FormationEval: an open benchmark for oil and gas geoscience MCQ evaluation

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

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 concept-based approach that avoids verbatim copying of copyrighted text. Each question includes provenance metadata and a contamination risk label to support reproducible evaluation. 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%. Among open-weight models, GLM-4.7 leads at 98.6%. 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 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 [23] cover broad knowledge but lack depth in technical disciplines. For petroleum geoscience and subsurface engineering—fields requiring understanding of well logging physics, reservoir characterization and geological interpretation—no widely adopted public benchmark exists.

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 MCQs from technical sources without verbatim copying; (2) a curated dataset with provenance 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 [23] and ARC [7] evaluate broad knowledge and reasoning capabilities but provide limited coverage of specialized domains. MMLU includes professional

exam questions across 57 subjects but only a few relate to earth sciences or engineering. Domain-specific benchmarks exist for medicine, law and computer science, but petroleum geoscience lacks an equivalent evaluation resource.

MCQ generation from text has been explored using both rule-based and neural approaches. The concept-based methodology presented here differs by emphasizing original question formulation rather than transformation of source sentences, prioritizing both copyright compliance and assessment of understanding over 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 may belong to multiple domains when the topic spans disciplines.

Each question is assigned a difficulty level (easy, medium or hard) based on cognitive demand: easy questions test definitions and direct recall, medium questions require applying concepts to scenarios, and hard questions involve integrating multiple concepts or analyzing edge cases.

3.2 Source selection and licensing policy

The benchmark draws from three sources: Ellis & Singer's *Well Logging for Earth Scientists* [11] (219 questions), Bjørlykke's *Petroleum Geoscience* [6] (262 questions) and TU Delft OpenCourseWare [41] (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 provenance 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. This enables analysis of model performance stratified by contamination likelihood.

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

3.4 Generation pipeline

The generation pipeline consists of four stages:

1. **Text extraction**: Source PDFs are converted to Markdown using OCR, preserving structure and mathematical notation.
2. **Chunking**: Documents are split by chapter or section, with each chunk sized for model context (typically one chapter, approximately 10,000–15,000 tokens).
3. **Candidate generation**: GPT-5.2 with high reasoning effort 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**: Generated questions undergo consistency validation (schema compliance, no duplicate choices, answer index in range) and evidence-based verification (confirming 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 475-line system prompt is available in the repository at `src/prompts/mcq_generator_system_prompt.tx`

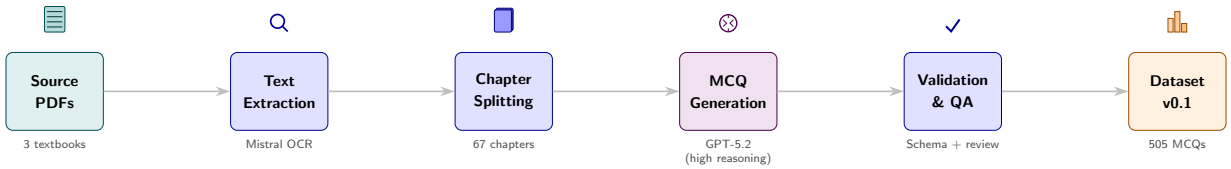

Figure 1: MCQ generation pipeline: source PDFs are converted to Markdown via OCR, split into chapter chunks, processed by GPT-5.2 to generate candidate questions, and verified for schema compliance and source evidence.

3.5 Quality assurance and audit

Human spot-checking is essential despite high LLM instruction-following reliability. For each source chapter, 5–10 questions were manually 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.

The full dataset (505 questions across 45+ chapters) was reviewed in batches grouped by source and chapter. The audit log records batch status and any issues requiring correction. One question required fixing during the final review pass.

The required rationale field accelerated verification: when a rationale contradicts the answer it claims to support, the error is immediately visible without re-reading source material.

3.6 Bias analysis and mitigation

Initial analysis revealed two exploitable patterns:

Length bias: Correct answers were longest in 64.6% of questions (expected: 25%), with correct answers averaging 86.6 characters versus 69.8 for distractors.

Qualifier word bias: 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).

A model using only surface heuristics could achieve approximately 70% accuracy without domain knowledge.

Mitigation applied: 136 distractors were expanded with technical context to reduce length imbalance (from 64.6% to 51.5% 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 with "no-effect" claims to balance hedging language.

Residual issues: Length bias remains above the 25% baseline. The absolute-word synonyms all have 0% correct rate, making a combined "any-absolute-word=wrong" heuristic still partially exploitable. These limitations are documented for transparency. See Appendix C for detailed before/after metrics.

3.7 Human-readable export

To facilitate human 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 [15, 14].

4 Dataset summary

Version 0.1 of FormationEval contains 505 questions in English covering 811 unique topics. Table 1 summarizes key metrics and Table 2 shows the distribution by domain and difficulty. Table 3 provides the breakdown by source. 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%.

Table 1: Dataset summary (v0.1).

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

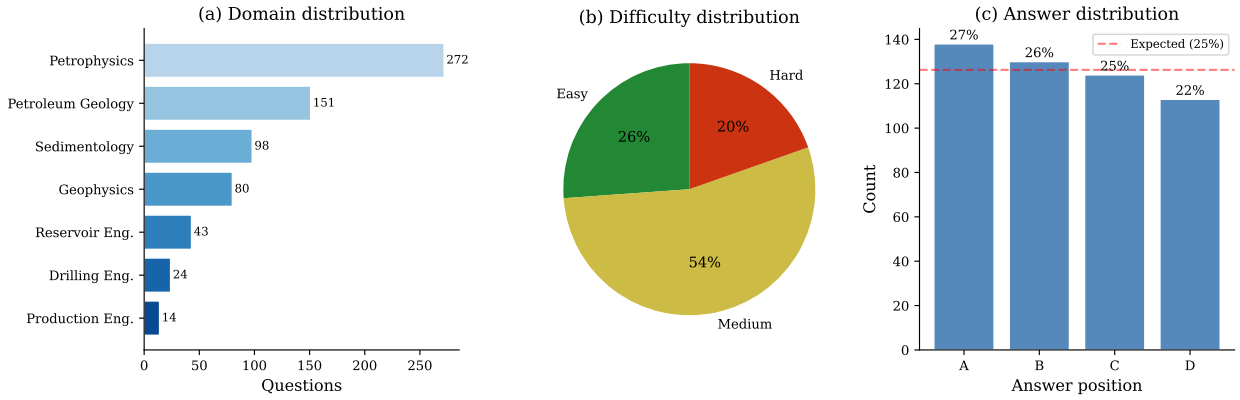

Figure 2: Dataset composition: (a) questions by domain, with Petrophysics dominating due to source coverage; (b) difficulty distribution close to 30/50/20 targets; (c) answer position distribution near the expected 25% baseline.

5 Evaluation setup

5.1 Models and providers

The evaluation covers 72 language models across two providers: Azure OpenAI [28] and OpenRouter [39]. This includes models from OpenAI [33, 34, 36, 35, 37] (GPT-4o, GPT-4.1, GPT-5 series, o3-mini, o4-mini), Anthropic [2, 3, 4, 5] (Claude 3.5 Haiku, Claude 3.7 Sonnet, Claude Opus 4.5), Google [17, 20, 18, 21, 22] (Gemini 2.0, 2.5, 3 series, Gemma 3), Meta [24, 25, 26, 27] (Llama 3.1, 3.2, 4), DeepSeek [8, 10, 9] (R1, V3.2), Mistral [30] (Small, Medium, Nemo, Ministral), Alibaba [40] (Qwen3 series), Zhipu [19, 44] (GLM-4, GLM-4.7), xAI [38, 43] (Grok 3, 4), Moonshot [31] (Kimi K2), MiniMax [29] (M2), Microsoft [1] (Phi-4) and Nvidia [32] (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 tokens (Llama-3.2-3b) to $25/M tokens (Claude Opus 4.5 output).

5.2 Prompting and answer extraction

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

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: 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 (`<thinking>`, `<think>`) from models that expose chain-of-thought traces (o1, o3, DeepSeek-R1). The extraction logic prioritizes explicit patterns ("The answer is B", "Answer: C") over positional heuristics. 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.

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: Source breakdown.

Source	Questions
Ellis & Singer – Well Logging for Earth Scientists (2007)	219
Bjørlykke – Petroleum Geoscience (2010)	262
TU Delft OCW – Petroleum Geology (2008)	24

5.3 Metrics and confidence intervals

The primary metric is overall accuracy: correct answers divided by total questions (505). For each accuracy value, 95% Wilson score confidence intervals [42] are computed, which provide better coverage than normal approximation intervals when accuracy is near 0 or 1.

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 longest in 51.5% of questions) provides a reference point for interpreting model length bias.

5.4 Caching and reproducibility

API responses are cached per model and question in a structured directory (`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, human-readable Markdown tables and CSV exports with per-question breakdowns including raw model responses (truncated to 500 characters).

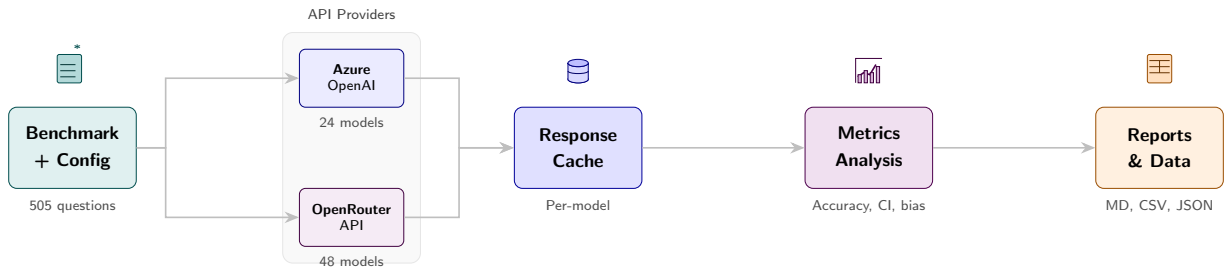

Figure 3: Evaluation pipeline: models are configured via YAML, questions sent to Azure OpenAI or OpenRouter APIs, responses cached per model/question, and analyzed to generate leaderboard and analysis reports.

6 Results

6.1 Overall results

Table 4 presents the top 25 models by accuracy [16]. 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 cost-effective 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.

6.2 By difficulty

Performance decreases with difficulty level, as expected. For top models, easy questions are nearly saturated (100% for several models), while hard questions remain discriminative. Gemini 3 Pro Preview achieves 100% on both easy and hard questions, with a single error on a medium question. GLM-4.7 shows 100% easy, 98.5% medium and 97.0% hard.

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.

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 specificity 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.

Production Engineering and Drilling Engineering show the highest average accuracy, though this may reflect smaller question counts (14 and 24 questions respectively) rather than domain difficulty. Figure 7 shows accuracy patterns across domains for the top 15 models.

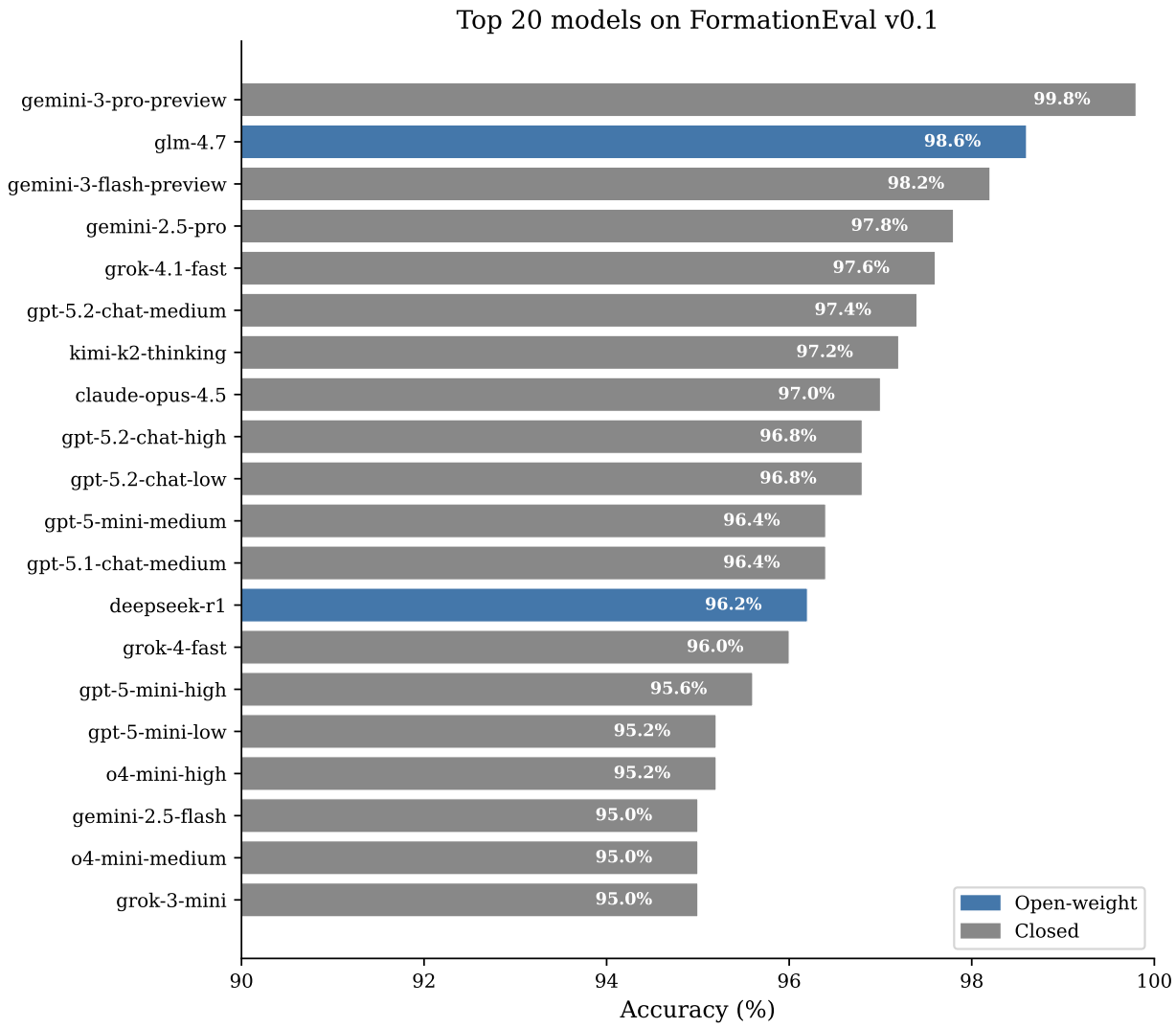

Figure 4: Top 20 models on FormationEval v0.1 by accuracy. Blue bars indicate open-weight models. GLM-4.7 (98.6%) leads among open-weight models, ranking second overall.

6.4 Hardest questions

The ten hardest questions (by model failure rate) reveal systematic knowledge gaps [13]. 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 [13].

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 (LWD propagation) also appear in the top 10.

Model agreement analysis shows: 21.6% of questions were answered correctly by all 72 models, 0% were missed by all models and 78.4% showed mixed results. This distribution suggests the

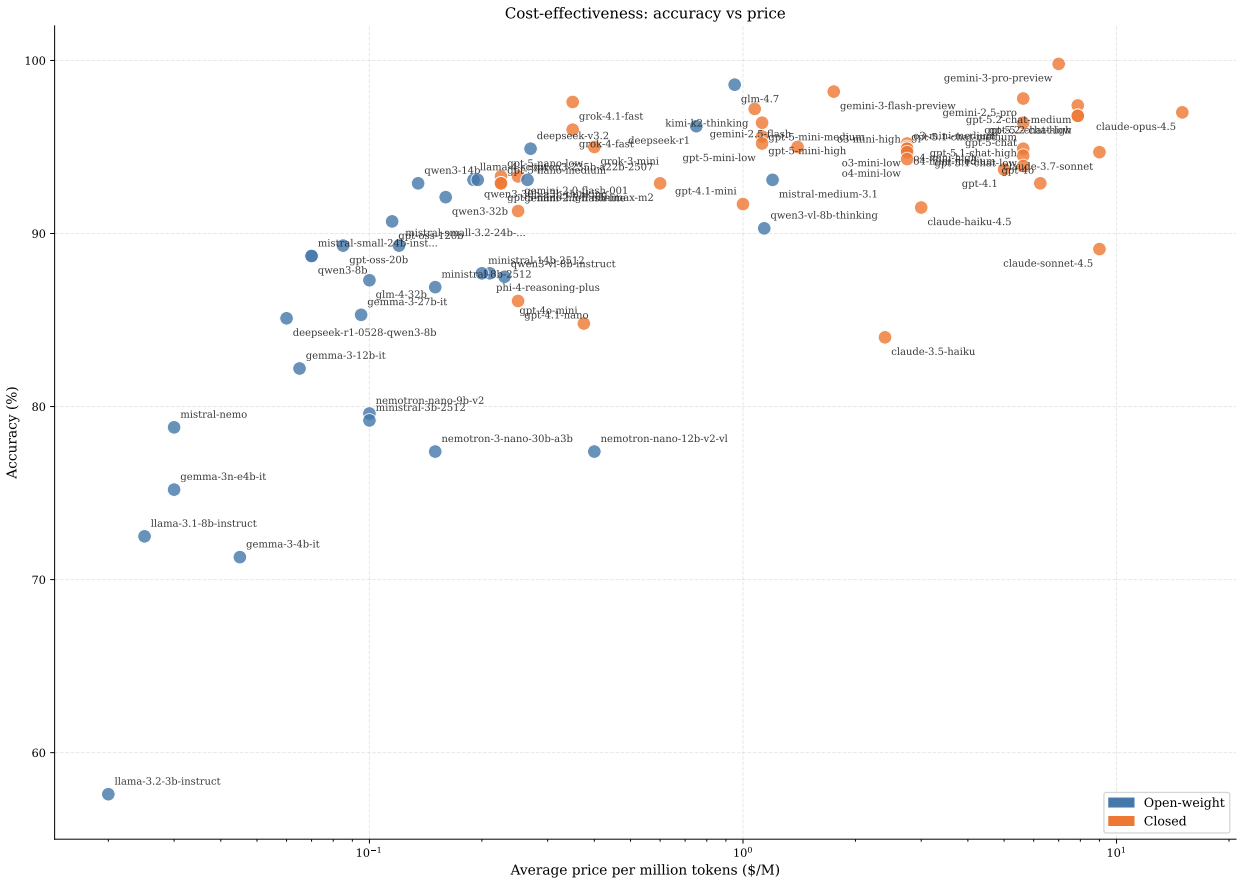

Figure 5: Cost-effectiveness analysis: accuracy versus average token price. Several high-accuracy models (Grok-4.1-fast, DeepSeek-R1) offer strong performance at lower cost. Open-weight models (blue) provide cost-effective alternatives to closed models (orange).

benchmark effectively discriminates between model capabilities.

6.5 Open-weight models

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 5 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 is tied for the top open-weight score in every domain and is the sole top model in Geophysics, Petroleum

Table 4: Top 25 models by accuracy. See Appendix E for the complete 72-model leaderboard.

Rank	Model	Open	Price ($/M)	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

Geology, Petrophysics, and Sedimentology. The lowest scores across all domains come from Llama-3.2-3b-instruct, including 45.8% in Drilling and 51.1% in Petrophysics.

Table 5: Open-weight models by accuracy.

Overall rank	Model	Price ($/M)	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

Continued on next page

Table 5 – continued from previous page

Overall rank	Model	Price ($/M)	Accuracy	Correct
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: 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 specificity of well logging concepts. The full 72-model leaderboard is provided in Appendix E.

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. However, this rate (38–47% across models) is close to or below the benchmark's residual bias (correct answer is longest in 51.5% of questions), suggesting models are largely tracking content rather than exploiting length as a proxy.

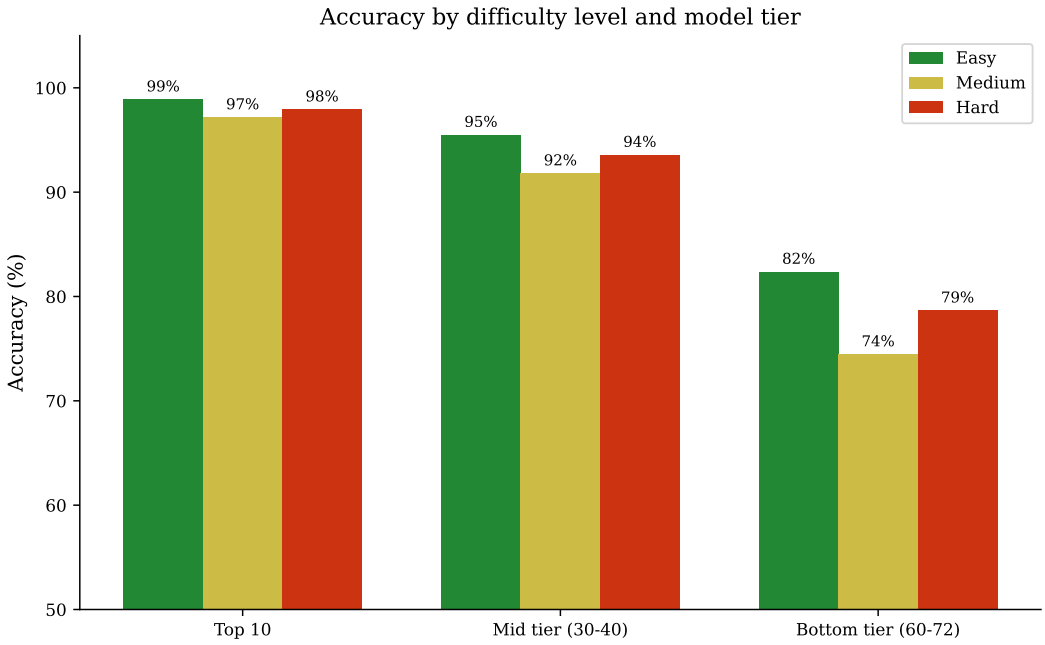

Figure 6: Accuracy by difficulty level across model tiers. Top-tier models show consistent performance across difficulty levels, while bottom-tier models exhibit larger gaps between easy and harder questions.

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: Despite mitigation efforts, correct answers remain longest in 51.5% of questions (versus 25% expected). This may inflate scores for models sensitive to answer length.

Contamination risk: 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 limits: Human verification covered spot checks rather than exhaustive review. One question was corrected during the final pass; additional errors may remain.

Domain coverage: Petrophysics dominates (54% of questions) due to source availability, which 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.

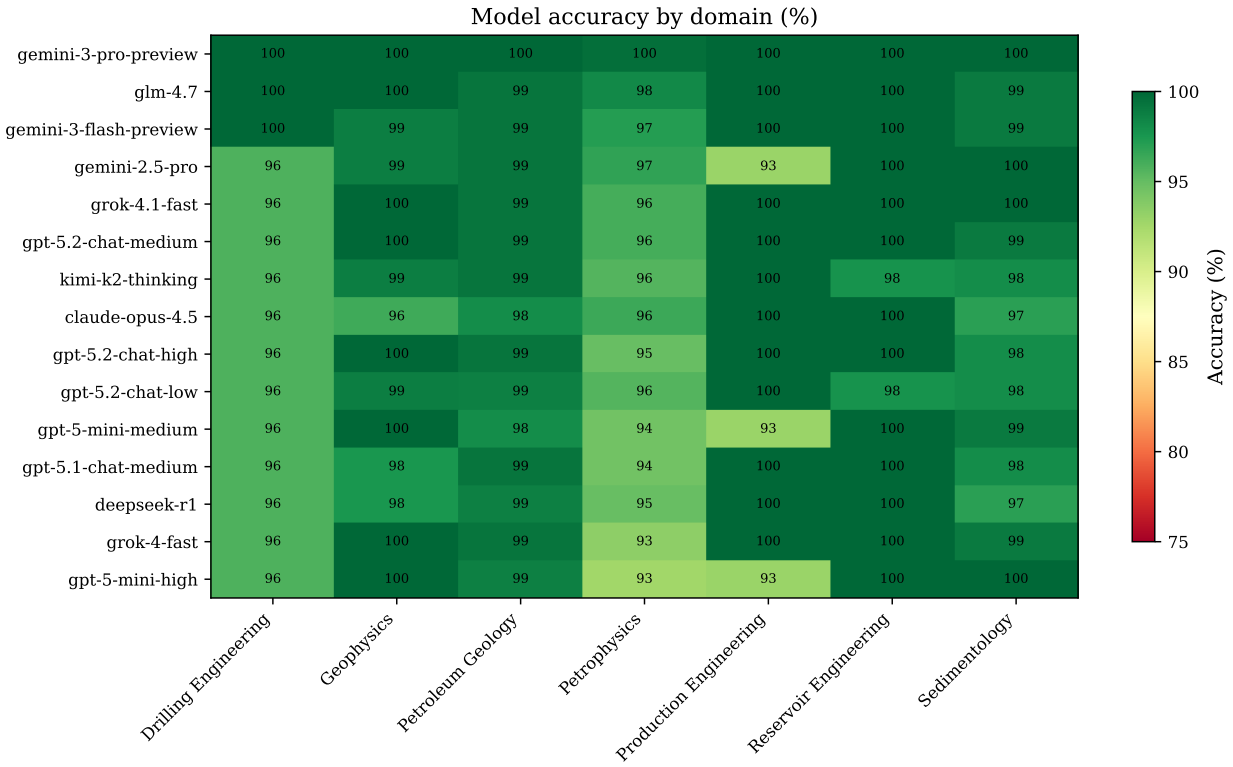

Figure 7: Model accuracy by domain for top 15 models. Petrophysics shows consistently lower accuracy than other domains across all models, reflecting the technical depth of well logging concepts.

8 Release and reproducibility

The benchmark artifacts are organized for reproducibility:

Tracked outputs (version-controlled): Dataset JSON and PDF, leaderboard and analysis reports in Markdown, per-question CSV with raw model responses (truncated to 500 characters).

Cached responses (gitignored): Raw API responses per model/question, enabling re-analysis without API calls.

Human-readable formats: PDF export of the full dataset with 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 [12].

- Repository: `github.com/AlmazErmilov/FormationEval-an-Open-Benchmark...`
- Dataset JSON [15]: `data/benchmark/formationeval_v0.1.json`
- Dataset PDF [14]: `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
`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
`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 Generation prompt summary

The MCQ generation system prompt instructs the model to:

1. Generate questions that test **understanding of concepts**, not recognition of phrases
2. Never copy sentences or descriptive phrases from source text
3. Create standalone questions answerable from domain knowledge without access to the source chapter
4. Distribute difficulty: 30% easy, 50% medium, 20% hard
5. Balance answer options in length and structure
6. Distribute correct answers evenly across positions A/B/C/D
7. Avoid exploitable patterns with qualifier words
8. Include provenance metadata and contamination risk assessment

The full system prompt (475 lines) is available in the repository at `src/prompts/mcq_generator_system_prompt`

C Bias mitigation summary

Tables 6 and 7 summarize the bias mitigation efforts.

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.

Table 6: Length bias before and after mitigation.

Metric	Original	After fixes
Correct answer is longest choice	64.6%	51.5%
Correct answer avg length	86.6 chars	86.6 chars
Distractor avg length	69.8 chars	74.0 chars

Table 7: 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	12	0%	New (from "always")
"inherently"	0	11	0%	New (from "always")
"consistently"	0	8	0%	New (from "always")
"may"	13	0	100%	Original (pre-fix)
"may"	13	13	50%	After fix

D Audit summary

Human verification covered 5–10 questions per source chapter (spot-check approach). The full dataset (505 questions across 45+ chapters) was reviewed in batches:

- Schema compliance: All questions validated
- Answer correctness: Verified against source material
- Phrase copying: Checked for verbatim lifts
- Rationale consistency: Confirmed rationales support marked answers

Several questions were corrected during the final review pass. The required rationale field accelerated verification by making logical inconsistencies immediately visible.

E Full model leaderboard

Table 8 presents all 72 evaluated models ranked by accuracy.

Table 8: Complete leaderboard: all 72 models by accuracy.

Rank	Model	Open	Price ($/M)	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

Continued on next page

Table 8 – continued from previous page

Rank	Model	Open	Price ($/M)	Accuracy	Correct
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

Continued on next page

Table 8 – continued from previous page

Rank	Model	Open	Price ($/M)	Accuracy	Correct
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

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] 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.

[3] 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.

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

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

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

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

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

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

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

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

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

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

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

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

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

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

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

[37] 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.

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

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

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

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

[42] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi:10.1080/01621459.1927.10502953.

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

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

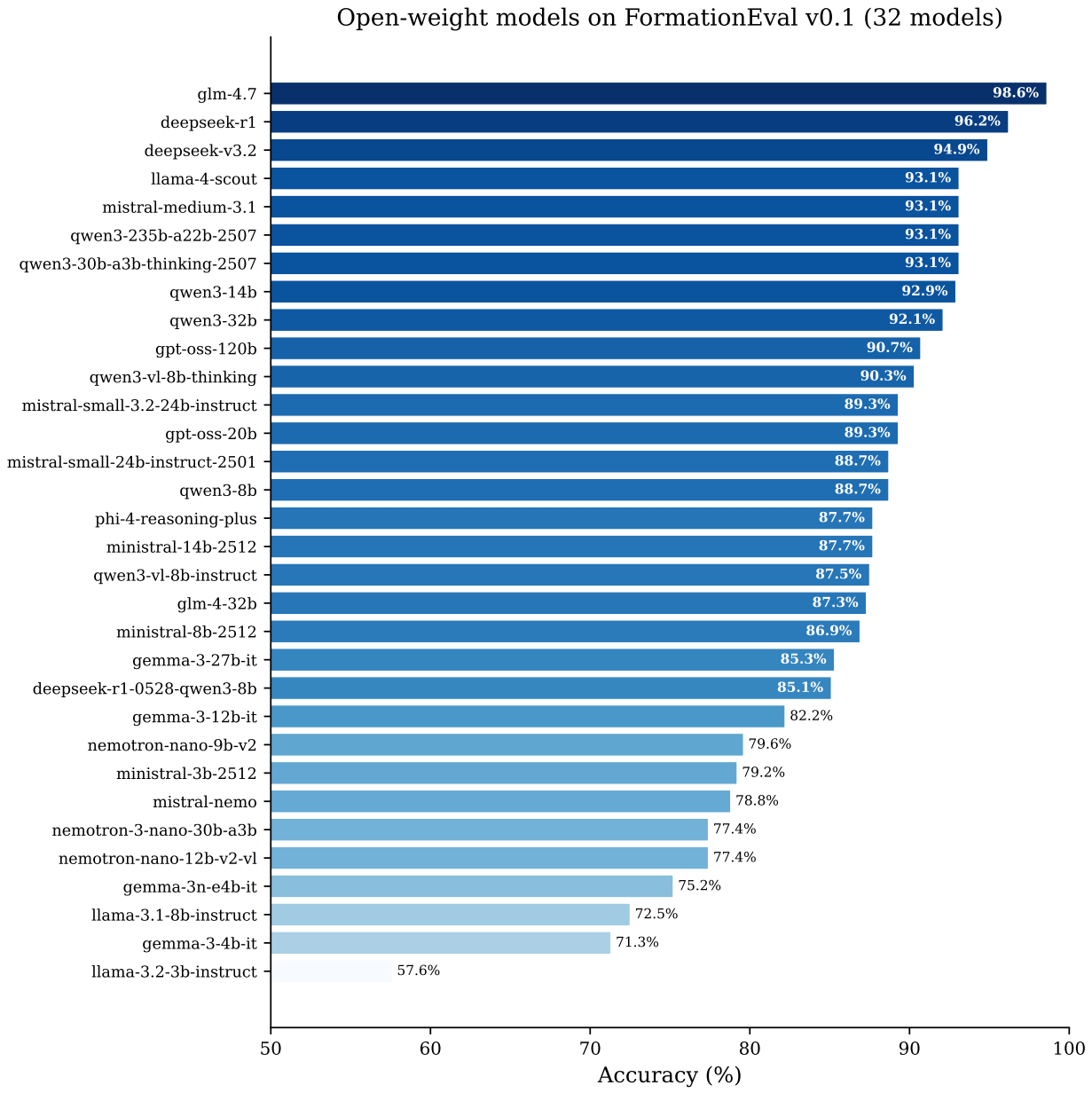

Figure 8: All 32 open-weight models ranked by accuracy. GLM-4.7 leads at 98.6%, followed by DeepSeek-R1 (96.2%) and DeepSeek-V3.2 (94.9%). Color intensity indicates accuracy level.