UpstreamBench: A 7,440-Question LLM Benchmark for Expert Upstream Petroleum and Subsurface Reasoning

Rong Lu *

Posted Date: 22 June 2026

doi: 10.20944/preprints202606.1410.v1

Keywords: large language models; LLM benchmark; multiple-choice question answering; petroleum engineering; geoscience; subsurface reasoning; domain-specific evaluation

 doi:10.20944/preprints202606.1410.v1

Article

UpstreamBench: A 7,440-Question LLM Benchmark for Expert Upstream Petroleum and Subsurface Reasoning

Rong Lu

Independent Researcher, USA; rlu@mines.edu

Abstract

UpstreamBench v0.1 is an expert-style multiple-choice benchmark for evaluating LLMs on upstream petroleum engineering, geoscience, geophysics, petrophysics, drilling, completions, production, geomechanics, and adjacent subsurface workflows. The benchmark contains exactly 7,440 source-linked questions: 20 mechanically validated accepted items from each of 372 technical reference books. UpstreamBench scales the earlier FormationEval benchmark and complements PetroBench with broader book-level coverage, per-source answer balancing, line-level evidence traceability, and an exact nine-model locally hosted evaluation. Its design follows the *Citius, Altius, Fortius* principle of a faster reproducible pipeline, higher full-corpus coverage, and stronger validation and failure accounting.

Keywords: large language models; LLM benchmark; multiple-choice question answering; petroleum engineering; geoscience; subsurface reasoning; domain-specific evaluation

---

1. Introduction

Large language models are now deployed as engineering assistants in petroleum and subsurface work, but most public benchmarks are either broad academic exams or small domain-specific datasets. Upstream petroleum work is unusually sensitive to mechanism reasoning: a plausible answer can be wrong because it violates a pressure-regime assumption, ignores a measurement limitation, confuses seismic acquisition geometry with processing geometry, or applies a rule of thumb outside its operating envelope.

UpstreamBench v0.1 is built to stress those distinctions. It contains 7,440 four-choice questions generated from 372 technical reference books. Every source file contributes exactly 20 accepted questions, and every complete source has a 5/5/5/5 A/B/C/D answer-key distribution. Each item includes source identifiers, SHA-256 hashes, logical source URIs, line spans, evidence hashes, rationale text, domain labels, question-type labels, difficulty labels, and contamination-risk metadata.

The paper's organizing frame is *Citius, Altius, Fortius*. Faster means a reproducible, cache-backed build and evaluation pipeline that can run on locally hosted models without paid external API calls. Higher means full coverage of a broad upstream and subsurface corpus rather than a small handpicked subset. Stronger means explicit validation gates, model-roster enforcement, red/green/blue audit checks, duplicate-content disclosure, and reporting of service errors, parse failures, and macro-by-book accuracy rather than only a single micro-average.

The contributions are:

1. A 7,440-item upstream petroleum and subsurface MCQ benchmark generated from 372 technical reference books, with exactly 20 mechanically validated accepted questions per source.
2. A deterministic validation pipeline for schema, source coverage, answer balance, evidence references, standalone wording, domain labels, question types, rejected-item fingerprints, and manifest consistency.
3. A locked nine-model locally hosted evaluation protocol with per-item response caching, strict answer extraction, macro-by-book metrics, latency reporting, and failure diagnostics.

 doi:10.20944/preprints202606.1410.v1

4. A transparent release posture for private or copyrighted sources: source chunks and source paths are not redistributed, while source ids, hashes, line references, question metadata, and aggregate results remain auditable by the author.

2. Related Work

General benchmarks such as MMLU [3], MMLU-Pro [4], GPQA [5], ARC [6], and SciBench [7] measure broad academic or scientific competence, but they do not isolate upstream petroleum tasks. Domain-specific benchmarks such as MedQA [8] in medicine and LegalBench [9] in law established field-specific evaluation where correctness depends on professional context.

In geoscience and petroleum, FormationEval [1] introduced a 505-question petroleum geoscience MCQ benchmark with source metadata and contamination-risk labels. PetroBench [2] separately targets petroleum engineering with 1,200 questions across production, reservoir, and drilling engineering. UpstreamBench differs in scale and coverage: it is built from 372 source books, covers upstream and adjacent subsurface disciplines, and uses per-book balancing so that every source contributes equally to macro-by-book accuracy.

Geoscience language modeling and evaluation work such as K2 [10], GeoGalactica [11], and surveys of generative AI in geoscience [12] motivate domain-aware evaluation beyond generic STEM performance. Automatic multiple-choice question generation has a long history [13], and recent work studies distractor generation and iterative refinement [14,15]. UpstreamBench uses LLM-assisted generation, but the accepted artifact is governed by deterministic validators, source evidence pointers, manual rejection fingerprints, and a fixed evaluation allowlist.

3. Benchmark Design

3.1. Scope

UpstreamBench v0.1 covers upstream and adjacent subsurface domains: geophysics and seismology, reservoir engineering, petrophysics, geomechanics, petroleum geology and stratigraphy, drilling and completions, production and facilities, geochemistry and mineralogy, economics and safety, near-surface and marine methods, and machine-learning/data workflows.

3.2. Task Format

The primary task is four-choice multiple choice with exactly one best answer. MCQ is chosen for local, auditable scoring, cheap evaluation across many models, and reliable single-letter answer extraction. Each complete source book contributes exactly 20 accepted questions, with five correct answers in each of A, B, C, and D.

Items are designed to be harder than direct definitions. The generator is instructed to prefer mechanism reasoning, diagnosis, workflow design, failure modes, comparative reasoning, calculations, assumption checks, edge cases, and counterfactual judgments. Trivial source copying is disallowed; questions must be standalone and concept-based.

3.3. Schema

Each item includes:

- `id`, `version`, `question_format`, and `language`
- `domains`, `topics`, `difficulty`, and `question_type`
- `question`, four `choices`, `answer_index`, and `answer_key`
- `rationale`, `requires_calculation`, `trap_type`, and `contamination_risk`
- `source_book` with source id, title, logical source URI, rights status, corpus category, and SHA-256
- `evidence_refs` with evidence id, line span, heading, page hint when available, and excerpt hash

 doi:10.20944/preprints202606.1410.v1

4. Corpus and Construction Pipeline

4.1. Corpus

The source corpus snapshot contains 372 technical reference books totaling 389 MB in the benchmark build (Table 1). Source paths are intentionally not published; public artifacts use source ids, logical `source://` URIs, source hashes, and evidence hashes. The 372 source books are not added as 372 bibliography entries because many are restricted or rights-uncleared references; the manifest is the appropriate hash-based inventory. The manifest documents four exact duplicate-content groups. Duplicates are kept because the benchmark requirement is per source file, but duplicate groups are disclosed so that downstream analyses can choose file-level or content-deduplicated views.

Table 1. Frozen corpus and dataset status.

Metric	Value
Technical source books	372
Source corpus size	389 MB
Accepted MCQ items	7,440
Questions per source file	20
Source files with accepted items	372
Answer-key count per letter	1,860
Exact duplicate-content groups	4

4.2. Generation

The builder scans every technical source, removes low-value sections such as tables of contents, references, glossaries, deliverables, exercise-answer blocks, biographies, navigation fragments, and approximate trend statements that cannot support precise calculation questions. It then extracts paragraph-level evidence packets across the full line range of each book. Packets store line spans, headings, page hints when available, compact excerpts, and excerpt hashes.

Question generation uses the approved local generator `qwen-3-coder` [16]. Raw model output is cached by source. A normalizer converts generated JSON into the benchmark schema, checks the four-choice structure, rejects self-correction artifacts, normalizes labels, assigns source evidence references, and deterministically moves the correct answer into the target answer slot for that item index. This yields per-book answer balance without changing which answer text is correct.

4.3. Validation

The strict validator checks unique ids, benchmark version, MCQ format, English language, standalone wording, four distinct choices, answer-index/key consistency, valid domains, valid difficulty labels, valid question types, rationale length, self-correction artifacts, normalized trap types, contamination-risk labels, calculation flags, logical source metadata, source SHA-256 fields, non-empty evidence references, low-value evidence headings, generator roster membership, source coverage, per-source item counts, per-source answer balance, unknown sources, rejected-item fingerprints, and manifest consistency.

The frozen dataset passes strict validation with no missing, underfilled, overfilled, imbalanced, malformed, or rejected rows. The final validation statistics report 7,440 items, 372 covered sources, zero missing sources, zero underfilled sources, zero overfilled sources, zero imbalanced sources, and zero validation issues. Table 2 and Figure 1 summarize the difficulty, question-type, and domain composition of the frozen dataset.

 doi:10.20944/preprints202606.1410.v1

Table 2. Dataset composition. Domain counts are non-exclusive because an item may carry two domain labels.

Difficulty	Count	Question type	Count	Domain	Count
Medium	519	Mechanism	3,976	Geophysics and seismology	3,915
Hard	4,296	Diagnosis	1,371	Reservoir engineering	1,582
Expert	2,625	Interpretation	819	Petrophysics	1,158
		Workflow design	321	Geomechanics	969
		Comparative reasoning	259	Drilling and completions	839
		Calculation	242	Petroleum geology and stratigraphy	834
		Assumption check	202	Geochemistry and mineralogy	596
		Failure mode	128	Production and facilities	530
		Edge case	63	Economics, management, and safety	238
		Counterfactual	59	Near-surface, environmental, and marine	202
				Machine learning and data methods	153

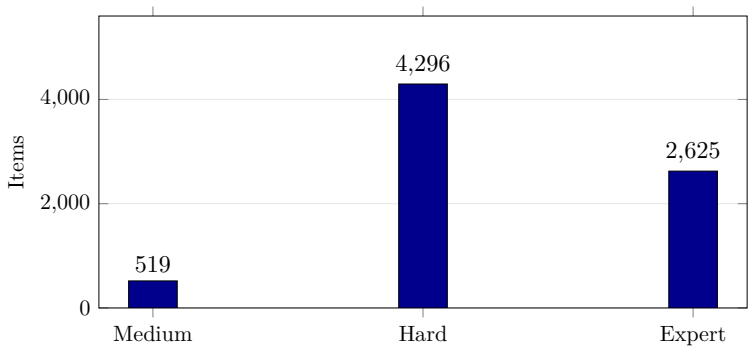

Figure 1. Difficulty distribution for the frozen 7,440-item dataset. The benchmark is dominated by hard and expert items rather than direct-definition questions.

5. Evaluation Protocol

5.1. Model Allowlist

Evaluation is reported for the exact nine locally hosted model IDs in Table 3. The reported set is intentionally narrow: it is the author's local no-paid-API run, not a broad public model leaderboard. The `qwen-3.6-27b` candidate had pilot and partial full-run artifacts, but it did not reach validated full-cache completeness after the scope narrowed; no full-suite score is reported for it.

Table 3. Frozen UpstreamBench local model set reported in the paper. No other model is used in the final nine-model evaluation.

Model id	Hosted family
`qwen-3-coder`	Qwen/Qwen3-Coder-Next
`qwen-3-vl-235b-fp8`	Qwen/Qwen3-VL-235B-A22B-Instruct-FP8
`qwen-3-vl-30b`	Qwen/Qwen3-VL-30B-A3B-Instruct
`qwen-3-vl-8b`	Qwen/Qwen3-VL-8B-Instruct
`qwen-3-vl-4b`	Qwen/Qwen3-VL-4B-Instruct
`qwen-3.6-35b`	Qwen/Qwen3.6-35B-A3B
`gpt-oss-120b`	openai/gpt-oss-120b
`gemma-4-31b`	google/gemma-4-31B-it
`minimax-m2.7`	MiniMaxAI/MiniMax-M2.7

 doi:10.20944/preprints202606.1410.v1

5.2. Prompting and Metrics

Models receive a zero-shot MCQ prompt and must answer in the format `ANSWER: <letter>`. The system message identifies the task as a difficult multiple-choice exam in petroleum geoscience, upstream engineering, and subsurface science. The user message includes domain labels, difficulty, the question, choices A–D, and the required output format.

Answer extraction first removes visible `<think>` or `<thinking>` blocks and then searches for explicit answer patterns such as `ANSWER: A`, `ANSWER IS A`, `FINAL ANSWER: A`, bare letters, leading letters, and trailing letters. Failed extraction, timeout, null visible content, or service error counts as incorrect. Reports include micro accuracy, macro-by-book accuracy, domain and difficulty breakdowns, parse failures, visible-empty counts, completion-budget exhaustion counts, service-error counts, finish reasons, extraction patterns, and mean latency.

The final run uses temperature 0, maximum 2,048 completion tokens, a 75-second client timeout, no OpenAI SDK retries, and a per-model concurrency of 8. Per-item responses are cached and final result files are reconstructed from cache only after every model has a valid cache record for every frozen benchmark row.

6. Results

Within this exact nine-model locally hosted run, the evaluation covers 7,440 questions with temperature 0.0, 2,048 maximum completion tokens, a 75-second timeout, no SDK retries, and per-model concurrency 8. Scores are reconstructed from validated per-question cache records.

Within this run, `qwen-3-vl-235b-fp8` leads at 84.70%, with `gemma-4-31b` close behind at 84.13%. The first-to-last spread is 59.07%. The largest service-error count is for `qwen-3.6-35b` (4,876 errors), while the largest visible-empty count is for `minimax-m2.7` (2,060 empty responses).

Because each source contributes exactly 20 items, macro-by-book accuracy is close to micro accuracy by construction; this does not remove duplicate-content or domain-composition effects. Operational failures are nevertheless material for slower reasoning-heavy models: visible-empty responses, length finishes, and service errors lower their practical benchmark scores even when some returned answers are correct.

Figure 2 plots the micro-accuracy ranking; Table 5 gives the per-difficulty breakdown.

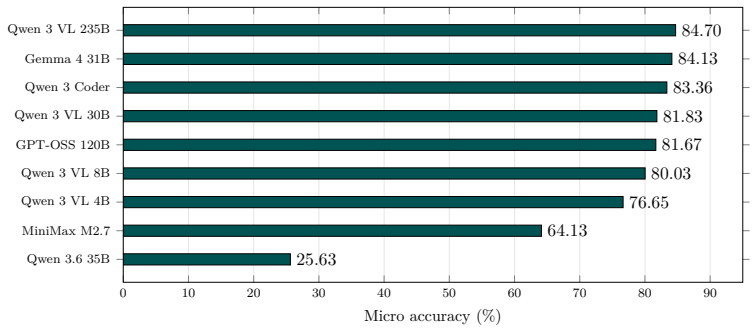

Figure 2. Micro accuracy in the exact local nine-model run. Exact model ids and failure counts are in Table 4.

 doi:10.20944/preprints202606.1410.v1

Table 4. UpstreamBench exact local evaluation results. Parse failures, empty responses, length finishes, and service errors are counted as incorrect.

Rank	Model	Micro	Macro	Correct	Parse	Empty	Length	Errors	Mean (s)
1	qwen-3-vl-235b-fp8	84.70%	84.70%	6,302/7,440	31	0	0	31	0.95
2	gemma-4-31b	84.13%	84.13%	6,259/7,440	26	0	0	26	6.74
3	qwen-3-coder	83.36%	83.36%	6,202/7,440	86	0	87	0	0.76
4	qwen-3-vl-30b	81.83%	81.83%	6,088/7,440	94	0	96	0	1.30
5	gpt-oss-120b	81.67%	81.67%	6,076/7,440	50	50	50	0	4.69
6	qwen-3-vl-8b	80.03%	80.03%	5,954/7,440	45	0	16	30	1.05
7	qwen-3-vl-4b	76.65%	76.65%	5,703/7,440	44	0	45	0	0.79
8	minimax-m2.7	64.13%	64.13%	4,771/7,440	2,144	2,060	2,066	77	22.49
9	qwen-3.6-35b	25.63%	25.63%	1,907/7,440	5,322	429	446	4,876	45.76

Table 5. Accuracy by difficulty. Difficulty labels are assigned during benchmark generation and checked by the validator.

Model	Medium	Hard	Expert
qwen-3-vl-235b-fp8	86.71%	84.19%	85.14%
gemma-4-31b	85.93%	84.31%	83.47%
qwen-3-coder	83.43%	83.12%	83.73%
qwen-3-vl-30b	80.92%	81.63%	82.32%
gpt-oss-120b	82.85%	81.40%	81.87%
qwen-3-vl-8b	80.54%	79.31%	81.10%
qwen-3-vl-4b	76.11%	76.56%	76.91%
minimax-m2.7	74.37%	64.46%	61.56%
qwen-3.6-35b	47.01%	23.35%	25.14%

7. Audit and Reproducibility

The main public release artifacts are:

- `data/benchmark/upstreambench_v0.1.jsonl`: frozen 7,440-row benchmark.
- `data/benchmark/upstreambench_v0.1_manifest.json`: source coverage, duplicate groups, local model roster, and generation metadata.
- `data/working/upstreambench/validation_stats.json`: strict validation output.
- `data/working/upstreambench/model_roster_verification.json`: sanitized local model-roster snapshot.
- `eval/results/upstreambench_runs.json`: canonical exact nine-model result summary rebuilt from validated private local cache records.
- `eval/results/upstreambench_leaderboard.md`: generated local evaluation table.

Private raw-response caches are retained locally for author-side reconstruction only; they are not redistributed in public artifacts.

For final paper quality control, the build is checked by three audit perspectives. The red check looks for overclaims, outdated draft language, missing caveats, non-reproducible claims, and leakage of private source text. The green check verifies positive claims against frozen artifacts: row counts, source counts, answer balance, model membership, cache completeness, and compile success. The blue check stresses interpretation: whether results distinguish model knowledge from serving availability, whether generator/evaluator overlap is disclosed, whether duplicates are handled honestly, and whether latency/failure metrics are reported alongside accuracy.

8. Limitations

The source corpus includes private or uncleared copyrighted material. UpstreamBench does not redistribute source chunks, long excerpts, source files, or source paths. Public rows use logical source identifiers, source hashes, and evidence hashes; broader redistribution decisions should remain subject to rights review. Evidence references use hashes and line spans for author-side audit.

The questions are LLM-assisted generations from `qwen-3-coder`. Although strict validators and manual rejection fingerprints remove many mechanical defects, they do not replace expert human

 doi:10.20944/preprints202606.1410.v1

review of every item. Since `qwen-3-coder` is also one evaluated model, its score must be interpreted with generator/evaluator overlap in mind.

The model set is intentionally narrow: exactly nine locally hosted no-paid-API models. This makes the study cost-controlled and reproducible for the author. Counting service timeouts, null content, and completion-budget exhaustion as incorrect reflects practical usability, but it also means some score gaps mix model ability with serving reliability under the fixed protocol.

The manifest contains four exact duplicate-content groups. They are retained to satisfy the per-file coverage requirement, but content-deduplicated analyses should account for them. Finally, contamination cannot be proven absent for modern models; UpstreamBench reports contamination-risk labels rather than claiming zero overlap with pretraining data.

9. Conclusion

UpstreamBench is a 7,440-item, 372-source benchmark for upstream petroleum and subsurface reasoning. It scales FormationEval from a compact public MCQ benchmark into a broader full-corpus benchmark with per-source balancing, line-level evidence metadata, strict validation, and an exact nine-model local evaluation snapshot.

Appendix A. Representative Items and Capability Splits

Appendix A.1. Complete Example Items

1. **Calculation: gas-processing heat duty.** Question: A gas-processing heat duty is 2.5 MJ/s. Using 1 kW = 1.341 hp, what horsepower is equivalent to that duty? Choices: A. About 2,500 hp; B. About 1,860 hp; C. About 3,100 hp; D. About 3,350 hp. Ground truth: D. Rationale: 2.5 MJ/s is 2,500 kW, and $2{,}500 \times 1.341 \approx 3{,}350$ hp. Provenance: source id prefix `handbook-of-natural-gas-...-afb3b220`, evidence hash prefix `c2bc3ed34326`, lines 14972–14977.
2. **Assumption check: P&ID documentation.** Question: Which item must be excluded from a P&ID per documentation criteria? Choices: A. Relief valve set pressure; B. Controller action, for example direct or reverse acting; C. Analyzer tubing size and specification; D. Local hand switch tags on control panels. Ground truth: B. Rationale: Controller actions, set points, and configuration details are excluded, while operational parameters such as set pressure and tubing specifications must be shown. Provenance: source id prefix `pid-doc-...-aa1ccb02`, evidence hash prefix `b3f4ea868c4f`, lines 874–879.
3. **Failure mode: sulfide stress cracking.** Question: Which failure mode is characteristic of sulfide stress cracking in sour environments? Choices: A. Ductile deformation with necking and elongation before fracture; B. Intergranular cracking occurring only at elevated temperatures above 150 °C; C. Brittle fracture with minimal plastic deformation, high crack propagation velocity, and transcrystalline cracking; D. Fatigue failure initiated by cyclic pressure fluctuations in the tubing string. Ground truth: C. Rationale: Sulfide stress cracking manifests as brittle fracture with high crack velocity and little visible warning under tensile stress in H_2S environments. Provenance: source id prefix `advanced-well-completion-...-c4f8e8a1`, evidence hash prefix `8847f8abf732`, lines 14302–14313.
4. **Geophysics: thin-bed seismic resolution.** Question: In seismic attribute analysis for channel detection, why might a 20-Hz time-frequency slice derived from CWT processing be insufficient to fully resolve the vertical extent and internal architecture of a thin-bedded channel system? Choices: A. Fixed-frequency slices cannot distinguish lateral facies changes from vertical stratigraphy; B. CWT suppresses high-frequency components needed for thin-bed tuning; C. Phase distortion obscures bed boundaries; D. 20 Hz is below the tuning frequency for typical thin beds, causing constructive interference that masks vertical resolution. Ground truth: D. Rationale: The dominant wavelength at 20 Hz can exceed thin-bed thickness, so reflections from top and base

 doi:10.20944/preprints202606.1410.v1

interfere and prevent vertical resolution of the channel architecture. Provenance: source id prefix `seismic-attributes-...-dca26ca5`, evidence hash prefix `8eee3fcfb4ec`, line 2963.

5. **Reservoir engineering: mobility anisotropy.** Question: During a formation test, a near probe records horizontal mobility of 490 md/cp and vertical mobility of 13 md/cp, with a skin factor of 3.0. What does this mobility contrast most strongly suggest? Choices: A. Severe near-wellbore damage affecting vertical flow more than horizontal; B. Conductive horizontal fractures enhancing lateral flow; C. Significant vertical compartmentalization with limited interlayer connectivity; D. Anisotropic permeability with $k_h/k_v \approx 38$, consistent with laminated sandstone. Ground truth: C. Rationale: The large horizontal-to-vertical mobility contrast indicates strong vertical flow resistance and limited vertical communication between layers despite lateral connectivity. Choice D cites the correct k_h/k_v ratio but attributes it to grain-scale lamination, whereas the suppressed vertical mobility most directly indicates interlayer flow barriers, hence C. Provenance: source id prefix `formation-testing-fluid-analysis-...-3c578615`, evidence hash prefix `f3cf68a68dce`, line 477.
6. **Structural geology: fault-bounded trap spillpoint.** Question: In structural interpretation across faults, what does the spillpoint of a fault-bounded trap represent? Choices: A. The shallowest depth at which hydrocarbons can migrate across the fault plane via juxtaposition of permeable layers; B. The deepest point of structural closure on the downthrown side; C. The point of maximum fault seal capacity; D. The elevation where the fault plane intersects the top reservoir horizon on the upthrown side. Ground truth: A. Rationale: The spillpoint is the shallowest level at which hydrocarbons can leak or migrate across the fault, controlled by reservoir juxtaposition across the fault plane. Provenance: source id prefix `3-d-seismic-interpretation-...-aedee953`, evidence hash prefix `5dd9458c2bc3`, line 1177.

Appendix A.2. Valid-Output Capability Splits

The following examples compare models that both returned valid `ANSWER:` outputs; the split therefore reflects answer choice rather than parse failure or empty content.

- **Coiled tubing RIH precaution in high-pressure zones.** Ground truth: D. VL-235B answered D correctly, while Coder answered C. The split distinguishes surge-pressure formation-fracture risk from swab-kick control.
- **Layered-earth integral-equation geometry limit.** Ground truth: B. VL-30B answered B correctly, while Coder answered A. The split distinguishes lateral homogeneity assumptions from the lack of Green's functions for non-parallel interfaces.

References

1. Ermilov, A. FormationEval, an open multiple-choice benchmark for petroleum geoscience. *arXiv preprint arXiv:2601.02158* **2026**. https://doi.org/10.48550/arXiv.2601.02158.
2. Wang, X.; Zhang, T.; Wang, S.; Wu, Y.; Meng, H.; Zhou, P.; Li, P. PetroBench: A Benchmark for Large Language Models in Petroleum Engineering. *arXiv preprint arXiv:2605.28032* **2026**. https://doi.org/10.48550/arXiv.2605.28032.
3. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring Massive Multitask Language Understanding. In Proceedings of the International Conference on Learning Representations (ICLR), 2021. https://doi.org/10.48550/arXiv.2009.03300.
4. Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2024. https://doi.org/10.48550/arXiv.2406.01574.
5. Rein, D.; Hou, B.L.; Stickland, A.C.; Petty, J.; Pang, R.Y.; Dirani, J.; Michael, J.; Bowman, S.R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In Proceedings of the Conference on Language Modeling (COLM), 2024. https://doi.org/10.48550/arXiv.2311.12022.

 doi:10.20944/preprints202606.1410.v1

6. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457* **2018**. https://doi.org/10.48550/arXiv.1803.05457.
7. Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A.R.; Zhang, S.; Sun, Y.; Wang, W. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In Proceedings of the International Conference on Machine Learning (ICML), 2024. https://doi.org/10.48550/arXiv.2307.10635.
8. Jin, D.; Pan, E.; Oufattole, N.; Weng, W.H.; Fang, H.; Szolovits, P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences* **2021**, *11*, 6421. https://doi.org/10.3390/app11146421.
9. Guha, N.; Nyarko, J.; Ho, D.E.; Ré, C.; Chilton, A.; et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2023. https://doi.org/10.48550/arXiv.2308.11462.
10. Deng, C.; Zhang, T.; He, Z.; Xu, Y.; Chen, Q.; Shi, Y.; Fu, L.; Zhang, W.; Wang, X.; Zhou, C.; et al. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. In Proceedings of the Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM), 2024, pp. 161–170. https://doi.org/10.1145/3616855.3635772.
11. Lin, Z.; Deng, C.; Zhou, L.; Zhang, T.; Xu, Y.; Xu, Y.; He, Z.; Shi, Y.; Dai, B.; Song, Y.; et al. GeoGalactica: A Scientific Large Language Model in Geoscience. *arXiv preprint arXiv:2401.00434* **2024**. https://doi.org/10.48550/arXiv.2401.00434.
12. Hadid, A.; Chakraborty, T.; Busby, D. When Geoscience Meets Generative AI and Large Language Models: Foundations, Trends, and Future Challenges. *Expert Systems* **2024**, *41*, e13654. https://doi.org/10.1111/exsy.13654.
13. Ch, D.R.; Saha, S.K. Automatic Multiple Choice Question Generation from Text: A Survey. *IEEE Transactions on Learning Technologies* **2020**, *13*, 14–25. https://doi.org/10.1109/TLT.2018.2889100.
14. Alhazmi, E.; Sheng, Q.Z.; Zhang, W.E.; Zaib, M.; Alhazmi, A. Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2024, pp. 14437–14458. https://doi.org/10.18653/v1/2024.emnlp-main.799.
15. Yao, Z.; Parashar, A.; Zhou, H.; Jang, W.S.; Ouyang, F.; Yang, Z.; Yu, H. MCQG-SRefine: Multiple Choice Question Generation and Evaluation with Iterative Self-Critique, Correction, and Comparison Feedback. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Association for Computational Linguistics, 2025, pp. 10728–10777. https://doi.org/10.18653/v1/2025.naacl-long.538.
16. Cao, R.; et al. Qwen3-Coder-Next Technical Report. *arXiv preprint arXiv:2603.00729* **2026**. https://doi.org/10.48550/arXiv.2603.00729.