FormationEval, an open multiple-choice benchmark for petroleum geoscience

Almaz Ermilov
UiT The Arctic University of Norway
almaz.ermilov@gmail.com

Preprint

Abstract

This paper presents FormationEval, an open multiple-choice question benchmark for evaluating language models on petroleum geoscience and subsurface disciplines. The dataset contains 505 questions across seven domains including petrophysics, petroleum geology and reservoir engineering, derived from three authoritative sources using a reasoning model with detailed instructions and a concept-based approach that avoids verbatim copying of copyrighted text. Each question includes source metadata to support traceability and audit. The evaluation covers 72 models from major providers including OpenAI, Anthropic, Google, Meta and open-weight alternatives. The top performers achieve over 97% accuracy, with Gemini 3 Pro Preview reaching 99.8%, while tier and domain gaps persist. Among open-weight models, GLM-4.7 leads at 98.6%, with several DeepSeek, Llama, Qwen and Mistral models also exceeding 93%. The performance gap between open-weight and closed models is narrower than expected, with several lower-cost open-weight models exceeding 90% accuracy. Petrophysics emerges as the most challenging domain across all models, while smaller models show wider performance variance. Residual length bias in the dataset (correct answers tend to be longer) is documented along with bias mitigation strategies applied during construction. The benchmark, evaluation code and results are publicly available.

1 Introduction

Large language models (LLMs) are increasingly applied to domain-specific tasks in science and engineering, yet their capabilities in specialized fields remain difficult to assess. General benchmarks like MMLU [28] cover broad knowledge but offer limited focus on specialized fields. For petroleum geoscience and subsurface engineering (fields requiring understanding of well logging physics, reservoir characterization and geological interpretation), publicly available benchmarks remain limited.

This work addresses the gap with FormationEval, a 505-question multiple-choice benchmark covering seven domains: petrophysics, petroleum geology, geophysics, reservoir engineering, sedimentology, drilling engineering and production engineering. Questions are derived from authoritative textbooks and open courseware using a concept-based methodology that tests understanding rather than phrase recognition, while respecting copyright constraints.

The contributions are: (1) a methodology for generating multiple-choice questions (MCQs) from technical sources without verbatim copying; (2) a curated dataset with source metadata and contamination risk labels; and (3) an evaluation of 72 language models across multiple providers, revealing performance patterns by domain and difficulty level.

2 Related work

General-purpose benchmarks like MMLU [28] evaluate broad knowledge across 57 subjects but provide limited coverage of specialized domains. MMLU-Pro [52] addresses some limitations with harder reasoning questions yet remains domain-general. For science, ARC [9] covers grade-school science questions and SciBench [51] targets college-level problem solving in physics, chemistry and mathematics. GPQA [49] provides 448 graduate-level questions where even domain experts achieve only 65% accuracy. Domain-specific benchmarks exist for medicine (MedQA [29], 12,723 questions from medical licensing exams) and law (LegalBench [26]). However, no public MCQ benchmark exists specifically for petroleum geoscience or subsurface engineering disciplines.

Recent work on large language models for geoscience has produced domain-adapted models such as K2 [13], which further pre-trains LLaMA on 5.5 billion tokens of geoscience text and introduces GeoBench for evaluation, and GeoGalactica [30], a 30-billion parameter model trained on 65 billion tokens of geoscience literature. Hadid et al. [27] survey generative AI applications across geoscience disciplines. These efforts demonstrate growing interest in AI for earth sciences but focus on model development rather than standardized evaluation of existing models across petroleum-specific knowledge areas.

Automatic MCQ generation from text is a well-studied problem. Ch and Saha [8] survey methods spanning sentence selection, key extraction, question formation and distractor generation. Recent LLM-based approaches include MCQG-SRefine [54], which uses iterative self-critique to generate medical exam questions, and various distractor generation techniques surveyed by Alhazmi et al. [2]. Most existing methods transform source sentences into questions through syntactic manipulation or paraphrasing. The concept-based methodology presented here differs by generating questions from extracted concepts rather than from specific sentences, which avoids close paraphrasing of copyrighted text while testing understanding rather than phrase recognition.

3 Benchmark design and construction

3.1 Task definition and scope

FormationEval uses a four-choice multiple-choice question format with exactly one correct answer per question. This format is compatible with standard evaluation frameworks and enables straightforward accuracy computation.

Questions cover seven domains: Petrophysics (well logging, formation evaluation), Petroleum Geology (source rocks, migration, trapping), Sedimentology (depositional environments, diagenesis), Geophysics (seismic interpretation, rock physics), Reservoir Engineering (fluid flow, recovery mechanisms), Drilling Engineering (wellbore stability, operations) and Production Engineering (completions, artificial lift). Questions can belong to more than one domain.

Difficulty levels reflect educational background. Easy questions (undergraduate) test definitions and direct recall. Medium questions (graduate or professional) require applying concepts to scenarios. Hard questions (specialist) involve integrating multiple concepts or edge cases.

3.2 Source selection and licensing policy

The benchmark draws from three sources: Ellis & Singer's *Well Logging for Earth Scientists* [14] (219 questions), Bjørlykke's *Petroleum Geoscience* [7] (262 questions) and TU Delft OpenCourseWare [50] (24 questions).

The benchmark uses a concept-based derivation approach. Questions are written from scratch based on concepts extracted from source material, without copying sentences or closely paraphrasing distinctive problem structures. This respects the legal distinction between ideas (not copyrightable) and expression (protected). Standard technical terms (porosity, Archie equation, neutron-density crossplot) may appear as-is since terminology is not copyrightable.

All generated items are tagged with `derivation_mode: concept_based` and include source tracking fields that enable verification without reproducing protected text.

3.3 Schema and metadata

Each question includes required fields: unique identifier, question text, four choices, answer index (0–3), answer key (A–D), difficulty level, domains, topics and a rationale explaining the correct answer. The rationale serves dual purposes. It aids human verification during development and provides educational value to benchmark users.

Each item also includes a `contamination_risk` label indicating likelihood that similar questions exist in LLM training data: *low* for novel questions specific to the source, *medium* for common concepts where similar questions may exist and *high* for standard introductory topics almost certainly present in training data. Assessing contamination risk is hard without access to training data; this label is an estimate based on topic commonality and cannot be directly checked. Still, approximate labels help when comparing results by risk level.

Provenance metadata includes source identifier, title, chapter reference, license and retrieval date. See Appendix A for the complete schema reference.

3.4 Multiple-choice question generation pipeline

Questions are generated using a reasoning model (GPT-5.2 [43] with extra high reasoning effort) that processes source chapters (Section 3.2) through extended chain-of-thought before producing output. The pipeline consists of four stages:

1. **Text extraction** converts source PDFs to Markdown using optical character recognition (OCR), preserving structure and mathematical notation.

2. **Chunking** splits documents by chapter or section, with each chunk sized for model context (typically one chapter, approximately 10,000–15,000 tokens).

3. **Candidate generation** uses GPT-5.2 with extra high reasoning effort, which receives the chapter text along with a detailed system prompt specifying schema requirements, concept-based derivation rules, difficulty targets and output format. The model generates 5–12 questions per chapter.

4. **Verification** checks schema compliance (no duplicate choices, answer index in range) and confirms support in source text, flagging ambiguous items.

Figure 1 illustrates this pipeline from source ingestion to final dataset.

The system prompt emphasizes that questions must be standalone, answerable from domain knowledge without access to the source chapter. Phrases like “according to the chapter” or “the text describes” are explicitly prohibited. A summary of the generation prompt is provided in Appendix B; the full system prompt is available in the repository at `src/prompts/mcq_generator_system_prompt.txt`.

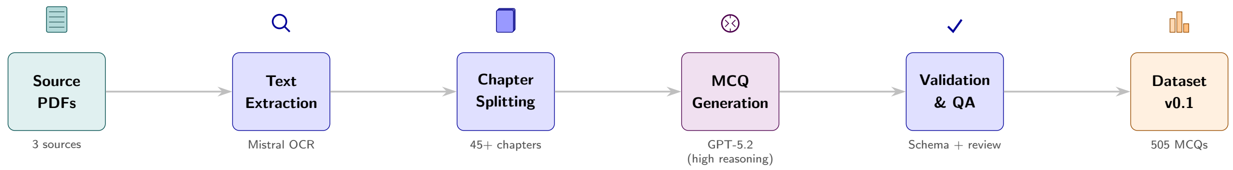

Figure 1: MCQ generation pipeline. Source PDFs are converted to Markdown via OCR, split into chapter chunks, processed by GPT-5.2 to generate candidate questions and verified for schema compliance and source evidence.

3.5 Quality assurance and audit

Human spot-checking is essential despite high LLM instruction-following reliability. All manual verification was performed by the author, a petrophysicist with background in well logging, formation evaluation and petroleum geology. For each source chapter, 2–5 questions were verified against the source material to confirm: (1) the marked answer is correct and unambiguous; (2) distractors are plausible but clearly wrong; (3) no copied phrases appear in questions or answers; (4) the rationale supports the marked answer. This domain expertise was particularly important for petrophysics questions (54% of the dataset), where assessing answer correctness requires familiarity with logging tool physics and interpretation methods.

The full dataset (505 questions across 45+ chapters) was reviewed in batches grouped by source and chapter, with each batch cross-checked against the source material. The final audit flagged one issue out of 505 questions (a question where the source text contained internally inconsistent units). All other batches passed with no issues. In addition to human spot-checks, LLM-based batch review flagged potential inconsistencies between rationales and correct answers.

Post-generation corrections addressed two categories of issues. First, 55 questions required rewriting to remove chapter self-references (phrases like “according to the chapter” or “the text describes”) that violated the standalone requirement. These were replaced with domain context so that each question is answerable from general petroleum geoscience knowledge. Second, 12 grammar corrections were applied after batch bias-mitigation edits introduced broken sentences.

The generation prompt evolved through eight major iterations over the course of the project. The list of explicit prohibitions (covering self-references, negative phrasing, length balance and qualifier word patterns) grew from 5 to 15 rules. Few-shot examples were also corrected after analysis revealed that two of three examples had the longest option as the correct answer, contradicting the length balance guidance in the prompt text.

The required rationale field accelerated verification throughout this process. When the model produced a plausible-sounding answer but could not write a coherent rationale, the answer was often incorrect or fabricated. This made the rationale field an effective early indicator of generation errors during both automated and manual review.

3.6 Bias analysis and mitigation

Initial analysis revealed two exploitable patterns.

Length bias was significant. Correct answers were uniquely longest in over 55% of questions (expected 25%), averaging 86.6 characters versus 69.8 for distractors.

Qualifier word bias was also present. Absolute words like “always” appeared only in distractors (49 instances, 0% correct rate), while hedged words like “may” appeared only in correct answers (13 instances, 100% correct rate).

Combined, these patterns could be exploited well above random chance.

Mitigation involved expanding 136 distractors with technical context to reduce length imbalance (from over 55% to 43.2% uniquely-longest-is-correct). All "always" instances were replaced with varied synonyms (invariably, necessarily, inherently) to break the single-word exploit. The word "may" was added to 13 distractors (e.g., "has no effect" → "may have no effect") to balance hedging language.

Residual issues remain. Length bias is still above the 25% baseline. The absolute-word synonyms all have 0% correct rate, making a combined "any-absolute-word=wrong" heuristic still partially exploitable.

For benchmarking purposes, these residual biases should not significantly affect relative comparisons. Since the same patterns apply to all questions, any bias-based advantage would affect all models equally, preserving the validity of model-to-model comparisons. The major exploits (length as a proxy for correctness, "always" as a marker for wrong answers, "may" as a marker for correct answers) have been addressed. After mitigation, correct answers average 87 characters versus 74 for distractors, a difference of roughly 13 characters (2–3 words). With the uniquely longest answer correct only 43.2% of the time, a length-based strategy would fail more often than it succeeds. The remaining issues with absolute-word synonyms affect fewer than 10% of questions. These limitations are documented for transparency; see Appendix C for detailed before/after metrics.

3.7 PDF export for review

For easier review, the dataset is exported to PDF with a cover page, question cards showing all metadata and bookmarks for navigation. This format is more accessible than raw JSON for domain experts performing quality checks and enables browsing questions by domain or topic without programming tools. An accompanying PDF rendering, generated from the JSON release, provides a readable format for browsing and spot checks [18, 17].

4 Dataset summary

Version 0.1 of FormationEval contains 505 questions in English covering 811 unique topics (as questions can address multiple topics). Table 1 summarizes key metrics, Table 2 shows the distribution by domain and difficulty and Table 3 provides the difficulty breakdown within each domain. Questions are distributed across three sources: Bjørlykke's textbook contributes the largest share (262 questions, 52%), followed by Ellis & Singer (219 questions, 43%) and TU Delft open courseware (24 questions, 5%). Figure 2 visualizes these distributions.

Petrophysics represents the largest domain (54% of questions) reflecting the depth of coverage in the well logging textbook. The difficulty distribution targets 30% easy, 50% medium and 20% hard; the actual distribution (26%/54%/20%) is close to these targets. Answer positions are balanced: A=27%, B=26%, C=25%, D=22%.

Petrophysics has the lowest share of easy questions (18%) and the highest share of hard questions (24%), consistent with the technical depth of well logging physics covered in Ellis & Singer [14]. Smaller domains such as Drilling Engineering and Production Engineering contain few hard questions, reflecting the limited coverage of these topics in the selected sources.

Table 1: Dataset summary (v0.1).

Metric	Value
Questions	505
Sources	3
Domains	7
Unique topics	811
Language	English

Table 2: Domain and difficulty distribution. Domain counts are non-exclusive (questions may belong to multiple domains).

Category	Count	Share
By domain		
Petrophysics	272	54%
Petroleum Geology	151	30%
Sedimentology	98	19%
Geophysics	80	16%
Reservoir Engineering	43	9%
Drilling Engineering	24	5%
Production Engineering	14	3%
By difficulty		
Easy	132	26%
Medium	274	54%
Hard	99	20%

Table 3: Difficulty distribution by domain. Domain counts are non-exclusive (questions may belong to multiple domains). Percentages show within-domain distribution.

Domain	Easy	Medium	Hard	Total
Petrophysics	49 (18%)	159 (58%)	64 (24%)	272
Petroleum Geology	50 (33%)	70 (46%)	31 (21%)	151
Sedimentology	30 (31%)	54 (55%)	14 (14%)	98
Geophysics	19 (24%)	44 (55%)	17 (21%)	80
Reservoir Engineering	13 (30%)	25 (58%)	5 (12%)	43
Drilling Engineering	10 (42%)	11 (46%)	3 (12%)	24
Production Engineering	5 (36%)	8 (57%)	1 (7%)	14

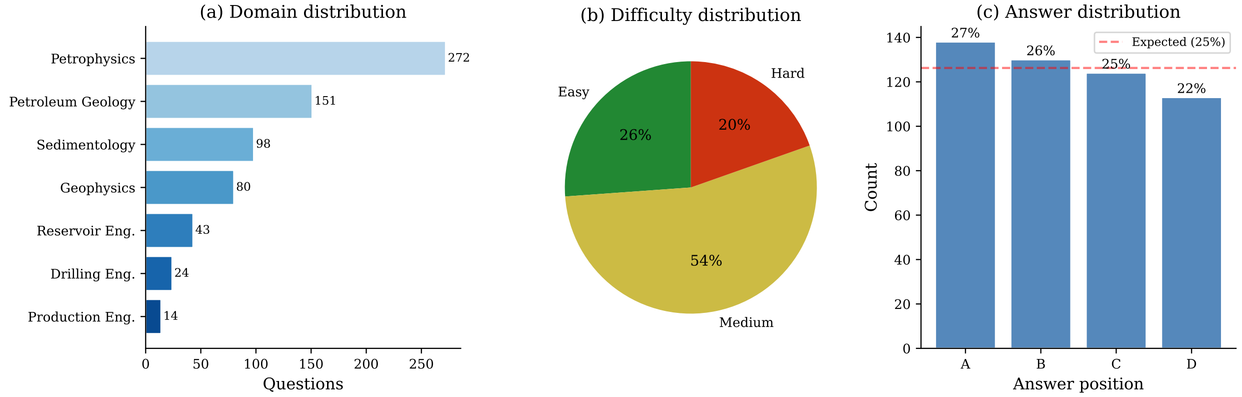

Figure 2: Dataset composition. (a) Questions by domain, with Petrophysics dominating due to source coverage; (b) difficulty distribution close to 30/50/20 targets; (c) answer position distribution near the expected 25% baseline.

5 Evaluation setup

5.1 Models and providers

The evaluation covers 72 models accessed through two API providers: Azure OpenAI [35] and OpenRouter [47]. The evaluation ran in December 2025. Models come from OpenAI [40, 41, 44, 42, 45] (GPT-4o, GPT-4.1, GPT-5 series, o3-mini, o4-mini), Anthropic [3, 4, 5, 6] (Claude 3.5 Haiku, Claude 3.7 Sonnet, Claude Opus 4.5, Claude Sonnet 4.5, Claude Haiku 4.5), Google [20, 23, 21, 24, 25] (Gemini 2.0, 2.5, 3 series, Gemma 3), Meta [31, 32, 33, 34] (Llama 3.1, 3.2, 4), DeepSeek [10, 12, 11] (R1, V3.2), Mistral [37] (Small, Medium, Nemo, Ministral), Alibaba [48] (Qwen3 series), Zhipu [22, 55] (GLM-4, GLM-4.7), xAI [46, 53] (Grok 3, 4), Moonshot [38] (Kimi K2), MiniMax [36] (M2), Microsoft [1] (Phi-4) and Nvidia [39] (Nemotron).

Models range from compact 3B parameter variants to frontier reasoning models. Open-weight models (32 of 72) include GLM-4.7, DeepSeek-R1, Llama-4-Scout, Qwen3 variants, Mistral models and Gemma 3. Pricing spans from \$0.02/M input tokens (Llama-3.2-3b-instruct) to \$25/M output tokens (Claude Opus 4.5).

5.2 Prompting and answer extraction

The evaluation uses a zero-shot prompt format to assess model knowledge without providing examples.

The **system prompt** tells the model "You are taking a multiple-choice exam on Oil & Gas geoscience. For each question, select the single best answer from the options provided. State your final answer as a single letter: A, B, C or D."

The **user prompt** presents the question text followed by four labeled choices (A–D) and "Answer:" as the final line.

Answer extraction uses flexible regex patterns to handle varied response formats. Preprocessing removes reasoning tags (`<think>`, `<thinking>`) from models like DeepSeek-R1 that expose chain-of-thought traces. The extraction logic prioritizes explicit patterns ("The answer is B", "Answer: C"). Failed extractions (where no A/B/C/D letter can be identified) are counted as incorrect answers. Figure 3 shows the evaluation flow from configuration to report generation.

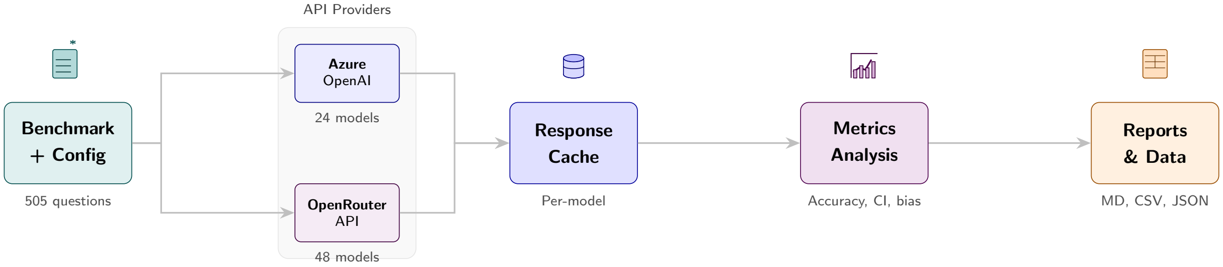

Figure 3: Evaluation pipeline. Models are configured via YAML, questions sent to Azure OpenAI or OpenRouter APIs, responses cached per model/question and analyzed to generate leaderboard and analysis reports.

5.3 Metrics

The primary metric is overall accuracy, defined as correct answers divided by total questions (505).

Secondary metrics include accuracy by difficulty level (easy, medium, hard) and by domain (seven categories). The analysis also covers bias patterns: position bias (deviation from uniform A/B/C/D selection) and length bias (tendency to select the longest answer choice). The benchmark's residual length bias (correct answer is uniquely longest in 43.2% of questions) provides a reference point for interpreting model length bias.

5.4 Caching and reproducibility

API responses are cached per model and question (`eval/cache/{model}/{question_id}.json`). This enables resuming interrupted evaluation runs, re-analyzing results without additional API costs and debugging extraction failures.

Evaluation can be re-run in analyze-only mode to regenerate reports from cached responses. Output files include machine-readable JSON, Markdown tables and CSV exports with per-question breakdowns including raw model responses (truncated to 500 characters).

6 Results

6.1 Overall results

Table 4 presents the top 50 models by accuracy [19]. The highest-performing model is Gemini 3 Pro Preview at 99.8% (504/505 correct), followed by GLM-4.7 at 98.6% and Gemini 3 Flash Preview at 98.2%. Among open-weight models, GLM-4.7 leads, followed by DeepSeek-R1 (96.2%) and DeepSeek-V3.2 (94.9%).

Accuracy spans a wide range, from 99.8% (Gemini 3 Pro Preview) to 57.6% (Llama-3.2-3b-instruct). Models from Google, OpenAI and Zhipu dominate the top positions (Figure 4). Pricing varies considerably, with some lower-cost models like Grok-4.1-fast ($0.20/M input) achieving 97.6% accuracy. Figure 5 shows the cost-effectiveness trade-off across all models.

Table 4: Top 50 models by accuracy (prices in USD per million tokens). See Appendix D for the complete 72-model leaderboard.

Rank	Model	Open	\$/M (in/out)	Accuracy	Correct
1	gemini-3-pro-preview	No	2.00/12.00	99.8%	504/505
2	glm-4.7	Yes	0.40/1.50	98.6%	498/505
3	gemini-3-flash-preview	No	0.50/3.00	98.2%	496/505
4	gemini-2.5-pro	No	1.25/10.00	97.8%	494/505
5	grok-4.1-fast	No	0.20/0.50	97.6%	493/505
6	gpt-5.2-chat-medium	No	1.75/14.00	97.4%	492/505
7	kimi-k2-thinking	No	0.40/1.75	97.2%	491/505
8	claude-opus-4.5	No	5.00/25.00	97.0%	490/505
9	gpt-5.2-chat-high	No	1.75/14.00	96.8%	489/505
10	gpt-5.2-chat-low	No	1.75/14.00	96.8%	489/505
11	gpt-5-mini-medium	No	0.25/2.00	96.4%	487/505
12	gpt-5.1-chat-medium	No	1.25/10.00	96.4%	487/505
13	deepseek-r1	Yes	0.30/1.20	96.2%	486/505
14	grok-4-fast	No	0.20/0.50	96.0%	485/505
15	gpt-5-mini-high	No	0.25/2.00	95.6%	483/505
16	gpt-5-mini-low	No	0.25/2.00	95.2%	481/505
17	o4-mini-high	No	1.10/4.40	95.2%	481/505
18	gemini-2.5-flash	No	0.30/2.50	95.0%	480/505
19	o4-mini-medium	No	1.10/4.40	95.0%	480/505
20	grok-3-mini	No	0.30/0.50	95.0%	480/505
21	deepseek-v3.2	Yes	0.22/0.32	94.9%	479/505
22	gpt-5.1-chat-low	No	1.25/10.00	94.9%	479/505
23	o3-mini-low	No	1.10/4.40	94.9%	479/505
24	o3-mini-medium	No	1.10/4.40	94.9%	479/505
25	claude-3.7-sonnet	No	3.00/15.00	94.7%	478/505
26	o3-mini-high	No	1.10/4.40	94.7%	478/505
27	gpt-5-chat	No	1.25/10.00	94.5%	477/505
28	o4-mini-low	No	1.10/4.40	94.3%	476/505
29	gpt-5.1-chat-high	No	1.25/10.00	93.9%	474/505
30	gpt-4.1	No	2.00/8.00	93.7%	473/505
31	gemini-2.0-flash-001	No	0.10/0.40	93.3%	471/505
32	gpt-5-nano-low	No	0.05/0.40	93.3%	471/505
33	llama-4-scout	Yes	0.08/0.30	93.1%	470/505
34	mistral-medium-3.1	Yes	0.40/2.00	93.1%	470/505
35	qwen3-235b-a22b-2507	Yes	0.07/0.46	93.1%	470/505
36	qwen3-30b-a3b-thinking-2507	Yes	0.05/0.34	93.1%	470/505
37	gpt-4o	No	2.50/10.00	92.9%	469/505
38	gpt-5-nano-high	No	0.05/0.40	92.9%	469/505
39	gpt-5-nano-medium	No	0.05/0.40	92.9%	469/505
40	minimax-m2	No	0.20/1.00	92.9%	469/505
41	qwen3-14b	Yes	0.05/0.22	92.9%	469/505
42	qwen3-32b	Yes	0.08/0.24	92.1%	465/505
43	gpt-4.1-mini	No	0.40/1.60	91.7%	463/505
44	claude-haiku-4.5	No	1.00/5.00	91.5%	462/505
45	gemini-2.5-flash-lite	No	0.10/0.40	91.3%	461/505
46	gpt-oss-120b	Yes	0.04/0.19	90.7%	458/505
47	qwen3-vl-8b-thinking	Yes	0.18/2.10	90.3%	456/505
48	mistral-small-3.2-24b-instruct	Yes	0.06/0.18	89.3%	451/505
49	gpt-oss-20b	Yes	0.03/0.14	89.3%	451/505
50	claude-sonnet-4.5	No	3.00/15.00	89.1%	450/505

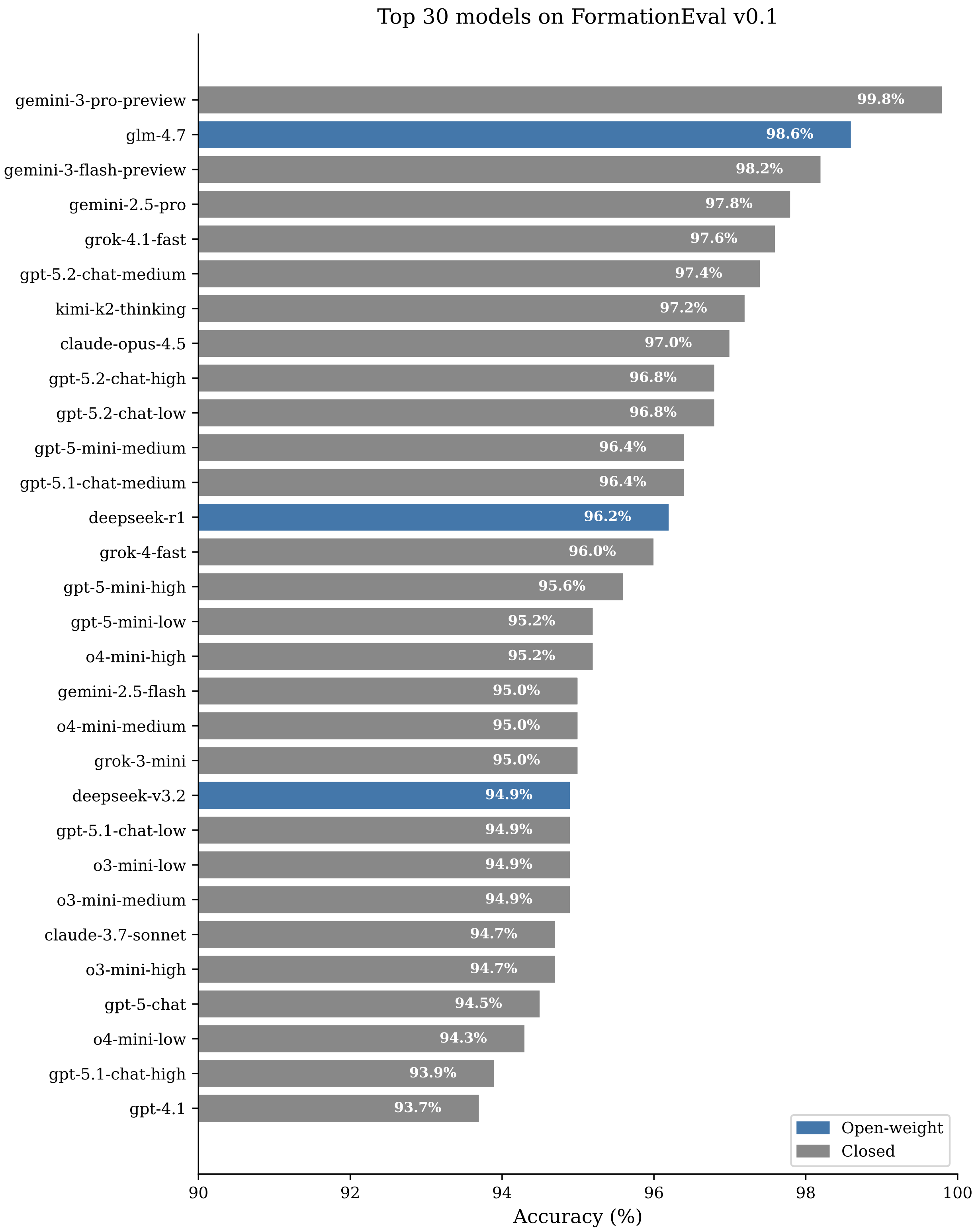

Figure 4: Top 30 models on FormationEval v0.1 by accuracy. Blue bars indicate open-weight models. GLM-4.7 (98.6%) leads among open-weight models, ranking second overall.

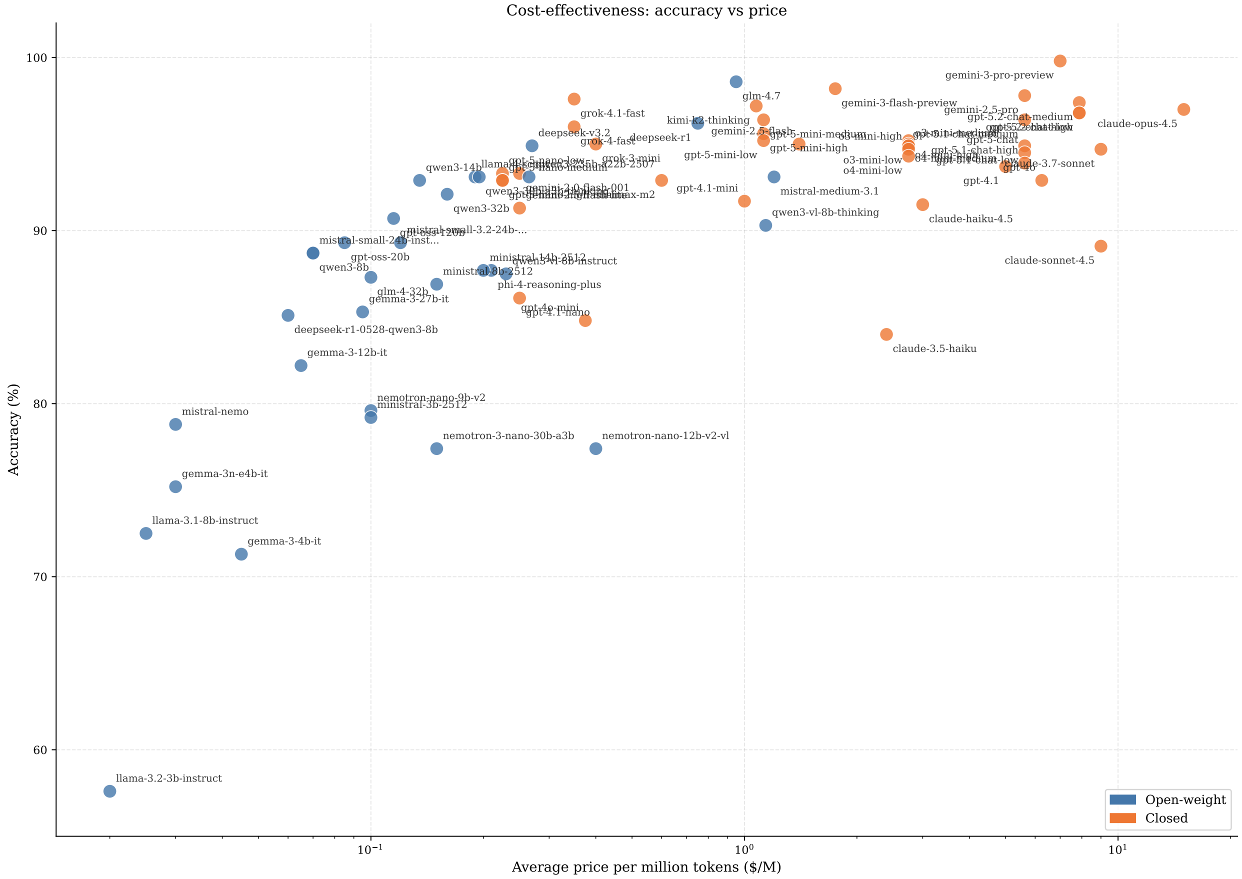

Figure 5: Cost-effectiveness analysis. Accuracy versus average token price (mean of input and output prices). Several high-accuracy models (Grok-4.1-fast, DeepSeek-R1) offer strong performance at lower cost. Open-weight models (blue) provide lower-cost alternatives to closed models (orange).

6.2 By difficulty

Across all model tiers, medium questions show the lowest accuracy, below both easy and hard (Figure 6). Even Gemini 3 Pro Preview made its single error on a medium question. Medium questions are more comparison-heavy and tool-specific, have a higher calculation rate (8.8% versus 5.1% for hard) and have less length-bias cueing than hard (correct answer is uniquely longest in 40.5% of medium versus 55.6% of hard, per the benchmark dataset [18]).

Smaller models show wider variance by difficulty. Llama-3.2-3b-instruct achieves 62.9% on easy questions but only 54.7% on medium, indicating difficulty labels correlate with model performance patterns. Figure 6 compares accuracy by difficulty across model tiers.

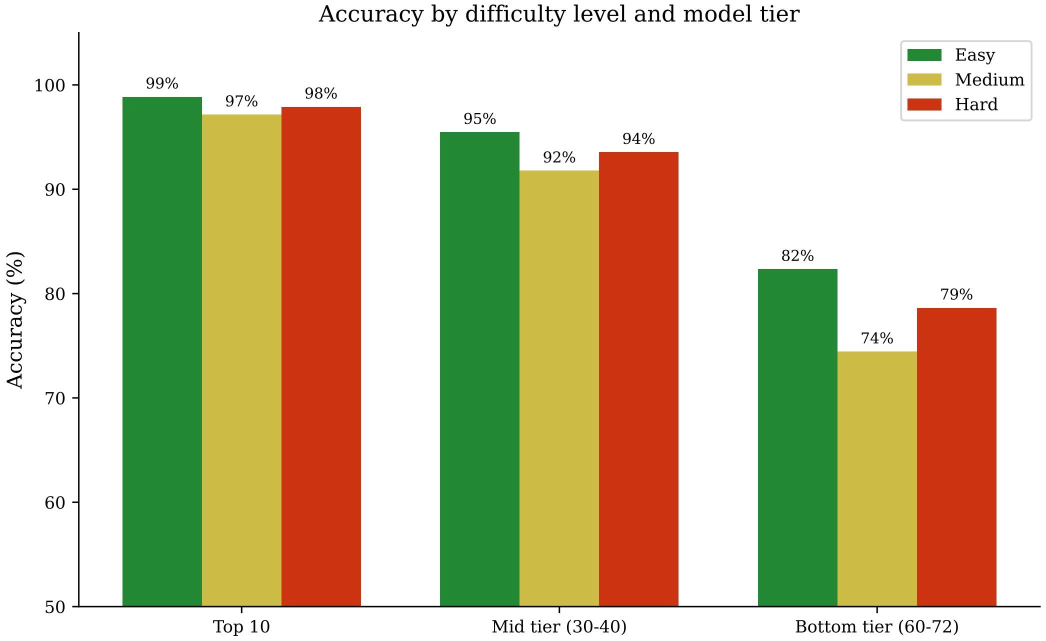

Figure 6: Accuracy by difficulty level across model tiers. Top-tier models show consistent performance across difficulty levels, while bottom-tier models exhibit larger gaps between easy and harder questions.

6.3 By domain

Petrophysics emerges as the most challenging domain across all models, with typical accuracy 3–5 percentage points lower than other domains. This reflects the technical detail of well logging physics and formation evaluation concepts.

The top model (Gemini 3 Pro Preview) achieves near-perfect scores across all domains: 99.6% Petrophysics, 100% for all other domains. Open-weight leader GLM-4.7 shows 98.2% Petrophysics versus 99–100% for other domains.

Table 5 shows average accuracy per domain across all 72 models. Reservoir Engineering leads at 95.6%, while Petrophysics is consistently lowest at 87.5%. Top models often achieve 100% on Production and Drilling (which have fewer questions: 14 and 24 respectively), but overall averages place these domains in the middle of the ranking. Figure 7 shows accuracy patterns for the top 15 models.

Table 5: Average accuracy by domain across all 72 models.

Domain	Average accuracy
Reservoir Engineering	95.6%
Petroleum Geology	93.9%
Sedimentology	93.6%
Geophysics	93.2%
Production Engineering	91.5%
Drilling Engineering	91.3%
Petrophysics	87.5%

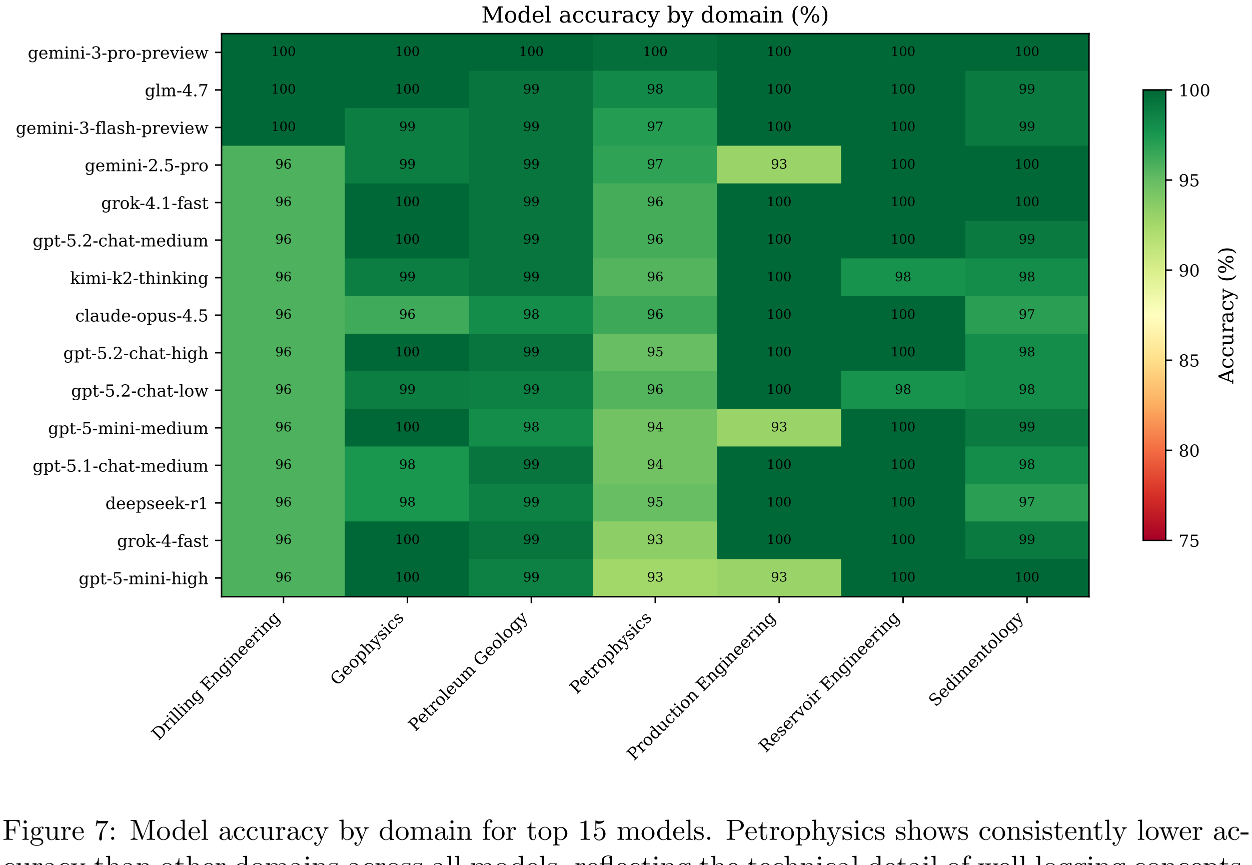

Figure 7: Model accuracy by domain for top 15 models. Petrophysics shows consistently lower accuracy than other domains across all models, reflecting the technical detail of well logging concepts.

6.4 Hardest questions

The ten hardest questions (by model failure rate) reveal systematic knowledge gaps [16]. The hardest question (on strike-slip fault stepovers and pull-apart basin formation) was answered incorrectly by 61 of 72 models (85%). Most models selected option A (left-stepping arrangement) instead of the correct answer D (right-stepping arrangement for dextral motion). The top three hardest questions are documented in the analysis report [16].

Eight of the ten hardest questions are from the Petrophysics domain, covering specialized topics like neutron-density crossplot interpretation, invasion profiles and tool calibration. One geology question (strike-slip stepovers) and one easy-labeled question (logging-while-drilling (LWD) propagation) also appear in the top 10.

Model agreement analysis shows that 21.6% of questions were answered correctly by all 72 models, 0% were missed by all models and 78.4% showed mixed results, indicating most questions help distinguish model performance levels.

6.5 Open-weight models

A key question for this evaluation was whether open-weight models, especially smaller and cheaper ones, have enough domain knowledge for petroleum geoscience. While larger closed models were expected to perform well, the capabilities of open-weight alternatives in this specialized field were less clear.

Of the 72 models evaluated, 32 have publicly available weights. Open-weight accuracy ranges from 57.6% (Llama-3.2-3b-instruct) to 98.6% (GLM-4.7), with mean 85.7% and median 87.7%. Eleven open-weight models reach at least 90% and 22 reach at least 85%. Figure 8 visualizes all open-weight models by accuracy; Table 6 provides the detailed breakdown.

The highest-accuracy open-weight models are GLM-4.7, DeepSeek-R1, DeepSeek-V3.2, Llama-4-Scout and Mistral Medium 3.1, followed by large Qwen3 variants at 93.1%. Qwen3 base variants span 88.7–93.1%, Qwen3-VL variants 87.5–90.3% and GPT-OSS 89.3–90.7%. Mistral and Ministral models span 78.8–93.1%; GLM-4-32B reaches 87.3%; Phi-4 reasoning-plus reaches 87.7%; DeepSeek-R1-0528-Qwen3-8B reaches 85.1%. DeepSeek-R1 at \$0.30/M input offers similar accuracy to GPT-5.1 variants at \$1.25/M input; Qwen3-14b reaches 92.9% at \$0.05/M input.

Across domains, open-weight averages are highest in Reservoir Engineering (93.3%) and lowest in Petrophysics (82.0%), with other domains between 86.4% and 90.2% (Drilling 87.3%, Geophysics 89.8%, Petroleum Geology 90.2%, Production 86.4%, Sedimentology 89.6%). GLM-4.7 achieves the top open-weight score in every domain, leading alone in four (Geophysics, Petroleum Geology, Petrophysics, Sedimentology) and tying with other models in three. The lowest scores across all domains come from Llama-3.2-3b-instruct, including 45.8% in Drilling and 51.1% in Petrophysics.

These results suggest that petroleum geoscience knowledge is well-represented in open-weight models, with 11 of 32 exceeding 90% accuracy and 22 exceeding 85%. Even models with fewer parameters and lower costs show useful domain understanding, making them viable options for applications where open weights or cost efficiency matter. Smaller closed models also perform well (GPT-5-nano variants achieve 92.9–93.3%), suggesting that compact models of either type can handle this domain.

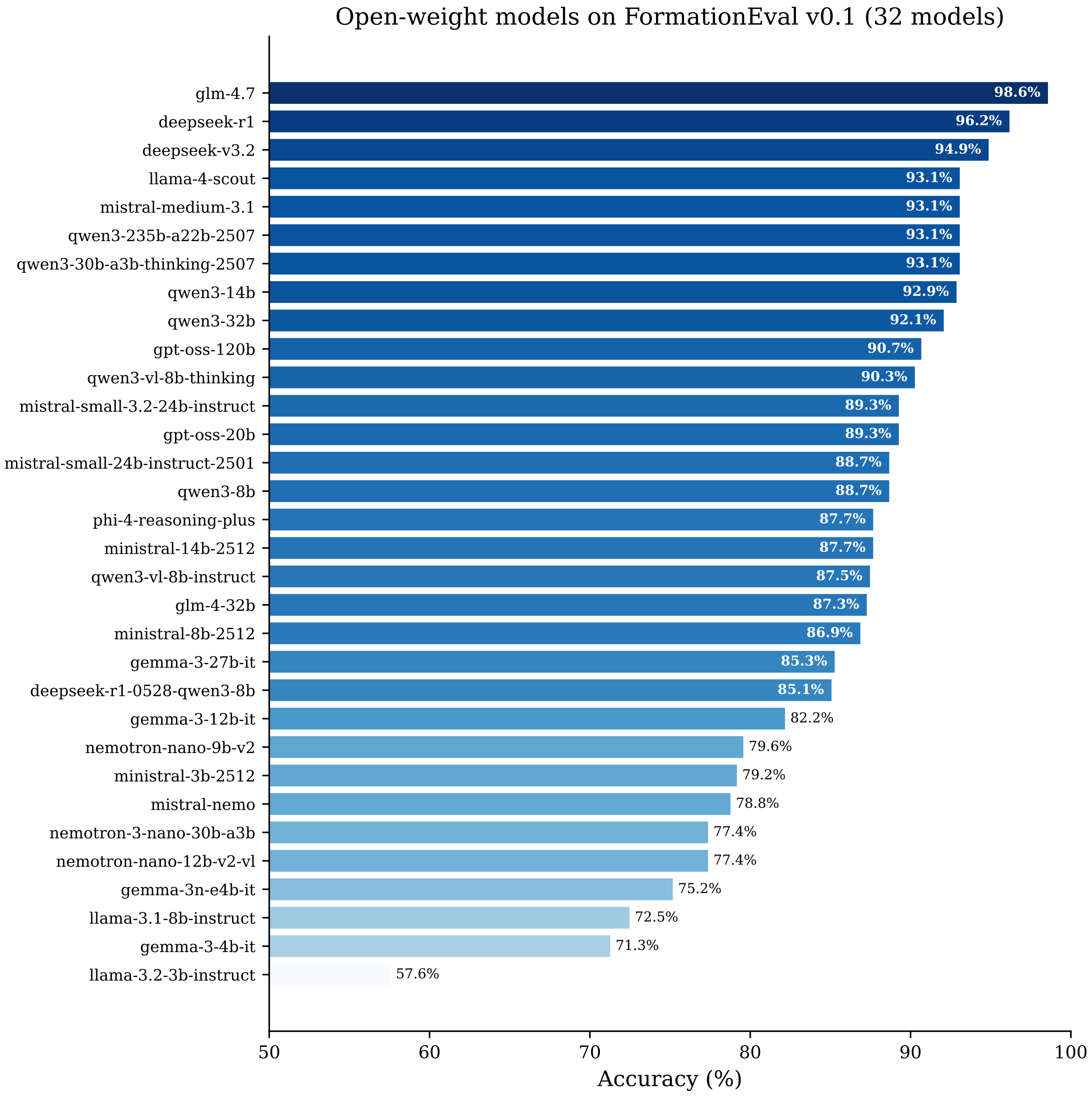

Figure 8: All 32 open-weight models ranked by accuracy. GLM-4.7 leads at 98.6%, followed by DeepSeek-R1 (96.2%) and DeepSeek-V3.2 (94.9%). Color intensity indicates accuracy level.

Table 6: Open-weight models by accuracy (prices in USD per million tokens).

Overall rank	Model	$/M (in/out)	Accuracy	Correct
2	glm-4.7	0.40/1.50	98.6%	498/505
13	deepseek-r1	0.30/1.20	96.2%	486/505
21	deepseek-v3.2	0.22/0.32	94.9%	479/505
33	llama-4-scout	0.08/0.30	93.1%	470/505
34	mistral-medium-3.1	0.40/2.00	93.1%	470/505
35	qwen3-235b-a22b-2507	0.07/0.46	93.1%	470/505
36	qwen3-30b-a3b-thinking-2507	0.05/0.34	93.1%	470/505
41	qwen3-14b	0.05/0.22	92.9%	469/505
42	qwen3-32b	0.08/0.24	92.1%	465/505
46	gpt-oss-120b	0.04/0.19	90.7%	458/505
47	qwen3-vl-8b-thinking	0.18/2.10	90.3%	456/505
48	mistral-small-3.2-24b-instruct	0.06/0.18	89.3%	451/505
49	gpt-oss-20b	0.03/0.14	89.3%	451/505
51	mistral-small-24b-instruct-2501	0.03/0.11	88.7%	448/505
52	qwen3-8b	0.03/0.11	88.7%	448/505
53	phi-4-reasoning-plus	0.07/0.35	87.7%	443/505
54	ministral-14b-2512	0.20/0.20	87.7%	443/505
55	qwen3-vl-8b-instruct	0.06/0.40	87.5%	442/505
56	glm-4-32b	0.10/0.10	87.3%	441/505
57	ministral-8b-2512	0.15/0.15	86.9%	439/505
59	gemma-3-27b-it	0.04/0.15	85.3%	431/505
60	deepseek-r1-0528-qwen3-8b	0.02/0.10	85.1%	430/505
63	gemma-3-12b-it	0.03/0.10	82.2%	415/505
64	nemotron-nano-9b-v2	0.04/0.16	79.6%	402/505
65	ministral-3b-2512	0.10/0.10	79.2%	400/505
66	mistral-nemo	0.02/0.04	78.8%	398/505
67	nemotron-3-nano-30b-a3b	0.06/0.24	77.4%	391/505
68	nemotron-nano-12b-v2-vl	0.20/0.60	77.4%	391/505
69	gemma-3n-e4b-it	0.02/0.04	75.2%	380/505
70	llama-3.1-8b-instruct	0.02/0.03	72.5%	366/505
71	gemma-3-4b-it	0.02/0.07	71.3%	360/505
72	llama-3.2-3b-instruct	0.02/0.02	57.6%	291/505

6.6 Bottom performers

The ten lowest-scoring models are all compact variants (3B–12B parameters) or older architectures. Llama-3.2-3b-instruct achieves 57.6%, only marginally above random guessing (25%). The bottom tier includes Gemma-3-4b-it (71.3%), Llama-3.1-8b-instruct (72.5%), Gemma-3n-e4b-it (75.2%) and Nemotron variants (77–80%).

These models show consistent patterns. Accuracy on easy questions (63–87%) exceeds medium (55–78%) and hard (59–82%), but the gap between easy and medium is larger than for top models. Llama-3.2-3b-instruct shows the widest spread, at 62.9% easy versus 54.7% medium.

Petrophysics accuracy for bottom-tier models falls to 51–74%, while other domains remain at

62–98%. This suggests that smaller models particularly struggle with the technical detail of well logging concepts. The full 72-model leaderboard is provided in Appendix D.

7 Discussion

7.1 Interpretation of scores

The benchmark provides relative comparisons between models rather than absolute measures of domain competence. High scores indicate that a model can answer concept-based questions derived from authoritative sources, but do not guarantee expertise in practical applications.

All models exhibit elevated length bias in this analysis. They select the longest answer choice more often than the 25% baseline. This rate (38–47% across models) is close to the benchmark's residual bias (correct answer is uniquely longest in 43.2% of questions), making it difficult to distinguish length exploitation from genuine reasoning that produces longer correct answers.

Position bias is generally low across models, with most showing near-uniform A/B/C/D distributions. Two Nvidia Nemotron variants showed elevated position bias ("High" level), potentially indicating instruction-following limitations.

7.2 Limitations and threats to validity

Residual length bias persists despite mitigation efforts. Correct answers remain uniquely longest in 43.2% of questions (versus 25% expected), which may inflate scores for models sensitive to answer length.

Contamination risk cannot be fully ruled out. While questions are generated from concepts rather than copied, the underlying topics appear in textbooks that may be in model training data. Contamination risk labels are provided but actual training data overlap cannot be verified.

Quality assurance combined batch review of all 505 questions with spot checks against source material (Section 3.5), but did not include independent expert review. Despite 67 corrections identified during the process, additional errors may remain.

Domain coverage is uneven, with petrophysics dominating (54% of questions) due to source availability. This may not reflect the breadth of petroleum geoscience equally.

7.3 Provider and pricing variability

Model pricing fluctuates and may not reflect values at time of reading. OpenRouter and Azure pricing differ for the same models. Some models are available through multiple providers at different costs.

Provider reliability varies. Rate limits, latency and availability differ across Azure OpenAI and OpenRouter endpoints. The caching strategy mitigates this for reproducibility but may not reflect real-world API behavior.

8 Release and reproducibility

The benchmark artifacts are organized for reproducibility. Version-controlled outputs include the dataset JSON and PDF, leaderboard and analysis reports in Markdown, and per-question CSV with raw model responses (truncated to 500 characters). Cached responses (gitignored) store raw API responses per model and question, enabling re-analysis without additional API calls. A PDF export of the full dataset provides question cards, metadata and bookmarks for domain expert review.

Evaluation scripts support an analyze-only mode that regenerates reports from cached responses without making API calls, ensuring reproducibility even when model APIs change or become unavailable.

9 Data and code availability

The benchmark, evaluation reports and code are available in the project repository [15].

- Repository: `github.com/AlmazErmilov/FormationEval-an-Open-Benchmark...`
- Dataset JSON [18]: `data/benchmark/formationeval_v0.1.json`
- Dataset PDF [17]: `data/benchmark/formationeval_v0.1.pdf`
- Leaderboard: `eval/results/leaderboard.md`
- Analysis: `eval/results/analysis.md`
- Per-question results: `eval/results/questions.csv`
- Generation scripts: `src/`
- Evaluation scripts: `eval/`

10 Conclusion and future work

This paper introduced FormationEval, a 505-question multiple-choice benchmark for evaluating language models on petroleum geoscience. The benchmark covers seven domains derived from authoritative sources using a concept-based methodology that respects copyright while testing domain knowledge.

Evaluation of 72 models reveals that frontier models achieve over 97% accuracy, with Gemini 3 Pro Preview leading at 99.8%. Open-weight models perform competitively, with GLM-4.7 achieving 98.6%. Petrophysics emerges as the most challenging domain across all models.

Future work includes expanding to additional languages (Norwegian, Russian), adding questions from more sources to balance domain coverage and developing contamination detection methods. The benchmark is intended as a living resource, with new models added to the leaderboard as they become available.

A Schema reference

Each question includes the following fields:

Field	Description
`id`	Unique identifier
`version`	Version string (e.g., "0.1")
`question`	Question text
`choices`	Array of 4 options (A–D)
`answer_index`	Correct answer index (0–3)
`answer_key`	Correct answer letter (A–D)
`rationale`	Explanation of correct answer
`difficulty`	easy | medium | hard
`language`	Language code: en | ru | no
`domains`	Array of broad categories
`topics`	Array of specific subjects
`sources`	Provenance metadata array
`derivation_mode`	Always "concept_based"
`metadata.calc_required`	Boolean (calculation needed)
`metadata.contamination_risk`	low | medium | high

Contamination risk indicates likelihood that similar questions exist in LLM training data: *low* for questions unique to this source, *medium* for common concepts and *high* for standard textbook topics.

B Prompt templates

B.1 Evaluation prompt

System prompt:

```
You are taking a multiple-choice exam on Oil & Gas geoscience.
For each question, select the single best answer from the options provided.
State your final answer as a single letter:  A, B, C or D.
```

User prompt format:

```
{question}

A) {choice_a}
B) {choice_b}
C) {choice_c}
D) {choice_d}

Answer:
```

B.2 Multiple-choice question generation prompt summary

The full system prompt for MCQ generation (475 lines) is available in the repository. Key instructions include:

1. Generate questions that test **understanding of concepts**, not recognition of phrases
2. Test one concept per question
3. Never copy sentences or descriptive phrases from source text
4. Create standalone questions answerable from domain knowledge without access to the source chapter
5. Cover key concepts without repetition
6. Distribute difficulty: 30% easy, 50% medium, 20% hard
7. Balance answer options in length and structure
8. Distribute correct answers evenly across positions A/B/C/D
9. Avoid "All of the above" and "None of the above"
10. Avoid negative phrasing ("Which is NOT...")
11. Avoid exploitable patterns with qualifier words
12. Provide rationale explaining why the answer is correct
13. Include source metadata and contamination risk assessment

C Bias mitigation summary

Tables 7 and 8 summarize the bias mitigation efforts.

Table 7: Length bias before and after mitigation.

Metric	Original	After fixes
Correct answer is uniquely longest	>55%	43.2%
Correct answer avg length	86.6 chars	86.6 chars
Distractor avg length	69.8 chars	74.0 chars

Table 8: Qualifier word patterns before and after mitigation.

Word	In correct	In distractor	Correct rate	Status
"always"	0	49	0%	Replaced with synonyms
"invariably"	0	12	0%	New (from "always")
"necessarily"	0	14	0%	New (from "always")
"inherently"	0	12	0%	New (from "always")
"consistently"	0	9	0%	New (from "always")
"may"	13	0	100%	Original (pre-fix)
"may"	13	13	50%	After fix

Mitigation applied: (1) All 49 "always" instances replaced with varied synonyms to break single-word exploit; (2) Added "may" to 13 distractors with "no-effect" claims.

Residual issues: Absolute word synonyms still have 0% correct rate. Combined “any-absolute-word=wrong” heuristic remains partially exploitable.

D Full model leaderboard

Table 9 presents all 72 evaluated models ranked by accuracy.

Table 9: Complete leaderboard. All 72 models by accuracy (prices in USD per million tokens).

Rank	Model	Open	$/M (in/out)	Accuracy	Correct
1	gemini-3-pro-preview	No	2.00/12.00	99.8%	504/505
2	glm-4.7	Yes	0.40/1.50	98.6%	498/505
3	gemini-3-flash-preview	No	0.50/3.00	98.2%	496/505
4	gemini-2.5-pro	No	1.25/10.00	97.8%	494/505
5	grok-4.1-fast	No	0.20/0.50	97.6%	493/505
6	gpt-5.2-chat-medium	No	1.75/14.00	97.4%	492/505
7	kimi-k2-thinking	No	0.40/1.75	97.2%	491/505
8	claude-opus-4.5	No	5.00/25.00	97.0%	490/505
9	gpt-5.2-chat-high	No	1.75/14.00	96.8%	489/505
10	gpt-5.2-chat-low	No	1.75/14.00	96.8%	489/505
11	gpt-5-mini-medium	No	0.25/2.00	96.4%	487/505
12	gpt-5.1-chat-medium	No	1.25/10.00	96.4%	487/505
13	deepseek-r1	Yes	0.30/1.20	96.2%	486/505
14	grok-4-fast	No	0.20/0.50	96.0%	485/505
15	gpt-5-mini-high	No	0.25/2.00	95.6%	483/505
16	gpt-5-mini-low	No	0.25/2.00	95.2%	481/505
17	o4-mini-high	No	1.10/4.40	95.2%	481/505
18	gemini-2.5-flash	No	0.30/2.50	95.0%	480/505
19	o4-mini-medium	No	1.10/4.40	95.0%	480/505
20	grok-3-mini	No	0.30/0.50	95.0%	480/505
21	deepseek-v3.2	Yes	0.22/0.32	94.9%	479/505
22	gpt-5.1-chat-low	No	1.25/10.00	94.9%	479/505
23	o3-mini-low	No	1.10/4.40	94.9%	479/505
24	o3-mini-medium	No	1.10/4.40	94.9%	479/505
25	claude-3.7-sonnet	No	3.00/15.00	94.7%	478/505
26	o3-mini-high	No	1.10/4.40	94.7%	478/505
27	gpt-5-chat	No	1.25/10.00	94.5%	477/505
28	o4-mini-low	No	1.10/4.40	94.3%	476/505
29	gpt-5.1-chat-high	No	1.25/10.00	93.9%	474/505
30	gpt-4.1	No	2.00/8.00	93.7%	473/505
31	gemini-2.0-flash-001	No	0.10/0.40	93.3%	471/505
32	gpt-5-nano-low	No	0.05/0.40	93.3%	471/505
33	llama-4-scout	Yes	0.08/0.30	93.1%	470/505
34	mistral-medium-3.1	Yes	0.40/2.00	93.1%	470/505
35	qwen3-235b-a22b-2507	Yes	0.07/0.46	93.1%	470/505

Continued on next page

Table 9 – continued from previous page

Rank	Model	Open	$/M (in/out)	Accuracy	Correct
36	qwen3-30b-a3b-thinking-2507	Yes	0.05/0.34	93.1%	470/505
37	gpt-4o	No	2.50/10.00	92.9%	469/505
38	gpt-5-nano-high	No	0.05/0.40	92.9%	469/505
39	gpt-5-nano-medium	No	0.05/0.40	92.9%	469/505
40	minimax-m2	No	0.20/1.00	92.9%	469/505
41	qwen3-14b	Yes	0.05/0.22	92.9%	469/505
42	qwen3-32b	Yes	0.08/0.24	92.1%	465/505
43	gpt-4.1-mini	No	0.40/1.60	91.7%	463/505
44	claude-haiku-4.5	No	1.00/5.00	91.5%	462/505
45	gemini-2.5-flash-lite	No	0.10/0.40	91.3%	461/505
46	gpt-oss-120b	Yes	0.04/0.19	90.7%	458/505
47	qwen3-vl-8b-thinking	Yes	0.18/2.10	90.3%	456/505
48	mistral-small-3.2-24b-instruct	Yes	0.06/0.18	89.3%	451/505
49	gpt-oss-20b	Yes	0.03/0.14	89.3%	451/505
50	claude-sonnet-4.5	No	3.00/15.00	89.1%	450/505
51	mistral-small-24b-instruct-2501	Yes	0.03/0.11	88.7%	448/505
52	qwen3-8b	Yes	0.03/0.11	88.7%	448/505
53	phi-4-reasoning-plus	Yes	0.07/0.35	87.7%	443/505
54	ministral-14b-2512	Yes	0.20/0.20	87.7%	443/505
55	qwen3-vl-8b-instruct	Yes	0.06/0.40	87.5%	442/505
56	glm-4-32b	Yes	0.10/0.10	87.3%	441/505
57	ministral-8b-2512	Yes	0.15/0.15	86.9%	439/505
58	gpt-4.1-nano	No	0.10/0.40	86.1%	435/505
59	gemma-3-27b-it	Yes	0.04/0.15	85.3%	431/505
60	deepseek-r1-0528-qwen3-8b	Yes	0.02/0.10	85.1%	430/505
61	gpt-4o-mini	No	0.15/0.60	84.8%	428/505
62	claude-3.5-haiku	No	0.80/4.00	84.0%	424/505
63	gemma-3-12b-it	Yes	0.03/0.10	82.2%	415/505
64	nemotron-nano-9b-v2	Yes	0.04/0.16	79.6%	402/505
65	ministral-3b-2512	Yes	0.10/0.10	79.2%	400/505
66	mistral-nemo	Yes	0.02/0.04	78.8%	398/505
67	nemotron-3-nano-30b-a3b	Yes	0.06/0.24	77.4%	391/505
68	nemotron-nano-12b-v2-vl	Yes	0.20/0.60	77.4%	391/505
69	gemma-3n-e4b-it	Yes	0.02/0.04	75.2%	380/505
70	llama-3.1-8b-instruct	Yes	0.02/0.03	72.5%	366/505
71	gemma-3-4b-it	Yes	0.02/0.07	71.3%	360/505
72	llama-3.2-3b-instruct	Yes	0.02/0.02	57.6%	291/505

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used GPT-5.2 (OpenAI) with extra high reasoning effort to transform textbook chapter content into multiple-choice questions following a structured generation pipeline (Section 3.4). All domain knowledge was derived from the source materials.

The model served as a generation and reformatting tool, not as a knowledge source. GenAI tools were also used for language editing. The author reviewed and edited all generated content and takes full responsibility for the publication.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. URL: `https://arxiv.org/abs/2412.08905`.

[2] Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *Proceedings of EMNLP*, 2024. URL: `https://arxiv.org/abs/2402.01512`.

[3] Anthropic. The Claude 3 model family: Opus, sonnet, haiku. `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`, 2024. Model card. Accessed December 2025.

[4] Anthropic. Model card addendum: Claude 3.5 haiku and upgraded Claude 3.5 sonnet. `https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf`, 2024. Model card addendum. Accessed December 2025.

[5] Anthropic. Claude 3.7 sonnet system card. `https://www.anthropic.com/claude-3-7-sonnet-system-card`, 2025. System card. Accessed December 2025.

[6] Anthropic. Claude opus 4.5 system card. `https://www.anthropic.com/claude-opus-4-5-system-card`, 2025. System card. Accessed December 2025.

[7] Knut Bjørlykke. *Petroleum Geoscience: From Sedimentary Environments to Rock Physics*. Springer, Berlin, 2010. `doi:10.1007/978-3-642-02332-3`.

[8] Dhawaleswar Rao Ch and Sujan Kumar Saha. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25, 2020. `doi:10.1109/TLT.2018.2889100`.

[9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL: `https://arxiv.org/abs/1803.05457`.

[10] DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL: `https://arxiv.org/abs/2412.19437`.

[11] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL: `https://arxiv.org/abs/2501.12948`.

[12] DeepSeek-AI. DeepSeek-V3.2 model card. `https://huggingface.co/deepseek-ai/DeepSeek-V3.2`, 2025. Hugging Face model card. Accessed December 2025.

[13] Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuzhong Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, 2024. URL: `https://arxiv.org/abs/2306.05064`, `doi:10.1145/3616855.3635772`.

[14] Darwin V. Ellis and Julian M. Singer. *Well Logging for Earth Scientists*. Springer, Dordrecht, 2nd edition, 2007. `doi:10.1007/978-1-4020-4602-5`.

[15] Almaz Ermilov. FormationEval: An open benchmark for oil and gas geoscience MCQ evaluation. `https://github.com/AlmazErmilov/FormationEval-an-Open-Benchmark-for-Oil-Gas-Geoscience-MCQ-Evaluation`, 2025. Repository with benchmark data and evaluation code. Accessed December 2025.

[16] Almaz Ermilov. FormationEval v0.1 analysis. `https://github.com/AlmazErmilov/FormationEval-an-Open-Benchmark-for-Oil-Gas-Geoscience-MCQ-Evaluation/blob/v0.1/eval/results/analysis.md`, 2025. Hardest questions and bias analysis. Accessed December 2025.

[17] Almaz Ermilov. FormationEval v0.1 benchmark pdf. `https://github.com/AlmazErmilov/FormationEval-an-Open-Benchmark-for-Oil-Gas-Geoscience-MCQ-Evaluation/blob/v0.1/data/benchmark/formationeval_v0.1.pdf`, 2025. PDF export of the benchmark for reading and review. Accessed December 2025.

[18] Almaz Ermilov. FormationEval v0.1 dataset. `https://github.com/AlmazErmilov/FormationEval-an-Open-Benchmark-for-Oil-Gas-Geoscience-MCQ-Evaluation/blob/v0.1/data/benchmark/formationeval_v0.1.json`, 2025. 505 MCQs in JSON format with source metadata. Accessed December 2025.

[19] Almaz Ermilov. FormationEval v0.1 leaderboard. `https://github.com/AlmazErmilov/FormationEval-an-Open-Benchmark-for-Oil-Gas-Geoscience-MCQ-Evaluation/blob/v0.1/eval/results/leaderboard.md`, 2025. Evaluation leaderboard results. Accessed December 2025.

[20] Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. URL: `https://arxiv.org/abs/2312.11805`.

[21] Gemini Team, Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL: `https://arxiv.org/abs/2507.06261`.

[22] GLM Team, Zhipu AI. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. URL: `https://arxiv.org/abs/2406.12793`.

[23] Google DeepMind. Gemini 2 flash model card. `https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf`, 2025. Model card. Accessed December 2025.

[24] Google DeepMind. Gemini 3 pro model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf`, 2025. Model card. Accessed December 2025.

[25] Google DeepMind. Gemma 3 model card. https://ai.google.dev/gemma/docs/core/model_card_3, 2025. Accessed December 2025.

[26] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL: https://arxiv.org/abs/2308.11462.

[27] Abdenour Hadid, Tanujit Chakraborty, and Daniel Busby. When geoscience meets generative AI and large language models: Foundations, trends, and future challenges. *Expert Systems*, 41(10), 2024. URL: https://arxiv.org/abs/2402.03349, doi:10.1111/exsy.13654.

[28] 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*, 2021. URL: https://arxiv.org/abs/2009.03300.

[29] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. URL: https://arxiv.org/abs/2009.13081, doi:10.3390/app11146421.

[30] Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuzhong Shi, Beiya Dai, Yunchong Song, et al. GeoGalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*, 2024. URL: https://arxiv.org/abs/2401.00434.

[31] Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL: https://arxiv.org/abs/2407.21783.

[32] Meta. Llama 3.1 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md, 2024. Model card. Accessed December 2025.

[33] Meta. Llama 3.2 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md, 2024. Model card. Accessed December 2025.

[34] Meta. Llama 4 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md, 2025. GitHub model card. Accessed December 2025.

[35] Microsoft. Azure OpenAI Service documentation. https://learn.microsoft.com/en-us/azure/ai-services/openai/, 2025. Accessed December 2025.

[36] MiniMax. MiniMax-M2: A model built for max coding and agentic workflows. https://huggingface.co/MiniMaxAI/MiniMax-M2, 2025. Hugging Face model card. Accessed December 2025.

[37] Mistral AI. Mistral AI models. https://docs.mistral.ai/getting-started/models/, 2025. Model documentation. Accessed December 2025.

[38] Moonshot AI. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. URL: https://arxiv.org/abs/2507.20534.

[39] NVIDIA. Nemotron: Foundation models for agentic AI. https://developer.nvidia.com/nemotron, 2025. Accessed December 2025.

[40] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL: https://arxiv.org/abs/2303.08774.

[41] OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL: https://arxiv.org/abs/2410.21276.

[42] OpenAI. GPT-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025. System card. Accessed December 2025.

[43] OpenAI. GPT-5.2 system card. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf, 2025. System card. Accessed January 2026.

[44] OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025. Accessed December 2025.

[45] OpenAI. OpenAI o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, 2025. System card. Accessed December 2025.

[46] OpenRouter. Grok 3 model page. https://openrouter.ai/x-ai/grok-3, 2025. Model page. Accessed December 2025.

[47] OpenRouter. OpenRouter: Unified api for AI models. https://openrouter.ai/docs, 2025. Accessed December 2025.

[48] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL: https://arxiv.org/abs/2505.09388.

[49] David Rein, Betty Li Hou, et al. GPQA: A graduate-level Google-Proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL: https://arxiv.org/abs/2311.12022.

[50] TU Delft OpenCourseWare. Petroleum geology. https://ocw.tudelft.nl/courses/petroleum-geology/, 2008. Accessed December 2025.

[51] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shicheng Zhang, Yixin Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023. URL: https://arxiv.org/abs/2307.10635.

[52] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL: https://arxiv.org/abs/2406.01574.

[53] xAI. Grok 4 model card. https://data.x.ai/2025-08-20-grok-4-model-card.pdf, 2025. Model card. Accessed December 2025.

[54] Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Fei Ouyang, Zhichao Yang, and Hong Yu. MCQG-SRefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. *Proceedings of NAACL*, 2025. URL: https://arxiv.org/abs/2410.13191.

[55] Zhipu AI. GLM-4.7 model card. `https://huggingface.co/zai-org/GLM-4.7`, 2025. Hugging Face model card. Accessed December 2025.