WellLogBench: A Domain-Expert Curated Benchmark for Evaluating LLM Reasoning on Subsurface Well Log Data

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown promising capabilities in text summarization, reasoning, and time-series analysis, offering significant potential to simplify workflows in the oil and gas industry by enhancing and accelerating subsurface exploration/development. However, their ability to reason over multivariate well-log data remains largely unexplored. Accurate interpretation requires integrating domain expertise with numerical reasoning over long, depth-indexed sequences—challenges unmet by existing NLP or tabular benchmarks. To the best of our knowledge, we provide the first expert-curated benchmark focused on well log reasoning, using well log curves as the primary input modality. We introduce **WellLogBench**, the first comprehensive benchmark for well log reasoning, comprising 1,085 expert-annotated QA pairs based on well log data curated from multiple wells across three geological basins in different parts of the world. The benchmark spans diverse petrophysical categories and reasoning complexities, from single-formation to multi-well analysis. Addressing the ultra-long context of raw logs and domain gap, we propose a context-engineering framework that transforms high-dimensional well log data into compact, petrophysics-rich context optimized for LLM processing. We evaluate state-of-the-art open and closed-source models; the top performer (Gemini 2.5 Pro) achieves ∼74% and the closest open-source alternative is Qwen3-235B-Instruct (∼57%) on our composite metric, highlighting substantial opportunities for improvement. WellLogBench establishes a rigorous foundation for developing and evaluating AI systems for petrophysics.

1 Introduction

Large Language Models (LLMs) have emerged as powerful tools for scientific reasoning, capable of performing multi-step logical deductions, and generating coherent technical narratives (Wei et al., 2022; Kojima et al., 2022).Their ability to act as reasoning assistants (Qiu et al., 2025; Dang et al., 2025) has transformed workflows across various disciplines by automating the integration of diverse data sources and supporting expert decision-making. In the oil and gas industry, well logs are vital for characterizing subsurface formations and guiding reservoir evaluation and drilling decisions. In a well log interpretation task, the goal is to provide accurate answers to a user query given the user-provided well log data. While LLMs offer a promising foundation for such reasoning assistants, applying them for well log interpretation presents unique challenges:

- **High-Dimensionality:** Well logs contain 30+ interdependent curves across thousands of depth steps, linking measurements to complex geological factors.
- **Ultra-Long Contexts:** Raw LAS files often exceed **2 million tokens**, far surpassing the context windows of most current LLMs.
- **Multi-Scale Reasoning:** Interpretation requires

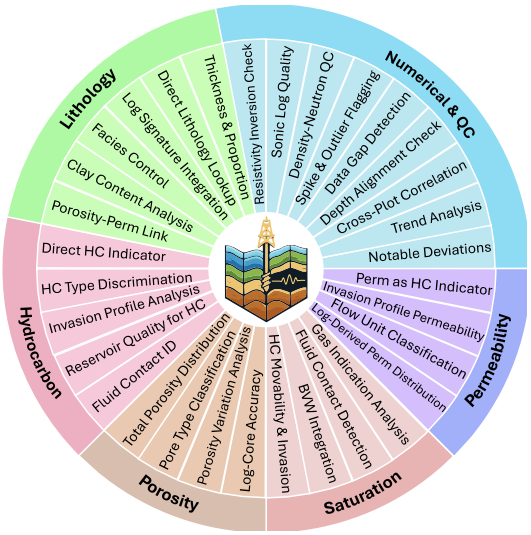


Figure 1: WellLogBench category map with six top-level petrophysical categories in the outer ring and their corresponding example subcategories in the inner ring.

both local (interval-level) decisions and global (cross-formation/multi-well) synthesis.

- **Tool-Augmented Computation:** Accuracy relies on derived quantities (e.g V_{sh}, ϕ, S_w, Pay) and Quality Control functions, not just textual generation.

These factors render direct LLM application to raw logs infeasible. Moreover, evaluation is sensitive to **petrophysical correctness**; plausible-sounding answers that violate physical constraints (e.g., invasion effects, lithology bounds) are dangerous. To bridge this gap, we introduce **WellLogBench**, a domain-expert curated benchmark comprising 1,085 QA pairs for over **62 real-world wells** from **three diverse basins**. It covers six core petrophysical categories (Figure 1) and three reasoning levels (single-formation, cross-formation, and multi-well). We pair this with a context-engineering pipeline that transforms raw LAS files[1] into lithology-aware, structured contexts via dynamic truncation, preserving decision-critical information within practical LLM limits. To ensure rigor, we collaborated with **six senior petrophysicists** to manually verify all data, ensuring the benchmark reflects professional standards. Our contributions are:

1. **WellLogBench:** The first expert-curated benchmark for well log reasoning (1,085 QA pairs, 62 wells, 3 basins), validated by senior petrophysicists.
2. **Unified Evaluation:** The first framework for formulating petrophysics reasoning by integrating well logs and LLMs, spanning six domain categories and three reasoning complexities.
3. **Context Engineering:** A domain grounded pipeline transforming raw logs (>2M tokens) into compact, LLM-ready structured contexts using a custom petrophysics-aware context generation framework and dynamic feature selection method for token and cost reduction.
4. **Public Leaderboard:** We partition the dataset into a public release (60%) for model development and a held-out private test set (40%) to host a leaderboard competition, driving community innovation and reliable LLM adoption in the oil and gas industry.
5. **Baseline Analysis:** Comprehensive evaluation of state-of-the-art LLMs and Multimodal LLMs, identifying persistent gaps in grounded petrophysical reasoning.

[1]See Appendix B for an example of a raw LAS file.

2 Related Work

We categorize prior work into scientific reasoning benchmarks, AI for petrophysics, and long-context time-series reasoning.

Scientific and Geoscientific Reasoning. General benchmarks such as ScienceQA (Lu et al., 2022) and MMLU (Hendrycks et al., 2021) test broad scientific knowledge but omit petrophysics and the specialized data formats needed for subsurface analysis. Recent geoscience-focused efforts (e.g., K2 (Deng et al., 2023), GeoGalactica (Lin et al., 2024)) pre-train models on geological corpora to improve text-based reasoning and support workflows like hazard assessment (Hadid et al., 2024). However, they largely focus on textual understanding and do not address the quantitative, multi-signal well-log interpretation required for petrophysical reasoning, which is central to our work.

AI in Petrophysics. AI in petrophysics has evolved from supervised estimation of reservoir properties from conventional logs (Hassaan et al., 2024; Rahmati et al., 2024; Zhang et al., 2025) to deep learning for lithofacies/lithology interpretation (Shi et al., 2024; Jiang et al., 2024; Mohammadi et al., 2025) and physics-informed methods that embed petrophysical constraints (Shao et al., 2024; Pothana and Ling, 2025). More recently, foundation-model efforts pre-train transformers on large well log corpora for multi-task interpretation (Qi et al., 2025; TGS, 2025), while generative models address curve reconstruction and imputation (Zhangxing et al., 2025; Al-Fakih et al., 2025). In parallel, LLMs are being adapted to petroleum workflows via domain-specific models trained on SPE/OnePetro content (Eckroth et al., 2025) and applications such as automated logging report drafting (Li et al., 2025) and tool/workflow integration (Xu and Hulea, 2024; Srivastav and Kothe, 2025). However, most work focuses on short-window prediction, document QA, or proprietary pipelines, and does not benchmark multi-step *reasoning* over raw LAS file, ultra-long, multivariate well log contexts. WellLogBench fills this gap with expert-validated questions targeting reasoning and explanation over raw well log data.

Long-Context Time-Series Reasoning. Well logs are *ultra-long, multivariate depth-indexed time series* requiring interval aggregation, trend/change reasoning, and cross-curve consistency checks. Existing long-context benchmarks (e.g., LongBench

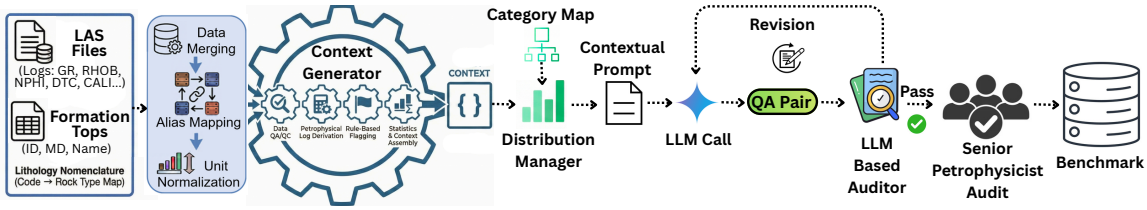


Figure 2: End-to-end QA generation pipeline showing the flow from raw well log data through context generation, LLM-based QA creation, automated auditing, and domain expert verification.

(Bai et al., 2024), RULER (Hsieh et al., 2024)) are largely text-centric, while time-series evaluations report persistent zero-shot temporal reasoning gaps (Merrill et al., 2024) and motivate time-series LLM/QA frameworks (Chow et al., 2024; Kong et al., 2025). WellLogBench addresses this gap by benchmarking expert-validated *petrophysical reasoning and explanation* directly over ultra-long well log contexts.

3 WellLogBench: Design and Data Creation

Overview. WellLogBench comprises 1,085 expert-validated QA pairs derived from public LAS well logs spanning three diverse basins covering six petrophysical categories and three reasoning levels (single-formation, cross-formation, multi-well). We construct QA pairs via a four-stage pipeline: (1) ingest and clean LAS files, (2) perform lithology-aware context engineering with derived quantities, (3) generate questions/answers using Gemini 2.5 Pro (Comanici et al., 2025) with category-specific templates, and (4) audit with GPT-5 (OpenAI, 2025a), followed by validation by domain experts. Figure 2 illustrates the end-to-end process.

3.1 Design Philosophy

WellLogBench is guided by three principles:
1) Domain Authenticity: Questions are drawn from real petrophysical interpretation and reporting workflows rather than synthetic toy settings. Six senior petrophysicists curated and validated the content to ensure the required assumptions, terminology, and outputs align with industry practices.
2) Reasoning Complexity: The benchmark spans a spectrum from direct numerical retrieval and basic calculations to multi-step interpretation that synthesizes multiple curves and formation context. Harder items require interval-level aggregation, cross-curve consistency checks, and explanation of decisions, pushing models beyond local pattern matching.

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.

3) Geological Diversity: To test generalization, we include logs from multiple basins and depositional settings with varied lithologies (carbonate, clastic, and mixed), tool suites, and logging conditions, reducing overfitting to a single geology or acquisition style.

3.2 Data Sources

We curate well log data from three publicly available repositories that jointly span mixed clastic–carbonate, chalk-dominated carbonate, and Middle Eastern carbonate settings. Table 1 presents the distribution across reasoning levels and data sources.

Geolink: A collection of North Sea wells with mixed clastic–carbonate successions. This forms our largest subset with 677 QA pairs (Gama et al., 2024).
COSTA: Offshore wells from a carbonate-prone Middle Eastern setting, based on the COSTA open-source carbonate benchmarking case study, contributing 271 QA pairs (Costa Gomes et al., 2022).
Danish Basin: Wells from the Danish sector of the North Sea, characterized by Upper Cretaceous–Danian chalk and associated carbonate formations, providing 137 QA pairs (Geological Survey of Denmark and Greenland (GEUS)).

3.3 Petrophysical Categories

Questions are organized into six categories aligned with core subsurface interpretation workflows (Figure 1), targeting distinct reasoning skills: numerical computation, multivariate integration, causal inference, and decision-making under uncertainty.

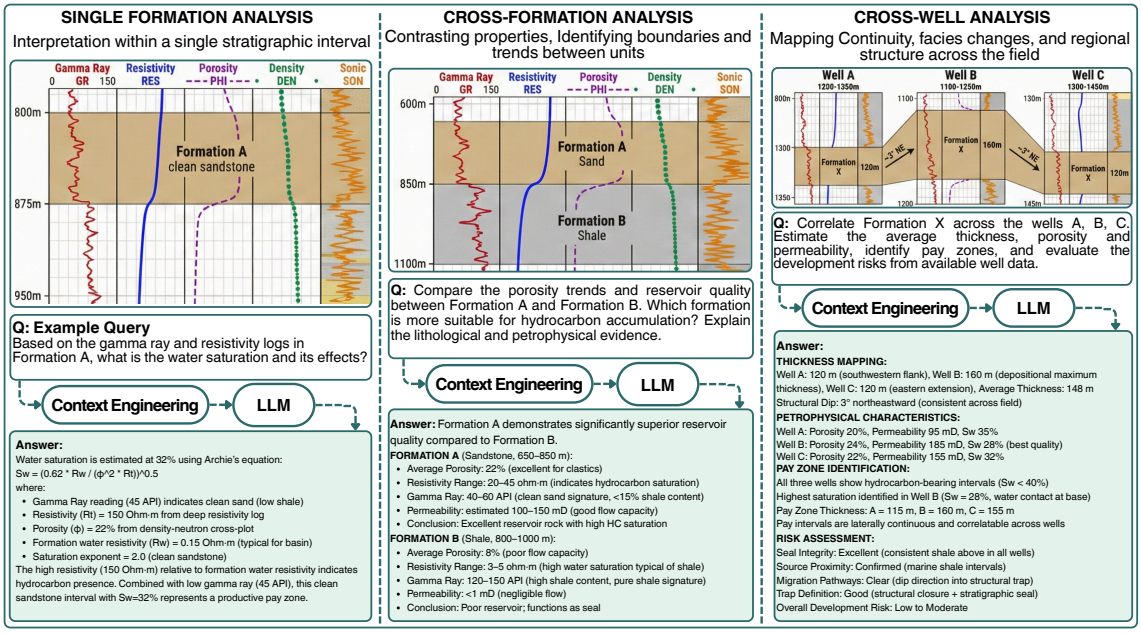


Figure 3: Hierarchical Reasoning Complexity. WellLogBench evaluates models across three scales of analysis: localized Single-Formation evaluation, vertical Cross-Formation comparison within a wellbore, and regional Multi-Well correlation across fields.

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.

1) Numerical & QC: Evaluate data quality, identify artifacts, and judge interpretation reliability.
2) Porosity Analysis: Estimate porosity, apply lithology corrections, and link to reservoir behavior.
3) Saturation Analysis: Infer fluid saturation using resistivity-based methods across depth profiles.
4) Hydrocarbon Analysis: Identify hydrocarbon zones, distinguish fluid phases, and assess producibility.
5) Permeability Analysis: Estimate flow capacity using porosity–permeability trends and log signatures.
6) Lithology Analysis: Classify rock types from logs and maintain stratigraphic consistency.

Table 2 shows the balanced distribution across categories and reasoning levels in the entire dataset.

3.4 Reasoning Levels

We define three levels of reasoning complexity:

- **Single-Formation Analysis:** Questions focus on interpretation within a single stratigraphic interval. The model must analyze log responses, compute derived properties, and draw conclusions about a specific formation.
- **Cross-Formation Analysis:** Questions require comparison and contrast across multiple formations within a single well. The model must identify formation boundaries, compare properties, and explain geological variations.
- **Multi-Well Analysis:** This requires regional-scale reasoning by integrating data from multiple wells for the same geological formation across multiple wells. Questions test correlation, trend identification, and synthesis across multiple wells.

Figure 3 illustrates the three reasoning levels.

3.5 QA Generation Pipeline

The creation of WellLogBench follows a systematic 4-stage pipeline that integrates automated auditing with domain expertise to transform raw well log data into expert-validated QA pairs. Ground truth answers are expert-corrected and rubric-validated (not raw model outputs).

Stage 1: Data Ingestion and Preprocessing. We parse LAS curves and merge them with auxiliary geology (formation tops and lithology labels) to form depth-indexed records. To harmonize the heterogeneous GEOLINK, COSTA, and Danish

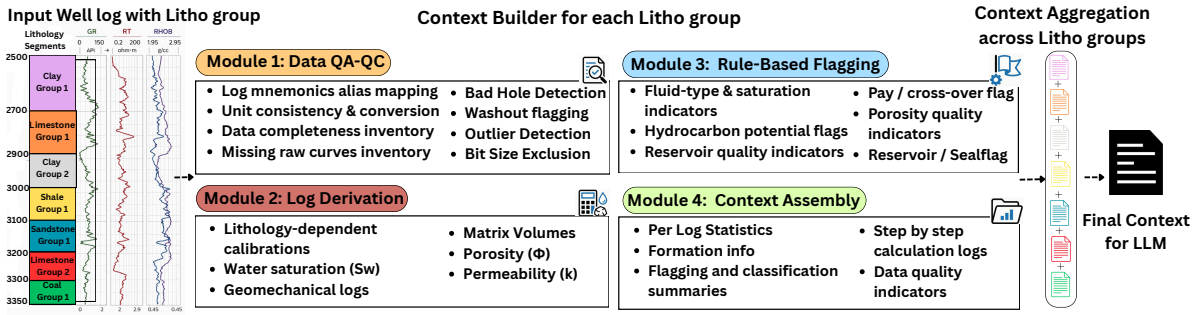


Figure 4: Context Generation Engine architecture. Input data sources are merged and segmented by lithology group. Each segment passes through four processing modules—Data QA/QC, Petrophysical Log Derivation, Rule-Based Flagging, and Statistics Assembly—producing structured contexts per lithology group.

Basin sources, we apply two standardization steps: **(a) Mnemonic mapping**: ∼200 vendor mnemonics are mapped to 25 canonical curve types (e.g., GR/SGR/GRAM → gamma ray); **(b) Unit normalization**: measurements are converted to consistent units (depth: m; porosity: fraction v/v; resistivity: Ωm; density: g/cm^3; sonic: μs/ft).

Stage 2: Context Generation and QA Pair Creation. We convert each preprocessed LAS well into a structured petrophysical context (Section 3.6) and use it to generate QA pairs with an LLM guided by domain constraints. A **Category Map** defines template type (required curves, derived quantities, example questions, and difficulty etc), while a **Distribution Manager** enforces balanced coverage across categories (15–20% each), reasoning levels, basins, and lithologies via weighted sampling. Given a sampled specification and context, the **generation module** produces a question and answer with prompts that enforce petrophysical terminology and numerical consistency.

Stage 3: Automated Auditing and Refinement. A GPT-5-based auditor checks each QA pair against the ground-truth context under a **strict-pass policy**. Pairs with issues are revised using auditor feedback; unresolved pairs are discarded and regenerated, ensuring only high-quality items reach expert review.

Stage 4: Domain Expert Verification. QA pairs that pass undergo final review by a panel of six senior petrophysicists (15+ years of experience). Experts verify technical accuracy, assess difficulty calibration, ensure alignment with industry practices, and correct or reject remaining errors.

3.6 Context Engineering Framework

A critical challenge in applying LLMs to well logs is transforming high-dimensional, depth-indexed measurements into structured contexts that preserve petrophysical relationships while remaining tractable for LLMs. Figure 4 summarizes our context engineering framework.

3.6.1 Lithology-Based Segmentation

Formations contain multiple lithology transitions (e.g., Clay → Limestone → Shale), yet averaging properties across mixed lithologies yields misleading statistics. Our framework segments formations into lithologically homogeneous intervals by: **(1)** merging log data with formation tops via depth-based alignment **(2)** mapping lithology codes to standardized groups (see Appendix C) **(3)** identifying contiguous same-lithology intervals and assigning sequential identifiers; **(4)** processing each segment independently with lithology-specific parameters.

3.6.2 Processing Modules

As shown in Figure 4, each lithology segment passes through four processing modules in the context builder:

Module 1 (Data QA/QC). Identifies unreliable data via bad-hole detection, missing-data inventory, outlier filtering, and propagates quality flags to the output contexts.

Module 2 (Petrophysical Log Derivation). Computes task-relevant derived quantities (shale volume, porosity, water saturation, permeability) from raw measurements using domain-specific formulae with lithology-dependent calibrations, exposing decision-critical variables for downstream reasoning (see Appendix I).

Module 3 (Rule-Based Flagging). Applies expert-validated heuristics to add predefined interpretations, including binary flags (e.g., RESFLAG, SEALFLAG, PAY_FLAG) and discrete bins (e.g., hydrocarbon potential, porosity quality, permeability), providing interpretable anchors beyond raw

logs and summary statistics (see Appendix I).

Module 4 (Statistics and Context Assembly). Computes per-log statistics (mean, min, max, std, percentiles) and assembles the final context containing formation metadata, lithology group information, statistics, classification intervals, and data quality metrics, yielding a complete and concise context for each lithology group.

3.6.3 Context Output Modes

The framework supports two configurations: **Single-Formation Mode** generates one context per lithology group for detailed formation analysis, which is then aggregated to form a single formation context. **Cross-Formation Mode** and **Multi-Well Mode** aggregate contexts across formations for comparative multi-formation or multi-well questions.

4 Evaluation Metrics

We evaluate model responses with an LLM-as-judge framework using GPT-5 under default parameters as the judge. Each answer is graded on a set of fine-grained metrics and then aggregated into four rubric categories that emphasize correctness, grounding, communication efficiency, and independent reasoning. All metric scores are normalized to a 0–100 scale, where higher is better:

- **Response Adequacy** — correctness and coverage:
  - *Accuracy:* overall correctness vs. reference.
  - *Petrophysical Claims (Ref.):* petrophysics facts and claims match the reference.
  - *Conclusion (Ref.):* final answer conclusion matches the reference outcome.
  - *Coverage:* addresses all required subparts/intervals/comparisons.
- **Trustworthiness** — contextual grounding and uncertainty handling:
  - *Faithfulness:* claims supported by provided logs/statistics (no fabrication).
  - *Numerical Accuracy:* computations consistent with provided inputs/derived quantities (e.g., S_w, BVW, ϕ, net pay).
  - *Data Limitations (Ref.):* correctly notes missing logs, QC/bad-hole flags, and uncertainty per reference.
  - *Log Utilization (Ref.):* uses appropriate curves/tools consistent with the reference workflow.
- **Style** — readability and efficiency:
  - *Clarity:* clear, readable, well-structured.
  - *Conciseness:* sufficient without over-generation.
  - *Repetition:* avoids redundant statements.
- **Petrophysical Reasoning Quality** — reference-free technical soundness:
  - *Petrophysics Claims:* technically plausible and consistent with petrophysical principles.
  - *Petrophysics Chain of Thought:* coherent reasoning from evidence → interpretation → conclusion.

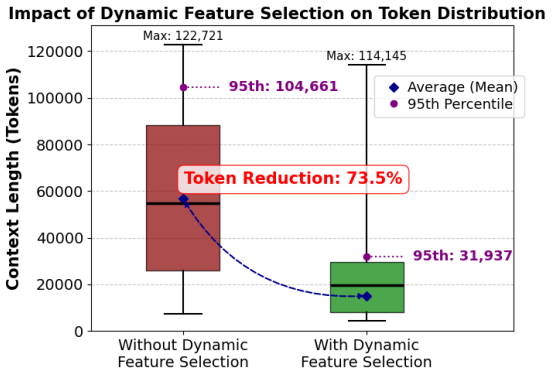


Figure 5: Impact of Dynamic Feature Selection on context length distribution. The method achieves a 73.5% reduction in average token usage.

Weighted score. The final score is calculated by combining the four rubric categories, with 40% for Response Adequacy, 20% for Trustworthiness, 20% for Style, and 20% for Reasoning Quality.

5 Addressing Long-Context

Well log data presents significant context length challenges. A single raw log is more than 2M tokens. However, after passing a raw log through our context engineering pipeline, single formation context can be on average 50k tokens, while cross-formation or multi-well questions reach a greater number of tokens, depending on the number of formations compared, surpassing the limits of current LLMs (see Appendix G).

5.1 Dynamic Feature Selection Pipeline

To operate within practical context limits while retaining task-relevant information, we implement a dynamic feature selection mechanism: an intermediate LLM call receives the full context plus the target question, then extracts only task-relevant statistics and analysis context into a compact context. This approach enables query-specific context reduction rather than static truncation rules. LLM model used for answer generation is also used for

Dynamic Feature selection. For answer generation:

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

This dynamic selection preserves task-relevant relationships while reducing token overhead, allowing models to operate efficiently within their context windows. On average for a single formation it can reduce input context tokens by 73.5% with minimal accuracy loss and providing higher inference speed.

5.2 Cross-Formation and Multi-Well Analysis

For cross-formation and multi-well questions, we employ a hierarchical decomposition: each formation or well is processed independently by a single LLM call, generating a local response. These intermediate responses are then aggregated by a second LLM call that performs comparative analysis, synthesizing relationships and trends across scales. This approach avoids concatenating contexts (which would exceed token limits) while enabling scalable multi-scale reasoning

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

6 Results and Analysis

WellLogBench employs a public-private leaderboard split to ensure robust evaluation and reduce overfitting. We partition our 1,085 expert-validated QA pairs into a **60% public test set** (646 questions) for model development and comparison, and a **40% private test set** (430 questions) held out for leaderboard competition evaluation. Benchmark results for private test set are provided in Appendix E. This section reports results on the public test set.

We evaluate models across three distinct fronts:

1. **Text-only:** Models process structured statistical summaries (with and without context feature selection).
2. **Text + Well Log Plot (Multimodal):** Vision-language models analyze both textual context and visual well log displays (with truncated text context).
3. **Well Log Plot Only (Single-Formation):** Vision-language models analyze only visual well log displays without textual context.

All models are evaluated across four rubrics: Response Adequacy (40%), Trustworthiness (20%), Style (20%), and Petrophysical Reasoning Quality (20%) and averaged across categories and question types (Single Formation, Cross Formation, and Multi-well).

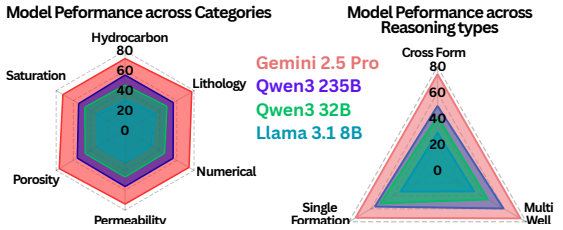


Figure 6: Model performance comparison across categories and reasoning types.

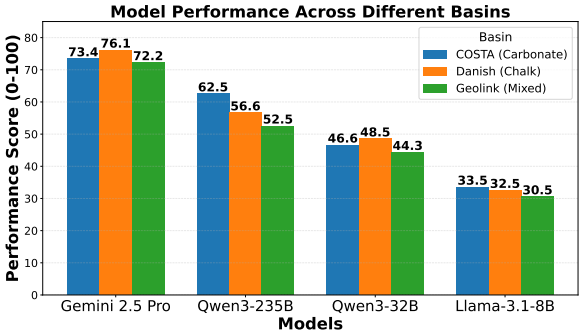


Figure 7: Model performance comparison across basins.

6.1 Text-Only Evaluation

Table 3 presents the performance of 15 base model families (30 variants with truncated/non-truncated contexts).

Closed-source models lead overall: Gemini 2.5 Pro scores highest (74.1), followed by GPT-5 (65.5) and Gemini Flash (60.5). The strongest open-source baseline, Qwen3-235B-Instruct reaches 57.1, trailing Gemini 2.5 Pro by 17.0%.

Performance across basins, categories, and reasoning types. We analyze models across different model sizes on text-only context without any dynamic feature selection. Figure 6 reveals that model performance remains remarkably consistent across both petrophysical categories (6 topics) and reasoning complexity levels (3 question types), with most models exhibiting tight performance bands. This consistency validates WellLogBench's balanced design across the petrophysical domain. Basin characteristics introduce more meaningful variation (Figure 7): Danish yields the highest scores for top-tier models (Gemini 2.5 Pro: 76.3), likely due to homogeneous lithology, while Geolink proves universally challenging across all models due to its high lithology variations (refer to Appendix C). Refer to Appendix D.2 for more detailed ablation.

Context length effects. The majority of models benefit from full context (average: +1.0 point), with top models showing larger gains: Qwen3-

Model	With Dynamic Feature Selection					Without Dynamic Feature Selection				
	RA	Trust.	Style	RQ	Overall	RA	Trust.	Style	RQ	Overall
Gemini 2.5 Pro (Comanici et al., 2025)	**61.6**	**65.1**	76.1	78.4	**68.6**	**69.7**	**71.3**	**78.4**	81.3	**74.1**
GPT-5 (OpenAI, 2025a)	58.1	56.9	58.3	**79.1**	62.1	64.1	60.1	59.5	**79.7**	65.5
Gemini Flash (Doshi, 2025)	57.5	57.8	57.4	71.9	60.4	56.9	58.1	59.4	71.1	60.5
Qwen3-235B-Instruct (Yang et al., 2025)	43.7	46.7	59.9	65.0	51.8	52.7	53.0	59.7	67.2	57.1
Qwen3-Next-80B (Team, 2025)	38.0	40.6	58.0	58.6	46.6	42.7	45.3	58.9	60.0	49.9
Llama-4-109B (Meta, 2025a)	35.2	42.7	64.1	52.4	45.9	36.5	43.9	64.3	52.1	46.7
Qwen3-32B (Yang et al., 2025)	34.9	38.3	56.8	53.3	43.6	39.8	41.8	58.9	54.7	47.0
Llama-3.3-70B (Meta, 2024b)	30.3	42.1	64.1	50.0	43.3	29.9	41.4	62.2	48.9	42.5
Nemotron-49B (NVIDIA, 2025)	36.4	40.6	58.7	56.2	45.6	31.9	38.7	58.5	55.3	43.3
Qwen3-8B (Yang et al., 2025)	33.8	36.9	53.1	50.2	41.6	33.7	38.1	54.5	51.3	42.2
GPT-OSS-120B (OpenAI, 2025b)	43.9	46.4	58.6	62.3	51.0	24.0	31.6	57.3	62.6	39.9
OLMo3-32B (OLMo et al., 2025)	23.8	28.4	52.8	44.5	34.7	18.4	32.0	51.3	44.4	32.9
Llama-3.1-8B (Meta, 2024a)	20.5	27.4	48.0	36.4	30.6	21.8	29.6	49.3	37.4	32.0
OLMo3-7B (OLMo et al., 2025)	19.1	22.8	50.9	39.7	30.3	12.5	19.8	51.2	35.5	26.3
Ministral-3-8B (AI, 2024)	21.6	25.3	26.7	17.8	22.6	24.2	28.0	28.0	19.2	24.7

Note: RA = Response Adequacy; Trust. = Trustworthiness; RQ = Petrophysical Reasoning Quality

Table 3: Text-only model performance (0–100) on the public test set. Models process structured statistical summaries without visual information. Best value per column in **bold**.

Model	Score
Gemini 2.5 Pro (Comanici et al., 2025)	**68.7**
Qwen3 VL 235B (Bai et al., 2025)	49.1
Mistral Large 2512 (AI, 2025a)	49.0
Mistral Small 3.2-24B (AI, 2025b)	43.9
Llama-4 Scout (Meta, 2025b)	41.2
Qwen3 VL 30B (Bai et al., 2025)	40.3
Qwen3 VL 8B (Bai et al., 2025)	33.6

Table 4: Multimodal (Text + Visual) performance on the public test set (60% split).

235B-Instruct (+5.3), Gemini 2.5 Pro (+5.5), GPT-5 (+3.4). However, GPT-OSS-120B degrades sharply (-11.1), followed by OLMo3-7B (-4.0) and Nemotron-49B (-2.3). Context management does not correlate strictly with model size: Llama-3.1-8B (+1.4) and Ministral-3-8B (+2.1) outperform some larger models. Critically, truncated variants achieve performance within 5–6 % of non-truncated versions (Gemini 2.5 Pro: 68.6 vs. 74.1), demonstrating that carefully selected summaries capture critical information with significantly fewer tokens and inference speed as given in Figure 5.

6.2 Multimodal and Visual-Only Evaluation

Table 4 presents multimodal vision-language model (VLM) performance when processing both textual context and visual well log plots. Visual-only results and complete rubric-level analysis are provided in Appendix D.1.

The performance hierarchy—Text-Only Truncated (68.6) $\approx$ Multimodal (68.7) $>$ Visual-Only (39.7)—reveals minimal visual contribution. Gemini 2.5 Pro achieves 68.7 multimodal versus 68.6 text-only (+0.1%), while visual-only drops to 39.7 (-29.0 %). Qwen3-VL-235B follows this pattern: 49.1 multimodal versus 51.8 text-only (-2.7%) and 36.3 visual-only (-12.8%). Rubric-level analysis (Appendix D.1) shows that Response Adequacy and Trustworthiness collapse to 10–24 range in visual-only mode, revealing systematic perceptual failures: VLMs struggle to extract quantitative data from such visual plots (example plot in Appendix B). In multimodal contexts, models fall back to text with negligible visual benefit. For production workflows, text-only models remain superior tradeoffs—multimodal pipelines add computational cost without performance gains.

7 Conclusion

We present WellLogBench, the first comprehensive benchmark for evaluating LLM reasoning on well log data. Our contributions include a carefully curated dataset of 1,085 expert-validated QA pairs, a structured evaluation framework spanning multiple petrophysical categories and reasoning levels, and extensive baseline evaluations revealing significant challenges for current models. Results demonstrate that well log reasoning represents a challenging frontier for AI systems, with even the best models achieving only 74.1% of maximum scores. We hope WellLogBench catalyzes scientific AI research and establishes a rigorous foundation for petrophysics decision-making.

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.

8 Ethical Considerations

Data & Privacy. We utilize strictly public datasets (Geolink, COSTA, GEUS) adhering to open-data licenses. No proprietary or personally identifiable information is included.

Safety & Deployment. Subsurface interpretation carries high operational and environmental risks. With current top models achieving only ∼74% accuracy, AI systems based on this benchmark must function solely as expert assistants with rigorous human oversight; we advise against autonomous deployment.

Fair Labor. All domain experts involved in the curation and validation process were fairly compensated for their contributions.

References

Mistral AI. 2024. Un ministral, des ministraux. Mistral AI Blog.

Mistral AI. 2025a. Mistral-large-3-675b-instruct-2512 (model card). Hugging Face.

Mistral AI. 2025b. Mistral-small-3.2-24b-instruct-2506 (model card). Hugging Face.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 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*.

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.

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.

Tulsee Doshi. 2025. Start building with gemini 2.5 flash. Google Developers Blog.

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. Includes Geolink dataset description and external resource link to the original Geolink well-log download.

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.

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.

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.

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

Meta. 2025b. Llama-4-scout-17b-16e (model card). Hugging Face.

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.

NVIDIA. 2025. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*.

Team OLMo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, et al. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

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.

A Annotation and Experts

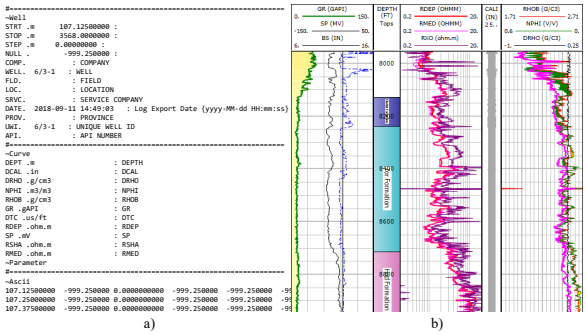
To construct ground-truth labels for our evaluation set, we designed an expert-facing annotation workflow that guided domain specialists through a consistent review of LLM-generated formation interpretations. Experts were presented with each QA item as a self-contained unit: the original question, the 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.

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.

B Modality definitions

LAS (Log ASCII Standard) files (Figure 8-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



a) b)

Figure 8: 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.

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 8-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 9) were created for WellLogBench's evaluation from LAS files using Python. These displays span over the total depth of a formation with 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 log-

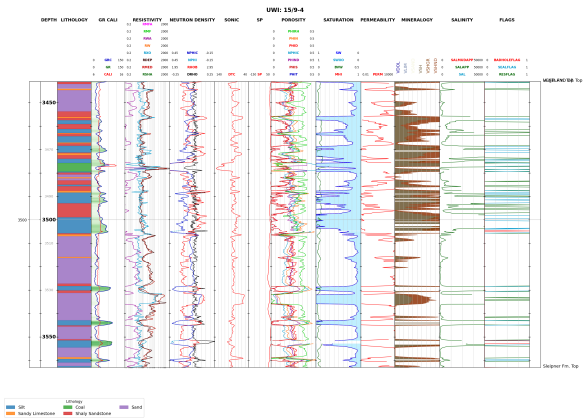


Figure 9: Python generated well log visualization format used in WellLogBench's multimodal evaluation.

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.

arithmic 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.

C 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.

C.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*.

C.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.

D Benchmarking Ablations

D.1 Vision-Language Model Analysis

This section presents comprehensive VLM evaluation across three modalities: text-only (with dynamic feature selection), multimodal (text + visual), and visual-only. Multimodal models use truncated textual context to isolate visual contribution.

D.1.1 Multimodal Performance

Table 6 presents VLM performance when processing both textual context and well log plots. Gemini 2.5 Pro achieves 68.7 overall, essentially matching text-only performance (68.6, +0.1%), indicating minimal visual contribution. Trustworthiness degrades by 9.3 % (65.1 → 55.8), suggesting visual features introduce noise rather than improving data grounding. Open-source VLMs follow this pattern: Qwen3-VL-235B scores 49.1 multimodal versus 51.8 text-only (-2.7%). Style shows modest improvements (Gemini 2.5 Pro: +6.4), but this does not compensate for degraded factual rubrics weighted at 60% of total score.

Model	Resp. Adeq.	Trust.	Style	Reason. Quality	Ovr.
Gemini 2.5 Pro	61.8	55.8	82.5	81.6	68.7
Qwen3-VL-235B	41.4	44.8	56.7	61.1	49.1
Mistral-Large-2512	42.4	43.7	52.9	63.4	49.0
Mistral-Small-3.2-24B	35.4	42.6	53.6	52.6	43.9
Llama-4-Scout	29.9	36.5	61.8	48.1	41.2
Qwen3-VL-30B	33.2	36.2	47.3	51.5	40.3
Qwen3-VL-8B	29.2	33.4	33.6	42.6	33.6

Table 6: Multimodal (Text + Visual) VLM performance (public test set, 60% split).

D.1.2 Visual-Only Performance

Table 7 presents visual-only performance. Response Adequacy and Trustworthiness collapse to 10–24 range, representing -60% degradation from multimodal. In contrast, Style and Reasoning Quality remain viable (56–65 range), demonstrating -20% degradation. This reveals *perceptual* rather than *conceptual* failures: VLMs cannot extract quantitative data from scientific plots but retain logical reasoning capabilities.

Model	Resp. Adeq.	Trust.	Style	Reason. Quality	Ovr.
Gemini 2.5 Pro	24.3	24.4	62.8	62.8	39.7
OpenAI GPT-5	22.6	24.2	60.1	65.1	38.9
Qwen3-VL-235B	20.1	21.5	61.4	58.3	36.3
Mistral-Large-2512	19.8	23.0	56.2	60.9	35.9
Mistral-Medium-3	18.1	20.1	60.2	51.1	33.5
Mistral-Small-3.2-24B	17.9	19.3	63.0	50.0	33.6
Llama-4-Scout	17.9	20.2	62.5	44.6	32.6
Qwen3-VL-30B	16.7	17.4	58.7	52.1	32.3
Qwen3-VL-8B	17.1	18.0	56.0	46.6	30.9
Qwen3-VL-32B	10.7	10.9	33.7	31.3	19.5

Table 7: Visual-only VLM performance (public test set, 60% split).

D.1.3 Rubric-Level Degradation

Figure 10 shows rubric-level comparison between multimodal and visual-only modes across 6 VLMs. Response Adequacy and Trustworthiness show severe degradation (30–40 point drops), while Style and Reasoning Quality exhibit moderate drops (15–20 %). This pattern demonstrates systematic perceptual failures across all architectures—VLMs cannot extract quantitative data from scientific charts—explaining why multimodal approaches fail to improve over text-only baselines. Future VLM development should prioritize scientific chart perception before investing in multimodal fusion architectures.

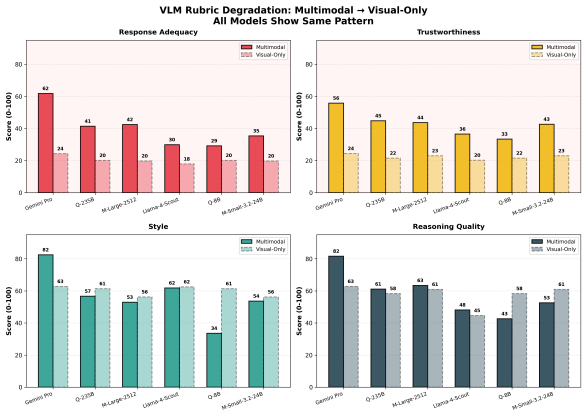


Figure 10: VLM rubric degradation. Each panel shows one rubric with multimodal (dark, solid) vs visual-only (light, dashed) bars. Response Adequacy and Trustworthiness (top row, pink backgrounds) show 30–40 point drops universally. Style and Reasoning Quality (bottom row) show 15–20 point drops.

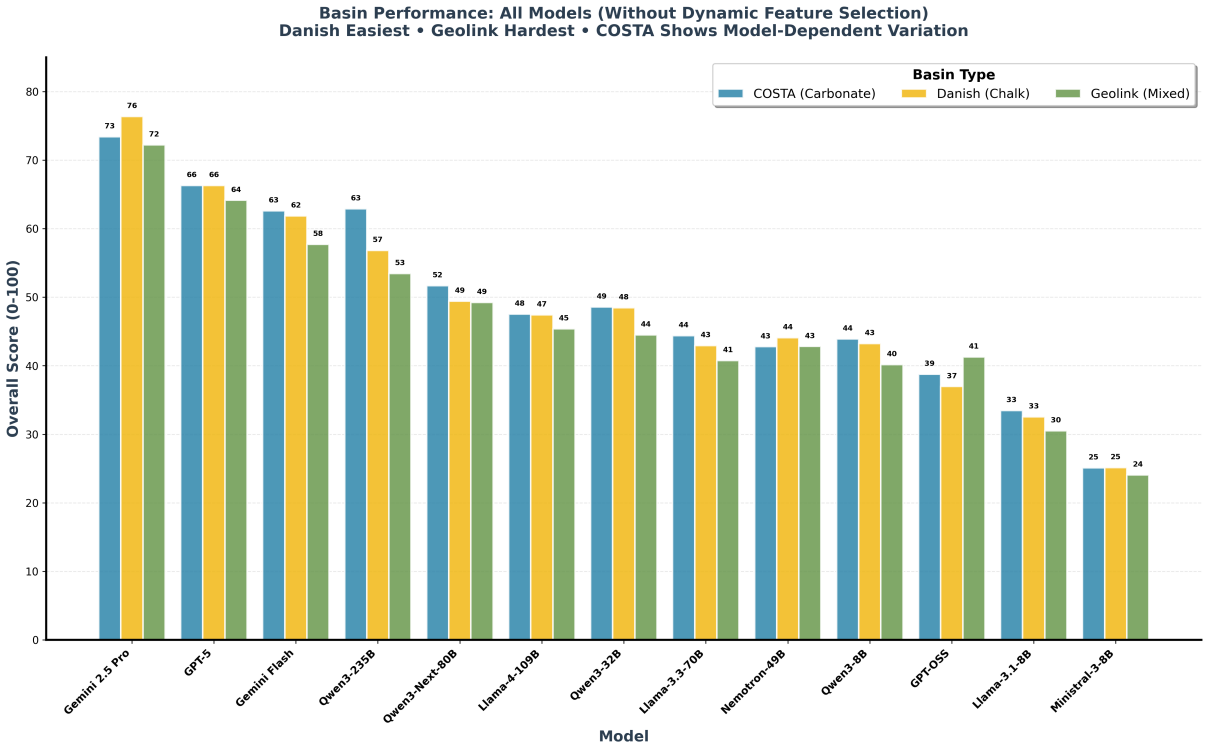


Figure 11: Basin performance for all models (without dynamic feature selection). Danish (Chalk, yellow) is easiest, Geolink (Mixed, green) is hardest. COSTA (Carbonate, blue) shows inverse scaling: open-source models gain disproportionately. Pattern validates balanced benchmark design without systematic geological bias.

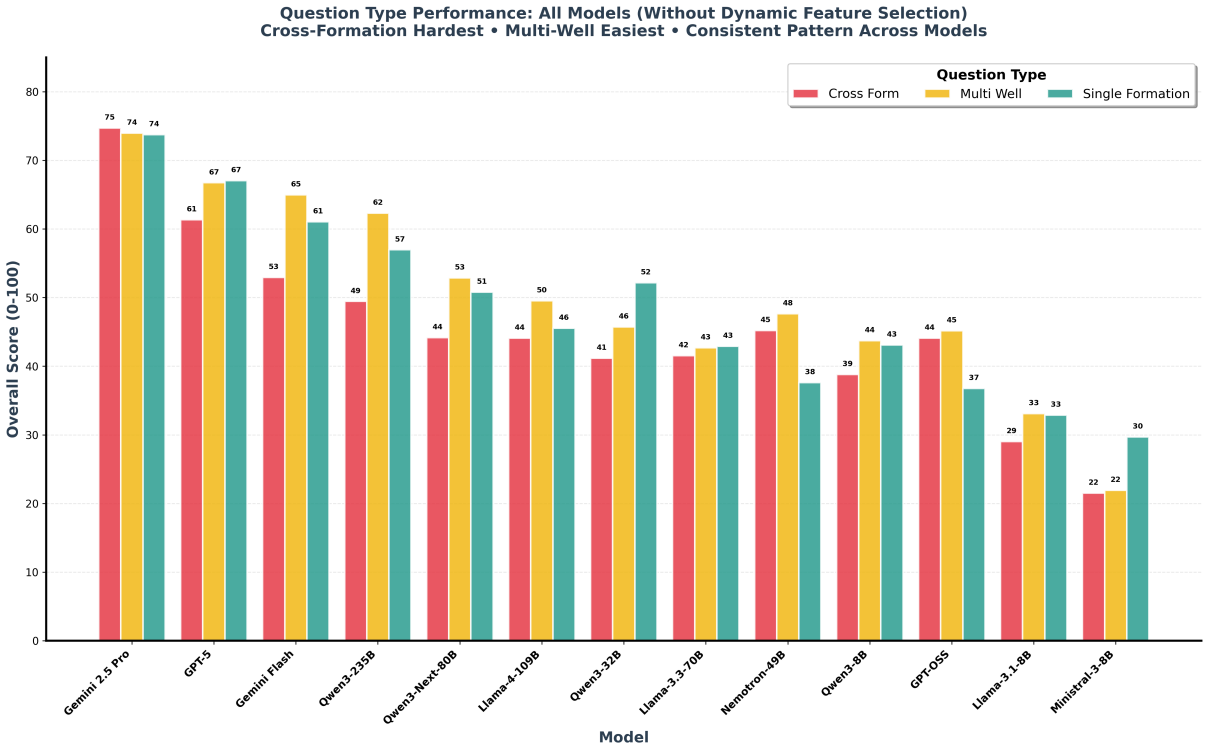


Figure 12: Question type performance for all models (without dynamic feature selection). Cross-Formation (red) is hardest, Multi-Well (yellow) is easiest. Top models show consistent performance across types (narrow spread), while mid-tier models show larger variation. Pattern demonstrates that question complexity affects all models but impacts weaker systems more severely.

D.2 Performance Analysis: Basins and Question Types

This section analyzes model performance across geological basins and question complexity levels using the public test set (60% split, without dynamic feature selection).

D.2.1 Basin-Level Performance

Figure 11 compares all 15 models across three basins. Danish (Chalk) yields highest scores (average: 54.5), likely due to homogeneous lithology. Geolink (Mixed) proves most challenging (average: 50.1) with high lithological variation. COSTA (Carbonate) demonstrates model-dependent behavior: Qwen3-235B-Instruct shows 10.7% boost over Danish (62.9 vs. 56.8), while Gemini 2.5 Pro shows 3.8% decline (73.4 vs. 76.3), suggesting carbonate-specific features favor different reasoning strategies. Model rankings remain stable across basins, but performance gaps vary basin-dependently (10–20 point variations).

D.2.2 Question Type Performance

Figure 12 compares performance across question types. Cross-Formation questions prove hardest universally (Gemini: 74.7, Qwen3-235B-Instruct: 49.4), requiring comparative reasoning across geological contexts. Multi-Well questions are easiest (Gemini: 74.0, Qwen3-235B-Instruct: 62.3), likely due to similar formation patterns enabling pattern recognition. Single-Formation questions fall between (Gemini: 73.7, Qwen3-235B-Instruct: 57.0). Performance remains remarkably consistent across question types for top models (Gemini spread: 1.0 point), while mid-tier models show larger variation (Qwen3-235B-Instruct spread: 12.9 %), indicating that question-type robustness correlates with overall capability.

Key insights. Basin and question-type analysis reveals two patterns: (1) Geological characteristics introduce meaningful variation (Danish vs. Geolink: 4.4-point gap), but model rankings remain stable, validating robust evaluation framework. (2) Question complexity shows universal difficulty ordering (Cross-Formation $>$ Single-Formation $>$ Multi-Well), with top models demonstrating type-invariant performance while weaker models exhibit sensitivity to reasoning requirements. These patterns confirm WellLogBench's balanced design: no single basin or question type dominates difficulty, preventing benchmark overfitting while enabling diagnostic profiling of model capabilities across diverse geological settings and reasoning demands.

D.2.3 Category-Level Performance

Figure 13 presents model performance across six petrophysical categories using spider plots. The dual-panel layout separates top-tier models (left) from mid-tier models (right), revealing distinct performance stratification. Top models (Gemini 2.5 Pro, GPT-5, Gemini Flash) demonstrate large, consistent hexagonal shapes with scores spanning 57–77, indicating robust performance across all petrophysical domains. Mid-tier models (Llama-4-109B through Qwen3-8B) exhibit smaller polygons with scores ranging 42–55, showing proportional capability reduction but maintaining similar shape profiles.

Critically, all models exhibit similar polygon shapes despite different sizes, demonstrating that category-level difficulty remains consistent across capability tiers. No model excels disproportionately on specific categories (e.g., Lithology vs. Saturation), and no category proves universally easier or harder. Top models maintain 2–5 point spreads across categories (Gemini: 72.2–77.4), while mid-tier models show comparable 3–6 point spreads (Qwen3-32B: 45.3–51.5). This pattern validates WellLogBench's balanced category distribution: the benchmark evaluates general petrophysical reasoning rather than narrow specialization, preventing models from achieving high scores through category-specific optimization while failing on others.

E Private Test Set Results

To validate generalization and prevent overfitting, WellLogBench employs a held-out private test set comprising 40% of the benchmark (430 questions). Table 8 presents the complete results on this private split, which was not accessible during model development or hyperparameter tuning.

Public-Private consistency. Performance on the private test set closely mirrors the public set results, with minimal score variations (mean absolute difference: 0.7 %). Gemini 2.5 Pro maintains the top position (73.9 vs. 74.1 public, -0.2 %), followed by GPT-5 (66.2 vs. 65.5 public, $+0.7$ %) and Gemini Flash (61.8 vs. 60.5 public, $+1.3$ %). The best open-source model, Qwen3-235B-Instruct, achieves 56.4 (vs. 57.1 public, -0.7 %). Model rankings remain identical across both splits,

Model	With Dynamic Feature Selection					Without Dynamic Feature Selection				
	Resp. Adeq.	Trust.	Style	Reason. Quality	Overall	Resp. Adeq.	Trust.	Style	Reason. Quality	Overall
Gemini 2.5 Pro	**60.4**	**63.7**	75.5	77.9	**67.6**	**69.4**	**70.9**	**77.9**	81.8	**73.9**
GPT-5	57.1	56.8	58.3	**79.5**	61.8	64.6	61.6	59.9	**80.4**	66.2
Gemini Flash	55.5	55.7	58.3	68.9	58.8	59.0	60.2	59.4	71.4	61.8
Qwen3-235B-Instruct	44.4	46.7	59.4	64.2	51.8	51.4	52.1	60.6	66.7	56.4
Qwen3-Next-80B	36.9	41.5	58.0	58.6	46.4	45.0	46.5	58.4	60.8	51.1
Llama-4-109B	34.9	41.2	63.9	51.9	45.4	36.4	44.4	64.9	52.1	46.8
Qwen3-32B	35.3	38.0	55.7	53.8	43.6	39.0	41.8	58.7	55.5	46.8
Nemotron-49B	34.6	39.3	59.4	55.4	44.7	32.2	38.0	58.7	55.2	43.2
Qwen3-8B	34.8	37.6	54.2	52.5	42.8	33.0	39.1	55.2	51.4	42.4
Llama-3.3-70B	30.9	41.9	63.7	49.7	43.4	29.4	41.3	62.0	48.1	42.0
K2-V2-Instruct	29.6	29.6	55.1	53.5	39.5	32.9	33.6	55.9	52.0	41.5
GPT-OSS-120B	43.2	45.6	57.8	61.5	50.3	19.7	29.1	57.2	56.1	36.4
OLMo3-32B	25.2	28.5	52.8	45.2	35.4	17.4	31.6	51.2	44.7	32.5
Llama-3.1-8B	20.4	27.1	46.8	35.7	30.1	21.7	29.3	49.4	37.1	31.9
OLMo3-7B	18.4	21.5	49.9	41.8	30.0	13.0	20.5	52.5	37.8	27.4
Ministral-3-8B	21.8	26.0	26.6	18.4	22.9	23.7	28.1	27.9	19.3	24.5

Note: Resp. Adeq. = Response Adequacy; Trust. = Trustworthiness; Reason. Quality = Petrophysical Reasoning Quality

Table 8: Text-only model performance (0–100) on the private test set (40% split). Models process structured statistical summaries without visual information. Best value per column in **bold**.

and the tight performance correlation (differences $< \pm 1.3$ % for all top models) indicates robust generalization without overfitting. The consistent behavior across splits validates WellLogBench's reliability as an evaluation framework for geoscientific reasoning.

F Example QA pairs

Q: Where does neutron-density crossover coincide with elevated resistivity within the GRID FM. TOP formation, potentially indicating gas-bearing zones?

F.1 Ground Truth

Formation Overview

In the GRID FM. TOP formation, logs are dominated by shaly siltstones with poor reservoir quality and low resistivity. One upper shaly sandstone interval exhibits anomalously high deep resistivity and low water saturation, indicating hydrocarbon presence. However, a definitive neutron-density gas crossover is not observed, and fluid type remains uncertain.

Detailed Analysis of Gas Indicators

Neutron-Density Response

Across the formation, neutron porosity (PHIN) generally exceeds density-derived porosity (PHID) in the silty, high-shale-volume sections, consistent with a shale effect that masks any gas response. In the only hydrocarbon-indicative zone at the top, the interval-average still shows PHIN > PHID, so a classic gas crossover is not present on average.

- Upper shaly sandstone (1506.7–1519.3 m): Mean PHIN (0.372 V/V) > mean PHID (0.328 V/V), with low shale volume (VSH = 0.142 V/V). Despite favorable lithology and very high resistivity, the average neutron-density response does not display a clear gas crossover. This is compatible with oil charge or a gas effect suppressed by residual shale and/or borehole conditions.
- Non-prospective silt intervals: Large PHIN–PHID separation persists (PHIN » PHID) with high shale volumes (typical mean VSH > ~0.68 V/V), indicating bound-water effects and water saturation dominance.

Neutron-density summary:

Depth Interval (m)	Lithology	Mean PHID (V/V)	Mean PHIN (V/V)	Mean VSH (V/V)	Interpretation
1506.7–1519.3	Shaly Sandstone	0.328	0.372	0.142	No classic gas crossover on average; lithology is relatively clean for the unit.
1519.5–1629.5	Silt	0.298	0.523	0.725	Strong shale effect (PHIN $\gg$ PHID); masks gas response; non-reservoir.
1669.7–1726.9	Silt	0.270	0.502	0.721	Strong shale effect; non-reservoir.
1728.9–1767.9	Silt	0.262	0.498	0.745	Strong shale effect; non-reservoir.

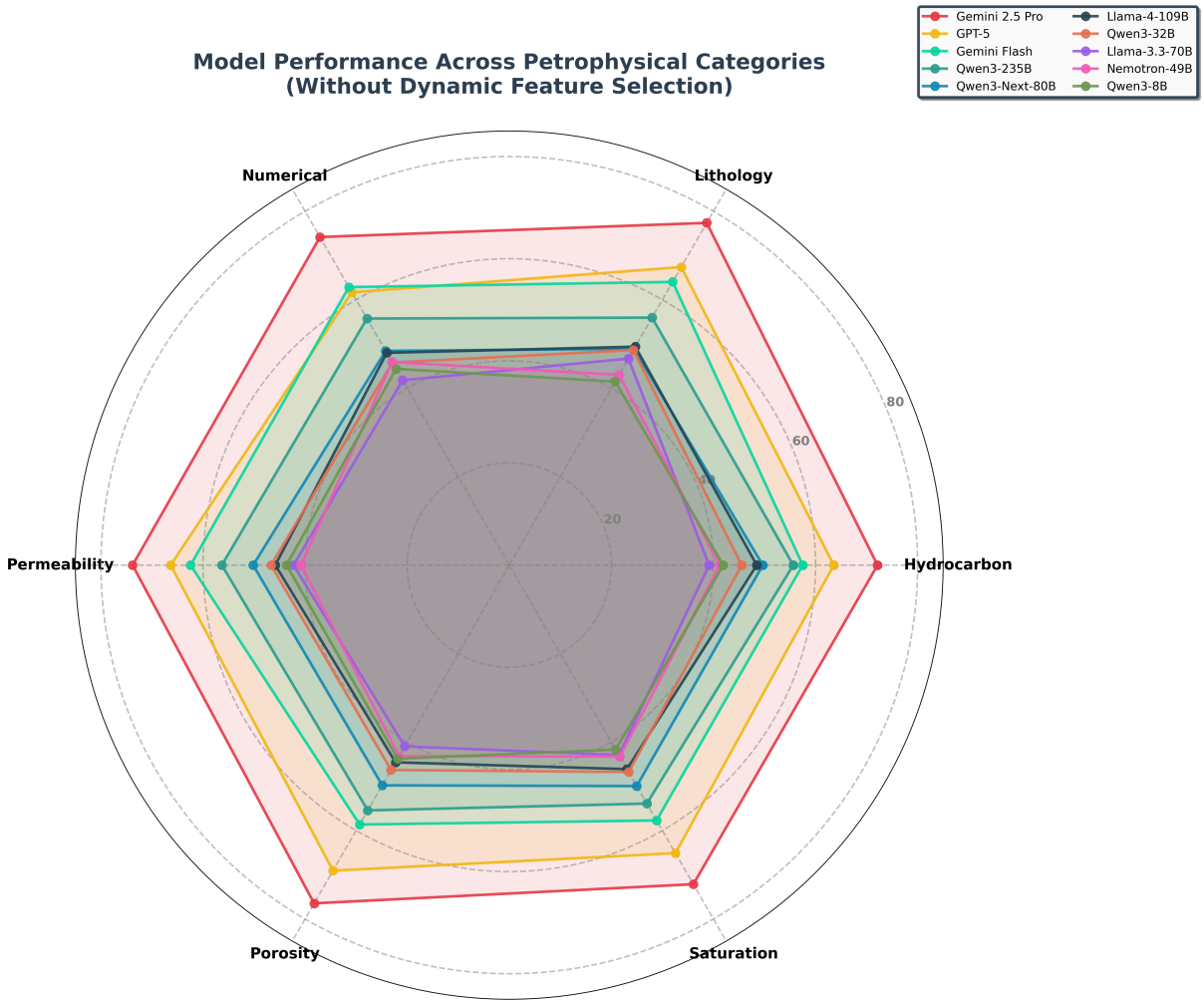


Figure 13: Category-level performance comparison. Left: Top 5 models show large, consistent hexagons (57–77 range). Right: Models 6–10 display smaller but similarly-shaped polygons (42–55 range). All models maintain balanced performance across categories—no single category dominates difficulty—validating WellLogBench's domain coverage without systematic bias toward specific petrophysical topics.

Additional non-prospective intervals referenced: 1629.7–1631.8 m (shaly sandstone), 1632.0–1642.5 m (silt), 1644.0–1646.7 m (silt), 1648.0–1667.8 m (silt).

Resistivity Profile

Resistivity differentiates the only hydrocarbon-bearing candidate from the remainder of the formation:

- Upper shaly sandstone (1506.7–1519.3 m): Deep resistivity is anomalously high (mean RDEP 3.42 ohm.m; maximum 84.54 ohm.m). A minimum calculated water saturation (SW) of 0.13 V/V supports hydrocarbon presence. The best reservoir quality is concentrated within the 1506.7–1517.3 m sub-interval.
- Non-prospective intervals: Deep resistivity is uniformly low (mean RDEP generally < 0.7 ohm.m), consistent with water-bearing, shaly siltstones with high water saturation (mean SW > 0.89 V/V as a group).

Resistivity summary:

Depth Interval (m)	Lithology	Mean RDEP (ohm.m)	Max RDEP (ohm.m)	Interpretation
1506.7–1519.3	Shaly Sandstone	3.42	84.54	Anomalously high resistivity; hydrocarbon-bearing. SW(min) = 0.13 V/V.
1519.5–1629.5	Silt	0.61	–	Consistently low; water-bearing, non-reservoir.
1669.7–1726.9	Silt	0.55	–	Consistently low; water-bearing, non-reservoir.
1728.9–1767.9	Silt	0.62	–	Consistently low; water-bearing, non-reservoir.

Data Limitations and Implications

- Shear sonic (DTS) is absent, preventing VP/VS-based gas discrimination.
- No direct shows/samples (e.g., mud gas, formation fluid sampling) are cited; fluid typing in 1506.7–1519.3 m remains uncertain (oil vs. gas or suppressed gas effect).

Conclusion

A single upper shaly sandstone interval (1506.7–1519.3 m) in GRID FM. TOP exhibits strong hydrocarbon indication from deep resistivity and low water saturation, with best reservoir quality over 1506.7–1517.3 m. However, a formation-wide neutron-density gas crossover is not evident; in the upper interval, mean PHIN remains higher than PHID, consistent with either an oil charge or a gas response masked by lithology. All remaining intervals are high-shale, low-resistivity, water-bearing silts and shaly sandstones with no gas potential.

F.2 Gemini 2.5 Pro

A:

Gas Indication Analysis in GRID FM. TOP Formation

Based on the petrophysical analysis of log data, no intervals within the GRID FM. TOP formation exhibit the classic simultaneous signature of a significant neutron-density crossover and elevated deep resistivity that would strongly indicate a commercial gas-bearing zone. While one interval shows high resistivity, it does not have an accompanying average crossover effect.

Shaly Sandstone Interval: 1506.7-1519.3 m

This upper shaly sandstone unit is the most prospective interval within the formation, characterized by a significant resistivity anomaly, which suggests the presence of hydrocarbons. However, the average porosity log responses do not indicate the presence of gas.

- **Resistivity:** Deep resistivity is anomalously high, reaching a maximum of 84.54 ohm.m, a strong indicator for hydrocarbons when contrasted with surrounding zones.
- **Neutron-Density Response:** There is no average crossover observed. The mean density porosity (PHID) is lower than the mean lithology-corrected neutron porosity (PHIN), which is contrary to the expected response in a gas zone. This suggests the hydrocarbons are likely liquid (oil), or any potential gas effect is suppressed by shale content (average VSH of 0.14 V/V).

Depth Interval (m)	Parameter	Mean Value	Maximum Value	Unit
1506.7–1519.3	RDEP	3.42	84.54	ohm.m
	PHID	0.3278	0.3839	V/V
	PHIN	0.3715	0.5560	V/V

Silt and Deeper Shaly Sandstone Intervals

The remainder of the GRID FM. TOP formation, composed predominantly of silt and a thin shaly sandstone, does not show any indication of gas. These intervals are characterized by low resistivity and a consistent lack of neutron-density crossover.

- **Resistivity:** Deep resistivity values are uniformly low, with mean values typically between 0.5 and 0.6 ohm.m, indicating these zones are likely water-saturated.
- **Neutron-Density Response:** In all these deeper intervals, the average neutron porosity (PHIN) is significantly higher than the average density porosity (PHID). This response is typical of water-filled, shaly formations and confirms the absence of a gas effect.

Depth Interval (m)	Lithology	Mean RDEP (ohm.m)	Mean PHID (V/V)	Mean PHIN (V/V)
1519.5–1629.5	Silt	0.61	0.2975	0.5232
1629.7–1631.8	Shaly Sandstone	0.54	0.3127	0.5306
1632.0–1642.5	Silt	0.51	0.2879	0.5087
1644.0–1646.7	Silt	0.49	0.2931	0.5241
1648.0–1667.8	Silt	0.50	0.2890	0.5029
1669.7–1726.9	Silt	0.55	0.2698	0.5016
1728.9–1767.9	Silt	0.62	0.2617	0.4979

Data Limitations and Further Recommendations

The interpretation is constrained by the absence of key measurements.

- **Shear Sonic Data (DTS):** This log was not available, preventing the calculation of the Vp/Vs ratio. The Vp/Vs ratio is a highly effective fluid indicator, as low values are diagnostic of gas. Its absence reduces the confidence in definitively ruling out thin, gassy stringers that may not produce a pronounced crossover effect.
- **Direct Hydrocarbon Indicators:** No mud log gas shows or formation fluid samples were available for this analysis. This information would be crucial for calibrating the log response and confirming the fluid type within the resistive zone at 1506.7-1519.3 m.

Conclusion

Within the GRID FM. TOP formation, no zones satisfy the combined criteria of neutron-density crossover and elevated resistivity, the classic indicators for a gas-bearing reservoir. The highly resistive shaly sandstone at 1506.7-1519.3 m indicates hydrocarbons, but the lack of crossover suggests these are more likely oil. The remainder of the formation appears to be non-reservoir or water-bearing silt.

F.3 Qwen3-235B-Instruct

A:

Formation Overview

The **GRID FM. TOP** formation exhibits distinct lithological variations with depth, primarily transitioning between **shaly sandstone** and **silt** intervals. Hydrocarbon indication detection and commercial evaluation are centered on identifying zones where neutron-density crossover coincides with elevated resistivity—key indicators of potential gas-bearing reservoirs. Integrated log response analysis reveals that such conditions are most prominent within the **shaly sandstone interval at 1506.7–1519.3 ft**, where favorable porosity, low shale volume, and high resistivity align to suggest the presence of hydrocarbons. In contrast, silt intervals across the formation (1519.5–1629.5 ft, 1632.0–1642.5

ft, 1644.0–1646.7 ft, 1648.0–1667.8 ft, 1669.7–1726.9 ft, and 1728.9–1767.9 ft) exhibit consistently high shale volume (VSH > 0.68), low permeability, and water saturation near 1.0, indicating limited reservoir potential and no significant hydrocarbon indications.
Neutron-density crossover, a classic gas detection signature, occurs when the neutron porosity (PHIN) reads lower than the density-derived porosity (PHID) due to the low hydrogen index of gas. This effect is most pronounced in clean, porous zones. When combined with elevated deep resistivity (RDEP), it strengthens the case for hydrocarbon presence. Commercial pay determination requires not only hydrocarbon indication but also sufficient reservoir quality—defined here by field-specific cutoffs of VSH < 0.40 and effective porosity > 0.15.

Hydrocarbon Indication Detection and Commercial Evaluation

The **shaly sandstone interval at 1506.7–1519.3 ft** stands out as the primary zone of interest. Within this interval, particularly from **1506.7–1517.3 ft**, neutron-density crossover is observed alongside significantly elevated resistivity, suggesting the presence of movable hydrocarbons. The average deep resistivity reaches **3.42 ohm.m**, with peak values exceeding **84 ohm.m**, far above the surrounding shale and silt zones. This resistivity anomaly, coupled with a mean total porosity of **0.330 V/V** and average VSH of **0.142 V/V**, satisfies commercial reservoir quality criteria.
Invasion profile analysis, inferred from the lack of shallow resistivity (RSHAL) and RXO logs, is limited. However, the derived movable hydrocarbon index (MHI) shows values up to **0.87**, with a mean of **0.10**, indicating zones of potential hydrocarbon mobility. Water saturation (SW) averages **0.896**, but with a wide range down to **0.13**, confirming variable hydrocarbon saturation within this interval.
All other intervals, dominated by silt, show no significant neutron-density crossover or resistivity elevation. Their high VSH (>0.68), low permeability (geometric mean < 0.2 mD), and SW near 1.0 indicate water-bearing, non-commercial lithologies.

Fluid Contact and Hydrocarbon Type Discrimination

No definitive gas-water contact (GWC) is identified within the formation due to the isolated nature of the hydrocarbon-bearing zone and lack of continuous saturation trends across lithological boundaries. However, the sharp transition from the high-resistivity shaly sandstone (1506.7–1519.3 ft) to the overlying silt interval (1519.5–1629.5 ft) with uniformly high water saturation suggests a possible fluid contact near **1519.3 ft**.
Hydrocarbon type discrimination is inferred primarily through neutron-density crossover and resistivity behavior. The observed crossover, combined with high RDEP and low SW in the shaly sandstone, strongly suggests **gas** rather than oil, as oil typically does not produce significant neutron-density separation. The absence of sonic compressional and shear logs (DTC and DTS) prevents VP/VS ratio analysis, limiting direct fluid typing via elastic properties.

Commercial Pay Determination and Reservoir Quality

Commercial pay is determined using field-specific cutoffs: **VSH < 0.40**, **PHIT > 0.15**, and **RDEP > 1.0 ohm.m**. Only the **1506.7–1517.3 ft** interval meets these criteria, with 71 out of 84 data points classified as excellent reservoir quality. This zone exhibits high porosity, low shale content, and elevated resistivity, making it the sole commercial pay interval.
The remainder of the formation fails to meet these cutoffs due to excessive shale content and low permeability, rendering them non-commercial despite minor porosity development.

Statistical Analysis

The following table summarizes key petrophysical parameters across the main lithological intervals within **GRID FM. TOP**, focusing on zones relevant to hydrocarbon indication and commercial evaluation.

Depth Interval (m)	Lithology	Parameter	Value	Unit
1506.7–1517.3	Shaly Sandstone	RDEP (mean)	3.42	ohm.m
1506.7–1517.3	Shaly Sandstone	PHIT (mean)	0.330	V/V
1506.7–1517.3	Shaly Sandstone	VSH (mean)	0.142	V/V
1506.7–1517.3	Shaly Sandstone	SW (mean)	0.896	V/V
1506.7–1517.3	Shaly Sandstone	MHI (mean)	0.104	V/V
1506.7–1517.3	Shaly Sandstone	PERM (geometric mean)	0.055	mD
1519.5–1629.5	Silt	RDEP (mean)	0.613	ohm.m
1519.5–1629.5	Silt	PHIT (mean)	0.298	V/V
1519.5–1629.5	Silt	VSH (mean)	0.725	V/V
1519.5–1629.5	Silt	SW (mean)	0.896	V/V
1669.7–1726.9	Silt	RDEP (mean)	0.554	ohm.m
1669.7–1726.9	Silt	PHIT (mean)	0.270	V/V
1669.7–1726.9	Silt	VSH (mean)	0.721	V/V
1669.7–1726.9	Silt	SW (mean)	0.978	V/V

Lithological Variations

Lithology varies systematically with depth:

- **1506.7–1519.3 ft**: Shaly sandstone with excellent reservoir quality in the upper portion (1506.7–1517.3 ft). This interval shows the only significant neutron-density crossover and resistivity elevation.
- **1519.5–1629.5 ft**: Silt with high VSH and low permeability. No hydrocarbon indications.
- **1629.7–1631.8 ft**: Thin shaly sandstone with high VSH (0.69) and low reservoir quality.
- **1632.0–1642.5 ft, 1644.0–1646.7 ft, 1648.0–1667.8 ft**: Silt intervals with consistent log response—high GR, high VSH, low resistivity.
- **1669.7–1726.9 ft and 1728.9–1767.9 ft**: Thicker silt packages with increasing resistivity at depth but still water-saturated and non-commercial.

Neutron-density crossover is only meaningful in porous, clean zones. It is observed in the **1506.7–1517.3 ft** interval where PHIN is systematically lower than PHID, indicating gas effect. In silt zones, PHIN and PHID are compressed due to high clay content, masking any crossover.

Conclusions

- **Neutron-density crossover coincides with elevated resistivity only in the 1506.7–1517.3 ft interval** of the shaly sandstone unit, indicating a potential gas-bearing zone.
- This interval meets commercial pay criteria (VSH < 0.40, PHIT > 0.15, RDEP > 1.0 ohm.m) and exhibits the only significant hydrocarbon indications in the formation.
- All silt intervals show no meaningful crossover or resistivity anomaly and are water-saturated, non-commercial.
- **Critical data limitations** include the absence of shallow resistivity (RSHAL), invaded zone resistivity (RXO), and shear sonic (DTS) logs, which limits invasion profile analysis, mobility assessment, and VP/VS-based fluid typing.
- **Additional external information** such as mud log gas shows, formation fluid samples, or cased hole production logs would be essential to confirm hydrocarbon presence and type in the 1506.7–1517.3 ft interval and assess deliverability.

G Feature List With/Without Dynamic Feature Selection

Example Question: Does the product of porosity and water saturation (BVW) remain constant across different porosity ranges within the GRID FM. TOP, suggesting the reservoir is at irreducible water saturation?

G.1 Without Dynamic Feature Selection

Statistics Features:

PHIN_overall	SWXO_overall
PHIT_overall	CALI_overall
RMED_overall	RW_overall
TMPC_overall	BVW_overall
NPHIC_overall	VSHND_overall
VDOL_overall	VSHGR_overall
RHOB_overall	DRHO_overall
DTC_overall	RMF_overall
PHID_overall	NPHI_overall
VLIM_overall	RDEP_overall
VSH_overall	RESFLAG_overall
GRC_overall	PHIRH_overall
SW_overall	GR_overall
ROP_overall	PHIND_overall
SAL_overall	PHIS_overall
SALAPP_overall	DTS_overall
VSAND_overall	PERM_overall
MHI_overall	BS_overall
PEF_overall	SEALFLAG_overall
lithology_distribution	

Analysis Summary Features:

Water Saturation Estimation Analysis
Bulk Volume Water Porosity Integration
Hydrocarbon Movability Invasion Characteristics
Fluid Contact Detection Indicators
Gas Indicator Analysis Methods
Integrated Saturation Interpretation

G.2 With Dynamic Feature Selection

Statistics Features:

BVW_overall	SW_overall
PHIT_overall	PHID_overall
NPHI_overall	PERM_overall
VSH_overall	GR_overall
RDEP_overall	lithology_distribution

Analysis Summary Features:

Water Saturation Estimation Analysis
Bulk Volume Water Porosity Integration
Integrated Saturation Interpretation

H Prompt Formulations

For the QA Generation pipeline, we craft three separate prompt builders with specialized instructions to ensure quality QA pair creation. We also craft a system prompt for our LLM-as-Judge evaluator, consisting of the fine-grained metrics we use as part of our main rubrics.

H.1 Question Generation

This builder creates a system + user prompt to generate one generic analytical question for a given well and formation, guided by example questions and requirements. The context data is provided, along with optional analysis configuration and lithology intervals. The output will always be of the format `Q: ....`

```
def _build_question_generation_prompt(
    well_name,
    formation_name: str,
    question_categories: Dict[str, Any],
    main_category: str, sub_category: str,
    context_data: Dict[str, Any], analysis_config, use_plots,
) -> Tuple[str, str, bool]:
    """Build the prompt for single-formation question
    ↪  generation with lithology context."""

    # Get example question styles
    {EXAMPLE_Q_STYLE} - Example style of question to generate

    # Extract category data requirements
    {MIN_REQS_LIST_STR} - Minimum data required for
    ↪  generating/answering questions in this category.
    {IDEAL_REQS_LIST_STR} - Ideal Data for generating/answering
    ↪  questions in this category.
```

```
    {ADDITIONAL_INFO_STR} - Additional External Info available.

    # Analysis configuration (controls the answer focus and
    ↪  wording)
    {ANALYSIS_FOCUS}           # e.g., "petrophysical"

    # Enhanced internal guidance when plots are available
    ↪  (COMPLETELY HIDDEN from question output)
    {INTERNAL_VISUAL_GUIDANCE} - Well log plots can be provided
    ↪  to the LLM for generating more sophisticated and
    ↪  insightful questions, but THIS IS NOT REVEALED IN THE
    ↪  QUESTION.

    # Use context data to help generate questions
    {CONTEXT_DATA_JSON_STR}

    # Build system message
    system_message = f"You are an expert petrophysicist
    ↪  generating standard analytical 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
    ↪  analytical questions that any petrophysicist would
    ↪  pose for single-formation {ANALYSIS_FOCUS} analysis
    ↪  considering lithological variations. Never hint at
    ↪  having analyzed the data or knowing specific formation
    ↪  characteristics."

    # Build user prompt
    user_prompt = f"""You are an expert petrophysicist.
    ↪  Generate ONLY ONE analytical question primarily
    ↪  focused on **{ANALYSIS_FOCUS} analysis for the
    ↪  specific formation** **FORMATION: {FORMATION_NAME}**.
{INTERNAL_VISUAL_GUIDANCE}

* (Your Internal Knowledge Only) Minimum Data for this
↪  Category: [{MIN_REQS_LIST_STR}]
* (Your Internal Knowledge Only) Ideal Data for this Category:
↪  [{IDEAL_REQS_LIST_STR}]
* (Your Internal Knowledge Only) Additional External Info:
↪  [{ADDITIONAL_INFO_STR}]

**COMPREHENSIVE FORMATION CONTEXT DATA (Use this rich context
↪  to understand the formation, but DO NOT reveal this context
↪  in your question):**
```json
{CONTEXT_DATA_JSON_STR}
```

{EXAMPLE_Q_STYLE}

{GUIDELINES_BLOCK_FOR_QUESTION_GENERATION}

{EXAMPLES OF GOOD VS BAD QUESTIONS}

{CRITICAL_RESTRICTIONS}

Output ONLY the generated question, prefixed with "Q: ".

Q: """

    return system_message, user_prompt, formation_with_plots
```

H.2 Feature Selection

This builder creates a system + user prompt that asks the LLM to select the most relevant features from a specified section of context data to answer a given question. It provides the list of available features and requires the model to return only a valid JSON array of selected features, including both primary and supportive features, while never inventing features outside the data section.

```
def _build_feature_selection_prompt(
    question: str, section_name: str, available_features:
    ↪  List[str]
) -> Tuple[str, str]:
    """Build the prompt to select features for generating
    ↪  answers based on question and available lithology
    ↪  context."""

    # Convert available features to a formatted string
    available_features_str = "\n".join([f"  - {key}" for key in
    ↪  sorted(available_features)])

    # System message
    system_message = f"""You are an expert in well log and
    ↪  lithology analysis.
Your task is to select the most relevant features from the
↪  {section_name} section of a dataset.
Respond strictly with a valid JSON array of features, with no
↪  extra formatting, text, or explanations.
Example output: ["statistics.PHIS_overall",
↪  "statistics.PHID_overall"]

Selection principle:
- Include all features that could plausibly contribute to
↪  answering the question.
- Do not restrict yourself to only the single most obvious
↪  feature; include secondary or supporting features if they
↪  may provide useful context.
- Stay within the {section_name} section — never invent
↪  features or include from other sections.
"""

    # User prompt
    user_prompt = f"""
You are analyzing a well log analysis JSON dataset to determine
↪  which keys from the {section_name} section are useful for
↪  answering a specific question.

Available Features in {section_name} section:
{available_features_str}

Question to Answer: {question}

Guidelines for selection:
1. Consider what the question is asking and which
↪  lithology-related measurements or derived statistics could
↪  help answer it.
2. Select both direct indicators and supportive features that
↪  might improve interpretability or accuracy.
3. Prefer a reasonably broad but still focused selection —
↪  avoid extreme minimalism, but do not include irrelevant
↪  features.
4. Always return ONLY a valid JSON array of features, nothing
↪  else.

Example format:
["statistics.PHIS_overall",
↪  "statistics.lithology_distribution", "derived.GR_cluster"]
"""

    return system_message, user_prompt
```

H.3 Answer Generation

This builds a system + user prompt to generate a formation-level, data-driven answer to a given question, using formation context data, lithology/depth-level summaries and optional revision instructions. The prompt enforces strict markdown formatting (headings + tables with units), requires depth-interval references, and appends a separate REASONING section. If plots are available, they can be used internally to improve analysis, but must never be mentioned in the final answer. The output format is fixed: `A: ...` followed by `REASONING: ...` and metadata blocks.

```
def _build_answer_generation_prompt(
    question: str,
    formation_name: str,
```

```
    question_categories: Dict[str, Any],
    main_category: str,
    sub_category: str,
    context_data: Dict[str, Any],
    difficulty: str,
    revision_instructions: Optional[str] = None,
    use_plots: bool = False,
    analysis_configs: Optional[Dict[str, Any]] = None,
) -> Tuple[str, str, bool]:
    """
    Build the prompt for single-formation answer generation with lithology-aware depth interval analysis.
    """

    # ---- Extract / derive core prompt variables (placeholders in your snippet) ----
    FORMATION_NAME = formation_name
    WELL_NAME = context_data.get("well_name", "Unknown")

    # Category requirements strings (placeholders from your pipeline)
    MIN_REQS_LIST_STR = "{MIN_REQS_LIST_STR}"
    IDEAL_REQS_LIST_STR = "{IDEAL_REQS_LIST_STR}"
    ADDITIONAL_INFO_STR = "{ADDITIONAL_INFO_STR}"

    # Analysis configuration (use your real config logic)
    ANALYSIS_FOCUS = "{ANALYSIS_FOCUS}"
    KEY_CONCEPTS = "{KEY_CONCEPTS}"
    TECHNICAL_AREAS = "{TECHNICAL_AREAS}"

    # Data sections (placeholders from your pipeline)
    LITHOLOGY_INTERVAL_DATA_SECTION_STR = "{LITHOLOGY_INTERVAL_DATA_SECTION_STR}"
    LOG_UNITS_REFERENCE_STR = "{LOG_UNITS_REFERENCE_STR}"

    # Optional revision guidance (placeholder)
    REVISION_GUIDANCE_STR = revision_instructions or "{REVISION_GUIDANCE_STR}"

    # Plots: internal-only guidance, but keep it centralized and short
    INTERNAL_PLOT_GUIDANCE = ""
    formation_with_plots = bool(use_plots)
    if formation_with_plots:
        INTERNAL_PLOT_GUIDANCE = (
            f"(INTERNAL ONLY) If plots exist for {FORMATION_NAME}, use them only to refine internal "
            f"interpretation of depth-dependent variability; NEVER mention plots/images/visual inspection."
        )

    # ---- Shared constraints (avoid repeating throughout prompt) ----
    HARD_CONSTRAINTS = f"""
HARD CONSTRAINTS:
- Do NOT mention plots, images, charts, or visual inspection (even if used internally).
- Do NOT mention internal data structures, variable/column names, pipelines, workflows, files, or software.
- Do NOT use the forbidden phrases listed below.
"""

    FORBIDDEN_LANGUAGE = """
FORBIDDEN LANGUAGE (NEVER USE):
- "ground truth", "provided data", "analysis results", "data coverage", "input dataset", "computation pipeline"
- "confidence in lithological breakdown", "data structure", "processing workflow", "analysis files"
- "based on detailed analysis of provided information", "data quality from input sources"
- Any reference to input data structure, specific variable names, columns, etc.
""".strip()

    FORMATTING_RULES = """
FORMATTING RULES (MARKDOWN):
- Use markdown headings (##, ###) to organize sections.
- Put multi-value numeric outputs in markdown tables with clear headers that include units.
- Use real newlines in tables (no escaped \\n).
""".strip()

    CORE_ANALYSIS_RULES = f"""
CORE ANALYSIS RULES:
- Always anchor statements to specific depth intervals (no depth-free generalizations).
- If asked for statistics: report values directly per relevant depth interval(s).
- If hydrocarbon-related: emphasize {ANALYSIS_FOCUS}-derived indicators and their depth patterns.
- Prefer derived logs when they can express the answer; avoid raw logs when a derived log suffices.
- Limitations: explicitly note missing logs / gaps per affected depth interval, why they matter, and what external info would improve interpretation.
""".strip()

    OUTPUT_SCHEMA = f"""
OUTPUT FORMAT (STRICT):
A:
## Formation Overview
## Key Depth-Interval Findings
## Statistical Summary (if applicable)
## Limitations & Data Gaps
## Conclusions

REASONING:
- Bullet list explaining methods/notes used (only what is explicitly provided in the stats notes / analysis summary).
- Must reference depth intervals and units when citing values.

MINIMUM_DATA_REQUIREMENTS_FOR_THIS_QUESTION_CATEGORY:
[{MIN_REQS_LIST_STR if MIN_REQS_LIST_STR else 'None'}]

IDEAL_DATA_REQUIREMENTS_FOR_THIS_QUESTION_CATEGORY:
[{IDEAL_REQS_LIST_STR if IDEAL_REQS_LIST_STR else 'None'}]

ADDITIONAL_EXTRA_EXTERNAL_INFORMATION_USEFUL_FOR_THIS_CATEGORY:
[{ADDITIONAL_INFO_STR if ADDITIONAL_INFO_STR else 'None'}]

DIFFICULTY: {difficulty}
""".strip()

    # ---- System message (keep short; no repeated formatting rules) ----
    system_message = (
        "You are an expert petrophysicist providing detailed, data-driven formation evaluation. "
        f"Your focus is {ANALYSIS_FOCUS} / {KEY_CONCEPTS} / {TECHNICAL_AREAS} for formation {FORMATION_NAME}. "
        "Use only the supplied formation interval statistics/notes; do not reference internal tooling."
    )

    # ---- User prompt (single source of truth for rules) ----
    user_prompt = f"""
{REVISION_GUIDANCE_STR}
{INTERNAL_PLOT_GUIDANCE}

TASK:
Answer the question for formation **{FORMATION_NAME}** (well: **{WELL_NAME}**) with primary emphasis on:
- Focus: **{ANALYSIS_FOCUS}**
- Concepts: **{KEY_CONCEPTS}**
- Areas: **{TECHNICAL_AREAS}**

QUESTION:
Q: {question}

LOG UNITS:
{LOG_UNITS_REFERENCE_STR}

{HARD_CONSTRAINTS}
{FORBIDDEN_LANGUAGE}

{FORMATTING_RULES}

{CORE_ANALYSIS_RULES}

CATEGORY REQUIREMENTS:
MINIMUM:
{MIN_REQS_LIST_STR}
IDEAL:
{IDEAL_REQS_LIST_STR}
ADDITIONAL EXTERNAL INFO:
{ADDITIONAL_INFO_STR}

FORMATION INTERVAL DATA:
{LITHOLOGY_INTERVAL_DATA_SECTION_STR}

{OUTPUT_SCHEMA}
""".strip()
```

```
    return system_message, user_prompt, formation_with_plots
```

H.4 LLM-as-Judge: Evaluation Prompt

```
system_prompt = """
You are a domain-specific evaluator for petrophysical well log analysis. Your task is to assess how accurately and concisely the LLM-generated response matches the provided ground truth. Focus strictly on petrophysical reasoning, data alignment, and domain-specific terminology. Penalize over-generation and hallucinated interpretations, even if they appear plausible. The question, ground truth, and LLM response are all provided — consider all three during evaluation.

For each of the following metrics, assign a score from 1 to 10 based on the scoring guidelines AND provide a concise reasoning string explaining your score.

---

### Core Evaluation Metrics

1. **Accuracy (1-10):**
   - Measures correctness of facts and conclusions compared to the ground truth.
   - 9-10: Fully correct; aligns precisely with the ground truth.
   - 7-8: Mostly correct; minor factual or logical issues.
   - 5-6: Some correct elements, but notable inaccuracies.
   - 3-4: Largely incorrect with only a few correct \%.
   - 1-2: Fundamentally inaccurate.
   - **Reasoning should note**: exact \% of agreement/mismatch with ground truth, especially petrophysical interpretations (e.g., pay flags, fluid type, net/gross, shale effects).


2. **Faithfulness (1-10):**
   - Checks whether the response strictly adheres to the ground truth.
   - 9-10: Entirely faithful; no hallucinated or altered facts.
   - 7-8: Mostly faithful; one or two slight deviations.
   - 5-6: Some incorrect or inferred content not in the ground truth.
   - 3-4: Multiple fabricated elements.
   - 1-2: Largely unfaithful or made-up.
   - **Reasoning should note**: where the response extrapolates beyond provided data (e.g., cites curves, cutoffs, or lithologies not present).


3. **Numerical Accuracy (1-10):**
   - Focuses on correctness of all **numerical values common to both** the LLM response and the ground truth (depths, porosity, resistivity, etc.).
   - Additional or alternative statistics in the LLM response are **not penalized** unless they contradict the numerical trends in the ground truth.
   - 9-10: All shared values accurate and consistent.
   - 7-8: Minor rounding or format differences.
   - 5-6: Some mismatches in shared numericals.
   - 3-4: Multiple inaccuracies in overlapping data.
   - 1-2: Most shared numericals incorrect or inconsistent.
   - **Reasoning should note**: matching or conflicting numerical values and trends only for overlapping data \%.


4. **Repetition (1-10):**
   - Evaluates whether ideas or phrases are repeated unnecessarily.
   - 9-10: No repetition; each point stated once and clearly.
   - 7-8: Minor repetition, not distracting.
   - 5-6: Some redundant phrasing.
   - 3-4: Frequent repetition impairs clarity.
   - 1-2: Excessive repetition; poor readability.
   - **Reasoning should note**: where ideas/phrases are repeated and whether repetition adds any value.

5. **Clarity (1-10):**
   - Measures how clearly technical ideas are expressed.
   - 9-10: Very clear; easy to follow even for complex ideas.
   - 7-8: Mostly clear with minor confusion.
   - 5-6: Some unclear or ambiguous parts.
   - 3-4: Often hard to interpret.
   - 1-2: Poorly written or incoherent.
   - **Reasoning should note**: phrasing or structure that aids or hinders comprehension.
---

### Advanced "By Reference" Metrics

6. **Conclusion by Reference (1-10):**
    - Evaluates how well the response's overall conclusions align with the ground truth.
    - 9-10: Conclusions match exactly; consistent logic and end statements.
    - 7-8: Minor tone or phrasing differences.
    - 5-6: Some misaligned or contradictory conclusions.
    - 3-4: Major contradictions to the ground truth.
    - 1-2: Entirely divergent conclusions.
    - **Reasoning should note:** agreement or mismatch in final petrophysical interpretations (e.g., pay presence, fluid type, reservoir quality).

7. **Petrophysical Claims by Reference (1-10):**
    - Evaluates the accuracy of specific petrophysical claims (porosity, shale, saturation) against the ground truth.
    - 9-10: All claims correctly aligned with ground truth.
    - 7-8: Mostly aligned; one or two incorrect trends.
    - 5-6: Several mismatched or missing claims.
    - 3-4: Mostly incorrect or speculative.
    - 1-2: Contradicts ground truth entirely.
    - **Reasoning should note:** correctly or incorrectly reproduced claims.

8. **Data Limitations Score by Reference (1-10):**
    - Measures whether the LLM respects or acknowledges data limitations in the ground truth.
    - 9-10: Fully consistent; acknowledges uncertainties.
    - 7-8: Slightly overconfident or missed one limitation.
    - 5-6: Some disregard for data uncertainty.
    - 3-4: Often extrapolates beyond data limits.
    - 1-2: Ignores stated limitations.
    - **Reasoning should note:** whether uncertain or missing data zones were misused.

9. **Log Utilization by Reference (1-10):**
    - Evaluates whether the LLM used the same or similar log types as the ground truth.
    - 9-10: All key logs correctly referenced and applied (e.g., GR, RHOB, NPHI, Rt).
    - 7-8: Minor omission or one irrelevant log.
    - 5-6: Several missing or misapplied logs.
    - 3-4: Uses mostly incorrect or irrelevant logs.
    - 1-2: Entirely different or fabricated log set.
    - **Reasoning should note:** which logs are correctly or incorrectly referenced compared to the ground truth.

---

### Advanced "Free Thinking" Metrics

10. **Petrophysics Claims (Free Thinking) (1-10):**
    - Evaluates the **technical validity** of the petrophysical statements themselves, independent of the ground truth.
    - 9-10: All interpretations are scientifically sound and defensible.
    - 7-8: Minor conceptual gaps.
    - 5-6: Some plausible but weak or incorrect reasoning.
    - 3-4: Poor or confused technical understanding.
    - 1-2: Fundamentally incorrect physics or logic.
    - **Reasoning should note:** correctness of physical principles (e.g., Archie's law, saturation trends, porosity-resistivity relationships).

11. **Petrophysics Chain of Thought (Free Thinking) (1-10):**
    - Evaluates the **logical progression** from data to interpretation.
    - 9-10: Clear, sequential reasoning (e.g., GR → shale → porosity → saturation → pay).
    - 7-8: Mostly coherent but slightly disjointed.
    - 5-6: Some missing reasoning steps.
    - 3-4: Disorganized or inconsistent logic.
    - 1-2: Illogical or circular reasoning.
    - **Reasoning should note:** how well reasoning steps connect and justify conclusions.

### Additional Instructions
```

```
- Reread both the ground truth and the response carefully
↪  before scoring.
- Compare interpretations, trends, and numerical conclusions
↪  rigorously.
- Prioritize correctness and alignment over verbosity or
↪  surface fluency.
- If the response is empty, null, or contains only whitespace:
↪  set all scores to 0 and each reasoning string to "No
↪  content to evaluate."
- Keep reasoning strings concise (1-3 sentences, <=50 words).
- Avoid quoting large spans of text; identify mismatches or
↪  omissions directly.

---

### Output Format

Return ONLY the following JSON dictionary and ensure that it is
↪  valid JSON:

{
  "Accuracy": float,
  "Accuracy Reasoning": "string",
  "Faithfulness": float,
  "Faithfulness Reasoning": "string",
  "Numerical Accuracy": float,
  "Numerical Accuracy Reasoning": "string",
  "Repetition": float,
  "Repetition Reasoning": "string",
  "Clarity": float,
  "Clarity 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",
  "Overall Score": float
}
"""
```

I Petrophysical Pipeline and Context Engineering

I.1 Stage 1: Data Ingestion and Quality Control

Raw logs are standardized to ensure numerical stability across solvers.

I.1.1 Sentinel Handling

All NaN values or missing data points are mapped to a global sentinel value of -999.25 to ensure consistent exclusion from statistical aggregations.

I.1.2 Unit Standardization

- **Neutron Porosity (NPHI):** If values > 1.0 (percentage), divided by 100 to convert to decimal fraction (v/v).
- **Sonic (DTC/DTS):** Renamed to canonical mnemonics. DTE or DT automatically renamed to DTS.

I.1.3 Automated QC (Bad Hole Detection)

A BADHOLEFLAG is computed by the WellboreQC class. The flag triggers ($= 1$) when:

$$(CALI - BS) > 1.0 \text{ inches} \quad \text{OR} \quad |DRHO| > 0.05 \text{ g/cm}^3. \tag{3}$$

When triggered, all lithology-dependent calculations default to sentinel values to avoid artifacts.

I.2 Stage 2: Environmental and Fluid Modeling

To support resistivity-based saturation equations, environmental parameters are synthesized when measurements are absent.

I.2.1 Temperature Gradient

If borehole temperature log (TMPC) is missing, a North Sea regional gradient is applied:

$$T_{form}(^\circ C) = 0.0324 \times \text{Depth (m)} + 14.0. \tag{4}$$

I.2.2 Formation Water Salinity and Resistivity

Formation water salinity is depth-stratified:

$$Sal_w = \begin{cases} 80,000 \text{ ppm}, & \text{if } D < 1525 \text{ m}, \\ 80,000 \text{ ppm}, & \text{if } 1525 \leq D < 2300 \text{ m}, \\ 110,000 \text{ ppm}, & \text{if } D \geq 2300 \text{ m}. \end{cases}$$

Formation water resistivity via Arps equation:

$$R_w = \left(0.031 + \frac{3647.5}{Sal_w^{0.955}}\right) \times \frac{23.8 + 21.5}{T_{form} + 21.5}.$$

I.2.3 Mud Filtrate Resistivity

Mud filtrate resistivity computed with constant salinity of 50,000 ppm (Synthetic Oil-Based Mud):

$$R_{mf} = \left(0.031 + \frac{3647.5}{50000^{0.955}}\right) \times \frac{23.8 + 21.5}{T_{form} + 21.5}.$$

I.3 Stage 3: Lithology-Dependent Calculations

Calculations are performed row-by-row. Critical to this stage is the **Lithology Logic**: The physics engine dynamically loads constants based on ground-truth lithology labels (LITHOLOGY). If lithology labels are NaN or missing, the system defaults to **Sandstone** parameters to ensure computational continuity.

Lithology	ρ_{ma}	Δt_{ma}	m	n	Perm Slope	Perm Int
Sandstone	2.65	56.0	1.8	1.9	28.0	-2.0
Limestone	2.71	49.0	2.0	2.0	20.0	-3.0
Dolomite	2.85	44.0	2.2	2.2	18.0	-4.0
Anhydrite	2.98	50.0	2.0	2.0	1.0	-2.0
Coal	1.47	60.0	1.8	1.9	1.0	-4.0
Calc. Sand	2.68	52.5	1.9	1.95	24.0	-2.5

Table 9: Lithology-Specific Physics Constants (Defaults to Sandstone if NaN).

Depth Interval (m)	C_p (Compaction)	C_{sh} (Shale)
$1380 < D < 2656$	0.67	0.16
$2656 < D < 2784$	0.67	0.04
$D > 2784$	0.67	0.28
Default	0.67	0.24

Table 10: Depth-Dependent Sonic Scalars.

I.3.1 Lithology Parameter Selection

I.3.2 Shale Volume Calculation

Shale volume computed from hole-size corrected gamma ray and neutron-density separation:

$$GR_c = GR \cdot \frac{0.0009 \cdot CALI^2 + 0.046 \cdot CALI + 0.621}{1.08},$$

$$V_{sh,GR} = \min(1.0, \max(0.001, \frac{GR_c - 40}{140 - 40})),$$

$$V_{sh,ND} = \min(1.0, \max(0.001, \frac{\phi_N - \phi_D}{0.32})),$$

$$V_{sh} = \begin{cases} V_{sh,ND}, & \text{if } V_{sh,ND} \geq 0, \\ V_{sh,GR}, & \text{otherwise} \end{cases}$$

I.3.3 Porosity Calculation

Total porosity uses priority logic:

Density Porosity

$$\phi_D = \frac{\rho_{ma} - \rho_{corr}}{\rho_{ma} - 1.0}. \tag{5}$$

Neutron Porosity

$$\phi_N = \phi_{N,log} + \Delta\phi_{lith}. \tag{6}$$

Neutron-Density Crossplot

$$\phi_{ND} = \sqrt{\frac{\phi_D^2 + \phi_N^2}{2}}. \tag{7}$$

Selection Logic

$$\phi_t = \min(\phi_D, \phi_{ND}, 0.40). \tag{8}$$

Bad Hole Override: If `BADHOLEFLAG=1`, sonic porosity is used:

$$\phi_S = C_p \left(\frac{\Delta t - \Delta t_{ma}}{\Delta t_{fl} - \Delta t_{ma}} \right) - C_{sh} V_{sh} \tag{9}$$

I.3.4 Water Saturation

Standard Archie equation for uninvaded (*Sw*) and invaded (*Sxo*) zones:

$$S_w = \min \left(1.0, \left(\frac{a \cdot R_w}{\phi_t^m \cdot R_{dep}} \right)^{1/n} \right), \tag{10}$$

$$S_{xo} = \max \left(S_w, \min \left(1.0, \left(\frac{a \cdot R_{mf}}{\phi_t^m \cdot R_{xo}} \right)^{1/n} \right) \right), \tag{11}$$

where $a = 1.0$, m and n are lithology-dependent. If $R_{xo} < 0.01$, system substitutes *Rshal*, then *Rmed*.

I.3.5 Permeability Estimation

Lithology-calibrated exponential relationship with shale penalty:

$$\log_{10}(k) = \min\left(4.5, \text{Slope} \cdot \phi_t \cdot (1 - 1.2 \cdot V_{sh}) + \text{Intercept}\right)$$

Parameters vary by lithology. Maximum permeability capped at $10^{4.5}$ mD.

I.4 Stage 4: Advanced Attributes and Interpretation

Derived attributes quantify reservoir quality, geomechanics, and fluid distribution.

I.4.1 Bulk Volume Water

$$BVW = \phi_t \cdot S_w. \tag{12}$$

I.4.2 Moveable Hydrocarbon Index

$$MHI = \frac{S_{xo} - S_w}{S_{xo}}. \tag{13}$$

Only computed when *Rxo* data is available; not estimated artificially.

I.4.3 Apparent Fluid Salinity

Back-calculated from:

$$R_{wa} = R_{dep} \cdot \phi_t^m / a. \tag{14}$$

Temperature-corrected at 23.8°C reference:

$$R_{wa,238} = R_{wa} \times \frac{T_{form} + 21.5}{23.8 + 21.5}. \tag{15}$$

Salinity recovery:

$$Sal_{app} = \left(\frac{3647.5}{R_{wa,238} - 0.0123} \right)^{1/0.955}. \tag{16}$$

I.4.4 Rock Quality Indicators

$$RQI = 0.0314\sqrt{\frac{k}{\phi_t}}, \quad FZI = \frac{RQI}{\phi_t/(1-\phi_t)}. \tag{17}$$

I.4.5 Geomechanical Properties

- $V_p = 10^6/(DTC \times 0.3048)$ [ft/s] when DTC missing,
- $E = \rho_b \times V_p^2 \times 10^{-6}$ [GPa],
- $\nu = 0.25$ (default) when shear unavailable.

I.4.6 Electrofacies Clustering

K-Means clustering ($k = 5$) on standardized vectors:

$$\text{Features} = [\rho_b, \phi_N, GR, R_{dep}]. \tag{18}$$

I.5 Classification and Flagging Logic

I.5.1 Hydrocarbon Indicators

POTENTIAL_PAY: $V_{sh} < 0.30$ AND $\phi_e > 0.10$ AND $S_w < 0.60$

MOVABLE_HC: $MHI > 0.4$ AND $S_w < 0.6$

STRONG_GAS: Crossover ($\phi_D - \phi_N > 0.01$) AND $V_p/V_s < 1.75$

I.5.2 Reservoir Quality

EXCELLENT: $\phi > 0.20$ OR $k > 1000$ mD

VERY_GOOD: $100 < k \leq 1000$ mD

GOOD: $0.12 < \phi \leq 0.20$ OR $10 < k \leq 100$ mD

FAIR: $0.06 < \phi \leq 0.12$ OR $1 < k \leq 10$ mD

POOR: $\phi \leq 0.06$ OR $k \leq 1$ mD

I.5.3 Flow Units (FZI-Based)

MEGA_PORES: $FZI \geq 5$

MACRO_PORES: $2 \leq FZI < 5$

MESO_PORES: $0.5 \leq FZI < 2$

I.5.4 Data Quality

CARBONATE_CONFIRMED: Ground truth matches log signature ($PEF > 2.8, \rho_b > 2.6$)

LOG_CONSISTENCY: Score ≥ 0.75 indicates excellent match

DATA_COVERAGE: `EXCELLENT` if $\geq 80\%$ logs have valid data

TREND_ZONE: Triggered if ≥ 3 logs show gradients ($|dLog/dz| > 0.1$)

J Specialized Analysis Classes

J.1 HydrocarbonAnalysis

Extends `BasePetroAnalysis` with multi-dimensional hydrocarbon assessment.

J.1.1 Crossover Flag

Calculated from lithology-corrected neutron-density separation:

$$\text{Crossover} = \phi_D - \phi_N > 0.01 \tag{19}$$

Uses lithology-corrected neutron which automatically accounts for different lithologies.

J.1.2 Pay Flag

Standard petrophysical cutoffs:

$$\text{PAY_FLAG} = (V_{sh} < 0.30) \wedge (\phi_e > 0.10) \wedge (S_w < 0.60) \tag{20}$$

J.1.3 VP/VS Ratio Assessment

Calculated from measured DTS when available. For gas detection: $V_p/V_s < 1.75$ indicates gas.

J.1.4 Integrated Classification

PRIME_GAS_TARGET: Strong gas indication with excellent/good reservoir

PRIME_OIL_TARGET: Potential pay in excellent reservoir

POTENTIAL_PAY: Meets pay flag criteria

GAS_SHOW: Any gas indication

MARGINAL: All other combinations

J.2 PorosityAnalysis

Extends `BasePetroAnalysis` with porosity-focused assessment.

J.2.1 Porosity Quality Distribution

Classification based on total porosity values:

EXCELLENT_POROSITY: $\phi > 0.20$ (>20

GOOD_POROSITY: $0.12 < \phi \leq 0.20$ (12-20

FAIR_POROSITY: $0.06 < \phi \leq 0.12$ (6-12

POOR_POROSITY: $\phi \leq 0.06$ (<6

J.2.2 Pore System Architecture

Density-sonic separation analysis:

$$DS_\text{Separation} = \phi_D - \phi_S \tag{21}$$

PRIMARY_POROSITY: $|DS| \leq 0.05$

SECONDARY_POROSITY: $\phi_D > \phi_S + 0.05$

MATRIX_EFFECT: $\phi_S > \phi_D + 0.05$

J.2.3 Clay Content Impact

CLEAN_FORMATION: $V_{sh} < 0.25$

MODERATE_CLAY_IMPACT: $0.25 \leq V_{sh} \leq 0.50$

SHALY_FORMATION: $V_{sh} > 0.50$

J.2.4 Flow Unit Classification

EXCELLENT_FLOW_UNIT: $\phi > 0.15$ AND $k > 10$ mD

GOOD_FLOW_UNIT: $\phi > 0.10$ AND $k > 1$ mD

FAIR_FLOW_UNIT: $\phi > 0.06$ AND $k > 0.1$ mD

POOR_FLOW_UNIT: Below fair criteria

J.2.5 Integrated Assessment

EXCELLENT_RESERVOIR: Excellent flow unit with excellent/good porosity

GOOD_RESERVOIR: Good flow unit with good+ porosity

MARGINAL_RESERVOIR: Fair porosity with any flow unit

TIGHT_FORMATION: Poor flow unit

J.3 SaturationAnalysis

Extends `BasePetroAnalysis` with multi-dimensional saturation assessment.

J.3.1 Water Saturation Estimation

Primary saturation quality classification:

EXCELLENT_HC_SATURATION: $S_w < 0.20$

GOOD_HC_SATURATION: $0.20 \leq S_w < 0.40$

FAIR_HC_SATURATION: $0.40 \leq S_w < 0.60$

WATER_BEARING: $S_w \geq 0.60$

Resistivity context via R_{dep} values: > 50 Ωm (very high), 10–50 Ωm (high), < 2 Ωm (low, water-bearing).

J.3.2 Bulk Volume Water Analysis

BVW quality classification (formation-dependent):

EXCELLENT_MOVEABLE_HC: $BVW < 0.03$

GOOD_MOVEABLE_HC: $0.03 \leq BVW < 0.06$

FAIR_MOVEABLE_HC: $0.06 \leq BVW < 0.10$

J.3.3 Hydrocarbon Movability

Moveable Hydrocarbon Index classification:

HIGH_MOVEABLE_HC: $MHI > 0.50$

MODERATE_MOVEABLE_HC: $0.30 \leq MHI \leq 0.50$

LOW_MOVEABLE_HC: $0.10 \leq MHI < 0.30$

MINIMAL_MOVEABLE_HC: $MHI < 0.10$

Resistivity invasion ratio fallback (when MHI unavailable):

$$\text{Invasion Ratio} = \frac{R_{xo}}{R_{dep}}. \tag{22}$$

CONDUCTIVE_INVASION: Ratio < 0.7

MINIMAL_INVASION: $0.7 \leq$ Ratio ≤ 1.3

RESISTIVE_INVASION: Ratio > 1.3

J.3.4 Fluid Contact Detection

Methodology based on depth-dependent changes in ϕ_t and S_w:

POTENTIAL_CONTACT_ZONE: $\Delta S_w > 0.20$ with $\Delta\phi_t < 0.05$

TRANSITION_ZONE: $0.10 < \Delta S_w \leq 0.20$ with $\Delta\phi_t < 0.05$

STABLE_HC_ZONE: $S_w < 0.40$ consistent with depth

STABLE_WATER_ZONE: $S_w > 0.60$ consistent with depth

J.3.5 Gas Indicator Analysis

Neutron-density crossover using lithology-corrected neutron:

$$ND_\text{Separation} = \phi_D - \phi_N \tag{23}$$

STRONG_GAS_EFFECT: $ND > 0.15$

MODERATE_GAS_EFFECT: $0.08 \leq ND \leq 0.15$

WEAK_GAS_EFFECT: $0.03 \leq ND < 0.08$

NO_GAS_EFFECT: $ND < 0.03$

Additional indicators: $DRHO < -0.05$ (density response), $DTC > 120$ with $S_w < 0.50$ (sonic response).

Gas-corrected total porosity:

$$\phi_t^{\text{gas}} = \min\left(0.45, \sqrt{\frac{\phi_D^2 + \phi_N^2}{2}}\right). \quad (24)$$

J.3.6 Integrated Saturation Interpretation

Hierarchical classification combining all indicators:

GAS_ZONE: Strong gas effects

POSSIBLE_GAS_ZONE: Moderate gas effects

EXCELLENT_HC_ZONE: Excellent saturation with good BVW

GOOD_HC_ZONE: Good saturation with acceptable BVW

MARGINAL_HC_ZONE: Fair saturation

WATER_ZONE: Water-bearing

UNDETERMINED: Insufficient data

K Implementation Details

K.1 NaN Lithology Handling

When `LITHOLOGY` is `NaN`:

1. All lithology-dependent parameters default to sandstone
2. Calculations proceed using sandstone constants
3. Results are valid but represent sandstone-equivalent properties

K.2 Row-by-Row Processing

The `_calculation_engine` method processes each row independently with:

- Sentinel value replacement for missing inputs
- Lithology parameter lookup via `LithologyMapper`
- Sequential calculation of all properties
- Return of complete property set per row

K.3 Fallback Methods

When primary log unavailable:

- *RXO*: Substitute *Rshal* $\rightarrow$ *Rmed*
- *Porosity*: Sonic override in bad hole conditions
- *VP/VS*: Use measured DTS when available

K.4 Results Structure

Output dictionary contains:

- **derived_logs_computed**: List of logs actually created
- **data_completeness**: Missing primary logs, computation notes
- **statistics**: Per-log statistics (min, max, mean, median, stdev)
- **analysis_summary**: Classification intervals and quality assessments