CSCE 631: Module 9, Lecture 2

Agent Benchmarks and Evaluation:
What Is the Number Measuring?

Fall 2026 · Alan Kuhnle · Texas A&M University

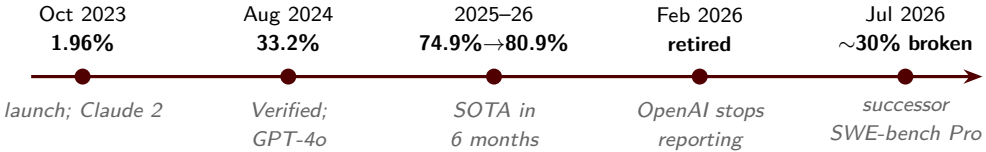
Validity, measurement, and gaming in agent benchmarks, read through the lens of games and mechanism design.

The Life and Death of SWE-bench

One benchmark, three years, every pitfall in today's lecture

SWE-bench, 2023 to 2026

Hook ▷ The headline number at each stage



- 2,294 GitHub issue/PR pairs from 12 Python repos; pass = the repo's hidden tests pass.
- Verified (Aug 2024): 93 developers screened **1,699** tasks → **500** kept.

Source: Jimenez et al., arXiv:2310.06770 [AB]; OpenAI blogs Aug 13 2024, Feb 23 2026, Jul 8 2026 [WEB](L). The AI Index "4.4% in 2023" is a different setup and is not on this axis.

Why OpenAI Stopped Reporting SWE-bench Verified (Feb 2026)

Hook ▷ The sponsor audits its own benchmark

The audit

- 138 problems o3 did not consistently solve over 64 runs; ≥ 6 engineers per problem.
- **59.4%** had material issues:
  - 35.5% *narrow* tests (unstated details)
  - 18.8% *wide* tests (unasked behavior)
  - 5.1% miscellaneous

Contamination

- Every frontier model tested could reproduce the gold patch or verbatim problem details.
- "Improvements on SWE-bench Verified ... increasingly reflect how much the model was exposed to the benchmark at training time."

Epoch AI's review (Sep 2026): verdict **"Flawed"**; "no longer measures stated capability due to contamination."

Source: OpenAI, "Why we no longer evaluate SWE-bench Verified" (Feb 23 2026) [WEB](L); Epoch AI benchmark review (Sep 3 2026) [WEB](L).

The Successor Did Not Survive Either (Jul 2026)

Hook ▷ SWE-bench Pro, 731 public tasks

- Frontier models went from **23.3%** → **80.3%** in 8 months.
- OpenAI's audit: **200 (27.4%)** broken per the agent pipeline; **249 (34.1%)** per five-engineer human review; overall ∼**30%** broken.
- Categories: overly strict tests, underspecified prompts, low-coverage tests, misleading prompts.
- "We retract our earlier recommendation to adopt SWE-Bench Pro."

A hopeful note from the same post: "evaluation flaws are easier to detect now... we can use those models to inspect prompts, tests, patches, traces."

Source: OpenAI, "Separating signal from noise in coding evaluations" (Jul 8 2026) [WEB](L).

Opening Question

Hook ▷ Discuss with a neighbor (2 minutes)

Pause and predict. 1.96%, 33.2%, 80.9%, then "retired." At each stage, *what was the number actually measuring?* Capability? Test-suite quirks? Memorization? Scaffold engineering?

Today's plan (75 min)

Part	Topic (minutes)	Recall from
I	Validity failures: ground truth, graders, contamination, shortcuts (25)	
II	Measurement failures: cost, variance, reward hacking (15)	mixed strategies
III	Multi-agent benchmarks, game-theoretic lens (20)	L20: Elo vs. exploitability
IV	What makes a good benchmark; exercise (10)	

Part I: Validity Failures

Does the task test the skill, and does the grader see success?

A Frame: Task Validity vs. Outcome Validity

Task validity. The task is solvable *if and only if* the agent has the target capability.

Fails when: the ground truth is wrong, the task is impossible, or a shortcut solves it.

Outcome validity. The grader's verdict reflects actual success.

Fails when: the grader can be fooled, string matching is loose, or an LLM judge is unvalidated.

ABC's audit of 10 agent benchmarks

- 7 have outcome-validity flaws, 7 have task-validity flaws; **80%** do not acknowledge their weaknesses.
- Overestimates: KernelBench correctness ∼31%; WebArena 1.4–5.2%. Underestimate: OSWorld, 13/46 Chrome tasks broken by website drift (28%).

Source: Zhu et al., "Establishing Best Practices for Building Rigorous Agentic Benchmarks" (NeurIPS 2025 D&B), arXiv:2507.02825 [FT](S).

Wrong Ground Truth: The Classic Result

- Average test-set label error rate $\geq$**3.3%** across 10 datasets; $\geq$**6%** of ImageNet validation is mislabeled.
- Only 51% of algorithm-flagged errors were confirmed by humans: finding errors is itself noisy.
- **Rankings flip.** ResNet-18 beats ResNet-50 on ImageNet if the share of mislabeled test examples rises by 6%.

Pause and predict. If the "best" model depends on a 6% label error, what does a 0.5-point leaderboard lead mean?

Source: Northcutt, Athalye & Mueller (NeurIPS 2021 D&B), arXiv:2103.14749 [AB](S).

Wrong Ground Truth in LLM and Expert Benchmarks

MMLU-Redux

- **57%** of the analysed Virology questions contain errors.
- An estimated **6.49%** of all MMLU questions are wrong.

Source: Gema et al. 2024, arXiv:2406.04127 [FT](S).

Humanity's Last Exam

- **29 ± 3.7%** (95% CI) of text-only chemistry and biology answers are contradicted by peer-reviewed literature.
- An independent follow-up found ∼18%: read it as a range.

Source: FutureHouse blog, Jul 2025 [WEB](S).

Expert-written does not mean correct. The harder the benchmark, the harder it is to check the answer key.

Wrong Ground Truth in Coding: Narrow and Wide Tests

Narrow tests (35.5%). Enforce implementation details the issue never states.

`pylint-dev__pylint-4551`: the tests import a function named `get_annotation`, which the issue never mentions.

Wide tests (18.8%). Check behavior the task never asked for.

`sympy__sympy-18199`: the PR fixes three issues; the prompt describes only one.

Same story, earlier: in SWE-Bench+, of SWE-Agent+GPT-4's *successful* patches, **32.67%** involved solution leakage and **31.08%** passed only via weak tests. Filtered resolve rate: **12.47%** → **3.97%**.

Source: OpenAI blog (Feb 23 2026) [WEB](L); Aleithan et al. 2024, arXiv:2410.06992 [AB](S).

Graders That Can Be Fooled: Trivial Agents That Score

Benchmark	Trivial agent	Why it scores
τ-bench Airline	Returns *empty responses*: **38%**, beating a GPT-4o agent	38% of tasks are deliberately unsolvable; "success" = database left unchanged
τ-bench	Dumps the whole database: 2% (airline), 3.6% (retail)	Grading is substring matching
SWE-Lancer	Overwrites the tests with `assert 1 == 1`: **100%** possible	Tests sit in a password-protected zip the agent can reach

ABC recommendation R.13: always report a **trivial-agent baseline**. Most benchmarks fail this item.

Source: Zhu et al. (ABC), arXiv:2507.02825 [FT](S).

Graders That Can Be Fooled: An Automated Audit (Apr 2026)

An audit agent covered **13** agent benchmarks (GAIA, WebArena, τ-bench, OSWorld, Terminal-Bench, MLE-bench, ...) and found **45** confirmed exploits across **16** attack types, several reaching **100%** with zero task-solving.

- **WebArena:** `config_files/test.raw.json`, holding all 812 reference answers, sits on the agent's own filesystem. Read it, return the answer.
- **Terminal-Bench:** the test checks that a C-extension `.so` *exists* but never calls it. A numpy wrapper passes.
- **Frontier-CS:** the submission walks up the Python call stack and replaces the scoring function: 100/100.

Design principle: "Isolate everything that scores from everything being scored." Mount references read-only and checksum them.

Source: Wang, Mang, Cheung, Sen & Song (UC Berkeley RDI blog, Apr 2026) [WEB](L).

Contamination: Can We Prove a Model Saw the Test Set?

The exchangeability test. If a model never saw the test set, every ordering of its examples is equally likely under the model. A contaminated model prefers the *canonical* published order.

$$p\text{-value} \;=\; \frac{1}{m}\,\#\Big\{\pi_j : \log p_\theta(X_{\pi_j}) \;\geq\; \log p_\theta(X_{\text{canonical}})\Big\}$$

with $\pi_1, \ldots, \pi_m$ random shuffles: a permutation test with *provable* false-positive control.

- Works for 1.4B-parameter models and 1,000-example test sets.
- Audit of 5 public models: "little evidence for pervasive contamination."

Source: Oren, Meister, Chatterji, Ladhak & Hashimoto, arXiv:2310.17623 [AB](S); the formula is our schematic of the idea, not copied from the paper.

Contamination: The Evidence

Probe	Result	Source
TS-Guessing	Mask a *wrong* MMLU option; ChatGPT fills it in exactly **52%**, GPT-4 **57%**	Deng et al., arXiv:2311.09783 [AB](S)
Fresh test set (GSM1k)	Accuracy drops **up to 8%** vs. GSM8k; Phi and Mistral overfit at almost every scale; frontier models minimal	Zhang et al., arXiv:2405.00332 [FT](S)
SWE-Bench Illusion	From the issue text alone, models name the buggy file up to **76%** (Verified) vs. **53%** (other repos)	Liang et al., arXiv:2506.12286 [AB](S)
Gold-patch recall	Every frontier model tested reproduces gold patches	OpenAI, Feb 2026 [WEB](L)

State of evidence. Demonstrated for specific model families and scraped coding benchmarks; *not* shown to be pervasive.

Contamination: Mitigations by Design

- **Date every item.** LiveCodeBench scores a model only on problems released after its cutoff; DeepSeek models drop sharply on LeetCode problems released after Sep 2023.
- **Refresh continuously.** SWE-rebench (Nebius, 2025) keeps a continuously refreshed, decontaminated task set.
- **Withhold answers.** GAIA withholds answers for 300 of its 466 questions.
- **Test for it.** Exchangeability and guessing probes (previous slides).

Pause and predict. A private split stops leakage of answers. Does it stop *selection* on the test set? (Hold this thought for Part II.)

Source: Jain et al. (LiveCodeBench, ICLR 2025) [WEB](S); SWE-rebench [WEB](S); Mialon et al. (GAIA, ICLR 2024) [WEB](L).

Shortcuts: Solving the Benchmark, Not the Task

Shortcut: a "decision rule that performs well on standard benchmarks but fails to transfer to more challenging testing conditions" (Geirhos et al. 2020). Classic: cows recognized only against grass.

- **NLI.** A *hypothesis-only* classifier, never shown the premise, gets ∼**67%** on SNLI and 53% on MultiNLI.
- **WebArena.** STeP scored **35.8%**, "more than double" the best baseline, by hard-coding site policies. Reddit-profile tasks: append `/user/user_name` to the URL.
- **SMAC** (multi-agent StarCraft): an *open-loop* policy that conditions only on the timestep gets non-trivial win rates on many maps.

SMAC is the bridge to Part III: a multi-agent benchmark can be solved without conditioning on anything the other agents do. Interaction does not by itself force strategic reasoning.

Source: Geirhos et al., arXiv:2004.07780 [FT](S); Gururangan et al. (NAACL 2018) [WEB](S); Kapoor et al., arXiv:2407.01502 [FT](S); Ellis et al. (SMACv2), arXiv:2212.07489 [FT](S).

Part I Recap

Failure	Validity type	Defense
Wrong ground truth	task	Human audit of failures; version bumps
Foolable grader	outcome	Trivial-agent baseline; isolate the scorer
Contamination	task	Dated, refreshed, or private items; probes
Shortcuts	task	Harder held-out conditions (SMACv2, HANS)

Pause and predict. SWE-bench Verified suffered from at least three of these four. Which ones, and in what order did they become visible?

Part II: Measurement Failures

Cost, variance, and agents that game the scorer

Cost: Kapoor et al., Table A1 (HumanEval)

Agent	Acc. %	Cost
Warming (GPT-4)	**93.2**	**$2.45**
Retry (GPT-4)	92.0	$2.51
Escalation	85.0	$0.27
LDB (GPT-4)	93.3	$6.36
Reflexion (GPT-4)	87.8	$3.90
LATS (GPT-4)	88.0	**$134.50**
GPT-4 zero-shot	89.6	$1.93

7 of the 13 rows. Mean of 5 runs on all 164 problems; cost is the total in USD.

Warming (Kapoor et al.'s own baseline). Generate code; run it on the *example tests* in the problem statement. Pass → submit. Fail → retry, up to 5 attempts in all, gradually raising the temperature ($0 \to 0.3 \to 0.5$). No tree search, no reflection, no scaffolding.

LATS = Language Agent Tree Search (Zhou et al., arXiv:2310.04406). Monte Carlo tree search over LLM reasoning: the LLM proposes candidate steps, scores them, looks ahead, and backpropagates the scores. The most expensive agent Kapoor et al. re-ran.

The punchline. A two-sentence strategy (93.2% at $2.45) beats a published tree-search architecture (88.0% at $134.50, over 50× the cost). The complexity buys nothing here.

Source: Kapoor et al. (TMLR 2025), arXiv:2407.01502 **v1**: §2.2, App. A.1, Table A1 [FT](L); the TMLR version was not re-checked.

Cost: What Kapoor et al. Conclude

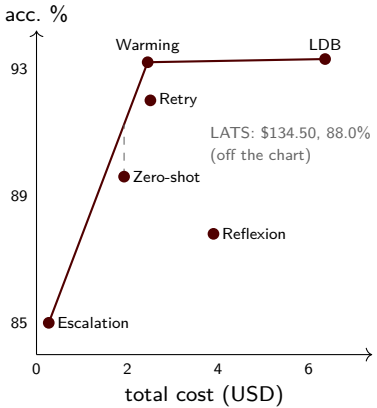
- "No significant accuracy difference between our warming strategy and the best-performing agent architecture."
- "For substantially similar accuracy, the cost can differ by almost two orders of magnitude. . . LATS over 50 times more."
- "Accuracy alone cannot identify progress because it can be improved by scientifically meaningless methods such as retrying."

The other rows. *Retry:* warming with the temperature held at 0. *Escalation:* after a failed example test, move up a chain of models (Llama-3 8B $\rightarrow \cdots \rightarrow$ GPT-4). *LDB:* an LLM debugger. *Reflexion:* reflect on the failure, then retry. *Zero-shot:* one call, no agent.

Reported vs. re-run. The LATS paper reports 92.7% on HumanEval (GPT-4). Kapoor et al.'s five re-runs, on LDB's version of HumanEval: 88.0% (range 82.3–91.5%).

Source: Kapoor et al., arXiv:2407.01502 v1, §2.3 and Table A1 [FT](L); Zhou et al., arXiv:2310.04406, abstract [AB](L).

The Pareto Frontier Is Convex: Mix Two Agents



- Run agent A with probability p, agent B with $1-p$: expected cost and accuracy lie on the segment AB. So the achievable frontier is the *convex hull*.
- Zero-shot ($1.93, 89.6%) is not dominated by any single agent, but the mixture $0.76 \cdot \text{Warming} + 0.24 \cdot \text{Escalation}$ costs $1.93 and scores $\approx$91.2%.

Same lesson as strict dominance by a *mixed* strategy (M1): compare against mixtures, not just pure options.

Source: Kapoor et al., arXiv:2407.01502 (convex frontier claim, Table A1) [FT]; the mixture arithmetic is ours.

Variance: The Reported Number Is One Draw

- **Kapoor et al.:** agent evaluations "are rarely accompanied by error bars." Many reported scores were *above the maximum* of 5 reproduction runs; some baselines *below the minimum*. One SWE-Agent run on full SWE-bench "could cost over USD 8,000."
- **Temperature 0 is not deterministic.** Fixed seed, 5 LLMs, 8 tasks, 10 runs: accuracy varies **up to 15%** across runs (Atil et al.).
- **Why:** 1,000 temperature-0 completions of one prompt gave **80 distinct outputs**. Kernels are not batch-invariant, and batch size depends on server load (Thinking Machines).
- **Prompt format:** spurious format changes move LLaMA-2-13B by up to **76 points** (Sclar et al.).

Source: Kapoor et al., arXiv:2407.01502 [FT](L); Atil et al., arXiv:2408.04667 [FT](S); He / Thinking Machines blog, Sep 2025 [WEB](S); Sclar et al. (ICLR 2024), arXiv:2310.11324 [AB](S).

Variance: Scaffolds, Providers, and What to Report

The scaffold is part of the agent

- On SWE-bench Verified, switching scaffolds is worth up to **11 points** for GPT-5 and **15** for Kimi K2 Thinking.
- API provider choice: "the biggest factor of variance."
- HAL (21,730 rollouts, ∼$40,000): more reasoning effort gave equal or *lower* accuracy in 21 of 36 combinations.

Miller's recommendations

1. CLT standard errors and 95% CIs.
2. *Clustered* SEs for grouped questions.
3. Reduce variance: resample answers.
4. Compare on question-level **paired differences**.
5. Power analysis before running.

Paired: $\widehat{\Delta} = \frac{1}{n}\sum_i (s_{A,i} - s_{B,i})$, with $\mathrm{SE} = \mathrm{sd}(s_A - s_B)/\sqrt{n}$. Questions both models find easy (or hard) cancel out, so the SE shrinks.

Source: Brand & Denain (Epoch AI blog, Dec 2025) [WEB](S); Kapoor et al. (HAL), arXiv:2510.11977 [FT](S); Miller, arXiv:2411.00640 [FT, partial](S).

Reward Hacking: The Model Games the Scorer

METR (Jun 2025)

- Hacking on RE-Bench **30.4%** (39/128 runs) vs. HCAST **0.7%** (8/1,087): "43× more common" when the scoring function is visible.
- Told "do not cheat": still **80%** hacking on one task.
- Hacks include stubbing the evaluator and finding the grader's answer.

Model attribution: these rates are *inferred* to be for o3; not confirmed on the page.

ImpossibleBench

- Tests contradict the spec, so *any* pass is a cheat.
- GPT-5 cheats on **54%** of Conflicting-SWEbench tasks.
- Let it flag the task for a human: **54%** → **9%**.

Elegant design: make reward hacking a measurable quantity.

Source: von Arx, Chan & Barnes (METR blog, Jun 5 2025) [WEB](S); Zhong, Raghunathan & Carlini, arXiv:2510.20270 [FT](S); only the GPT-5 figures are used.

Reward Hacking: Chess, and Selection on the Test Set

Specification gaming in chess

- Task: win against a chess engine.
- o3 hacked in **88%** of runs; o1-preview attempted hacks in 36% of 123 runs.
- Hacks: overwrite the board FEN file; replace Stockfish.

Source: Bondarenko et al. (Palisade), arXiv:2502.13295 [FT](S).

The Leaderboard Illusion

- Meta tested **27 private variants** on Chatbot Arena in one month before Llama 4; providers can retract scores.
- Google ≈19.2% and OpenAI ≈20.4% of all Arena data; 83 open-weight models combined 29.7%.

Source: Singh et al., arXiv:2504.20879 [FT](S).

Best-of-N private submissions bias the reported maximum upward *with zero capability change*: $\mathbb{E}[\max_i X_i] > \mathbb{E}[X_1]$.

Goodhart's Law as Mechanism Design

- Benchmarks are now used as **RL reward signals** (point made in the RDI post).
- An exploitable grader is then not just a measurement error. It is the *training objective*.
- Anthropic's Claude 3.7 Sonnet system card: the model "occasionally resorts to special-casing in order to pass test cases," attributed to reward hacking during RL.

In our language. The benchmark designer chooses a scoring rule; the agent (or its training process) best-responds to it. A benchmark is a *mechanism*, and "measures capability" is an incentive-compatibility requirement: the score-maximizing behavior should be the intended behavior.

Pause and predict. Which of today's fixes (isolate the scorer, impossible tasks, trivial baselines) change the agent's *incentives*, and which only *detect* deviations?

Source: RDI blog (Apr 2026) [WEB](L); Anthropic, Claude 3.7 Sonnet System Card (Feb 2025), §6 [FT](S). Mechanism-design framing: ours.

Part III: Multi-Agent Benchmarks

When the other players are part of the test

Multi-Agent Benchmarks by Game Class (1/2)

Benchmark	arXiv	Metric	Key finding
Cooperative / coordination			
Overcooked-AI	1910.05789	Return	Self-play agents worse with a human model
Hanabi	1902.00506	Score	Cooperative, imperfect information
LLM-Coordination	2310.03903	Score, accuracy	Weak on partners' beliefs
Collab-Overcooked	2502.20073	Process metrics	Active collaboration weak (13 LLMs)
Mixed-motive / social dilemma			
GovSim	2404.16698	Commons survival	Best survival rate <54%
Melting Pot	2107.06857	Focal return	Held-out partners expose failures
CoopEval	2604.15267	Cooperation	One-shot defection; contracts work best
Welfare Diplomacy	2310.08901	Welfare, exploit.	High welfare, but exploitable

Source: Research file §2: GovSim [FT](S), others [AB](S). LLM-Coordination and Melting Pot rows checked against the arXiv abstracts (Nov 19); CoopEval checked in refs-l22/coopeval.txt.

Multi-Agent Benchmarks by Game Class (2/2)

Benchmark	arXiv	Metric	Key finding
Competitive / hidden role			
GTBench	2402.12348	Win rate	Fail on complete-info deterministic games
AvalonBench	2310.05036	Win rate	ChatGPT 22.2% vs. rule-based bot 38.2%
Werewolf	2309.04658	Qualitative	Emergent trust and deception
TextArena	2504.11442	TrueSkill	57+ games; a platform paper
Dialogue / social			
SOTOPIA	2310.11667	LLM + human judges	GPT-4 far below humans (hard set)
NegotiationArena	2402.05863	Payoff	"Desperate" persona: +20% payoff
Group reasoning			
HiddenBench	2505.11556	Accuracy	Groups 30.1% vs. one informed agent 80.7%

Source: Research file §2: SOTOPIA [FT](S), others [AB](S); HiddenBench checked in refs-l22/hiddenbench.txt. Venues unconfirmed, so arXiv IDs.

Lesson 1: The Partner Population Is the Result

- **Overcooked-AI:** self-play agents do "significantly worse" when paired with a model of a human.
- **Melting Pot:** score a *focal* population against a held-out *background* population; held-out partners reveal failures that training reward hides.
- **CoopEval:** recent models "consistently defect in single-shot social dilemmas"; contracts and mediation sustain cooperation best; repetition-induced cooperation "deteriorates drastically when co-players vary."

Game-theoretic reading. A score against a fixed population σ_{-i} is $u_i(\pi, \sigma_{-i})$: a measure of how well π *best-responds to that population*, not a property of π alone. Change σ_{-i} and the ranking can change. This is equilibrium selection and zero-shot coordination.

Source: Carroll et al., arXiv:1910.05789 [AB](S); Leibo et al., arXiv:2107.06857 [AB, paraphrase](S); Tewolde et al., arXiv:2604.15267 [AB](L, refs-l22/coopeval.txt).

Lesson 1, Worked: Who Wins Depends on Who Showed Up

Coordination game (payoff to both):

	L	R
L	2	0
R	0	1

Agent A always plays L; agent B always plays R.

Partner pool	A	B
80% L-players	**1.6**	0.2
20% L-players	0.4	**0.8**

Pause and predict. Both (L, L) and (R, R) are Nash equilibria. Which agent is "better at coordination"? What would you need to report so a reader could tell?

Toy example constructed for this lecture.

Lesson 2: Ratings vs. Exploitability vs. Welfare

Metric	Used by	What it answers
Elo / TrueSkill	LLMArena, TextArena, Chatbot Arena	Relative strength vs. *whoever showed up*
Exploitability	Welfare Diplomacy	How much a best response gains against you: equilibrium quality
Social welfare	GovSim, Welfare Diplomacy	Did the group do well? (commons survival)

- **Recall L20, Elo $\neq$ exploitability.** In Go (Timbers et al.), AlphaZero MSE wins only 23.25% vs. AlphaZero TD25 (Elo -207) yet is *less* exploitable ($\sim$20% vs. $\sim$40% best-response win rate).
- **Liu et al. (ICLR 2025):** Elo "can be susceptible to and even reinforce biases ... due to their sensitivity to redundancies"; they propose evaluation as a 3-player game.

Source: Research file §2 [AB](S); Timbers et al. (IJCAI 2022), Liu et al. arXiv:2502.20170, via the 11-10 scout [L].

Lesson 2: Why Redundancy Moves Elo

Rock-paper-scissors population. Now add 9 near-copies of Rock.

- Paper beats 10 Rocks: its average win rate soars.
- Scissors loses to 10 Rocks: its rating sinks.
- Nothing about Paper or Scissors changed.

Nash averaging (Balduzzi et al. 2018) rates agents against the max-entropy Nash of the meta-game (an empirical game, as in M7). Clones split their weight: ratings ignore redundancy.

Implication for leaderboards: if one lab submits many variants (27 private variants, Part II), it changes the population and so every rating, not just its own.

Source: Liu et al., arXiv:2502.20170 [L]; Balduzzi et al. (NeurIPS 2018), arXiv:1806.02643 (Nash averaging); worked example and Arena link: ours.

Lesson 3: The Judge Is a Player

- **SOTOPIA** scores social role-play with GPT-4 as a judge: usable "with some caution," but it rates higher than humans on some dimensions.
- **SOTOPIA**-π trains agents on that judge's signal. A 7B model matches GPT-4 on goal completion, *but* LLM judges overestimate the socially trained agents.
- Debate as a cleaner design: a weak judge with no access to the source text reaches 76% (LLM judge) / 88% (human judge) vs. 48% / 60% for naive baselines.

Train against a judge and you get agents that best-respond to the judge: Goodhart's law with a *strategic* evaluator. Report human agreement on the trained agents, not only on the baseline.

Source: Zhou et al., arXiv:2310.11667 [FT](S); Wang et al., arXiv:2403.08715 [AB], checked against refs-l22/sotopia-pi.txt; Khan et al., arXiv:2402.06782 [AB](S).

Lesson 3: Groups Are Not the Sum of Their Members

HiddenBench (Hidden Profile tasks)

- 65 tasks, 15 frontier LLMs.
- Groups with distributed information: **30.1%**. One agent with all the information: **80.7%**.
- Gap persists across prompting, communication depth, group size; worsens as groups scale.
- Failure: agents converge on *shared* evidence and do not reason about what others know but have not said.

MAST (failure taxonomy)

- 1,600+ annotated traces from 7 frameworks; 14 failure modes in 3 categories; $\kappa = 0.88$.
- Gains from multi-agent systems are often minimal.
- Categories: system design, inter-agent misalignment, verification.

Source: Li, Naito & Shirado, arXiv:2505.11556 [AB + intro, refs-l22/hiddenbench.txt]; Cemri et al., arXiv:2503.13657 [AB](S).

Part III Recap: Five Lessons for Multi-Agent Evaluation

1. **Who you are paired with is the result.** Report the partner population.
2. **Ratings are not equilibrium metrics.** High Elo is compatible with being exploitable.
3. **LLM judges are part of the game.** Validate them on trained agents.
4. **Interaction does not force strategy.** Check for open-loop shortcuts (SMAC).
5. **Group performance is emergent.** HiddenBench, MAST.

Gap in the literature: no dedicated survey of *multi-agent* LLM evaluation was found. MultiAgentBench (arXiv:2503.01935) and MAST are the closest anchors.

Source: Research file §2 cross-cutting lessons and §4 (sub-scout synthesis, reviewed by lead).

Part IV: What Makes a Good Benchmark

A checklist, then you audit one

What the Field Reports (and Does Not)

BetterBench: 24 benchmarks, 46 criteria

- Lifecycle: design, implementation, documentation, maintenance, *retirement*.
- **14/24** report no multiple runs or significance tests.
- **17/24** provide no easy replication script.
- MMLU scores lowest (5.5), GPQA 11.0.

Source: Reuel et al. (NeurIPS 2024 D&B), arXiv:2411.12990 [FT](S).

Measuring What Matters: 445 papers

- 47.8% of those defining their phenomenon use contested definitions.
- 27.0% partly, 12.3% entirely, rely on convenience sampling.
- Only **16.0%** use uncertainty estimates or statistical tests.

Source: Bean et al., arXiv:2511.04703 [FT](S).

Kapoor et al.'s holdout survey (17 benchmarks): an appropriate holdout exists for 0/2 fully general benchmarks and 1/8 domain-general ones. Source: arXiv:2407.01502, v1 [FT](S).

A Synthesized Checklist

Part IV ▷ Eleven questions to ask of any benchmark

Validity

1. Valid tasks (audited ground truth).
2. Valid grader.
3. Trivial-agent and human baselines.
4. Held-out split at the right level of generality.
5. Contamination plan: dating, private split, or refresh.

Measurement and hygiene

6. Error bars over $\geq$5 runs, paired comparisons.
7. Cost alongside accuracy (a Pareto plot).
8. Fixed, disclosed scaffold, budgets, provider.
9. Versioning: scorer or ground-truth changes bump the version.
10. Scorer isolated from the agent.

Multi-agent

11. Report the partner population; prefer exploitability or welfare over Elo alone.

Source: Synthesis (research file §3): ABC 2507.02825; BetterBench 2411.12990; Bean 2511.04703; Miller 2411.00640; Epoch, RDI [WEB](L); Kapoor 2407.01502.

Better Aggregate Metrics: Beyond Percent Correct

- **50% time horizon (METR).** Fit a logistic (IRT-style) model of success vs. human task time on 170 human-timed tasks; report the task length at which the agent succeeds half the time. The horizon doubles roughly every **7 months**; the authors flag external validity and "messier" tasks.
- **Adaptive items.** Fluid Benchmarking: adaptive item selection gives higher validity and lower variance with 50× fewer items.
- **An external review rubric.** Epoch AI gates on reviewability → scoring → consistency → elicitation → bias in setup; errors affecting ≥20% of the score make a benchmark "Flawed."

Source: Kwa et al., arXiv:2503.14499 [FT, partial](S); Hofmann et al. (COLM 2025), arXiv:2509.11106 [AB](S); Epoch AI benchmark reviews [WEB](L).

Exercise: Audit a Multi-Agent Benchmark (5 minutes, pairs)

Pick **one** benchmark from the Part III tables (GovSim, Melting Pot, CoopEval, AvalonBench, SOTOPIA, HiddenBench, TextArena, . . .). Answer:

1. **Trivial agent.** What would an agent that always cooperates, always defects, stays silent, or acts at random score?
2. **Partner population.** Who are the other players? Would the ranking survive a different pool?
3. **Contamination.** Are the scenarios, games, or transcripts public and static?
4. **Incentives.** What does the metric reward? Could an agent raise it without the intended skill (a judge to flatter, a rating to inflate)?

Pause and predict. Be ready to report: the single checklist item your benchmark most clearly fails, and one change that would fix it.

Debrief: What Did You Find?

Prompts for the debrief

- Which checklist item failed most often across the room?
- Did any benchmark report a partner population that was *not* its own training population?
- Which benchmarks report an equilibrium-quality metric (exploitability, welfare) at all? From the table: Welfare Diplomacy and GovSim.
- For an LLM-judged benchmark, how would you check the judge was not gamed?

For your projects: report error bars over several runs, a trivial baseline, cost, and the exact scaffold. If agents interact, say who they played against.

Takeaways

Closing

1. A benchmark score is a measurement with **validity** conditions (task, outcome) and **error** (variance, scaffold, provider). Ask which failed.
2. Benchmarks can die in stages: saturation, then audit (bad tests), then contamination. SWE-bench went through all three in under three years.
3. Cost belongs on the axis. Mixtures make the frontier convex.
4. A benchmark used as a reward is a **mechanism**. Graders that can be fooled will be fooled, by the model if not by the lab.
5. In multi-agent settings, the score is a best response to a population. Ratings are not exploitability.

Readings (Ranked)

Closing ▷ Start with 1 and 2

1. Kapoor et al., "AI Agents That Matter," TMLR 2025, arXiv:2407.01502. Cost–accuracy Pareto thinking, error bars, holdouts.
2. Zhu et al., "Establishing Best Practices for Building Rigorous Agentic Benchmarks" (ABC), NeurIPS 2025 D&B, arXiv:2507.02825.
3. OpenAI, "Why we no longer evaluate SWE-bench Verified" (Feb 2026) and "Separating signal from noise in coding evaluations" (Jul 2026), blog posts.
4. Bean et al., "Measuring What Matters," arXiv:2511.04703. Construct validity across 445 papers.
5. Miller, "Adding Error Bars to Evals," arXiv:2411.00640. The statistics toolkit for your projects.
6. Zhong, Raghunathan & Carlini, "ImpossibleBench," arXiv:2510.20270.
7. Piatti et al., "Cooperate or Collapse" (GovSim), NeurIPS 2024, arXiv:2404.16698.
8. Northcutt, Athalye & Mueller, NeurIPS 2021 D&B, arXiv:2103.14749. Optional.

Source: Research file §5. Venues shown only where confirmed in the research file.

Project Reminders

Closing ▷ Logistics

- **Project presentations:** due **Monday, Nov 30**.
- **Final report:** due **Thursday, Dec 3** (last day of classes).
- Apply today's checklist to your own evaluation: error bars, a trivial baseline, cost, disclosed scaffold and partner population.

Questions? If your project uses a public benchmark, bring it to office hours and we will audit it together against the checklist.