A Domain-Expert-Curated Benchmark and Domain-Specialized LLM for Subsurface Well Log Interpretation

Anonymous ACL submission

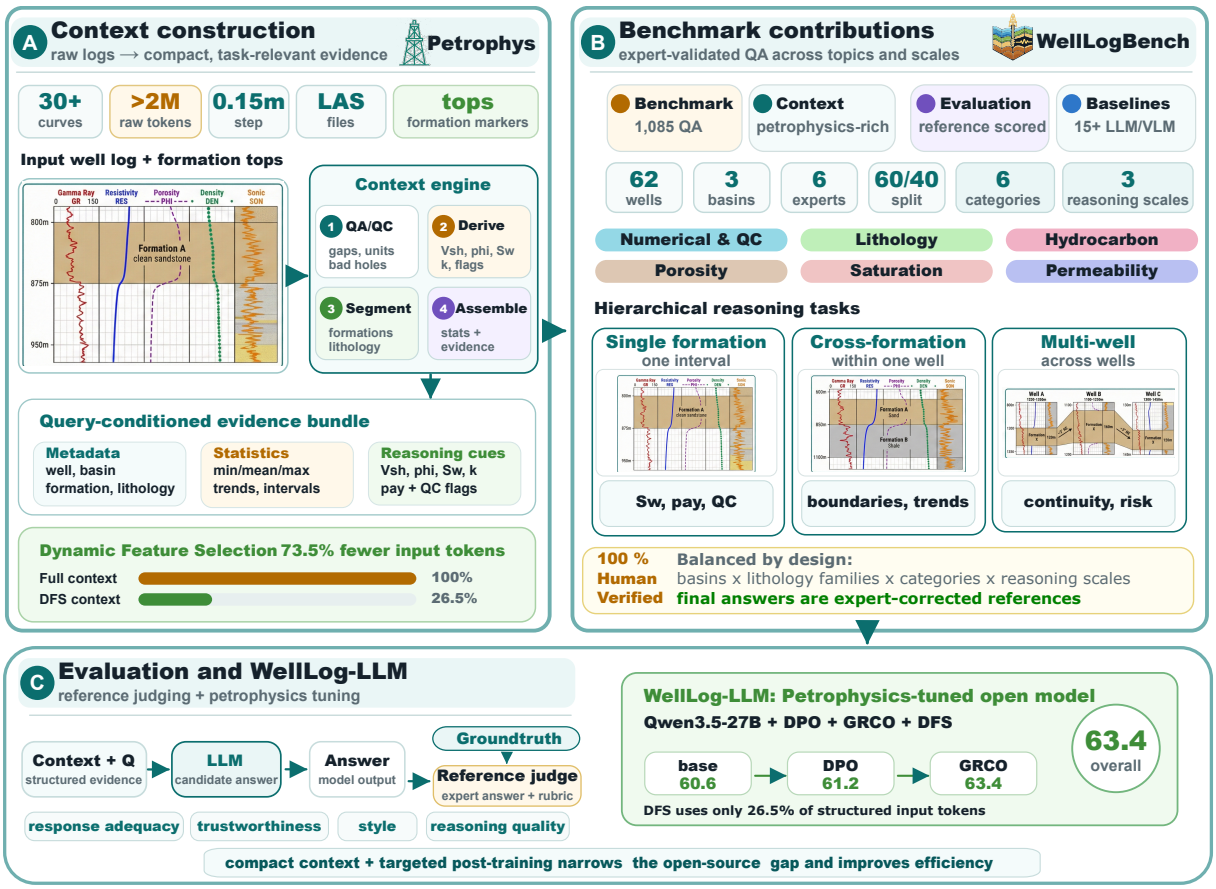


Figure 1: **Overview of WellLogBench and WellLog-LLM:** Well-log measurements, in the form of raw LAS files and formation tops, are cleaned, segmented by formation and lithology, and converted into compact petrophysical contexts containing metadata, summary statistics, derived properties, and quality-control flags. We curate a comprehensive benchmark with 1,085 expert-validated QA pairs spanning six petrophysical topics and three reasoning scales: single-formation interpretation, cross-formation comparison, and multi-well synthesis. The same structured context is used for reference-based model evaluation and for our open model, WellLog-LLM, which combines petrophysics tuning with Dynamic Feature Selection to reduce the input context.

Abstract

Subsurface well-log interpretation is a challenging task that requires the ability to reason over long-context multivariate well-log curves. Although large language models (LLMs) could accelerate this process, their effectiveness for this task remains largely unexplored. We introduce **WellLogBench**, the first expert-curated benchmark for well-log reasoning, comprising 1,085 expert-validated QA pairs covering six petrophysical categories and three levels of reasoning complexity, ranging from single-formation interpretation to cross-formation and multi-well synthesis. We introduce several contributions to enable LLM-based reasoning over raw LAS logs. First, we develop **Petrophys**, a Python petrophysics library that standardizes calculations and ensures numerical reproducibility. Building on this library, we introduce a context-engineering pipeline that transforms million-token depth-indexed raw well logs into compact lithology-aware representations, enabling LLM-based interpretation of well-log data. We further curate SFT, DPO, and context-learning training data and develop a domain-specialized model family, **WellLog-LLM**, that: (a) reduces the input context for efficient LLM inference through the proposed Dynamic Feature Selection (DFS),

and (b) introduces domain-specialized LLMs through supervised fine-tuning (SFT), direct preference optimization (DPO), and a context-optimization approach, Group Relative Context Optimization (GRCO). We perform comprehensive evaluations across more than 15 LLMs/VLMs. Gemini 3.5 Flash achieves the best overall performance at 74.0%, and Kimi-K2.6 is the strongest public open-source baseline at 64.2%. Our **WellLog-LLM-Qwen-27B** reaches 63.1% Overall (+2.5 over Qwen3.5-27B; within 1.1 points of Kimi-K2.6), while **WellLog-LLM-K2V2** reaches 56.0% (+13.5 over K2-V2 Instruct); both require only $\sim$27% of the full-context input tokens. Overall, WellLogBench, the Petrophys library, and WellLog-LLM provide a reproducible benchmark and model family for grounded petrophysical reasoning and facilitate future research on domain-specialized LLMs for subsurface interpretation. We will open-source our code, evaluation benchmark, the Petrophys library, and model weights.

1 Introduction

Large Language Models (LLMs) have emerged as powerful scientific reasoning assistants (Wei et al., 2022; Kojima et al., 2022; Qiu et al., 2025; Dang et al., 2025), but their ability to interpret *well logs*, the depth-indexed petrophysical measurements that drive subsurface characterization in oil and gas industry, remains unexplored. Applying LLMs here poses unique challenges that are less common in existing benchmarks: raw LAS files often exceed **2M tokens** across 30+ interdependent curves; correct interpretation requires both interval-level decisions and cross-formation/multi-well synthesis; accuracy depends on derived petrophysical quantities (e.g., V_{sh}, ϕ, S_w, pay) computed via lithology-aware pipelines rather than free-form text; and plausible-sounding outputs that violate physical constraints (invasion effects, lithology bounds) are dangerous in production use.

To bridge this gap, we introduce **WellLogBench**, a domain-expert curated benchmark comprising 1,085 QA pairs over 62 real-world wells from three basins, validated by **six senior petrophysicists**. It covers six petrophysical categories and three reasoning levels (single-formation, cross-formation, multi-well; Figure 1), paired with a context-engineering pipeline that compresses raw LAS files[1] into lithology-aware structured contexts within practical LLM context limits.

Our contributions are:

(i) Benchmark (§3.4): The first expert-curated benchmark for well-log reasoning, with 1,085 QA pairs over 62 wells from 3 basins, with a 60/40 public/private leaderboard split.

(ii) Context engineering & inference pipeline (§3.5–§5): We develop **Petrophys**, a comprehensive open-source petrophysics library[2] used to construct WellLogBench contexts from LAS files via QA/QC, lithology-aware segmentation, petrophysical derivation, rule-based flagging, and LLM-ready context assembly; inference then uses dynamic feature selection and hierarchical multi-well decomposition for practical token budgets. Appendix O shows the exported JSON schema.

(iii) Evaluation framework (§4):Dual reference-based LLM-as-a-Judge approach that scores each response against the expert-curated ground truth across four rubrics, validated by blinded expert ratings and cross-judge robustness.

(iv) WellLog-LLM variants (§6): A two-backbone petrophysics-tuned model family built with SFT, DPO, and Group Relative Context Optimization (GRCO); **WellLog-LLM-Qwen-27B** (Qwen3.5-27B backbone) reaches 63.1 Overall, approaching Kimi-K2.6 at an order-of-magnitude smaller scale.

(v) Baseline evaluation (§7): Evaluation of 15+ LLMs and VLMs across text, multimodal, and visual-only settings, identifying persistent gaps in grounded petrophysical reasoning.

2 Related Work

Scientific and Geoscientific Reasoning. General science benchmarks (ScienceQA (Lu et al., 2022), MMLU (Hendrycks et al., 2021)) omit petrophysics and the specialized data formats needed for subsurface interpretation. Geoscience-focused efforts (K2 (Deng et al., 2023), GeoGalactica (Lin et al., 2024), (Hadid et al., 2024)) pre-train on geological corpora but target textual understanding rather than quantitative, multi-signal well-log interpretation.

AI in Petrophysics. Prior work has progressed from supervised estimation of reservoir properties (Hassaan et al., 2024; Rahmati et al., 2024; Zhang et al., 2025) to deep lithofacies models (Shi et al., 2024; Jiang et al., 2024; Mohammadi et al., 2025), physics-informed architectures (Shao et al., 2024; Pothana and Ling, 2025), well-log foundation mod-

[1] See Appendix L for an example.

[2] Anonymous project page: petrophys.netlify.app.

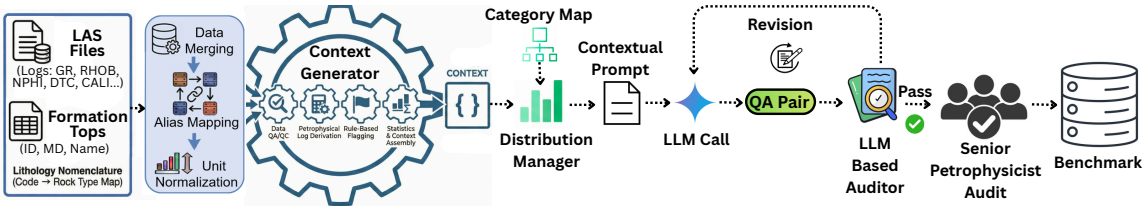


Figure 2: End-to-end QA generation pipeline showing the flow from raw well log data through context generation, LLM-based QA creation, automated auditing, and domain expert verification.

els (Qi et al., 2025; TGS, 2025), and generative imputation (Zhangxing et al., 2025; Al-Fakih et al., 2025). LLM adaptations target petroleum workflows via domain pre-training (Eckroth et al., 2025), report drafting (Li et al., 2025), and tool integration (Xu and Hulea, 2024; Srivastav and Kothe, 2025). To our knowledge, no prior benchmark evaluates multi-step *reasoning* over raw, ultra-long, multivariate well-log contexts.

Long-Context Time-Series Reasoning. Well logs are ultra-long, multivariate depth-indexed time series requiring interval aggregation, trend reasoning, and cross-curve consistency. Existing long-context benchmarks (LongBench (Bai et al., 2024), RULER (Hsieh et al., 2024)) are text-centric; time-series evaluations report persistent zero-shot temporal-reasoning gaps (Merrill et al., 2024; Chow et al., 2024; Kong et al., 2025). WellLogBench fills this gap with expert-validated petrophysical reasoning over ultra-long well-log contexts.

3 WellLogBench: Design and Data Creation

Overview. WellLogBench comprises 1,085 expert-validated QA pairs from 62 public LAS wells across three basins, covering six petrophysical categories and three reasoning levels. Items are produced by a four-stage pipeline (Figure 2): (1) LAS ingest and cleanup, (2) lithology-aware context engineering with derived quantities, (3) QA generation with Gemini 2.5 Pro (Comanici et al., 2025), (4) GPT-5 (OpenAI, 2025a) auditing followed by expert validation. Three design principles guide construction: (i) *domain authenticity*: questions drawn from real interpretation/reporting workflows, not synthetic settings; (ii) *reasoning complexity*: from direct retrieval and computation to multi-step interpretation with cross-curve consistency checks; (iii) *geological diversity*: multiple basins, lithologies (carbonate, clastic, mixed), and acquisition conditions to mitigate overfitting.

Dataset	Raw logs used	Cross	Multi	Single	Total
GEOLINK	42	251	205	221	677
COSTA	15	0	195	76	271
DANISH	5	55	32	50	137
Total	**62**	306	432	347	**1,085**

Table 1: Distribution of QA pairs across datasets, raw logs used, and reasoning levels.

Category	Cross	Multi	Single	Total
Porosity	59 (19.3%)	71 (16.4%)	48 (13.8%)	178
Permeability	55 (18.0%)	79 (18.3%)	66 (19.0%)	200
Saturation	55 (18.0%)	76 (17.6%)	59 (17.0%)	190
Lithology	52 (17.0%)	82 (19.0%)	57 (16.4%)	191
Hydrocarbon	49 (16.0%)	77 (17.8%)	56 (16.1%)	182
Numerical	36 (11.8%)	47 (10.9%)	61 (17.6%)	144
Total	306	432	347	1085

Table 2: Combined category distribution across all three basins.

3.1 Data Sources

We curate three publicly available repositories spanning the two dominant well-logging lithology families plus chalk (Table 1): **Geolink** (North Sea, mixed clastic–carbonate, 677 pairs (Gama et al., 2024)), **COSTA** (Middle Eastern carbonate, 271 pairs (Costa Gomes et al., 2022)), and **Danish Basin** (Upper Cretaceous–Danian chalk, 137 pairs (Geological Survey of Denmark and Greenland (GEUS))). Detailed formations, lithologies, and basin-specific coverage are provided in Appendix A.

3.2 Petrophysical Categories

Questions span six categories that exercise complementary reasoning skills (numerical computation, multivariate integration, causal inference, 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). Table 2 shows the balanced distribution.

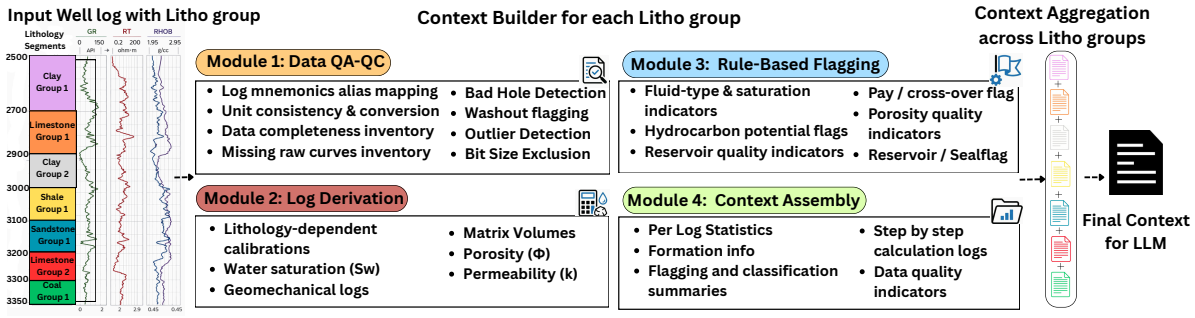


Figure 3: Context Generation Engine architecture. Input data sources are merged and segmented by lithology group. Each segment passes through four processing modules: Data QA/QC, Petrophysical Log Derivation, Rule-Based Flagging, and Statistics Assembly—producing structured contexts per lithology group.

3.3 Reasoning Levels

Three levels of complexity (Figure 1): **Single-Formation** (analyse log responses and derive properties within one stratigraphic interval), **Cross-Formation** (compare/contrast across formations within one well, identifying boundaries and explaining variations), and **Multi-Well** (regional-scale correlation and trend synthesis across wells for the same formation).

3.4 QA Generation Pipeline

Items follow a 4-stage pipeline; ground-truth answers are expert-corrected and rubric-validated, not raw model outputs.

Stage 1: Ingestion & Preprocessing. Parse LAS curves and merge with formation tops and lithology labels. We harmonize heterogeneous sources via (a) mnemonic mapping (∼200 vendor mnemonics → 25 canonical curves) and (b) unit normalization (m, v/v, Ωm, g/cm^3, μs/ft). **Stage 2: Context & QA Generation.** Each well is converted to a structured petrophysical context (§3.5). A *Category Map* defines templates (required curves, derived quantities, example questions, difficulty); a *Distribution Manager* enforces balanced coverage (15–20% per category) across reasoning levels, basins, and lithologies via weighted sampling; a *generation module* produces (Q, A) pairs under domain-constrained prompts.[3] **Stage 3: Auditing.** A GPT-5 auditor validates each pair, revises failures, and rejects unresolved cases for regeneration. Revision rates are 71.8% / 77.4% / 76.0% for single/cross/-multi, so the final benchmark is not a direct mirror of generator outputs. **Stage 4: Expert Verification.** Six senior petrophysicists (15+ yrs) verify technical accuracy and correct or reject errors; expert corrections occur on 59.2% / 61.0% / 59.5% of single/cross/multi items, with auditor+expert union reaching 81.4% / 86.4% / 85.2% (Appendix C); the full annotation workflow and rubric used by the expert panel are described in Appendix B.

[3] Question-generation and answer-generation prompt templates are provided in Appendix Q.

3.5 Context Engineering Framework

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 comprehensive petrophysics library for benchmark context generation that will be released with Well-LogBench; anonymous documentation is available at petrophys.netlify.app. Figure 3 summarizes the framework, and Appendix O shows a representative JSON schema exported by the pipeline.

Lithology-based segmentation. Averaging properties across mixed lithologies yields misleading statistics, so we segment each formation into contiguous lithologically homogeneous intervals by aligning log data to formation tops, mapping lithology codes to standardized groups (Appendix A), and assigning sequential segment identifiers; each segment is processed independently with lithology-specific parameters.

Four processing modules. Each segment passes through: (1) **QA/QC** (bad-hole detection, missing-data inventory, outlier filtering, quality-flag propagation); (2) **Petrophysical Derivation** (shale volume, porosity, water saturation, permeability with lithology-dependent calibrations); (3) **Rule-Based Flagging** (expert-validated binary flags such as RESFLAG, SEALFLAG, PAY_FLAG and discrete quality bins); (4) **Statistics & Context Assembly** (per-log statistics and final context with metadata, classifications, and data-quality fields).

Output modes. *Single-Formation Mode* produces one context per lithology group, aggregated to a

per-formation context; *Cross-Formation* and *Multi-Well Modes* aggregate further for comparative questions. A raw LAS file exceeds 2M tokens; the resulting structured single-formation context averages ∼50k tokens, while cross-formation and multi-well contexts scale with the number of formations compared. The next section describes the inference-time pipeline that operates over these contexts within practical LLM context budgets.

4 Evaluation

Each item includes a question-specific, expert-corrected *ground-truth answer*. We use a dual, reference-based LLM-as-a-Judge protocol.[4] The primary judge, GPT-5, scores the candidate answer against the ground truth across the main evaluation dimensions. In parallel, a hosted Qwen3.5-27B judge scores *Numerical Accuracy* using the question, ground truth, candidate answer, and the lithological/log context. Both judges are grounded in the expert reference; the added context for the numerical judge allows verification of values that are derivable from the logs even when not stated explicitly in the reference. All scores are reported on a 0–100 scale.

(1) **Response Adequacy (RA)**. Does the answer match the expert reference?
(i) *Accuracy*: Agreement with the ground-truth conclusion, (ii) *Petrophysical Claims*: Each cited log or quantity is supported by GT, (iii) *Conclusion*: Final interpretation matches GT, (iv) *Coverage*: No required GT element is omitted.
(2) **Trustworthiness (Trust)**. Is every claim anchored in the cited context?
(i) *Faithfulness*: Claims are verifiable against the provided context, (ii) *Numerical Accuracy*: Depths, cutoffs, and derived quantities are arithmetically correct, (iii) *Data Limitations*: Uncertainty acknowledged where the GT does, (iv) *Log Utilization*: The logs the GT relies on are actually invoked.
(3) **Style**. Is the answer well-communicated? *Clarity*, *Conciseness*, and *Repetition*, scored per-section.
(4) **Petrophysical Reasoning Quality (RQ)**. Does the reasoning hold up without the reference? GT is hidden from the judge to probe independent reasoning.
(i) *Petrophysics Claims*: Each claim is supported by the logs, (ii) *Chain of Thought*: The derivation is internally valid.

[4]Primary and numerical judge prompt templates are provided in Appendix Q.

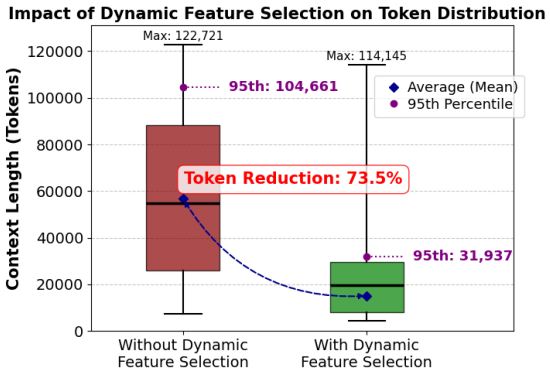


Figure 4: Impact of Dynamic Feature Selection on context length distribution; 73.5% mean token reduction at single-formation scale.

Weighted score. 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=0.995, ρ=0.994; Appendix F).

Judge validation. (i) *Cross-judge:* re-scoring the shared baseline cohort with Qwen3 235B A22B Instruct, Gemma-4 31B IT, and Qwen3.5-27B yields Pearson $r \geq 0.92$ on RA/Trust/RQ and overall rank ρ=0.939 (Appendix E). (ii) *Expert-LLM alignment:* senior petrophysics experts blindly scored 50 Gemini-2.5-Pro vs. GPT-5 answer pairs. Expert and LLM-as-judge ratings show good absolute agreement, with $\mathrm{ICC}(A,1)$=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 D.

5 Inference Pipeline

Even after context engineering, single-formation contexts (∼50k tokens) and multi-formation contexts can exceed practical LLM budgets and inflate inference cost. We adopt two complementary inference-time strategies[5]: query-conditioned *Dynamic Feature Selection* for single-formation items,[6] and *hierarchical decomposition* for cross-formation and multi-well items.

[5]The general inference prompt template is provided in Appendix Q.
[6]The dynamic feature-selection key-path prompt is provided in Appendix Q.

5.1 Dynamic Feature Selection (DFS)

An intermediate LLM call receives the full structured context plus the target question, then extracts only task-relevant statistics into a compact context that is fed to the same model for answer generation:

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

This query-conditioned compression preserves task relationships while cutting single-formation input tokens by 73.5% on average (Figure 4).

Cost-accuracy trade-off. DFS is used in the deployable WellLog-LLM pipeline, where it cuts input tokens by ∼73% while retaining enough task evidence for WellLog-LLM-Qwen-27B to remain competitive with much larger baselines. The main results therefore report baseline models with full structured context and WellLog-LLM variants with DFS; per-rubric DFS-vs-full-context deltas are retained as a sensitivity study in Appendix H.

5.2 Hierarchical Cross-Formation & Multi-Well Decomposition

Concatenating N formation/well contexts would blow past every LLM's budget. We instead decompose the query: each formation/well is processed independently, producing a local response, and a second LLM call synthesizes the locals into a comparative answer:

$$\{\text{Context}_i\}_{i=1}^{N} + Q \xrightarrow{\text{LLM}} \{R_i\}_{i=1}^{N} \xrightarrow{\text{LLM}} A. \quad (2)$$

This map-then-synthesise pattern scales linearly in N and keeps every per-formation context within native limits.

6 WellLog-LLM: Petrophysics-Tuned Variants

To test whether targeted post-training can close the gap between mid-scale open-source and frontier closed-source models, we instantiate **WellLog-LLM** as two named variants: **WellLog-LLM-Qwen-27B** uses the **Qwen3.5-27B** backbone, and **WellLog-LLM-K2V2** uses the **K2-V2 Instruct** backbone. The recipe separates three roles—domain adaptation, preference sharpening, and inference-time experience refinement—all on 65k-token DFS contexts:

- **SFT:** ∼30k WellLogBench-style QA triples (10k per reasoning level), drafted by Kimi-K2.6 (Moonshot AI, 2025), to teach domain vocabulary, answer structure, and use of DFS-formatted petrophysical evidence. The long-context SFT hyperparameters and infrastructure are detailed in Appendix I.1.
- **DPO:** ∼15k preference pairs (5k per reasoning level), using Kimi-K2.6 responses as chosen samples and Qwen3-8B responses as rejected samples. The goal is contrastive grounding: DPO increases the likelihood margin for answers that better use the provided logs, support conclusions from the reference context, and avoid unsupported petrophysical claims under the delta-learning hypothesis (Geng et al., 2025). DPO training details are in Appendix I.2.
- **Group Relative Context Optimization (GRCO)** (Cai et al., 2025): ∼18k prompts (1k per category × 6 categories × 3 reasoning levels). We use GRCO as an offline stage that optimizes an external buffer of instructions and experiences: the current DPO checkpoint — K2-V2 Instruct or Qwen3.5-27B, depending on the backbone — generates grouped rollouts; a Qwen3.5-27B rubric judge/meta model scores them and distills the contrast between stronger and weaker responses into reusable domain experiences. At inference, retrieved experiences are inserted into the otherwise fixed prompt as an experience prior; checkpoint weights are not updated. Representative learned experience families and rules are shown in Appendix I.4.

Recipe details, including the long-context SFT configuration and compute setup, are in Appendix I; benchmark performance and a stage-by-stage attribution are reported in Section 7.

7 Results

We evaluate on a **60% public test set** (651 unique questions); the held-out private set is reserved for the leaderboard and described in Appendix N. Three evaluation fronts probe complementary capabilities: *(i)* the structured petrophysical context as primary input, with baseline full-context scores in Table 3 and DFS sensitivity in Appendix H; *(ii)* multimodal: context paired with standardised well-log plots; and *(iii)* plot-only on the rendered well-log image alone. All models are scored under the rubric of §4.

7.1 Context-Based Evaluation

Table 3 reports the full cohort under the primary context-based setting. Baseline models use the full structured petrophysical context, while WellLog-LLM variants use the deployable DFS pipeline

Model	Params	Response Adequacy	Trust-worthiness	Style	Reasoning Quality	Overall
GPT-5 (OpenAI, 2025a)	—	62.5	63.8	69.4	77.9	67.2
Gemini 2.5 Pro (Comanici et al., 2025)	—	68.9	69.9	78.1	81.1	73.4
Gemini 3.5 Flash (Kavukcuoglu et al., 2026)	—	**69.5**	**70.3**	**79.1**	**81.7**	**74.0**
LLaMA-3.1 8B Inst. (Meta, 2024a)	8B	22.0	29.3	48.4	37.3	31.8
LLaMA-3.3 70B (Meta, 2024b)	70B	29.8	40.9	61.7	49.3	42.3
LLaMA-4 109B (Meta, 2025)	109B	36.3	43.9	64.5	52.3	46.7
K2-V2 Instruct (K2 Team, 2025)	70B	34.7	33.6	55.8	53.6	42.5
GPT-OSS-120B (OpenAI, 2025b)	117B	44.0	45.7	54.5	59.7	49.6
Gemma-4 31B IT (DeepMind, 2025)	31B	49.4	55.1	63.7	67.3	57.0
Qwen3-8B (Yang et al., 2025)	8B	33.5	38.5	54.5	51.0	42.2
Qwen3.5-27B	27B	52.3	60.3	66.5	71.6	60.6
Qwen3 Next 80B A3B (Team, 2025)	80B	43.8	45.4	58.3	60.0	50.3
Qwen3 235B A22B Inst. (Yang et al., 2025)	235B	48.1	49.4	60.1	65.9	54.3
Kimi-K2.6 (Moonshot AI, 2025)	1.1T	58.9	61.3	67.9	74.1	64.2
WellLog-LLM-K2V2	70B	49.5	54.0	65.2	61.8	56.0
WellLog-LLM-Qwen-27B	27B	55.8	62.4	69.3	72.1	63.1

Table 3: Model performance on the public test set (0–100). **Params** = reported total parameter count ("—" when undisclosed). Baselines use full structured context; WellLog-LLM variants use DFS, requiring only $\sim$27% of full-context input tokens. **Overall** $= 0.4\cdot$Response Adequacy$+0.2\cdot$Trustworthiness$+0.2\cdot$Style$+0.2\cdot$Reasoning Quality. Best per column in **bold**; per-stage attribution in Table 4.

(§5.1), which reduces input tokens by roughly 73% relative to full-context inference while preserving task-relevant petrophysical evidence. Appendix G lists all model variants and modality settings.

Headline findings. The closed-source frontier leads: Gemini 3.5 Flash reaches 74.0 Overall, followed by Gemini 2.5 Pro at 73.4 and GPT-5 at 67.2. Within the open-source group, the $\sim$1.1T-parameter Kimi-K2.6 sets the ceiling at 64.2, ahead of Qwen3.5-27B (60.6), Gemma-4 31B IT (57.0), and Qwen3 235B A22B Inst. (54.3). **WellLog-LLM-Qwen-27B reaches 63.1 Overall** with the DFS pipeline—within 1.1 points of Kimi-K2.6 at roughly $40\times$ fewer parameters and $\sim$73% fewer input tokens. **WellLog-LLM-K2V2** shows the same monotone progression ($51.6 \rightarrow 54.4 \rightarrow 56.0$), confirming the recipe transfers across backbones.

7.2 Stage attribution

Rather than a generic stack of tuning tricks, the ablations show that our recipe drives targeted improvements across the pipeline (full per-rubric and per-cell breakdown in Appendix K.1):

(i) SFT builds the petrophysical interface, but only when the base needs it. On the K2-V2 Instruct backbone, SFT is the largest jump (+6.9 Overall), driven by Trust (+11.7), RA (+6.2), and Style (+6.2): the model learns the vocabulary, answer format, and log-grounding conventions needed to read DFS contexts. The cell ablation shows this is broad, with every covered K2V2 basin–reasoning cell gaining at least +9 Overall and Geolink Single-Formation gaining +13.4. On the stronger Qwen3.5-27B base, however, pure imitation degrades performance by −9.0 (teacher-student distribution mismatch (Huang et al., 2026; Li et al., 2024; Dong et al., 2024)), so the deployable Qwen chain uses the SFT checkpoint only as the DPO reference policy.

(ii) DPO adds contrastive grounding. The preference pairs are not a style-only signal: matched diagnostics show Kimi-K2.6 chosen responses exceed Qwen3-8B rejected responses by roughly 20–22 Overall points, with especially large gaps in faithfulness, hallucination control, log utilization, and reference-supported petrophysical claims. After DPO, the gains concentrate in the same ground-truth-aligned dimensions: on K2-V2 Instruct, Accuracy, Coverage, Conclusion-by-Reference, and Petrophysical-Claims-by-Reference each rise by about +3 points; on Qwen3.5-27B, the same dimensions rise by about +1.2. Thus, DPO is the evidence-alignment stage before GRCO adds reusable inference-time experience.

(iii) GRCO adds reusable inference-time experience. Its gains are rubric-specific: RA-led on K2-V2 Instruct (+1.8) and RQ-led on Qwen3.5-27B (+3.7), with the largest Qwen gains landing on harder cells such as COSTA Multi-Well (+4.7) and Danish Cross-Formation (+4.4).

Backbone	Stage	RA	Trust	Style	RQ	Ovr.	ΔOvr
K2-V2 Instruct	base	38.5	38.0	55.5	53.1	44.7	—
	+ SFT	44.7	49.6	61.6	57.1	51.6	+6.9
	+ DPO	47.6	52.3	63.6	60.7	54.4	+2.8
	+ GRCO	49.5	54.0	65.2	61.8	56.0	+1.6
Qwen3.5-27B	base	52.0	60.0	67.9	68.5	60.1	—
	+ SFT†	47.5	46.9	51.9	61.8	51.1	−9.0
	+ DPO	53.2	60.2	68.0	68.4	60.6	+0.6
	+ DPO + GRCO	55.8	62.4	69.3	72.1	63.1	+2.5

Table 4: Stage attribution: each stage applied incrementally on top of the previous checkpoint. †From a prior evaluation pass; see Table 17 for the paired diagnostic.

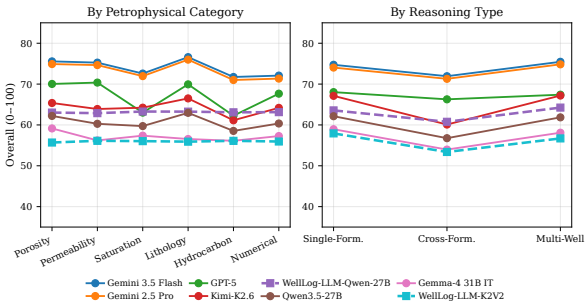
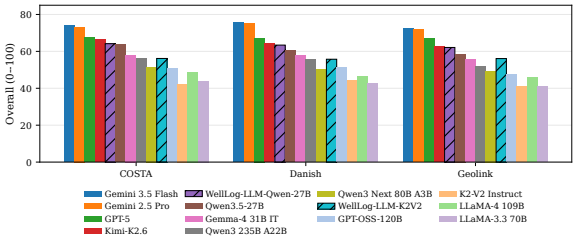


Figure 5: Top-tier model performance across the six petrophysical categories (left) and three reasoning types (right), public split. Baselines use full context and WellLog-LLM variants (dashed) use DFS, matching Table 3. Category and reasoning averages use the same corrected basin weighting as the headline table.

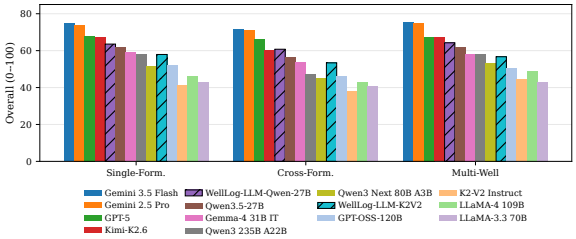
The learned buffer contains general petrophysical rules—depth-unit checks, MHI/RXO validity, pay-flag priority, crossover/SW/RDEP reconciliation—that are inserted into later prompts (examples in Appendix I.4). The result is better-evidenced reasoning: the severe Faithfulness complaint *"claims outrun the available support"* drops from 80% to 14% for K2-V2 Instruct, and Qwen3.5-27B gains a new majority class, *"Most claims stay within the benchmark frame"* (35%). Appendix diagnostics show the same shift at item level: a representative stage-by-stage progression is shown in Figure 13, additional answer behaviours are shown in Figure 15, WellLog-LLM variants move into the top-right grounding cluster (Figure 19), and WellLog-LLM-Qwen-27B removes Qwen3.5-27B's severe below-50 tail ($106/625 \rightarrow 0/624$; Figure 24).

7.3 Performance across basins, categories, and reasoning types

Figures 5–6 show tight category bands while exposing the main sliced failures. Geolink's mixed clastic–carbonate setting penalizes grounding (Qwen3.5-27B Trust 64.0 COSTA → 57.9 Geolink; Qwen3 Next 80B 46.7 Danish → 44.2 Geolink), while WellLog-LLM-Qwen-27B is more basin-robust, with a 2.0-point Overall spread (64.2 COSTA, 63.4 Danish, 62.1 Geolink) vs. 5.0 for



(a) Per-basin Overall. COSTA/Danish are easiest, Geolink (mixed clastic–carbonate) hardest; rankings are stable across basins.



(b) Per-reasoning-type Overall. Cross-Formation is hardest for mid-tier models; top systems are reasoning-type invariant.

Figure 6: Sliced performance on the public split; hatched bars are WellLog-LLM variants.

its base. Cross-Formation is hardest: Qwen3.5-27B Trust drops 9.9 points from Single to Cross, whereas Gemini 2.5 Pro drops 5.4. Unlike most baselines, WellLog-LLM-Qwen-27B peaks on Multi-Well (64.2), suggesting the experience prior helps multi-formation synthesis. Appendix J expands the per-rubric breakdown.

7.4 Multimodal and Visual Analysis

As a complementary probe, we evaluated rendered well-log plots paired with the structured context and plot-only inputs. Context+plot remains close to the headline text setting (Gemini 2.5 Pro 73.0 vs. 73.4), whereas plot-only is much weaker (38.5), indicating that current VLMs do not recover the structured petrophysical evidence already exposed by Petrophys (V_{sh}, ϕ, S_w, pay flags, . . .). We therefore use structured context as the headline setting; modality definitions and VLM failure analysis are in Appendices L–M.

8 Conclusion

We present WellLogBench, a benchmark for LLM-based well-log reasoning with 1,085 expert-validated QA pairs, a Petrophys context pipeline, a structured evaluator, and WellLog-LLM. Our experimental results demonstrate the effectiveness of the proposed approach. We hope WellLogBench and our opensource model will catalyzes rigorous, grounded petrophysical reasoning.

Limitations

WellLogBench is built from 62 wells across three basins, so it does not cover the full range of geology, tool suites, and logging conditions seen in practice. 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 interpretation.

9 Ethical Considerations

Data & privacy. We use only public datasets (Geolink, COSTA, GEUS) 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 curation and validation were fairly compensated.

10 Release and Reproducibility

We will release, under permissive open-data/open-source licenses, the public benchmark split (651 expert-validated QA pairs with context, ground truth, rubric annotations, and metadata), source-data snapshots/URLs, the Petrophys context-engineering pipeline (§3.5; petrophys.netlify.app), LLM-as-Judge prompts and scoring protocol, and WellLog-LLM-Qwen-27B/WellLog-LLM-K2V2 weights with SFT/DPO/GRCO training details (Appendix I). We will also host the 434-question private split through an online leaderboard/evaluation server so the community can track progress without exposing held-out answers.

References

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

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. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*.

Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, et al. 2025. Training-free group relative policy optimization. *arXiv preprint*.

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

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

Jorge Costa Gomes, Sebastian Geiger, and Daniel Arnold. 2022. Costa model: Hierarchical carbonate reservoir benchmarking case study. Dataset, Heriot-Watt Research Portal.

Hai Dang, Ben Lafreniere, Tovi Grossman, Kashyap Todi, and Michelle Li. 2025. Authoring llm-based assistance for real-world contexts and tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 211–230, New York, NY, USA. Association for Computing Machinery.

Google DeepMind. 2025. Gemma-4 31b instruct (model card). Hugging Face.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Le Zhou, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2023. K2: A foundation language model for geoscience knowledge understanding and utilization. https://arxiv.org/abs/2306.05064. *Preprint*, arXiv:2306.05064.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. *Preprint*, arXiv:2310.05492.

J. Eckroth, J. Boden, Lannie D. Hough, H. Gatewood, S. Gipson, E. Schoen, and B. Gunderson. 2025. Building "energyllm": A domain-specific large language model trained on SPE content. In *SPE Annual Technical Conference and Exhibition*. Paper Number: SPE-228097-MS.

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. 2024. Processed datasets – imputation in well log data: A benchmark. Zenodo. Version v1.

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. 2025. The delta learning hypothesis: Preference

tuning on weak data can yield strong gains. *arXiv preprint*.

Geological Survey of Denmark and Greenland (GEUS). Subsurface data denmark. Web portal.

Abdenour Hadid, Tanujit Chakraborty, and Daniel Busby. 2024. When geoscience meets generative AI and large language models: Foundations, trends, and future challenges. *Expert Systems*, 41(10):e13654.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? https://arxiv.org/abs/2404.06654. *Preprint*, arXiv:2404.06654.

Zixian Huang, Kaichen Yang, Xu Huang, Feiyang Hao, Qiming Ge, Bowen Li, He Du, Kai Chen, and Qipeng Guo. 2026. How to fine-tune a reasoning model? a teacher-student cooperation framework to synthesize student-consistent SFT data. *Preprint*, arXiv:2604.14164.

Shiyi Jiang, Panke Sun, Fengqing Lyu, Sicheng Zhu, Ruifeng Zhou, Bin Li, Taihong He, Yujian Lin, Yining Gao, Wendan Song, and Huaimin Xu. 2024. 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.

K2 Team. 2025. K2-V2: A 360-open, reasoning-enhanced LLM. *arXiv preprint arXiv:2512.06201*.

Koray Kavukcuoglu, Jeff Dean, Oriol Vinyals, and Noam Shazeer. 2026. Gemini 3.5: Frontier intelligence with action. Google Blog.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. 2025. Time-MQA: Time series multi-task question answering with context enhancement. https://arxiv.org/abs/2503.01875. *Preprint*, arXiv:2503.01875.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308. Association for Computational Linguistics.

Pin Li, Yubo Liu, Jun Zhou, Juan Zhang, and Yuange Zhou. 2025. A report generation system for well logging industry based on large language model and retrieval augmented generation. In *International Petroleum Technology Conference*, page D012S004R002. IPTC.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, et al. 2024. Geogalactica: A scientific large language model in geoscience. https://arxiv.org/abs/2401.00434. *Preprint*, arXiv:2401.00434.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Meta. 2024a. Llama-3.1-8b-instruct (model card). Hugging Face.

Meta. 2024b. Llama-3.3-70b-instruct (model card). Hugging Face.

Meta. 2025. Llama 4 model card. GitHub (meta-llama/llama-models).

Abbas Mohammadi, SanLinn I. Kaka, A. Koeshidayatullah, et al. 2025. Transforming petrophysical well-logs to images: Leveraging deep learning for lithology recognition. *Earth Science Informatics*, 18(3):457.

Moonshot AI. 2025. Kimi-k2.6 (model card). Moonshot AI.

OpenAI. 2025a. Gpt-5 system card. System card. PDF: https://cdn.openai.com/gpt-5-system-card.pdf.

OpenAI. 2025b. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*. Companion page: https://openai.com/index/gpt-oss-model-card/.

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

Zhenyu Qi, Qing Yu, Jichen Wang, Yun-Bo Zhao, Zerui Li, and Wenjun Lv. 2025. WLFM: A well-logs foundation model for multi-task and cross-well geological interpretation. *arXiv preprint*. ArXiv September 2025.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv preprint arXiv:2503.04691*.

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

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

Yu Shi, Junqiao Liao, Lu Gan, and Rongjiang Tang. 2024. Lithofacies prediction from well log data based on deep learning: A case study from southern sichuan, china. *Applied Sciences*, 14(18):8195.

Sumit Srivastav and Sagar Kothe. 2025. Catalysing new value streams in oil and gas with agentic AI. PwC India report.

Qwen Team. 2025. Qwen3-next-80b-a3b-instruct (model card). Hugging Face.

TGS. 2025. Well log foundation model: Making promptable AI models for well log interpretation. Technical report, TGS.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Chicheng Xu and Iulian N. Hulea. 2024. Bridging data and algorithms in the petrophysics domain: Introducing the open petrophysics data & utilities (opdu) platform. SPWLA Houston Chapter Lunch Seminar (Northside Houston), abstract.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Chen Zhangxing, Yongan Zhang, LI Jian, HUI Gang, SUN Youzhuang, LI Yizheng, CHEN Yuntian, and Dongxiao ZHANG. 2025. Artificial intelligence large model for logging curve reconstruction. *Petroleum Exploration and Development*, 52(3):842–854.

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

Table 5: Summary of geological formations and lithologies contained within WellLogBench.

Appendix Roadmap

The supplement is organized by purpose so readers can move from dataset provenance to evaluation robustness, model/training details, diagnostic analyses, multimodal extensions, examples, and prompts.

- **Benchmark construction** (Appendices A–C): basin geology, expert annotation workflow, and QA audit transparency.
- **Evaluation methodology** (Appendices D–F): expert–judge alignment, cross-judge robustness, and scoring-weight sensitivity.
- **Models and training** (Appendices G–I): model roster, DFS sensitivity, and SFT/DPO/GRCO recipe details.
- **Results analyses** (Appendices J–K): fine-grained basin/category/reasoning trends and the ablation suite.
- **Multimodal extensions** (Appendices L–M): well-log modality reference and visual/multimodal failure analysis.
- **Reference, schema, examples, and prompts** (Appendices N–Q): private-test protocol, example Petrophys context, complete answer, and prompt templates.

A Detailed Basin Information

To ensure **WellLogBench** evaluates LLM reasoning capabilities across diverse geological environments, we aggregated and curated data from three distinct sources: the Geolink (North Sea) dataset, the COSTA (Carbonate) dataset, and the Subsurface Data Denmark (GEUS) repository. This combination ensures the benchmark covers clastic, carbonate, and evaporite depositional environments, each presenting unique challenges for petrophysical interpretation.

A.1 Data Sources and Stratigraphy

Geolink Dataset (North Sea) Sourced from the Geolink repository (Gama et al., 2024), this subset primarily represents the Norwegian Continental Shelf, including the Volve field. It represents the most heterogeneous subset in our benchmark, characterized by complex intercalation of clastics and evaporites.

- **Stratigraphic Coverage:** The dataset spans major stratigraphic Groups including the *Boknfjord, Cromer Knoll, Dunlin, Hordaland, Nordland, Rogaland, Rotliegendes, Shetland, Tyne, Vestland, Viking,* and *Zechstein* Groups.
- **Lithology:** This subset introduces high-variance lithological distinctness, containing extensive lists of *Sandstone, Shale, Limestone, Dolomite, Anhydrite, Halite (Salt), Gypsum, Coal, Conglomerate, Chert,* and intrusive/extrusive igneous rocks.

COSTA Model (Carbonate Reservoir) The COSTA subset (Costa Gomes et al., 2022) focuses on hierarchical carbonate reservoir benchmarking. Unlike the North Sea datasets, this subset is focused on a specific, high-heterogeneity carbonate environment.

- **Stratigraphic Coverage:** Data is concentrated on the *Kharaib-2* formation. This formation is geologically significant as it serves as a distinct analog for major carbonate reservoirs in the Middle East.
- **Lithology:** The lithology is strictly distinct from clastic environments, composed primarily of *Limestone* and *Dolomite*. This tests the LLM's ability to reason about porosity and fluid saturation in complex pore systems distinct from sandstone reservoirs.

Subsurface Data Denmark (GEUS) Derived from the Geological Survey of Denmark and Greenland (Geological Survey of Denmark and Greenland (GEUS)), this subset covers the Danish sector of the North Sea. To ensure benchmark reliability, we filtered this source for data quality, sampling only stable formations with 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:** This dataset is heavily weighted towards chalks and marls, including *Chalk, Claystone, Marl, Salt, Basalt, Siltstone,* and *Sandstone*.

A.2 Lithological Diversity Summary

Table 5 summarizes the lithological classes present in WellLogBench. The presence of overlapping lithologies (e.g., Zechstein salts in both Geolink and GEUS, versus pure carbonates in COSTA) requires the model to generalize physical rock properties across different basinal contexts.

B Annotation Workflow and Expert Panel

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. We also provided the raw LAS files to allow experts to cross-check values in formation metadata.

For each QA item, experts completed a structured rubric covering multiple aspects of answer quality. The rubric emphasized (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.

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 ∪ expert)	81.4%	86.4%	85.2%

Table 6: Proportion of QA pairs touched by automated auditing and by human expert review. The high union rates show the final benchmark is substantially shaped by independent review.

This annotation process served two goals: (1) producing high-quality ground truth labels for benchmarking, and (2) surfacing systematic failure modes (e.g., missing-data assumptions, unit inconsistencies, or overconfident fluid interpretations) that informed iterative improvements to our data-generation pipeline.

C 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) *revalidation*; 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 6 reports the intervention rates by reasoning level.

Why Gemini 2.5 Pro as drafter. Gemini 2.5 Pro was selected for drafting because of its reported 1M-token context window and strong long-context retrieval/reasoning, suitable for items grounded in very long source contexts. To mitigate generator-style bias, both auditing (GPT-5) and expert correction were intentionally decoupled from the generator; the high revision/correction rates above show this decoupling is effective in practice.

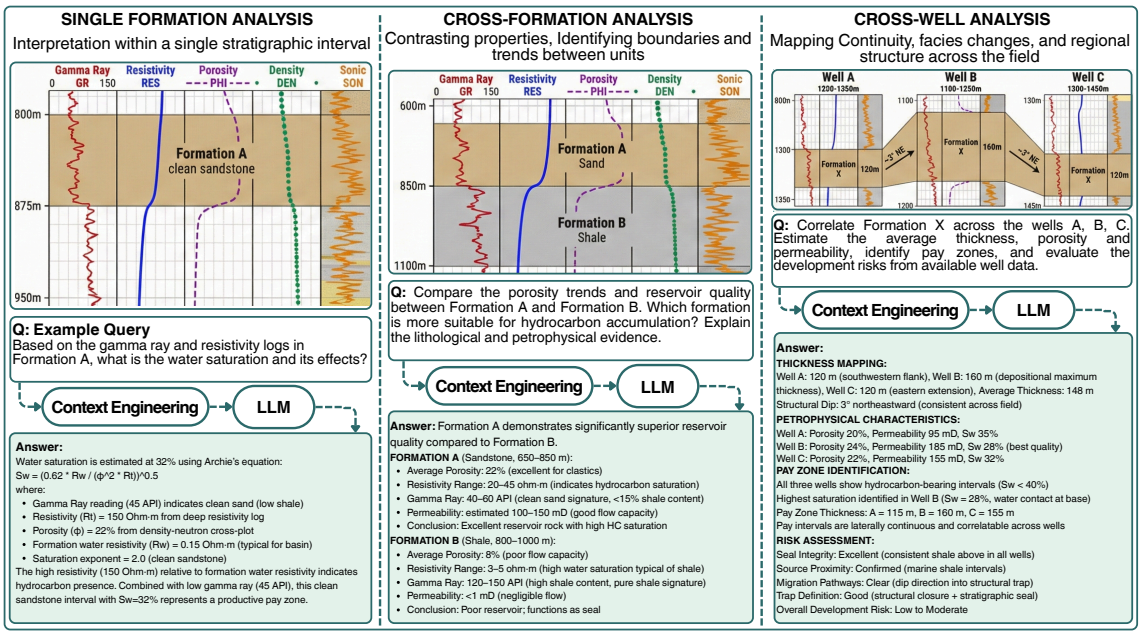


Figure 7: Hierarchical Reasoning Complexity. WellLogBench evaluates models across three scales of analysis: localized Single-Formation evaluation, vertical Cross-Formation comparison within a wellbore, and regional Multi-Well correlation across fields.

D 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).

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, $\mathrm{ICC}(A,1)$. Across 50 QA pairs, two answers, and 13 shared sub-metrics, expert and LLM-as-judge ratings achieve $\mathrm{ICC}(A,1)$=0.721, indicating good agreement. Agreement is slightly higher for Answer 1/Gemini 2.5 Pro ($\mathrm{ICC}(A,1)$=0.697) than for Answer 2/GPT-5 ($\mathrm{ICC}(A,1)$=0.683).

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.

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 8 reports per-metric correlations.

Rubric	W.	Exp. A1	Exp. A2	LLM A1	LLM A2
Resp. 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
Petrophys. Reason.	20%	81.7	76.8	82.2	77.6
Wgt. Overall	—	**76.2**	**64.7**	**75.8**	**65.8**

Table 7: Expert vs. LLM-as-judge rubric scores. Answer 1 = Gemini 2.5 Pro; Answer 2 = GPT-5. Expert $\Delta_{\text{A1-A2}}$=+11.6; LLM $\Delta_{\text{A1-A2}}$=+9.9.

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

Table 8: Pearson correlations between expert and LLM-as-judge per-metric scores, separately for A1 (Gemini 2.5 Pro) and A2 (GPT-5).

Discussion. The ICC results indicate good absolute agreement between expert and LLM-as-judge ratings, supporting the use of the automated judge for aggregate model comparison. Pearson correla-

tions show particularly strong alignment on technical, evidence-grounded 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 for more subjective style-oriented criteria and some GPT-5 petrophysical reasoning metrics. We therefore treat the LLM-as-judge as reliable for aggregate trends and technical scoring, while continuing to report sub-metrics separately so readers can inspect or re-aggregate dimensions with different levels of expert–judge alignment.

E Cross-Judge Robustness

Setup. We re-scored the shared baseline cohort, replacing GPT-5 in our dual-judge protocol with three different judges: Qwen3 235B A22B Instruct, Gemma-4 31B IT, and Qwen3.5-27B. The second judge remains the same (Qwen3.5-27B) in these ablations. Table 9 reports per-rubric Pearson and Spearman agreement with GPT-5; Table 10 reports the resulting model rankings.

Discussion. Conclusions are invariant to the choice of judge: the top-3 placements never change and overall rank correlation is $\rho \geq 0.93$. The only systematically weaker rubric is Style, consistent with the limitation discussed in Appendix D.

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

Table 9: Cross-judge agreement vs. GPT-5 by rubric (shared baseline cohort).

Model	GPT-5 (Reference)	Qwen3-235B	Gemma-4-31B-IT	Qwen3.5-27B
Gemini 3.5 Flash	1	1 (0)	1 (0)	1 (0)
Gemini 2.5 Pro	2	2 (0)	2 (0)	2 (0)
GPT-5	3	3 (0)	3 (0)	3 (0)
Kimi-K2.6	4	5 (+1)	4 (0)	4 (0)
Qwen3.5-27B	5	4 (-1)	6 (+1)	5 (0)
Gemma-4-31B-IT	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 109B	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)

Table 10: Model ranking by judge over the 11 baseline models present in the final benchmark cohort. Top-3 placements are identical across all judges; overall Spearman $\rho \geq 0.93$ across judges; max rank shift is 2 positions.

Rubric	Mean (all models)
Response Adequacy	47.7
Trustworthiness	52.1
Style	63.7
Petrophys. Reasoning	63.8

Table 11: Mean rubric levels across the benchmarked models. Style and Reasoning are systematically higher than Response Adequacy; uniform weighting therefore upweights the dimensions on which most models already score well.

F Weighting Sensitivity Analysis

Default vs. equal weighting. The default weighting (40/20/20/20) and a uniform (25% each) alternative yield near-identical model orderings: Pearson r=0.995, Spearman ρ=0.994. Where they differ is informative.

Why we keep weighted scoring. The benchmark's intended use is to rank systems by correct, complete petrophysical interpretation. Equal weighting rewards "well-written but wrong" outputs in this regime; the default weighting (40% RA, 20% each) better matches this intent while still allocating 60% to grounding, communication, and reasoning quality. We report all sub-metrics, so readers may re-aggregate under any preferred scheme.

Model	Resp. Adeq.	Trust	Style	Reason. Quality	Weighted Final	Equal-weight Final	Δ (Eq. – Wgt.)
Weaker models: equal weighting inflates score							
LLaMA-3.3 70B	29.8	40.9	61.7	49.3	42.3	45.4	+3.1
LLaMA-4 109B	36.4	43.9	64.5	52.3	46.7	49.3	+2.6
LLaMA-3.1 8B Inst.	22.0	29.3	48.4	37.3	31.8	34.2	+2.5
Qwen3-8B	33.5	38.5	54.5	51.0	42.2	44.4	+2.2
Top-tier models: ranking essentially unchanged							
Gemini 3.5 Flash	69.5	70.3	79.1	81.7	74.0	75.1	+1.1
Gemini 2.5 Pro	68.9	69.8	78.1	81.1	73.4	74.5	+1.1
Kimi-K2.6	58.9	61.3	67.9	74.1	64.2	65.6	+1.3
GPT-5	62.5	63.8	69.4	77.9	67.2	68.4	+1.2

Table 12: Effect of equal vs. weighted scoring. Weaker models with low Response Adequacy but moderate-to-high Style/Reasoning gain disproportionately under uniform weights; top-tier rankings are essentially unchanged.

Model	Size/Variant	Txt	MM	Vis
Closed-source				
Gemini 3.5 Flash	API (closed)	✓	–	–
Gemini 2.5 Pro	API (closed)	✓	✓	✓
GPT-5	API (closed)	✓	–	–
Open-source baselines				
Kimi-K2.6	1.1T (MoE)	✓	✓	✓†
Qwen3.5-27B	27B	✓	✓	✓
Gemma-4 31B IT	31B	✓	✓	✓
Qwen3 235B A22B Inst.	235B (MoE)	✓	✓	✓
Qwen3 Next 80B A3B	80B (MoE)	✓	–	–
GPT-OSS-120B	117B (MoE)	✓	–	–
K2-V2 Instruct	70B	✓	–	–
LLaMA-4 109B	109B (MoE)	✓	✓	✓
Mistral Small 4 119B	119B	–	–	✓
LLaMA-3.3 70B Inst.	70B	✓	–	–
Qwen3-8B	8B	✓	–	–
LLaMA-3.1 8B Inst.	8B	✓	–	–
WellLog-LLM and stage ablations (§6)				
WellLog-LLM-Qwen-27B	27B	✓	–	–
Qwen3.5-27B - SFT	27B	✓	–	–
Qwen3.5-27B + DPO	27B	✓	–	–
WellLog-LLM-K2V2	70B	✓	–	–
K2-V2 Instruct SFT + DPO	70B	✓	–	–
K2-V2 Instruct SFT	70B	✓	–	–

Table 13: Model roster by experimental setting. Txt = text-only; MM = multimodal (text + visual); Vis = visual-only. †Kimi-K2.6 visual-only is carried over from a prior evaluation and is marked separately in Appendix M.

G Model Roster

Table 13 lists every model evaluated, its scale, and the experimental settings in which it was used. The primary multimodal models also appear in the text-only experiments, enabling direct cross-modality comparison; visual-only evaluation additionally includes a Mistral probe.

H DFS Delta Analysis

We report $\Delta = (\text{With DFS}) - (\text{Without DFS})$ on the 0–100 scale, averaged across basins and reasoning levels, for every model with both configurations in the current evaluation cohort.

Model	ΔRA	ΔTrust.	ΔStyle	ΔRQ	ΔOvr.
Gemini 2.5 Pro	−7.1	−5.8	−1.5	−3.4	−5.0
Gemini 3.5 Flash	−6.7	−5.6	−1.3	−3.2	−4.7
Qwen3 Next 80B A3B	−6.3	−3.6	+0.9	−1.7	−3.4
GPT-5	−3.3	−4.8	−6.5	+0.6	−3.5
LLaMA-3.1 8B Inst.	−1.8	−2.4	−4.7	−1.8	−2.5
LLaMA-4 109B	−3.4	−5.5	+0.9	−2.2	−2.7
GPT-OSS-120B	−5.8	−3.8	+0.1	+1.2	−2.8
Kimi-K2.6	−4.9	−3.3	+1.2	−7.2	−3.8
Qwen3.5-27B	−0.3	−0.3	+1.3	−3.1	−0.5
LLaMA-3.3 70B	−0.7	−2.6	+2.9	−1.9	−0.6
Gemma-4 31B IT	+0.4	+1.6	−0.1	+0.8	+0.6
Qwen3 235B A22B Inst.	+1.8	+5.0	+1.6	−4.5	+1.1
Qwen3-8B	+1.3	−1.1	−0.1	−0.5	+0.2
WellLog-LLM-Qwen-27B	**−5.7**	**−8.3**	**0.0**	**+1.6**	**−3.6**
K2-V2 Instruct	+3.8	+4.4	−0.3	−0.5	+2.2

Table 14: Rubric deltas under DFS ($\Delta = \text{DFS} - \text{Full}$), computed with the same corrected basin weighting as Table 3 for models with both configurations. Negative deltas dominate, especially for the strongest models—DFS modestly trades correctness for context efficiency. Our headline table uses full structured context for baselines and the DFS pipeline for WellLog-LLM variants; this table is retained as a sensitivity study.

Interpretation. On the current model set, DFS *slightly hurts on average* ($\bar{\Delta} \approx -1.9$ Overall). The largest costs are on the strongest full-context systems: Gemini 2.5 Pro (-5.0), Gemini 3.5 Flash (-4.7), Kimi-K2.6 (-3.8), GPT-5 (-3.5), and Qwen3 Next 80B (-3.4). A few models gain from attention focusing, especially K2-V2 Instruct ($+2.2$), Qwen3 235B A22B Inst. ($+1.1$), and Gemma-4 31B IT ($+0.6$). The largest rubric-level penalty is on Response Adequacy and Trustworthiness for top systems, consistent with statistical-anchor loss. DFS remains the right default when token/cost budgets are binding: it trades a small average score cost for roughly 73% fewer input tokens, and WellLog-LLM is explicitly trained/evaluated under this compressed-context regime.

I Petrophysics-Tuned Model Training Details

This appendix expands Section 6. WellLog-LLM is instantiated as **WellLog-LLM-Qwen-27B** (Qwen3.5-27B backbone) and **WellLog-LLM-**

K2V2 (K2-V2 Instruct backbone). All three stages operate on contexts produced by our Dynamic Feature Selection pipeline (§5.1) with a maximum context length of **65,536 tokens**. SFT targets are drafted by Kimi-K2.6 (Moonshot AI, 2025); DPO uses Kimi-K2.6 responses as chosen samples and Qwen3-8B responses as rejected samples; GRCO uses the current DPO checkpoint (K2-V2 Instruct or Qwen3.5-27B) as the policy model and Qwen3.5-27B as the rubric-aligned judge/meta model to optimize an external buffer of instructions and experiences offline.

I.1 Stage 1: Supervised Fine-Tuning (SFT)

We construct ∼30,000 (context, question, answer) triples by sampling WellLogBench-style items from a pool drawn from the training basins, deriving DFS contexts as in the benchmark pipeline, and using **Kimi-K2.6** to draft answers. The split is balanced across reasoning level:

- 10,000 *Single-Formation* items,
- 10,000 *Cross-Formation* items,
- 10,000 *Multi-Well* items,

with the per-category proportion mirroring the benchmark distribution (15–20% per category). Table 15 shows the resulting context-token distribution per petrophysical category: mean lengths cluster around 21–24k tokens with the 99th percentile bounded by the 65k context limit, validating the choice of DFS-summarized contexts as an SFT format that fits comfortably under most open-source backbones' attention windows.

Category	Mean tokens (context + thinking + answer)	p50	p90	p95	p99	Max
hydrocarbon	23,418	17,642	45,387	57,914	64,836	104,731
lithology	21,367	15,928	41,706	53,482	61,942	91,615
permeability	23,846	18,731	47,963	62,417	64,684	101,293
porosity	20,924	15,441	39,628	51,736	60,487	88,904
numerical	22,671	16,884	43,972	56,318	63,731	94,276
saturation	22,983	17,319	44,816	58,642	64,284	97,835

Table 15: SFT data context-token statistics per petrophysical category. Mean lengths sit at ∼21–24k tokens, p99 stays under the 65k DFS cap, and max length reaches up to ∼105k for the longest hydrocarbon items (further truncated during training)

Long-context SFT configuration. Both SFT backbones use the same long-context recipe; only the base checkpoint changes (K2-V2 Instruct checkpoint `6ef3e31` for WellLog-LLM-K2V2 and Qwen3.5-27B for WellLog-LLM-Qwen-27B). Each run uses a 10k-sample shard from the SFT pool described above, assistant-only loss, no packing, and keep-end truncation at 65,536 tokens. Table 16 summarizes the concrete optimization and infrastructure settings.

Group	Setting	Value
Data/model	Base models	K2-V2 Instruct; Qwen3.5-27B
Data/model	Samples	10k-sample SFT shard
Data/model	Context	65,536 tokens; keep-end truncation
Data/model	Loss / packing	Assistant-only; packing disabled
Optimization	Epochs / batch	10; per-device batch 8; grad. accum. 1
Optimization	LR schedule	1×10^{-4} cosine; warmup 0.03
Optimization	Regularization	Weight decay 0.0; BF16; gradient checkpointing
LoRA	Adapter	r=16, α=32, dropout 0.05
LoRA	Target modules	`q/k/v/o_proj`; `gate/up/down_proj`
Infrastructure	GPUs	16 nodes × 8 GPUs (128 total)
Infrastructure	Parallelism	SP size 16; variable length; FlashAttention-2
Infrastructure	Training engine	DeepSpeed ZeRO-3 + CPU offload; Liger kernel
Infrastructure	Checkpoints	Every epoch; keep last 10

Table 16: Long-context SFT configuration used for the K2-V2 Instruct and Qwen3.5-27B SFT runs. The SFT shard is drawn from the larger synthetic SFT pool; the backbone checkpoint changes across variants.

SFT-only negative control (prior evaluation). In an earlier evaluation pass we also trained **Qwen3.5-27B - SFT**, a checkpoint fine-tuned directly on Kimi-K2.6 drafted answers before preference optimization. That checkpoint degraded across every rubric group (Table 17, −9.0 Overall), with the largest losses on Trustworthiness and Style. We interpret this as a teacher–student distribution mismatch rather than evidence against domain supervision itself: recent teacher–student SFT work shows that stronger-teacher synthetic data can reduce reasoning performance when teacher style diverges from the student distribution (Huang et al., 2026), and broader LLM tuning studies document forgetting and performance conflicts from SFT data composition (Li et al., 2024; Dong et al., 2024). This motivates using Kimi data as an initialization signal, then applying DPO and GRCO to recover grounded behavior. We did not re-evaluate the SFT-only checkpoint under the present public split.

Model	RA	Trust.	Style	RQ	Ovr.	Ovr.$_{-S}$
Qwen3.5-27B	52.0	60.0	67.9	68.5	60.1	58.1
Qwen3.5-27B - SFT†	47.5	46.9	51.9	61.8	51.1	50.9
Δ (SFT – base)	−4.5	−13.1	−16.0	−6.7	−9.0	−7.2

Table 17: Paired diagnostic for SFT-only Qwen3.5-27B tuning on Kimi-K2.6 answers. †The SFT-only row is from a prior evaluation pass (not re-evaluated under the current public split); the base row uses the same evaluation setting for fair comparison. **Ovr.** is the headline weighted metric and **Ovr.**$_{-S}$ is a diagnostic variant excluding Style.

I.2 Stage 2: Direct Preference Optimization (DPO)

We build ∼15,000 preference pairs—5,000 per reasoning level (Single, Cross, Multi)—using Kimi-K2.6 responses as the *chosen* samples and Qwen3-8B responses as the *rejected* samples. The DPO signal follows the chosen–rejected delta hypothesis: the policy is optimized to increase the relative likelihood gap between preferred and rejected answers against the SFT reference policy. Pairs are drawn from items *disjoint* from the SFT corpus to avoid overfitting to the supervisor's surface style on already-seen contexts. DPO is implemented as long-context Megatron-RLHF LoRA fine-tuning with the SFT checkpoint as the reference policy; the configuration is summarized in Table 18.

Group	Setting	Value
Runs	Backbones	Qwen3.5-27B; K2-V2 SFT checkpoint
Runs	Context / data	65,536 tokens; same DPO pair set
Runs	Backbone-specific setup	Qwen default template; K2-specific chat template
Distributed	GPUs	4 nodes $\times$ 8 GPUs (32 total)
Distributed	Parallelism	Tensor parallel 8; sequence parallel enabled
Objective	RLHF / loss	DPO; sigmoid loss; β=0.1; RPO α=0.1
Optimization	Epochs / LR	3; LR 1×10^{-5}; min LR 1×10^{-6}
Optimization	Batch / warmup	Micro batch 1; global batch 32; warmup 0.05
LoRA	Adapter	r=8, α=32; target modules = all-linear
Memory	Long-context stack	FlashAttention; full activation recomputation
I/O	Evaluation / saving	Eval every 100 steps; save every 200 (Qwen) / 100 (K2)

Table 18: DPO training configuration for WellLog-LLM-Qwen-27B and WellLog-LLM-K2V2. Both runs use the same preference data, DPO objective, 65k context length, and 32-GPU distributed LoRA setup; the main differences are the base checkpoint, chat-template configuration, and checkpoint cadence.

I.3 Stage 3: Group Relative Context Optimization (GRCO)

We use Group Relative Context Optimization (GRCO), adapted from training-free GRPO (Cai et al., 2025), as an offline stage for optimizing a natural-language buffer of instructions and experiences. It does not update model weights or rewrite the full prompt online. For each prompt, the policy model—the current DPO checkpoint, either K2-V2 Instruct or Qwen3.5-27B—proposes a group of K rollouts, and a Qwen3.5-27B rubric-aligned judge scores each response. Mixed-quality groups are then used for a three-step experience update: rollout summaries, per-query critique over correct vs. incorrect attempts, and a batch update that consolidates add/modify operations into the next experience library. The resulting library is injected as accumulated experience during later policy calls.

- **Data:** 1,000 items per petrophysical category $\times$ 6 categories $\times$ 3 reasoning levels = 18,000 prompts. All prompts are drawn from a held-out training pool disjoint from the public benchmark.
- **Policy model:** the served DPO checkpoint for the target backbone (K2-V2 Instruct or Qwen3.5-27B) generates the rollout group.
- **Judge & meta model:** Qwen3.5-27B is used for rollout scoring, rollout summarization, critique generation, and batch experience updates in the offline GRCO loop, isolating GRCO from any closed-source dependency at inference time.
- **Context length:** the same 65k DFS contexts as SFT/DPO.
- **Infrastructure:** no gradient update is performed during GRCO.

I.4 Representative GRCO experiences

The experience buffer stores general rules distilled from judge-scored rollout groups, not item-specific answers. At inference, the relevant subset is retrieved and prepended to the prompt as guidance. Table 19 paraphrases representative buffer entries.

I.5 Recipe summary

Why this recipe works. The three stages target complementary failure modes documented in the evaluation rubric (Section 4) and sliced results (Appendix J): SFT closes the petrophysical-vocabulary gap, DPO sharpens evidence-aligned preferences for accurate, complete, reference-supported claims, and GRCO adds reusable experience rules for grounded numerical reasoning. The cumulative effect is most visible on Cross-Formation questions, where the tuned Qwen3.5-27B variant adds +5.6 RA and +6.7 Trust over its base (Appendix J).

Experience family	Representative learned rule	General effect at inference
Depth/unit fidelity	Prefer depth units in interval labels (ft/m) over global metadata when they conflict, and preserve stratigraphic order before comparing intervals, contacts, or pay zones.	Prevents spatial misalignment in formation comparisons, fluid-contact placement, and pay-zone ranking.
Derived-log validity	Before using MHI, SW, FZI, or CROSSOVER_FLAG, verify that required logs and substitutions were valid. If RXO is missing and SWXO defaults to 1.0, treat MHI as $1 - \mathrm{SW}$ rather than true invasion.	Stops the model from treating placeholder derived values as physical measurements.
Fluid typing hierarchy	Combine neutron-density crossover with RDEP and SW; use VP/VS only when DTS is present or quality-controlled, and treat missing strong-gas flags as non-detection rather than proof of no gas.	Reduces unsupported oil/gas calls when one diagnostic channel is missing or unreliable.
Crossover definition checks	Reconcile CROSSOVER_FLAG with the benchmark definition and interval statistics before concluding gas; explicitly check whether $\mathrm{PHID} - \mathrm{PHIN}$ or $\mathrm{PHIN} - \mathrm{PHID}$ is diagnostic in the provided context.	Catches sign-definition errors and prevents generic petrophysics heuristics from overriding benchmark-specific computations.
Pay-zone reasoning	Prioritize explicit depth-specific flags such as PAY_FLAG_PETROPHYSICAL, MOVABLE_HC, and EXCELLENT_RESERVOIR, then verify co-located VSH/PHIT/SW/PERM support rather than relying on interval averages.	Improves pay-zone ranking and avoids dismissing localized pay evidence diluted by interval means.
Reservoir-quality proxies	Use computed FZI only when present; if unavailable, reason from PERM and PHIT with caveats, and avoid computing bulk FZI from interval means.	Keeps reservoir-quality explanations traceable to available data.
Evidence discipline	Follow supplied computation methodology and analysis summaries when they conflict with generic rules; distinguish data insufficiency from confirmed absence of hydrocarbons.	Produces grounded, caveated answers and reduces faithfulness errors.

Table 19: Representative GRCO experience-buffer entries. The actual buffer contains compact natural-language rules of this form; retrieved rules are inserted into the inference prompt as an experience prior rather than used to update model weights.

Stage	Supervisor / Judge	#Items	Context	Artifact
SFT	Kimi-K2.6 (drafter)	30,000	65k	Qwen3.5-27B, K2-V2 Inst.
DPO	Kimi-K2.6 chosen / Qwen3-8B rejected	15,000	65k	+ SFT ckpt
GRCO	DPO ckpt policy / Qwen3.5-27B judge	18,000	65k	exp. buffer

Table 20: Petrophysics-tuned training recipe at a glance. SFT/DPO update model weights on 65k-token DFS contexts; GRCO leaves weights fixed and builds an inference-time experience buffer.

J Fine-grained Per-Rubric Patterns

This appendix expands on the per-rubric and sliced-performance findings summarized in §7.1–§7.3. Numbers use the corrected basin-weighted public-split aggregation, reporting rubric scores (RA, Trust, Style, RQ, Overall) under the same context convention as Table 3: full structured context for baselines and the DFS pipeline for WellLog-LLM variants.

Basin-level analysis. (B1) Basin penalty on Trust. Trust (grounding) is the most basin-sensitive rubric. Danish $\rightarrow$ Geolink drops: Qwen3 Next 80B $46.7 \rightarrow 44.2$ (-2.5), Qwen3-8B $43.8 \rightarrow 34.4$ (-9.4), GPT-OSS-120B $48.0 \rightarrow 43.8$ (-4.2), and Qwen3.5-27B $60.2 \rightarrow 57.9$ (-2.3). Gemini 3.5 Flash remains higher but still drops $72.8 \rightarrow 68.6$, and Gemini 2.5 Pro drops $72.4 \rightarrow 68.0$. Geolink's heterogeneous clastic–carbonate mix exposes models that rely on lithology-parameter shortcuts. **(B2) Response Adequacy is also basin dependent.** RA is comparatively stable for some models but drops for others, including Qwen3.5-27B ($53.2 \rightarrow 48.3$ from Danish to Geolink) and Gemini 3.5 Flash ($71.2 \rightarrow 67.0$). **(B3) WellLog-LLM-Qwen-27B is the most basin-invariant open-source variant.** Its Overall spread across basins is only 2.0 points (COSTA 64.2, Danish 63.4, Geolink 62.1), vs. 5.0 for its untuned Qwen3.5-27B base (COSTA 63.6, Danish 60.6, Geolink 58.6). Tuning targets exactly the heterogeneous-lithology setting that hurts baselines. **(B4) Rankings remain stable across basins.** The top closed-source systems remain ahead on every basin, and Kimi-K2.6 remains the strongest public open-source baseline; the open-source order shuffles by at most 1–2 positions in the 6–12 range.

Reasoning-level analysis. (R1) Cross-Formation collapses grounding (Trust) most. Single $\rightarrow$ Cross drops are largest for mid-tier baselines: GPT-OSS-120B $50.8 \rightarrow 39.9$

(-10.9), Qwen3 Next 80B $49.8 \rightarrow 40.2$ (-9.6), and Qwen3.5-27B $63.5 \rightarrow 53.6$ (-9.9). Stronger systems are more stable but still lose grounding under cross-formation prompts: Kimi-K2.6 $66.2 \rightarrow 54.2$ and Gemini 2.5 Pro $72.2 \rightarrow 66.8$. **(R2) Response Adequacy falls in Cross too**, especially for mid-tier baselines: Qwen3 Next 80B Single $\rightarrow$ Cross drop is sharpest among open-source models, and even closed-source systems lose some RA under cross-formation prompts. **(R3) Multi-Well partially recovers what Cross-Formation loses.** For most mid-tier models, Multi-Well Overall sits between Cross-Formation and Single-Formation: Kimi-K2.6 Cross 60.1 / Multi 67.2 / Single 67.1; Qwen3 Next 80B Cross 45.3 / Multi 53.5 / Single 51.4. **(R4) Top models are reasoning-type-invariant.** Gemini 3.5 Flash has a 3.6-point Overall spread across Single/Multi/Cross ($74.7/75.5/71.9$), and Gemini 2.5 Pro is similarly stable ($74.0/74.8/71.3$). Qwen3.5-27B spreads by 5.3 points ($62.1/61.9/56.8$), while mid-tier baselines swing by roughly 5–12 points. **(R5) WellLog-LLM-Qwen-27B uniquely peaks on Multi-Well.** Its Overall profile is Single 63.6 / Cross 60.8 / Multi 64.2 – the Multi-Well peak suggests the experience prior helps most on cross-formation/multi-well synthesis, exactly where baselines weaken.

WellLog-LLM vs. base. The DPO + GRCO recipe lifts both backbones monotonically under the staged evaluation: the Qwen3.5-27B chain reaches $60.1 \rightarrow 60.6 \rightarrow 63.1$ Overall (base$\rightarrow$+DPO$\rightarrow$+DPO+GRCO), with RA climbing $52.0 \rightarrow 53.2 \rightarrow 55.8$ and Trust $60.0 \rightarrow 60.2 \rightarrow 62.4$. The K2-V2 Instruct chain (SFT$\rightarrow$+DPO$\rightarrow$+GRCO) reaches $51.6 \rightarrow 54.4 \rightarrow 56.0$, well above the un-tuned K2-V2 Instruct baseline in the same stage setting (44.7). WellLog-LLM-Qwen-27B approaches the open-source ceiling (Kimi-K2.6, 64.2) at an order-of-magnitude smaller parameter count than Qwen3 235B A22B Inst. (54.3).

J.1 Category-Level Performance

Figure 8 presents model performance across the six petrophysical categories using spider plots. The dual-panel layout separates the top-tier (Gemini 3.5 Flash, Gemini 2.5 Pro, GPT-5, Kimi-K2.6, WellLog-LLM-Qwen-27B, Qwen3.5-27B; left) from the mid-tier (Gemma-4 31B IT, Qwen3 235B A22B, Qwen3 Next 80B, WellLog-LLM-K2V2, GPT-OSS-120B, K2-V2 Instruct; right), with WellLog-LLM variants (dashed) shown in their respective tiers. Top models cover larger, more uniform hexagons; the next tier preserves the shape at a smaller scale, indicating proportional capability reduction without category-specific specialization.

Critically, no category proves universally easier or harder—per-model spreads across categories are tight relative to model-to-model gaps (Gemini 3.5 Flash 71.7–76.6, Gemini 2.5 Pro 71.0–76.0, GPT-5 62.3–70.4, Qwen3.5-27B 58.5–63.0). Tuning further flattens category variance: Qwen3.5-27B's category spread is 4.4 points before tuning and only 0.4 after DPO+GRCO; the K2-V2 Instruct family shows the same collapse from 7.1 to 0.4. This validates WellLogBench's balanced category distribution: the benchmark evaluates general petrophysical reasoning rather than narrow specialization, and gains from petrophysics-tuning (§6) appear uniformly across all six categories rather than only on a subset.

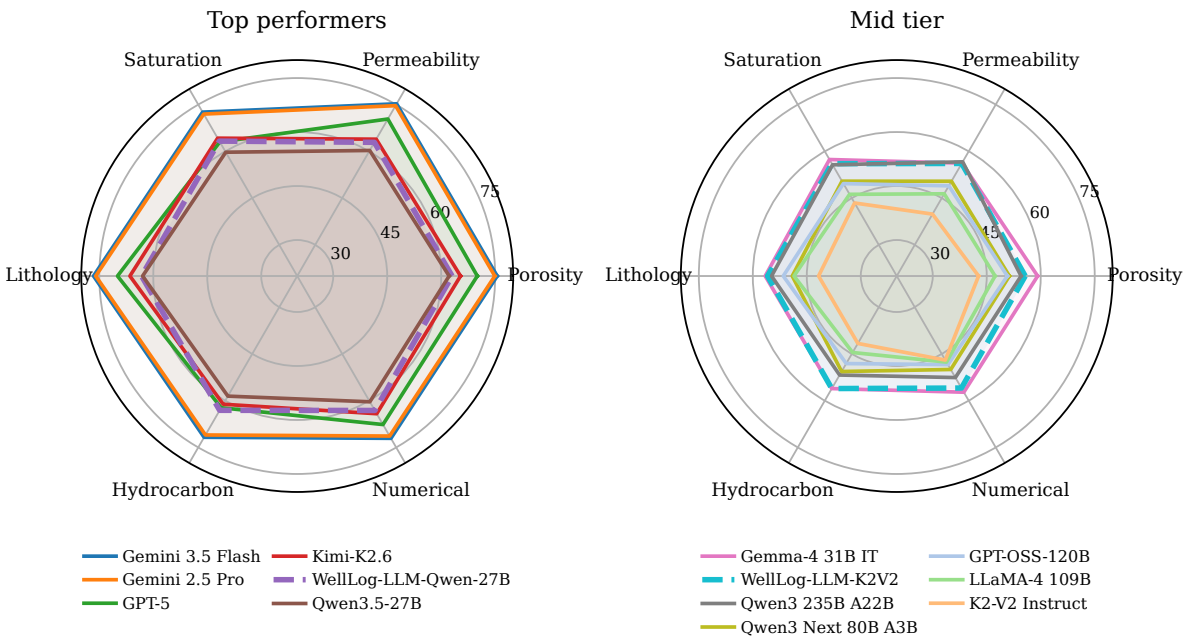


Figure 8: Category-level performance comparison. Left: top-performing closed-source, open-source, and WellLog-LLM systems show large, consistent polygons. Right: the next tier displays smaller but similarly shaped polygons. All models maintain balanced performance across categories, with no single category dominating difficulty, validating WellLogBench's domain coverage without systematic bias toward specific petrophysical topics.

K Ablations

This appendix expands the staged-recipe analysis in §7.2. We probe *where* our recipe gains come from, *how* they distribute across reasoning types, and *what* kind of remaining gap separates WellLog-LLM-Qwen-27B from frontier closed-source systems. Recipe ablations use the same DFS-compressed context as the WellLog-LLM rows in the headline table.

Coverage audit. Figure 9 summarizes the non-overlapping diagnostic axes in the public-split analysis: context compression, reasoning complexity, severe failure tails, and category balance. The remaining appendix figures then expand each visual cue into the corresponding stage, basin, category, and failure-mode breakdown.

The dashboard is an index into the detailed ablations rather than a replacement for them. The main pattern is consistent across views: the hardest failures are evidence-selection failures under compressed or cross-formation contexts, and the tuned recipe helps most when it either adapts the policy to DFS or injects GRCO experiences that remind the model which petrophysical signals must be checked before making a claim.

K.1 Recipe-stage attribution

This subsection expands §7.2 along three axes: (a) which *rubric* carries each stage, (b) how each stage *distributes its lift* across the weighted Overall, and (c) where each stage's gain *lands geographically and by reasoning complexity*.

Per-rubric trajectory. Figure 10 traces the K2-V2 Instruct four-stage recipe (base $\rightarrow$ +SFT $\rightarrow$ +DPO $\rightarrow$ +GRCO); Figure 11 reports the same information as per-stage rubric deltas, including the Qwen3.5-27B chain for comparison. The K2V2 trajectory is monotone on every rubric: SFT alone delivers the largest single jump (+11.7 Trust, +6.2 RA, +6.2 Style, +4.0 RQ), closing the domain-vocabulary gap. DPO contributes another $\sim +2$ across the rubrics, consistent with the contrastive chosen-vs.-rejected signal improving multiple answer-quality axes (Geng et al., 2025). GRCO contributes its final +1.8 on RA and +1.7 on Trust *without any gradient update*—purely by conditioning rollouts on the accumulated experience prior (Cai et al., 2025). For Qwen3.5-27B the DPO step is dominated by RA (+1.2 vs. $<$+0.3 on the other rubrics); adding GRCO is RQ-led (+3.7). Table 21 reports the K2V2 four-stage numbers in 0–100 scale; cumulative Δ(full recipe $-$ base) for K2V2 is +11.3 Overall, with Trust gaining the most (+16.0).

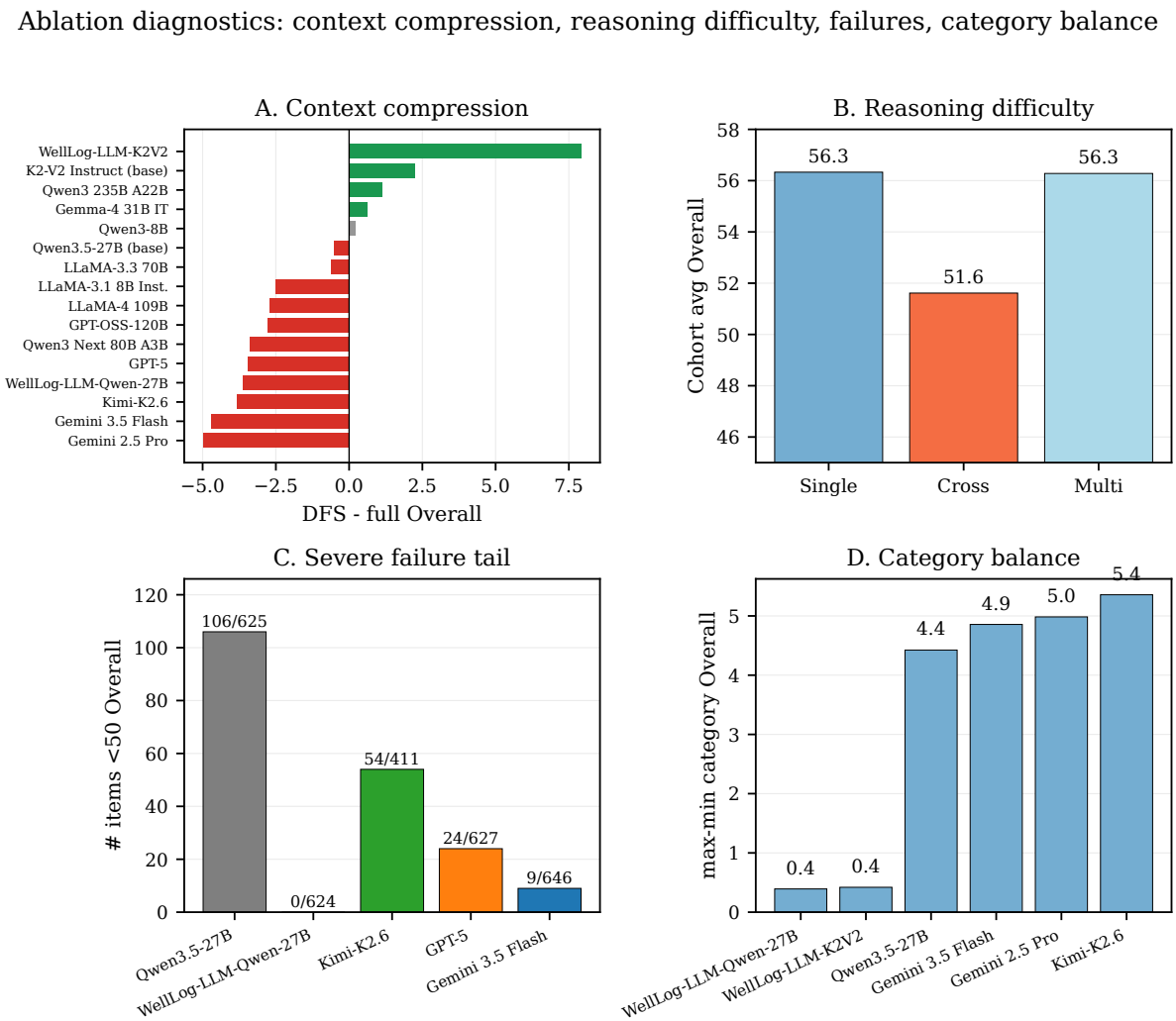


Figure 9: Ablation dashboard. A: DFS usually trades a small amount of Overall for context efficiency. B: Cross-Formation is the hardest reasoning level. C: WellLog-LLM-Qwen-27B removes the severe below-50 tail. D: tuned variants have very small category spread, showing that gains are not concentrated in a single petrophysical category.

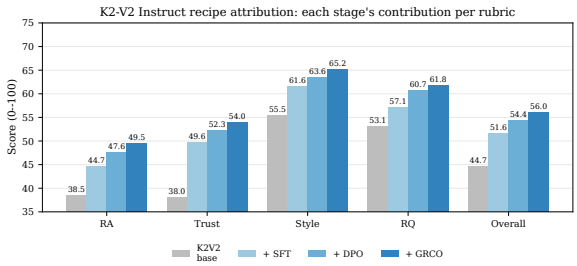


Figure 10: K2-V2 Instruct four-stage recipe attribution. SFT carries the bulk of the lift on Trust; DPO and GRCO contribute incremental but consistent gains across every rubric.

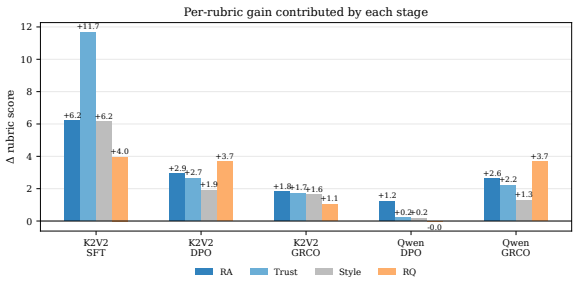


Figure 11: Per-rubric gain at each tuning stage for both backbones. K2-V2 Instruct SFT is dominated by Trust (+11.7); K2-V2 Instruct DPO is broadly positive; K2-V2 Instruct GRCO is RA-led; Qwen3.5-27B DPO is RA-led (+1.2); Qwen3.5-27B GRCO is RQ-led (+3.7).

Where each stage's lift goes. Figure 12 re-expresses the per-stage deltas as their share of the *weighted* Overall gain ($0.4{\cdot}\Delta\text{RA} + 0.2{\cdot}\Delta\text{Trust} + 0.2{\cdot}\Delta\text{Style} + 0.2{\cdot}\Delta\text{RQ}$), with the absolute Overall lift overlaid. The visualization separates two effects often conflated by raw ΔOverall: (1) the *magnitude* of each stage's improvement, and (2) the *rubric profile* of where the improvement lives. K2V2 SFT is large (+6.9 weighted) and jointly RA/Trust-heavy; K2V2 DPO and GRCO are smaller (+2.8 and +1.6) but remain broad rather than style-only. On Qwen3.5-27B, DPO contributes a thin +0.6 that is mostly RA, while GRCO is a stronger +2.5

in which RQ and RA lead.

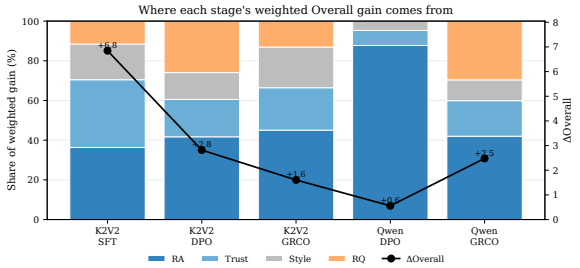


Figure 12: Share of weighted Overall gain by rubric at each stage, with the total weighted ΔOverall overlaid as a black line. Same lifts, different shapes: K2V2 SFT is RA/Trust-heavy, K2V2 GRCO is RA-heavy, and Qwen GRCO is the stage where RQ contributes most strongly.

Stage	**RA**	**Trust**	**Style**	**RQ**	**Overall**
K2V2 base	38.5	38.0	55.5	53.1	44.7
+ SFT	44.7	49.6	61.6	57.1	51.6
+ DPO	47.6	52.3	63.6	60.7	54.4
+ GRCO	49.5	54.0	65.2	61.8	56.0
Δ (final – base)	+11.0	+16.0	+9.7	+8.7	+11.3

Table 21: K2-V2 Instruct recipe-stage attribution (0–100). Stages compose: each row applies the labelled stage *on top of* the previous row's checkpoint.

Representative stage progression. A Geolink Cross-Formation hydrocarbon item asks the model to compare commercial-pay potential across BLODOEKS, EKOFISK, ROEDBY, and HOD. The expert answer ranks HOD strongest, then ROEDBY, while conditioning the call on SW/RDEP evidence and the non-diagnostic status of MHI when RXO/RMF support is missing.

Stage	Overall	Qualitative change in the answer
Base	41.7	Gives a plausible pay ranking, but skips required anchors (SW, RDEP/RMED separation, MHI validity), so the judge flags claims that *"outrun the available support."*
+SFT	47.5	Adopts the expected well-log answer structure and domain vocabulary; the response is easier to score, but still treats some derived indicators as direct evidence.
+DPO	49.6	Turns the teacher signal into contrastive grounding: claims narrow toward cited logs, unsupported pay language is reduced, and the conclusion becomes more reference-conditioned.
+GRCO	53.6	Adds the reusable inference-time rule learned from rollouts: verify RXO/RMF before trusting MHI, reconcile pay flags with SW/RDEP, and keep the final ranking conditional rather than overstated.

Figure 13: Representative qualitative answer progression across the K2-V2 Instruct recipe. The same item improves monotonically from base to SFT, DPO, and GRCO while moving from plausible but under-grounded interpretation toward conditional, evidence-aware petrophysical reasoning.

Where the gain lands (per-cell). Figure 14 shows the cumulative ΔOverall (base → full recipe) by basin × reasoning-type cell, and Table 22 reports the underlying numbers. K2-V2 Instruct gains are large and broadly distributed: every covered cell improves by ≥+9 Overall, and the largest cell gain is on Geolink Single-Formation (+13.4). Qwen3.5-27B gains are smaller and concentrated on the harder cells: COSTA Multi-Well (+4.7), Danish Cross-Formation (+4.4), Geolink Multi-Well (+3.7); the COSTA Single-Formation cell is essentially flat (−0.4)—little room to improve where the base model is already strong.

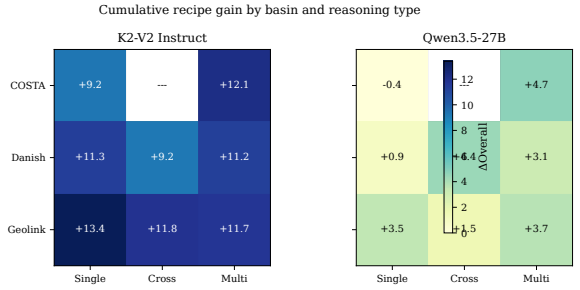


Figure 14: Per-(basin × reasoning-type) cumulative Overall gain (base → full recipe). K2-V2 Instruct (left): broad, large gains everywhere. Qwen3.5-27B (right): gains concentrate on harder cells (Cross-Formation, Multi-Well) and saturate where the base is already strong.

Basin	Reasoning	K2V2 base	K2V2 final	Qwen base	Qwen final
COSTA	Single-Form.	46.9	56.2 (+9.2)	63.0	62.7 (−0.4)
COSTA	Multi-Well	44.1	56.2 (+12.1)	60.0	64.7 (+4.7)
Danish	Single-Form.	46.5	57.8 (+11.3)	62.1	63.1 (+0.9)
Danish	Cross-Form.	44.0	53.2 (+9.2)	58.5	62.8 (+4.4)
Danish	Multi-Well	46.2	57.4 (+11.2)	61.2	64.3 (+3.1)
Geolink	Single-Form.	45.8	59.2 (+13.4)	61.2	64.7 (+3.5)
Geolink	Cross-Form.	41.8	53.6 (+11.8)	57.2	58.7 (+1.5)
Geolink	Multi-Well	44.7	56.3 (+11.7)	60.2	63.9 (+3.7)

Table 22: Per-(basin × reasoning-type) Overall scores for base vs. final tuned variant.

Additional qualitative answer behaviours. Figure 15 complements the stage and per-cell evidence with representative high-scoring WellLog-LLM-Qwen-27B behaviours. The examples are intentionally diverse: they show refusal to answer an invalid numerical comparison, conditional gas interpretation under bad-hole conditions, multi-well contrastive synthesis, and lithology discrimination from paired PEF/RHOB evidence. Appendix P then provides one complete answer in its original structured style.

Representative qualitative behaviours from WellLog-LLM-Qwen-27B.

Question type	Overall	What the answer gets right
Invalid numerical comparison	64.7	For a KHARAIB-2 multi-well permeability question asking for sandstone-vs.-shale PERM, the answer first refuses the requested subtraction because all three intervals are limestone, then reports the available limestone PERM contrast instead of inventing absent lithologies.
Gas-sensitive saturation	64.7	For BURTON FM. TOP, the answer separates a thin silt gas-effect interval (2698.5–2698.8 m) from a lower sand interval with low SW but no neutron-density crossover, and explicitly downgrades confidence because the silt parent interval is 88.26% bad hole.
Multi-well hydrocarbon contrast	64.7	For KHARAIB-2 wells HW-32 vs. HW-5, the answer preserves the key contrast: HW-32 is log-indicated gas-bearing with commercial potential, while HW-5 is predominantly water-bearing; it also notes that missing shallow resistivity and DTS limit invasion and elastic-fluid typing.
Lithology discrimination	64.7	For ROGALAND GP. TOP, the answer does not force a carbonate interpretation. It uses paired PEF and RHOB responses to describe extrusive/volcaniclastic and basal silt intervals, concluding that no clean carbonate-dominated interval is supported.

Figure 15: Additional qualitative examples showing the kinds of answer behaviours rewarded by the benchmark. The common pattern is not longer prose, but better evidence selection: refusing impossible calculations, separating hydrocarbon indicators by mechanism, and making conclusions conditional on missing or unreliable logs.

Qualitative shift in judge reasoning. Figure 16 re-bins each item's Faithfulness judge reasoning into four phrasing buckets, ordered from severe to lenient. The base K2-V2 Instruct model receives the severe *"claims outrun the available support"* phrasing on 80% of items; after the full recipe, that bucket drops to 14%, with most items reclassified into the milder *"some claims extend beyond what the cited logs clearly support."* On Qwen3.5-27B, the post-GRCO majority shifts to a strictly stricter bucket the base never receives, *"Most claims stay within the benchmark frame and avoid major unsupported extensions"* (35%, 217/624). This is the same effect the rubric numbers capture (Trust improves, Hallucination decreases), but observed in the judge's natural-language framing rather than the score.

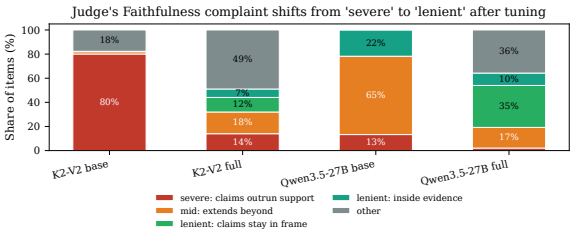


Figure 16: Distribution of judge's Faithfulness reasoning across four phrasing buckets (severe → lenient), for each of K2-V2 Instruct base, WellLog-LLM-K2V2, Qwen3.5-27B base, and WellLog-LLM-Qwen-27B. After tuning, the severe *"claims outrun the available support"* bucket collapses and items migrate to lenient framings the base models essentially never receive.

K.2 Recipe transfer across backbones

Figure 17 shows the same recipe applied to two different backbones (K2-V2 Instruct and Qwen3.5-27B). Both trajectories are monotone in every rubric; final Overall scores are 56.0 for K2V2 and 63.1 for Qwen, tracking base capability rather than flattening model differences. This confirms the recipe is backbone-agnostic.

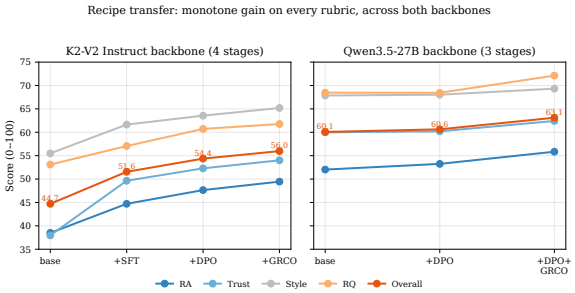


Figure 17: Recipe transfer: per-rubric trajectory for both backbones. K2-V2 Instruct (left, 4 stages) and Qwen3.5-27B (right, 3 stages) show the same monotone shape on every rubric.

K.3 Where the experience prior helps most

Figure 18 plots Overall against reasoning type. Every baseline shows a characteristic V–shape: highest on Single-Formation, lowest on Cross-Formation, partial recovery on Multi-Well. **WellLog-LLM-Qwen-27B is the only model in the cohort whose peak is on Multi-Well rather than Single-Formation** (64.2 vs. 63.6), with the smallest Single–Cross gap among open-source variants in this cohort. The GRCO experience prior

most directly targets the multi-formation synthesis behaviour that base models lose under cross-formation prompts.

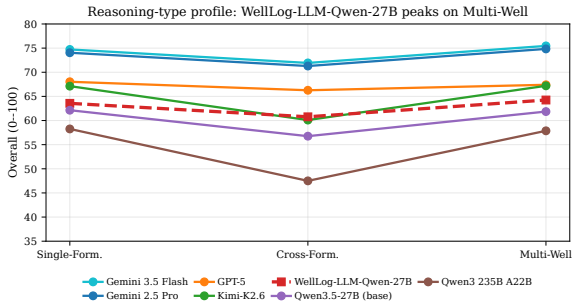


Figure 18: Reasoning-type profile of selected models. The tuned variant (red dashed) is the only one whose Multi-Well score exceeds its Single-Formation score, and it has the flattest Single → Cross dip among open-source baselines.

K.4 Faithfulness vs. Hallucination: grounding map

Figure 19 maps each model on Faithfulness (claims the judge can verify against the context) against the Hallucination judge score (high = fewer unsupported claims). The two are strongly correlated, and the WellLog-LLM variants land in the top-right cluster alongside the closed-source frontier—they are not just larger or more verbose; they are more grounded. WellLog-LLM-K2V2 sits noticeably above un-tuned K2-V2 Instruct on both axes, confirming that the recipe pushes the model toward correct grounding rather than confident extrapolation.

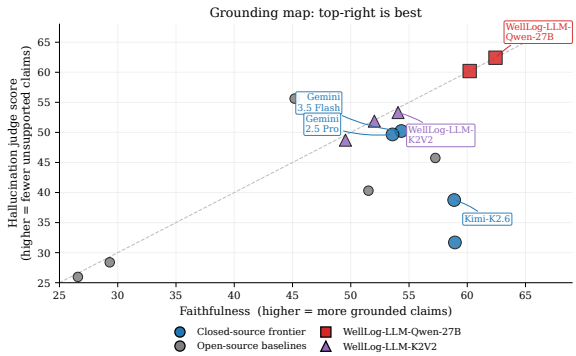


Figure 19: Faithfulness vs. Hallucination judge scores. Top-right is best. Tuned variants (red squares, purple triangles) cluster with frontier systems; un-tuned mid-tier baselines sit in the lower-left.

K.5 Where the remaining gap with frontier lives

Figure 20 compares WellLog-LLM-Qwen-27B to the two strongest baselines (Gemini 3.5 Flash, Kimi-K2.6) on eight diagnostic sub-rubrics. WellLog-LLM-Qwen-27B remains close to Kimi-K2.6 across the rubric and the gap with Gemini 3.5 Flash is consistently in two places: Numerical Accuracy and Petrophysical-Claims-by-Reference. These are the metrics that reward correct arithmetic on derived petrophysical quantities and exact alignment with the expert-curated reference claims—the same axes where mid-tier models fail under Cross-Formation prompts (§J). This points to compute-faithful arithmetic and reference-anchored claim selection as the most productive next directions.

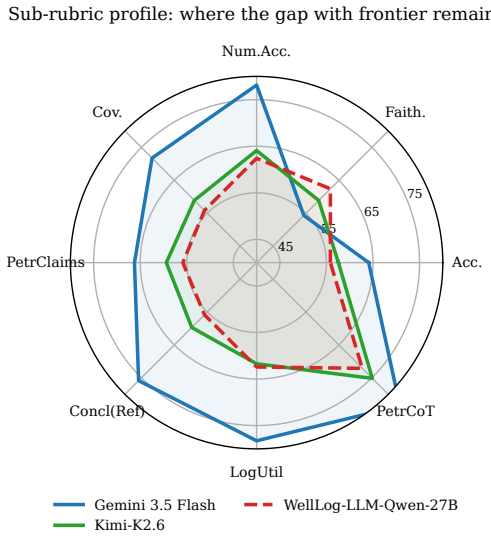


Figure 20: Sub-rubric profile: WellLog-LLM-Qwen-27B (red dashed) remains close to Kimi-K2.6 (green) across the rubric. Remaining gap with Gemini 3.5 Flash (blue) localises on Numerical Accuracy and Petrophysical-Claims-by-Reference.

K.6 Parameter–Overall Pareto

Figure 21 places models on a (parameter count, Overall) plane. Open-source baselines (gray) trace a clear capability–scale frontier that flattens above 100B. Closed-source frontier systems (blue stars) define the upper envelope. **WellLog-LLM-Qwen-27B (red square)** sits in the upper-left quadrant: at 27B parameters it approaches Kimi-K2.6, while open-source models at 10× the parameter count (Qwen3 235B A22B, 54.3 Overall) sit well below. WellLog-LLM-K2V2 (purple triangle) shows the same effect on the larger backbone. The Pareto picture is consistent with our headline claim: the recipe *shifts* the open-source frontier rather than chasing closed-source scale.

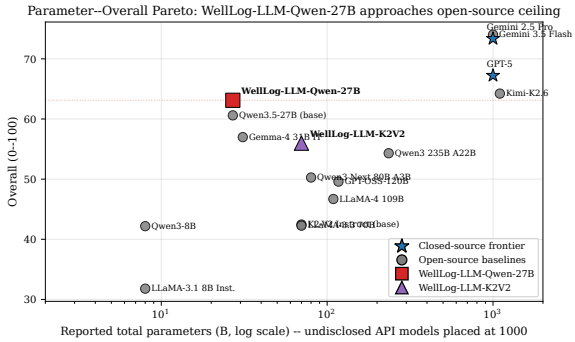


Figure 21: Parameter–Overall Pareto. WellLog-LLM-Qwen-27B approaches the open-source ceiling at an order-of-magnitude smaller scale than the strongest public baseline.

K.7 Per-basin tuning gains

Figure 22 decomposes the WellLog-LLM-Qwen-27B gain over the Qwen3.5-27B base by basin. The Overall profile is positive across all basins: COSTA gains +0.6, Danish gains +2.8, and Geolink gains +3.5. The per-rubric profile explains the pattern: Geolink improves strongly on RA (+5.9), Trust (+3.8), and Style (+3.0), while COSTA mostly gains on RQ (+2.0) and RA (+0.7). Thus the recipe is not simply raising every score uniformly; it helps most where the base model struggles to select and ground evidence.

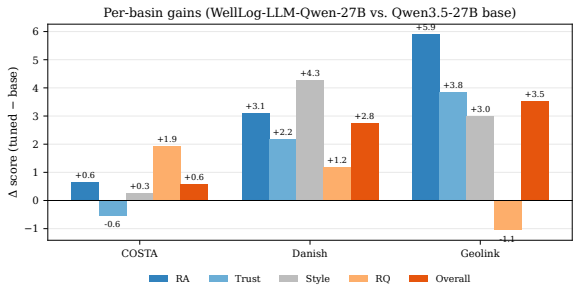


Figure 22: Per-basin gain of WellLog-LLM-Qwen-27B over the Qwen3.5-27B base. Overall is flat-to-positive, with the largest gains on Danish and Geolink.

K.8 Per-reasoning-type tuning gains

Figure 23 repeats the decomposition along reasoning type. Single-Formation, the easiest category, gains the least Overall (+1.4); Cross-Formation gains the most (+4.0), followed by Multi-Well (+2.4). The Cross-Formation lift is specifically a grounding lift: RA rises +5.6 and Trust rises +6.7, even though RQ is slightly lower. This is the desired profile—the recipe helps most when answers must reconcile evidence across formations rather than simply restate local interval statistics.

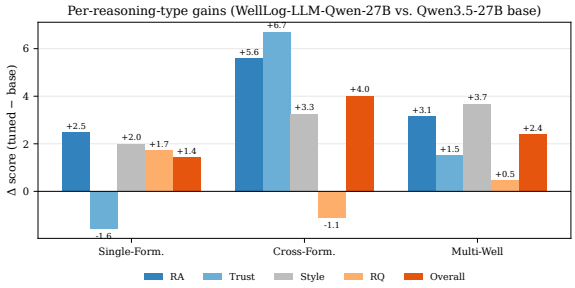


Figure 23: Per-reasoning-type gain of WellLog-LLM-Qwen-27B over the Qwen3.5-27B base. Cross-Formation gains the most, driven by RA and Trust.

K.9 Qualitative failure modes

The dominant failure pattern, sampled from the judge's free-text reasoning on the worst-scoring Qwen3.5-27B base items under the baseline context setting, is *extrapolation beyond the cited evidence*—i.e., adding plausible-sounding petrophysical detail that the ground truth does not justify:

- **Faithfulness gaps**: *"Several claims outrun the available support and are not well anchored to the cited logs."*
- **Hallucination patterns**: *"The answer introduces claims that go beyond what the cited logs and the main conclusion actually support."*
- **Numerical-Accuracy failures**: *"Quantitative support is limited relative to "Cross-Formation Overview", and key depth/value relationships are not carried through cleanly."*

Figure 24 visualizes the same tail. The left panel counts severe failures under the headline context convention; the right panel extracts the evidence channels most often named as missing by the judge in Qwen3.5-27B base failures.

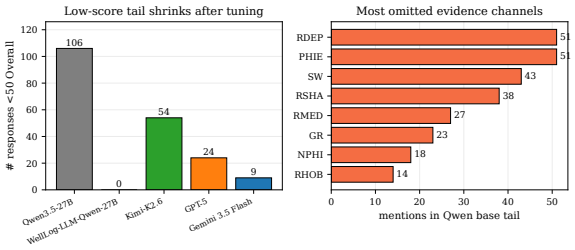


Figure 24: Severe failure-tail diagnostic. WellLog-LLM-Qwen-27B removes the below-50 tail, and the remaining base-model failures are concentrated in omitted grounding evidence such as RDEP, PHIE, SW, RSHA, and RMED.

This pattern is consistent across Faithfulness and Hallucination low-score items and matches the residual radar gap in §K.5: even after tuning, the model still occasionally introduces logs (RDEP,

RHOB, V_{sh}, etc.) and conclusions beyond the supported ground truth. The item-level tail makes this concrete: Qwen3.5-27B has 106/625 responses below 50 Overall, concentrated in Cross-Formation (64.2%), while WellLog-LLM-Qwen-27B removes that tail entirely (0/624). The judge's coverage text most often names missing RDEP, PHIE, SW, RSHA, and RMED, showing that failures are concentrated in omitted evidence channels rather than merely weak prose. Tightening the remaining gap requires reference-aware decoding constraints, not just more training data.

L Modality Definitions

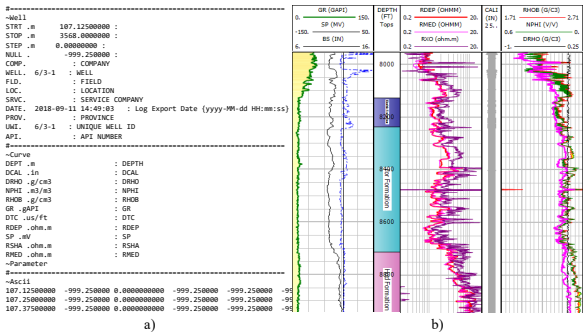


Figure 25: Well log data formats: (a) LAS file structure showing header sections and depth-indexed ASCII data section; (b) CPI display for a well showing multi-track layout across 8000-8950m depth interval with three formation boundaries.

LAS (Log ASCII Standard) files (Figure 25-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 visual displays, exemplified by Computer Processed Interpretation (CPI) panels taken from the software (Figure 25-b), present multi-track layouts that simultaneously render several vertical tracks spanning thousands of meters, each serving distinct petrophysical properties and derived parameters (water saturation, clay volume, pay flags, etc.), and color-coded lithology bars. Petrophysical interpretation of these displays requires precise quantitative extraction (reading curve values at specific depths against fine-grained linear or logarithmic scales), alongside multi-step analytical reasoning such as recognizing diagnostic patterns and trends, correlating features across all curves while maintaining depth alignment, and integrating evidence to identify formation boundaries and hydrocarbon zones.

Standardized visualizations of well logs (Figure 26) were created for WellLogBench's evaluation from LAS-derived measurements. These displays span over the total depth of a formation with

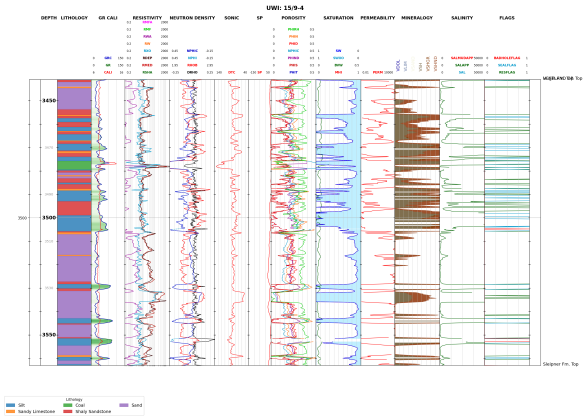


Figure 26: Standardized well-log visualization format used in WellLogBench's multimodal evaluation.

10-15 tracks showing raw measurements (gamma ray, resistivity, density, neutron, sonic), derived petrophysical properties (porosity, water saturation, clay volume, permeability), interpretation flags (pay zones, reservoir quality, seal flags), and color-coded lithology bars. These displays preserve the complexity of professional CPI. They include overlaid curves with various colors and line styles, as well as logarithmic and linear scales that span multiple orders of magnitude. Additionally, they ensure depth-aligned correlation requirements across all tracks. This consistency in formatting allows for a systematic evaluation of VLMs, assessing their ability to extract numerical values, recognize diagnostic patterns, and perform multi-curve reasoning for hydrocarbon identification and reservoir characterization.

M Extended Vision-Language Model Analysis

This appendix consolidates the full VLM evaluation behind §7.4: headline multimodal and visual-only scores, a four-pattern probe that localises failure to visual-to-quantitative extraction, and a rubric-level degradation chart. Text-context comparisons refer back to the headline public scores in Table 3. We evaluate across three input configurations—structured context only, multimodal (context + plot), and plot only; multimodal runs use truncated context to isolate the visual contribution.

M.1 Multimodal Performance

Table 23 reports VLM performance with both the structured context and well-log plots. Gemini 2.5 Pro reaches 73.0, essentially tying its headline text-context score (73.4), confirming minimal additive visual contribution. Other VLMs show the same

pattern (Kimi-K2.6: 63.6 multimodal vs. 64.2 text-context; Qwen3 235B A22B: 55.2 multimodal vs. 54.3 text-context).

Model	RA	Trust.	Style	RQ	Ovr.
Gemini 2.5 Pro	68.2	69.6	77.8	81.0	**73.0**
Kimi-K2.6	57.6	59.9	68.4	74.7	63.6
Qwen3.5-27B	50.7	59.3	67.3	72.1	60.0
Gemma-4 31B IT	48.0	53.7	64.7	68.1	56.5
Qwen3 235B A22B Inst.	48.5	50.9	61.7	66.4	55.2
LLaMA-4 109B	33.7	39.9	66.6	52.3	45.2

Table 23: Multimodal (Text + Visual) VLM performance (public test set, 60% split).

M.2 Visual-Only Performance

Table 24 reports visual-only performance. Response Adequacy and Trustworthiness collapse into the ∼17–23 range (roughly a 60% drop vs. multimodal), while Style and Reasoning Quality remain partly viable (about 47–63), revealing *perceptual* rather than *conceptual* failure: VLMs cannot extract quantitative data from scientific plots but retain some logical/reasoning capacity.

Model	RA	Trust.	Style	RQ	Ovr.
Gemini 2.5 Pro	23.2	22.2	62.6	61.6	38.5
Qwen3 235B A22B Inst.	18.9	19.6	57.5	52.8	33.5
Gemma-4 31B IT	18.6	19.3	57.9	55.4	34.0
Qwen3.5-27B	18.7	18.4	57.5	54.5	33.6
LLaMA-4 109B	16.7	17.1	53.9	47.4	30.4
Mistral Small 4 119B	19.5	19.6	56.9	51.6	33.4
Kimi-K2.6†	21.5	20.5	62.2	58.1	36.7

Table 24: Visual-only VLM performance (public test set, 60% split). †Kimi-K2.6 visual-only carried over from prior evaluation; not included in the current public-split evaluation.

M.3 Rubric-Level Degradation

Figure 27 shows rubric-level comparison between multimodal and visual-only modes across comparable VLM families. Response Adequacy and Trustworthiness show severe degradation (roughly 25–47 point drops), while Style and Reasoning Quality exhibit smaller but still visible drops (roughly 5–20 points). This pattern demonstrates systematic perceptual failures across architectures—VLMs cannot extract quantitative data from scientific charts—explaining why pairing plots with the structured context fails to improve over context-only baselines. Future VLM development should prioritize scientific chart perception before investing in multimodal fusion architectures.

Figure 27: VLM rubric degradation. Each panel shows one rubric with multimodal (dark, solid) vs visual-only (light, dashed) bars. Response Adequacy and Trustworthiness (top row, pink backgrounds) show the largest drops; Style and Reasoning Quality (bottom row) degrade more moderately.

M.4 Four-Pattern Failure Probe

We expand the four-pattern probe summarised in §7.4 with full tables.

Pattern 1: Category-level dissociation in vision-only mode. Across VLMs, Response Adequacy and Trustworthiness floor at ∼17–23, while Style and Reasoning Quality remain moderate (47–63), localizing failure to quantitative extraction rather than domain knowledge.

Model	RA	Trust.	Style	RQ
Gemini 2.5 Pro	23.2	22.2	62.6	61.6
Qwen3 235B A22B Inst.	18.9	19.6	57.5	52.8
Mistral Small 4 119B	19.5	19.6	56.9	51.6
Gemma-4 31B IT	18.6	19.3	57.9	55.4
Qwen3.5-27B	18.7	18.4	57.5	54.5
LLaMA-4 109B	16.7	17.1	53.9	47.4
Kimi-K2.6†	21.5	20.5	62.2	58.1

Table 25: Visual-only rubric breakdown. Grounding metrics (RA, Trust.) collapse while independent reasoning (Style, RQ) is preserved. †Kimi-K2.6 visual-only carried over from prior evaluation.

Pattern 2: Vision-only → multimodal rubric deltas. Adding structured text alongside the image bypasses the vision channel: grounding rubrics (RA, Trust.) jump ∼2.5–3.1× while Style and Reasoning rise only ∼1.1–1.3×—domain reasoning is already present in vision-only; only the plot-reading channel fails.

Pattern 3: Text/Vision/Multimodal three-way ablation. For comparable model families with text, visual, and multimodal measurements, V/T quantifies per-rubric degradation using the headline text-context scores from Table 3: grounding

Model	RA (V→M)	Trust.	Style	RQ
Gemini 2.5 Pro	23.2→68.2	22.2→69.6	62.6→77.8	61.6→81.0
Kimi-K2.6[†]	21.5→57.6	20.5→59.9	62.2→68.4	58.1→74.7
Qwen3.5-27B	18.7→50.7	18.4→59.3	57.5→67.3	54.5→72.1
Gemma-4 31B IT	18.6→48.0	19.3→53.7	57.9→64.7	55.4→68.1
Qwen3 235B A22B Inst.	18.9→48.5	19.6→50.9	57.5→61.7	52.8→66.4
LLaMA-4 109B	16.7→33.7	17.1→39.9	53.9→66.6	47.4→52.3
Avg. ratio	$\sim$2.6$\times$	$\sim$2.8$\times$	$\sim$1.16$\times$	$\sim$1.25$\times$

Table 26: Vision-only (V) → Multimodal (M) rubric-level deltas. Grounding rubrics improve substantially when text is added; independent reasoning (Style/RQ) is less affected. [†]Kimi-K2.6 visual-only is carried over from a prior evaluation.

(RA, Trust.) retains only $\sim$0.3–0.5 of text-mode performance; style/reasoning retain $\sim$0.75–0.96; multimodal essentially recovers text-mode grounding.

Model	Response Adequacy		Trustworthiness		Style		Reasoning Quality	
	T / V / M	V/T	T / V / M	V/T	T / V / M	V/T	T / V / M	V/T
Gemini 2.5 Pro	68.9 / 23.2 / 68.2	0.34	69.9 / 22.2 / 69.6	0.32	78.1 / 62.6 / 77.8	0.80	81.1 / 61.6 / 81.0	0.76
Kimi-K2.6†	58.9 / 21.5 / 57.6	0.37	61.3 / 20.5 / 59.9	0.33	67.9 / 62.2 / 68.4	0.92	74.1 / 58.1 / 74.7	0.78
Qwen3.5-27B	52.3 / 18.7 / 50.7	0.36	60.3 / 18.4 / 59.3	0.31	66.5 / 57.5 / 67.3	0.86	71.6 / 54.5 / 72.1	0.76
Gemma-4 31B IT	49.4 / 18.6 / 48.0	0.38	55.1 / 19.3 / 53.7	0.35	63.7 / 57.9 / 64.7	0.91	67.3 / 55.4 / 68.1	0.82
Qwen3 235B A22B Inst.	48.1 / 18.9 / 48.5	0.39	49.4 / 19.6 / 50.9	0.40	60.1 / 57.5 / 61.7	0.96	65.9 / 52.8 / 66.4	0.80
LLaMA-4 109B	36.3 / 16.7 / 33.7	0.46	43.9 / 17.1 / 39.9	0.39	64.5 / 53.9 / 66.6	0.84	52.3 / 47.4 / 52.3	0.91

Table 27: Three-way ablation: Text (T) / Visual (V) / Multimodal (M) per rubric. Text values are the headline public text-context scores from Table 3. Grounding (RA, Trust.) collapses in V (V/T $\approx$ 0.3–0.5) but Multimodal recovers near-T; Style and RQ degrade only modestly in V—the vision encoder is the bottleneck. †Kimi-K2.6 visual-only is carried over from a prior evaluation.

Model	Params	Response Adequacy	Trust-worthiness	Style	Reasoning Quality	Overall
GPT-5 (OpenAI, 2025a)	—	65.6	61.8	63.7	78.3	67.0
Gemini 2.5 Pro (Comanici et al., 2025)	—	68.1	70.5	76.1	80.4	72.6
Gemini 3.5 Flash (Kavukcuoglu et al., 2026)	—	**69.0**	**71.3**	**77.1**	**81.3**	**73.5**
LLaMA-3.1 8B Inst. (Meta, 2024a)	8B	17.8	21.1	52.3	32.5	28.3
LLaMA-3.3 70B (Meta, 2024b)	70B	30.5	39.2	69.0	48.0	43.4
LLaMA-4 109B (Meta, 2025)	109B	33.1	39.1	65.7	51.5	44.5
K2-V2 Instruct	70B	39.1	39.1	55.4	55.7	45.7
GPT-OSS-120B (OpenAI, 2025b)	117B	43.7	45.4	54.3	59.4	49.3
Gemma-4 31B IT (DeepMind, 2025)	31B	49.8	55.0	63.6	67.8	57.2
Qwen3-8B (Yang et al., 2025)	8B	36.8	39.6	55.0	53.0	44.2
Qwen3.5-27B	27B	52.9	60.6	66.8	71.5	60.9
Qwen3 Next 80B A3B (Team, 2025)	80B	42.5	46.1	61.6	61.9	50.9
Qwen3 235B A22B Inst. (Yang et al., 2025)	235B	49.6	52.3	60.8	63.5	55.2
Kimi-K2.6 (Moonshot AI, 2025)	1.1T	58.6	60.5	68.3	73.5	63.9
WellLog-LLM-Qwen-27B	27B	56.2	62.4	69.0	72.5	63.3
WellLog-LLM-K2V2	70B	49.8	55.3	65.3	62.7	56.6

Table 28: Model performance on the private test set (0–100). **Params** = reported total parameter count ("—" when undisclosed). Baselines use the full structured context; shaded WellLog-LLM rows use the DFS pipeline (§5.1), reducing input tokens by $\sim$70% relative to full-context inference. **Overall** $= 0.4\cdot$Response Adequacy $+$ $0.2\cdot$Trustworthiness $+ 0.2\cdot$Style $+ 0.2\cdot$Reasoning Quality. Best per column in **bold**; per-stage attribution in Table 4.

N Private Test Set Results

To validate generalization and prevent overfitting, WellLogBench employs a held-out private test set comprising 40% of the benchmark (434 questions) that is not exposed during model development or hyperparameter tuning. Table 28 mirrors the protocol of Table 3: baseline models are evaluated *without* dynamic feature selection (DFS), i.e. with full structured context, while WellLog-LLM models are evaluated *with* DFS.

The private split closely tracks the public results in Table 3. Across the 16-model cohort, the mean absolute change in **Overall** is only 0.19 points, the maximum absolute change is 1.2 points, and the model ranking remains the same between public and private splits.

Our DFS-trained models also retain their public-split advantages under hidden evaluation. **WellLog-LLM-Qwen-27B** reaches 63.3 Overall on the private split, remaining within 0.6 points of Kimi-K2.6 while using the DFS pipeline with $\sim$70% fewer input tokens than full-context baselines, and outperforming the untuned Qwen3.5-27B baseline by 2.4 points. **WellLog-LLM-K2V2** reaches 56.6 Overall, outperforming its K2-V2 Instruct backbone by 11.0 points. These results indicate that the gains observed on the public split transfer cleanly to the held-out private split rather than reflecting overfitting to the released portion of the benchmark.

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 Representative Complete Answer

The example below shows the style of a full WellLog-LLM-Qwen-27B answer after DFS context construction and GRCO experience injection. It is a single-formation saturation question from Geolink; the answer is reproduced in its structured form, with only typography normalized for page layout.

```
Question:
Which neutron-density separation intervals align with significant density
corrections and low water saturation in BURTON FM. TOP?

Answer:
## Formation Overview
The BURTON FM. TOP formation spans approximately 37.8 m (2696.1-2733.9 m)
and contains two interpreted intervals: a dominant Silt interval
(2696.1-2728.4 m) and a thinner Sand interval (2731.1-2733.7 m). The question
asks where neutron-density separation coincides with density correction and
low water saturation, so the answer evaluates ND_SEPARATION, DRHO, SW, and
bad-hole flags together.

## Statistical Analysis
The Silt interval has broader ND_SEPARATION, including positive gas-effect
flags, whereas the Sand interval remains negative. Both intervals have low SW,
but the Silt interval is much less reliable because 88.26% is flagged bad hole.

| Depth Interval (m) | Lithology | ND_SEPARATION Mean | DRHO Mean | SW Mean | Bad Hole Flag |
|---|---|---:|---:|---:|---:|
| 2696.1-2728.4 | Silt | -0.2251 V/V | -0.0122 g/cm3 | 0.2149 V/V | 88.26% |
| 2731.1-2733.7 | Sand | -0.1815 V/V | -0.0152 g/cm3 | 0.3169 V/V | 0.00% |

## Lithological Variations
### Silt Interval (2696.1-2728.4 m)
The Silt interval dominates the formation thickness and shows the most
significant variation in fluid indicators.

- Neutron-density separation: the interval shows a maximum ND_SEPARATION of
  0.1616 V/V, with specific sub-intervals flagged for gas effects.
- Density correction: DRHO values range up to 0.113 g/cm3, suggesting
  significant density variation that may reflect gas response or borehole
  washout.
- Water saturation: mean SW is 0.2149 V/V, with multiple sub-intervals
  classified as excellent hydrocarbon saturation.
- Specific intervals: 2698.5-2698.8 m exhibits strong-to-weak gas effects
  alongside low water saturation; additional low-SW zones occur deeper in the
  same silt interval.
- Data quality: the high bad-hole fraction (88.26%) is a critical limitation
  and reduces confidence in DRHO and neutron-density separation.

### Sand Interval (2731.1-2733.7 m)
The Sand interval is thinner but has better borehole conditions.

- Neutron-density separation: ND_SEPARATION is consistently negative
  (mean -0.1815 V/V), indicating no gas crossover effect.
- Density correction: DRHO is small (mean -0.0152 g/cm3), with no significant
  correction observed.
- Water saturation: mean SW is 0.3169 V/V. Low-SW zones occur at
  2731.1-2731.4 m and 2733.0-2733.3 m.
- Data quality: the bad-hole flag is 0.00%, making the log measurements more
  reliable than in the upper silt interval.

## Conclusions
The neutron-density separation intervals that align with low water saturation
are primarily within the Silt interval. The key sub-interval is
2698.5-2698.8 m, where positive ND_SEPARATION occurs with low SW. However, the
very high bad-hole fraction means the gas interpretation should remain
conditional rather than definitive.

The Sand interval (2731.1-2733.7 m) shows reliable low water saturation, but it
lacks neutron-density crossover and significant density correction. It should
therefore be treated as a non-gas low-SW hydrocarbon indication rather than a
gas-sensitive interval.

## Data Limitations and Recommendations
- RXO, DTS, PEF, TEMP, RMF, and RW are missing as primary logs. This limits
  flushed-zone saturation and VP/VS-style fluid typing.
- The 88.26% bad-hole fraction in the Silt interval materially affects DRHO
  reliability, so caliper/density-correction validation is required before
  making a firm gas-zone pick.
- NMR, pressure, or production data would improve confidence in the
  2698.5-2698.8 m gas-effect interpretation.
```

Q Prompt Templates

This appendix records the prompt templates used by the benchmark pipeline. Runtime placeholders such as `{formation_name}`, `{question}`, and `{context_data_json}` are filled programmatically. For space, repeated rubric anchors are condensed, but the scoring dimensions, output contracts, and domain constraints are preserved.

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:
{
  "Accuracy": float,
  "Accuracy Reasoning": "string",
  "Precision": float,
  "Precision Reasoning": "string",
  "Faithfulness": float,
  "Faithfulness Reasoning": "string",
  "Coverage": float,
  "Coverage Reasoning": "string",
  "Conciseness": float,
  "Conciseness Reasoning": "string",
  "Hallucination": float,
  "Hallucination Reasoning": "string",
  "Repetition": float,
  "Repetition Reasoning": "string",
  "Clarity": float,
  "Clarity Reasoning": "string",
  "Completeness": float,
  "Completeness Reasoning": "string",
  "Conclusion by Reference": float,
  "Conclusion by Reference Reasoning": "string",
  "Petrophysical Claims by Reference": float,
  "Petrophysical Claims by Reference Reasoning": "string",
  "Data Limitations Score by Reference": float,
  "Data Limitations Score by Reference Reasoning": "string",
  "Log Utilization by Reference": float,
  "Log Utilization by Reference Reasoning": "string",
```

```
  "Petrophysics Claims (Free Thinking)": float,
  "Petrophysics Claims (Free Thinking) Reasoning": "string",
  "Petrophysics Chain of Thought (Free Thinking)": float,
  "Petrophysics Chain of Thought (Free Thinking) Reasoning": "string",
  "Internal Consistency": float,
  "Internal Consistency Reasoning": "string",
  "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
- Truncated Reference Context
- LLM Response

Use the combined reference:
1. Ground Truth
2. Truncated 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}"
```