InquiryBench: Evaluating LLM Agents Across the Full Scientific Inquiry Cycle

Anonymous ACL submission

Abstract

Benchmarks for scientific language agents—DS-1000, ScienceAgentBench, BIRD—evaluate a single phase of inquiry: *given a dataset and question, produce the correct result*. We introduce **InquiryBench**, to our knowledge the first benchmark spanning all six NGSS Science & Engineering Practices (SEPs), from formulating investigable hypotheses (SEP1) to communicating findings (SEP6), enabling systematic measurement of how agent capabilities vary across the *full* inquiry cycle. Across 120 questions $\times$ 2 agent configurations (structured tool-use vs. ungrounded LLM), we find that **scaffolding effectiveness is strongly heterogeneous**: oracle-verified advantages are largest on data quality assessment (SEP3, $\Delta{+}1.38$ LP points) and data analysis (SEP4, $\Delta{+}0.88$); hypothesis formulation (SEP1, $\Delta{+}1.14$) and explanation quality (SEP5, $\Delta{+}0.32$) show large gaps via exploratory keyword grading—all six proxy-score gaps are significant (Mann-Whitney, Bonferroni-corrected $p{<}0.001$, $r{=}0.26$–0.52), though generative phases remain exploratory pending human calibration. Extending to 182 questions across three openness levels, we find a **non-monotonic openness pattern**: structured agent advantage is largest on constrained tasks ($+20.0$ pp) and collapses completely on open inquiry tasks where the agent must autonomously determine what to investigate (L3: **0.000** for both architectures, LLM-judge confirmed and trace-supported)—an *open inquiry cliff* traceable to the absence of question-generation capability. We further document a **contamination effect**: apparent tool-use binary correctness advantages are $2.5\times$ larger on publicly-known datasets than on post-cutoff synthetic datasets, consistent with contamination or LLM memorization effects. As a diagnostic check, a 3-turn agentic extension (curiosity engine $\rightarrow$ feasibility gate $\rightarrow$ structured execution) achieves **46.7%** execution-oracle hit rate on 30 L3 Type-A questions—outperforming Raw-NL (0.0%, $p{<}0.001$) and AOHN variants (36.7–40.0%)—with 100% Turn-3 pass-through, supporting the diagnosis that question-generation capability is the enabling architectural change. InquiryBench v2 (182 questions, three openness levels, five post-cutoff synthetic datasets) will be released upon acceptance with evaluation code, oracle verification, and full answer traces.

1 Introduction

Scientific language agents—LLM-based systems that analyze structured datasets and communicate scientific reasoning in natural language—are increasingly deployed to assist with data-driven inquiry. Yet every major AI-for-science benchmark—DS-1000 (Lai et al., 2023), ScienceAgentBench (Chen et al., 2025), BIRD (Li et al., 2023)—evaluates only one step: *given a question and a dataset, produce the correct analytical result*. This corresponds to what the NGSS framework (NGSS Lead States, 2013) calls SEP4 (Analyzing and Interpreting Data). A system that scores 0.97 on SEP4 may simultaneously produce no testable hypothesis (SEP1), miss a key confounder (SEP2), ignore data quality problems (SEP3), overstate its conclusion (SEP5), and communicate results at the wrong level of abstraction (SEP6). None of these failures are visible in SEP4-only evaluations.

NLP scope. InquiryBench is an NLP benchmark: every task is solved by reading, analyzing, and narrating structured datasets without physical experimentation. It evaluates *scientific language agents* on a six-phase task taxonomy, positioning alongside DS-1000, ScienceAgentBench, and AgentBench (Liu et al., 2024) while extending the NLP evaluation ecosystem from single-step analysis to multi-phase scientific inquiry.

The reason prior benchmarks cluster at SEP4 is not that it is most important—it is that it is eas-

iest to evaluate. Auto-evaluation against known numeric oracles is cheap and reliable. Evaluating hypothesis formulation or conclusions requires LLM-as-judge calibrated against domain experts, which is tractable but has been avoided.

This paper introduces InquiryBench, to our knowledge the first benchmark measuring scientific language agent capability across all six inquiry phases. Our central contribution is an *evaluability spectrum* that explains why the field has been stuck at SEP4, and an evaluation protocol that covers the full cycle.

Principal findings.

- **SEP heterogeneity**: structured agent LP scores range 0.77–2.23 (factor 2.9×); all six SEP gaps significant (Kruskal-Wallis $H = 118.43$, $p<0.0001$).
- **Oracle-verified advantages**: data quality assessment (SEP3, Δ+1.38) and data analysis (SEP4, Δ+0.88) show the largest, fully reproducible structured-agent gains.
- **Exploratory advantages**: hypothesis formulation (SEP1, Δ+1.14, r=0.52) and communication (SEP6, Δ+0.93) show large gaps via keyword grader; human calibration in progress.
- **SEP5 reversal** (*exploratory*): the ungrounded LLM initially outperforms the structured agent on CER conclusions—a grader artifact fixed by adding a Synthesizer stage (AOHN SEP5 0.33→0.77, $p = 0.0001$).
- **Open inquiry cliff** (*LLM-judge confirmed, trace-supported*): both architectures score exactly 0.000 on L3 open inquiry tasks—neither generates an investigable hypothesis from data without a question scaffold.
- **Contamination** (*diagnostic; partially confounded*): +40.4 pp advantage on public datasets vs. +16.5 pp on post-cutoff synthetic data (2.5× observed difference, consistent with memorization; see §7).

Contributions.

1. **Evaluability spectrum** (§3): taxonomy of why prior benchmarks cluster at SEP4 and what evaluation each SEP requires.
2. **InquiryBench dataset** (§4): 182 questions with phase-specific evaluation protocols and post-cutoff synthetic datasets with verified oracle values.
3. **Evaluation protocol** (§5): LLM-as-judge methodology with explicit validity claim tiers (oracle-verified vs. exploratory) for each inquiry phase.
4. **Empirical results** (§6): to our knowledge, the first per-SEP profile across structured vs. ungrounded LLMs, including the open inquiry cliff and contamination analysis.

2 Background

NGSS Science & Engineering Practices. The Next Generation Science Standards (NGSS Lead States, 2013) define eight SEPs. We focus on the six most directly assessable by an LLM agent (Table 1), following the Pedaste et al. (2015) inquiry-cycle framework.

Table 1: NGSS SEPs and evaluability.

SEP	Practice	Evaluability
1	Hypothesis formulation	Hard (generative)
2	Investigation planning	Medium (DAG GT)
3	Data quality assessment	Easy (planted GT)
4	Data analysis	Easy (numeric oracle)
5	Constructing explanations (CER)	Hard (generative)
6	Communicating findings	Medium (style grader)

Prior AI-for-science benchmarks. DS-1000 (Lai et al., 2023), ScienceAgentBench (Chen et al., 2025), and BIRD (Li et al., 2023) each measure SEP4 only. Prior work using ReAct (Yao et al., 2023)-style tool-augmented agents has similarly focused on analysis correctness. To our knowledge, InquiryBench is the first benchmark to cover all six SEPs with distinct evaluation protocols per phase.

Language agent evaluation. AgentBench (Liu et al., 2024) evaluates LLMs as sequential decision-making agents across diverse task environments, establishing that agent capability profiles differ substantially across task types. InquiryBench applies this multi-task profiling lens to scientific inquiry, where the "task environments" correspond to the six NGSS practices and agent outputs are scientific language artifacts (hypotheses, analyses, explanations) rather than executable actions.

LLM-as-judge methodology. Zheng et al. (2023) established calibration requirements and failure modes for LLM-as-judge evaluation of open-ended responses. InquiryBench adopts this approach for SEP1 (hypothesis novelty) and SEP5 (CER quality), and introduces an explicit

validity claim tier framework—distinguishing oracle-verified (SEP3/SEP4), calibrated-judge, and exploratory (keyword proxy) claims—so that readers can identify which findings carry different evidential weight.

3 AI-Role Taxonomy and Evaluability Spectrum

Table 2 characterizes each SEP: the student action, the AI role, LLM mode required, and evaluability. The taxonomy makes explicit that prior benchmark clustering at SEP4 is a measurement selection effect, not an importance judgment.

Table 2: Per-SEP taxonomy with InquiryBench empirical results (n=120/arm). All six proxy-score gaps significant (Bonferroni α=0.0083); validity tier varies by SEP.

SEP	AI role	LLM mode	Result (Δ)
1	Surfaces patterns; proposes hypotheses	Gen + tool	1.54 vs 0.40 (+1.14)
2	`plan_analysis`: hypothesis + method before run	Causal + plan	2.15 vs 1.16 (+0.99)
3	Mandatory DQ pre-check; detects missing/outlier patterns	Analytic + tool	1.68 vs 0.30 (+1.38)
4	Typed cell system: 168 cells × 29 families, evidence bundles	Function-call	2.23 vs 1.34 (+0.88)
5	Synthesizer stage: CER verbalization of evidence bundle	Evaluative+gen	0.77 vs 0.45 (+0.32)
6	Pedagogue: level-adaptive narration (3 audience levels)	Generative	1.63 vs 0.71 (+0.93)

The evaluability spectrum. Auto-eval reliability decreases as generativity increases:

$$\underbrace{\text{SEP3}}_{\text{planted GT}} \rightarrow \underbrace{\text{SEP4}}_{\text{oracle}} \rightarrow \underbrace{\text{SEP2}}_{\text{DAG}} \rightarrow \underbrace{\text{SEP6}}_{\text{style}} \rightarrow \underbrace{\text{SEP5}}_{\text{CER}} \rightarrow \underbrace{\text{SEP1}}_{\text{novelty}}$$

Prior benchmarks occupy the leftmost positions; SEP1 and SEP5 require LLM-as-judge calibrated against experts—tractable but requiring human annotation time that prior work has avoided.

Failure-mode asymmetry. **Structured agents fail silently**: on failure, the system returns an empty string (LP= 0). Students receive no scaffolding and no signal of failure. **Ungrounded LLMs fail verbosely**: fluent refusal text ("Cannot determine because...") contains SEP-like keywords, inflating SEP5 scores. Both failure modes are problematic; neither is visible in SEP4-only benchmarks.

4 InquiryBench Dataset

Design principles. Each scenario must: (1) support $\geq$4 SEPs with phase-specific GT or evaluation protocol; (2) have known causal structure for SEP2; (3) contain planted data features for SEP3; (4) use synthetic datasets with analytically verified oracle values to eliminate contamination.

Benchmark versions. InquiryBench v1 (120Q, publicly-known datasets; main SEP comparison) and v2 (182Q, post-cutoff synthetic datasets; openness + contamination analysis).

Post-cutoff synthetic datasets (v2). Five datasets with deterministically planted structure and verified oracles, all generated after the model's training cutoff:

Table 3: InquiryBench v2 synthetic datasets (37/37 oracle checks pass).

Dataset	Domain	n	Key oracle
DS-4 Health	Public health	800	ATE $= -35.0$
DS-7 IV Fert.	Agriculture	500	LATE $= 0.40$
DS-8 Air Qual.	Environment	300	k^*=5, ARI= 1
DS-9 Bayes Vac.	Epidemiology	180	$\hat{\theta}$=0.214
DS-10 RDD Sch.	Education	1000	LATE $= 0.30$

Questions are distributed across three openness levels: **L1 constrained** (dataset + question given), **L2 guided** (dataset given; question open-ended), and **L3 open** (agent must decide what to investigate). Question prompts *never prescribe steps*—the autonomous inquiry principle ensures that ungrounded LLMs cannot score by following instructions.

5 Evaluation Protocol

Phase-specific methods. **SEP3/SEP4**: numeric oracle with tolerance; SEP3 issue-detection precision/recall vs. planted manifest. **SEP2/SEP6**: confounder recall + F1 vs. known DAG; deterministic audience-adaptation grader. **SEP1/SEP5**: LLM-as-judge (Zheng et al., 2023) with calibration requirement: $N_{\text{cal}} = 150$ items, expert $\kappa > 0.60$, judge-expert Pearson $r > 0.70$.

Validity claim hierarchy.

Tier	Condition	Strength
Strong	Auto-eval, known GT	Main quantitative
Calibrated	LLM-judge r>0.70 vs. expert	Valid for ranking
Exploratory	Uncalibrated keyword proxy	Directional only

Current results: SEP4 = Strong (numeric oracle); SEP3 binary = Strong; SEP1, SEP2, SEP3 LP, SEP5, SEP6 = Exploratory (keyword grader; human calibration is in progress).

6 Experiments

Setup. We evaluate two primary arms on 120Q (v1): **AOHN v3**—structured tool-use agent with 168 typed analysis cells, Synthesizer + Pedagogue pipeline, and `plan_analysis` hypothesis tool—and **Raw-NL**—same GPT backbone, natural language only, no tools. Extended baselines: **ReAct** (Yao et al., 2023) (tool-use without evidence bundles) and **GPT-4o CoT** (chain-of-thought, no tools). All four baselines are GPT-family models. Primary comparison: Mann-Whitney U, Bonferroni $\alpha = 0.0083$ ($n = 120$/arm).

6.1 Main SEP Results

Table 4: Per-SEP LP mean, AOHN v3 vs. Raw-NL (n=120/arm, same GPT backbone). All six proxy-score gaps significant at Bonferroni α=0.0083; validity tier varies by SEP (SEP3 binary/SEP4: Strong; SEP1/SEP5: Exploratory).

SEP	AOHN	Raw-NL	Δ	p	r
SEP1 Hypothesis	**1.54**	0.40	+1.14	<0.001	0.52
SEP2 Planning	**2.15**	1.16	+0.99	<0.001	0.33
SEP3 Data qual.	**1.68**	0.30	+1.38	<0.001	large
SEP4 Analysis	**2.23**	1.34	+0.88	<0.001	0.33
SEP5 CER	**0.77**	0.45	+0.32	0.0001	0.26
SEP6 Communic.	**1.63**	0.71	+0.93	<0.001	0.36
Correctness	**74.2%**	33.8%	+40.4pp	—	—

H1 (SEP heterogeneity within AOHN): Kruskal-Wallis $H = 118.43$, p<0.0001. AOHN scores range 0.77–2.23 (factor 2.9×): the system that scores highest on SEP4 does not score highest on SEP5, and neither ranking would be visible in a SEP4-only benchmark. **Evidence tiers**: all six proxy-score gaps are significant; however, only SEP3 binary detection and SEP4 numeric correctness are oracle-verified (Strong tier). SEP1, SEP2, SEP3 LP, SEP5, and SEP6 gaps are Exploratory (keyword grader; human calibration in progress) and should be read as directional until calibration is complete.

All-baselines LP comparison. Table 5 shows that ReAct achieves 73.9% oracle correctness—nearly matching AOHN (74.2%)—yet its LP mean is only 0.289 vs. AOHN's 0.556 (−48%). Oracle correctness and LP quality measure different constructs: a system can produce the right number while providing inquiry reasoning of substantially lower quality.

Table 5: LP across all baselines (n=120/arm). *SEP6 CoT artifact (verbose prose inflates keyword-based SEP6; see §7).

System	S1	S2	S3	S4	S5	S6	Avg
AOHN v3	**1.54**	**2.15**	**1.68**	**2.23**	**0.77**	1.63	**0.556**
GPT-4o CoT	0.59	0.91	0.58	1.35	0.41	**1.95***	0.321
ReAct	0.63	1.13	0.60	1.11	0.51	1.23	0.289
Raw-NL	0.40	1.16	0.30	1.34	0.45	0.71	0.242

6.2 The SEP5 Reversal

At v2 baseline (no Synthesizer), Raw-NL outperformed AOHN on SEP5. **Mechanism**: AOHN produced correct bare numbers without verbal conclusion; Raw-NL produced fluent refusal text with CER-like keywords (e.g., "Cannot determine because [causal explanation]")—scored LP2–3.

Fix: (1) Synthesizer stage converting evidence bundles to natural-language CER conclusions; (2) Rule 9 in system prompt requiring conclusion language in `final_answer`. Result: AOHN SEP5 0.33→0.77 (+0.44), $p = 0.0001$.

Implication: SEP5 quality is an architectural choice, not a fundamental limitation of structured tool-use. Without InquiryBench's per-SEP grading, this reversal—and its resolution—would be invisible.

6.3 Openness Level Analysis and the Open-Inquiry Cliff

We evaluate H5: AOHN's advantage scales monotonically with task openness (L1 < L2 < L3), predicting the largest advantage on fully open tasks.

Table 6: Binary correctness by openness level (182Q × 2 arms). L3 graded by LLM judge (GPT-4.1-mini); oracle otherwise.

Level	AOHN	Raw-NL	Δ	H5?
L1 Constrained	45.0%	25.0%	+20.0pp	—
L2 Guided	80.1%	62.3%	+17.8pp	↓✗
L3 Open	**0.0%**	**0.0%**	+0.0pp	—
Overall	69.2%	52.7%	+16.5pp	

H5 not supported. AOHN's largest advantage is at L1, not L3. The L1→L2 margin decrease (−2.2 pp) may reflect L2's wider range of valid inquiry paths. The dominant result is the L3 collapse: both systems score exactly **0.000** on open inquiry tasks (all 32 runs scored incorrect by LLM judge).

The cliff is not a grading artifact. Three confirmations: (1) the LLM judge is explicitly

permissive—any valid investigable question is credited; (2) AOHN tool-call traces show 0.56 tool calls/L3 vs. 0.72/L2—the systems do not engage with the data; (3) judge reasoning traces: "does not inspect the dataset", "does not identify a testable question." The cliff reflects an *architectural boundary*: neither structured tool-use nor ungrounded prompting has a mechanism for generating investigable research questions from data without a question scaffold. A 3-turn agentic extension achieves 46.7% oracle hit rate on the same question type (§6.4).

6.4 Bridging the Cliff: ECE-DIG 3-Turn Agent

As a diagnostic check on whether the observed cliff reflects a real architectural gap rather than a grading artifact, we implemented and evaluated a 3-turn extension: **Turn 1** runs the Epistemic Curiosity Engine (ECE) to generate candidate research questions from data; **Turn 2** applies an explicit feasibility gate—logging whether the highest-EIG candidate is operationalizable and re-selecting from the bundle if not; **Turn 3** runs AOHN-v2 execution only when Turn 2 passes. We evaluate on 30 Type-A L3 questions (all five domains, 6 questions each).

Metric note. Table 7 reports *execution oracle hit rate* (OH): whether the agent's final numerical answer matches the planted oracle. This is a different, weaker criterion than the L3 correctness used in Table 6, which requires both autonomous question formulation and grounded data inspection (LLM judge). AOHN variants can hit the oracle opportunistically—by traversing their cell library without formulating a research question—and thus score 36.7–40.0% OH here while scoring 0.000 on the full L3 criterion in Table 6.

Table 7: ECE-DIG 3-Turn arm vs. selected baselines on 30 Type-A L3 questions. OH = execution oracle hit rate (numerical answer vs. planted oracle; see text for distinction from Table 6 criterion). OH difference vs. AOHN-v4 is directional (p=0.80, n=30); ECE-DIG vs. Raw-NL: p<0.001 (Fisher exact).

Arm	AIS	OH	LP	T3
ECE-DIG 3-Turn	0.444	**46.7%**	0.719	100%
AOHN-v4 (ECE+v3)	0.522	40.0%	0.739	—
AOHN-v3 (no ECE)	0.367	36.7%	0.603	—
Raw-NL	0.063	0.0%	0.396	—

ECE-DIG achieves 46.7% oracle hit rate on 30 Type-A L3 questions, with all 30 tasks completing Turn 3. For comparison, Raw-NL achieves 0.0% (p<0.001, Fisher exact) and AOHN variants achieve 36.7–40.0% (directional; p=0.80). The AOHN result reflects opportunistic analysis without explicit question generation: the system sometimes produces the correct numerical answer while failing the autonomous-question-quality criterion that defines the cliff. ECE-DIG's improvement over AOHN and guaranteed Turn-3 pass-through support the interpretation that systematic question-generation is the enabling capability. The +6.7 pp advantage over AOHN-v4 is underpowered at n=30; reaching 80% power for this comparison would require n≈855/arm.

Domain-stratified results reveal an oracle-design tension. Four domains (HP, IV, AQ, RDD) achieve 58.3% oracle hit rate; the Bayesian Vaccine domain achieves 0%. The BV result is *not* an agent execution failure—Turn 3 completes successfully—but a mismatch between oracle-designed and scientifically interesting questions: the ECE generates prior-sensitivity and credible-interval questions (scientifically appropriate for Bayesian posteriors) while the planted oracle tests posterior-mean estimation ("read $\hat{\theta} = 0.214$"). This distinction illustrates a fundamental challenge for open-inquiry benchmarks: the oracle rewards the question the designer anticipated, not the most defensible scientific question given the data. RDD achieves 100% oracle hit (6/6), achieving perfect execution-oracle hit rate for this domain.

7 Discussion

Table 8: Binary correctness advantage by dataset type.

Bank	AOHN	Raw-NL	Δ
v1: public (IHDP, LaLonde, Iris, Wine)	74.2%	33.8%	+40.4pp
v2: post-cutoff synthetic	69.2%	52.7%	+16.5pp
Inflation factor			2.5×

Contamination and dataset choice. Using publicly-known datasets is associated with a 2.5× larger apparent tool-use binary correctness advantage (Table 8), consistent with contamination or LLM memorization effects. Because dataset family, task design, and benchmark version differ between v1 and v2, this is diagnostic evidence rather than a controlled causal estimate. The LP advantage (Δ0.88–1.38 per SEP) is *not* contamination-sensitive because LP scores measure reasoning

structure and depth, not answer recall. Future AI-for-science benchmarks should require post-cutoff synthetic datasets with analytically verified oracles.

Architecture predicts SEP profile. Each AOHN commitment predicts its own SEP gain: typed cell system (SEP4 $+0.88$), Planner (SEP2 $+0.99$), Pedagogue (SEP6 $+0.93$), DQ pre-check (SEP3 $+1.38$), `plan_analysis` + Synthesizer (SEP1 $+1.14$, SEP5 $+0.32$). SEP coverage should be a *first-class design criterion* for LLM-agent evaluation.

Domain choice as an evaluation lever. Causal inference tasks activate SEP2/SEP4/SEP6 at >1.5; basic statistics tasks suppress all SEPs regardless of agent. Domain selection is not neutral—it determines which capabilities are measurable in a benchmark.

SEP6 keyword artifact. GPT-4o CoT's SEP6 (1.95) exceeds AOHN's (1.63) because CoT generates verbose prose that triggers audience-adaptation keywords. This is a known limitation of keyword-based SEP6 scoring.

Limitations. *Exploratory evaluation* for SEP1, SEP2, SEP3 LP, SEP5, and SEP6 (keyword grader pending human calibration; only SEP3 binary detection and SEP4 numeric correctness are oracle-verified). *Single model family* (GPT). *No ecological validity*: InquiryBench measures AI-to-AI performance; whether AOHN's SEP advantages translate to measurable student learning gains requires a classroom study that is in preparation. *Open inquiry cliff as boundary, not ceiling*: the cliff is not evidence that L3 items are poorly designed—both architectures simply lack question-generation capability, as traced in tool-call logs. A 3-turn diagnostic extension (ECE-DIG; §6.4) achieves 46.7% oracle hit rate on 30 L3 Type-A questions (100% Turn-3 pass-through), supporting the interpretation that the cliff is architecturally bridgeable. The Bayesian Vaccine domain achieves 0% oracle hit not from execution failure but from oracle-design tension: the ECE proposes prior-sensitivity questions while the oracle rewards posterior-mean estimation, illustrating the challenge of designing oracles for open-inquiry tasks. *Single dataset version for LP analysis*: the main per-SEP LP comparison uses v1 (publicly-known datasets), which may inflate binary correctness baselines; LP scores are not contamination-sensitive (see Discussion).

8 Conclusion

We introduce InquiryBench, to our knowledge the first NLP benchmark evaluating scientific language agents across all six NGSS Science & Engineering Practices. Oracle-verified results show that structured tool-use advantages are largest on data quality assessment (SEP3, $\Delta{+}1.38$) and data analysis (SEP4, $\Delta{+}0.88$), not uniformly distributed. The SEP5 reversal—where ungrounded LLMs initially outperform structured agents on CER conclusions—is a fixable architectural flaw, not a fundamental trade-off. The open inquiry cliff ($L3{=}0.000$ for both architectures) reveals a concrete, measurable capability boundary traceable to the absence of question-generation capability; a 3-turn diagnostic extension supports the interpretation that the boundary is architectural rather than a grading artifact. The public-vs-post-cutoff comparison reveals a $2.5\times$ apparent advantage difference consistent with contamination or memorization effects. InquiryBench v2 will be released upon acceptance with evaluation code and oracle verification to support reproducible AI-for-science research.

References

Ziru Chen and 1 others. 2025. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations (ICLR)*. ArXiv:2410.05080.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Scott Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2023. DS-1000: A natural and reliable benchmark for data science code generation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. ArXiv:2211.11501.

Jinyang Li and 1 others. 2023. Can LLM already serve as a database interface? A BIg bench for large-scale database grounded text-to-SQLs. In *Advances in Neural Information Processing Systems*, volume 36. ArXiv:2305.03111.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hanglei Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2024. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*. ArXiv:2308.03688.

NGSS Lead States. 2013. *Next Generation Science Standards: For states, by states*. The National Academies Press.

Margus Pedaste, Mario Mäeots, Leo A. Siiman, Ton de Jong, Siswa A. N. van Riesen, Ellen T. Kamp, Costas C. Manoli, Zacharias C. Zacharia, and Eleftheria Tsourlidaki. 2015. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14:47–61.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*. ArXiv:2210.03629.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36. ArXiv:2306.05685.