EXPERT-REVIEWED BENCHMARK AND DOMAIN-SPECIALIZED LLM FOR SUBSURFACE WELL-LOG INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Well-log interpretation requires reasoning over long, multivariate, depth-indexed measurements. The oil and gas industry relies on it to identify hydrocarbon-bearing zones, estimate reserves, and plan well completions. Although large language models (LLMs) could accelerate this process, how well they handle it is largely unknown. We introduce **WellLogBench**, the first expert-reviewed benchmark for this task, built with over 1,000 hours of expert review: 1,075 validated QA pairs across six petrophysical categories and three reasoning levels, from single-formation interpretation to cross-formation and multi-well synthesis. To make raw LAS logs usable by LLMs, we develop **Petrophys**[1], a Python petrophysics library that turns million-token depth-indexed logs into compact, lithology-aware structured contexts with reproducible calculations; Dynamic Feature Selection (DFS) then cuts the context supplied to the model by 73.5%. We also present **WellLog-LLM**, a domain-specialized model trained with direct preference optimization (DPO) and group relative policy optimization (GRPO), then refined through inference-time Group Relative Context Optimization (GRCO). Across 15 evaluated LLMs and VLMs, `GPT-5.6 Luna` (high) leads the public and private splits at 70.4 and 69.3 Overall score. WellLog-LLM reaches overall score of **66.0 and 66.4**, outperforming the 1T-parameter open-weight `Kimi-K2.6` (65.0 and 65.2) despite having **approximately 37× fewer parameters**. Under the same DFS context, it raises the Overall score of its `Qwen3.5-27B` base model from 56.3 to 66.0. We will open-source the benchmark, code, and model weights, providing a reproducible foundation for grounded petrophysical reasoning.

1 INTRODUCTION

Large Language Models (LLMs) have become increasingly capable reasoning assistants, with strong performance across scientific question answering, multi-step reasoning, and domain-specific tasks (Wei et al., 2022; Kojima et al., 2022; Qiu et al., 2025), and they have been adapted to specialized domains such as chemistry (Bran et al., 2024), finance (Wu et al., 2023), and time-series analysis (Jin et al., 2024). Yet many scientific workflows require reasoning over long, multivariate numerical measurements with domain-specific structure and dependencies (Kong et al., 2025; Wu et al., 2026). Answering questions about these measurements requires identifying relevant intervals, combining signals, and applying domain-specific calculations while maintaining consistency with the underlying evidence. How to enable LLMs to perform these operations reliably over long measurement records remains a challenge (Merrill et al., 2024; Chow et al., 2024).

Well-log interpretation, a central task in petrophysics, exemplifies the challenges of reasoning over long, multivariate scientific measurements. Well logs are depth-indexed petrophysical measurements used to characterize the subsurface and estimate properties such as lithology, porosity, fluid saturation, permeability, and reservoir quality. A single Log ASCII Standard (LAS) file can contain more than 30 interdependent curves and exceed **2M tokens** when represented directly in its raw format. Interpretation goes beyond retrieving values: an analyst must identify relevant depth intervals, account for missing or unreliable measurements, apply lithology-dependent calculations, and reconcile

[1]Anonymous project page: petrophys-anon.netlify.app.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

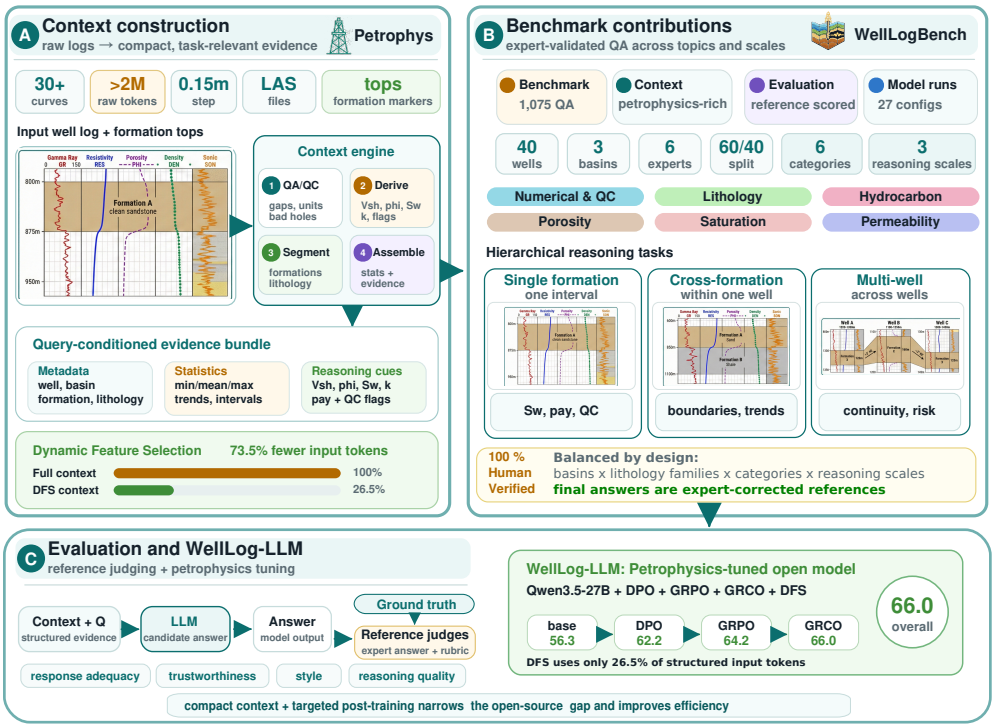


Figure 1: **Overview of WellLogBench and WellLog-LLM.** (A) Petrophys converts raw LAS logs and formation tops into quality-controlled, lithology-aware structured evidence; query-conditioned DFS reduces the mean context supplied to answer generation by 73.5%. (B) WellLogBench contains 1,075 expert-reviewed QA pairs from 40 wells and three basins, spanning six categories and three reasoning levels (worked examples in Appendix P). (C) Candidate answers are scored against expert-corrected references, and, with DFS context, the WellLog-LLM lineage progresses from 56.3 for the backbone to 62.2 after DPO, 64.2 after GRPO, and 66.0 with inference-time GRCO.

evidence across formations and wells. Several quantities of interest, such as shale volume, porosity, water saturation, and pay, must be derived through petrophysical calculations, and even seemingly plausible interpretations can still be inconsistent with the underlying measurements or physical constraints. Recent LLM systems for well logs orchestrate petrophysical tools or classify lithology (Chen et al., 2026; Cheng et al., 2026), and petroleum-geoscience benchmarks are knowledge exams (Ermilov, 2026); none evaluates long-form, interval-level and cross-well interpretations of LAS-derived evidence against expert references (Sec. 2). These tasks are a hard test of LLM reasoning.

To bridge this gap, we introduce **WellLogBench**, an expert-reviewed benchmark that contains 1,075 question–answer pairs derived from 40 public wells across three basins. It covers six petrophysical topics and three reasoning settings: Single-Formation, Cross-Formation, and Multi-Well (Figure 1). Initial questions and answers are generated under expert-designed templates, audited by a separate model, and manually reviewed or corrected by senior petrophysicists.

Our contributions are as follows:

1. **Benchmark** (Sec. 3): We introduce the first expert-reviewed benchmark for well-log reasoning: **1,075 QA pairs over 40 wells from three basins**, refined through more than **1,000 hours of collective expert review** and split 60/40 into public and private leaderboard sets.
2. **Evaluation** (Sec. 4): To reliably evaluate answers containing numerous numerical and derived quantities, we introduce a dual reference-based LLM-as-a-Judge protocol, including a dedicated numerical judge, and validate it against blinded expert ratings and alternative judges.
3. **Petrophys** (Sec. 3.2): We develop an open-source petrophysics library that transforms raw LAS logs into lithology-aware structured contexts through QA/QC, petrophysical derivations, segmentation, and rule-based flagging.

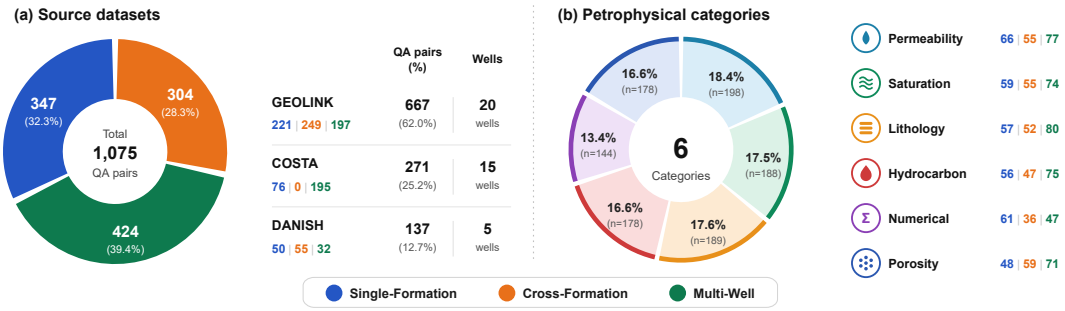


Figure 2: **WellLogBench composition.** The left panel summarizes the benchmark by reasoning level and source dataset; the right panel shows the six petrophysical categories and their Single-Formation, Cross-Formation, and Multi-Well counts.

4. **WellLog-LLM** (Secs. 5–6): We develop a 27B petrophysics-specialized model using DPO, GRPO, and GRCO. To enable efficient reasoning over long well-log contexts, we introduce DFS and hierarchical decomposition; the resulting model achieves an Overall score of 66.0, the highest among the evaluated open-weight models.

2 Related Work

Scientific and Geoscientific Reasoning. General science benchmarks (ScienceQA (Lu et al., 2022), MMLU (Hendrycks et al., 2021)) omit petrophysics and its data formats. Geoscience LLMs and benchmarks, such as K2 (Deng et al., 2024), GeoGalactica (Lin et al., 2024), OceanGPT (Bi et al., 2024), ClimaQA (Manivannan et al., 2025), and EarthSE (Xu et al., 2026), build questions from papers, textbooks, or figures, testing knowledge rather than reasoning over raw measurements.

AI in Petrophysics. Machine learning for well logs spans reservoir-property estimation (Hassaan et al., 2024), lithofacies classification (Shi et al., 2024; Jiang et al., 2024; Mohammadi et al., 2025), physics-informed models (Shao et al., 2024a; Pothana & Ling, 2025), well-log foundation models (Qi et al., 2025; Lasscock et al., 2025), and curve prediction and imputation (Rahmati et al., 2024; Zhang et al., 2025; Chen et al., 2025; Al-Fakih et al., 2025). LLMs enter petroleum workflows through domain pre-training (Eckroth et al., 2025), report drafting (Li et al., 2025), and, recently, agents that orchestrate log-processing tools or classify lithology (Chen et al., 2026; Luo & Xiao, 2026; Cheng et al., 2026); these are evaluated by task accuracy or case studies. Petroleum QA benchmarks such as FormationEval (Ermilov, 2026) and PetroBench (Wang et al., 2026) are multiple-choice or short-answer knowledge tests without measurement data; frontier models exceed 97% on the former. No prior benchmark evaluates long-form, multi-step *interpretation* of long, multivariate, LAS-derived contexts against expert references.

Long-Context Time-Series Reasoning. Well logs are ultra-long, multivariate depth-indexed series requiring interval aggregation, trend reasoning, and cross-curve consistency. Long-context benchmarks (LongBench (Bai et al., 2024), RULER (Hsieh et al., 2024)) are text-centric, time-series evaluations report persistent zero-shot reasoning gaps (Merrill et al., 2024; Chow et al., 2024; Kong et al., 2025), and SciTS (Wu et al., 2026) shows that serializing long scientific signals as text degrades LLMs, which motivates structured contexts. WellLogBench fills this gap with expert-reviewed reasoning tasks over ultra-long well-log contexts.

3 WellLogBench: Design and Data Creation

Overview. WellLogBench comprises 1,075 expert-reviewed QA pairs from 40 public LAS wells across three basins, covering six petrophysical categories and three reasoning levels. Items are generated with Gemini 2.5 Pro (Comanici et al., 2025) from lithology-aware contexts, audited by GPT-5 (OpenAI, 2025b), and validated by experts (Sec. 3.3). Three principles guide construction: (i) *domain authenticity*: questions drawn from real interpretation and reporting workflows; (ii) *reasoning complexity*: from direct retrieval and computation to multi-step interpretation with cross-curve

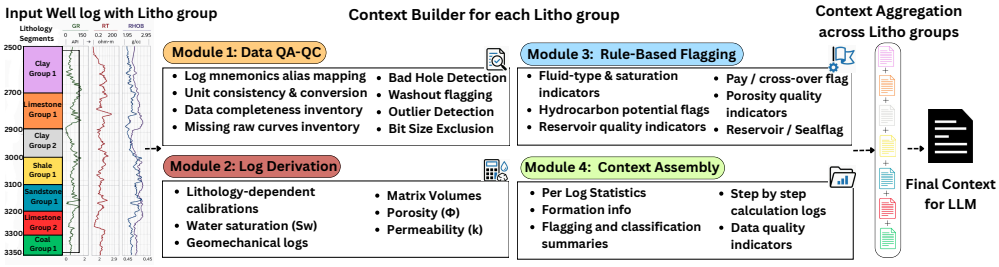


Figure 3: **Petrophys context-generation engine.** Input data sources are merged and segmented by lithology group. Each segment passes through Data QA/QC, Petrophysical Log Derivation, Rule-Based Flagging, and Statistics Assembly to produce the structured context.

consistency checks; and (iii) *geological diversity*: multiple basins, lithologies (carbonate, clastic, mixed), and acquisition conditions.

3.1 Data Sources, Categories, and Reasoning Levels

We curate three public repositories spanning clastic, carbonate, and chalk-dominated settings (Figure 2(a)): **Geolink** (North Sea, mixed clastic–carbonate, 667 pairs (Gama et al., 2024)), **COSTA** (Middle Eastern carbonate, 271 pairs (Costa Gomes et al., 2022)), and **Danish** (Upper Cretaceous–Danian chalk, 137 pairs (Geological Survey of Denmark and Greenland (GEUS), 2026)); formations, lithologies, and representative well-log plots are in Appendix B. Questions span six categories that exercise numerical computation, multivariate integration, causal inference, and decision-making under uncertainty: **Numerical & QC** (data-quality and artifact judgement), **Porosity** (estimation with lithology corrections), **Saturation** (resistivity-based fluid saturation), **Hydrocarbon** (zone identification and phase discrimination), **Permeability** (porosity–permeability flow capacity), and **Lithology** (rock-type classification with stratigraphic consistency). Each category is posed at three reasoning levels (Figure 2(b)): **Single-Formation** (analyse logs and derive properties within one stratigraphic interval), **Cross-Formation** (compare formations within one well, identifying boundaries and explaining variations), and **Multi-Well** (regional correlation and trend synthesis across wells for the same formation).

3.2 Context Engineering with Petrophys

Applying LLMs to well logs requires transforming high-dimensional, depth-indexed measurements into structured contexts that preserve petrophysical relationships within LLM context limits. To make this step reproducible, we developed **Petrophys**, a petrophysics library that will be released with WellLogBench. Figure 3 summarizes the framework; Appendix O shows an exported JSON schema.

Petrophys first applies **lithology-based segmentation**, because averaging properties across mixed lithologies yields misleading statistics. It aligns the log data to formation tops, maps lithology codes to standardized groups (Appendix B), and splits the formation into contiguous, lithologically homogeneous intervals with sequential identifiers, so that each segment is processed with its own lithology-specific parameters. Every segment then passes through four modules (Figure 3): **Data QA/QC** detects bad holes, inventories missing data, and filters outliers; **Petrophysical Log Derivation** computes shale volume, porosity, water saturation, and permeability with lithology-dependent calibrations; **Rule-Based Flagging** adds expert-validated flags such as RESFLAG, SEALFLAG, and PAY_FLAG; and **Statistics Assembly** summarizes each log with metadata and data-quality fields.

The segments are then assembled into one of three output modes. *Single-Formation Mode* aggregates the lithology groups into one context per formation, while *Cross-Formation* and *Multi-Well Modes* aggregate further for comparative questions (illustrated in Figure 23, Appendix P). A raw LAS file exceeds 2M tokens, whereas a structured single-formation context averages $\sim$60k tokens (Figure 5); cross-formation and multi-well contexts grow with the number of units compared (Sec. 5).

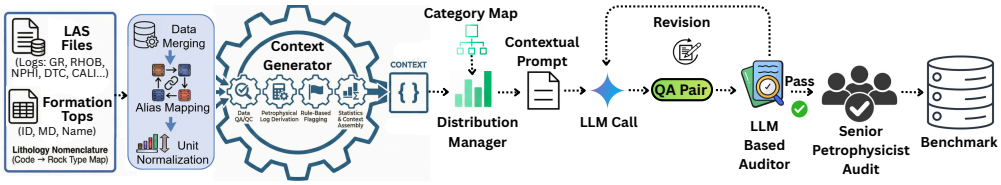


Figure 4: **QA generation and validation pipeline.** Raw LAS and formation data are converted into structured petrophysical contexts; model-generated QA pairs are then audited by GPT-5 and verified or corrected by senior petrophysicists before benchmark inclusion.

3.3 QA Generation and Expert Validation

Items follow the four-stage pipeline in Figure 4; ground-truth answers are expert-corrected and rubric-validated, not raw model outputs.[2] **(1) Ingestion.** LAS curves are merged with formation tops and lithology labels, with mnemonic mapping ($\sim$200 vendor mnemonics $\rightarrow$ 25 canonical curves) and unit normalization (m, v/v, Ωm, g/cm^3, μs/ft). **(2) Context and QA generation.** Each well becomes a structured context (Sec. 3.2); a *Category Map*, designed with the expert panel, defines templates (required curves, derived quantities, example questions, difficulty), and a *Distribution Manager* balances categories (13.4% Numerical to 18.4% Permeability), reasoning levels, basins, and lithologies by weighted sampling. **(3) Auditing.** A GPT-5 auditor validates each pair, revises failures, and rejects unresolved cases; it revises 71.8% / 77.4% / 76.0% of single/cross/multi items, so the benchmark is not a mirror of generator outputs. **(4) Expert verification.** Six senior petrophysicists (15+ years) verify technical accuracy and correct or reject errors, correcting 59.2% / 61.0% / 59.5% of items (auditor or expert: 81.4% / 86.4% / 85.2%; Appendices C and D).

4 Evaluation Protocol

4.1 Judges and Inputs

Each item includes a question-specific, expert-corrected *ground-truth answer*. Because answers are long-form interpretations that admit many valid phrasings, we score them against this reference with a dual LLM-as-a-Judge protocol (Zheng et al., 2023) that splits the evaluation by the evidence each dimension needs. A primary judge compares the candidate answer with the ground truth across the main evaluation dimensions, without the long log context. Numerical claims, however, often concern values that the reference omits but the logs determine, so a separate numerical judge scores *Numerical Accuracy* from the question, ground truth, candidate answer, and log context. Judge models and prompts are given in Appendices F and Q. All scores are reported on a 0–100 scale.

4.2 Rubrics and Aggregation

1. **Response Adequacy (RA).** Does the answer match the expert reference? *Accuracy* (agreement with the GT conclusion), *Petrophysical Claims* (cited logs and quantities are supported by GT), *Conclusion* (final interpretation matches GT), and *Coverage* (no required GT element omitted).
2. **Trustworthiness (Trust).** Is every claim anchored in the context? *Faithfulness* (claims verifiable against the context), *Numerical Accuracy* (depths, cutoffs, and derived quantities are correct), *Data Limitations* (uncertainty acknowledged where GT does), and *Log Utilization* (logs the GT relies on are invoked).
3. **Style.** Is the answer well communicated? *Clarity*, *Conciseness*, and *Repetition* are scored per section.
4. **Petrophysical Reasoning Quality (RQ).** Is the reasoning a coherent petrophysical argument, with each claim supported by the logs and an internally valid derivation, independent of the RA reference match?

Overall is 40% RA + 20% Trust + 20% Style + 20% RQ. RA is upweighted because task success outranks fluency; equal weighting (25% each) yields near-identical rankings (r=0.999, ρ=1.000; Appendix G).

[2] All prompt templates (question and answer generation, inference, DFS, and judges) are in Appendix Q.

4.3 Judge Validation

Two checks support the automated judge. Senior petrophysicists blindly scored 50 Gemini 2.5 Pro vs. GPT-5 answer pairs, and expert and LLM-judge ratings show good absolute agreement, with $\mathrm{ICC}(\mathrm{A},1)$=0.721 across shared rubric metrics (0.697 for Gemini 2.5 Pro; 0.683 for GPT-5). Experts and the LLM judge also agree on the overall preference, Gemini 2.5 Pro > GPT-5, with similar margins (Expert Δ=+11.6, LLM Δ=+9.9). Full ICC, Pearson, and per-metric analyses appear in Appendix E. Re-scoring the shared model cohort with three alternative judges, Qwen3 235B A22B Inst., Gemma 4 31B IT, and Qwen3.5-27B, yields Pearson $r \geq 0.91$ on RA/Trust/RQ and model-rank $\rho \geq 0.96$ (Appendix F).

5 Managing Long Contexts: DFS and Hierarchical Decomposition

Even after context engineering, single-formation contexts (∼60k tokens on average, some exceeding 120k) and multi-formation contexts can exceed practical LLM budgets. We use two complementary query-conditioned strategies in both training and inference: *Dynamic Feature Selection* for single-formation items and *hierarchical decomposition* for cross-formation and multi-well items.

5.1 Dynamic Feature Selection (DFS)

An intermediate LLM call receives only the structured context *schema* (its key paths, without values) together with the target question and selects the task-relevant key paths; the corresponding values form a compact context that is fed to the same model for answer generation:

$$\text{Context}_{\text{Schema}} + Q \xrightarrow{\text{LLM}} \text{Context}_{\text{compact}} \xrightarrow{\text{LLM}} A. \quad (1)$$

This query-conditioned compression preserves task relationships while cutting single-formation input tokens by 73.5% on average (Figure 5). Because the selector reads only the schema, neither call processes the full context. Among the Table 1 systems, only WellLog-LLM uses DFS; all others receive the full structured context. Appendix I applies DFS to them.

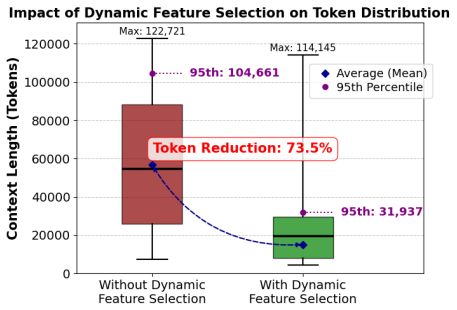


Figure 5: **DFS context compression** of single-formation contexts (boxes: IQR and median; diamonds: means; circles: 95th percentile). DFS lowers the mean context from 56.6k to 15.0k tokens (73.5%) and the 95th percentile from 104.7k to 31.9k.

5.2 Hierarchical Cross-Formation & Multi-Well Decomposition

Concatenating k formation/well contexts would exceed practical context budgets. We instead decompose the query into two explicitly question-conditioned stages. Stage 1 processes each formation or well independently and extracts a local finding F_i; Stage 2 receives the same question and the set of findings, then reconciles evidence across units into a comparative answer:

$$F_i = \text{LLM}(\text{Context}_i, Q) \;\; (i = 1, \ldots, k), \qquad A = \text{LLM}\left(\{F_i\}_{i=1}^{k}, Q\right). \quad (2)$$

The Stage 1 calls can be executed independently, while the synthesis input grows with compact findings rather than full contexts. This map-then-synthesise pattern therefore scales approximately linearly in k and keeps every per-unit context within the model's native limit.

6 WellLog-LLM: Petrophysics Post-Training

To test whether targeted post-training can close the gap between a mid-scale open-weight model and closed-source systems, we instantiate **WellLog-LLM** from Qwen3.5-27B. Figure 6 summarizes the complete lineage: DPO and GRPO successively update model weights, whereas GRCO keeps the GRPO checkpoint frozen and adds a learned library of experience rules to the prompt at inference time. All post-training data are built with the Petrophys pipeline from LAS files of wells outside WellLogBench, so no benchmark well, question, or reference answer is seen during training.

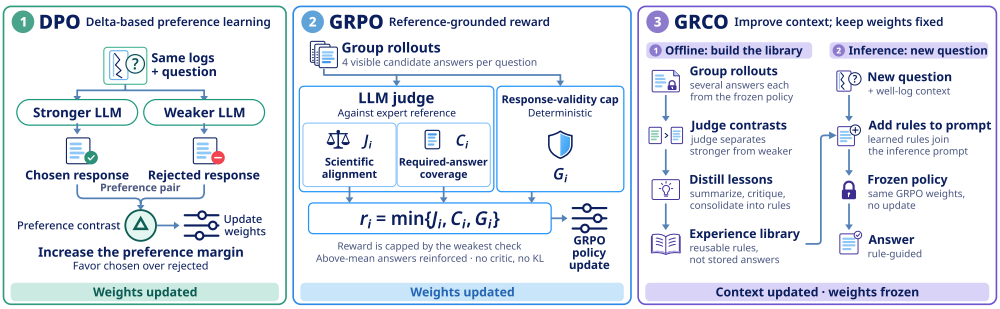


Figure 6: **WellLog-LLM post-training and inference pipeline.** DPO learns from preferred–rejected pairs; GRPO (Sec. 6.2) updates the policy with four rollouts per prompt and the reward in Eq. 3; GRCO freezes that checkpoint, distills rollout contrasts into an experience library, and adds the relevant rules to the prompt at inference.

6.1 DPO Initialization

We initialize WellLog-LLM with DPO (Rafailov et al., 2023) on the base model, following the delta-learning hypothesis (Geng et al., 2025): preference tuning improves a model through the quality *difference* between the chosen and rejected answers, so neither answer needs to be expert-written. We build ∼15k pairs (5k per reasoning level; Appendix J.2) in which both answers respond to the same DFS-formatted input, the chosen one from Kimi-K2.6 (Moonshot AI, 2026) and the rejected one from Qwen3-8B. Fixing the input ties the delta to how the evidence is interpreted rather than to what evidence is available, and pairing models that score 65.0 and 46.1 Overall on WellLogBench (Table 1) keeps it large. DPO thus raises the margin for answers that use the supplied logs, ground their conclusions, and avoid unsupported petrophysical claims, without imitating the chosen answers token by token. SFT on Kimi-K2.6 answers, tried first (Appendix J.1), instead lowered Overall to 51.1 (Table 2), in line with evidence that SFT can erode reasoning-tuned models (Chu et al., 2025; Huan et al., 2026).

6.2 GRPO Reward and Policy Update

Starting from the DPO checkpoint, we apply GRPO (Shao et al., 2024b) with the corrections of Liu et al. (2025), training for 480 steps with a batch size of four prompts on 2k QA pairs drafted with Kimi-K2.6 and balanced across reasoning types and categories. For each question the policy samples $K{=}4$ answers; hidden reasoning is discarded at the `</think>` boundary, so only the visible answer is rewarded. Each answer receives three scores: *scientific alignment* $J_i \in [0,1]$, a rubric score from a frozen LLM judge that sees only the question, the reference final answer, and the candidate answer; *required-answer coverage* $C_i \in [0,1]$, the fraction of ground-truth answer segments present, estimated by the same judge; and a deterministic *response guard* $G_i \in [0,1]$, which is 1 unless the answer is empty, malformed, truncated, repetitive, or leaks reasoning. The reward is the weakest of the three, and the advantage is the reward relative to the group mean:

$$r_i = \min\{J_i,\, C_i,\, G_i\}, \qquad A_i = r_i - \bar{r}. \tag{3}$$

Taking the minimum means fluent prose cannot hide a scientific, completeness, or formatting failure. A_i is not divided by the group standard deviation; it is applied to every response token and optimized with a clipped PPO-style objective (Schulman et al., 2017), with no critic and no KL term. Training dynamics, hyperparameters, and guard caps are in Appendix J.3.

6.3 GRCO: Inference-Time Experience Refinement

Much of what separates a strong log interpretation from a weak one comes down to general petrophysical rules: which log responses justify calling a zone hydrocarbon-bearing, how to weigh logs that disagree, and when a missing log is only missing evidence rather than a negative finding. A petrophysicist can state each rule in one sentence, and each holds from well to well. GRPO learns such rules only indirectly: its reward scores whole answers, never which rule made one answer better. Group Relative Context Optimization (GRCO) learns them directly: adapting training-free GRPO (Cai et al., 2025), it freezes the GRPO weights and learns a natural-language rule library.

Table 1: **WellLogBench public leaderboard** (0–100; rubrics in Sec. 4). **Params**: reported total parameters ("—" if undisclosed). Other models use full structured context; WellLog-LLM uses DFS, requiring only ∼27% of the input tokens. *Best per column in **bold**, second best underlined.*

Model	Params	Ctx.	RA	Trust	Style	RQ	Overall
Closed-source							
GPT-5.6 Luna (high) (OpenAI, 2026)	—	Full	**67.6**	**70.6**	64.1	**82.0**	**70.4**
Gemini 2.5 Pro (Comanici et al., 2025)	—	Full	57.2	68.0	**76.0**	75.4	66.8
GPT-5 (OpenAI, 2025b)	—	Full	59.1	65.1	60.5	78.3	64.4
Open-weight							
Kimi-K2.6 (Moonshot AI, 2026)	1T	Full	59.4	65.2	65.6	75.5	65.0
Gemma 4 31B IT (Google DeepMind, 2026)	31B	Full	51.3	62.5	66.6	69.5	60.2
Qwen3.5-27B (Qwen Team, 2026)	27B	Full	48.6	60.2	59.8	66.9	56.8
Qwen3 235B A22B Inst. (Yang et al., 2025)	235B	Full	43.0	55.6	57.7	63.0	52.5
GPT-OSS-120B (OpenAI, 2025a)	117B	Full	43.5	55.6	57.4	61.1	52.2
Qwen3 Next 80B A3B (Qwen Team, 2025)	80B	Full	40.0	53.1	58.5	58.1	49.9
LLaMA-4 Scout (Meta, 2025)	109B	Full	35.5	49.2	63.9	51.3	47.1
Qwen3-8B (Yang et al., 2025)	8B	Full	37.7	50.4	53.8	51.1	46.1
LLaMA-3.3 70B Inst. (Meta, 2024b)	70B	Full	28.1	45.5	60.2	46.5	41.7
Qwen2.5-7B Instruct (Qwen Team, 2024)	7B	Full	27.2	41.3	48.6	39.9	36.8
LLaMA-3.1 8B Instruct (Meta, 2024a)	8B	Full	20.8	35.8	43.7	34.7	31.2
WellLog-LLM	27B	DFS	61.3	66.2	64.8	76.3	66.0

The learning loop mirrors GRPO. On 900 training QA pairs drafted with GPT-5.6 Luna (high), the policy answers each question several times and a judge scores every answer against its reference. We keep only groups with both strong and weak answers, as their contrast plays the role of the advantage; a meta model explains why the strong answers won and turns this into library edits, the analogue of the parameter update. At inference, the library enters the prompt where the final interpretation is formed: Stage 1 for single-formation questions and the Stage-2 synthesis for multi-well and cross-formation questions. Although learned on wells outside the benchmark, the rules raise Response Adequacy by 2.3 points over the GRPO model (Sec. 7.2), showing that they transfer; as plain text, every rule can also be checked by a domain expert (Appendix J.4).

7 Results

7.1 Overall Public Leaderboard

We evaluate on the **60% public split** (646 questions); the 429-question private split is held out for the leaderboard (Appendix L). Table 1 reports all structured-context systems. Scores are question-deduplicated means over non-error rows with complete rubric aggregates (632–646 valid questions per model; roster in Appendix H); GPT-5 is scored from split-level rubric aggregates (private split: 64.5).

GPT-5.6 Luna (high) leads at 70.4 Overall, followed by Gemini 2.5 Pro at 66.8. **WellLog-LLM is the strongest evaluated open-weight system at 66.0**, 1.0 point above Kimi-K2.6 (65.0; significant under a paired bootstrap, Appendix L) despite using 27B rather than 1T parameters. GPT-5 reaches 64.4, and the untuned Qwen3.5-27B full-context baseline reaches 56.8. The ranking is essentially unchanged under equal rubric weights and on the held-out private split (Appendices G and L). Applying DFS to the other models lowers their Overall by only 0.9 points on average (Appendix I), so we report them with full context, their stronger setting. Structured context also matters: in a separately scored study with six VLM families, models reading only rendered log plots fall to 17–23 in Response Adequacy and Trust, whereas adding the Petrophys context restores grounding (Appendix N, Table 24, Figure 22).

Beyond aggregate scores, judge-coded failure rates show that post-training chiefly corrects interpretation: relative to its backbone under the same DFS context, WellLog-LLM more than halves wrong conclusions (43.4%→18.9%) and omissions (29.9%→11.5%), and its wrong-conclusion rate is lower than those of Kimi-K2.6 (21.1%) and Gemini 2.5 Pro (26.8%). Numerical errors remain

near zero (0.3%, vs. 0.8% for the full-context backbone), showing that the compact DFS context does not compromise quantitative accuracy (Appendix K.4).

7.2 Stage Attribution

Table 2 traces the Qwen3.5-27B lineage through DPO, GRPO, and GRCO. Because all stages share the same DFS context, the differences isolate post-training. Every stage carried forward improves Overall, and the gains hold when adjacent stages are compared on the same questions (Appendix K.2). DPO gives the largest gain, raising Overall from 56.3 to 62.2 with every rubric improving and Response Adequacy most. GRPO then adds 1.9 points, mainly through answers that better match the expert reference (higher RA and RQ). GRCO adds another 1.9 points without changing weights, with its largest gain in Trust: the experience rules help the model state data limitations instead of guessing. Appendix K.2 also breaks down each stage by sub-metric and reasoning type and traces one question through all stages.

Table 2: **Stage attribution** (public split, DFS context). Δ: unrounded Overall change from the preceding stage. †Not carried forward; DPO starts from the base model.

Stage	RA	Trust	Style	RQ	Overall	Δ
Qwen3.5-27B (DFS)	47.2	59.4	59.8	67.6	56.3	—
+ SFT†	47.5	46.9	51.9	61.8	51.1	−5.2
+ DPO	56.3	62.6	64.1	72.0	62.2	+6.0
+ GRPO	59.0	63.4	64.7	74.6	64.2	+1.9
+ GRCO	61.3	66.2	64.8	76.3	66.0	+1.9

7.3 Performance Across Basins, Categories, and Reasoning Types

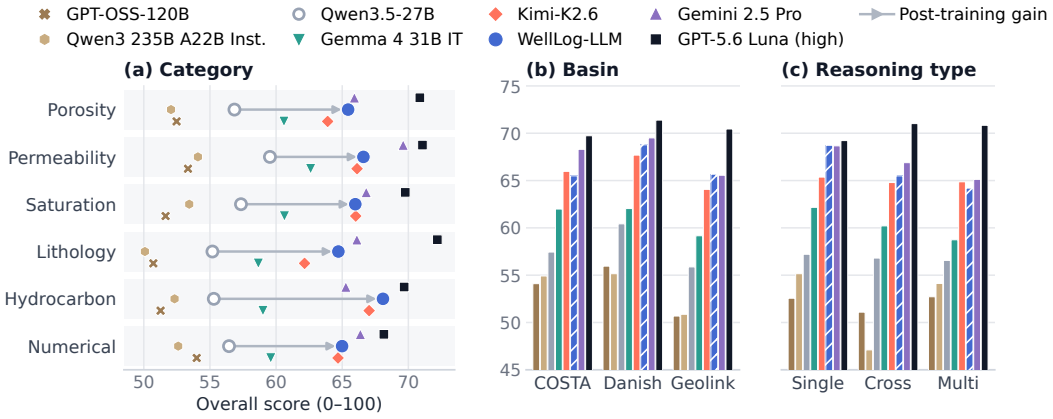


Figure 7: **Public-split performance slices.** Overall score (0–100) by (a) category, where arrows run from the Qwen3.5-27B backbone to WellLog-LLM, (b) basin, and (c) reasoning type. GPT-OSS-120B and Qwen3 235B A22B Inst. are larger open-weight models that score below the backbone. Other models use full context; WellLog-LLM (striped) uses DFS. Table 14 lists all models.

In Figure 7, WellLog-LLM performs consistently across categories (64.7–68.1), is the strongest open-weight system on most slices, and matches or exceeds the closed-source Gemini 2.5 Pro on Hydrocarbon, Geolink, and Single-Formation questions (all models: Appendix K.1).

8 Conclusion

We present WellLogBench, a benchmark for LLM-based well-log reasoning with 1,075 expert-reviewed QA pairs, together with the Petrophys context pipeline, a structured evaluator, and WellLog-LLM. GPT-5.6 Luna (high) leads the public leaderboard at 70.4 Overall, and our 27B system reaches 66.0 through DPO → GRPO → GRCO, surpassing the 1T-parameter Kimi-K2.6 (65.0). Each stage improves the model, and together they halve wrong conclusions and omissions relative to the backbone while keeping numerical errors near zero, showing that targeted post-training on structured petrophysical context is an effective route to grounded well-log interpretation. Our work lays a foundation for LLM-based well-log interpretation, a core task in the oil and gas industry.

Ethics Statement

Data and privacy. We use only public datasets (Geolink, COSTA, Danish) under open-data licenses; no proprietary or personally identifiable information is included.

Safety. Subsurface interpretation carries operational and environmental risk, so benchmarked systems should remain expert-assistance tools with human oversight, not autonomous decision-makers.

Fair labor. Domain experts involved in template design, review, and validation were fairly compensated.

Reproducibility Statement

Under permissive open-data/open-source licenses, the release will include:

- the 646-item public split with contexts, expert ground truths, rubric annotations, and metadata;
- versioned source-data snapshots/URLs and the Petrophys context-engineering pipeline (Sec. 3.2; petrophys-anon.netlify.app);
- LLM-as-a-Judge prompts, rubric definitions, and the complete scoring protocol;
- WellLog-LLM weights and the SFT, DPO, GRPO, and GRCO recipes; and
- an evaluation server for the 429-item private split, allowing progress tracking without exposing private answers.

The full SFT, DPO, GRPO, and GRCO configuration is documented in Appendix J.

AI Use Statement

In this work, we used generative AI tools for tasks with required disclosure: generating synthetic data (drafting benchmark QA pairs before GPT-5 auditing and expert review, and drafting post-training data), implementing methods (writing and debugging code for experiments, statistical analysis, and figures), and supporting qualitative data analysis (LLM judges score answers and code their failure modes; Sec. 4 and Appendix K.4). Generative models are also components of the evaluated methodology itself, serving as the auditor, the reward and evaluation judges, and the GRCO meta model. Theoretical models, mathematical claims, proofs, and translation are not applicable to this work. Additionally, we used generative AI tools for language editing and manuscript organization, LaTeX formatting, and checking references against their published records. We have reviewed all AI-assisted work: domain experts reviewed or corrected every benchmark item, the authors verified AI-assisted code by checking its outputs against the underlying evaluation records, and every citation was checked against its published source. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Abdulrahman Al-Fakih, A. Koeshidayatullah, Tapan Mukerji, Sadam Al-Azani, and SanLinn I. Kaka. Well log data generation and imputation using sequence based generative adversarial networks. *Scientific Reports*, 15:11000, 2025. doi: 10.1038/s41598-025-95709-0.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. OceanGPT: A large language model for ocean science tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3357–3372. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.184.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, Yong Mao, Ke Li, and Xing Sun. Training-free group relative policy optimization. *arXiv preprint arXiv:2510.08191*, 2025. URL https://arxiv.org/abs/2510.08191.

Zhangxing Chen, Yongan Zhang, Jian Li, Gang Hui, Youzhuang Sun, Yizheng Li, Yuntian Chen, and Dongxiao Zhang. Artificial intelligence large model for logging curve reconstruction. *Petroleum Exploration and Development*, 52(3):842–854, 2025. doi: 10.1016/S1876-3804(25)60607-0.

Zhangxing Chen, Ruichen Ding, Yang Meng, Yizheng Li, Junwei Zhang, Liu Cao, Jian Li, Wenqi Fan, Yiyuan Zhang, Liqiu Wang, Dongxiao Zhang, and Yuntian Chen. A multi-agent collaborative framework with human-in-the-loop for well log interpretation and applications. *Petroleum Exploration and Development*, 53(3):900–914, 2026. doi: 10.1016/S1876-3804(26)60734-3.

Mingyue Cheng, Yitong Zhou, Jiahao Wang, Qingyang Mao, and Qi Liu. GeoMind: An agentic workflow for lithology classification with reasoned tool invocation. arXiv preprint arXiv:2604.21501, 2026. URL https://arxiv.org/abs/2604.21501.

Winnie Chow, Lauren Gardiner, Haraldur T. Hallgrímsson, Maxwell A. Xu, and Shirley You Ren. Towards time series reasoning with LLMs. arXiv preprint arXiv:2409.11376, 2024. URL https://arxiv.org/abs/2409.11376.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10818–10838, 2025. URL https://proceedings.mlr.press/v267/chu25c.html.

Domenic V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290, 1994. doi: 10.1037/1040-3590.6.4.284.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. doi: 10.48550/arXiv.2507.06261. URL https://arxiv.org/abs/2507.06261.

Jorge Costa Gomes, Sebastian Geiger, and Daniel Arnold. COSTA model: Hierarchical carbonate reservoir benchmarking case study. Dataset, Heriot-Watt Research Portal, 2022. URL https://researchportal.hw.ac.uk/en/datasets/costa-model-hierarchical-carbonate-reservoir-benchmarking-case-st/.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, 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 '24)*, pp. 161–170. ACM, 2024. doi: 10.1145/3616855.3635772.

J. Eckroth, J. Boden, L. Hough, H. Gatewood, S. Gipson, E. Schoen, and B. Gunderson. Building "EnergyLLM": A domain-specific large language model trained on SPE content. In *SPE Annual Technical Conference and Exhibition*, pp. D021S020R009, Houston, Texas, USA, 2025. SPE. doi: 10.2118/228097-MS. Paper Number: SPE-228097-MS.

Almaz Ermilov. FormationEval, an open multiple-choice benchmark for petroleum geoscience. arXiv preprint arXiv:2601.02158, 2026. URL https://arxiv.org/abs/2601.02158.

Pedro H. T. Gama, Jackson Faria, Jessica Sena, Francisco Neves, Vinícius R. Riffel, Lucas Perez, André Korenchendler, Matheus C. A. Sobreira, and Alexei M. C. Machado. Processed datasets – imputation in well log data: A benchmark. Zenodo, 2024. URL https://doi.org/10.5281/zenodo.10987946. Version v1.

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. In *Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=9rwtezthwo.

Geological Survey of Denmark and Greenland (GEUS). Subsurface data Denmark. Web portal, 2026. URL https://eng.geus.dk/products-services-facilities/data-and-maps/subsurface-data-denmark/. Accessed September 2026.

Google DeepMind. Gemma 4 31B Instruct (model card). Hugging Face, 2026. URL https://huggingface.co/google/gemma-4-31B-it.

Said Hassaan, Abdulaziz Mohamed, Ahmed Farid Ibrahim, and Salaheldin Elkatatny. Real-time prediction of petrophysical properties using machine learning based on drilling parameters. *ACS Omega*, 9:17066–17075, 2024. doi: 10.1021/acsomega.3c08795.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Maggie Ziyu Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general LLM capabilities? Understanding transferability of LLM reasoning. In *Forty-third International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2507.00432. URL https://icml.cc/virtual/2026/poster/65125.

Shiyi Jiang, Panke Sun, Fengqing Lyu, Sicheng Zhu, Ruifeng Zhou, Bin Li, Taihong He, Yujian Lin, Yining Gao, Wendan Song, and Huaimin Xu. Machine learning (ML) for fluvial lithofacies identification from well logs: A hybrid classification model integrating lithofacies characteristics, logging data distributions, and ML models applicability. *Geoenergy Science and Engineering*, 233: 212587, 2024. doi: 10.1016/j.geoen.2023.212587.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01728.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-MQA: Time series multi-task question answering with context enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29736–29753. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1437. URL https://aclanthology.org/2025.acl-long.1437/.

Ben Lasscock, Altay Sansal, Keyla Gonzalez, and Alejandro Valenciano. Well log foundation model: Making promptable AI models for interpretation. In *Fifth International Meeting for Applied Geoscience & Energy (IMAGE 2025)*, pp. 2910–2914. Society of Exploration Geophysicists, 2025. doi: 10.1190/image2025-4311806.1.

Pin Li, Yubo Liu, Jun Zhou, Juan Zhang, and Yuange Zhou. A report generation system for well logging industry based on large language model and retrieval augmented generation. In *International Petroleum Technology Conference*, pp. D012S004R002, Kuala Lumpur, Malaysia, 2025. IPTC. doi: 10.2523/IPTC-24775-EA. Paper IPTC-24775-EA.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, Boyi Zeng, Qiyuan Chen, Yuxun Miao, Bo Xue, Shu Wang, Luoyi Fu, Weinan Zhang, Junxian He, Yunqiang Zhu, Xinbing Wang, and Chenghu Zhou. GeoGalactica: A scientific large language model in geoscience. arXiv preprint arXiv:2401.00434, 2024. URL https://arxiv.org/abs/2401.00434.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Gang Luo and Li-Zhi Xiao. Large language model-based AI agent for well logging data processing and interpretation. *Petroleum Science*, 2026. doi: 10.1016/j.petsci.2026.05.031. In press.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. ClimaQA: An automated evaluation framework for climate question answering models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.16701.

Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996. doi: 10.1037/1082-989X.1.1.30.

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3512–3533, 2024. doi: 10.18653/v1/2024.findings-emnlp.201. URL https://aclanthology.org/2024.findings-emnlp.201/.

Meta. Llama-3.1-8B-Instruct (model card). Hugging Face, 2024a. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Meta. Llama-3.3-70B-Instruct (model card). Hugging Face, 2024b. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Meta. Llama-4-Scout-17B-16E-Instruct (model card). Hugging Face, 2025. URL https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct.

Kamyar Mohammadi, Seyed Hamid Reza Mousavi, Seyed Mojtaba Hosseini-Nasab, and Mohammad Shahbazi. Transforming petrophysical well-logs to images: leveraging deep learning for lithology recognition. *Earth Science Informatics*, 18(3):457, 2025. doi: 10.1007/s12145-025-01953-3.

Moonshot AI. Kimi-K2.6 (model card). Hugging Face, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025a. doi: 10.48550/arXiv.2508.10925. URL https://arxiv.org/abs/2508.10925.

OpenAI. GPT-5 system card. System card, 2025b. URL https://cdn.openai.com/gpt-5-system-card.pdf.

OpenAI. GPT-5.6 Luna model. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116617–116637, 2024. doi: 10.52202/079017-3702.

Prasad Pothana and Kegang Ling. Physics-integrated neural networks for improved mineral volumes and porosity estimation from geophysical well logs. *Energy Geoscience*, 6(2):100410, 2025. doi: 10.1016/j.engeos.2025.100410.

Zhenyu Qi, Qing Yu, Jichen Wang, Yun-Bo Zhao, Zerui Li, and Wenjun Lv. WLFM: A well-logs foundation model for multi-task and cross-well geological interpretation. *arXiv preprint arXiv:2509.18152*, 2025. doi: 10.48550/arXiv.2509.18152. URL https://arxiv.org/abs/2509.18152.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, 16:9799, 2025. doi: 10.1038/s41467-025-64769-1.

Qwen Team. Qwen2.5-7B-Instruct (model card). Hugging Face, 2024. URL https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

Qwen Team. Qwen3-Next-80B-A3B-Instruct (model card). Hugging Face, 2025. URL https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct.

Qwen Team. Qwen3.5-27B (model card). Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.5-27B.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Mehdi Rahmati, Ghasem Zargar, and Abbas Ayatizadeh Tanha. Predicting density log from well log using machine learning techniques and heuristic optimization algorithm: A comparative study. *Petroleum Research*, 9(2):176–192, 2024. doi: 10.1016/j.ptlrs.2024.01.008.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Rongbo Shao, Hua Wang, and Lizhi Xiao. Reservoir evaluation using petrophysics informed machine learning: A case study. *Artificial Intelligence in Geosciences*, 5:100070, 2024a. doi: 10.1016/j.aiig.2024.100070.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Yu Shi, Junqiao Liao, Lu Gan, and Rongjiang Tang. Lithofacies prediction from well log data based on deep learning: A case study from Southern Sichuan, China. *Applied Sciences*, 14(18):8195, 2024. doi: 10.3390/app14188195.

Xiang Wang, Tingting Zhang, Sen Wang, Ying Wu, Heng Meng, Peng Zhou, and Peng Li. PetroBench: A benchmark for large language models in petroleum engineering. arXiv preprint arXiv:2605.28032, 2026. URL https://arxiv.org/abs/2605.28032.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023. doi: 10.48550/arXiv.2303.17564. URL https://arxiv.org/abs/2303.17564.

Wen Wu, Ziyang Zhang, Liwei Liu, Xuenan Xu, Jimin Zhuang, Ke Fan, Qitan Lv, Junlin Liu, Chen Zhang, Zheqi Yuan, Siyuan Hou, Tianyi Lin, Kai Chen, Bowen Zhou, and Chao Zhang. SciTS: Scientific time series understanding and generation with LLMs. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.03255.

Wanghan Xu, Xiangyu Zhao, Yuhao Zhou, Xiaoyu Yue, Ben Fei, Fenghua Ling, Wenlong Zhang, and Lei Bai. EarthSE: A benchmark evaluating Earth scientific exploration capability for large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jyYE06FL8G.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Ji Zhang, Guiping Liu, Zhen Wei, Shengge Li, Yeheya Zayier, and Yuanfeng Cheng. Machine learning-based prediction of well logs guided by rock physics and its interpretation. *Sensors*, 25 (3):836, 2025. doi: 10.3390/s25030836.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

APPENDIX ROADMAP

The supplement proceeds from benchmark provenance and evaluator validation to supplementary model analyses, training details, auxiliary modalities, and reproducibility material.

- **Limitations** (Appendix A): coverage, reference, and judging caveats.
- **Benchmark construction** (Appendices B–D): basin geology, expert annotation workflow, and QA audit transparency.
- **Evaluation methodology** (Appendices E–G): expert–judge alignment, cross-judge robustness, and scoring-weight sensitivity.
- **Configurations and training** (Appendices H–J): structured-context roster, DFS sensitivity for external models, and the complete SFT–DPO–GRPO–GRCO recipe.
- **Supplementary results** (Appendices K–L): stage and slice patterns, public qualitative diagnostics, and private-split stability.
- **Auxiliary modalities** (Appendices M–N): modality reference and a visual/multimodal study.
- **Reproducibility artifacts** (Appendices O–Q): an exported context schema, qualitative examples with complete answers, and prompt templates.

A LIMITATIONS

WellLogBench is built from 40 wells across three basins, so it does not cover the full range of geology, tool suites, and logging conditions seen in practice. This is largely because most well data are proprietary and not publicly available, and because expert review is costly: writing and validating reference answers requires experienced petrophysicists, which limits how many wells and questions can be curated. Some questions depend on assumptions (e.g., petrophysical parameters and cutoffs), and reasonable experts may make slightly different choices, so a single reference answer will not capture every valid perspective. A natural next step is to integrate well logs with complementary data sources: formation pressure measurements (e.g., to constrain fluid contacts and pressure regimes), seismic data (e.g., well-to-seismic ties for lateral continuity), and core and production data. A future benchmark could then evaluate whether models reconcile evidence across these sources. We expect this kind of multi-source integration to become a central part of evaluating petrophysical reasoning.

B DETAILED BASIN INFORMATION

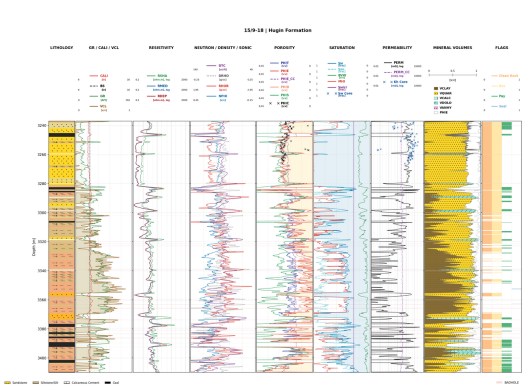
To test LLM reasoning in different geological settings, **WellLogBench** draws on three public sources: the Geolink dataset (North Sea), the COSTA model (carbonate), and the Danish dataset from the Subsurface Data Denmark portal. Together they cover clastic, carbonate, and chalk settings, each with its own interpretation problems.

B.1 DATA SOURCES AND STRATIGRAPHY

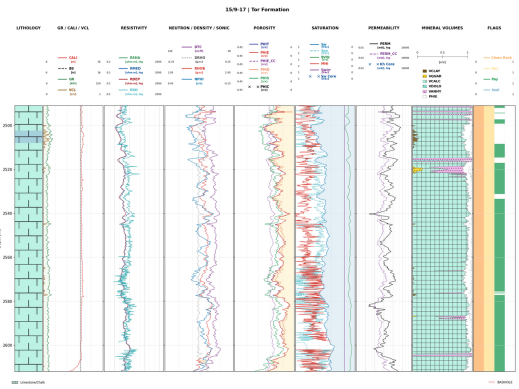
Geolink Dataset (North Sea) Sourced from the Geolink repository (Gama et al., 2024), this subset mostly covers the Norwegian Continental Shelf, including the Volve field. It is the most varied part of the benchmark, with clastics and evaporites often interbedded.

- **Stratigraphic Coverage:** The benchmark questions span the *Cromer Knoll, Dunlin, Hegre, Hordaland, Nordland, Rogaland, Shetland, Vestland,* and *Viking* Groups, including Brent-equivalent formations such as Etive and Ness.
- **Lithology:** The widest range of rock types in the benchmark: *Sandstone, Shale, Limestone, Dolomite, Anhydrite, Halite (Salt), Gypsum, Coal, Conglomerate, Chert,* and intrusive/extrusive igneous rocks.

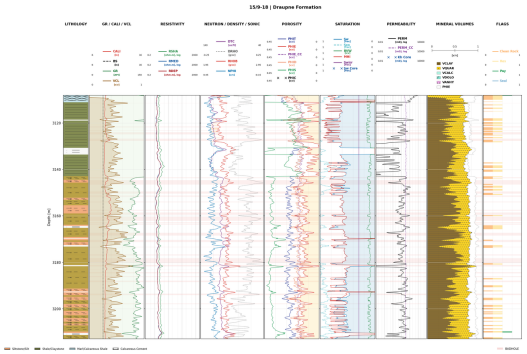
Figure 8 shows four Geolink formations that span this lithological range, from a sandstone reservoir with coal stringers to chalk and organic-rich shale.



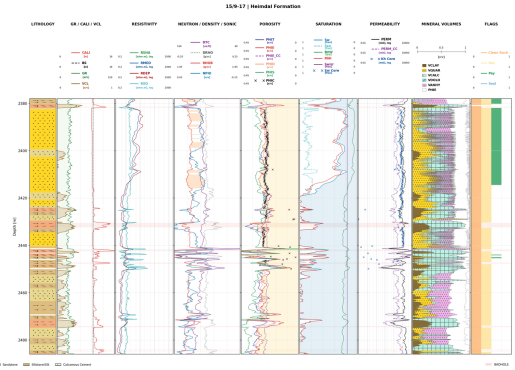
(a) Hugin Fm., 15/9-18 (3237–3410 m): sandstone with coal



(b) Tor Fm., 15/9-17 (2491–2612 m): chalk



(c) Draupne Fm., 15/9-18 (3108–3213 m): shale and siltstone



(d) Heimdal Fm., 15/9-17 (2378–2486 m): sandstone and siltstone

Figure 8: **Representative Geolink well logs.** Each plot shows, from left to right, the lithology column, gamma ray/caliper/clay volume, resistivity, neutron–density–sonic, porosity, saturation, permeability, mineral volumes, and petrophysical flags; shaded bands mark bad-hole intervals. These are the curves that Petrophys summarizes into structured context.

COSTA Model (Carbonate Reservoir) The COSTA subset (Costa Gomes et al., 2022) comes from a carbonate reservoir benchmarking case study. Unlike the North Sea data, it covers a single, highly heterogeneous carbonate setting.

- **Stratigraphic Coverage:** All data come from the *Kharaib-2* formation, a common analog for major Middle Eastern carbonate reservoirs.
- **Lithology:** Mainly *Limestone* and *Dolomite*, with no clastic intervals. These questions test reasoning about porosity and fluid saturation in complex carbonate pore systems, which behave differently from sandstones.

Danish Dataset (Danish North Sea) Taken from the Subsurface Data Denmark portal (Geological Survey of Denmark and Greenland (GEUS), 2026), this subset covers the Danish sector of the North Sea. We kept only formations with stable, consistent logging curves.

- **Stratigraphic Coverage:** The curated subset includes the *Chalk Group, Cromer Knoll Group, Ekofisk, Herring, Hidra, Hod, Kimmeridge Clay, Rotliegend, Skagerrak, Sola, Tor, Tuxen, Vaale,* and *Zechstein* formations.
- **Lithology:** Mostly chalks and marls: *Chalk, Claystone, Marl, Salt, Basalt, Siltstone,* and *Sandstone*.

B.2 Lithological Diversity Summary

Table 3 summarizes the rock types in WellLogBench. Some lithologies recur across basins (e.g., carbonates in Geolink's Shetland Group and in COSTA, and chalks in the Danish data), so a model must apply the same rock physics in different geological settings.

Table 3: **Geological formations and lithologies in WellLogBench.**

Dataset	Primary Formations/Groups	Dominant Lithologies
Geolink	Viking GP, Shetland GP, Dunlin GP, Rogaland GP, Brent GP analogs	Sandstone, Shale, Silt, Limestone, Dolomite, Anhydrite, Halite, Coal, Conglomerate, Igneous (Intrusive/Extrusive)
COSTA	Kharaib-2	Limestone, Dolomite
Danish	Ekofisk, Tor, Hod (Chalks), Kimmeridge Clay, Zechstein	Chalk, Marl, Claystone, Salt, Basalt, Siltstone, Sandstone

C Annotation Workflow and Expert Panel

The expert panel was involved from the start: the question templates in the Category Map (required curves, derived quantities, example questions, and difficulty levels for each category and reasoning level; Section 3.3) were designed with the experts before generation. To construct ground-truth labels for our evaluation set, we designed an expert-facing annotation workflow that guided domain specialists through a consistent review of LLM-generated formation interpretations. Experts were presented with each QA item as a self-contained unit: the original question, the LLM's answer, and the LLM's reasoning (when available), along with formation metadata (e.g., well/formation context) needed to judge whether the response was geologically and petrophysically plausible. Experts also loaded the raw LAS files into industry interpretation software (Petrel and Techlog) and re-checked values, depths, and interpretations directly against the source measurements, not against the extracted context that the auditor sees.

For each QA item, experts completed a structured 17-field rubric covering multiple aspects of answer quality, grouped into four aspects: (i) **clarity and structure** (e.g., whether the response provides an interpretable statement supported by relevant tables and a coherent reasoning flow), (ii) **geological context** (e.g., whether lithological variation is handled appropriately and whether missing logs or data limitations are acknowledged), (iii) **scope and specificity** (e.g., whether interpretations are interval-specific and whether cross-interval or cross-formation comparisons are made when required), and (iv) **correctness and sensibility** (e.g., numerical value correctness, units checking, and whether interpretations are reasonable given the provided logs). When experts judged an answer as incomplete or incorrect, they were asked to provide brief corrective comments and, when necessary, a corrected response in the same format expected from the system.

Expert panel. Annotation was carried out with an established third-party energy consulting firm by six senior petrophysics consultants with over 180 years of collective experience at leading operators and service companies, including SLB, Halliburton, Chevron, Shell, and BP. The firm and the individual experts will be acknowledged in the camera-ready version.

Pilot round and auditor checklist. Before full-scale annotation, several experts reviewed a small sample of generated QA pairs, and the recurring issues they flagged were turned into the checklist used by the GPT-5 auditor (Section 3.3). The auditor checks each drafted pair against this checklist, catching formatting errors, numerical inconsistencies with the formation statistics, and incomplete answers, and either revises the pair or recommends regeneration, for up to two revision cycles. The auditor is a pre-filter that lets experts concentrate on domain judgment such as geological plausibility and interpretation quality; the expert review that follows is an independent, full re-verification of every item, not a check of the auditor's output.

Corrections, rejection, and effort. Experts fixed factual errors, removed hallucinated claims, and corrected numerical values, units, lithology and fluid interpretations, uncertainty statements, and

unsupported conclusions. Answers that could not be repaired were marked *Can't Fix* and rejected. The final reference answer is the expert-approved answer, not the auditor's output. Reviewing one QA pair took an expert 45–60 minutes on average, and annotation, correction, and consensus review together required over 1,000 expert hours.

Consistency and disagreement resolution. Each QA item was assigned to one primary expert reviewer, and all experts used the same rubric. Ambiguous or disputed items were first discussed in a private expert-review Slack channel and then revisited in weekly meetings with all annotators and the authors; an item entered the benchmark only once the group converged on a technically valid reference answer, and recurring feedback from these discussions was fed back into the prompts, category requirements, and filtering rules. The full benchmark was not independently double-annotated item by item, so no dataset-wide inter-annotator statistic is reported; on the 50-item set that two experts scored independently, expert–expert agreement is $\mathrm{ICC}(\mathrm{A}, 1){=}0.765$ (Appendix E).

The annotations served two purposes: they provided ground-truth labels for the benchmark, and they exposed recurring failure modes (e.g., missing-data assumptions, unit inconsistencies, or overconfident fluid calls) that we used to improve the data-generation pipeline.

D QA Pipeline Audit Transparency

Two-pass GPT-5 auditing workflow. Each Gemini-drafted QA pair passes through (i) *Pass 1 validation* against quality and consistency criteria, (ii) auditor *revision* on failure, and (iii) *re-validation*; items failing the second pass are rejected and re-generated. This makes the final benchmark a product of independent review rather than a direct mirror of the drafting model. Table 4 reports the intervention rates by reasoning level.

Table 4: **Share of QA pairs changed by auditing and expert review.** The high union rates show that the final benchmark is substantially shaped by independent review.

Stage	Single	Cross	Multi-Well
GPT-5 auditor revision rate	71.8%	77.4%	76.0%
Expert correction rate	59.2%	61.0%	59.5%
Union (auditor $\cup$ expert)	81.4%	86.4%	85.2%

Why Gemini 2.5 Pro as drafter. We chose Gemini 2.5 Pro as the drafter for its 1M-token context window and good long-context retrieval, which items built on very long source contexts need. To limit generator-style bias, auditing (GPT-5) and expert correction were kept separate from the generator; the high revision and correction rates above show that this separation works.

E Expert–LLM Judge Alignment

Blinded expert evaluation setup. Two senior petrophysicists independently scored 50 QA items. For each item, experts received: (a) the ground truth, (b) Answer 1 (Gemini 2.5 Pro), and (c) Answer 2 (GPT-5), with model identifiers removed and answer order randomized. Experts rated both answers on the full rubric (all sub-metrics that compose Response Adequacy, Trustworthiness, Style, and Petrophysical Reasoning Quality).

Rubric mapping and score construction. The reported RA mean uses Accuracy, Petrophysical Claims by Reference, Conclusion by Reference, and Coverage. Trust combines Faithfulness, the separately hosted Numerical Accuracy score, Data Limitations by Reference, and Log Utilization by Reference. Style averages Clarity, Conciseness, and Repetition; RQ averages the reference-free Petrophysical Claims and Petrophysical Chain-of-Thought scores. Precision, Hallucination, Completeness, Internal Consistency, and the primary judge's proposed *Overall Score* are released as diagnostics but do not enter the four reported group means. The headline Overall is then computed deterministically from the four group means using 40/20/20/20 weights. Rows with a primary-judge error or an incomplete reported rubric aggregate are excluded from model means.

Table 5: **Expert vs. LLM-as-judge rubric scores.** A1 = Gemini 2.5 Pro; A2 = GPT-5. Expert $\Delta_{\text{A1-A2}}=+11.6$; LLM $\Delta_{\text{A1-A2}}=+9.9$ (computed from unrounded scores).

		Expert		LLM judge	
Rubric	**Weight**	**A1**	**A2**	**A1**	**A2**
Response Adequacy	40%	71.2	66.7	71.2	67.0
Trustworthiness	20%	76.2	66.5	74.3	65.9
Style	20%	81.1	46.8	80.0	51.7
Petrophysical Reasoning Quality	20%	81.7	76.8	82.2	77.6
Weighted Overall	—	**76.2**	**64.7**	**75.8**	**65.8**

Absolute agreement via ICC. Because judge validation requires agreement in the absolute score values, not only rank association, we report a two-way absolute-agreement single-measure intraclass correlation (McGraw & Wong, 1996), $\mathrm{ICC}(\mathrm{A},1)$. Across 50 QA pairs, two answers, and 13 shared sub-metrics, expert and LLM-as-judge ratings achieve $\mathrm{ICC}(\mathrm{A},1)=0.721$, indicating good agreement under the guidelines of Cicchetti (1994) (0.60–0.74). Agreement is slightly higher for Answer 1/Gemini 2.5 Pro ($\mathrm{ICC}(\mathrm{A},1)=0.697$) than for Answer 2/GPT-5 ($\mathrm{ICC}(\mathrm{A},1)=0.683$).

Expert–expert agreement. Because the two experts scored the items independently, their agreement is an empirical ceiling for judge agreement. Expert–expert agreement is $\mathrm{ICC}(\mathrm{A},1)=0.765$ across the shared sub-metrics, only 0.044 above expert–judge agreement (0.721). The metrics on which the judge agrees less with experts are also hard for experts: expert–expert $\mathrm{ICC}(\mathrm{A},1)$ is 0.639 for Accuracy, 0.507 for Coverage, and 0.473 for Petrophysical Claims by Reference, compared with 0.762 for Numerical Accuracy and 0.710 for Faithfulness.

Overall preference agreement. Experts and the LLM-as-judge agree on the overall preference direction (Gemini 2.5 Pro > GPT-5) and on its approximate magnitude.

Table 6: **Expert–judge Pearson correlations per metric**, separately for A1 (Gemini 2.5 Pro) and A2 (GPT-5).

Rubric	**Metric**	r_{A1}	r_{A2}	**Rubric**	**Metric**	r_{A1}	r_{A2}
Resp. Adeq.	Accuracy	0.549	0.553	Trust.	Faithfulness	0.763	0.412
	Petro. Claims (Ref.)	0.800	0.355		Numerical Accuracy	0.897	0.752
	Conclusion (Ref.)	0.606	0.551		Data Limitations (Ref.)	0.512	0.419
	Coverage	0.561	0.420		Log Utilization (Ref.)	0.567	0.415
Style	Clarity	0.423	0.311	Petro. RQ	Claims (Ref.-free)	0.656	0.347
	Conciseness	0.471	0.505		CoT (Ref.-free)	0.428	0.431
	Repetition	0.408	0.266				

Per-metric Pearson correlations. Pearson correlations are reported as a secondary association measure. Across all shared rubric scores, expert–LLM correlation is $r=0.723$ for Answer 1/Gemini 2.5 Pro and $r=0.689$ for Answer 2/GPT-5. Table 6 reports per-metric correlations.

Discussion. The ICC results show good absolute agreement between expert and LLM-as-judge ratings, which supports using the automated judge to compare models. Pearson correlations are highest on technical, evidence-based dimensions for Gemini 2.5 Pro, including Numerical Accuracy ($r=0.897$), Petrophysical Claims by Reference ($r=0.800$), and Faithfulness ($r=0.763$). Agreement is weaker on the more subjective style criteria and on some GPT-5 petrophysical reasoning metrics. We therefore rely on the LLM judge for aggregate trends and technical scoring, and report sub-metrics separately so readers can inspect or re-weight dimensions with different levels of expert–judge agreement.

F CROSS-JUDGE ROBUSTNESS

Setup. Our default protocol uses GPT-5 as the primary judge and Qwen3 Next 80B A3B as the numerical judge. We re-scored the shared model cohort, replacing GPT-5 with three different judges: Qwen3 235B A22B Inst., Gemma 4 31B IT, and Qwen3.5-27B. The numerical judge remains the same (Qwen3 Next 80B A3B) in these ablations. Table 7 reports per-rubric Pearson and Spearman agreement with GPT-5; Table 8 reports the resulting model rankings.

Table 7: **Cross-judge agreement with GPT-5** by rubric on the shared model cohort.

Judge	Resp. Adeq.	Trust	Style	Reasoning Q.
Pearson r				
Qwen3 235B A22B Inst.	0.977	0.942	0.848	0.922
Gemma 4 31B IT	0.964	0.954	0.838	0.911
Qwen3.5-27B	0.989	0.979	0.859	0.934
Spearman ρ				
Qwen3 235B A22B Inst.	0.976	0.916	0.661	0.879
Gemma 4 31B IT	0.964	0.892	0.653	0.868
Qwen3.5-27B	0.988	0.914	0.670	0.890

Table 8: **Model ranking by judge** over the frozen 11-model shared judge-validation cohort; rank shifts relative to GPT-5 are in parentheses. Top-3 placements are identical across all judges; model-rank Spearman $\rho \geq 0.96$ across judges; the maximum rank shift is 2 positions.

Model	GPT-5 (reference)	Qwen3 235B A22B Inst.	Gemma 4 31B IT	Qwen3.5-27B
GPT-5.6 Luna (high)	1	1 (0)	1 (0)	1 (0)
Gemini 2.5 Pro	2	2 (0)	2 (0)	2 (0)
Kimi-K2.6	3	3 (0)	3 (0)	3 (0)
GPT-5	4	5 (+1)	4 (0)	4 (0)
Gemma 4 31B IT	5	4 (−1)	6 (+1)	5 (0)
Qwen3.5-27B	6	6 (0)	5 (−1)	6 (0)
Qwen3 235B A22B Inst.	7	7 (0)	7 (0)	7 (0)
GPT-OSS-120B	8	8 (0)	8 (0)	8 (0)
LLaMA-4 Scout	9	9 (0)	10 (+1)	10 (+1)
Qwen3-8B	10	11 (+1)	11 (+1)	9 (−1)
LLaMA-3.3 70B Inst.	11	10 (−1)	9 (−2)	11 (0)

Discussion. Rankings were stable across the tested judges within the frozen 11-model validation set: the top-3 placements never change and model-rank correlation is $\rho \geq 0.96$ (0.964–0.991). The only systematically weaker rubric is Style, consistent with the limitation discussed in Appendix E.

G WEIGHTING SENSITIVITY ANALYSIS

The headline Overall score assigns 40% to Response Adequacy (RA) and 20% each to Trustworthiness, Style, and Petrophysical Reasoning Quality (RQ). Figure 9 recomputes the 17 structured-context configurations (the 15 systems in Table 1 plus the DPO and GRPO stages) with equal 25% weights. GPT-5 is recomputed from its aggregate component scores; the other 16 configurations are recomputed from per-question records. The two aggregations are almost identical at the model level (Pearson r=0.999, Spearman ρ=1.000). Equal weighting mainly raises systems whose Style and RQ are stronger than their RA; it does not change the leaderboard order in any meaningful way.

We retain the default weighting because answering the requested scientific question correctly and completely is the primary benchmark objective. Since all four component scores are released, other weightings are easy to recompute.

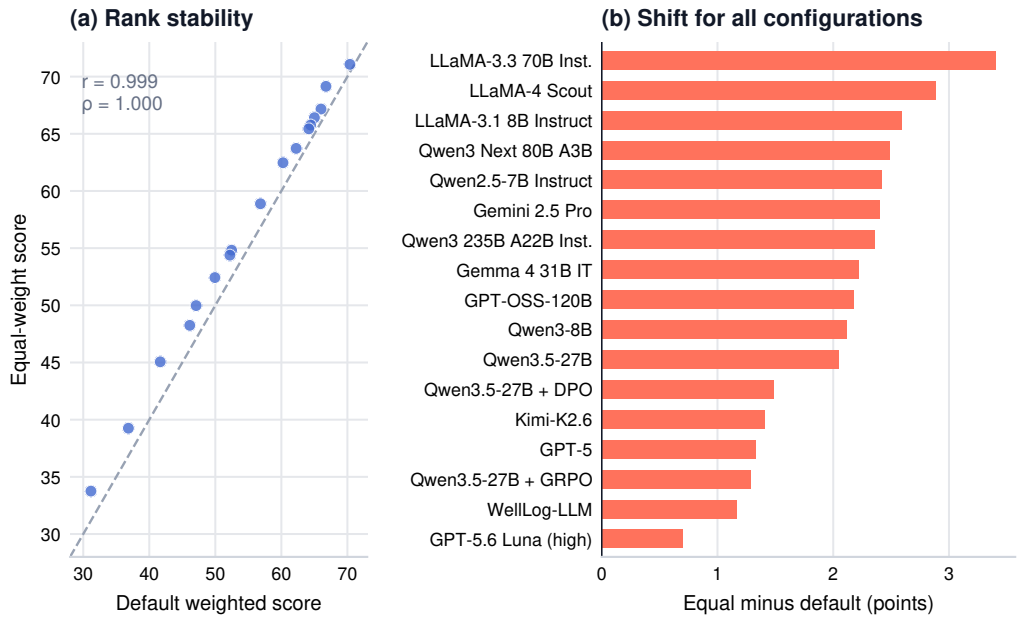


Figure 9: **Sensitivity to rubric weights on the public split.** (a) Default 40/20/20/20 Overall versus equal 25/25/25/25 aggregation for the 17 structured-context configurations (Table 1 plus the DPO and GRPO stages). The dashed line is equality. (b) Equal-weight minus default Overall for all 17 configurations; every configuration shifts upward. For the 16 configurations with per-question records, scores are question-deduplicated means over non-error rows with complete rubric aggregates; GPT-5 contributes its aggregate component scores.

H Evaluation Configurations

Table 9 separates model family, context condition, and post-training stage. "Full" denotes the full structured context and "DFS" the query-conditioned compact context of Section 5.1. Table 1 reports external systems and the untuned Qwen backbone with full context, and WellLog-LLM with DFS; every post-trained stage uses DFS. DFS configurations of the external models appear in the sensitivity study of Appendix I, and the DFS backbone anchors the lineage analyses of Appendix K.2. GPT-5 is scored from public and private component aggregates, so it is included where aggregate scores suffice and excluded from analyses that require per-question records. In total, the text leaderboard scores 27 configurations: 14 models with full context, nine of them again with DFS, and four post-trained Qwen3.5-27B stages (SFT, DPO, GRPO, and GRCO). The VLM study (Appendix N) evaluates six VLMs in two input settings and is scored separately.

Table 9: **Structured-context evaluation roster** (27 configurations). ✓ marks an evaluated context condition. The post-trained Qwen3.5-27B stages use DFS only (Table 2); †SFT is an ablation and is not carried forward. GPT-5 is scored from aggregate component scores.

Model family	Params	Full	DFS	Model family	Params	Full	DFS
GPT-5.6 Luna (high)	—	✓	—	GPT-OSS-120B	117B	✓	✓
GPT-5	—	✓	—	Qwen3 Next 80B A3B	80B	✓	✓
Gemini 2.5 Pro	—	✓	—	LLaMA-4 Scout	109B	✓	✓
Kimi-K2.6	1T	✓	—	Qwen3-8B	8B	✓	✓
Gemma 4 31B IT	31B	✓	—	LLaMA-3.3 70B Inst.	70B	✓	✓
Qwen3.5-27B	27B	✓	✓	Qwen2.5-7B Instruct	7B	✓	✓
Qwen3 235B A22B Inst.	235B	✓	✓	LLaMA-3.1 8B Instruct	8B	✓	✓
Post-trained Qwen3.5-27B (27B, DFS): + SFT†, + DPO, + GRPO, + GRCO (WellLog-LLM)							
Configurations: 14 Full + 9 DFS + 4 post-trained = 27							

I DFS SENSITIVITY FOR EXTERNAL MODELS

Nine model families (the Qwen3.5-27B backbone and eight external models) were also evaluated with DFS context instead of the full structured context. We compare their configuration-level aggregates; valid-response counts can differ between the two configurations, so these are not shared-question paired estimates. Across the nine families, DFS minus full-context Overall averages -0.90 points on the public split and -1.15 points on the private split. Seven families decline on both splits, while LLaMA-3.3 70B Inst. and LLaMA-3.1 8B Instruct show small gains. The direction and relative size of the aggregate difference are therefore stable across the disjoint splits, but remain model dependent. Table 10 gives the displayed scores; Figure 10 visualizes the same aggregate differences.

Table 10: **Full-context versus DFS Overall scores for external models.** Δ is DFS minus Full on the 0–100 scale and is computed from unrounded aggregates. The two configurations can have different valid response counts and are not paired on a shared-question subset.

	Public			Private		
Model	**Full**	**DFS**	**Δ**	**Full**	**DFS**	**Δ**
Qwen3.5-27B	56.8	56.3	−0.6	57.1	56.4	−0.7
Qwen3 235B A22B Inst.	52.5	51.8	−0.6	53.0	51.9	−1.0
GPT-OSS-120B	52.2	51.6	−0.6	52.9	51.1	−1.8
Qwen3 Next 80B A3B	49.9	47.8	−2.1	49.9	47.5	−2.3
LLaMA-4 Scout	47.1	44.9	−2.2	46.6	44.4	−2.2
Qwen3-8B	46.1	44.3	−1.9	45.4	43.4	−2.0
LLaMA-3.3 70B Inst.	41.7	42.0	+0.3	40.3	41.2	+0.9
Qwen2.5-7B Instruct	36.8	35.3	−1.6	36.8	35.2	−1.6
LLaMA-3.1 8B Instruct	31.2	32.2	+1.0	30.9	31.3	+0.4

Combined with the 73.5% reduction in answer-context tokens (Figure 5), DFS therefore trades about one Overall point for a much smaller context even for models that were never trained on DFS-formatted evidence. We report external models with full context in Table 1 because it is their stronger setting.

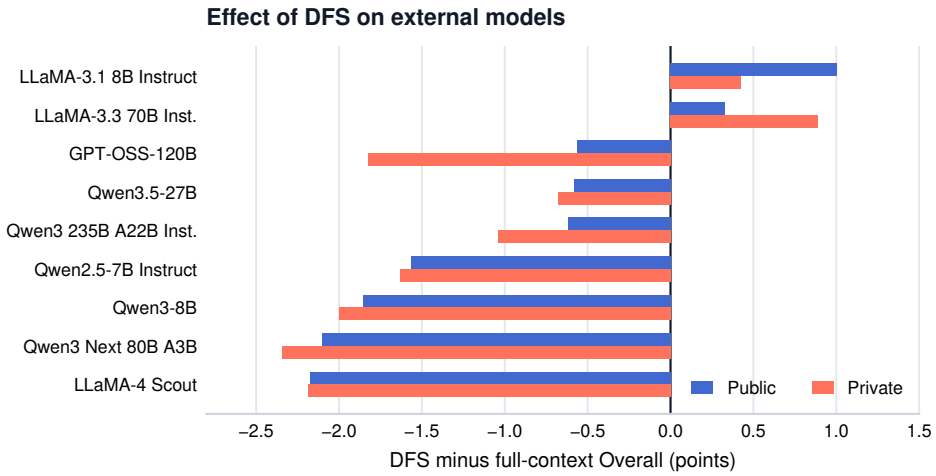


Figure 10: **Effect of DFS on external models.** Bars show each model family's DFS Overall minus its full-context Overall; differences are formed before rounding. Valid response counts can differ by configuration, so the bars are not shared-question paired estimates. Negative values favor full context.

J WellLog-LLM Post-Training Details

This section expands the Qwen3.5-27B lineage in Section 6. DPO and GRPO successively update the base model's weights; GRCO freezes the released GRPO checkpoint and adds a learned natural-language experience library to the prompt at inference. SFT was evaluated as a first stage but is not part of the final lineage. All training data for SFT, DPO, GRPO, and GRCO are built under the DFS setting, and for cross-formation and multi-well questions every training prompt is a Stage-1 prompt covering a single formation or well (Section 5.2). No stage learns from Stage-2 synthesis prompts; Stage 2 runs only at inference, where GRCO adds its rules to the synthesis prompt. Table 11 gives the complete stage order and separates training artifacts from inference-time state.

Table 11: **WellLog-LLM post-training recipe.** Counts refer to the data used by each stage, not the 1,075-item evaluation benchmark.

Stage	Supervision/data	Initialization	Output
SFT	∼30k drafted QA pool; 10k shard	Qwen3.5-27B	Evaluated, not carried forward (Overall 51.1)
DPO	∼15k disjoint chosen–rejected pairs	Qwen3.5-27B (base)	Step-424 checkpoint
GRPO	∼2k QA pairs drafted with Kimi-K2.6	DPO checkpoint	Step-480 policy (one pass)
GRCO	900 prompts with GPT-5.6 Luna (high) references (300 per reasoning type)	Frozen GRPO policy	Experience library in the prompt; no weight update

J.1 SFT

SFT lowers Overall from 56.3 to 51.1 on the public split (Table 2), with Trust and Style falling most, so its checkpoint is not carried forward and DPO starts from the base model. This matches reports that SFT can erode the abilities of reasoning-tuned models while preference and RL objectives preserve them (Chu et al., 2025; Huan et al., 2026). The SFT stage was meant to teach domain vocabulary, answer structure, and the use of DFS-formatted petrophysical evidence. Its pool contains approximately 30,000 DFS-formatted (context, question, answer) triples, balanced as 10,000 Single-Formation, 10,000 Cross-Formation, and 10,000 Multi-Well examples, where each Cross-Formation and Multi-Well example is a Stage-1 prompt covering one formation or well. Kimi-K2.6 drafts the target answers. The reported Qwen run uses a 10k-example shard, assistant-only loss, packing disabled, and keep-end truncation at 65,536 tokens. Optimization uses 10 epochs, cosine decay from a 10^{-4} learning rate with 0.03 warmup, BF16, gradient checkpointing, and LoRA (Hu et al., 2022) with r=16, α=32, and dropout 0.05 on the attention and MLP projection modules. The distributed run uses 16 eight-accelerator nodes with sequence parallelism and ZeRO-3 CPU offload.

J.2 DPO

DPO starts directly from the base Qwen3.5-27B checkpoint. The preference corpus contains approximately 15,000 pairs, 5,000 per reasoning level, drawn from items disjoint from the SFT corpus and from WellLogBench wells. Both answers in a pair respond to the same DFS-formatted input, which for Cross-Formation and Multi-Well items is a Stage-1 prompt covering one formation or well: the chosen answer comes from Kimi-K2.6 and the rejected answer from Qwen3-8B. Each pair therefore contrasts a stronger and a weaker model on identical evidence, the setting in which preference tuning on weak data yields strong gains (Geng et al., 2025). One percent of the pairs is held out for validation.

The objective is the sigmoid DPO loss with β=0.1 plus a small negative log-likelihood term on the chosen answer (RPO, weight α=0.1; Pang et al., 2024). Plain DPO can widen the preference margin while lowering the likelihood of both answers; the added term keeps chosen answers likely, and the chosen-answer log-likelihood rises throughout training. Pairs longer than 65,536 tokens are dropped rather than truncated, so no chosen answer loses its conclusion.

Training uses LoRA (Hu et al., 2022) (r=8, α=32, dropout 0.05) on all linear layers; Adam (β_1=0.9, β_2=0.95, weight decay 0.1, gradient clipping 1.0); learning rate 10^{-5} with 5% warm-up and cosine decay to 10^{-6}; and a global batch of 32 pairs (micro-batch 2). The schedule spans two epochs (886 steps), and the reported checkpoint is step 424, about one pass over the data. The run uses BF16 on 16 GPUs (two 8-GPU nodes) with tensor parallelism 8, sequence parallelism, full activation recomputation, and padding-free FlashAttention batches.

J.3 GRPO CONFIGURATION

GRPO starts from the DPO checkpoint. Its active reward is exactly the minimum composition in Eq. 3: the frozen judge's recommended semantic score J_i, the ground-truth-segment coverage C_i, and the deterministic response-contract guard G_i. The reward uses no weighted sub-rubric sum or additive category bonus. The judge is a frozen Qwen3.5-27B (thinking disabled, temperature 0), separate from the GPT-5 evaluator, and sees the question, authoritative ground-truth final answer, and visible candidate answer; hidden reasoning and the long source context are excluded. The training questions use DFS-formatted Petrophys contexts built exactly as for WellLogBench, from wells outside the benchmark; Kimi-K2.6 drafted the questions and their reference answers. For multi-well and cross-formation questions, each training prompt covers a single well or formation and its reference is a self-contained single-unit analysis, so GRPO rewards the Stage-1 analyses; the Stage-2 synthesis is performed at inference.

Coverage term. The judge estimates the fraction of required ground-truth answer segments that the candidate answer covers. C_i turns this fraction into a reward term with three simple rules. (i) An answer that covers at least 90% of the required segments receives $C_i = 1$, so it is not penalized for small omissions; this tolerance absorbs paraphrasing and the judge's own uncertainty when matching segments. (ii) Below 90%, C_i grows linearly with coverage ($C_i = 0.15 + 0.95\times$ coverage), so every additional covered segment still raises the reward and partially complete answers remain distinguishable within a group; the small offset of 0.15 keeps a weak but non-zero learning signal for answers that cover very little. (iii) Below the threshold, C_i never exceeds 0.98, so an incomplete answer always scores below a complete one with the same judge score. Because the reward takes the minimum of the three terms, C_i acts only as a ceiling: it can lower the reward of an answer that omits required content but never raise it above the judge score. In practice this ceiling was rarely active: it lowered the mean reward by less than 0.0002 over the whole run, so training was driven almost entirely by the judge score and the response guard, and the exact thresholds have little influence on the reported results.

Response guard. $G_i = 1$ unless a deterministic check fails; if several fail, the lowest cap applies. Caps are ordered by severity: answers that are unusable receive little or no reward, while answers that are usable but badly formatted keep part of their score. The caps are 0 for an empty answer; 0.08 for a missing answer boundary; 0.15 for trailing planning, repetition, or chat/garble artifacts; 0.20 for visible planning or solver/meta content; 0.35 for prompt/template echo or placeholder text; 0.40 for truncation or serialized output; and 0.55 for multiple answer boundaries or non-metric depth units. In the reported run, the guard was the lowest term for 6.5% of rollouts, mostly early in training when outputs were often truncated or leaked reasoning (Figure 12c).

Optimization. Each optimization step samples four prompts and four responses per prompt (16 rollouts). Advantages are centered by the four-response group mean without standard-deviation normalization, then optimized with clipped PPO-style updates and length-bias-free `seq-mean-token-sum-norm` aggregation. The run uses a constant 10^{-6} learning rate, an 8,192-token response budget, 16 training nodes, two identical eight-accelerator judge replicas for throughput/failover, no critic, and no KL term. Over 480 steps it processes 1,920 prompt groups and 7,680 sampled rollouts. Figure 11 reports training-batch diagnostics only; no validation-reward curve was logged.

Reward dynamics. Between the first and last 40 steps, mean reward rises from 0.60 to 0.87, and all three reward components rise with it (Figure 11a). The largest gains are on Cross-Formation ($0.53 \rightarrow 0.87$) and Multi-Well ($0.60 \rightarrow 0.89$) prompts, while Single-Formation prompts improve less ($0.71 \rightarrow 0.79$); every petrophysical category improves (Figure 11b–c).



Figure 11: **GRPO training dynamics (480 steps).** (a) Mean reward r and its three components: scientific alignment J, coverage C, and response guard G. (b) Reward by reasoning type. (c) Reward by petrophysical category. Curves are smoothed training-batch means, not validation scores; Figure 12 reports further diagnostics.

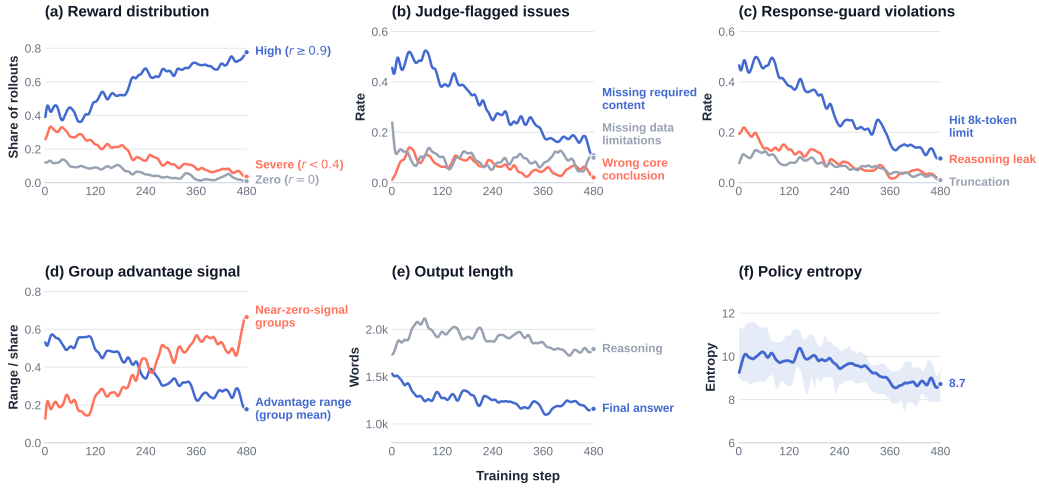


Figure 12: **Additional GRPO training diagnostics (480 steps).** (a) Shares of high-reward ($r \geq 0.9$), severe ($r < 0.4$), and zero-reward rollouts. (b) Judge-flagged issue rates. (c) Response-guard violation rates. (d) Mean within-group reward range and share of near-zero-signal groups (range < 0.1). (e) Words in the reasoning trace and final answer. (f) Policy entropy; the shaded band shows batch dispersion. Curves are smoothed training-batch statistics, not validation scores.

Training diagnostics. Figure 12 complements the reward curves in Figure 11; all values below compare the first and last 40 steps. The share of high-reward rollouts ($r \geq 0.9$) rises from 41% to 75%, while severe failures ($r < 0.4$) fall from 32% to 5% and zero-reward rollouts from 12.5% to 1.2% (a). The judge flags missing required content far less often (48% $\rightarrow$ 14%), with smaller drops in missing data-limitation caveats (12% $\rightarrow$ 8%) and wrong core conclusions (9% $\rightarrow$ 4%) (b). Response-guard violations fall in parallel: responses reaching the 8,192-token budget drop from 48% to 11%, visible reasoning leaks from 20% to 2%, and truncated answers from 12% to 2% (c). As rewards concentrate near the ceiling, the mean within-group reward range falls from 0.55 to 0.22 and the share of near-zero-signal groups (range < 0.1) rises from 20% to 59% (d), so later steps carry less group-relative learning signal. Outputs become shorter rather than longer: final answers fall from about 1,460 to 1,160 words and reasoning traces from about 1,860 to 1,780 words (e). Policy entropy declines from 9.9 to 8.7 (f).

J.4 GRCO Experience Refinement

GRCO adapts training-free GRPO (Cai et al., 2025) to well-log interpretation. The GRPO weights stay frozen. What is learned instead is a library of natural-language *experiences*: short, self-contained

rules that are placed in the policy prompt. The two ingredients of GRPO carry over directly. The contrast between stronger and weaker answers to the same question plays the role of the group-relative advantage, and an edit to the library plays the role of a parameter update. Figure 13 shows the procedure, and Table 12 lists every setting.

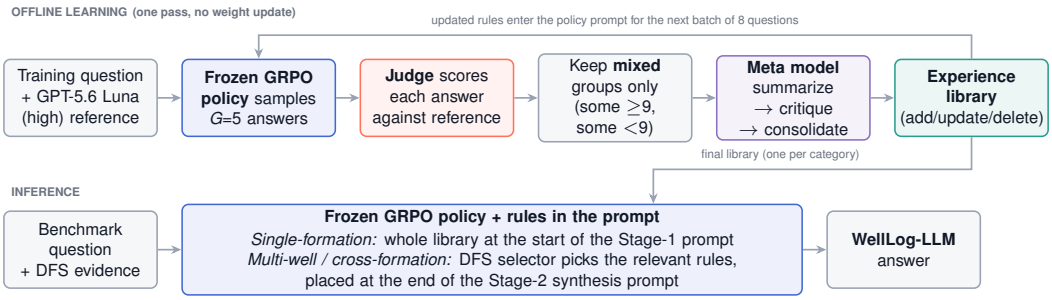


Figure 13: **GRCO learning loop and inference.** *Top:* offline learning; mixed groups are turned into library edits that condition the next batch. *Bottom:* inference, where the final library is added to the prompt of the unchanged policy.

Table 12: **GRCO settings** (one library per question category; no weight updates).

Component	Setting
Policy	Frozen GRPO checkpoint (step 480); temperature 0.7; only the answer after `</think>` is scored
Data	900 DFS-formatted prompts from wells outside WellLogBench (300 per reasoning type; cross-formation and multi-well prompts are Stage-1 prompts covering one formation or well); GPT-5.6 Luna (high) references
Judge, meta	Qwen3 235B A22B Inst. (temperature 0.1 / 0); reward = mean of six 1–10 criteria
Loop	One pass; 8 prompts per batch; G=5 answers per prompt; mixed groups at threshold 9.0
Edits	$\leq$1 per learnable group; batch consolidated into ADD / UPDATE / DELETE / NONE
Library	$\leq$100 rules per category (final single-formation libraries: 21–36 rules); never truncated
Inference	Single-formation: whole library in Stage 1; multi-well, cross-formation: DFS-selected rules in Stage 2

Why the loop is built this way. Only *mixed* groups, with at least one correct and one incorrect answer, are used. If all answers in a group are correct, or all are wrong, nothing separates a good strategy from a poor one, just as a GRPO group with identical rewards has zero advantage; the correctness threshold sits near the typical training reward, so most groups split into stronger and weaker answers. Each learnable group may propose only one edit, which ties every change to a single contrast, and a batch-level consolidation step merges or resolves overlapping edits before the library changes. The learning judge differs from the benchmark evaluator, so the library is not fitted to the evaluator's preferences (the two judges rank systems consistently; Appendix F), and references come from the strongest system on the leaderboard, giving a target above the policy's own level.

Three meta steps. The meta model first *summarizes* each answer (approach, claims, errors, strengths), then *critiques* the group, comparing correct and incorrect attempts with the reference and the current library and proposing one new or revised rule, and finally *consolidates* the batch's proposals into add, update, or delete operations. Rules must be self-contained strategies rather than question-specific calculations. The prompts are in Appendix Q.5.

Using the library at inference. Rules are placed where the final interpretation is formed. Single-formation questions receive the whole library before the interval evidence, so the rule that corrects a wrong conclusion is always present. For multi-well and cross-formation questions, the Stage-1 calls are unchanged; the DFS selector also picks the relevant rules, split into single-principle statements, and appends them to the Stage-2 synthesis prompt, keeping the comparison focused. Table 13 shows representative rules.

Table 13: **Representative experience-library rules** (abridged). Entries are general guidance distilled from rollout contrasts on training wells, not stored benchmark answers.

Experience family	Rule
Depth and units	Preserve stratigraphic order and prefer units attached to interval labels when global metadata conflicts.
Derived-log validity	Before using MHI, SW, FZI, or crossover flags, verify that required source logs and substitutions are valid.
Fluid typing	Combine neutron–density separation with RDEP and SW; treat a missing diagnostic as non-detection, not proof of absence.
Pay-zone reasoning	Prioritize depth-specific pay/movable-HC flags, then verify co-located VSH, PHIT, SW, and PERM evidence instead of relying on interval means.
Evidence discipline	Follow supplied computations when they conflict with generic heuristics, and distinguish data insufficiency from confirmed absence.

K Supplementary Public-Split Results

K.1 Performance Slices for All Models

Figure 7 in the main text shows six representative systems. Table 14 extends it to every system in Table 1 that has per-question records (GPT-5 has only aggregate scores). Values are question-deduplicated public-split Overall means over each model's valid responses.

Table 14: **Public-split Overall by category, basin, and reasoning type for all models.** Best per column in **bold**; WellLog-LLM uses DFS, all other systems full context.

	Category						Basin			Reasoning type		
Model	Por.	Perm.	Sat.	Lith.	HC	Num.	COSTA	Danish	Geolink	Single	Cross	Multi
GPT-5.6 Luna (high)	**70.9**	**71.1**	**69.8**	**72.2**	**69.7**	**68.1**	**69.7**	**71.4**	**70.4**	**69.2**	**71.0**	**70.8**
Gemini 2.5 Pro	65.9	69.6	66.8	66.1	65.3	66.4	68.3	69.5	65.6	68.7	66.9	65.1
Kimi-K2.6	63.9	66.1	66.0	62.2	67.0	64.7	66.0	67.7	64.1	65.4	64.8	64.9
Gemma 4 31B IT	60.6	62.6	60.6	58.7	59.0	59.6	62.0	62.0	59.2	62.2	60.2	58.7
Qwen3.5-27B	56.9	59.5	57.3	55.2	55.3	56.4	57.4	60.4	55.9	57.2	56.8	56.6
Qwen3 235B A22B Inst.	52.1	54.1	53.4	50.1	52.3	52.6	54.9	55.2	50.9	55.2	47.1	54.1
GPT-OSS-120B	52.5	53.3	51.6	50.7	51.3	54.0	54.1	56.0	50.7	52.6	51.1	52.7
Qwen3 Next 80B A3B	49.9	52.7	50.2	46.2	50.0	50.5	53.1	50.3	48.6	51.3	45.2	52.2
LLaMA-4 Scout	45.8	48.3	47.2	44.3	47.4	50.0	47.2	50.2	46.4	46.4	47.4	47.4
Qwen3-8B	45.4	46.3	45.6	46.0	46.6	47.0	46.2	48.7	45.6	47.2	44.5	46.5
LLaMA-3.3 70B Inst.	40.4	44.1	40.5	41.4	39.6	44.1	44.6	43.8	40.0	40.6	41.7	42.5
Qwen2.5-7B Instruct	34.3	40.7	36.6	36.7	35.4	36.8	36.7	40.9	36.1	37.3	35.3	37.5
LLaMA-3.1 8B Instruct	31.5	32.2	30.4	31.7	30.1	30.8	33.6	32.9	29.8	30.9	29.2	32.8
WellLog-LLM	65.4	66.6	66.0	64.7	68.1	65.0	65.5	68.8	65.7	68.7	65.5	64.2

K.2 Qwen-Lineage Stage Patterns

Figure 14 expands the Qwen lineage across all six categories, three basins, and three reasoning types. The table below provides the four aggregate rubric groups and the corresponding private Overall for the same checkpoints.

The displayed sequence is 56.3 $\rightarrow$ 62.2 $\rightarrow$ 64.2 $\rightarrow$ 66.0 on the public split and 56.4 $\rightarrow$ 62.9 $\rightarrow$ 64.2 $\rightarrow$ 66.4 on the private split. All four checkpoints use DFS, so the sequence reflects post-training rather than a change of context. The base-to-DPO transition gives the largest change in every rubric group; after DPO, RA and RQ gain most at GRPO, while GRCO's largest change is Trust.

Paired item-level transitions. The stage table reports aggregates over each checkpoint's own valid responses. Table 16 instead pairs adjacent checkpoints on the public items both answered, starting from the DFS backbone as in Table 2. All three transitions remain positive on shared items, with

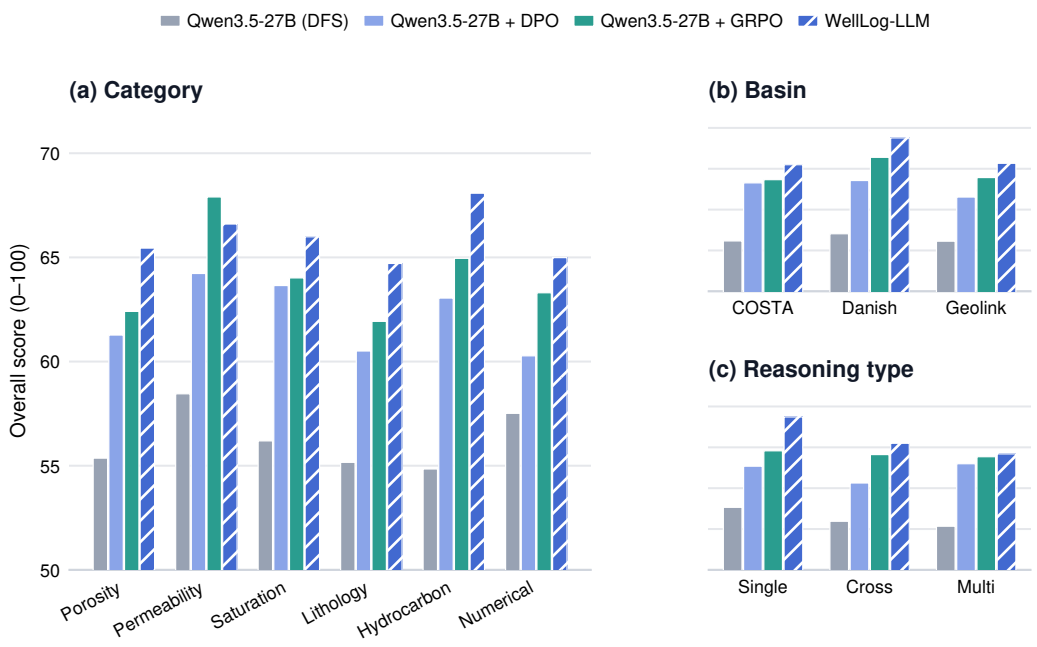


Figure 14: **Qwen-lineage public-split slices.** Bars show question-deduplicated Overall by category, basin, and reasoning type. All four checkpoints use DFS. Hatched bars mark WellLog-LLM. All panels use the same 50–72 displayed range on the 0–100 metric.

Table 15: **Qwen-lineage aggregates.** Public component scores and Overall are shown with private Overall. All rows use DFS; shading marks post-trained checkpoints.

Checkpoint	RA	Trust	Style	RQ	Pub.	Priv.
Qwen3.5-27B (DFS)	47.2	59.4	59.8	67.6	56.3	56.4
Qwen3.5-27B + DPO	56.3	62.6	64.1	72.0	62.2	62.9
Qwen3.5-27B + GRPO	59.0	63.4	64.7	74.6	64.2	64.2
WellLog-LLM	61.3	66.2	64.8	76.3	66.0	66.4

95% bootstrap intervals (1,000 resamples) above zero. The base-to-DPO step is the largest and most one-sided: more than half of the items improve by at least five points and about one in ten declines. The later steps are smaller and more mixed, with roughly one item in five declining by five points or more, so their aggregate gains are net effects rather than uniform improvements. The gains also land on different reasoning types: GRPO helps Cross-Formation items most (+3.8), whereas GRCO helps Single-Formation items most (+4.0) and leaves Multi-Well nearly unchanged (+0.4).

Table 16: **Paired stage transitions on shared public items.** Δ is the mean per-item change in Overall (0–100) with a 95% bootstrap interval; improved/flat/worse use a $\pm$5-point band. The last three columns give the mean change by reasoning type.

			Items (%)			Δ by reasoning type		
Transition	n	Δ Overall [95% CI]	Improved	Flat	Worse	Single	Cross	Multi
Qwen3.5-27B (DFS) $\rightarrow$ DPO	615	+6.0 [+5.1, +6.8]	53.3	36.1	10.6	+5.0	+4.7	+7.6
DPO $\rightarrow$ GRPO	583	+1.9 [+1.1, +2.8]	33.4	44.6	22.0	+1.7	+3.8	+0.8
GRPO $\rightarrow$ GRCO	613	+1.8 [+1.2, +2.5]	31.3	50.2	18.4	+4.0	+1.4	+0.4

Stage signatures by judge dimension. Table 17 decomposes the lineage into the judge's sub-metrics (0–10, higher is better). Each stage has a distinct signature. DPO lifts almost every dimension, led by Coverage (+1.02), reference-grounded conclusions and petrophysical claims, and Accuracy, while shortening answers by about a quarter; it barely changes data-limitation handling. GRPO concentrates on grounding: reference-grounded claims, Precision, Faithfulness, and Accuracy each rise by about 0.3 while style dimensions barely move. GRCO's largest change by far is data-limitation handling (+0.76, more than twice its next-largest gain), consistent with experience rules that separate missing evidence from negative findings.

Table 17: **Judge sub-metric means across the Qwen lineage** (public split, 0–10). Columns give the DFS backbone, the change at each stage, and WellLog-LLM; the largest gain per stage is in **bold**. †The judge's Hallucination score, where higher means fewer unsupported specifics.

Sub-metric	Qwen3.5-27B (DFS)	Δ DPO	Δ GRPO	Δ GRCO	WellLog-LLM
Accuracy	4.29	+0.83	+0.31	+0.18	5.61
Coverage	5.32	**+1.02**	+0.13	+0.20	6.67
Conclusion (ref.)	5.33	+0.90	+0.29	+0.29	6.81
Petro. claims (ref.)	3.94	+0.89	**+0.36**	+0.26	5.46
Precision	3.79	+0.37	+0.35	+0.06	4.57
Faithfulness	3.20	+0.45	+0.34	+0.15	4.14
Unsupported detail†	2.64	+0.40	+0.25	+0.14	3.44
Data limitations (ref.)	5.07	+0.16	−0.03	**+0.76**	5.97
Log utilization (ref.)	5.86	+0.49	+0.01	+0.18	6.54
Internal consistency	7.08	+0.31	+0.29	+0.14	7.83
Clarity	7.29	+0.39	−0.03	+0.13	7.77
Conciseness	2.97	+0.40	+0.09	−0.07	3.39
Numerical accuracy	9.66	+0.14	+0.03	+0.03	9.86
Median answer length (words)	1,685	−466	+150	+60	1,428

One question traced through the lineage. Table 18 follows the public item with the largest monotonic gain across all four checkpoints (83 of the 582 items answered by all four improve at every stage) and the item with the largest GRCO regression. In the traced item, the backbone and DPO misread matrix effects as primary porosity, GRPO recovers the reference conclusion but adds an unsupported secondary-porosity claim, and GRCO matches the reference. The regression shows the cost of the same rule: GRCO declares available flow-unit statistics missing. Figure 15 shows the logs behind the traced item.

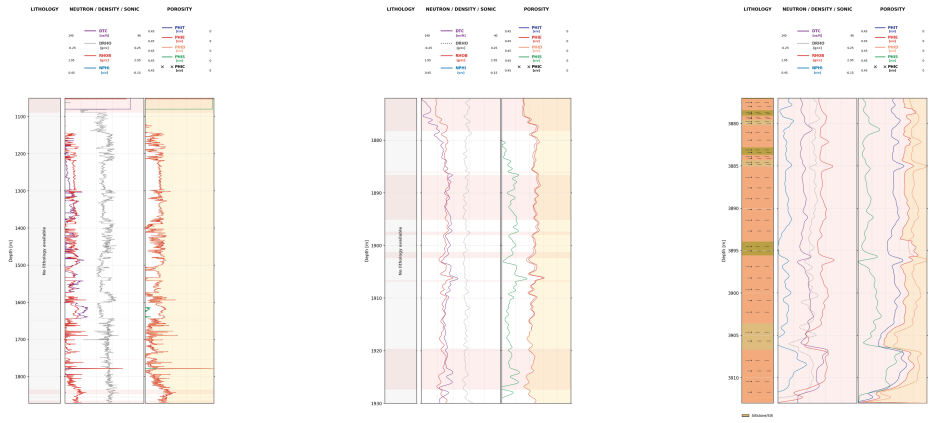
K.3 Qualitative Response and Judge Analysis

The public-split qualitative analysis uses candidate answers, ground truths, 17 primary-judge scores, 16 corresponding score explanations, numerical-judge reasoning, and the four reported rubric aggregates after an exact question-level join to the 646-question public split. Private-split items are excluded throughout. Figure 16 summarizes three transparent diagnostics computed directly from those fields. A "failure" is a judge subscore below 5 on its native 0–10 scale; this is an analysis threshold, not a new benchmark metric. The evidence-channel panel searches the free-text Coverage explanations for exact log mnemonics among rows with Coverage below 5 and reports the six most frequent base-model mentions from a predefined set of 11 common log channels. It describes what the judge mentions, not whether the underlying log is objectively causal. All four lineage checkpoints use DFS.

The severe-score tail contracts from 28.5% for the untuned Qwen3.5-27B (DFS) backbone to 13.8% after DPO, 10.6% after GRPO, and 7.1% with GRCO. Conversely, the share at or above 70 Overall rises from 9.0% to 25.5%, 30.0%, and 37.6%. Accuracy, Coverage, Faithfulness, and Completeness failure rates decline along the lineage. Data-caveat scores are less monotonic through DPO/GRPO but improve sharply after GRCO, consistent with the experience rules that distinguish missing evidence from negative findings.

Table 18: **Stage-by-stage qualitative trace.** Overall (0–100) per checkpoint; notes paraphrase the released GPT-5 judge explanations. Items are chosen by fixed rules and illustrate behavior rather than frequencies.

Stage	Overall	Judge-noted behavior
Largest monotonic gain: pore-system evidence (primary, secondary, or matrix effects) across five formations in well 34/10-35		
Qwen3.5-27B (DFS)	38.8	Classifies Rogaland and Sele as primary-porosity units, whereas the reference shows matrix effects (negative density–neutron separation) throughout; adds unsupported permeability claims.
+ DPO	42.0	Repeats the primary-porosity reading for Rogaland and Sele and recasts Hordaland as primary-dominated rather than matrix-dominated with local exceptions.
+ GRPO	65.8	Recovers the reference conclusion (matrix effects dominate, with localized primary/secondary evidence only in Hordaland) but adds a secondary-porosity trace in Heather.
+ GRCO	80.7	Matches the reference: matrix effects dominate every formation, Hordaland alone shows non-matrix intervals, and Heather stays matrix-dominated; minor unreferenced specifics remain.
Largest GRCO regression: permeability statistics per hydraulic flow unit, Kharaib-2 (HW-3)		
+ GRPO	83.2	Reports every per-unit permeability statistic and depth interval correctly, with minor unreferenced pore-class labels.
+ GRCO	48.2	Gives the same numbers but attributes them to lithology groups and wrongly states that flow-unit statistics are unavailable, over-applying the missing-evidence rule.



(a) Hordaland Gp. (1050–1872 m) (b) Rogaland Gp. (1872–1930 m) (c) Heather Fm. (3877–3913 m)

Figure 15: **Logs for the traced question** (Table 18): three of the five formations in well 34/10-35, cropped to the lithology, neutron–density–sonic, and porosity tracks. The question asks whether the log evidence indicates primary, secondary, or matrix-controlled porosity; the reference answer finds matrix effects throughout, with localized non-matrix intervals only in Hordaland. Rogaland lacks a neutron log in this well.

K.4 Failure Modes Across Models

Table 20 extends the lineage diagnostics to every system in Table 1 with row-level records. A failure is a judge sub-score below 5 on its native 0–10 scale: *wrong conclusion* (Conclusion by Reference), *omission* (Coverage), *unsupported detail* (Hallucination), *no caveats* (Data Limitations by Reference), *numeric error* (the separate numerical judge), *log misuse* (Log Utilization by Reference), *inconsistent* (Internal Consistency), and *weak chain* (reference-free Petrophysics Chain of Thought); *severe* is Overall below 50.

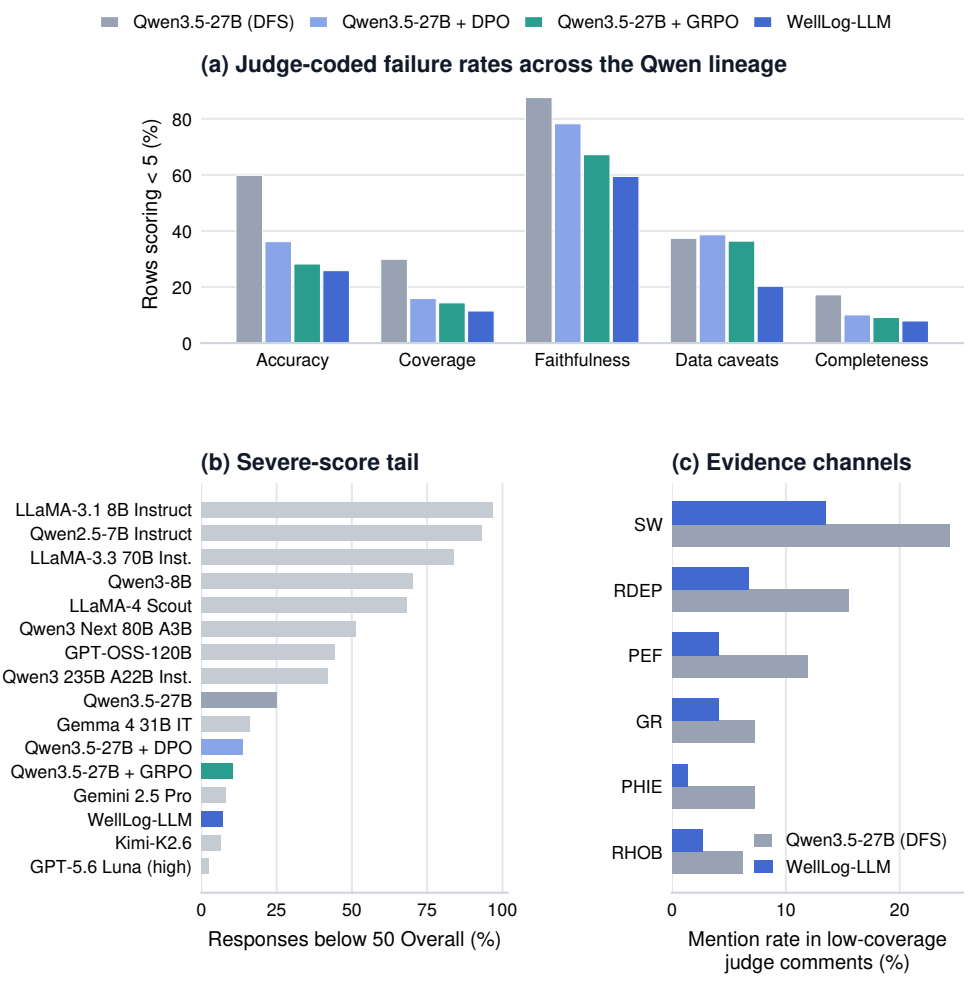


Figure 16: **Qualitative diagnostics from judge outputs.** (a) Fraction of Qwen-lineage responses below 5/10 on five reference-based judge fields. (b) Fraction of responses below 50 Overall for all 16 systems with row-level records (GPT-5 has none), sorted by rate; Qwen-lineage bars use the stage colours of panel (a). (c) Exact log-mnemonic mentions in low-Coverage judge explanations for Qwen3.5-27B (DFS) and WellLog-LLM. These are descriptive analyses of judge outputs, not independently annotated error labels. Panels (a) and (c) use the DFS backbone; panel (b) uses the configurations of Table 1 and the two intermediate stages.

Every failure mode rises with leaderboard position: across the 13 full-context systems, its rate has Spearman $\rho \geq 0.86$ with rank, except unsupported detail ($\rho = 0.77$). Unsupported detail is common for every system (64–100%) because the reference-based judge penalizes any specific that is absent from the expert answer, so it is informative only comparatively. Relative to its Qwen3.5-27B backbone under the same DFS context, WellLog-LLM more than halves wrong conclusions (43.4% $\rightarrow$ 18.9%) and omissions (29.9% $\rightarrow$ 11.5%) and reduces missing caveats (37.4% $\rightarrow$ 20.3%) and log misuse (20.5% $\rightarrow$ 9.1%), placing its profile close to Kimi-K2.6 and Gemini 2.5 Pro. The remaining gap to GPT-5.6 Luna (high) is concentrated in wrong conclusions (18.9% vs. 10.4%) and missing data-limitation caveats (20.3% vs. 7.7%). Numeric errors from the separate numerical judge are rare for the six strongest systems ($\leq 0.8\%$).

Where WellLog-LLM still fails. Its 46 severe failures (7.1% of 646 public items) are spread across reasoning types (Single-Formation 8.3%, Cross-Formation 8.2%, Multi-Well 5.4%), with Numerical the most frequent category (11 of 46). They are interpretive rather than arithmetic: 95.7% carry a

Table 19: **Public qualitative evidence at three reasoning scales.** Each row is the highest-Overall valid WellLog-LLM response within one public reasoning stratum. Scores are Overall on the 0–100 scale for the untuned Qwen3.5-27B base with DFS and WellLog-LLM. Summaries preserve both the corrected behavior and residual limitations identified by the released GPT-5 judge explanations. The examples illustrate behavior; the score-based selection is not an unbiased frequency estimate or stage-level efficacy test.

Case	Base → final	Observed change
Single: missing resistivity logs	71.3 → 92.5	The final answer correctly abstains because RSHA and RMED are absent, while noting that no anomalous zone can be confirmed or excluded. The judge flags only minor extra framing.
Cross: invasion and permeability contrast	66.7 → 84.8	The final answer contrasts resistive invasion, SW–SXO behavior, and permeability between tight HOD chalk and permeable ROTLIEGEND sandstone, and flags the missing RXO/RMF data. The judge notes a small HOD net-pay error and unreferenced NET_RES/DTS details.
Multi: TOR pay comparison	67.7 → 88.9	The final answer identifies oil-prone pay over about 29% of elna-1 and none in elin-1, with minor gas crossovers, the elin-1 MHI, and no fluid contacts. The judge flags added permeability, invasion-ratio, and NET_RES details absent from the reference.

Table 20: **Judge-coded failure rates (% of public-split responses) for all models.** Darker cells indicate higher rates. WellLog-LLM uses DFS; all other systems use full context.

Model	Severe	Wrong concl.	Omission	Unsupp. detail	No caveats	Numeric error	Log misuse	Incon-sistent	Weak chain
GPT-5.6 Luna (high)	2.6	10.4	6.2	78.5	7.7	0.2	5.0	0.2	0.2
Gemini 2.5 Pro	8.0	26.8	16.6	64.2	20.0	0.6	8.4	2.2	0.5
Kimi-K2.6	6.5	21.1	10.4	89.0	21.7	0.0	9.1	4.2	0.3
Gemma 4 31B IT	16.1	31.3	26.2	87.8	25.1	0.0	14.7	6.8	2.2
Qwen3.5-27B	25.0	38.1	26.5	97.1	32.2	0.8	15.7	8.5	1.1
Qwen3 235B A22B Inst.	41.9	48.3	39.9	98.4	47.5	2.4	31.6	23.6	7.4
GPT-OSS-120B	44.4	50.9	33.2	97.8	53.1	0.6	31.2	30.1	6.1
Qwen3 Next 80B A3B	51.2	55.9	48.5	99.1	57.4	1.9	40.9	37.6	9.4
LLaMA-4 Scout	68.3	74.5	86.2	92.3	80.5	3.3	62.1	26.5	43.7
Qwen3-8B	70.1	61.9	62.2	99.4	70.4	0.8	58.0	60.8	25.1
LLaMA-3.3 70B Inst.	83.9	89.5	94.7	94.6	74.9	12.5	69.2	24.6	60.7
Qwen2.5-7B Instruct	93.2	87.3	92.3	99.5	91.6	6.0	76.2	76.6	73.8
LLaMA-3.1 8B Instruct	96.9	95.0	98.3	98.9	93.7	18.7	88.7	59.8	92.3
WellLog-LLM	7.1	18.9	11.5	81.6	20.3	0.3	9.1	1.7	0.3

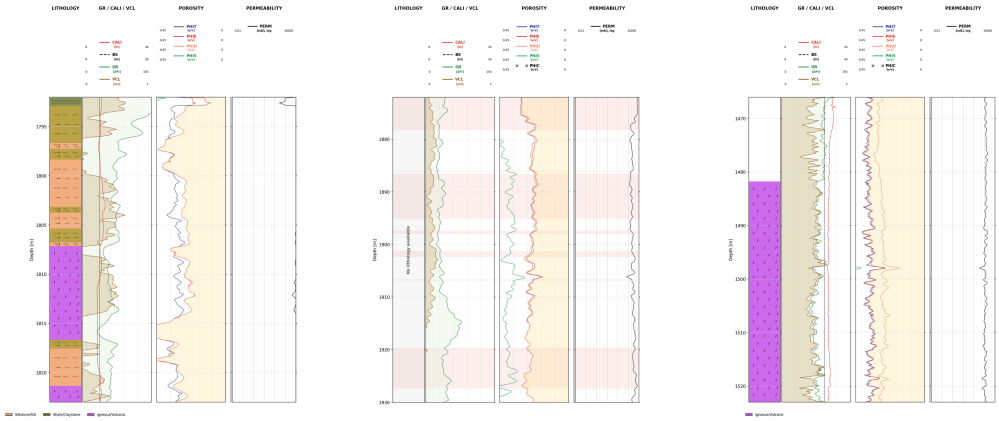
wrong-conclusion score and 87.0% an omission score, whereas only 2.2% involve a numeric error and 8.7% an internal inconsistency.

K.5 Cross-Model Qualitative Contrasts

Table 21 compares WellLog-LLM with the leaderboard leader on three public items chosen by fixed rules: WellLog-LLM's largest advantage over GPT-5.6 Luna (high), its largest deficit, and the item with the lowest mean Overall across the 14 Table 1 systems with per-question records (all but GPT-5). Summaries paraphrase the released GPT-5 judge explanations. Figures 17 and 18 show the logs for the advantage case and the hardest item.

Table 21: **Cross-model qualitative contrasts on the public split.** Scores are Overall (0–100) for WellLog-LLM / GPT-5.6 Luna (high) / mean of the 14 Table 1 systems with per-question records. The rule-based selection illustrates behavior and is not a frequency estimate.

Case	WL / Luna / mean	Judge-noted difference
Largest WellLog-LLM advantage over GPT-5.6 Luna (high). Multi-well: best porosity-based reservoir quality among three wells (Rogaland Gp.)	79.5 / 52.5 / 52.0	WellLog-LLM selects 25/11-19 S, as the reference does, because its low clay volume and high permeability outweigh the higher raw porosity of 35/11-11. GPT-5.6 Luna (high) selects 35/11-11 and omits the effective-porosity argument. The judge flags unreferenced specifics in both answers.
Largest WellLog-LLM deficit. Multi-well: lithology and mineralogy contrast between 15/9-18 and 16/2-11 A (Rogaland Gp.)	48.0 / 79.7 / 52.0	GPT-5.6 Luna (high) recovers the mixed silt–extrusive versus uniformly extrusive contrast and the RHOB, PEF, and GR differences. WellLog-LLM wrongly states that PEF is absent in both wells, misses the extrusive packages of 16/2-11 A, and asserts net pay that the reference does not support.
Hardest item (lowest mean over the 14 systems with per-question records). Single-formation: mobile-hydrocarbon zones in the Statfjord Fm. (30/6-5)	39.5 / 49.7 / 31.4	The reference flags silt intervals (primary 3315.1–3318.2 m) through MHI < 1 with conductive invasion. GPT-5.6 Luna (high) concludes that no zone shows mobility; WellLog-LLM highlights a different sandstone interval and misreports the MHI. Both miss the silt intervals.



(a) 25/11-19 S (1792–1823 m) (b) 34/10-35 (1872–1930 m) (c) 35/11-11 (1466–1523 m)

Figure 17: **Logs for the advantage case** (Table 21): Rogaland Gp. in three wells, cropped to the lithology, gamma ray/caliper/clay volume, porosity, and permeability tracks. The reference answer ranks 25/11-19 S highest because its low clay volume and high permeability outweigh the higher raw porosity of 35/11-11; WellLog-LLM reaches the same ranking, whereas GPT-5.6 Luna (high) selects 35/11-11.

L Private Test Set Results

The private split contains 429 questions disjoint from the 646-question public split. Table 22 applies the same scoring, deduplication, and context convention as Table 1. External systems and the untuned Qwen backbone receive full context; the three WellLog-LLM stage rows use DFS and remain grouped at the bottom. GPT-5 is scored from aggregate component scores and is therefore not included in per-question analyses.

Across the 17 configurations in Table 22, the public-to-private mean absolute Overall change is 0.48 points, the maximum absolute change is 1.31 points, and rank order is nearly identical (ρ=0.998). The only swap is at the top: WellLog-LLM scores 66.4 privately, compared with 66.0 publicly, and

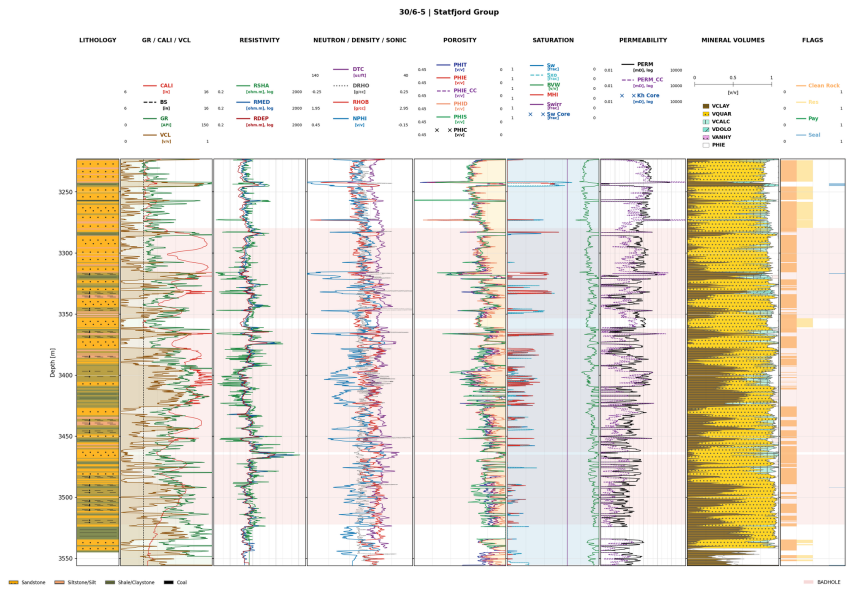


Figure 18: **Logs for the hardest item** (Table 21): Statfjord Fm. in well 30/6-5 (3223–3556 m). The question asks which zones indicate mobile hydrocarbons from permeability and resistivity-based invasion profiles; the reference flags thin silt intervals (primary 3315.1–3318.2 m) that both WellLog-LLM and GPT-5.6 Luna (high) miss.

Table 22: **Private-split structured-context performance** (0–100; higher is better). Overall $= 0.4$ RA $+0.2$ Trust $+0.2$ Style $+0.2$ RQ. Best values are bold; shaded stage rows use DFS. GPT-5 values are computed from aggregate component scores.

Model	Params	RA	Trust	Style	RQ	Overall
GPT-5.6 Luna (high)	—	**66.2**	**69.7**	63.4	**81.1**	**69.3**
Gemini 2.5 Pro	—	56.1	68.1	**75.3**	74.9	66.1
Kimi-K2.6	1T	59.7	65.2	65.5	76.0	65.2
GPT-5	—	58.9	64.9	60.3	79.3	64.5
Gemma 4 31B IT	31B	50.3	62.3	65.9	68.7	59.5
Qwen3.5-27B	27B	48.8	60.6	60.0	67.2	57.1
Qwen3 235B A22B Inst.	235B	43.5	55.9	57.9	63.9	53.0
GPT-OSS-120B	117B	44.5	55.9	57.5	62.0	52.9
Qwen3 Next 80B A3B	80B	39.8	53.0	58.2	58.6	49.9
LLaMA-4 Scout	109B	34.9	49.0	63.5	50.7	46.6
Qwen3-8B	8B	36.6	50.5	53.0	50.4	45.4
LLaMA-3.3 70B Inst.	70B	26.4	44.3	58.7	46.0	40.3
Qwen2.5-7B Instruct	7B	27.2	40.6	48.5	40.5	36.8
LLaMA-3.1 8B Instruct	8B	20.4	35.4	43.8	34.5	30.9
Qwen3.5-27B + DPO	27B	57.1	62.9	64.4	72.8	62.9
Qwen3.5-27B + GRPO	27B	59.2	63.7	64.7	74.3	64.2
WellLog-LLM	27B	62.0	66.7	64.5	77.1	66.4

edges past Gemini 2.5 Pro (66.1 privately); Kimi-K2.6 scores 65.2 privately, and the DPO and GRPO stages score 62.9 and 64.2. This split-level stability shows that the public-split ranking carries over to held-out questions.

Paired significance tests. Table 23 tests the two closest comparisons with a paired bootstrap over questions: for each question both systems answered, we take the difference in Overall and resample questions with replacement (10,000 resamples) to obtain a 95% interval for the mean difference. WellLog-LLM's lead over Kimi-K2.6 is significant on both splits, since both intervals exclude zero.

The difference from Gemini 2.5 Pro is not significant on either split: the two are statistically tied, and WellLog-LLM's 0.3-point private-split edge over Gemini 2.5 Pro is within noise.

Table 23: **Paired bootstrap comparisons** (Overall, 0–100). Δ is WellLog-LLM minus the comparison system, averaged over shared questions; the 95% interval is from 10,000 question-level resamples. Better / worse give the share of questions on which WellLog-LLM scores higher / lower.

WellLog-LLM vs.	Split	n	Δ	95% interval	Better / worse (%)
Kimi-K2.6	Public	646	+1.00	[+0.27, +1.73]	54.2 / 45.5
	Private	429	+1.22	[+0.30, +2.16]	55.0 / 44.5
Gemini 2.5 Pro	Public	646	−0.75	[−1.62, +0.11]	44.4 / 55.3
	Private	429	+0.34	[−0.72, +1.41]	50.6 / 49.0

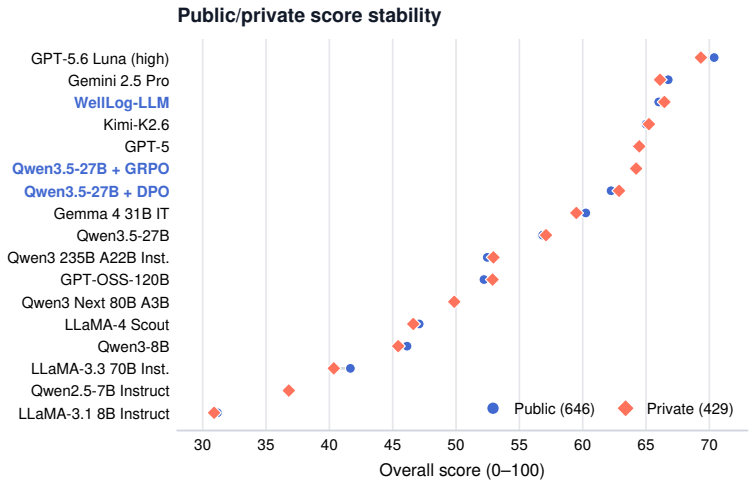


Figure 19: **Public/private stability for the reported structured-context cohort.** Each line connects the public and private Overall of the same checkpoint and context configuration. The public/private partition is disjoint; circles and diamonds are aggregate scores, not paired confidence intervals. GPT-5 uses public/private component aggregates; the other 16 configurations are computed from per-question records.

M Modality Definitions

LAS (Log ASCII Standard) files (Figure 20-a) are the industry-standard format for digital well log storage, consisting of structured header sections followed by an ASCII (American Standard Code for Information Interchange) Log Data section containing space-delimited numerical measurements that describe depth-indexed petrophysical properties across thousands of meters, recorded every 0.1524 m (6 in), resulting in tens of thousands of rows with 10+ curves per well (millions of individual values, exceeding 2 million tokens). This scale far surpasses current LLM context windows and lacks inherent structure for interval-based reasoning, lithology boundaries, or derived interpretations.

Well-log displays such as Computer Processed Interpretation (CPI) panels (Figure 20-b) show many vertical tracks side by side over thousands of meters, one per property or derived parameter (water saturation, clay volume, pay flags, etc.), plus color-coded lithology bars. Interpreting them means reading curve values at specific depths against fine linear or logarithmic scales, spotting diagnostic patterns and trends, correlating features across curves at the same depth, and combining the evidence to pick formation boundaries and hydrocarbon zones.

Figure 21 gives a page-scale overview of the standardized multi-track layout used in the auxiliary evaluation. Each released full-resolution plot spans one formation and preserves raw measurements, derived petrophysical properties, interpretation flags, lithology bars, and depth alignment; the

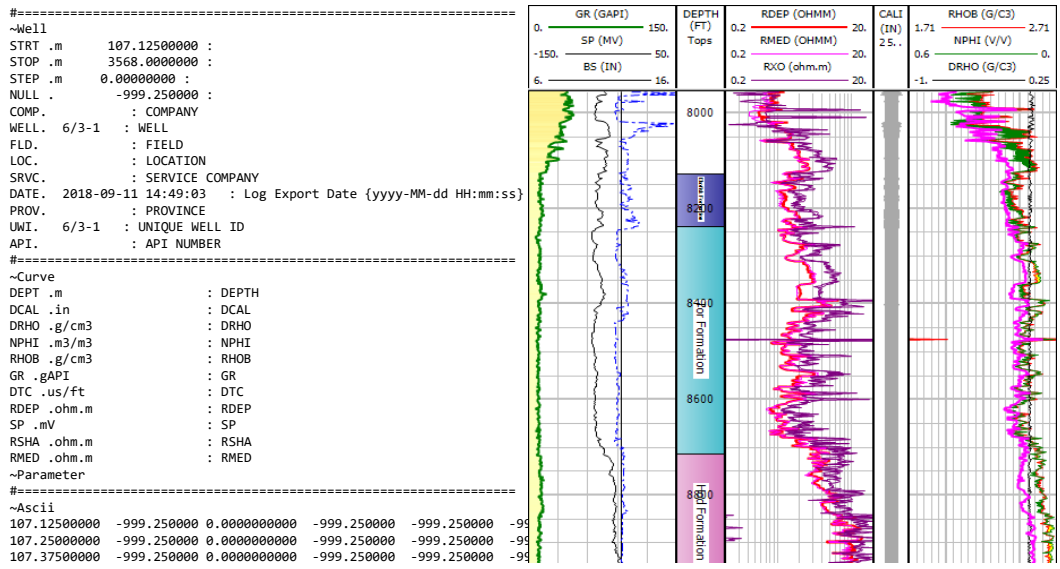


Figure 20: **Well-log data formats.** (a) LAS file structure showing header sections and depth-indexed ASCII data section; (b) CPI display for a well showing a multi-track layout across the 8000–8950 ft depth interval with three formation boundaries.

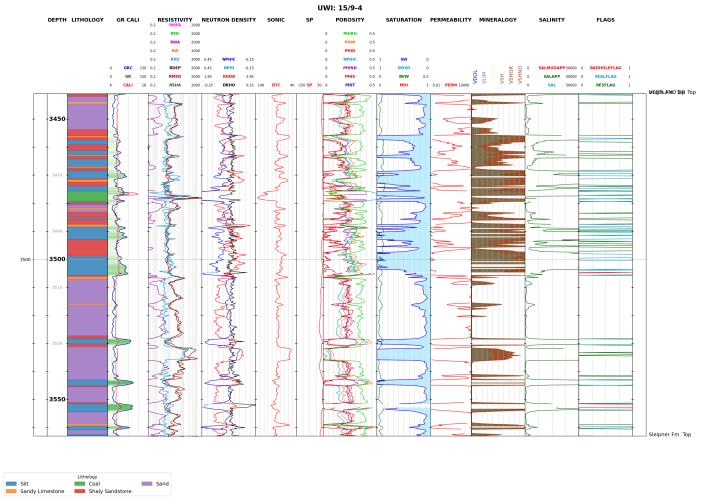


Figure 21: **Standardized well-log visualization layout.** The full-resolution auxiliary-evaluation plots contain 10–15 aligned tracks for raw logs, derived properties, interpretation flags, and lithology. This page-scale rendering is a layout overview; released source plots retain the labels and scales needed for quantitative inspection.

overview is intended to show layout rather than support reading individual track values at manuscript scale.

N Visual and Multimodal Study

Section 7.1 argues that structured context is what makes well-log reasoning tractable. This appendix tests that claim with vision-language models (VLMs). Six VLMs, including Qwen3-VL-235B-A22B (Bai et al., 2025), answer WellLogBench questions in two input settings: *text + plot*, where the model receives the Petrophys structured context together with the rendered

multi-track well-log plot of Figure 21 (Appendix M), and *plot only*, where it receives the plot image alone. Both settings use the rubric and judges of Section 4. The study is scored separately from the main leaderboard, so its numbers are compared with one another rather than with Table 1.

Table 24: **Text + plot versus plot-only performance** (0–100). Overall $= 0.4\,\mathrm{RA} + 0.2\,\mathrm{Trust} + 0.2\,\mathrm{Style} + 0.2\,\mathrm{RQ}$; Δ is plot-only minus text + plot Overall, with darker cells marking larger losses. *The best result per column is in **bold** and the second best is underlined.*

	Text + plot					Plot only					
Model	RA	Trust	Style	RQ	**Overall**	RA	Trust	Style	RQ	**Overall**	**Δ**
Gemini 2.5 Pro	56.8	**67.6**	**76.8**	74.2	**66.4**	**23.2**	**22.2**	**62.6**	**61.6**	**38.5**	−27.9
Kimi-K2.6	**57.6**	59.9	68.4	**74.7**	63.6	21.5	20.5	62.2	58.1	36.8	−26.8
Gemma 4 31B IT	50.7	59.3	67.3	72.1	60.0	18.7	18.4	57.5	54.5	33.6	−26.4
Qwen3.5-27B	48.0	53.7	64.7	68.1	56.5	18.6	19.3	57.9	55.4	34.0	−22.5
Qwen3-VL-235B-A22B	48.5	50.9	61.7	66.4	55.2	18.9	19.6	57.5	52.8	33.5	−21.7
LLaMA-4 Scout	33.7	39.9	66.6	52.3	45.2	16.7	17.1	53.9	47.4	30.4	−14.8

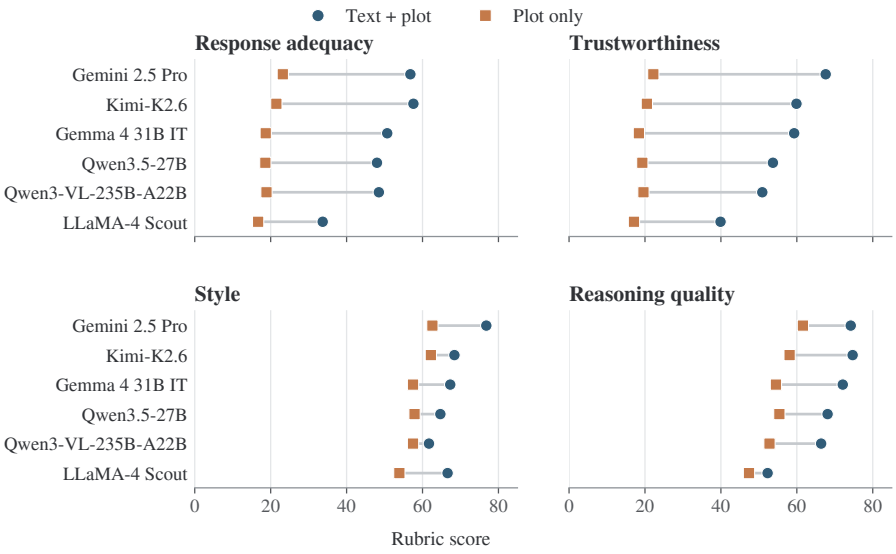


Figure 22: **Rubric change from text + plot to plot-only input.** Each connector links a model's plot-only score (square) to its text + plot score (circle). Response Adequacy and Trustworthiness collapse without the structured context, while Style and Reasoning Quality change far less.

Removing the structured context costs every model 15–28 Overall points (Table 24). The loss is concentrated in grounding: Response Adequacy falls by 17–36 points and Trust by 23–45 points, leaving both at 17–23 for every family, whereas Style and Reasoning Quality fall by only 4–18 points (Figure 22). Plot-only answers therefore remain fluent and plausibly organized but are rarely anchored in the measurements, which is the most costly failure mode in practice because such answers look credible. The model ranking is nearly unchanged across the two settings (only Gemma 4 31B IT and Qwen3.5-27B swap, 0.4 points apart), so the gap reflects the input representation rather than any particular model. These results support converting well logs into structured context, as Petrophys does, rather than relying on plot images alone.

O Example Structured Context JSON

To make a well-log segment usable by an LLM, Petrophys exports a structured JSON context that separates formation metadata, lithology-aware derived analysis, optional raw traces, and data-

completeness/QC metadata. The excerpt below shows the schema-level structure for a representative lithology-group context; long arrays and nested statistics are abbreviated.


```
{
  "formation_info": {
    "name": "BALDER FM. TOP",
    "min_depth": 2287.0698,
    "max_depth": 2386.8918,
    "thickness": 99.822
  },

  "analysis_results": {
    "well_name": "15_9-2",
    "formation_name": "BALDER FM. TOP",
    "lithology group": "SHALY SANDSTONE_GROUP_5",
    "lithology": "SHALY SANDSTONE",

    "lithology_group_info": {
      "start_depth": 2348.9441,
      "end_depth": 2354.7353,
      "thickness": 5.7913,
      "data_points": 39,
      "median_dz": 0.1523
    },

    "ground_truth_lithology_present": true,
    "core_phi_present": false,
    "core_k_present": false,

    "derived_logs_computed": [
      "BAD_HOLE", "BVW", "FZI", "INVASION_RATIO",
      "NET_PAY", "NET_RES", "PERM", "PHIE", "SW", "SXO",
      "..."
    ],

    "lithology_distribution": {
      "SHALY SANDSTONE": {
        "count": 39,
        "proportion": 1.0
      }
    },

    "statistics": {
      "GR_overall": {
        "minimum": 52.577,
        "maximum": 67.5193,
        "arithmetic_mean": 57.688,
        "..."
      },
      "RDEP_overall": {
        "minimum": 0.563,
        "maximum": 0.737,
        "arithmetic_mean": 0.6547,
        "..."
      },
      "PHIE_overall": { "..." : "same summary-statistics schema" },
      "SW_overall":   { "..." : "same summary-statistics schema" },
      "NET_PAY_overall": { "..." : "same summary-statistics schema" }
    },

    "analysis_summary": {
      "HYDROCARBON_ANALYSIS": {
        "petrophysical_hydrocarbon_indicators": {
          "description": "water saturation, mobility, and pay-flag assessment",
```

```
          "methodology": "uses SW, existing MHI when available, and NET_PAY",
          "calculation_thresholds": "POTENTIAL_PAY / MOVABLE_HC / LOW_SW /
          ↪  NO_INDICATION",
          "NO_INDICATION": { "..." : "interval-level summary" }
        },
        "geophysical_hydrocarbon_indicators": {
          "description": "neutron-density crossover and Vp/Vs-style screening",
          "methodology": "uses PHID-PHIN, VCL, PHIE, BAD_HOLE, and sonic
          ↪  availability",
          "calculation_thresholds": "STRONG_GAS_INDICATION / GAS_INDICATION /
          ↪  LOW_VPVS",
          "NO_INDICATION": { "..." : "interval-level summary" }
        },
        "reservoir_quality_for_hydrocarbons": {
          "description": "porosity-permeability reservoir-quality assessment",
          "methodology": "prioritizes effective porosity and permeability curves",
          "calculation_thresholds": "EXCELLENT_RESERVOIR / GOOD_RESERVOIR /
          ↪  FAIR_RESERVOIR",
          "GOOD_RESERVOIR": { "..." : "interval-level summary" }
        },
        "integrated_hydrocarbon_potential": {
          "description": "hierarchical integration of petrophysical, geophysical, and
          ↪  reservoir-quality signals",
          "methodology": "promotes prime oil/gas targets only when stricter criteria
          ↪  are met",
          "calculation_thresholds": "PRIME_GAS_TARGET / PRIME_OIL_TARGET /
          ↪  POTENTIAL_PAY / GAS_SHOW",
          "MARGINAL": { "..." : "interval-level summary" }
        }
      }
    },

    "data_completeness": {
      "missing_primary_logs": ["DTS", "PEF", "TEMP", "RMF", "RW"],
      "derived_log_computation_notes": [
        "dynamic petrophysical workflow with per-well routing",
        "temperature correction and NaN handling",
        "..."
      ]
    },

    "lithology_info": {
      "column_present": true,
      "total_rows": 39,
      "valid_lithology_percentage": 100.0,
      "unique_lithologies": ["SHALY SANDSTONE"]
    },

    "input_logs_present": [
      "BS", "CALI", "DRHO", "DTC", "DTE", "GR", "NPHI",
      "RDEP", "RHOB", "RMED", "SP", "..."
    ],

    "BADHOLE_QC_CHECK": {
      "total_points": 39,
      "bad_hole_points": 27,
      "bad_hole_percentage": 69.2308,
      "qc_method": "washout_or_large_hole",
      "summary_text": "bad-hole screening based on CALI, BS, and DRHO"
    }
  },

  "raw_log_traces_in_segment": {},
  "include_raw_traces_in_context": false
```

```
}
```

The exported context therefore gives the model both numerical evidence (summary statistics, derived logs, depth ranges) and provenance constraints (missing logs, QC checks, computation notes). Raw depth-by-depth traces can be attached under `raw_log_traces_in_segment` when needed, but are omitted in this example to keep the prompt compact.

P QUALITATIVE EXAMPLES

This appendix shows what interpretation questions and scored answers look like. Section P.1 follows one Single-Formation Hydrocarbon question from the logs to three models' answers; Section P.2 adds a Cross-Formation and a Multi-Well item from the other two basins; Section P.3 prints a complete WellLog-LLM answer. The stage-by-stage trace of one question is in Appendix K.2, and cross-model contrasts chosen by fixed rules are in Appendix K.5.

Figure 23 first illustrates the three reasoning levels on schematic logs. A Single-Formation question concerns one stratigraphic interval; a Cross-Formation question contrasts units within one well, so the answer must compare properties across a boundary; a Multi-Well question follows one formation across wells, so the answer must correlate thickness, properties, and pay laterally. In the hierarchical pipeline (Section 5.2), each formation or well is first analyzed on its own and the comparison is formed in the synthesis step.

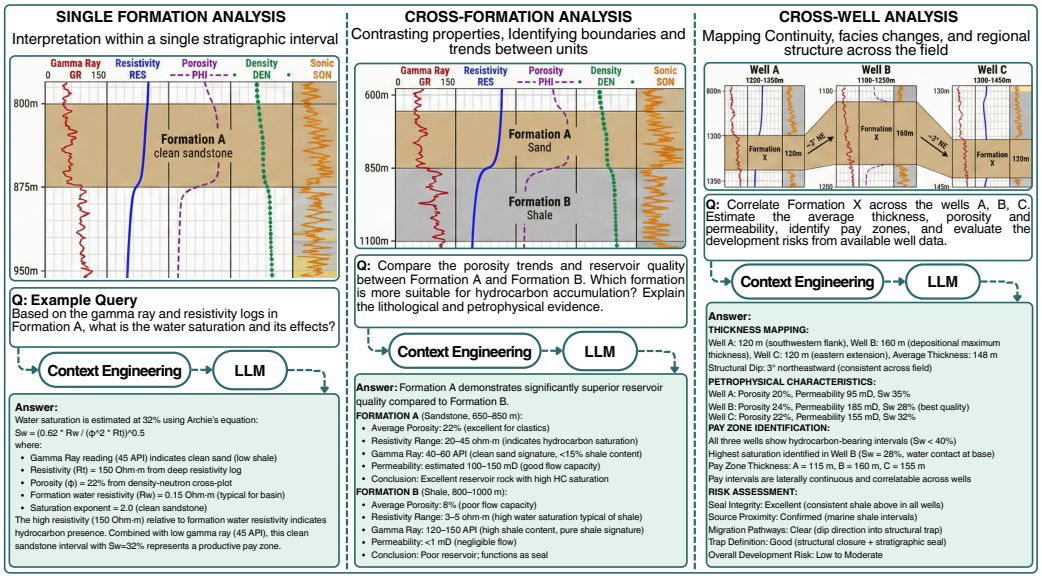


Figure 23: **The three reasoning levels in WellLogBench**, shown on schematic logs with an example question and answer for each: Single-Formation (left), Cross-Formation (middle), and Multi-Well (right; labelled cross-well analysis in the diagram). Values are illustrative, not benchmark items.

P.1 A HYDROCARBON QUESTION FROM LOGS TO SCORED ANSWERS

The question below is a public Single-Formation Hydrocarbon item from well 35/11-7. It asks the model to separate gas from liquid hydrocarbons by combining neutron–density separation with the Vp/Vs ratio, but the well has no shear sonic, so the second half of the requested test cannot be run. A good answer has to rank the crossover evidence interval by interval and then stop short of a fluid call.

Question. Analyze the relationship between neutron-density log separation and compressional-to-shear velocity ratios to distinguish between gas and liquid hydrocarbon signatures within the SOGNEFJORD FM. TOP.

Expert reference (abridged). The evaluated interval (1800.5–1961.7 m) is interbedded sand, shaly sandstone, silt, and shale. The strongest possible gas effect is at 1844.8–1850.6 m (mean PHID−PHIN separation 0.096 V/V, crossover flag active in most samples); 1831.1–1833.7 m is a weaker secondary candidate (0.044 V/V). Several other sands show only minor crossover, and some show none. Shear sonic (DTS) is absent, so Vp/Vs cannot be computed and the crossover cannot be attributed to gas rather than light oil, lithology, or clay effects. *Conclusion:* a ranked, log-based list of possible gas effects, not a validated gas or liquid classification.

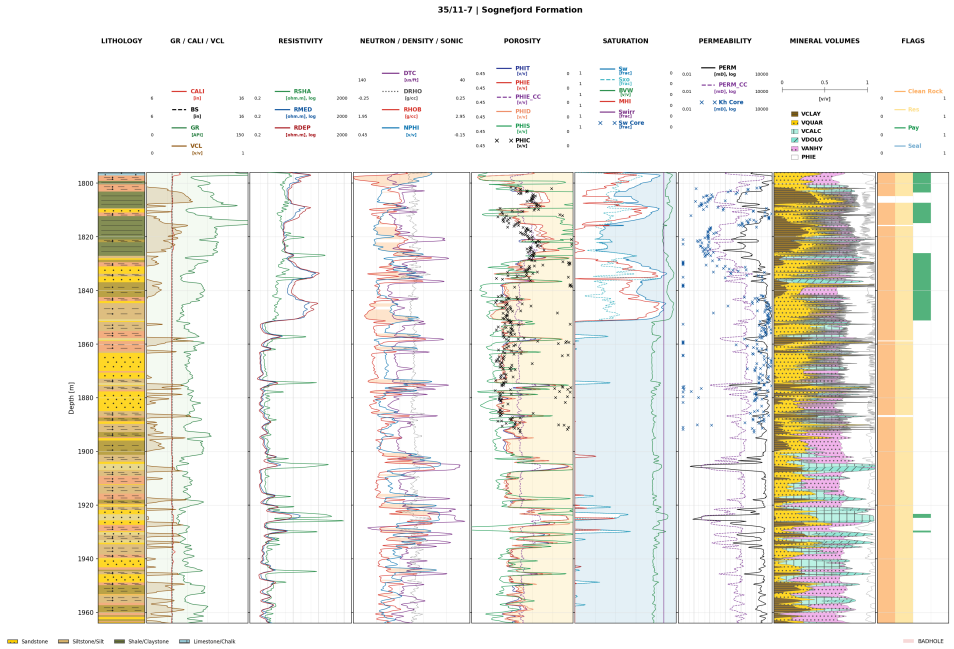


Figure 24: **Logs behind the Hydrocarbon example**: Sognefjord Fm. in well 35/11-7 (1796–1964 m). The formation alternates between sandstone, silt, and shale (lithology column), with sharp changes in resistivity and water saturation. Neutron–density crossover (shaded in the neutron–density–sonic track) is strongest around 1845–1851 m and weaker near 1831–1834 m; only compressional sonic (DTC) is present, so the Vp/Vs check the question asks for is impossible. Pay flags (right) mark several low-Sw sands, which is why models are tempted to call oil or gas outright.

Table 25 compares the final calls of three systems on this item. All three recognize that shear sonic is missing, and all three point to the same two crossover intervals. They differ in what they conclude from them. The DFS backbone turns the intervals into "Prime Oil Targets" and commercial pay, and GPT-5.6 Luna (high) reframes them as probably oil-prone; both go beyond what the logs can support. WellLog-LLM ranks the same intervals as possible gas shows and leaves the fluid type undetermined, as the reference does. This is the behavior that the GRCO rules on missing evidence target (Section 6.3), and it is typical of the Trust gains in Table 2.

P.2 Cross-Formation and Multi-Well Examples

Tables 26 and 27 show two more public items, one comparing formations within a well (Danish basin) and one comparing wells within a formation (COSTA). Both were chosen because the DFS backbone's conclusion contradicts the reference while WellLog-LLM's matches it; they illustrate typical failures rather than typical scores. In the Cross-Formation item the backbone recognizes the invasion evidence but explains part of it away; in the Multi-Well item it misses data that are present

Table 25: **Three answers to the Hydrocarbon example.** Overall (0–100) and the judge's Conclusion-by-Reference score (0–10); quotations are from the answers, and judge notes paraphrase the released GPT-5 explanations.

System	Overall / Concl.	Final call	Judge note
Qwen3.5-27B (DFS)	39.3 / 2	"a viable hydrocarbon prospect with distinct Prime Oil Targets" at 1831–1834 and 1845–1851 m	Declares prime oil and commercial pay instead of the reference's unvalidated, ranked gas-effect intervals.
GPT-5.6 Luna (high)	44.8 / 4	"supports a stronger oil/liquid interpretation than a confirmed gas interpretation"	Agrees gas cannot be confirmed without shear sonic, but reframes the gas-effect intervals as oil-prone and adds unsourced values.
WellLog-LLM	68.3 / 8	"the gas signature is unconfirmed" … crossover intervals "must be classified as GAS_SHOW rather than confirmed gas"	Matches: Vp/Vs cannot validate the crossover; 1844.8–1850.6 m strongest, 1831.1–1833.7 m secondary. Adds some unsupported saturation and pay metrics.

for one of the wells. In both cases WellLog-LLM, like GPT-5.6 Luna (high), keeps the comparison and reaches the reference conclusion.

Table 26: **Cross-Formation example** (Danish basin). Scores are Overall (0–100) / the judge's Conclusion-by-Reference score (0–10); Backbone is Qwen3.5-27B (DFS) and Luna is GPT-5.6 Luna (high). Reference findings are abridged; model behavior paraphrases the released GPT-5 judge explanations.

Cross-Formation, Permeability (Danish, well skjold-flank-1). "How do the mud filtrate invasion profiles, interpreted from resistivity log separations, compare between the TUXEN and HIDRA formations for assessing qualitative permeability?"

Reference		Both formations show resistive invasion (shallow resistivity above deep; mean invasion ratio 1.54 in HIDRA, 1.56 in TUXEN), so neither is a seal. The key difference is permeability: the TUXEN siltstone averages 105.9 mD against 4.3 mD in the HIDRA chalk, making TUXEN the better reservoir. Both intervals are flagged for bad hole, which limits confidence.
Backbone	54.8 / 4	Gets the shared invasion and the higher TUXEN permeability, but labels the HIDRA invasion an artifact and declares the two profiles non-comparable, contradicting the reference.
WellLog-LLM	80.0 / 9	Both formations show resistive invasion, so invasion alone does not separate them; TUXEN's much higher permeability makes it the stronger reservoir. Adds a few unreferenced saturation specifics.
Luna	81.2 / 9	Same conclusion as the reference; adds interval details not in the reference.

Table 27: **Multi-Well example** (COSTA). Scores and abbreviations as in Table 26.

Multi-Well, Permeability (COSTA, wells HW-29 and HW-6). "How many distinct hydraulic flow units (HFUs) are identified in the KHARAIB-2 formation for well HW-29 compared to well HW-6, and what are the average FZI and RQI values … ?"

Reference		Four HFUs in each well (mega-, macro-, meso-, and micro-pore). HW-29 has mean FZI 0.99 and RQI 0.30; HW-6 has mean FZI 1.46 and RQI 0.24. HW-29 has more storage and flow capacity, but HW-6 has the more efficient pore throats on average. Missing caliper, bit size, DRHO, and PEF limit borehole QC.
Backbone	42.7 / 2	Reports HW-6 correctly but claims there are no flow-unit data for HW-29, so it never makes the comparison; also labels depths in feet as meters.
WellLog-LLM	76.2 / 9	Four HFUs in both wells, with mean FZI and RQI matching the reference (HW-6 higher FZI, HW-29 higher RQI).
Luna	71.7 / 9	Same counts and values; adds unsupported claims about a deeper water-bearing interval.

P.3 A COMPLETE ANSWER

Shown below is the highest-scoring valid public Single-Formation WellLog-LLM response (Overall 92.5), condensed for space. This maximum-score example is illustrative, not representative.

Question. Are there zones in the KHARAIB-2 formation where medium resistivity is not numerically bracketed by shallow and deep resistivity, suggesting potential issues with environmental corrections or tool response?

Answer: Resistivity-ordering assessment in KHARAIB-2

Data availability. Displayed depths are converted from source feet at $0.3048\,\mathrm{m/ft}$. Across all four characterized lithological intervals, the only resistivity measurement present is RDEP. RSHA and RMED are absent from each interval's primary-log inventory. The standard test

$$\min(\mathrm{RSHA}, \mathrm{RDEP}) \leq \mathrm{RMED} \leq \max(\mathrm{RSHA}, \mathrm{RDEP})$$

therefore cannot be executed.

Interval	Depth (m)	Present	Missing
Limestone group 1	2551.48–2559.41	RDEP	RMED, RSHA, RXO
Limestone group 2	2559.89–2583.34	RDEP	RMED, RSHA, RXO
Limestone group 3	2583.13–2610.31	RDEP	RMED, RSHA, RXO
Limestone group 5	2613.05–2615.18	RDEP	RMED, RSHA, RXO

Interpretation. No depth-specific or interval-level RMED value is available to compare against RSHA and RDEP. Consequently, no zone can be confirmed as exhibiting out-of-order RMED behavior; nor can such behavior be definitively excluded. The missing measurements prevent direct assessment of radial resistivity profiling, invasion anomalies, or shoulder-bed effects.
Implication. In a complete three-resistivity suite, an unbracketed RMED response could indicate environmental-correction, shoulder-bed, or tool- response problems. Here, however, the absence of RSHA and RMED is a data gap, not evidence of a measured anomaly. Reliable ordering analysis requires finite, depth-aligned RSHA, RMED, and RDEP measurements.

The released GPT-5 judge notes that the response correctly identifies the missing logs and the resulting impossibility of the bracketing test. Its Coverage explanation flags that the candidate omits the reference answer's explicit 8371.0–$8580.0\,\mathrm{ft}$ formation range and interval-level RDEP statistics; other explanations note minor extra group labels and framing, without finding a conflict with the ground-truth conclusion.

Q PROMPT TEMPLATES

This appendix records abridged versions of the prompt templates used by the benchmark pipeline. Runtime placeholders such as `{formation_name}`, `{question}`, and `{context_data_json}` are filled programmatically. Repeated rubric anchors and runtime boilerplate are condensed for readability; these excerpts document the interfaces rather than serving as byte-for-byte prompt artifacts.

Q.1 LLM-AS-JUDGE PROMPTS

```
PRIMARY_SYSTEM_PROMPT
You are a domain-specific evaluator for petrophysical well-log interpretation.
Assess how accurately and concisely the LLM-generated response matches the
ground truth. Focus strictly on petrophysical reasoning, data alignment, and
domain-specific terminology. Penalize over-generation and hallucinated
interpretations, even if plausible. The question, ground truth, and LLM
response are all provided; consider all three.

For each metric, assign a score from 1 to 10 and provide concise reasoning
(1-3 sentences, <=50 words). If the response is empty, set all scores to 0
and use "No content to evaluate."

Core metrics:
```

```
1. Accuracy: correctness of facts and conclusions vs. ground truth.
2. Precision: focus; penalize unrelated/speculative details.
3. Faithfulness: strict adherence to ground truth; penalize unsupported facts.
4. Coverage: inclusion of key intervals, values, trends, and insights.
5. Conciseness: efficient answer without redundant elaboration.
6. Hallucination: fabricated curves, parameters, intervals, or interpretations.
7. Repetition: unnecessary repeated ideas or phrasing.
8. Clarity: clear expression of technical ideas.
9. Completeness: fully addresses the question/task.

Advanced by-reference metrics:
10. Conclusion by Reference: final interpretation matches the ground truth.
11. Petrophysical Claims by Reference: porosity, shale, saturation, pay, fluid,
    and reservoir-quality claims align with the reference.
12. Data Limitations Score by Reference: uncertainty/missing logs are respected.
13. Log Utilization by Reference: same or similar logs are used appropriately
    (e.g., GR, RHOB, NPHI, Rt, SW, PHIE, PERM).

Advanced free-thinking metrics:
14. Petrophysics Claims (Free Thinking): scientific validity independent of
    the ground truth.
15. Petrophysics Chain of Thought (Free Thinking): logical progression from
    data to interpretation.

Internal consistency metric:
16. Internal Consistency: no contradictions among porosity, Sw, Rt, lithology,
    pay, or fluid-type claims.

Additional instructions:
- Reread the ground truth and response before scoring.
- Compare interpretations, trends, and numerical conclusions rigorously.
- Prioritize correctness and alignment over surface fluency.
- Avoid quoting long spans; identify mismatches or omissions directly.

Return ONLY valid JSON:
{
  "<metric name>": float,
  "<metric name> Reasoning": "string",
  ... one score/reasoning pair for each of the 16 metrics above ...,
  "Overall Score": float
}
```

```
NUMERICAL_JUDGE_SYSTEM_PROMPT
You are a domain-specific evaluator for petrophysical well-log interpretation.
Score ONLY Numerical Accuracy for an LLM response.

Inputs:
- Question
- Ground Truth
- Reference Context
- LLM Response

Use the combined reference:
1. Ground Truth
2. Reference Context

Numerical Accuracy (1-10):
- Focus on correctness of numerical values and numerical trends that can be
  checked against the combined reference.
- Compare depths, thicknesses, porosity, saturation, resistivity, permeability,
  counts, percentages, and similar quantitative claims.
- Do not penalize missing numerics that are not necessary to answer the question.
- Minor formatting or rounding differences can still score highly.
- Penalize contradictory, fabricated, or materially incorrect numerics.
- If the response is empty, return 0 and "No content to evaluate."
- If the combined reference contains no useful numerics for the question, score
  based on whatever numerical overlap is verifiable and explain that limitation.

Return ONLY valid JSON:
{
  "Numerical Accuracy": float,
  "Numerical Accuracy Reasoning": "string"
}
```

Q.2 Question Generation Prompt

```
system_message =
"You are an expert petrophysicist generating standard interpretation questions
focused on {analysis_focus} for individual geological formations. Generate
questions that follow the provided example style closely without revealing any
internal context, analysis results, or data characteristics. Ask generic
```

```
interpretation questions that any petrophysicist would pose for single-formation
{analysis_focus} interpretation considering lithological variations. Never hint at
having analyzed the data or knowing specific formation characteristics."

user_prompt =
"Generate ONLY ONE interpretation question primarily focused on {analysis_focus}
for FORMATION: {formation_name}.

Internal context is provided for question formulation only and must not be
revealed:
- Question Category: {display_category_for_qgen}
- Well and Formation: {well_name}, {formation_name}
- Minimum Data: [{min_reqs_list_str}]
- Ideal Data: [{ideal_reqs_list_str}]
- Additional External Info: [{additional_info_str}]
- Comprehensive formation context JSON: {context_data_json}
- Lithology context: {lithology_context_json_str}

Question generation guidelines:
1. Focus on the specified formation.
2. Stay close to the example style: {example_q_style}
3. Consider lithological variations implicitly, but do not mention lithology
   group names or explicit lithological details.
4. Never reveal data availability, missing data, depths, thicknesses,
   statistical values, or internal analysis results.
5. Never mention visual/apparent/observed/plots/charts/images.
6. Do not mention specific logs or variable names from the context.
7. Do not start with 'should' or 'can'.
8. Output only the generated question, prefixed with 'Q: '."
```

Q.3 Dynamic Feature Selection Prompt

```
system_message =
"You are an expert in well-log and lithology interpretation.
Your task is to select the most relevant key paths from the {section_name}
section of a dataset. Respond strictly with a valid JSON array of key paths,
with no extra formatting, text, or explanations.

Selection principle:
- Include all keys that could plausibly contribute to answering the question.
- Include direct indicators plus secondary/supporting features.
- Stay within the {section_name} section; never invent keys or include keys
  from other sections."

user_prompt =
"Available Key Paths in {section_name}:
{available_keys_str}

{log_units_reference}

Question to Answer: {question}

Guidelines:
1. Use the log measurement unit reference to decide which measurements or
   derived statistics can help answer the question.
2. Prefer a broad but focused selection; avoid extreme minimalism and irrelevant
   keys.
3. If the question asks about petrophysical properties, include lithology-related
   measurements or mineral-volumetric curves where relevant (e.g., VCLAY, VLIM,
   VDOL).
4. Return ONLY a valid JSON array, e.g.:
   ["statistics.PHIS_overall", "statistics.lithology_distribution"]"
```

Q.4 Inference Prompt

```
system_message =
"You are an expert petrophysicist providing detailed, data-driven interpretive
answers about {analysis_focus}, {key_concepts}, {technical_areas}, and reservoir
implications for formation {formation_name}. Focus on formation-level interpretation
with lithological variation by depth. For statistical lookup questions, report
values from provided statistics. For hydrocarbon questions, focus on derived
indicators and patterns across intervals. Use proper units and markdown with
headings, tables, and clear conclusions. Never mention visual data, plots,
charts, analysis files, or computational workflows."

user_prompt =
"Answer the question about formation {formation_name} with strong focus on
{analysis_focus}, {key_concepts}, {technical_areas}, and reservoir implications.

Question:
```

```
Q: {question}

Context:
- Log units: {log_units_reference}
- Lithology interval data: {LITHOLOGY_INTERVAL_DATA_PLACEHOLDER}
- Minimum data requirements: [{min_reqs_list_str}]
- Ideal data requirements: [{ideal_reqs_list_str}]
- Additional useful external information: [{additional_info_str}]
- Revision/plot guidance used internally only: {revision_guidance_section}
  {internal_plot_guidance}

Forbidden language:
- 'ground truth', 'provided data', 'analysis results', 'data coverage',
  'input dataset', 'computation pipeline', 'analysis files', 'visual',
  'plot', 'chart', or references to data-generation processes.

Answer guidelines:
- Be concise and analytical.
- Center the answer on formation-level interpretation.
- Always reference specific depth intervals when discussing properties.
- Use proper units for all values.
- Use markdown headings and tables for numerical data.
- Acknowledge missing logs/data limitations and explain how they affect
  reliability.
- Use derived logs when they are more appropriate than raw logs.
- Make only claims supported by the supplied interval data and summary.

Required output:
A: [structured markdown answer with headings, tables, interval-specific
evidence, limitations, and conclusions]

REASONING:
[Bulleted list citing the available data sources, calculations, depth intervals,
missing data, and confidence limitations. Do not mention hidden workflow or
generation artifacts.]

MINIMUM_DATA_REQUIREMENTS_FOR_THIS_QUESTION_CATEGORY:
[{min_reqs_list_str}]

IDEAL_DATA_REQUIREMENTS_FOR_THIS_QUESTION_CATEGORY:
[{ideal_reqs_list_str}]

ADDITIONAL_EXTRA_EXTERNAL_INFORMATION_USEFUL_FOR_THIS_CATEGORY:
[{additional_info_str}]

DIFFICULTY: {difficulty}"
```

Q.5 GRCO PROMPTS

The GRCO judge and meta steps (Appendix J.4) use the following templates. `{experiences}` renders the current library as numbered rules (`[G0]`, `[G1]`, ...).

```
POLICY PROMPT WITH EXPERIENCES (learning rollouts)
**Task:** Solve the input problem by leveraging relevant insights from your
accumulated experiences.
1. Understand the Problem: analyse the CURRENT PROBLEM to identify key aspects.
2. Review Relevant Experiences: determine which insights, strategies, or
   patterns in the ACCUMULATED EXPERIENCES apply.
3. Apply Insights Thoughtfully: incorporate matching experiences; otherwise
   consider whether general principles can still guide the approach.
**ACCUMULATED EXPERIENCES:** {experiences}
**CURRENT PROBLEM:** {problem}

JUDGE (reward)
You are a domain-specific evaluator for petrophysical well log analysis. Assess
how accurately and concisely the response matches the ground truth. Penalise
over-generation and hallucinated interpretations. Score 1-10: Accuracy,
Faithfulness, Coverage, Numerical Accuracy, Completeness, Format. Overall Score
= average of the six metrics. Return only JSON.

STEP 1 - SUMMARIZE EACH ANSWER
Summarise the response: (1) approach/methodology; (2) given the evaluation and
correct answer, errors, inaccuracies, or missed information; (3) all core
findings and conclusions, even if flawed. Sections: Approach, Key Claims,
Errors/Issues, Strengths, Overall Strategy.
Input: question, correct answer, response, "Overall Score: {reward}/10".

STEP 2 - CRITIQUE A LEARNABLE GROUP
You are reviewing an LLM across multiple attempts at the same petrophysical
question. (1) Classify each attempt as correct/incorrect with justification.
(2) Diagnose root causes of failure. (3) Identify key success factors.
(4) Compare successful vs unsuccessful approaches. (5) Extract generalisable
insights about strategic reasoning patterns, not specific calculations.
```

```
(6) Every experience must be a complete, self-contained instruction.
Input: <Problem>, <Correct Answer>, <Attempts> (summaries labelled correct or
wrong), <Current Experiences>. Output: JSON list with at most 1 operation,
{"option": "add", "experience": ...} or
{"option": "modify", "experience": ..., "modified_from": "G17"}.

STEP 3 - CONSOLIDATE THE BATCH
You are an experience manager for an LLM petrophysical well-log system. Process
a batch of proposed operations into one revision plan (ADD, UPDATE, DELETE,
NONE). Multiple UPDATEs for the same ID -> synthesise into one. DELETE overrides
UPDATE for the same ID. Merge similar ADDs. Each experience must be clear,
generalisable, and self-contained, begin with brief context, and focus on
strategic patterns. Never truncate an experience.
Output: [{"operation": ..., "id": existing_id_or_null, "content": ...}]
```