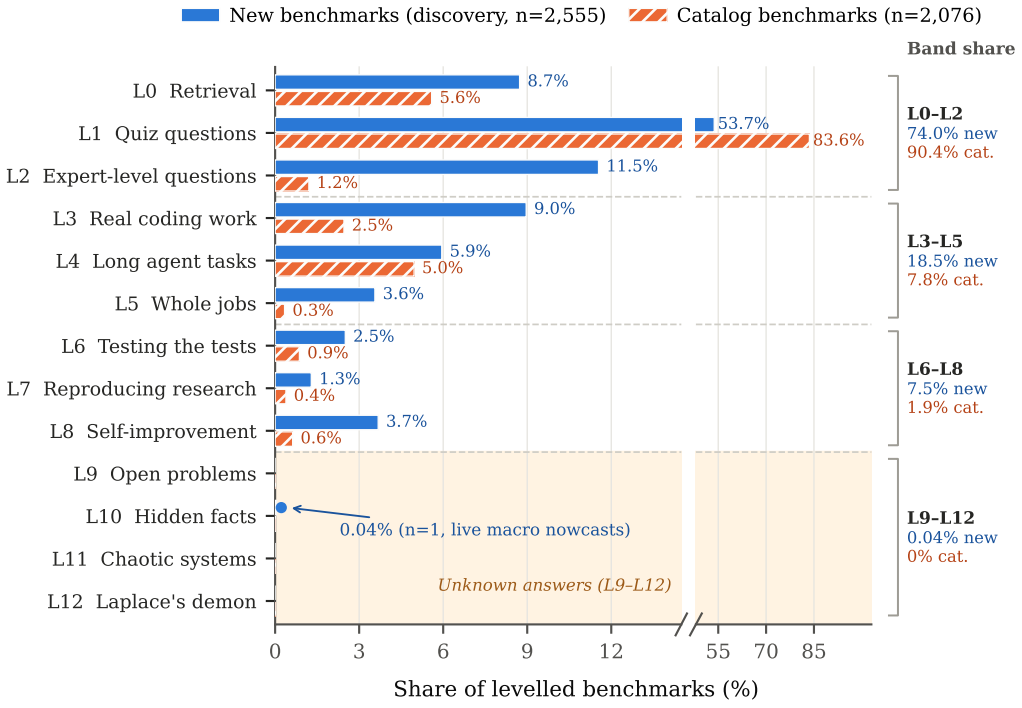

Figure 1: Most levelled benchmarks score one answer to a question. Share of levelled benchmarks at each BFE 1.0 level, for new benchmarks found from July to November 2026 and for catalog benchmarks. The x-axis breaks between 15% and 48% so levels L0 and L2–L12 remain readable alongside L1; right-hand annotations report band shares. Unclassified records and meta-evaluations are excluded from the shares and reported in the text.

Table 2: Level bands of levelled new benchmarks by month of discovery, restricted to the 11 sources present in every month. November ends on 2026-11-23.

Month	Records	Benchmarks	Levelled	L0–L2 (%)	L3–L5 (%)	L6–L8 (%)
August	4,217	1,762	336	69	24	7
September	6,938	2,292	492	63	28	8
October	8,174	2,628	606	70	21	9
November	5,337	1,738	425	77	15	9

81.1% at L0 to L2, 13.8% at L3 to L5, 5.1% at L6 to L8, and 1 at L9 or above. Exactly 1 benchmark reaches L9 or above: a live evaluation of macroeconomic nowcasts, scored when each official figure is released (Zhao et al., 2026). New benchmarks therefore reach beyond the leaderboard canon toward agentic work, but not toward unknown answers.

The share did not move in four months. Holding the 11 discovery sources that report in every month fixed, the L0 to L2 share of levelled new benchmarks was 69% in August, 63% in September, 70% in October and 77% in November (Table 2). The L6 to L8 share stayed under ten percent throughout.

Where the labels are stable. The labellers differ most on whether a record discloses enough for a level. On the same 1,500 records, the second labeller levels 87 of its 603 benchmarks, while the first levels 232 of 664; on catalog records the gap is 28 against 146. Exact levels on the 570 records both call benchmarks therefore agree poorly ($\kappa = 0.19$), mostly because one labeller says `unclassified`. Where both assign a level, the level band agrees on 87% of 55 records ($\kappa = 0.66$), and the distributions match: the second labeller puts 76% of its levelled benchmarks at L0 to

L2 against 82% for the first on the same records, and places 0 at L9 or above. The Claude check points the same way: every L7 and L8 label in its sample moved down to L2 or L4. The share of levelled records and the counts at L6 to L8 therefore depend on the labeller, while the dominance of L0 to L2 and the absence of L9 and above do not.

5 Related work

Ordered scales for AI mostly rank systems: by performance and generality (Morris et al., 2023), by the autonomy of research agents (Ding et al., 2026), or by autonomy in a domain such as medicine (Ye et al., 2025). METR places models on a continuous scale of task length using its own task suites (Kwa et al., 2025). The AI4S Benchmark Index Tree is the closest work: it orders benchmarks by answer status and pairs each stage with a verifier type, over 361 listings in AI for science (AItonomy Foundation, 2026). We extend that ordering to all of AI evaluation, add the task scope that separates quizzes from whole jobs, and measure a full multi-source corpus. Recent benchmarks for scientific agents score project-level research while progressively withdrawing methodological guidance (Zhou et al., 2026), or commit-level changes to Earth system model code (Wu et al., 2026a), and ScienceIDE turns scientific repositories into executable environments for training and evaluating agents (Geng et al., 2026). Studies of benchmark validity (Desai et al., 2026) and saturation (Akhtar et al., 2026) ask whether a benchmark measures what it claims; ESM-Bench shows the stakes, since three of its text-pattern checkers passed patches that did not apply to their starting files (Wu et al., 2026a). We ask which ability a benchmark can claim at all.

6 Limitations

The labels come from a model reading titles and abstracts, and two other models checked samples; no human has audited them. Three quarters of new benchmarks remain unclassified because their abstracts do not state the scored outcome or answer status, the levelled quarter may differ from them, and how many records receive a level depends on the labeller. L6 to L8 labels are the least stable. The catalog shares rest on the first labeller alone, because the second levelled too few catalog records to test them. The corpus reflects what Benchmark Radar's sources surface over four months; benchmarks that appear only in blogs or private evaluations are missing. The author co-authored ResearchClawBench (Xu et al., 2026), ASI-Bench (Zhou et al., 2026), ScienceIDE (Geng et al., 2026) and ESM-Bench (Wu et al., 2026a). The first three are in the corpus and are labelled from their own text like every other record; ESM-Bench was posted on 2026-11-21 but is not among the records Benchmark Radar collected, so it carries no label.

7 Conclusion

Across 33,735 records, the benchmarks released in late 2026 are harder than their predecessors in scope, but they still score answers that someone already holds. Benchmarks that wait for the world to reveal the answer remain almost absent. We release the scale, every label with its evidence and rationale, and the code to rebuild each number from Benchmark Radar v0.16.0.

References

AItonomy Foundation. The AI4S benchmark index tree. `https://aitonomy.org/projects/benchmark`, 2026.

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, et al. When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation. *arXiv preprint arXiv:2602.16763*, 2026. URL `https://arxiv.org/abs/2602.16763`.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. *arXiv preprint arXiv:2410.07095*, 2024. URL `https://arxiv.org/abs/2410.07095`.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Meera Desai, Sang T. Truong, Hanna Wallach, Alex Chouldechova, A. Feder Cooper, Jean Garcia-Gathright, Daniel E. Ho, Abigail Z. Jacobs, Sanmi Koyejo, Nicholas Pangakis, et al. What AI Benchmarks Actually Measure: Adapting Convergent and Discriminant Validity to Interrogate Fifty-Six AI Benchmarks. *arXiv preprint arXiv:2609.08812*, 2026. URL https://arxiv.org/abs/2609.08812.

Tianyu Ding, Aditya Nannapaneni, Bingfan Liu, and Ling Zhang. Autonomous Research Agents: A Survey of AI Scientists and the Verification Gap. *arXiv preprint arXiv:2608.05179*, 2026. URL https://arxiv.org/abs/2608.05179.

Hejia Geng, Zesen Huang, Haoyang Li, Wenbin Li, Koutian Wu, Zihan Zhou, Yuanbo Pang, Weihao Liu, Zigong Xu, Zhiping Li, et al. ScienceIDE: Turning World's Scientific Codebase into Agent Learnable Environments. *arXiv preprint arXiv:2609.19134*, 2026. URL https://arxiv.org/abs/2609.19134.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024. URL https://arxiv.org/abs/2411.04872.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020. URL https://arxiv.org/abs/2009.03300.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *arXiv preprint arXiv:2310.06770*, 2023. URL https://arxiv.org/abs/2310.06770.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. *arXiv preprint arXiv:2409.19839*, 2024. URL https://arxiv.org/abs/2409.19839.

Brian Keating. The artificial einstein test. https://keating.ai/einstein-test/, 2026.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, et al. Measuring AI Ability to Complete Long Software Tasks. *arXiv preprint arXiv:2503.14499*, 2025. URL https://arxiv.org/abs/2503.14499.

METR. Q4 2026 capability update. https://metr.org/blog/2026-11-20-q4-2026-capability-update/, 2026.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of AGI for Operationalizing Progress on the Path to AGI. *arXiv preprint arXiv:2311.02462*, 2023. URL https://arxiv.org/abs/2311.02462.

Jakub Pachocki. Perplexity and intelligence. https://openai.com/index/perplexity-and-intelligence/, 2026.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, et al. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. *arXiv preprint arXiv:2510.04374*, 2025. URL https://arxiv.org/abs/2510.04374.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*, 2025. URL https://arxiv.org/abs/2501.14249.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. PaperBench: Evaluating AI's Ability to Replicate AI Research. *arXiv preprint arXiv:2504.01848*, 2025. URL https://arxiv.org/abs/2504.01848.

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, et al. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024. URL https://arxiv.org/abs/2411.15114.

Koutian Wu, Yuxuan Cao, Gengchen Mai, and Zeping Liu. ESM-Bench: A Benchmark for Evaluating Whether AI Agents Understand Earth System Model Physics and Code. *arXiv preprint arXiv:2611.28477*, 2026a. URL https://arxiv.org/abs/2611.28477.

Koutian Wu, Wenjun Zhang, Junjie Zhou, Pengqian Han, Qiming Bao, Ergan Shang, Jiayu Wang, Junkai Wang, Wanghan Xu, Songyuanyi Lu, et al. Benchmark Radar: A Living Database and Search Engine for AI Benchmarks and Evaluation. *arXiv preprint arXiv:2609.11115*, 2026b. URL https://arxiv.org/abs/2609.11115.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. *arXiv preprint arXiv:2404.07972*, 2024. URL https://arxiv.org/abs/2404.07972.

Wanghan Xu, Shuo Li, Tianlin Ye, Qinglong Cao, Yixin Chen, Hengjian Gao, Yiheng Wang, Qi Li, Kun Li, Sheng Xu, et al. ResearchClawBench: A Benchmark for End-to-End Autonomous Scientific Research. *arXiv preprint arXiv:2606.07591*, 2026. URL https://arxiv.org/abs/2606.07591.

Xiao Ye, Jacob Dineen, Zhaonan Li, Zhikun Xu, Weiyu Chen, Shijie Lu, Yuxi Huang, Ming Shen, Phu Tran, Ji-Eun Irene Yum, et al. Evaluating Medical LLMs by Levels of Autonomy: A Survey Moving from Benchmarks to Applications. *arXiv preprint arXiv:2510.17764*, 2025. URL https://arxiv.org/abs/2510.17764.

Xinyue Zhao, Ruiyi Zhang, Liqin Ye, Rui Cao, Pengtao Xie, and Sudheer Chava. Can LLMs Take the Pulse of the Economy? A Real-Time Evaluation of LLM Nowcasts on Macroeconomic Indicators. *arXiv preprint arXiv:2608.30110*, 2026. URL https://arxiv.org/abs/2608.30110.

Junwei Zhou, Zhen Sun, Binyu Li, Jiangyu Zhou, Yuexi Pan, Hengyu Wang, Honghe Ren, Xiaohan Jia, Xueyang Zhou, Xiaoyu Cao, et al. ASI-Bench: At the Dawn of Artificial Superintelligence. *arXiv preprint arXiv:2608.17271*, 2026. URL https://arxiv.org/abs/2608.17271.

A BFE 1.0 RUBRIC: BOUNDARY TESTS AND EVIDENCE SCHEMA

To make the scale self-contained, this appendix reproduces the boundary tests (Table 3), required evidence fields (Table 4), and task card schema (Figure 2) of the BFE 1.0 rubric. A labeller reads the task instruction and verifier, identifies the requirement that decides a passing score, and tests levels from L12 down to L0, assigning the first level whose boundary test passes and whose required evidence fields are complete. Mixed suites with separate tracks receive separate track levels rather than an average; harnesses, leaderboards, and validity studies that test no ability of the system under test receive `meta-evaluation` without a level (L6 applies only when the system itself must judge validity).

Table 3: BFE 1.0 boundary tests. Levels are tested from L12 down to L0; a record receives the highest level whose boundary test passes and whose evidence fields are complete.

Boundary	The higher level passes when
L0 → L1	The answer is derived from information, not copied from a source. Search that returns a copied answer stays at L0.
L1 → L2	The items need domain expertise, and experts wrote or checked them to be hard. Both levels score only a returned answer.
L2 → L3	Running tests on code the system changed inside a real or realistic project decides the score. A standalone contest problem stays at L1 or L2.
L3 → L4	The task takes many tool or environment actions toward one goal, over a long human-time horizon. The first labelling pilot sets the threshold. Read-only browsing scored by its final answer stays at L1 or L2.
L4 → L5	The task comes from real paid work, and the complete deliverable is judged at professional quality.
L5 → L6	The score depends on the system correctly judging a test, a grader, or a claimed result as valid or invalid.
L6 → L7	The score is a research result that the evaluator already holds. Replication names the target; rediscovery hides it. A research task may state its question and still be L7. A closed question with one fixed answer stays at L1 or L2.
L7 → L8	The score measures an improvement to an AI system over a stated baseline, such as a faster kernel or a lower training loss. Reproducing an ML paper without such an improvement stays at L7.
L8 → L9, L10, L11	Nobody holds the answer at the run-start cutoff. L9 scores a result the system creates and needs a novelty check. L10 scores a prediction the world later resolves. L11 scores a prediction of a chaotic system against observations made after the prediction. Forecasts scored against data that existed at the cutoff, as WeatherBench 2 does for 2020, stay below L9. A prior result found before the run caps the task at L7 or below.
L11 → L12	The task predicts any state of the world. No finite benchmark passes.

Table 4: BFE 1.0 evidence fields. Every levelled assignment records fields 1–6; L9, L10, and L11 additionally require fields 7–8, without which the record stays `unclassified`.

#	Field	Required for	Definition
1	Scored outcome	L0–L12	The state or result that decides pass or fail.
2	Agent-visible target	L0–L12	The goal, finding, or protocol given to the agent.
3	Evaluator knowledge	L0–L12	What the evaluator knew before the run.
4	Verifier modality	L0–L12	Exact, executable, simulation, model judge, human expert, or hybrid.
5	Verifier procedure	L0–L12	The checks, equivalence criteria, and evidence link.
6	Level rationale	L0–L12	One sentence naming the requirement that sets the level.
7	Evaluation cutoff	L9–L11	The run-start timestamp showing the answer was unknown.
8	Novelty check / resolution	L9–L11	For L9: prior-art scope, sources, search dates, and domain-expert attestation. For L10–L11: the event that revealed the answer and its date.

```
TASK: <id>
BFE VERSION: 1.0
LEVEL: L0 / L1 / ... / L12 / meta-evaluation / unclassified

SCORED OUTCOME: <what decides pass or fail>
AGENT-VISIBLE TARGET: <what the agent receives>
EVALUATOR KNOWLEDGE: <what was known before the run>
EVALUATION CUTOFF: <run-start timestamp; required for L9 to L11>
NOVELTY CHECK OR RESOLUTION EVENT: <required for L9 to L11>
VERIFIER MODALITY: <exact, executable, simulation, model judge,
                    human expert, or hybrid>
VERIFIER PROCEDURE: <checks, equivalence criteria, and evidence link>
LEVEL RATIONALE: <one sentence>

TAGS
  horizon:     single-turn | multi-turn | cross-session |
               streaming | lifelong
  state:       stateless | workspace | memory |
               persistent learning state
  interaction: text | tools | browser | desktop | code |
               embodied | multi-agent
  evaluation:  exact | executable | simulation | model judge |
               human expert
  risk:        privacy | security | policy compliance |
               destructive action
  resources:   latency | cost | compute | memory | network

SLIM ASSESSMENT: <link or pending>
CLASSIFIED BY: <name>
DATE: <date>
```

Figure 2: BFE 1.0 task card schema and descriptive tags. Tags describe a benchmark without setting its level, except that `horizon` informs the L3 $\rightarrow$ L4 boundary.