Review

Not peer-reviewed version

Large Language Models and Foundation Models for Petroleum Engineering: A Survey

Rong Lu *

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1814.v1

Keywords: large language models; foundation models; petroleum engineering; survey; upstream oil and gas; generative AI; agentic AI; retrieval-augmented generation; subsurface foundation model; domain adaptation

 doi:10.20944/preprints202604.1814.v1

Review

Large Language Models and Foundation Models for Petroleum Engineering: A Survey

Rong Lu

Independent Researcher, USA; rlu@mines.edu

Abstract

Large language models (LLMs) and foundation models (FMs) are reshaping petroleum engineering at a pace no previous wave of artificial intelligence has matched. Between 2022 and 2026 the field went from zero petroleum-specific LLMs to eighteen domain-specialized models, more than a dozen subsurface foundation models, and more than twenty commercial industry platforms, while annual publication counts grew more than five-fold from 2020 to 2024. This survey integrates those developments into a single framework. We analyze **296 verified references** spanning **2003–2026** across **14 thematic areas** and **six petroleum sub-disciplines plus one cross-cutting category** (geophysics, drilling, reservoir, production, petrophysics, completions, and cross-cutting), from classical natural-language-processing baselines through today's vision–language models, retrieval-augmented generation stacks, and autonomous agents. Our organizing contributions include (i) a positioning matrix against 25 prior surveys, (ii) a bubble-plot taxonomy of sub-disciplines against AI paradigms, (iii) seven application-category tables, six additional thematic tables, and a dedicated maturity-model table (fourteen tables in total), (iv) a catalog of public petroleum AI systems and enabling substrate, and (v) the **PetroLLM Maturity Model** — a five-level scaffold (L1 Conversational Q&A, L2 Document Intelligence and Retrieval, L3 Domain-Specialized LLMs, L4 Autonomous Agents and Copilots, L5 Self-Improving Foundation-Model Ecosystems) that situates every surveyed system on a common ladder. The paper closes with a bibliometric snapshot (trends, sub-discipline distribution, method distribution, institutional footprint) and an open-research agenda spanning data, benchmarks, physics integration, safety, multilinguality, and standards. Our headline findings: geophysics leads, reservoir and production lag, petroleum benchmarks are scarce, industry deployments outpace academic publication, and L5 self-improving ecosystems remain aspirational but within a realistic 2030 horizon.

Keywords: large language models; foundation models; petroleum engineering; survey; upstream oil and gas; generative AI; agentic AI; retrieval-augmented generation; subsurface foundation model; domain adaptation

1. Introduction

1.1. The AI Revolution in Petroleum Engineering

In October 2025, at the Society of Petroleum Engineers Annual Technical Conference and Exhibition in Houston, Saudi Aramco, SPE, and i2k Connect released EnergyLLM (SPE-228097-MS) — a Llama 3-derivative adapted on SPE technical content — and reported, through double-blind subject-matter-expert evaluation, statistically significant wins over GPT-4o on petroleum question answering (Eckroth et al., 2025). Nearly two years earlier, the Chinese Academy of Sciences and Shanghai Jiao Tong University had released K2 at WSDM 2024, the first general-purpose geoscience LLM, built by continuing pretraining of LLaMA-7B on 5.5 billion tokens of geoscience literature and evaluating against a purpose-built benchmark called GeoBench (Deng et al., 2024). By early 2025, ADNOC and AIQ had completed a proof-of-concept for ENERGYai — a seventy-billion-parameter sector LLM whose first scalable release was planned for H1 2025 with five agents to be test-deployed across selected upstream assets, while a March 2025 contract set a staged rollout across ADNOC's upstream value

 doi:10.20944/preprints202604.1814.v1

chain (ADNOC, 2025; AIQ, 2025) — and Saudi Aramco announced its 250 B-parameter METABRAIN system trained on publicly described multi-decade Aramco data (Aramco, 2024,2; Aramco Europe, 2025), it had become untenable to describe the arrival of large language models (LLMs) and foundation models (FMs) in petroleum engineering as marginal or speculative any longer.

The shift unfolded in three waves. The *chat wave* (late 2022 through 2023) brought ChatGPT to the desk of every petroleum engineer; the first rigorous empirical probe, Ogundare et al. (2023), established both the promise (GPT-4 answered conceptual petroleum questions surprisingly well) and the risk (it confidently hallucinated quantitative physics). The *domain-adaptation wave* (2024–2025) saw the first petroleum-specialized LLMs — K2 (Deng et al., 2024), GeoGalactica (Lin et al., 2024), JiuZhou (Chen et al., 2025b), and, by 2025, EnergyLLM (Eckroth et al., 2025) — together with the first subsurface foundation models — SFM of Sheng et al. (2025), SeisCLIP (Si et al., 2024), StorSeismic (Harsuko and Alkhalifah, 2022) — built by transferring masked-modelling (MAE (He et al., 2022) and BERT-style trace masking) and contrastive (Radford et al., 2021) recipes from general computer vision and NLP to seismic volumes and well logs. The *industrial-deployment wave* (2025–2026) is now unfolding: SLB's Tela agentic assistant (SLB, 2025), Baker Hughes's Repsol Leucipa autonomous-production platform (Baker Hughes, 2025), Stone Ridge's ENVOY reservoir-simulation copilot (Wiegand et al., 2024a,2), and Geo-RAG for exploration archives (Dong et al., 2024) are being integrated into operator workflows that, a decade earlier, would have been unrecognisable.

This is not an incremental step. Petroleum engineering has embraced machine learning before — classical support-vector machines for lithology in the 2000s (Rahmanifard and Plaksina, 2019), convolutional networks for seismic facies in the 2010s (Dramsch and Lüthje, 2018; Koroteev and Tekic, 2021; Kuang et al., 2021; Sircar et al., 2021; Tariq et al., 2021; Wu et al., 2019; Zhu and Beroza, 2019) — but each of those waves produced task-specific models trained from scratch on task-specific labels. LLMs and FMs invert that premise: they pretrain once, on scales that no single petroleum company can muster, and then adapt cheaply to every downstream task the enterprise cares to pose.

1.2. Why LLMs and Foundation Models Are Different

Classical petroleum machine learning, surveyed comprehensively by Rahmanifard and Plaksina (2019), Tariq et al. (2021), Koroteev and Tekic (2021), Sircar et al. (2021), Kuang et al. (2021), Yu and Ma (2021), Mousavi and Beroza (2022), Bahaloo et al. (2023), and Xu et al. (2022), shares three chronic pathologies. First, it is *task-specific*: a fault-segmentation network cannot classify facies, a lithology classifier cannot summarise a drilling report, and transferring between basins often requires retraining. Second, it is *data-hungry*: every new task demands thousands to millions of labelled examples, and label cost in petroleum — where a geophysicist's time is expensive and an expert interpretation may be proprietary — is the binding constraint. Third, it is *brittle* to distribution shift: models that excel on Gulf of Mexico salt fail on Norwegian chalk, and models trained on one operator's data rarely transfer to another.

LLMs and FMs break all three. The foundation-model framing of Bommasani et al. (2021) captures the move: any model, once pretrained on a sufficiently broad corpus, becomes a foundation that a thousand downstream adaptations rest on. Vaswani et al. (2017)'s Transformer architecture, which enabled this move, replaces recurrence with self-attention, allowing direct token–token interactions across a sequence, although vanilla attention still carries quadratic cost in sequence length. Brown et al. (2020)'s GPT-3 added *in-context learning*: the same 175-billion-parameter model, without any weight updates, adapts to new tasks from three to eight exemplars in its prompt. OpenAI (2023)'s GPT-4 strengthened the conversational, instruction-following paradigm that InstructGPT (Ouyang et al., 2022) and ChatGPT had already made mainstream. Three distinguishing capabilities follow and are what make petroleum's FM moment different from every prior wave:

- **Pretrain once, adapt many ways.** Supervised fine-tuning (SFT), low-rank adaptation (Dettmers et al., 2023; Hu et al., 2022), and retrieval-augmented generation (Lewis et al., 2020) let an operator

 doi:10.20944/preprints202604.1814.v1

specialise the same base model to dozens of downstream tasks for a fraction of the pretraining cost.

- **Multimodal by design.** Vision–language backbones — CLIP (Radford et al., 2021), LLaVA (Liu et al., 2023), PaliGemma (Beyer et al., 2024; Steiner et al., 2024), ColPali (Faysse et al., 2025), Qwen-VL (Bai et al., 2023; Wang et al., 2024b) — ingest seismic sections, log panels, core photos, drilling reports, and scanned completions together, whereas classical pipelines treated each modality in a silo.
- **Agentic and interactive.** ReAct-style tool-use (Schick et al., 2023; Yao et al., 2023b), multi-agent orchestration (Li et al., 2023a; Shen et al., 2023; Wu et al., 2024), and plan–act–reflect loops (Shinn et al., 2023; Wang et al., 2023b) let an LLM orchestrate a reservoir simulator, a seismic-processing package, or a WITSML rig feed from natural language. The seismic-processing agent of Kanfar et al. (2025), running Madagascar operators end-to-end without a line of Python, illustrates the endpoint.

A historical benchmark helps calibrate the pace. Medical imaging's deep-learning wave crested between 2015 and 2020; finance's between 2017 and 2022. In both, pretrain-then-fine-tune eventually displaced task-specific models, but the transition took five years. Petroleum — arriving five years later — is leapfrogging. Where Word2Vec (Mikolov et al., 2013) needed about three years to surface in operator-scale petroleum NLP (the NGDD DDR classifier of Antoniak et al. (2016) in 2016) and nine years to produce a dedicated domain encoder (PetroBERT (Rodrigues et al., 2022) in 2022), Llama 3 (Grattafiori et al., 2024) was adapted into EnergyLLM within roughly a year. The CS-to-petroleum latency is compressing, and the 2024–2026 window is the inflection.

1.3. Scope and Contributions of This Survey

This paper is the first survey to integrate the full LLM / FM / VLM / RAG / agent / industry-platform stack for petroleum engineering into a single analytical framework. Prior surveys — Liu et al. (2024b) (88 references on oil-and-gas LLMs and multimodal), Hadid et al. (2024) (geoscience LLMs), Zhang et al. (2025) (an IEEE GRSM geoscience perspective), Zhang et al. (2024b) (scientific LLMs broadly), and Menon et al. (2026) (scientific FMs broadly) — each illuminate one slice of the picture. Our scope is upstream oil and gas plus adjacent geoscience, from 2003 to 2026, with deliberate emphasis on the 2022–2026 LLM era. Figure 1 is our taxonomy; the PetroLLM Maturity Model of Figure 3 is our unifying scaffold.

Methodologically, the corpus was assembled from twelve parallel literature-search streams spanning fourteen thematic areas, then deduplicated and verified against DOI records, arXiv identifiers, OnePetro paper numbers, or canonical publisher URLs. We retain grey literature only when it documents public industry systems, benchmarks, workshops, or platform announcements that materially shape deployment; unverifiable and headline-only items were removed. The final retained corpus contains 296 cited entries.

 doi:10.20944/preprints202604.1814.v1

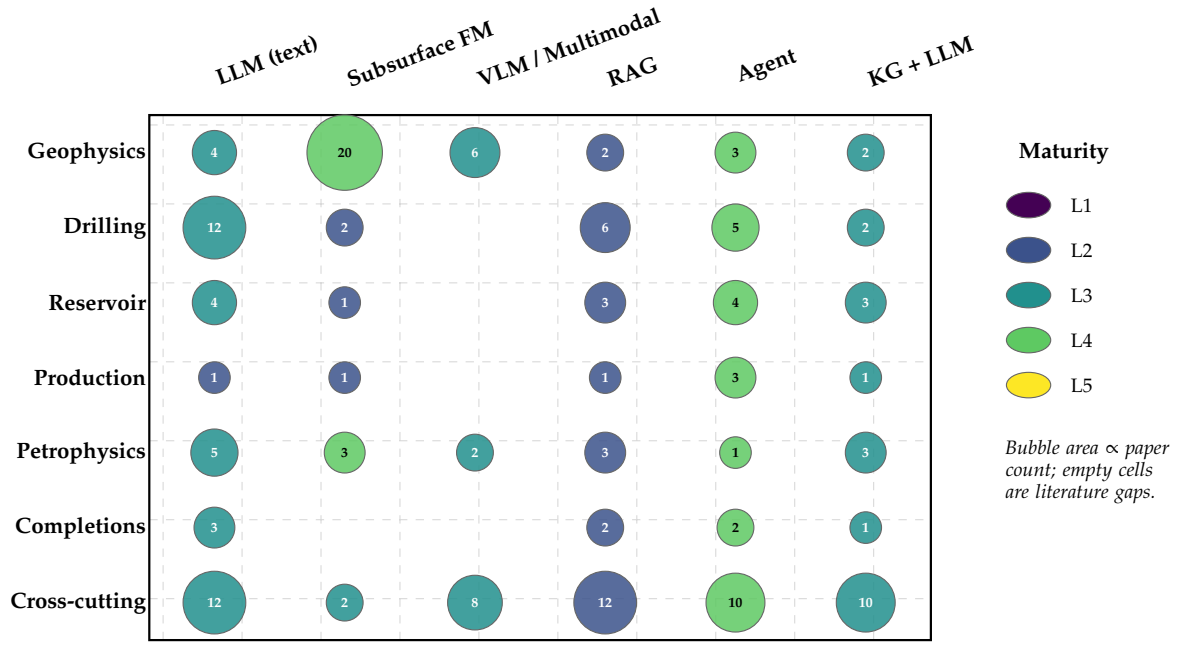

Figure 1. Taxonomy of LLM and foundation-model research for petroleum engineering. Six petroleum sub-disciplines plus one cross-cutting category (rows) cross six AI paradigms (columns). Bubble area is proportional to approximate paper count; colour encodes the dominant PetroLLM Maturity Level (L1 purple → L5 yellow), consistent with Figure 3. Small or empty cells mark thinner parts of the literature, especially petroleum VLM / multimodal work outside geophysics, sparse KG-coupled deployments outside cross-cutting and reservoir contexts, and limited agentic depth in several sub-disciplines.

Concretely, we make eight explicit contributions.

1. **First unified survey** integrating LLMs, foundation models, vision–language models, retrieval-augmented generation, agents, and commercial industry platforms for petroleum engineering in a single positioning framework (Table 1).
2. **Catalog of 296 references** spanning 2003–2026 and 14 thematic areas, integrated directly into the body text and summary tables — among the largest petroleum-specific bibliographic compilations published to date.
3. **First taxonomy figure** for the field (Figure 1), mapping six petroleum sub-disciplines plus one cross-cutting category against six AI paradigms with bubble-area paper counts, which doubles as a gap-finder for future research.
4. **The PetroLLM Maturity Model** (Figure 3, Table 14), a five-level scaffold (L1 Conversational Q&A, L2 Document Intelligence and Retrieval, L3 Domain-Specialized LLMs, L4 Autonomous Agents and Copilots, L5 Self-Improving Foundation-Model Ecosystems) inspired by CMMI and SAE-J3016 but grounded in petroleum realities and populated with canonical exemplars at every level.
5. **Catalog of 20+ public petroleum AI systems and enabling substrate** from major operators, oilfield-services companies, hyperscalers, and AI startups (Table 11) — the first survey to treat industry grey-literature announcements as a first-class scholarly object rather than a footnote.
6. **First compilation of petroleum-specific benchmarks** with explicit gap analysis (Table 12), naming FormationEval (Ermilov, 2026) as currently the only open petroleum-specific LLM benchmark and enumerating five dimensions along which the community must close the evaluation gap.
7. **First bibliometric snapshot** of the petroleum-LLM field (Figures 4–7, Table 13): publication trends, sub-discipline distribution, method distribution, institutional and geographic footprint, top-cited papers.

 doi:10.20944/preprints202604.1814.v1

8. **A concrete open-research agenda** across eight dimensions (§6, §7) anchored to the Maturity Model, naming under-served sub-disciplines (completions, production, well-test), under-served modalities (petroleum VQA, DAS, PVT), and under-served languages (Arabic, Russian, Portuguese, Chinese).

1.4. Related Surveys and Positioning

Two generations of petroleum-AI survey precede this work. The *classical-ML generation* — Rahmanifard and Plaksina (2019), Tariq et al. (2021), Koroteev and Tekic (2021), Sircar et al. (2021), Kuang et al. (2021), Noshi and Schubert (2018), Zhong et al. (2022), Wang and Chen (2023), Samnioti and Gaganis (2023a,2), Mousavi and Beroza (2022), Yu and Ma (2021), Xu et al. (2022), Bahaloo et al. (2023), Chen et al. (2025a), Latrach et al. (2024) — exhaustively cover support-vector machines, random forests, and convolutional-network baselines, but pre-date the LLM/FM era or treat it only as a postscript. The *LLM-era generation* is smaller but growing: Liu et al. (2024b) is the most comprehensive predecessor, with 88 references focused on LLMs and multimodal models in oil and gas, with only brief treatment of RAG and agents and no sustained treatment of industry platforms or benchmarks; Hadid et al. (2024) surveys geoscience LLMs without a petroleum focus; Zhang et al. (2025) offers an IEEE GRSM geoscience perspective; Zhang et al. (2024b) and Menon et al. (2026) frame scientific LLMs and FMs broadly without petroleum specificity; Fuchs et al. (2025) and Liu and Ma (2024) cover FMs in geophysics but not petroleum operations; and Wang et al. (2023a) and Han et al. (2024) provide the method-side primers we draw on. Table 1 rolls this landscape into a positioning matrix, mapping 25 prior surveys against the LLM/FM/VLM/RAG/agent/ benchmark/industry axes; our survey is the first to cover all eight.

Table 1. Positioning of this survey against 25 prior petroleum-, geoscience-, and method-adjacent AI surveys. Columns indicate whether each survey treats the given topic substantively (●= dedicated section), briefly (○= passing mention), or not at all (–). The bottom row summarises the present survey. A cell marked ● without qualification indicates at least a dedicated subsection or table; ○ indicates scattered inline mention only.

Survey	Year	#Refs	LLMs	FMs	VLMs	RAG	Agents	Bench.	Industry
Rahmanifard & Plaksina, petroleum AI (Rahmanifard and Plaksina, 2019)	2019	123	–	–	–	–	–	–	○
Sircar *et al.*, ML in O&G (Sircar et al., 2021)	2021	74	–	–	–	–	–	–	○
Koroteev & Tekic, upstream AI (Koroteev and Tekic, 2021)	2021	77	–	–	–	–	–	–	○
Tariq *et al.*, systematic ML (Tariq et al., 2021)	2021	225	–	–	–	–	–	○	○
Kuang *et al.*, AI in petroleum (Kuang et al., 2021)	2021	19	–	–	–	–	–	–	●
Noshi & Schubert, drilling ML (Noshi and Schubert, 2018)	2018	56	–	–	–	–	–	–	○
Zhong *et al.*, drilling review (Zhong et al., 2022)	2022	105	–	–	–	–	–	–	○
Wang & Chen, reservoir ML (Wang and Chen, 2023)	2023	71	–	–	–	–	–	–	○
Samnioti & Gaganis, reservoir sim I (Samnioti and Gaganis, 2023a)	2023	163	–	–	–	–	–	–	–
Samnioti & Gaganis, reservoir sim II (Samnioti and Gaganis, 2023b)	2023	217	–	–	–	–	–	–	–
Mousavi & Beroza, seismology ML (Mousavi and Beroza, 2022)	2022	177	–	–	–	–	–	●	–
Yu & Ma, geophysics deep learning (Yu and Ma, 2021)	2021	171	–	–	–	–	–	○	–
Xu *et al.*, petrophysics ML (Xu et al., 2022)	2022	23	–	–	–	–	–	–	–
Bahaloo *et al.*, petroleum ops (Bahaloo et al., 2023)	2023	146	–	–	–	–	–	–	○
Chen *et al.*, unconventional (Chen et al., 2025a)	2025	158	–	–	–	–	–	–	○
Latrach *et al.*, PIML (Latrach et al., 2024)	2024	163	–	–	–	–	–	–	–
Liu *et al.*, O&G LLMs (Liu et al., 2024b)	2024	88	●	○	○	○	○	–	○
Hadid *et al.*, geoscience LLMs (Hadid et al., 2024)	2024	120	●	●	○	–	–	●	○
Zhang *et al.*, geoscience FMs (Zhang et al., 2025)	2025	299	●	●	●	–	–	–	○
Zhang *et al.*, scientific LLMs (Zhang et al., 2024b)	2024	407	●	●	●	●	○	●	–
Menon *et al.*, scientific FMs (Menon et al., 2026)	2026	210	●	●	○	–	○	○	–
Fuchs *et al.*, FM review (Fuchs et al., 2025)	2025	52	–	●	–	–	–	●	–
Liu and Ma, FMs in geophysics (Liu and Ma, 2024)	2024	96	●	●	●	–	–	●	–
Wang *et al.*, text mining & KG (Wang et al., 2023a)	2023	130	–	–	–	–	–	–	–
Han *et al.*, PEFT (Han et al., 2024)	2024	261	●	–	○	–	–	●	–
This survey (Lu, 2026)	2026	**296**	●	●	●	●	●	●	●

● = dedicated section or table; ○ = brief mention only; – = not covered.

 doi:10.20944/preprints202604.1814.v1

1.5. Organisation of the Paper

Section 2 traces the computer-science lineage from word embeddings to foundation models and supplies the technical vocabulary on which the remainder of the paper relies. Section 3 analyses the domain-adaptation strategies (pretraining, continued pretraining, supervised fine-tuning, parameter-efficient fine-tuning, prompt engineering, retrieval-augmented generation, agentic orchestration) by which general LLMs become petroleum LLMs. Section 4 is the heart of the paper, with seven subsections covering geophysics (§4.1), drilling (§4.2), reservoir (§4.3), production (§4.4), petrophysics (§4.5), completions (§4.6), and cross-cutting applications (§4.7), each with its own summary table. Section 5 provides the bibliometric snapshot; Section 6 enumerates open challenges; Section 7 lays out the research agenda; and Section 8 closes.

How to read this paper.

The survey is written for a dual audience, and not every reader needs every section. A *petroleum engineer* arriving with a specific sub-discipline in mind can skim §2 for a common vocabulary, skip to the relevant §4 subsection and its summary table, and then proceed to the PetroLLM Maturity Model in §7 to place the sub-discipline on a common ladder with peer operators. A *machine-learning or CS reader* new to the petroleum domain should read §2 and §3 in full, then skim §4 for domain context, and concentrate on the evaluation-gap discussion in §6.3 and the open research agenda in §7. A *program manager or CIO* can start with the positioning matrix (Table 1), jump to the industry-platform catalogue (Table 11) and the PetroLLM Maturity Model (§7), and use the bibliometric snapshot in §5 to calibrate the pace and institutional footprint of the field. Readers looking for a specific system can locate it through the per-sub-discipline summary tables in §4 or the thematic tables in §4.7.

2. Background: From Word Embeddings to Foundation Models

This section supplies the computer-science vocabulary on which the rest of the survey depends. Our target audience is dual: petroleum engineers who want a precise, minimal-jargon walk through the architectural lineage, and computer-science readers who want a compressed refresher anchored to the petroleum applications that will appear in Sections 3 and 4. We therefore prioritize the concepts that later sections reuse and explain them in enough detail for the downstream discussion. That level of detail lets the reader, for example, distinguish why Harsuko and Alkhalifah (2022) adopts a masked-autoencoder pretext task rather than a contrastive one, or why Eckroth et al. (2025) couples supervised fine-tuning with parameter-efficient adapters rather than full-weight training. A timeline of the key milestones we cover, together with the petroleum-specific artifacts they enable, is provided in Figure 2.

 doi:10.20944/preprints202604.1814.v1

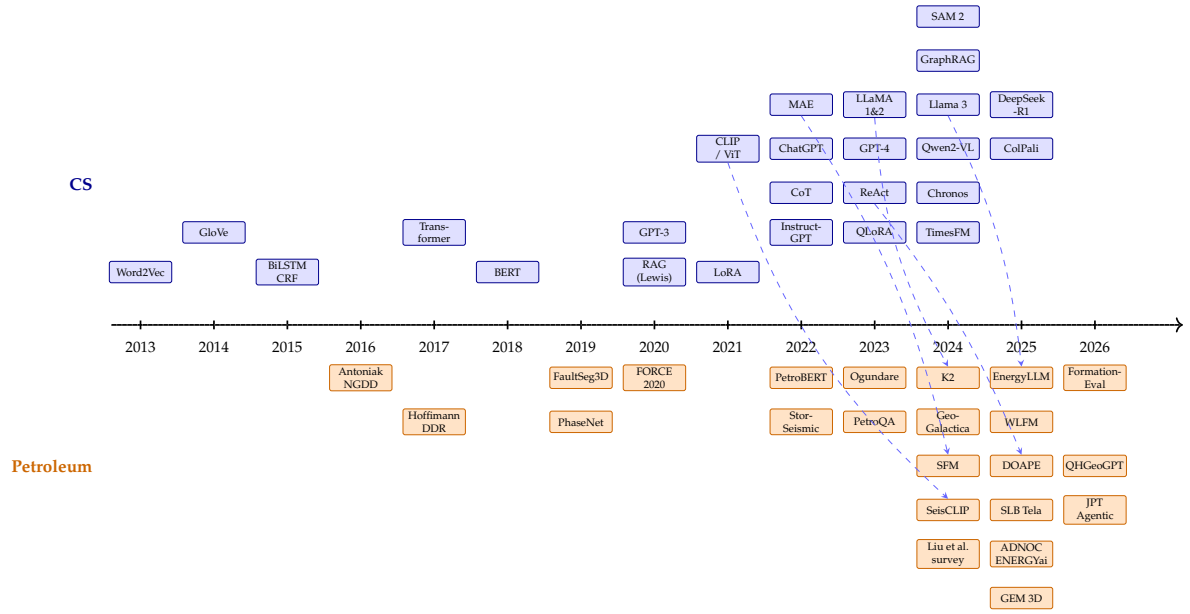

Figure 2. Evolution of key computer-science foundations (top, blue) and petroleum and adjacent geoscience LLM/FM milestones (bottom, orange) from 2013 to 2026. Dashed arrows mark representative technical lineages: LLaMA to K2 (Deng et al., 2024; Touvron et al., 2023a,2), MAE to SFM (He et al., 2022; Sheng et al., 2025), Llama 3 to EnergyLLM (Eckroth et al., 2025; Grattafiori et al., 2024), CLIP to SeisCLIP (Radford et al., 2021; Si et al., 2024), and ReAct-era agentic tool-use patterns to DOAPE (Yao et al., 2023b; Zejli et al., 2025). The CS-to-petroleum latency has compressed from about three years (Word2Vec to NGDD) and roughly nine years for a dedicated domain encoder (Word2Vec to PetroBERT) to roughly one year in the fastest recent case (Llama 3 to EnergyLLM).

2.1. From Word Embeddings to Transformers

The modern era of natural-language processing (NLP) began when discrete tokens were replaced by dense, learned real-valued vectors. Word2Vec (Mikolov et al., 2013) showed that a shallow neural network trained with a skip-gram or continuous-bag-of-words objective produced 300-dimensional vectors in which semantic relationships correspond to linear algebraic relationships — the canonical "king $-$ man $+$ woman $\approx$ queen" example. GloVe (Pennington et al., 2014) reached similar embeddings through global co-occurrence factorisation. These static embeddings displaced earlier statistical paradigms such as Latent Dirichlet Allocation (Blei et al., 2003), which the first wave of petroleum NLP — daily drilling-report clustering and morning-report topic extraction — had used extensively. Dense word embeddings were paired with bidirectional LSTMs and conditional random fields to produce the canonical sequence-labelling stack (Huang et al., 2015; Lample et al., 2016), which remained the dominant recipe for petroleum named-entity recognition until the late 2010s.

The decisive break came with the Transformer of Vaswani et al. (2017), which replaced recurrence entirely with a self-attention mechanism. Given a sequence of token embeddings projected into queries $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{n \times d_k}$ and values $\mathbf{V} \in \mathbb{R}^{n \times d_v}$, scaled dot-product attention computes

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \;=\; \mathrm{softmax}\!\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

with a multi-head variant that applies Eq. (1) in parallel on h linear projections. Two properties made this operator displace recurrent networks almost overnight. First, every token interacts with every other token in $\mathcal{O}(1)$ sequential steps, removing the vanishing-gradient bottleneck that limited LSTM depth. Second, the computation is dense matrix multiplication, which maps perfectly onto the GPU and TPU hardware accelerators that had just become commodity. Within five years these same properties would enable the pretraining of trillion-token language models and billions-of-image vision models on a single hardware stack.

 doi:10.20944/preprints202604.1814.v1

2.2. The Pretrain–Finetune Paradigm

Self-attention decoupled *what* a model does (any sequence task) from *how* it learns to do it. The key architectural families that exploit this decoupling are all transformers: BERT (Devlin et al., 2019) uses only the encoder and a bidirectional masked language-modelling (MLM) objective; GPT-style models (Brown et al., 2020) use only the decoder and an autoregressive left-to-right objective; and T5 (Raffel et al., 2020) keeps the full encoder–decoder and frames every task as text-to-text. BERT proved that unsupervised pretraining on large unlabelled corpora, followed by lightweight supervised fine-tuning, reaches or exceeds task-specific models while needing orders of magnitude less labelled data. This was the *pretrain–finetune paradigm*.

The paradigm quickly proved domain-transferable. BioBERT (Lee et al., 2020) continued pretraining on PubMed and PMC, and SciBERT (Beltagy et al., 2019) pretrained a new WordPiece vocabulary on a 1.14M-paper scientific corpus; both established that absorbing the vocabulary drift of a technical field — `p53`, `apoptosis`, `porosity`, `permeability` — before fine-tuning is a more data-efficient use of compute than training larger general models. PetroBERT (Rodrigues et al., 2022) transported this recipe to petroleum, continuing pretraining of BERT-multilingual and BERTimbau on the Portuguese-language *Petrolês* corpus and private Petrobras daily drilling reports, and it remains the most-cited BERT-era domain model for our industry. An earlier geoscience-adapted encoder of Lawley et al. (2022) is the direct ancestor of this lineage and established the first rigorous intrinsic evaluation protocol for domain LMs in the earth sciences. We will return to PetroBERT's descendants — SeisBERT, GEOBERTje, and refinery BERT — in Section 4.

2.3. Scaling Laws and Large Language Models

The jump from BERT-base (110M parameters) to GPT-3 (175B, Brown et al. 2020) is conventionally summarised by three empirical claims: (i) negative-log-likelihood on held-out text falls smoothly as a power law in parameters, data, and compute; (ii) some benchmarked capabilities — most visibly few-shot "in-context learning" where a model is conditioned on three-to-eight input–output exemplars at inference time with no weight updates — were reported to appear abruptly once models crossed certain scale thresholds; and (iii) this behavior helped motivate billion-plus-parameter systems. Hoffmann et al. (2022) refined claim (i) with the *compute-optimal* Chinchilla relation: for a fixed training-compute budget, parameters and training tokens should grow at roughly the same rate, implying that GPT-3-scale models were undertrained. This observation reshaped subsequent LLM families. PaLM (Chowdhery et al., 2022) and GPT-4 (OpenAI, 2023) pushed the closed-weight frontier, while the Meta LLaMA program (Grattafiori et al., 2024; Touvron et al., 2023a,2) made high-quality open-weight language models broadly available and, in doing so, directly seeded several petroleum-domain adaptations in this survey, including K2, EnergyLLM, EnergyGPT, and Ma *et al.*'s Llama-3.1-405B orphan-well pipeline. Parallel open-weight families — Mistral and its mixture-of-experts successors (Jiang et al., 2023), Gemma (Gemma Team, 2024), Qwen (Qwen Team, 2024a,2), and DeepSeek (DeepSeek-AI, 2024a,2,2) — now provide a diverse Pareto front of size-versus-licence trade-offs, while closed-weight families such as Gemini (Gemini Team, Google, 2023,2) continue to push the hosted-model frontier. The foundation-model framing of Bommasani et al. (2021) gave the field its shared language: any model, once pretrained, is a *foundation* that a thousand downstream adaptations rest on. The collective surveys of Minaee et al. (2024); Zhao et al. (2023) document the taxonomy in exhaustive detail.

2.4. Instruction Tuning, Alignment, and RLHF

A pretrained language model is a general next-token predictor, not a helpful assistant. Closing this gap is the task of *alignment*. The canonical pipeline proposed by InstructGPT (Ouyang et al., 2022) has three stages: supervised fine-tuning (SFT) on a curated set of instruction–response pairs; reward-model training on human preference comparisons; and reinforcement learning from human feedback (RLHF) using proximal policy optimisation against the reward model. Variants scale the SFT stage by automatically generating instructions (Self-Instruct, Wang et al. 2023d) and by mixing

 doi:10.20944/preprints202604.1814.v1

thousands of annotated NLP tasks (FLAN, Chung et al. 2022; Super-NaturalInstructions, Wang et al. 2022c). Direct Preference Optimization (DPO, Rafailov et al. 2023) later showed that the explicit RL stage can be replaced by a closed-form preference-optimization objective, although the approach still depends on curated preference data.

Alignment matters acutely in petroleum. A model that hallucinates a kick-tolerance formula or misquotes a completions tally is not a quirky inconvenience; in the control-room deployments documented by Ferrigno et al. (2024) and the well-construction copilots of Yi et al. (2024) it is a safety issue. We will see in Section 3 that, where publicly documented, many petroleum LLMs include SFT, SME evaluation, or both. For several commercial platforms, however, the exact alignment pipeline has not been publicly disclosed.

2.5. Parameter-Efficient Fine-Tuning

Full-weight fine-tuning of a 7B-parameter base model already requires tens of gigabytes of GPU memory; at 70B it requires a multi-node cluster. Parameter-efficient fine-tuning (PEFT, Han et al. 2024) sidesteps this by training only a small subset of weights. Low-Rank Adaptation (LoRA, Hu et al. 2022) is the dominant instance. It freezes the pretrained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ and learns a low-rank update $\Delta\mathbf{W} = (\alpha/r)\mathbf{BA}$ with $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times k}$, and $r \ll \min(d,k)$, where α is a scaling hyperparameter, so that the effective weight at inference is

$$\mathbf{W} \;=\; \mathbf{W}_0 + \Delta\mathbf{W} \;=\; \mathbf{W}_0 + \frac{\alpha}{r}\mathbf{BA}. \tag{2}$$

Typical settings use $r \in \{8, 16, 32\}$, reducing trainable parameters by three to four orders of magnitude while matching or nearly matching full fine-tuning quality. QLoRA (Dettmers et al., 2023) pushes this further by quantising $\mathbf{W}_0$ to 4-bit NormalFloat, paging optimiser state through CPU memory, and attaching LoRA adapters to the quantised base; in practice this brings 65B-parameter fine-tuning within a single 48 GB-GPU workstation, a configuration that operator IT departments can actually procure.

Both properties — the single-GPU memory footprint and the possibility of swapping LoRA adapters without touching the base model — map directly onto the petroleum industry's operational constraints. An on-premise operator can train LoRA adapters on confidential reservoir-management data without ever shipping weights outside the firewall; adapters for different basins or business units can be hot-swapped at serving time. This is the mechanism by which EnergyGPT (Chebbi and Kolade, 2025), the Lin rock-mechanics LLM (Lin et al., 2025), GeoFactory (Chen et al., 2025c), and the QH-GeoGPT exploration assistant (Ma et al., 2026) are deployable at all.

2.6. Vision-Language Models and Multimodal Foundations

Petroleum is not a text-only discipline. Seismic sections, core photographs, rock-thin-section microscopy, well-log panels, P&IDs, scanned completion reports and HSE screenshots all live in images. Three branches of the vision literature are directly relevant to Section 4.

Contrastive vision–language pretraining.

CLIP (Radford et al., 2021) jointly trains an image encoder and a text encoder on 400M (image, caption) pairs such that matched pairs produce similar embeddings and mismatched pairs do not. This yields an image embedding space that is *indexable by natural language*, and it is the essential precondition for retrieval, zero-shot classification, and the text-prompted segmentation and captioning pipelines that follow. SigLIP (Zhai et al., 2023) replaces the softmax-over-batch loss with a pairwise sigmoid, improving training efficiency at smaller batch sizes. In the petroleum FM literature, SeisCLIP (Si et al., 2024) ports the CLIP recipe to seismology, using earthquake spectra as images and event metadata as captions.

 doi:10.20944/preprints202604.1814.v1

Self-supervised visual pretraining.

Three recipes dominate. Masked Autoencoders (MAE, He et al. 2022) split an image into non-overlapping patches, mask 75% of them, and train a vision transformer to reconstruct the missing patches in pixel space; a closely related variant is SimMIM (Xie et al., 2022). Contrastive self-supervision — SimCLR (Chen et al., 2020), MoCo (He et al., 2020), BYOL (Grill et al., 2020), DINO and DINOv2 (Caron et al., 2021; Oquab et al., 2024) — maximises agreement between two augmented views of the same image without any labels. These self-supervised recipes are *the* mechanism by which the seismic-foundation-model family, SFM of Sheng et al. (2025), WLFM for well logs (Qi et al., 2025), StorSeismic (Harsuko and Alkhalifah, 2022), and the TGS 1.8B-parameter seismic FM (Kainkaryam et al., 2019) all acquire their representations without needing the (near-unobtainable) expert-labelled interpretation of every volume they ingest. Swin Transformer (Liu et al., 2021) provides the hierarchical backbone that several of these subsurface FMs adopt; the MAE-ViT adaptation of Li et al. (2023d) is a petroleum-adjacent precursor for MAE applied to subsurface velocity modelling.

Multimodal LLMs and document-level retrieval.

LLaVA (Liu et al., 2023) connected a frozen CLIP-ViT image encoder to a LLaMA backbone through a small projection layer, creating the first open-weight conversational vision-language model. Qwen-VL and Qwen2-VL (Bai et al., 2023; Wang et al., 2024b), and PaliGemma and its second generation (Beyer et al., 2024; Steiner et al., 2024), refined this architecture with stronger encoders and better instruction tuning. The Segment Anything Model (SAM, Kirillov et al. 2023) and SAM 2 (Ravi et al., 2024) added prompted segmentation over points, boxes or masks; they are the direct ancestors of the promptable Geological Everything Model (Dou et al., 2025). ColPali (Faysse et al., 2025) and M3DocRAG (Cho et al., 2024) further showed that, for document-centric retrieval over slide decks, scanned reports and well logs, it is more effective to embed *page images* directly than to OCR the text first — a finding especially relevant to petroleum, where a typical well file mixes tables, figures, handwritten annotations and stamped covers. The broader MLLM landscape is surveyed by Yin et al. (2024).

2.7. Retrieval-Augmented Generation

A standard LLM can only answer from parameters learned at training time, which makes it structurally hallucination-prone on any information it did not memorise. Retrieval-Augmented Generation (RAG, Lewis et al. 2020) factorises the task: given a query x, a retriever returns a set of supporting passages z, and a generator answers conditioned on both. Formally,

$$p(y \mid x) = \sum_{z \in \mathcal{Z}_x} p(z \mid x)\, p(y \mid x, z), \tag{3}$$

where $\mathcal{Z}_x$ is the top-k retrieval output and the inner generator $p(y \mid x, z)$ is typically a decoder LLM. The retriever in the original RAG paper is Dense Passage Retrieval (Karpukhin et al., 2020), a BERT bi-encoder trained with contrastive in-batch negatives. Subsequent retrievers include ColBERT's late-interaction scoring (Khattab and Zaharia, 2020), Sentence-BERT's pooled-embedding Siamese network (Reimers and Gurevych, 2019), the efficient billion-scale FAISS index (Johnson et al., 2021), BGE-M3's multi-lingual dense–lexical–multi-vector hybrid (Chen et al., 2024a), and the ColPali page-image late-interaction retriever (Faysse et al., 2025) mentioned above.

A flourishing subfamily of *advanced* RAG techniques improves either the retrieval or the generation side. HyDE (Gao et al., 2023) first drafts a hypothetical answer, embeds it, and retrieves against *that*. Self-RAG (Asai et al., 2024) teaches the model to emit reflection tokens that decide when to retrieve and when to cite. CRAG (Yan et al., 2024) introduces a lightweight retrieval-quality classifier and triggers web search when local evidence is unreliable. RAPTOR (Sarthi et al., 2024) hierarchically clusters and summarises a corpus to support multi-granular retrieval. GraphRAG (Edge et al., 2024)

 doi:10.20944/preprints202604.1814.v1

constructs an entity–relation graph first and retrieves community summaries, excelling at global questions. Comprehensive synthesis is provided by Gao et al. (2024).

For petroleum, RAG is not simply one method among many: it is the dominant *deployment* pattern. Petroleum knowledge is disproportionately tied up in private and proprietary documents (daily drilling reports, end-of-well reports, morning reports, PI files, PVT studies, well-test reports, geosteering logs), is legally obliged to be cited (AFE justifications, regulatory filings, HSE incident root-cause analyses), and turns over quickly (rig-state telemetry is fresh within minutes; log-based formation tops refresh with each well). These three constraints — private data, citation obligation, and knowledge freshness — are the properties that RAG addresses and that full fine-tuning does not. Every operator-scale petroleum copilot surveyed in Section 4 uses RAG at its core, often in combination with SFT-based style control.

2.8. Agentic AI: Reasoning, Tool Use, and Multi-Agent Systems

An *agent* extends an LLM from a single-turn text generator to a stateful act–observe–control loop over external tools and data. The foundational prompting pattern is Chain-of-Thought (CoT, Wei et al. 2022), which elicits step-by-step reasoning by prepending "let us think step by step"-style prefixes (Kojima et al., 2022) or few-shot exemplars that show the decomposition. Self-Consistency (Wang et al., 2022b) samples multiple CoT trajectories and votes; Tree-of-Thoughts (Yao et al., 2023a) explores them as a search tree. ReAct (Yao et al., 2023b) interleaves *thought*, *action*, and *observation* steps so the LLM can call external tools (calculators, search, simulators, SQL endpoints) mid-reasoning. Toolformer (Schick et al., 2023) shows the LLM can learn when to call which tool in a self-supervised way. Reflexion (Shinn et al., 2023) adds verbal self-critique between trials; Voyager (Wang et al., 2023b) accumulates a library of successful action sequences as reusable skills; and the generative-agents framework of Park et al. (2023) demonstrates long-horizon coherent behaviour in societies of agents with memory, planning, and reflection modules. Multi-agent orchestration frameworks such as AutoGen (Wu et al., 2024), HuggingGPT (Shen et al., 2023), and CAMEL (Li et al., 2023a) compose roles (planner, executor, critic, retriever) into workflows; broad surveys are provided by Guo et al. (2024); Wang et al. (2023c); Xi et al. (2023).

Agent architectures are especially consequential in petroleum because core workflows are multi-tool by nature. Seismic processing in Madagascar or SeisSpace, reservoir simulation in ECLIPSE, CMG or tNavigator, drilling analysis over WITSML streams, and production optimisation against a Leucipa-style surveillance system all require chaining heterogeneous tools under SME supervision. The seismic-processing agent of Kanfar et al. (2025) — running Madagascar operators with zero written code — the ENVOY reservoir-simulation assistant (Wiegand et al., 2024a) and the physics-informed DOAPE production-agent (Zejli et al., 2025) are the exemplars we will study in Section 4. The survey by Singh et al. (2025) captures the point of convergence between this subsection and the previous one: in practice, production petroleum copilots are *agentic-RAG* systems, not pure RAG and not pure agents.

2.9. Time-Series Foundation Models

The final CS subfield on which this survey leans is time-series foundation models (TSFMs). TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), Lag-Llama (Rasul et al., 2024), Moirai (Woo et al., 2024), and MOMENT (Goswami et al., 2024) are decoder- or encoder-only transformers pretrained on hundreds of billions of time-series points drawn from weather, traffic, retail, energy, epidemiology and other public corpora. All five demonstrate that a single backbone can deliver zero-shot or few-shot forecasting at or above the level of classical Box–Jenkins and task-specific deep models, across horizon length and sampling frequency.

This capability is structurally important for petroleum engineering, whose core data modalities — production rate, bottom-hole pressure, drill-string telemetry, mud-weight logs, rig-state streams, well-log curves — are all irregularly sampled time series. Koeshidayatullah's adaptation of TimeGPT for well-log imputation and depth-trend extrapolation (Koeshidayatullah et al., 2024) is the first petroleum-

 doi:10.20944/preprints202604.1814.v1

facing published instance. We discuss downstream uses — including the drilling-ROP-forecasting and production-surveillance FMs — in Sections 4. Whether TSFMs can fuse with multimodal petroleum FMs over seismic + log + rig-state + PVT is, as we argue in Section 7, the single most important subsurface-FM research question for 2026–2030.

3. Domain Adaptation Strategies for Petroleum Engineering

Section 2 surveyed the computer-science primitives. The practical question petroleum teams face is which combination of those primitives to use, at what stage of the model lifecycle, and against which cost constraint. This section answers that question along a ladder of increasing domain signal — from pretraining entire 250B-parameter models on proprietary corpora down to zero-shot prompting of a hosted frontier model — and indicates which subsections of Section 4 each strategy feeds. The organising taxonomy follows the *continued-pretraining, supervised fine-tuning, PEFT, prompting, and retrieval-augmented* lineage that Gururangan et al. (2020), Han et al. (2024) and Ponomareva et al. (2024) have formalised for general scientific domains, and to which we add the agentic patterns that dominate the 2025–2026 petroleum literature.

3.1. Pretraining from Scratch on Petroleum Corpora

At one extreme, a handful of national and international operators have elected to pretrain foundation models from initialisation on their own data. The Saudi Aramco METABRAIN programme (Aramco, 2024), announced at the 2024 Global AI Summit, is the most visible example: a 250B-parameter industrial LLM initiative trained on publicly described multi-decade Aramco data, with a larger future system on the roadmap (Aramco, 2024,2; Aramco Europe, 2025). Comparable closed Chinese operator programs have also been reported publicly, but sufficiently detailed model-card-level disclosures remain limited. These systems are rare and expensive: the Chinchilla relation (Hoffmann et al., 2022) implies tens of millions of GPU-hours even at Llama-3-scale, and full pretraining requires a legal-and-data-sovereignty posture that only national operators and a few supermajors can sustain.

Pretraining from scratch therefore makes sense only under three joint conditions: (i) a proprietary corpus large enough (on the order of 10^{11}–10^{12} tokens) to recover the scaling-law benefit; (ii) regulatory or confidentiality constraints that preclude even parameter-efficient fine-tuning of an external base; and (iii) an internal platform roadmap long enough (multi-year) to amortise the investment. Even for the largest operators, the dominant pattern remains the next stage of the ladder.

3.2. Continued / Domain-Adaptive Pretraining

Continued pretraining (CPT) — sometimes called domain-adaptive pretraining (Gururangan et al., 2020) or continued pretraining — takes an existing open-weight base and continues its self-supervised objective on a domain corpus. This has become the default recipe for geoscience and petroleum LLMs because it retains the linguistic and world-knowledge prior of the base while injecting domain vocabulary and reasoning patterns at a fraction of the cost of training from scratch. Recent open geoscience CPT examples span several base-model lineages rather than a single family: K2 continues pretraining of LLaMA-7B, GeoGalactica continues Galactica, and JiuZhou adapts Mistral-7B.

K2 (Deng et al., 2024), the first foundation LLM for geoscience, continues pretraining of LLaMA-7B on 5.5B tokens of geoscience literature and OpenStreetMap-augmented text, then applies instruction tuning with the authors' GeoSignal dataset and introduces the GeoBench evaluation suite. GeoGalactica (Lin et al., 2024) scales the recipe to 30B parameters by continuing the Galactica model (Taylor et al., 2022) on 65B geoscience tokens, producing the largest open geoscience LLM at time of release. JiuZhou (Chen et al., 2025b) applies the same pattern to Mistral-7B with a domain-balanced corpus that includes Chinese geoscience text and outperforms GPT-3.5 on GeoBench objective tasks. PreparedLLM (Chen et al., 2024b) formalises pre-task corpus preparation for geoscience CPT, emphasising deduplication, license tracing and contamination control. Commercial managed-service pipelines, most visibly AWS Bedrock's O&G-terminology customisation (Amazon Web Services, 2024), replicate the same structural idea for Titan, Claude and Llama bases.

 doi:10.20944/preprints202604.1814.v1

The key engineering trade-off in CPT is catastrophic forgetting: continued pretraining on a narrow domain corpus erases fractions of the base model's general knowledge. The standard mitigation is a *mixed-data* schedule — typically 10–30% general-domain replay against 70–90% domain data — plus a small learning rate and vocabulary extension only for the highest-frequency unseen tokens. These implementation details, usually tucked into appendices, are what separates a working petroleum CPT from a model that has forgotten how to write English.

3.3. Supervised Fine-Tuning with Petroleum Instruction Data

Continued pretraining gives a petroleum LLM vocabulary and topical awareness; it does not teach the model to follow engineering instructions or answer in an SPE-style format. Supervised fine-tuning (SFT) on curated instruction–response pairs bridges this gap. EnergyLLM (Eckroth et al., 2025) is the best-documented current example: a joint Aramco + SPE + i2kConnect system adapted from a Llama 3-family base on SPE technical content, with v1.0 SME blind evaluations reporting statistically significant wins over GPT-4o on petroleum QA. EnergyGPT (Chebbi and Kolade, 2025) fine-tunes LLaMA-3.1-8B on a broader energy corpus using both full SFT and LoRA, making it one of the few public petroleum studies that contrasts the two regimes directly. The two-stage template of Gharieb et al. (2024) first adapts a base model to conversational style and then to petroleum QA, a blueprint for on-premise operator deployments that must preserve data confidentiality. PetroQA (Eckroth et al., 2023) anticipated this pattern by combining PetroWiki grounding with GPT-3.5/GPT-4 prompting.

The bottleneck for SFT in petroleum is the same one that constrains every technical field: instruction data is scarce, because authoring a correct question–answer pair requires a subject-matter expert whose hourly rate is high. Three partial solutions have emerged. First, *society-authored content* (SPE OnePetro papers, PetroWiki, EAGE proceedings) can be recomposed into QA pairs with LLM assistance and expert review; this is the approach of EnergyLLM. Second, *self-instruct* pipelines (Wang et al., 2023d,2) generate candidate instructions from seed exemplars and filter by model self-consistency; scientific self-instruct with self-reflection has been explored at NeurIPS 2024 by SciInstruct (Zhang et al., 2024a). Third, *multi-agent* annotation, where a *critic* agent judges a *proposer* agent's QA pair, is explored by Sabbagh et al. (2024). Even with all three, the total annotated petroleum instruction-data pool in the open literature is on the order of 10^5–10^6 pairs — three orders of magnitude below the general-domain FLAN-T5 corpus — and remains the single largest bottleneck for the subfield.

3.4. Parameter-Efficient Fine-Tuning: LoRA/QLoRA in Practice

Where SFT dictates *what* a petroleum model learns, LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) dictate *at what cost*. The low-rank decomposition introduced in Eq. (2) makes fine-tuning tractable on operator-scale hardware. Public petroleum deployments now cluster around three patterns.

Pattern 1: LoRA on a mid-size open base. EnergyGPT (Chebbi and Kolade, 2025) attaches rank-r adapters to LLaMA-3.1-8B, reporting petroleum-QA gains at a fraction of full-SFT compute. The rock-mechanics LLM of Lin et al. (2025) applies the same recipe to an oil-and-gas geomechanics corpus and enumerates the data-standardisation, security, and physics-vs-data trade-offs. GeoFactory (Chen et al., 2025c) is the rare systematic ablation: four enhancement methods (RAG, prompt engineering, fine-tuning, agents) across 14 algorithms on Mistral-7B, with LoRA + GeoFactory beating GPT-4 on the authors' geoscience suite. QH-GeoGPT (Ma et al., 2026) combines a LoRA-tuned DeepSeek-R1 with a retrieval stage. LithoGPT-Mini (Li et al., 2025c) and BB-GeoGPT (Zhang et al., 2024c) target lightweight lithology identification and geographic-information QA respectively.

Pattern 2: QLoRA on a large open base. The 4-bit quantisation plus paged optimiser of QLoRA enables 65B-scale fine-tuning on a single 48 GB GPU workstation. This is the configuration most compatible with operator IT procurement rules, and it is the operational story behind Gharieb et al. (2024)'s on-premise roadmap and several industry deployments whose technical details are confidential.

 doi:10.20944/preprints202604.1814.v1

Pattern 3: Adapter routing. Because LoRA adapters are small and modular, an operator can train one adapter per basin, per business unit, or per classification level, and hot-swap them at serving time without touching the base model. This *LoRA-as-a-plugin* pattern is nascent in the petroleum literature but is implicit in the multi-tenant industry-platform architectures of SLB, ADNOC and their peers.

The practical trade-off is accuracy versus rank: $r = 8$ suffices for narrow stylistic adaptation, $r = 32$–64 is common for quantitative petroleum QA, and full SFT still wins by a few percentage points on the hardest benchmarks. The cost ratio, however, is rarely worth those percentage points once one accounts for iteration velocity and hardware lock-in.

3.5. Prompt Engineering and In-Context Learning for Petroleum

Even without any weight updates, a hosted frontier model can be steered by prompting. The relevant mechanics — zero-shot (Kojima et al., 2022), few-shot (Brown et al., 2020), Chain-of-Thought (Wei et al., 2022), Self-Consistency (Wang et al., 2022b), Tree-of-Thoughts (Yao et al., 2023a) — were covered in Section 2.8. Their petroleum applicability is dictated by one feature of the problems: quantitative petroleum engineering requires multi-step, unit-consistent, physics-aware reasoning. A material-balance calculation, a Darcy-flow derivation, or a kick-tolerance check is exactly the kind of decomposable reasoning task on which CoT and its successors offer the largest gains over direct prompting.

Ogundare et al. (2023) provide the canonical empirical probe: ChatGPT (GPT-3.5/GPT-4) is tested on oil-and-gas physics problems and failure modes are enumerated (unit errors, formula-misremembrance, implausible PVT correlations). The paper remains the baseline against which every subsequent petroleum-LLM comparison is implicitly or explicitly measured. Subsequent studies have layered CoT and Self-Consistency onto GPT-4 and Claude for well-log interpretation (Pacis et al., 2024) and for historical-well extraction (Ma et al., 2024), the latter achieving 100% extraction on text-PDF records using Llama-3.1-405B with CoT prompting. GeoFactory (Chen et al., 2025c) offers the most systematic petroleum-side ablation comparing prompt-only, RAG, SFT, and agent regimes; outside petroleum, GeoLLM of Manvi et al. (2024) at ICLR 2024 provides the canonical benchmark for prompt-engineering on geospatial tasks.

The general lesson from these studies is consistent with the general NLP literature: prompting alone is competitive when the base model is large (GPT-4 class), the task is knowledge-retrieval-dominated, and the correct answer can be justified by a short chain of reasoning. When any of those conditions fails — for example, when the task requires citing internal reports, when the budget cannot support a GPT-4 call, or when the reasoning chain is longer than a few steps — one of the next two strategies is preferred.

3.6. Retrieval-Augmented Generation as an Alternative to Fine-Tuning

For many operator deployments, RAG is not a complement to fine-tuning but a substitute. The decision is driven by three axes. First, *knowledge freshness*: if the canonical answer changes weekly (field-development plans updated with new well results, rig-state interpretations tied to the current bit run), RAG retrieves the latest documents while a fine-tuned model serves a stale snapshot. Second, *citation obligation*: a well-instrumented RAG system can return the passages it retrieved, whereas SFT alone cannot provide passage-level provenance. Third, *data confidentiality*: the retrieval corpus can be held behind the operator firewall and queried by a hosted generator without the sensitive text ever entering a pretraining set. RAG's factorised decomposition, Eq. (3), makes passage-level provenance tractable when retrieval results and citations are preserved, although standard RAG does not by itself guarantee sentence-level attribution.

Petroleum RAG has matured rapidly. Geo-RAG (Dong et al., 2024) is a well-documented deployment on unstructured geological documents; the "Revolutionizing Drilling Operations" WITSML real-time RAG of Matheus et al. (2025), and the drilling well-data orchestration of Reddicharla et al. (2025) are operator-scale exemplars. GeoGPT-RAG (Huang et al., 2025) open-sources a technical report on a geoscience RAG platform; GraphRAG for reservoirs (Jiang et al., 2025) shows how the

 doi:10.20944/preprints202604.1814.v1

graph-indexed retrieval of Edge et al. (2024) can support the global, multi-document questions typical of field-development decisions. ASK Thamama of Braik et al. (2021) provides the earliest operator-scale QA chatbot in the ADNOC onshore asset. The mud-invasion RAG of Aliyev et al. (2025), the seismic assistant of Zwartjes et al. (2024), the EAGE synthetic-data RAG benchmark (Chang et al., 2025), and the fine-tuning-plus-RAG study of Elyas et al. (2025) together establish a solid base for Section 4 Table 8. The companion survey on agentic RAG by Singh et al. (2025) is a useful synthesis between this subsection and the next.

3.7. Agent Construction over Tools, Simulators and Data Lakes

The last strategic pattern is agentic deployment, where the LLM orchestrates external tools and specialised sub-models. The enabling CS primitives — ReAct (Yao et al., 2023b), Toolformer (Schick et al., 2023), Reflexion (Shinn et al., 2023), AutoGen (Wu et al., 2024), CAMEL (Li et al., 2023a), HuggingGPT (Shen et al., 2023) — are covered in Section 2.8; the petroleum-specific instantiations we single out here anticipate the extended treatment of Section 4. The seismic-processing agent of Kanfar et al. (2025) runs Madagascar operators from natural language with a RAG layer for tool pre-selection. ENVOY (Wiegand et al., 2024a) and its segmentation companion (Wiegand et al., 2024b) orchestrate reservoir-simulation workflows. DOAPE (Zejli et al., 2025) is a physics-informed production agent. The JPT-level system-paper of Sharma (2026) captures the autonomous-well-modelling agenda. Autonomous-drilling agents are documented by Cayeux et al. (2021,2); Jacobs (2025); Osman et al. (2025), and the multi-agent well-construction architectures of Sabbagh et al. (2025,2) quantify the truthfulness gains and cost overheads of multi-agent versus single-agent deployments.

3.8. Evaluation Challenges for Domain-Specific LLMs

A recurring difficulty shared across every adaptation strategy above is evaluation. General-purpose LLM benchmarks (MMLU, GSM8K, HumanEval, BIG-bench) do not probe petroleum reasoning; SME-generated evaluation data is expensive; and commercial deployments rarely disclose their internal test sets. Three partial responses have emerged. First, FormationEval (Ermilov, 2026), released in January 2026, is the first open petroleum-specific LLM benchmark and is likely to become the community reference point; it spans formation evaluation, well-log interpretation, and reservoir-engineering QA. Second, GeoBench, introduced alongside K2 (Deng et al., 2024), supplies a geoscience-oriented evaluation suite that multiple subsequent models (Chen et al., 2025b; Lin et al., 2024) report against. Third, SME *blind evaluation* — where domain experts rate anonymised model outputs without knowing which system produced them — was used successfully by the EnergyLLM team (Eckroth et al., 2025) and is increasingly the gold standard for new petroleum-LLM releases. The gap between these three and what the field actually needs — open drilling, reservoir-simulation, and petroleum-multimodal benchmarks — is taken up in Section 6.

4. Applications

This section surveys concrete applications of large language models (LLMs) and foundation models (FMs) across petroleum engineering sub-disciplines. For each area we frame the problem in terms a computer-science reader can follow, survey the representative systems, their methods, data, and headline results, note the dominant methodological themes, and conclude with a summary table cataloging the key works. The organization follows the sub-discipline taxonomy established in Section 1: geophysics and seismic interpretation (§4.1), drilling engineering (§4.2), reservoir engineering (§4.3), production engineering (§4.4), petrophysics and well logging (§4.5), well completion and stimulation (§4.6), and a cross-cutting bucket covering document intelligence, agents, knowledge graphs, benchmarks, and industry platforms (§4.7). Within each subsection we trace the migration from classical NLP and deep-learning baselines (largely cross-referenced from the reviews of Koroteev and Tekic (2021); Kuang et al. (2021); Rahmanifard and Plaksina (2019); Sircar et al. (2021); Tariq et al. (2021) and the direct LLM-era predecessor of Liu et al. (2024b)) to the foundation-model era whose scope and velocity motivated this survey.

 doi:10.20944/preprints202604.1814.v1

The per-subsection summary tables are the main reference device: following the template of Liu et al. (2024b) and the medical-imaging survey by Litjens et al. (2017), each table groups works by sub-theme so that a reader can locate, for instance, all DDR-NLP works or all subsurface foundation models at a glance. All citation keys below resolve against the master `references.bib` catalogued in our bibliographic audit. Forward references to the PetroLLM Maturity Model (L1 Conversational Q&A → L2 Document Intelligence and Retrieval → L3 Domain-Specialized LLMs → L4 Autonomous Agents and Copilots → L5 Self-Improving Foundation-Model Ecosystems) situate each work on the maturity pyramid so that by the end of §4.7 the reader can locate any LLM-or-FM deployment in the field on a consistent scaffold. Figure 1, introduced in §1.3, visualises the full landscape: six petroleum sub-disciplines plus one cross-cutting category crossed with six AI paradigms, bubble area proportional to paper count, colour encoding the dominant PetroLLM Maturity Level; Figure 3 shows the maturity-model pyramid we will reference throughout.

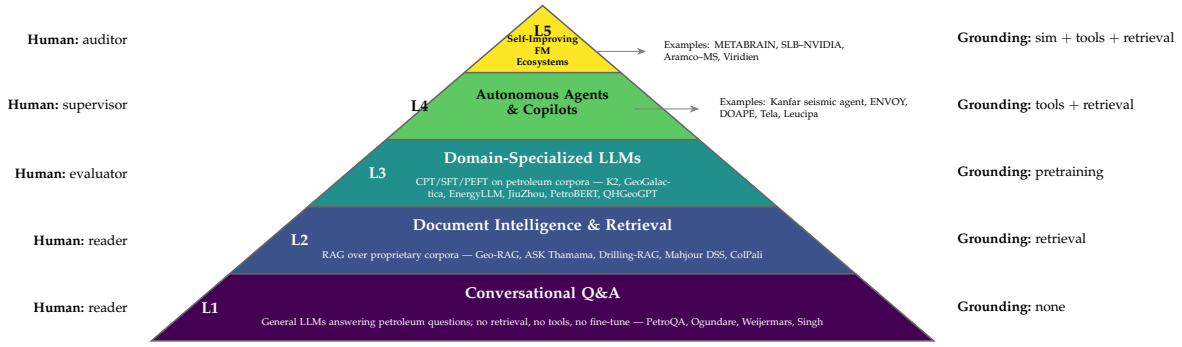

Figure 3. The PetroLLM Maturity Model, a five-level scaffold describing where any petroleum LLM or foundation-model deployment sits on the spectrum from conversational question answering (L1, bottom) to self-improving foundation-model ecosystems (L5, apex). Colours follow a `viridis` palette consistent with Figure 1. Right-hand annotations show the dominant *grounding* mechanism at each level; left-hand annotations show the corresponding human role.

4.1. LLMs and Foundation Models in Geophysics and Seismic Interpretation

Seismic data are the upstream industry's largest data modality (single field volumes routinely exceed 10 TB) and have accordingly attracted the most intense foundation-model effort of any petroleum sub-discipline. The central challenge from a machine-learning standpoint is that seismic cubes are high-dimensional, heavily noise-corrupted, and domain-shift-prone (salt, channel, deep-water, land, DAS acquisitions all exhibit markedly different statistics), which makes the pretrain-then-fine-tune paradigm of Dosovitskiy et al. (2021); He et al. (2022); Radford et al. (2021) especially attractive. Classical deep-learning baselines focused narrowly on a single task (fault detection, horizon tracking, facies classification) and were reviewed by Liu and Ma (2024); Mousavi and Beroza (2022); Yu and Ma (2021); the 2023–2026 wave surveyed here instead targets a *shared encoder* that any downstream task can tap.

Seismic image foundation models.

The foundational work in this lineage is the Seismic Foundation Model (SFM) of Sheng et al. (2025), which pretrains a ViT encoder with the masked-autoencoder (MAE) objective on 2,286,422 two-dimensional seismic images extracted from 192 global 3-D volumes. Published in *Geophysics* after two years of arXiv iteration, SFM established the empirical scaling laws for seismic pretraining and transferred to facies classification, geobody segmentation, interpolation, denoising, and impedance inversion. Pham et al. (2025) adopt a BERT-style bidirectional masked-image objective to produce SeisBERT; the associated *Leading Edge* paper emphasises cross-volume similarity search as a new downstream. The TGS-led program of Sansal et al. (2025) took the scaling argument to industrial extremes, training 660M- and 1.8B-parameter 3-D ViT-MAE models on 20 TB / 444,000 km^2 of field data via AWS SageMaker HyperPod and demonstrating zero-shot generalization from Gulf of Mexico

 doi:10.20944/preprints202604.1814.v1

and Brazil to West African salt. At the opposite end of the compute spectrum, Dong et al. (2025) match SFM on denoising, interpolation, and first-break picking with a dataset-distilled lightweight ViT, and Cheng et al. (2025) package a single generative FM for a complete processing flow. An alternate, contrastive-self-distillation thread adapts DINOv2-style objectives (Caron et al., 2021; Oquab et al., 2024) with handcrafted seismic attributes as auxiliary supervision. Harsuko and Alkhalifah (2022) is the earliest BERT-style seismic pretraining paper in the lineage and is now routinely cited as the "proto-SFM."

A parallel and less often discussed thread is *cross-modal* or *cross-domain* adaptation of natural-image foundation models to seismic data. Guo et al. (2025) systematically evaluates SAM, DINO, CLIP, and MAE-ViT on seismic, distributed acoustic sensing (DAS), and lunar data using lightweight adapters, and Fuchs et al. (2025) conducted an extensive comparison of natural-image FMs on three seismic tasks, concluding that hierarchical backbones (Swin Transformer (Liu et al., 2021)) often outperform plain ViT when the receptive-field requirements of seismic horizons and faults are taken into account. Kainkaryam et al. (2019) provides the precursor infrastructure argument: that an integrated oilfield-services data lake is a prerequisite to any trillion-parameter subsurface FM.

Promptable 3-D models.

A second cluster of works treats subsurface understanding as a *promptable* task in the spirit of the Segment Anything Model (SAM) (Kirillov et al., 2023; Ravi et al., 2024). The Geological Everything Model 3-D (GEM) of Dou et al. (2025) combines sparse-decoder MAE pretraining with adversarial mixed-prompt fine-tuning on masks, sketches, and well-log curves, yielding a single model that performs structural interpretation, stratigraphic analysis, geobody segmentation, and property modelling. The related prompt-engine seismic geobody model (Gao et al., 2026) enables cross-survey transfer for channels, salt, karst, and faults. A promptable fault-specific adaptation uses frozen SAM weights with lightweight adapters and a 2.5-D orthogonal scheme to match FaultSeg3D (Wu et al., 2019) at a fraction of the trainable parameters. In a related contrastive-pretraining thread, Si et al. (2024) builds SeisCLIP on a CLIP-style spectrum / metadata alignment objective and demonstrates strong transfer to event classification, focal-mechanism analysis, and source localization.

Fault and facies interpretation.

The most mature downstream task remains 3-D fault segmentation. Wu et al. (2019) set the classical-U-Net-on-synthetic benchmark (FaultSeg3D) that every subsequent model is measured against on F3, Kerry, and Opunake. Zhang et al. (2024d) show that Swin-UNETR with SimMIM pretraining (Liu et al., 2021; Xie et al., 2022) outperforms plain ViT with MAE, and Li et al. (2025a) push further with a hybrid ViT+CNN encoder. Analogous progress in semi-supervised facies classification is reported by Li et al. (2023b), whose CONSS model reaches state-of-the-art on the F3 and Parihaka benchmarks with 1% labelled data. Closer to the foundation-model era, Chikhaoui and Alfarraj (2024) benchmarks BYOL (Grill et al., 2020), SimSiam, and MAE pretraining on unlabelled F3 and shows that 5–10% labelled data now suffice to beat supervised baselines—an emblematic labeled-data-economy result that recurs across subsurface FMs (Dramsch and Lüthje, 2018).

Seismic waveform and passive-seismic FMs.

The passive-seismic community pioneered waveform-level FMs that are now informing the reflection-seismic literature. Mousavi et al. (2020) (EQTransformer) established the attention-based template for simultaneous earthquake detection and P-/S-phase picking, building on the Zhu and Beroza (2019) U-Net benchmark and the 1.2M-event STEAD corpus (Mousavi et al., 2019). Liu et al. (2024c) train SeisLM with a Wav2Vec2-style SSL objective on single-station waveforms and show strong zero- and few-shot transfer for phase picking. Shi et al. (2025) pretrains an MAE on offshore DAS cables to denoise and enhance earthquake detection—evidence that subsurface FMs generalize across traditional seismic, DAS fibre, and microseismic sensing.

 doi:10.20944/preprints202604.1814.v1

Full-waveform inversion with generative priors.

A third and methodologically distinct thread uses diffusion models and generative priors to accelerate or regularize full-waveform inversion (FWI). Wang et al. (2024a) trains a conditional diffusion model on velocity models, conditioning on geological class, well logs, or reflectivity, thereby augmenting the FWI training set with plausible subsurface structures. A related line of work embeds learned diffusion priors inside a Bayesian FWI loop to return posterior samples with calibrated uncertainty. The OpenFWI benchmark of Deng et al. (2022) is now the *de facto* evaluation suite for these methods, offering twelve dataset families spanning interfaces, faults, CO_2, and 3-D models.

Seismic agents and LLM-assisted workflows.

One major methodological shift is the arrival of *seismic agents*: LLM-orchestrated systems that translate natural language into executable processing workflows. The agentic system of Kanfar et al. (2025), reported in *The Leading Edge*, is described by the authors as the "industry's first guardrails-equipped AI agent performing seismic processing with zero code input," wrapping the Madagascar open-source processing stack. A parallel EAGE 2024 effort (Zwartjes et al., 2024) builds a Seismic Unix chatbot on the ReAct framework (Yao et al., 2023b), LoRA-fine-tuning Llama-2 on query-tool JSON pairs to pre-select the right tool for each user prompt. These systems exemplify L4 on the PetroLLM Maturity Model: they do not supplant the processing stack, but they collapse the code-authoring latency between a geophysicist's intent and a processed volume from hours to minutes. Geo-RAG (Dong et al., 2024) complements the agentic thread with a retrieval layer over unstructured geological documents (well reports, lithology descriptions, FWI run logs), giving the seismic interpreter a knowledge-grounded co-pilot.

Critical observations.

Two tensions deserve emphasis for a CS audience. First, the largest public seismic FMs (Dou et al., 2025; Sansal et al., 2025; Sheng et al., 2025) are pretrained predominantly on 3-D marine towed-streamer volumes, which leaves land, DAS, and OBN acquisitions systematically under-represented; generalization claims should be read with this caveat. Second, most fault-detection FMs continue to rely on *synthetic* fault-pretraining in the style of Wu et al. (2019), and the gap between synthetic noise statistics and field-noise statistics (swell noise, multiples, ground-roll) is presently a leading source of inference-time error. The maturing diffusion-prior and SSL works address part of this, but the field still lacks an open multi-acquisition field-noise benchmark. Table 2 consolidates the representative works in this subsection.

Table 2. Summary of representative works applying LLMs and foundation models to geophysics and seismic interpretation. Rows are grouped by sub-theme: seismic image foundation models, promptable/multimodal subsurface FMs, fault and facies interpretation, passive-seismic and DAS waveform FMs, FWI with generative priors, and seismic agents.

Reference	Year	Method / Model	Task	Data	Key result
Seismic image foundation models (pretrain-then-fine-tune)					
Sheng et al. (2025)	2025	SFM: ViT-MAE	Facies, geobody, interpolation, denoising, impedance inversion	2.286 M 2-D seismic images from 192 3-D volumes	Large-scale seismic FM in *Geophysics*; reports transfer across multiple downstream tasks and scaling behavior
Harsuko and Alkhalifah (2022)	2022	StorSeismic: BERT-style trace masking	Denoising, velocity estimation, FB picking, NMO	Synthetic + field pre-stack traces	Earliest MLM-style seismic pre-training; SFM precursor
Pham et al. (2025)	2025	SeisBERT: masked image modelling	Facies / salt / fault; cross-volume similarity	2-D seismic image patches	*TLE* 44(2) industry FM with cross-survey retrieval
Sansal et al. (2025)	2025	660 M / 1.8 B 3-D ViT-MAE on AWS SageMaker HyperPod	Salt segmentation; cross-basin transfer	20 TB / 444,000 km^2 TGS field data	Largest published seismic FM; 6 mo $\rightarrow$ 5 d training
Dong et al. (2025)	2025	Lightweight ViT with dataset distillation	Denoising, interpolation, FB picking	Distilled subset of SFM corpus	Matches SFM at a fraction of the compute
Cheng et al. (2025)	2025	Generative seismic FM	Unified processing workflow	Field seismic	*Surveys in Geophys.*: one-FM-for-all-processing
Kainkaryam et al. (2019)	2019	Industry data-lake argument	Salt-segmentation precursor	TGS salt Kaggle dataset	Foundational infrastructure paper

 doi:10.20944/preprints202604.1814.v1

Table 2. *Cont.*

Reference	Year	Method / Model	Task	Data	Key result
Promptable and multimodal subsurface FMs					
Dou et al. (2025)	2025	GEM 3-D: SAM-style promptable + MAE	Structure, stratigraphy, geobody, property modelling	Field 3-D seismic, logs, masks / sketches	Unified zero-shot subsurface understanding
Gao et al. (2026)	2026	Multimodal prompt-engine on pretrained VFM	Universal seismic geobody interpretation	Pseudo-labelled seismic surveys	Cross-survey geobody (channels / salt / karst / faults)
Si et al. (2024)	2024	SeisCLIP: CLIP-style spectrum + metadata	Event classification, focal mechanism, localization	Passive-seismic global waveforms	First CLIP-style seismology FM
Guo et al. (2025)	2025	Adapters on SAM / DINO / CLIP / MAE	Seismic, DAS, lunar cross-domain transfer	Multi-modality geophysical corpora	Natural-image FMs transfer to geophysics
Fault, facies, and self-supervised interpretation					
Wu et al. (2019)	2019	FaultSeg3D: 3-D U-Net on synthetic	3-D fault segmentation	Synthetic fault vols.; F3 / Kerry / Opunake eval	Canonical synthetic-pretrain baseline
Zhang et al. (2024d)	2024	Swin-UNETR + SimMIM pretraining	3-D fault segmentation	Unlabelled + synthetic fault volumes	Outperforms plain ViT + MAE
Li et al. (2025a)	2025	FaultVitNet: hybrid ViT + CNN	3-D fault segmentation	F3 benchmark	Beats FaultSeg3D on F3
Li et al. (2023b)	2023	CONSS: semi-supervised contrastive	Seismic facies classification	F3, Parihaka	SOTA with 1% labels
Dramsch and Lüthje (2018)	2018	ImageNet-pretrained VGG / ResNet / Inception	Seismic facies on F3	Netherlands F3	Classical transfer-learning baseline
Passive-seismic and DAS waveform FMs					
Mousavi et al. (2020)	2020	EQTransformer: attention-based encoder	Detection + P/S phase picking	Global seismic waveforms; validated on STEAD and continuous Japan data	Canonical attention-based seismic detection and picking model
Zhu and Beroza (2019)	2019	PhaseNet: U-Net picker	P/S arrival-time picking	600 k+ labelled waveforms	Benchmark precursor to SeisLM
Liu et al. (2024c)	2024	SeisLM: Wav2Vec2-style SSL	Phase picking, foreshock / aftershock	Single-station waveforms	Strong zero / few-shot transfer
Shi et al. (2025)	2025	MAE on DAS records	Denoising; offshore earthquake detection	Offshore DAS cable data	First DAS MAE
FWI with generative priors					
Wang et al. (2024a)	2024	Conditional diffusion over velocity	FWI augmentation, velocity synthesis	Synthetic velocity corpus	Class- / log- / reflectivity-conditioned priors
Deng et al. (2022)	2022	*Benchmark*: 12 FWI datasets (2.1 TB)	FWI evaluation	Interfaces, faults, CO_2, 3-D	Standard FWI-FM benchmark (NeurIPS D&B)
Seismic agents and LLM-assisted workflows					
Kanfar et al. (2025)	2025	Guardrails-equipped LLM agent	Full-waveform sonic data-processing assistance	User queries + seismic-processing tools	*TLE*: agent selects tools and executes filtering, clipping, and spectral-analysis tasks with safeguards
Zwartjes et al. (2024)	2024	ReAct-based LLM with Seismic Unix tools	Seismic-processing workflow assembly	User prompts + Python wrappers around Seismic Unix	EAGE 2024 assistant for operation and parameter selection

4.2. LLMs and Foundation Models in Drilling Engineering

Drilling is the sub-discipline in which LLM adoption has moved fastest from prototype to field pilot, for two reasons that help a CS reader orient. First, the drilling data lake is overwhelmingly textual and tabular: *daily drilling reports* (DDRs), *morning reports*, tool specifications, bit-forensics records, HAZOP and HSE narratives, end-of-well reports, and *non-productive time* (NPT) after-action reviews, all of which map naturally to classical NLP pipelines and now to LLMs. Second, drilling operators live under real-time decision pressure on rigs that cost US$200,000–$1,000,000 per day, so a 1–5% reduction in NPT or rate of penetration (ROP) uplift has direct economic consequences—a condition that makes drilling managers unusually receptive to LLM-assisted decision support. The evolution from keyword search to agentic co-pilots reviewed here spans roughly a decade.

DDR and mud-report text mining: the classical lineage.

Pre-LLM petroleum NLP was effectively synonymous with DDR mining. The earliest modern-NLP contribution is that of Antoniak et al. (2016), which applied logistic-regression classifiers over rig-crew natural-language fields to predict realized well risk. The Halliburton-authored work of Hoffimann et al. (2017) then established the deep-learning pipeline that dominated the next five years: EVENT/SYMPTOM/ACTION sequence labelling with LSTM encoders, reaching ∼83% accuracy on hundreds of wells. Sidahmed et al. (2015) and Arumugam et al. (2017) extended these techniques to symptom-cause mining with topic models and a drilling ontology, and Kowalchuk (2019) showed that

 doi:10.20944/preprints202604.1814.v1

sentiment-classification heuristics could identify dysfunction-bearing reports. Crucially, this entire lineage was pre-Transformer and relied on bespoke features and hand-curated ontologies—a point against which the LLM-era works should be read.

DDR classification and extraction with LLMs.

The LLM era opened with Kumar and Kathuria (2023), who used GPT-3.5/4 via prompt engineering for DDR NLP, and was formalised by Asif et al. (2024), who report a multi-label ADIPEC study (SPE-221870-MS) classifying DDR narratives across drilling, evaluation, completion, and workover activities with a single fine-tuned backbone. The breakthrough production paper is Wang et al. (2025), whose OTC 2025 submission demonstrates automatic daily drilling mud report processing with a generative AI pipeline that achieves a *∼99% reduction* in engineer processing time—an industry-realistic efficiency claim backed by explicit time-motion studies. The parallel strand of Ma et al. (2024) (published in *Scientific Reports*) tackles the adjacent problem of extracting location and depth from historical orphan-well documents using Llama 2 / Llama 3.1-405B / Mixtral, reporting 100% extraction on clean text PDFs and ∼70% on image-only archival records. Abdelgawad and El Ghattas (2025) applies the same methodology to Egyptian legacy well archives with a JSON-schema extraction target. Collectively, these works operationalise the L2 (document intelligence) level of the PetroLLM Maturity Model for drilling data.

Real-time drilling assistants and RAG.

A third wave targets real-time decision support in the well-construction control room. Ferrigno et al. (2024) (SPE-220798-MS, ATCE 2024) reports a GPT-3/GPT-4 assistant that classifies WITSML streams and log records in real time and reduces search/background-analysis time by more than 50×, all under zero-shot prompting. Yi et al. (2024) details a GPT+RAG blueprint deployed over more than 200 wells spanning sensor streams, morning reports, end-of-well reports, NPT reviews, and bit-forensics records. Alfarisi et al. (2024) documents what the authors describe as the first drilling-dedicated ChatGPT pilot (SPE GOTECH 2024), with a five-stage loading, scanning, indexing, training, and knowledge-extraction pipeline, and Matheus et al. (2025) (OTC-35742-MS) extends the pattern to offset-well lessons-learned and BHA-specification retrieval. Reddicharla et al. (2025) reports a Middle-Eastern operator case study in which LLM-enabled well-data access delivers measurable KPI uplift. Pacis et al. (2024) benchmarks commercial LLMs on zero-shot drilling QA and identifies aggregation-query strengths and factual-query weaknesses that motivate grounded RAG; the personalization roadmap of Gharieb et al. (2024) (SPE-220716-MS) addresses the confidentiality and cost-efficiency constraints that emerge when such systems are deployed on operator data.

Drilling copilots and multi-agent systems.

The most rigorous methodological contribution in the drilling LLM literature is Sabbagh et al. (2024), which appeared in *SPE Journal* and contrasts single-agent with multi-agent architectures on well-construction tasks, reporting a *28% truthfulness gain* for the multi-agent design at higher inference cost—one of the few peer-reviewed petroleum evaluations of the agentic paradigm. The companion Sabbagh et al. (2025) (OTC Brasil 2025) operationalizes the multi-agent stack as a drilling-and-completion knowledge-management RAG system. The *Journal of Petroleum Technology* editorial of Jacobs (2025) surveys eight autonomous-drilling papers from the 2025 SPE/IADC conference and concludes that the field is "headed in the right direction," foreshadowing the next generation of L4 systems. Cayeux et al. (2021,2) represent the physics-plus-AI lineage out of SINTEF/NORCE: autonomous decision-making while drilling and a revised digital drilling programme that now routinely serves as a ground-truth simulator for LLM-based copilots. Osman et al. (2025) (SPE-226893-MS) provides an autonomous directional-drilling case study on lumpsum turnkey rigs, illustrating the maturation of agentic drilling into commercial contracting.

Industry trials complete the picture. The Stone Ridge ENVOY system (Wiegand et al., 2024a,2) is discussed in detail in §4.3 but first shipped as a drilling–reservoir bridge, and the SLB Tela agentic assistant (SLB, 2025) pairs LLMs with domain FMs in a five-step observe-plan-generate-act-learn loop

 doi:10.20944/preprints202604.1814.v1

on the Lumi Data & AI platform (SLB, 2024c). Recent ADIPEC 2025 submissions on agentic RAG for drilling unify structured drilling data via knowledge graphs with conversational reasoning, producing real-time visualizations and narrative insights.

Rock mechanics, ROP, and time-series FMs.

Orthogonal to the text-mining thread, Lin et al. (2025) (*Natural Gas Industry B*) proposes the first workflow paper for a rock-mechanics LLM, calling out the three bottlenecks—data standardization, security, and physics-vs-data trade-offs—that any operator deployment must navigate. On the time-series side, the ROP benchmark of Tunkiel et al. (2021) now functions as the gold-standard benchmark suite against which time-series foundation models (Ansari et al., 2024; Das et al., 2024; Goswami et al., 2024; Rasul et al., 2024; Woo et al., 2024) provide a natural reference family for drilling-dynamics forecasting. These TSFM works were not designed for drilling and are best treated here as candidate baselines rather than petroleum-validated drilling models; the Koeshidayatullah et al. (2024) well-log study (discussed in §4.5) is the closest peer-reviewed analog in this survey.

Data standards.

The Energistics WITSML rig-telemetry standard and OSDU-compatible data platforms (Abolhassani et al., 2023; Microsoft, 2024) have emerged as the dominant integration layers for drilling-data workflows; Ferrigno et al. (2024); Matheus et al. (2025); Yi et al. (2024) all either ingest directly from WITSML or stage through an OSDU-style data platform. Table 3 consolidates representative works.

Critical observations.

Despite the sub-discipline's maturity, three gaps stand out. First, the DDR-classification literature still lacks a *shared open benchmark* comparable to FORCE 2020 (Bormann et al., 2020) in petrophysics; every operator builds on proprietary DDR corpora, which makes cross-paper accuracy comparison impossible. Second, few papers publish prompt templates or evaluation protocols, so the methodologies of Asif et al. (2024); Kumar and Kathuria (2023); Wang et al. (2025) cannot be reproduced. Third, safety-critical claims (NPT reduction, ROP uplift) are frequently reported without blinded control-group comparisons; the Sabbagh et al. (2024) peer review is an encouraging counter-example.

Table 3. Summary of representative works applying LLMs and foundation models to drilling engineering. Rows grouped by sub-theme. DDR = daily drilling report; NPT = non-productive time; ROP = rate of penetration; BHA = bottom-hole assembly; WITSML = Wellsite Information Transfer Standard Markup Language.

Reference	Year	Method / Model	Task	Data	Key result
Classical DDR and mud-report text mining					
Antoniak et al. (2016)	2016	Logistic regression + feature engineering	Risk hypothesis vs. realized risk classification	Two free-text drilling-risk datasets	Early SPE drilling-NLP study; domain adaptation achieved F1 0.84 on out-of-domain data
Hoffimann et al. (2017)	2017	LSTM sentence labelling (EVENT / SYMPTOM / ACTION)	DDR narrative classification	Hundreds of wells from an actual field	arXiv technical report: LSTM reached 82.7% mean 5-fold accuracy
Sidahmed et al. (2015)	2015	Rule-based + IR pipeline	Operations monitoring via DDR mining	Unstructured DDR archives	SPE-173429 operational NPT insights
Arumugam et al. (2017)	2017	LDA + drilling ontology	Symptom–cause association for DDRs	Field DDR corpus	Topic-modelling + ontology integration
Kowalchuk (2019)	2019	Sentiment-lexicon classifier	Flagging dysfunction-bearing reports	MEOS DDR corpus	Light-weight SPE-194961 baseline
DDR and mud-report classification / extraction with LLMs					
Kumar and Kathuria (2023)	2023	GPT-3.5 / 4 prompt engineering	DDR NLP, entity extraction	DDR text snippets	First SPE-ATCE LLM-on-DDR paper
Asif et al. (2024)	2024	Fine-tuned LLM multi-label classifier	Drilling, completion, workover activity	Operator DDR corpus	SPE-221870 end-to-end multi-label pipeline
Wang et al. (2025)	2025	Generative-AI mud-report parser	Automatic daily mud-report processing	Operator mud reports	OTC-35625: ∼99% reduction in engineer time
Ma et al. (2024)	2024	Llama 2 / 3.1-405B + Mixtral pipeline	Location & depth extraction from legacy docs	160 orphan-well documents	*Sci. Rep.*: 100% text / ∼70% image-only
Abdelgawad and El Ghattas (2025)	2025	Fine-tuned generative AI	Structured-data extraction from legacy reports	Egyptian oilfield archives	JSON-schema extraction on operator archive

 doi:10.20944/preprints202604.1814.v1

Table 3. *Cont.*

Reference	Year	Method / Model	Task	Data	Key result
Real-time drilling assistants and RAG					
Ferrigno et al. (2024)	2024	GPT-3 / GPT-4 zero-shot classifier	WITSML streams + log records, control-room	Real-time WITSML feeds	SPE-220798: $>50\times$ search-time reduction
Yi et al. (2024)	2024	GPT + RAG over wells, reports, NPT	Well-construction planning & real-time ops	$>$ 200 wells, multi-source	SPE-217700 practical drilling-RAG blueprint
Alfarisi et al. (2024)	2024	ChatGPT pilot, 5-stage pipeline	Drilling-knowledge capture	Operator knowledge base	First drilling-dedicated ChatGPT pilot
Matheus et al. (2025)	2025	LLM + RAG over offset wells / BHA specs	Real-time drilling efficiency	Morning reports, lessons learned, BHAs	OTC-35742 operator-grade RAG pattern
Reddicharla et al. (2025)	2025	LLM-enabled well-data access	KPI access and analysis	Middle-East major operator	SPE-229370 operator case study
Pacis et al. (2024)	2024	Commercial-LLM benchmark	Drilling-information retrieval QA	Curated drilling-query set	Aggregation-strong / factual-weak profile
Gharieb et al. (2024)	2024	Two-stage fine-tuning + RAG + local deployment roadmap	Personalized petroleum-engineering assistant	University of Houston petroleum-engineering materials	SPE-220716 roadmap for secure, locally hosted domain assistants
Drilling copilots and multi-agent systems					
Sabbagh et al. (2024)	2024	Single- vs. multi-agent LLMs	Well-construction Q&A and Text-to-SQL tasks	Domain-specific well-construction task set	*SPE Journal*: multi-agent improved Q&A truthfulness by 28%, while single-agent GPT-4 led Text-to-SQL
Sabbagh et al. (2025)	2025	Multi-agent RAG	Drilling & completion KM	Operator D&C knowledge	OTC-36203 productionized D&C RAG
Cayeux et al. (2021)	2021	Physics-based decision modules	Autonomous decision while drilling	SINTEF / NORCE physics stack	*Energies*: classical pre-LLM precursor
Cayeux et al. (2025)	2025	LLM-augmented digital drilling program	Drilling automation	SINTEF digital-drilling	SPE / IADC-223774 updated programme
Jacobs (2025)	2025	*JPT* editorial over 8 2025 papers	Autonomous drilling outlook	SPE / IADC 2025 cohort	Field "headed in right direction"
Osman et al. (2025)	2025	Agentic AI + directional drilling	Lumpsum-turnkey rigs autonomous DD	Operator DD job records	SPE-226893 autonomous DD case study
ROP, rock mechanics, and time-series FMs					
Tunkiel et al. (2021)	2021	Benchmark dataset for ROP	ROP modelling benchmark	NCS reference dataset	*JPSE* reference ROP benchmark
Lin et al. (2025)	2025	Workflow for rock-mechanics LLM	Data-to-deployment recipe	O&G rock-mech corpus	*NGIB* identifies three blocking challenges
Aliyev et al. (2025)	2025	Naive / vector-store / tree-index RAG over LAS + manuals	Mud-invasion zone detection and guidance retrieval	LAS well logs + domain manuals	SPE-227590: vector-store RAG outperformed naive and tree-index variants on relevance, ROUGE, and processing time

4.3. LLMs and Foundation Models in Reservoir Engineering

Reservoir engineering is the discipline concerned with fluid flow, pressure behaviour, and recovery optimisation in the subsurface over decadal time scales. Its computational core is the numerical reservoir simulator—standard commercial simulators include ECLIPSE, CMG, tNavigator, and open-source stacks such as OPM and ECHELON—whose input decks can exceed 100 000 lines of keyword-driven text. This text-heavy, workflow-heavy character makes the reservoir engineer's day unusually well matched to an LLM-based co-pilot, and the simulator's deterministic structure makes it an attractive tool-use surface for an agent.

Reservoir-simulation copilots.

The canonical reference is the Stone Ridge Technology ENVOY system (Wiegand et al., 2024a), a generative-AI reservoir-simulation assistant built on Amazon Bedrock with Titan embeddings and Claude callbacks. ENVOY helps engineers interpret input decks, diagnose runtime errors, and navigate multi-realization ensembles on the ECHELON GPU simulator, and the authors report a *>80% reduction* in forecasting cycle time on benchmark studies. The companion IMAGE 2024 submission (Wiegand et al., 2024b) extends the system to expert-model building, quality control, and post-run interpretation. A parallel Chevron-funded program, documented by Tharayil et al. (2024) (SPE-219324-MS), develops an LLM interface to transactional oil-and-gas screens so that engineers can interact with existing Petrel/DELFI dashboards via natural language—an especially relevant pattern for legacy-UI operators.

At the foundation-model physics interface, Kumar et al. (2023) proposes MyCrunchGPT, an LLM orchestrator over scientific-computing modules that has been used for well-testing and reservoir-coupled applications, and Li et al. (2025d) frames PDE solving itself as an LLM code-generation task,

 doi:10.20944/preprints202604.1814.v1

matching tailored numerical solvers on representative PDEs. Both works point at the longer-horizon agenda in which an LLM not only *drives* an existing simulator but *writes* its solver on demand.

Physics-informed agentic AI for production and reservoir decisions.

The Permian case study of Zejli et al. (2025) (SPE-230253-MS) introduces DOAPE (Digital Operations And Production Engineering), a multi-agent LLM system that combines physics-based nodal analysis with RLHF-fine-tuned agents, agentic RAG, cache-augmented generation (CAG), and in-context learning for sub-optimal-well identification. Although primarily a production system, DOAPE is pivotal to reservoir management because its well-by-well economic rankings close the loop with recovery forecasting. The *JPT* case study of Sharma (2026) (April 2026) reports a small-language-model (SLM) agent that orchestrates well simulators via Model Context Protocol (MCP) servers for India's ONGC, running 600+ wells through 370 tubing-sensitivity scenarios and compressing months of engineering work into hours—an illustrative L4 industrial result. Mahjour and Mahjour (2025) (arXiv:2509.11376) presents a more research-style prototype: GPT-4o, Claude 4 Sonnet, and Gemini 2.5 Pro combined with a domain-specific RAG over 50,000+ petroleum documents. The authors report gains on 15 field cases, but because the study is preprint-only and the evaluation corpus is not public, it should be read as preliminary evidence rather than as a benchmark anchor.

Field development planning.

The *field development plan* (FDP) is the industry artifact that synthesises geoscience, well, facilities, and economic modelling into a single multi-decade commitment, and it is one of the most promising targets for agentic AI. Rodriguez Torrado et al. (2025) (SPE-229333-MS) introduces an FDP optimisation-under-uncertainty pipeline that combines physics-informed neural networks with a generative-AI orchestrator on a North American real-field case. At higher methodological altitude, Cicconeto et al. (2022) formalises a reservoir ontology (GeoReservoir) that is now being used as a schema target for LLM-to-KG extraction pipelines such as Jiang et al. (2025), an application of GraphRAG (Edge et al., 2024) to a dual-layer process-and-knowledge reservoir graph published in *Processes*.

Domain LLMs targeting reservoir workflows.

Three continued-pretraining LLMs are directly usable in reservoir contexts. JiuZhou (Chen et al., 2025b) is a Mistral-7B continued on the JiuZhou-Corpus plus SFT on geoscience instructions, and outperforms GPT-3.5 on the GeoBench (Deng et al., 2024) objective-geoscience subset. GeoGalactica (Lin et al., 2024) is a 30B Galactica adaptation and K2 (Deng et al., 2024) a 7B LLaMA adaptation, both benchmarked on geoscience QA that spans reservoir, petrology, and basin modelling. Mahjour and Mahjour (2025) reports author-measured gains from a domain-specialised RAG over monolithic LLMs on reservoir-characterisation queries, but the evidence remains preprint-only.

Subsurface forecasting and CO_2 storage.

A separate and methodologically rich thread uses diffusion and generative models for subsurface forecasting, anchored by Wang et al. (2024a) (velocity models) and related diffusion multiphysics-monitoring prototypes that produce video-style CO_2 saturation evolutions conditioned on time-lapse seismic. These works sit at the boundary between reservoir engineering and geophysics and preview the *L5* vision of multimodal subsurface FMs that we discuss in §4.7. Finally, two systematic reviews anchor the discipline for a CS reader: Samnioti and Gaganis (2023a,2) survey machine-learning applications in reservoir simulation across two parts, and Wang and Chen (2023) surveys the broader reservoir-ML landscape.

Critical observations.

Reservoir engineering has a small number of well-documented LLM systems (ENVOY, DOAPE, Sharma's ONGC case) but few peer-reviewed benchmarks. Most published LLM work here concerns simulator copilots and decision support; core reservoir tasks such as history matching, material-balance analysis, well testing, and EOR screening remain largely open. The benchmark deficit is the single

 doi:10.20944/preprints202604.1814.v1

most urgent gap: no equivalent of OpenFWI (Deng et al., 2022) exists for reservoir simulation, and the FDP optimization literature in particular relies on proprietary field cases that outside researchers cannot reproduce.

Table 4. Summary of representative works applying LLMs and foundation models to reservoir engineering. Rows grouped by sub-theme. FDP = field development plan.

Reference	Year	Method / Model	Task	Data	Key result
Reservoir simulation copilots					
Wiegand et al. (2024a)	2024	ENVOY: GenAI assistant for ECHELON decks and outputs	Interpret decks, translate models, and analyze results	ECHELON input decks + manuals	SPE-221987 architecture paper for deck interpretation, model translation, and simulation-result analysis
Wiegand et al. (2024b)	2024	ENVOY extension	Expert model building; QC; interpretation	ECHELON workflows	IMAGE 2024 extension to expert modelling
Tharayil et al. (2024)	2024	LLM over transactional screens	NL interface to Petrel/DELFI dashboards	Aramco UI transaction logs	SPE-219324 GUI-to-LLM bridge
Physics-informed agentic systems					
Zejli et al. (2025)	2025	DOAPE: multi-agent copilot + physics-based nodal analysis	Well / field performance diagnosis and optimization	Two Permian pilot fields with high-quality production data	SPE-230253: reports < 30 s responses, $> 90\%$ accuracy, and gas-lift / ESP analysis
Sharma (2026)	2026	SLM agent + Model Context Protocol over simulators	Well modelling at field scale	ONGC 600+ wells, 370 tubing scenarios	*JPT*: author-reported months $\rightarrow$ hours and 1000+ engineering hours saved
Mahjour and Mahjour (2025)	2025	GPT-4o + Claude 4 + Gemini 2.5 + RAG	Reservoir decision support	50,000+ petroleum documents	Preprint-only framework; authors report validation across 15 reservoir environments
Field development planning and optimization					
Rodriguez Torrado et al. (2025)	2025	PINN + generative AI orchestrator	FDP optimization under uncertainty	North American real field	SPE-229333: real-field uncertainty-aware FDP
Knowledge graphs, ontologies, and GraphRAG for reservoirs					
Cicconeto et al. (2022)	2022	Reservoir ontology (GeoReservoir)	Schema target for LLM extraction	Reservoir-engineering vocabulary	Referenced by GraphRAG pipelines
Jiang et al. (2025)	2025	GraphRAG + dual-layer reservoir KG	Decision support (process + knowledge)	Offshore-field reservoir knowledge base; 40 expert Q&A pairs	*Processes*: dual-layer GraphRAG outperformed the single-layer graph baseline
Scientific computing and PDE copilots					
Kumar et al. (2023)	2023	LLM orchestrator over scientific modules	PDE/simulation driver	Reservoir-coupled codes	Early "LLM orchestrates numerical code"
Li et al. (2025d)	2025	CodePDE inference-time search	LLM-driven PDE solver generation	PDE problem suite	arXiv: matches tailored numerical solvers
Domain LLMs targeting reservoir workflows					
Chen et al. (2025b)	2025	Mistral-7B continued pretrain + SFT	Geoscience QA and instruction following	JiuZhou-Corpus + GeoBench	Geoscience-specialized LLM reporting gains over general baselines on objective tasks
Reservoir-ML context reviews					
Samnioti and Gaganis (2023a)	2023	Review, part I	Reservoir simulation ML survey	Mixed	Classical-to-DL reservoir-sim landscape
Samnioti and Gaganis (2023b)	2023	Review, part II	Reservoir simulation ML survey	Mixed	Complement to part I

4.4. LLMs and Foundation Models in Production Engineering

Production engineering covers everything downstream of first-oil: artificial lift (electric submersible pumps, beam pumps, gas lift), flow assurance (hydrate, asphaltene, paraffin), well

 doi:10.20944/preprints202604.1814.v1

testing, facilities, and the allocation and optimization of production across a field. The discipline is sensor-rich (pressures, temperatures, multiphase flow meters, SCADA historians) and event-label-poor, which is why the dominant deep-learning works have historically targeted *anomaly detection* and *event classification*. The LLM era has added a conversational layer on top of these models and has begun to absorb time-series forecasting into general-purpose foundation models. The published LLM evidence here is still concentrated in surveillance, anomaly detection, and operator copilots rather than the full production-engineering canon.

Anomaly and event detection.

The 3W dataset of Vargas et al. (2019) is the reference benchmark in this subsection—a Petrobras-released labelled dataset of rare, undesirable real events in offshore wells (hydrate formation, slugging, scaling) that now supports virtually every peer-reviewed production-anomaly method. Zhang et al. (2022) (SPE-208779-MS) targets the drilling-remarks complement with NLP-based event detection. On the industrial side, C3.ai's joint venture with Baker Hughes (Baker Hughes and C3.ai, 2019) productionized predictive-maintenance models on operator asset fleets, and the Aker BP / Cognite / NVIDIA stack (documented in several industry technical reports) deployed anomaly detection across the Norwegian Continental Shelf.

Production-LLM copilots and autonomous control.

Public production-facing systems remain relatively sparse. Baker Hughes' Leucipa and DOAPE are the best-documented public examples in this subsection. Chevron's ApEX (Chevron, 2025b) and APOLO (Chevron, 2025a) are better treated as cross-cutting industry systems: ApEX is an exploration prospecting platform, while APOLO sits closer to development planning and drilling-location optimization with production forecasting.

The Leucipa platform of Baker Hughes, extended with a Repsol partnership in Baker Hughes (2025), illustrates an L3–L4 transition in production: a production-optimization SaaS with LLM-enabled natural-language interaction on top of a classical surrogate-model core. The DOAPE system of Zejli et al. (2025), which we discussed in §4.3, is architecturally a production system first: its stated objective is sub-optimal-well identification across Permian assets with physics-informed agent reasoning. The Tharayil et al. (2024) natural-language screens system (SPE-219324-MS) similarly targets transactional production screens in Aramco fields.

Time-series foundation models for production forecasting.

Decline-curve analysis (DCA), arguably the most economically consequential forecasting task in the industry, has been transformed by the 2024 wave of time-series foundation models: TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), Lag-Llama (Rasul et al., 2024), Moirai (Woo et al., 2024), and MOMENT (Goswami et al., 2024). None of these were designed for petroleum, and the open petroleum literature remains thin: we did not find a peer-reviewed, apples-to-apples benchmark showing that these models consistently outperform Arps hyperbolic or exponential decline-curve baselines on production forecasting. The closest directly relevant petroleum evidence in this survey is Koeshidayatullah et al. (2024), which targets well-log prediction and anomaly detection rather than production-rate forecasting.

Autonomous maintenance and PdM.

Park et al. (2023) established the conceptual template (Generative Agents) on which autonomous PdM architectures are now being built. Industry prototypes target whole-life-cycle downhole maintenance with multi-agent coordination, and ExxonMobil's Vantage (ExxonMobil, 2025) and Shell AI (Shell, 2024) platforms illustrate the operator-scale direction of this evolution at operator scale.

Critical observations.

Production engineering benefits from a reference benchmark (3W)—unlike reservoir—but lacks a peer-reviewed *LLM-for-production* canon comparable to DDR NLP in drilling. Much of the public record

 doi:10.20944/preprints202604.1814.v1

is press releases and conference grey literature (Shell, Aker BP, Baker Hughes); systematic evaluation of the TSFMs on DCA would fill a visible hole. Refinery and downstream HSE NLP precursors — RefineryBERT (Macêdo et al., 2022) and the HAZOP NLP work of Feng et al. (2021) — offer templates that upstream HSE LLMs have yet to formally adopt. Table 5 catalogs the representative works.

Table 5. Summary of representative works applying LLMs and foundation models to production engineering. Rows grouped by sub-theme. TSFM = time-series foundation model; PdM = predictive maintenance; SCADA = supervisory control and data acquisition.

Reference	Year	Method / Model	Task	Data	Key result
Anomaly and event detection					
Vargas et al. (2019)	2019	3W labelled benchmark	Rare-event detection (hydrate, slug, scale)	Petrobras offshore-well archives	Canonical public production-anomaly benchmark
Zhang et al. (2022)	2022	NLP on drilling remarks	Event detection in reports	Operator remarks	SPE-208779: drilling-to-production NLP bridge
Production copilots and autonomous control					
Baker Hughes (2025)	2025	Leucipa press-release announcement with generative-AI assistant	Production optimization + conversational data access	Repsol production operations / Leucipa Lift Optimizer context	Baker Hughes and Repsol announced planned deployment; technical performance metrics were not publicly disclosed
Zejli et al. (2025)	2025	DOAPE: multi-agent physics-informed	Sub-optimal-well identification	Permian assets; production history	RLHF agentic RAG for production engineering
Tharayil et al. (2024)	2024	LLM over transactional screens	Natural-language production screens	Aramco operator logs	SPE-219324 GUI-to-LLM bridge
Time-series foundation models for forecasting					
Das et al. (2024)	2024	TimesFM: 200M decoder-only TSFM	Zero-shot time-series forecasting	Google time-series corpus	General TSFM baseline; no peer-reviewed petroleum production benchmark in this survey
Ansari et al. (2024)	2024	Chronos: T5-based tokenised TSFM	Probabilistic forecasting	Amazon time-series	General TSFM baseline; petroleum transfer remains sparsely documented
Rasul et al. (2024)	2024	Lag-Llama: LLaMA-style TSFM	Probabilistic TS forecasting	Unseen-domain probabilistic eval	Decoder-only TSFM family
Woo et al. (2024)	2024	Moirai: encoder-only masked TSFM	Multivariate TS forecasting	LOTSA (27B observations)	Unified training for universal forecasting
Goswami et al. (2024)	2024	MOMENT: open T5 TSFM family	Forecast, classify, anomaly, impute	Open time-series corpora	Open reference for TSFM evaluation
Autonomous-agent patterns adopted in production					
Park et al. (2023)	2023	Generative agents with memory/planning	Template for autonomous PdM agents	Simulacra of human behaviour	UIST 2023 foundational agent template
Ferrigno et al. (2024)	2024	Real-time LLM control-room support	Log-record and WITSML event classification	Real-time WITSML feeds + log records	SPE-220798: well-construction control-room support with $>$ 50$\times$ faster search; production-specific validation not reported

4.5. LLMs and Foundation Models in Petrophysics and Well Logging

Petrophysics—the quantitative interpretation of well logs and core data to derive porosity, permeability, saturation, and lithology—is the sub-discipline whose data modality most closely resembles the waveform-and-image stack that drove the foundation-model revolution in natural science. Every well in a field yields multi-curve log sequences (gamma ray, resistivity, density, neutron, sonic) over the complete depth interval, frequently supplemented by core images, thin sections, and NMR traces. The petrophysicist's interpretation task is structurally a multi-task transfer-learning problem, which the last two years have begun to treat as such.

Well-log foundation models.

 doi:10.20944/preprints202604.1814.v1

The inaugural well-log FM in the open literature is WLFM (Qi et al., 2025) (arXiv:2509.18152), a transformer-based FM pretrained on multi-curve logs from 1,200 wells via a three-stage pipeline: log-patch tokenization, masked-token pretraining combined with stratigraphy-aware contrastive learning, and multi-task fine-tuning. The authors report a porosity mean-squared-error of 0.0041 and 74.13% cross-well lithology accuracy on a held-out basin—the first numbers by which any future well-log FM will be judged. An industry counterpart with a promptable, SAM-style interface was announced by TGS at IMAGE 2025 under the name WLFM and is reported (in *First Break* coverage) to support lithology, stratigraphy, and facies prompts. The Koeshidayatullah et al. (2024) study takes a complementary route, fine-tuning the TimeGPT time-series foundation model on multi-well logs and reporting R^2 up to 87%, MAPE 1.95%, and 93% zero-shot anomaly-detection accuracy. Together these works establish that the classical log-cross-well transfer problem (Li et al., 2021; Romanenkova et al., 2022) is now squarely in the FM regime.

Core, thin-section, and lithology classification.

A parallel and older literature applies vision foundation models to core and thin-section imagery. Koeshidayatullah et al. (2022) introduced FaciesViT, a ViT-based core lithofacies classifier that reached 95% F1 without exhaustive augmentation. Cao et al. (2024) extended the approach with CoreViT, fine-tuning a ViT-B/16 on a broader core corpus, and Alzubaidi et al. (2021) provided the corresponding ResNet CNN baseline on an international core dataset with 96.7% test accuracy. Boiger et al. (2024) pushed toward a more physically meaningful task: direct prediction of continuous mineral fractions from drill-core images via transfer learning, in effect avoiding expensive XRF/XRD analysis.

Classical facies benchmarks.

The *de facto* benchmarks remain Hall (2016), whose *Leading Edge* facies-classification competition provided the Kansas Hugoton dataset now used for pedagogy, and Bormann et al. (2020), the FORCE 2020 North Sea well-log lithology contest that has supplied the most widely cited well-log labelled set of the last five years. Dramsch and Lüthje (2018) is the earliest widely-cited transfer-learning baseline on F3. These remain the evaluation datasets against which WLFM, CoreViT, and TimeGPT-for-logs should be read.

Formation evaluation with LLMs.

The document-intelligence side of petrophysics sits at the intersection of §4.7 and this subsection. The Egyptian oilfields case study of Abdelgawad and El Ghattas (2025) extracts structured petrophysical data from legacy well reports, and Ghorbanfekr et al. (2025) classifies Flemish borehole descriptions using a domain-adapted LLM (GEOBERTje), beating GPT-4 by +11% on lithology-label assignment. Elyas et al. (2025) (SPE-224128-MS) compares RAG against fine-tuning for drilling-petrophysical QA, finding that a small-budget RAG deployment matches a fine-tuned backbone at a fraction of the training cost—exactly the kind of cost/quality study operators need before committing to compute budgets. The ASK Thamama system (Braik et al., 2021) is the pre-LLM ADNOC precursor to all of these RAG systems: a pre-2023 production conversational subsurface knowledge engine with 90% intent-classification accuracy and 80% specific-retrieval accuracy.

Knowledge graphs for logs.

Liu et al. (2022) constructed a well-logging knowledge graph for hydrocarbon-bearing formation identification; together with the carbonate-rock classification transfer-learning study (Dawson et al., 2023) and the carbonate knowledge-base of He et al. (2021), it anchors the KG leg of the petrophysics literature. Xu et al. (2022) surveys the broader ML-for-petrophysics landscape and is a useful orientation for a CS reader.

Critical observations.

WLFM and CoreViT have begun to establish the multi-task FM paradigm in petrophysics, but three issues dominate. First, all current well-log FMs are trained on basin-specific or operator-specific

 doi:10.20944/preprints202604.1814.v1

log suites, and cross-basin generalization—especially across unconventional plays, carbonates, and deepwater siliciclastics—remains unvalidated outside of Qi et al. (2025). Second, core-image FMs have yet to integrate upstream petrophysical logs in a single multimodal backbone; the GEM 3-D work (Dou et al., 2025) points at what such a fusion could look like. Third, no open petroleum-specific formation-evaluation benchmark exists—FORCE 2020 is a lithology benchmark, not a saturation or permeability benchmark, and Ermilov (2026) is the first petroleum-LLM benchmark but does not yet cover numerical petrophysical prediction. Table 6 consolidates the representative works for this subsection.

Table 6. Summary of representative works applying LLMs and foundation models to petrophysics and well logging. Rows grouped by sub-theme.

Reference	Year	Method / Model	Task	Data	Key result
Well-log foundation models					
Qi et al. (2025)	2025	WLFM: three-stage MLM + stratigraphy-aware contrastive	Porosity regression; cross-well lithology; stratigraphy	Multi-curve logs, 1,200 wells	arXiv:2509.18152: porosity MSE 0.0041; 74.13% lithology
Koeshidayatullah et al. (2024)	2024	TimeGPT fine-tune for well logs	Multi-curve prediction; anomaly detection	Multi-well subsurface logs	R^2 up to 87%, 93% zero-shot anomaly
Core and thin-section transferred ViT / CNN models					
Koeshidayatullah et al. (2022)	2022	FaciesViT: ViT core classifier	Core lithofacies prediction	Core images, multi-basin	95% F1 without exhaustive augmentation
Cao et al. (2024)	2024	CoreViT: ViT-B/16 fine-tune	Core lithofacies identification	Broader core corpus	Matches FaciesViT on unseen formations
Alzubaidi et al. (2021)	2022	ResNet/Inception CNN	Core lithology classification	International core dataset	96.7% test accuracy (PLOS ONE)
Boiger et al. (2024)	2024	ImageNet-to-core transfer learning	Continuous mineral-fraction prediction	Drill core images	Swiss J. Geosci. avoids XRF/XRD
Classical facies benchmarks and legacy baselines					
Hall (2016)	2016	ML facies classification baseline	SEG-ML contest / Hugoton facies	Kansas Hugoton log dataset	*Leading Edge* benchmark anchor
Bormann et al. (2020)	2020	FORCE 2020 labelled log dataset	North Sea lithology classification	North Sea well logs (12 classes)	*Zenodo*: canonical lithology benchmark
Dramsch and Lüthje (2018)	2018	ImageNet-pretrained CNN	Seismic facies on F3	F3 volume	Classical transfer-learning baseline
Cross-well similarity and transfer					
Romanenkova et al. (2022)	2022	Similarity learning on logs	Cross-well similarity for analogues	LAS-style log collection	*JPSE* classical non-FM baseline
Li et al. (2021)	2021	Multi-scale sensor knowledge transfer	Cross-oilfield reservoir classification	Multi-field log data	AAAI: pre-FM cross-oilfield transfer
LLM-assisted formation evaluation and RAG					
Ghorbanfekr et al. (2025)	2025	GEOBERTje domain-adapted LLM	Borehole-description classification	283k Flemish borehole descriptions	App. Comp. Geosciences: +11% over GPT-4
Abdelgawad and El Ghattas (2025)	2025	Fine-tuned generative AI (JSON schema)	Legacy petrophysical extraction	Egyptian oilfield archives	JSON-schema extraction deployed
Elyas et al. (2025)	2025	RAG vs fine-tune comparison	Drilling/petrophysical QA	Operator knowledge base	SPE-224128: RAG matches fine-tune at lower cost
Braik et al. (2021)	2021	ASK Thamama pre-LLM KM engine	Subsurface Q&A (pre-ChatGPT)	ADNOC knowledge base	90% intent / 80% specific-retrieval accuracy
Knowledge graphs and context					
Liu et al. (2022)	2022	Well-logging KG + HC-identification method	Hydrocarbon-bearing formation identification	Logging ontology + well-log knowledge graph	*PED*: well-logging KG application for hydrocarbon-bearing formation identification
Dawson et al. (2023)	2023	CNN transfer learning	Carbonate rock classification	Carbonate image datasets	Dataset-size / CNN-architecture ablation
Xu et al. (2022)	2022	Petrophysics ML review	Landscape review	Mixed	Advantages/limitations framing

 doi:10.20944/preprints202604.1814.v1

4.6. LLMs and Foundation Models in Well Completion and Stimulation

Well completion (cementing, perforation, sand control, casing design) and stimulation (hydraulic fracturing and acidizing) are the operations that turn a drilled wellbore into a producer. The discipline is characterised by *high economic consequence per job* (a typical horizontal frac on a Permian well runs between US$2M and $5M), *strong geomechanical coupling* (stress anisotropy, fracture initiation, proppant transport), and *data that overlaps with both drilling and petrophysics*. The LLM literature here is the thinnest of any sub-discipline we review.

Completion-design copilots.

The best-documented completion-focused LLM application is the multi-agent well-construction copilot of Sabbagh et al. (2024), whose SPE Journal study—discussed in depth in §4.2—spans both drilling and completion tasks and measures a 28% truthfulness gain for a multi-agent over a single-agent architecture. The companion Sabbagh et al. (2025) OTC Brasil submission operationalises the system as a drilling-and-completion knowledge-management service, and Matheus et al. (2025) is architecturally a completions-adjacent RAG on offset-well lessons-learned corpora.

Fracturing knowledge LLMs.

The most specifically stimulation-focused paper is the fracturing knowledge-management LLM of Li et al. (2025b), published in *Water* with the DOI `10.3390/w17223317`: the authors move fracturing knowledge from traditional document governance to a natural-language Q&A pipeline backed by a fine-tuned LLM. This remains, to our knowledge, one of the few peer-reviewed LLM applications in the current corpus explicitly targeted at hydraulic-fracturing operations.

Rock-mechanics and geomechanics LLMs.

The rock-mechanics LLM workflow paper of Lin et al. (2025) is completion-adjacent rather than completion-specific, but it is the closest existing system for generative-AI-driven stimulation design: the authors document a data-to-deployment pipeline for an oil-and-gas rock-mechanics LLM and name the three open problems (data standardization, security, physics-vs-data trade-off) that a fracture- or geomechanics-specific successor must solve. The specialized geotechnical-design LLM of Fan et al. (2026) (*Solid Earth Sciences*, 2026) is the first domain-adaptation study targeted at quantitative geotechnical design and is directly relevant to casing, cementing, and completion-pressure calculations. The unconventional-AI review of Chen et al. (2025a) (*Energies* 18(2)) surveys the broader AI literature in unconventional oil and gas, much of which focuses on completion and stimulation.

Completion cross-references.

Completion data overlap substantially with drilling (DDRs record cementing and perforation events) and petrophysics (well logs drive stage selection). Works catalogued in §4.2 and §4.5—Asif et al. (2024); Elyas et al. (2025); Gharieb et al. (2024); Yi et al. (2024)—are therefore completion-adjacent and re-appear in the summary table.

Critical observations.

The thinness of the completions literature is itself a finding. Three plausible reasons, in order of decreasing speculative content: (i) frac data are *extremely* confidential, so operator-funded work tends to remain internal; (ii) the high economic consequence of each frac job punishes model hallucination more severely than the drilling or reservoir domains, making risk-averse operators slower to deploy LLM systems; (iii) completions are heavily regulated (Texas RRC, Alberta AER), and an LLM that returns a stimulation recipe that violates a permit condition is a legal exposure that an offset-well lessons-learned system does not create. We expect this gap to narrow by 2027 as Li et al. (2025b), Fan et al. (2026), and the multi-agent lineage of Sabbagh et al. (2025,2) expand. The summary table is deliberately compact (Table 7) to reflect the limited literature.

 doi:10.20944/preprints202604.1814.v1

Table 7. Summary of representative works applying LLMs and foundation models to well completion and stimulation. The literature is visibly thin, especially for hydraulic fracturing.

Reference	Year	Method / Model	Task	Data	Key result
Fracturing knowledge LLMs					
Li et al. (2025b)	2025	Document parsing + knowledge dictionary + Qwen3-14B Q&A pipeline	Fracturing-knowledge governance → NL Q&A	Fracturing documents and structured knowledge assets	*Water* 17:3317 proposes a practical fracturing knowledge-management and Q&A workflow
Completion/well-construction copilots and D&C multi-agent					
Sabbagh et al. (2024)	2024	Single-vs-multi-agent LLM	Well-construction QA (completion-adjacent)	D&C knowledge corpus	*SPE Journal*: +28% truthfulness (multi-agent)
Sabbagh et al. (2025)	2025	Multi-agent RAG	Drilling & completion knowledge management	Operator D&C knowledge base	OTC-36203 RAG KM system with source traceability and access-profile controls
Matheus et al. (2025)	2025	LLM+RAG over offset wells	Real-time drilling/completion support	Morning reports, lessons learned	OTC-35742 completion-adjacent lessons-learned
Rock mechanics, geomechanics, and unconventional					
Lin et al. (2025)	2025	Workflow for rock-mechanics LLM	Stimulation geomechanics data-to-deployment	O&G rock-mech corpus	*NGIB* identifies three adoption bottlenecks
Fan et al. (2026)	2026	Domain-adapted geotechnical LLM	Quantitative geotechnical design	Geotechnical documents and codes	*Solid Earth Sciences*: specialized framework
Chen et al. (2025a)	2025	Review of AI in unconventionals	Review (tight-oil/gas development)	Mixed: simulation + field	*Energies* review covering completion & frac
Personalization and RAG over confidential completion data					
Gharieb et al. (2024)	2024	Two-stage fine-tuning + RAG + local deployment roadmap	Personalized petroleum-engineering assistant	University of Houston petroleum-engineering materials	SPE-220716 secure, locally hosted domain-assistant roadmap rather than a completion-specific field deployment

4.7. Cross-Cutting Applications

Not every LLM or FM application in petroleum engineering localises to a single sub-discipline. Five cross-cutting clusters—document intelligence, agents, knowledge graphs, benchmarks, and industry platforms—appear throughout the corpus and deserve their own narrative thread. These clusters also populate the upper levels of the PetroLLM Maturity Model (L2 retrieval, L4 agents, L5 ecosystems) and therefore anchor the discussion in §6 and §7. The reference architecture stack used by these cross-cutting deployments — from a base LLM through domain adaptation, retrieval, agents/copilots, and the underlying petroleum data and tools — is rendered in Figure 8 at the head of §6. Table 8 synthesises the representative works discussed in this subsection; Tables 9–12 catalogue the domain-specific petroleum LLMs, subsurface foundation models, commercial platforms, and benchmarks that instantiate the upper tiers of this stack.

 doi:10.20944/preprints202604.1814.v1

Table 8. Summary of cross-cutting LLM and foundation-model works for petroleum engineering: document intelligence and RAG, agentic systems, knowledge graphs, multimodal retrieval, and evaluation benchmarks.

Reference	Year	Method / Model	Task	Data	Key result
Document intelligence and RAG (§4.7.1)					
Ma et al. (2024)	2024	Llama 2 / 3.1-405B + Mixtral pipeline	Location & depth extraction	160 orphan-well documents	*Sci. Rep.* LLM document intelligence
Hou et al. (2025a)	2025	Multimodal LLMs (GPT-4V, Claude 3.5)	Geologic document digitalization	Basin-study documents	*TLE* industry multimodal-LLM case study
Menezes (2023)	2023	Generative GPT vs hybrid extractive	Unstructured exploration search	Exploration archives	EAGE 2023: LLM vs extractive hybrid
Vimercati et al. (2022)	2022	NLP on reservoir management plans	Surveillance / corrective / lessons learned	Eni reservoir plans	SPE-209961 operator NLP deployment
Cordeiro et al. (2024)	2024	Petro NLP resource suite	Petrolês, PetroGold, PetroNER, PetroRE	Portuguese O&G corpus + Petro KGraph	Open Portuguese petroleum NLP/IE resource suite linking annotated corpora with a knowledge graph
Hutahaean and Simon (2025)	2025	CV + NLP deep-learning pipeline	Drilling-tool failure investigation	Drilling-tool imagery + reports	IPTC-24706 multimodal failure analysis
Chang et al. (2025)	2025	Embedding fine-tune with synthetic QA	Generic petroleum RAG uplift	Synthetic operator QA pairs	EAGE 2025: +19.46% context relevancy
Singh et al. (2023)	2023	Generative-AI conversational chatbot	Drilling + production analytics	Operator data + QA	SPE-216267 early GenAI chatbot
Multimodal and layout-aware components					
Xu et al. (2020)	2020	LayoutLM: text+layout MLM	Form/receipt/table understanding	RVL-CDIP / FUNSD / etc.	SOTA on scanned-DDR stacks
Kim et al. (2022)	2022	Donut: OCR-free Swin+BART	End-to-end doc understanding	Scanned doc corpora	ECCV 2022 OCR-free baseline
Blecher et al. (2023)	2023	Nougat: encoder-decoder transformer	Academic-PDF → Markdown	arXiv-like scientific PDFs	Critical for SPE/OnePetro PDFs
Nassar et al. (2022)	2022	TableFormer: Table-structure Transformer	Row/column/cell recovery	Table images	OTSL tokenisation cuts tokens by ∼80%
Li et al. (2023c)	2023	TrOCR: Transformer OCR	Handwriting / printed OCR	Core-log handwriting	Pretrained OCR for noisy archives
Faysse et al. (2025)	2025	ColPali: multi-vector on PaliGemma patches	Visual document retrieval	Visually-rich PDFs	ICLR 2025; outperforms text pipelines
Cho et al. (2024)	2024	M3DocRAG: ColPali + Qwen2-VL reader	Multi-page multi-doc QA	MP-DocVQA	SOTA multimodal RAG
Chen et al. (2024a)	2024	M3-Embedding: multilingual multi-granular	Dense retrieval for RAG	Multilingual corpora	Backbone embedding in petroleum RAGs
Knowledge graphs and ontologies (§4.7.3)					
Tang et al. (2023)	2023	Petroleum exploration & development KG	Ontology-based E&P KG construction	Petroleum E&P ontology and upstream documents	*Geoscience Frontiers* petroleum E&P ontology-to-KG framework with downstream knowledge services
Garcia et al. (2020)	2020	GeoCore ontology	Core geological ontology	General geology	*C&G* reusable general ontology
Santos et al. (2024)	2023	O3PO production-plant ontology	Offshore production plant	Offshore production data	Expert Systems ontology for offshore ops
Wang et al. (2022a)	2022	DL NLP + KG construction	Geological-report understanding	Geological report corpora	DL-based KG population of geological reports
Agentic systems and exploration chat (§4.7.2)					
Mosser et al. (2024)	2024	Conversational LLM over exploration archive	Decades-long exploration KM	Equinor exploration corpus	SPE-218439 exploration-robot chat

 doi:10.20944/preprints202604.1814.v1

Table 9. Announced and published petroleum- and geoscience-specific domain language models and adjacent encoders (2022–2026). All rows are Level 3 (Domain-Specialized LLMs) of the PetroLLM Maturity Model; lighter tint denotes academic open-weight releases, heavier tint denotes operator-scale industrial platforms that have graduated toward Level 4 or L5 deployment. n/d = not disclosed.

Name	Parameters	Base model	Training data	Access	Venue	Year
K2 (Deng et al., 2024)	7 B	LLaMA-7B	5.5 B geoscience tokens + GeoSignal instructions	Open	WSDM Resource	2024
GeoGalactica (Lin et al., 2024)	30 B	Galactica-30B	65 B geoscience tokens + 1 M instruction pairs	Open	arXiv	2024
JiuZhou (Chen et al., 2025b)	7 B	Mistral-7B	Balanced Chinese + English geoscience corpus	Open	Int. J. Digital Earth	2025
PetroBERT (Rodrigues et al., 2022)	110 M	BERT-mul. / BERTimbau	Petrolês + Petrobras DDR (Portuguese)	Public	PROPOR	2022
EnergyLLM (Eckroth et al., 2025)	n/d	Llama 3 family (publicly described)	SPE technical content	Hosted / proprietary	SPE ATCE	2025
EnergyGPT (Chebbi and Kolade, 2025)	8 B	LLaMA-3.1-8B	Curated energy-sector corpus; full-SFT and LoRA variants	Academic / arXiv	arXiv	2025
QHGeoGPT (Ma et al., 2026)	7 B	DeepSeek-R1-7B family	Qin-Hang geological corpus with LoRA/RAG workflow	Academic	Eng. Appl. AI	2026
GeoFactory (framework) (Chen et al., 2025c)	n/d	General LLM backbones	Geoscience factual/inferential enhancement framework; 14-algorithm evaluation	Academic	Big Earth Data	2025
LithoGPT-Mini (Li et al., 2025c)	n/d	n/d	Lithology-identification corpus	Academic	ADIPEC	2025
BB-GeoGPT (Zhang et al., 2024c)	7 B	LLaMA-2	Geographic-information QA	Academic	Inf. Process. Manage.	2024
GeoCode-GPT (Hou et al., 2025b)	7 B	Code Llama	Geospatial code-generation corpus	Academic	Int. J. Appl. Earth Obs.	2025
PreparedLLM (framework) (Chen et al., 2024b)	n/d	Model-agnostic	Domain-corpus preparation / pretraining framework	Academic	Big Earth Data	2024
Rock-Mech LLM (Lin et al., 2025)	n/d	LLaMA / Mistral families discussed	Rock-mechanics / geomechanics corpus	Academic	Nat. Gas Ind. B	2025
Geotech LLM (Fan et al., 2026)	n/d	Llama-family derivative	Geotechnical corpus	Academic	Solid Earth Sci.	2026
METABRAIN (Aramco, 2024; Aramco Europe, 2025)	250 B	Proprietary	90 years of Aramco data (publicly described); full training recipe undisclosed	Closed	Aramco announcement / Global AI Summit	2024
ENERGYai (ADNOC, 2025; AIQ, 2025)	70 B (publicly announced)	Partner-developed / proprietary	ADNOC operating data; staged rollout across operations	Closed	AIQ / ADNOC releases	2025
Aramco Dhahran LLM (Aramco, 2026; Aramco Europe, 2025)	n/d	Proprietary	Internal operator data; public source supports an AI/supercomputing program, not a full standalone LLM spec	Closed	Aramco news / JPT	2025

Table 10. Foundation models published or announced for subsurface data — seismic volumes, well logs, core imagery, and distributed acoustic sensing. Architectures cluster around MAE / ViT and BERT-style masked-modelling, CLIP-style contrastive, and Wav2Vec2-style self-supervised objectives.

Name	Modality	Architecture	Pretraining data	Downstream tasks	Venue	Year
SFM (Sheng et al., 2025)	2-D seismic	ViT-MAE	2.28 M 2-D seismic images from 192 3-D volumes	Facies, geobody, denoising, interpolation, impedance inv.	*Geophysics*	2025
SeisCLIP (Si et al., 2024)	Passive seismic	CLIP (spectra + metadata)	Global passive-seismic waveforms + event metadata	Event classification, focal mechanism, localization	IEEE TGRS	2024
GEM 3D (Dou et al., 2025)	3-D seismic + labels	SAM-style + MAE	Field 3-D seismic volumes with masks / sketches	Structure, stratigraphy, geobody, property modelling	arXiv	2025
StorSeismic (Harsuko and Alkhalifah, 2022)	Pre-stack traces	BERT-style trace masking	Synthetic + field pre-stack seismic traces	Denoising, velocity est., FB picking, NMO	IEEE TGRS	2022
SeisBERT (Pham et al., 2025)	2-D seismic	Masked image modelling	2-D seismic image patches	Facies / salt / fault; cross-volume similarity	*TLE*	2025
SeisLM (Liu et al., 2024c)	Waveform	Wav2Vec2-style SSL	Single-station raw waveforms	Phase picking, foreshock / aftershock classification	arXiv	2024
Light-FM (Dong et al., 2025)	2-D seismic	Dataset-distilled ViT	Distilled subset of SFM corpus	Denoising, interpolation, FB picking	*Geophysics*	2025
TGS 1.8B FM (Kainkaryam et al., 2019; Sansal et al., 2025)	3-D seismic	3-D ViT-MAE (660 M / 1.8 B)	20 TB / 444,000 km^2 TGS field data (SageMaker HyperPod)	Salt segmentation, cross-basin transfer	industry / *TLE*	2025
WLFM (Qi et al., 2025)	Well logs	Transformer masked modelling	Multi-well LAS log corpora	Log imputation, lithology, petrophysics	arXiv	2025
TimeGPT-Logs (Koeshidayatullah et al., 2024)	Well-log time series	Decoder-only TSFM	Cross-field LAS time series	Log forecasting / imputation	arXiv	2024
FaciesViT (Koeshidayatullah et al., 2022)	Thin-section / facies images	ViT	Thin-section image corpus	Automated facies / sedimentology	Front. Earth Sci.	2022
DAS-MAE (Shi et al., 2025)	DAS records	MAE	Offshore DAS cable data	Denoising, offshore earthquake detection	*GJI*	2025
CoreViT (Cao et al., 2024)	Core images	ViT	Core-image corpora	Core image classification	*Geoenergy Sci. Eng.*	2024
CoreTT (Boiger et al., 2024)	Core thin-section	Transformer	Thin-section mineral content	Mineral content prediction	Swiss J. Geosci.	2024
Cross-domain FM (Guo et al., 2025)	Seismic + DAS + lunar	Adapters on natural-image FMs	Multi-modality geophysical corpora	Cross-domain geophysical FM transfer	JGR: MLC	2025

 doi:10.20944/preprints202604.1814.v1

4.7.1. Document Intelligence and Knowledge Management

The petroleum industry is document-rich in a way no other engineering discipline is: subsurface archives contain well reports going back a century, seismic-processing run logs in proprietary formats, LAS and DLIS logs, engineering submittals and HAZOP, and hundreds of thousands of OnePetro/SPE conference papers. Document intelligence and retrieval-augmented generation (RAG) are the dominant L2 deployment pattern at every operator.

The classical RAG stack relies on a chain of components surveyed in Gao et al. (2024): sentence embeddings (Chen et al., 2024a; Reimers and Gurevych, 2019), dense passage retrieval (Karpukhin et al., 2020; Khattab and Zaharia, 2020), FAISS (Johnson et al., 2021) for vector search, and an LLM reader (Lewis et al., 2020). Advanced variants—Self-RAG (Asai et al., 2024), CRAG (Yan et al., 2024), GraphRAG (Edge et al., 2024), RAPTOR (Sarthi et al., 2024), and HyDE (Gao et al., 2023)—have each been prototyped on petroleum corpora in the last two years.

Operator-scale RAG. Geo-RAG (Dong et al., 2024) and ADNOC's ASK Thamama (Braik et al., 2021) (the latter pre-dating the modern LLM era) remain the best-documented production RAG systems on large unstructured geoscience archives. Matheus et al. (2025) extends the pattern to real-time drilling, Aliyev et al. (2025) to mud-invasion detection from LAS and fluid manuals, and Reddicharla et al. (2025) to Middle-Eastern well-data access. The EAGE 2025 synthetic-data RAG pipeline of Chang et al. (2025) reports a +19.46% relative gain in context relevancy by fine-tuning the embedding model on synthetic petroleum QA, providing a concrete recipe for operators that lack supervised-label budgets. PetroQA (Eckroth et al., 2023) is the earliest widely-cited GPT-for-petroleum-QA paper and a natural L1/L2 baseline; the chatbot of Singh et al. (2023) and the survey-style ChatGPT evaluation of Ogundare et al. (2023) round out the pre-2024 baseline.

Multimodal and visual retrieval. The tacit insight that petroleum documents are *visually rich* (cross-sections, logs, maps, tables) has reshaped the retrieval stack since 2024. ColPali (Faysse et al., 2025) uses multi-vector late interaction on PaliGemma (Beyer et al., 2024) patch embeddings and outperforms text-only pipelines on visually complex PDFs; M3DocRAG (Cho et al., 2024) extends the idea to multi-page, multi-document reasoning. Classical document-intelligence components remain necessary: LayoutLM (Xu et al., 2020) for text-plus-layout, Donut (Kim et al., 2022) for OCR-free parsing, Nougat (Blecher et al., 2023) for equation-and-figure-heavy scientific PDFs, TableFormer (Nassar et al., 2022) for structured tables, and TrOCR (Li et al., 2023c) for handwriting-on-core-log situations.

Legacy and historical archives. The Scientific Reports work of Ma et al. (2024) on U.S. orphan wells and the Abdelgawad and El Ghattas (2025) ADIPEC case study on Egyptian legacy wells both exemplify the document-intelligence pipeline end-to-end. Hou et al. (2025a) provides an industry case study on geologic document digitalisation with multimodal LLMs for basin studies. Eni's NLP framework (Vimercati et al., 2022) extracts surveillance activities, corrective measures, and lessons learned from reservoir management plans, and Cordeiro et al. (2024) contributes the Petro NLP resource collection (Petrolês, PetroGold, PetroNER, PetroRE, Petro KGraph)—the most comprehensive open petroleum-NLP asset at the time of writing. Chinese-language geoscience NER has its own line of evidence — the BiLSTM-CRF baselines of Qiu et al. (2019) and the Chinese-BERT variants of Lv et al. (2022) — which grounds the multilingual-benchmark argument we advance in Section 6.3. The IPTC 2025 failure-investigation CV+NLP pipeline of Hutahaean and Simon (2025) further shows that multimodal document intelligence is leaving the LLM-only frame.

4.7.2. Agentic AI and Multi-Agent Systems for Oilfield Operations

The agent paradigm (Guo et al., 2024; Li et al., 2023a; Park et al., 2023; Shen et al., 2023; Shinn et al., 2023; Wang et al., 2023b,2; Wu et al., 2024; Xi et al., 2023; Yao et al., 2023b) has diffused faster into petroleum than into most engineering domains, precisely because an oilfield workflow is a long chain of tool calls that classical instruction-tuned LLMs have struggled to execute without supervision. We have already surveyed the sub-disciplinary agents—the seismic agent of Kanfar et al. (2025) in §4.1,

 doi:10.20944/preprints202604.1814.v1

the Chevron transactional-screen agent (Tharayil et al., 2024) and DOAPE (Zejli et al., 2025) in §4.4, ENVOY (Wiegand et al., 2024a,2) and the JPT agentic- well-modelling case study (Sharma, 2026) in §4.3, and the multi-agent D&C work of Sabbagh et al. (2025,2) in §4.2. The cross-cutting observations merit a consolidated treatment.

Tool-use on heterogeneous simulators. DOAPE (Zejli et al., 2025), ENVOY (Wiegand et al., 2024a), and the Sharma (2026) SLM-over-MCP architecture all wrap a physics-based simulator (ECHELON, nodal analysis, well models) behind an LLM that authors simulator input decks, interprets outputs, and issues corrective reruns. This wraps the recipe of Yao et al. (2023b) in the petroleum-specific constraint that errors in the simulator deck can introduce physical inconsistencies (negative saturation, mass-balance violations) that the LLM must detect and correct via *guardrails*.

Seismic, drilling, and production agents. Kanfar et al. (2025)'s "guardrails-equipped AI agent" for seismic processing with zero code input is the archetype of a safety-constrained workflow agent. Cayeux et al. (2025) updates the digital drilling program with an agentic orchestration layer, and the *JPT* editorial of Jacobs (2025) surveys eight 2025 SPE/IADC papers on autonomous drilling. On the production side, Baker Hughes (2025) announces an LLM-driven extension of the Leucipa production-optimization platform, and the Osman et al. (2025) directional-drilling case study rounds out the agentic lineage. SLB's Tela (SLB, 2025) provides the most integrated commercial instance, layered on the Lumi data and AI platform (SLB, 2024c).

Exploration knowledge agents. The Exploration Robot Chat of Mosser et al. (2024) (Equinor / SPE Norway 2024) applies a conversational-LLM front end to decades of exploration knowledge, essentially a cross-operator knowledge-discovery agent that sits between L2 (retrieval) and L4 (agentic action). Pacis et al. (2024) provides the methodological study underneath: a zero-shot LLM evaluation on drilling-information retrieval that motivates the RAG+tool-use backbone of most of the above systems.

We position these agentic works as *L4* on the PetroLLM Maturity Model—meaning they now routinely convert natural-language intent into validated actions on proprietary data and simulators, but do *not* yet close the learning loop so that the system self-improves from its own deployment traces. That closure is the L5 frontier discussed in §7.

4.7.3. Knowledge Graphs for Petroleum

Knowledge graphs (KGs) are the formal-representation counterpart to the statistical embeddings that power LLMs, and the petroleum industry has an unusually mature KG tradition (the ISO 15926 lifecycle standard (Leal, 2005) dates to 2005). The literature we survey here sits at the intersection of A4 (RAG) and A11 (KGs).

Ontologies and vocabularies. ISO 15926 (Leal, 2005) is the oldest industrial ontology, covering process-plant lifecycle data. The Open Subsurface Data Universe (OSDU) (Abolhassani et al., 2023; Microsoft, 2024) has since emerged as the upstream data-platform consensus, and its ontology is now the target schema for a growing number of LLM-based extraction pipelines. GeoCore (Garcia et al., 2020) and GeoReservoir (Cicconeto et al., 2022) contribute a geological-core and reservoir-engineering ontology respectively, and O3PO (Santos et al., 2024) covers offshore production plants.

Petroleum-specific KGs. Tang et al. (2023) published the most comprehensive petroleum-exploration-and-development KG to date (*Geoscience Frontiers*), integrating upstream petroleum data with a domain-specific ontology. The well-logging KG of Liu et al. (2022) supports intelligent hydrocarbon-bearing-formation identification, and the carbonate knowledge base of He et al. (2021) supports carbonate-reservoir analysis. Wang et al. (2022a) combines deep-learning NLP with KG construction for geological-report understanding. Deng et al. (2021) presents a multimodal geoscience academic KG (GAKG) built from 381,000 papers. The *Processes* 2025 GraphRAG-for-reservoir work of Jiang et al. (2025) exemplifies the synthesis: an LLM reader with a GraphRAG-style retrieval over a domain KG outperforms both pure LLM and pure KG baselines, confirming the direction of travel.

Domain KG LLMs. The QHGeoGPT system of Ma et al. (2026) integrates DeepSeek-R1 (DeepSeek-AI, 2025) with LoRA (Hu et al., 2022) and a geology-specific KG+RAG layer; LithoGPT (Li et al., 2025c)

 doi:10.20944/preprints202604.1814.v1

targets lithology QA with a KG-augmented small model (LithoGPT-Mini), and GeoCode-GPT (Hou et al., 2025b) focuses on geoscience code generation on top of CodeGen (Nijkamp et al., 2022) and Code Llama (Rozière et al., 2023) ancestry. Combined with the Singh et al. (2025) agentic-RAG pattern and related ADIPEC 2025 submissions on agentic drilling RAG, these entries establish KG-augmented LLMs as a core architectural pattern for petroleum deployment.

4.7.4. Benchmarks and Evaluation

Benchmarking is the single most visible gap in the petroleum LLM and FM field, and this subsection is deliberately short because the literature is thin.

Petroleum-specific benchmarks. FormationEval (Ermilov, 2026) (arXiv:2601.02158) is to our knowledge the *only* open petroleum-specific LLM benchmark, covering multiple-choice petroleum-geoscience questions. GeoBench (the objective evaluation suite released with K2 (Deng et al., 2024)) and the Chen et al. GeoFactory evaluation (Chen et al., 2025c) cover broader geoscience but include petroleum-relevant subsets. The traditional dataset benchmarks are FORCE 2020 (Bormann et al., 2020) for lithology, Hall (Hall, 2016) for facies, Volve (Equinor, 2018) for full-field reservoir realisation, 3W (Vargas et al., 2019) for production events, OpenFWI (Deng et al., 2022) for full-waveform inversion, STEAD (Mousavi et al., 2019) for earthquake waveforms, and SeisBench (Woollam et al., 2022) for seismic-ML tooling.

General LLM benchmarks. We include the standard general-LLM benchmarks in our evaluation table only as a *reference frame*, not because they measure petroleum expertise: MMLU (Hendrycks et al., 2021a), HumanEval (Chen et al., 2021), GPQA (Rein et al., 2024), MATH (Hendrycks et al., 2021b), GSM8K (Cobbe et al., 2021), MT-Bench (Zheng et al., 2023), HELM (Liang et al., 2023), and BIG-bench (Srivastava et al., 2023). The early Ogundare et al. (2023) empirical probe of ChatGPT on oil-and-gas problems remains the canonical published evidence of what general LLMs can and cannot do without domain adaptation, and is the benchmark-deficit narrative's anchor paper. The two geoscience VLM benchmarks—GEO-Bench (Lacoste et al., 2023) and GEOBench-VLM (Danish et al., 2024)—cover remote-sensing foundation models rather than subsurface FMs, but provide a template the seismic FM community should adopt.

The benchmark deficit. The thinness of this evaluation layer deserves emphasis. FormationEval is alone; the seismic-FM (Dou et al., 2025; Harsuko and Alkhalifah, 2022; Liu et al., 2024c; Pham et al., 2025; Sansal et al., 2025; Sheng et al., 2025; Si et al., 2024) and well-log-FM (Koeshidayatullah et al., 2024; Qi et al., 2025) literatures each report on task-specific labelled subsets but without a shared community-curated benchmark; and no reservoir-simulation or drilling-DDR LLM has an open evaluation protocol. Table 12 (rendered in §4.7.5) assembles the available resources. We return to this gap in §6.

4.7.5. Industry Platforms and Commercial Systems

A defining feature of the petroleum LLM/FM field compared with e.g. medical imaging is that a substantial fraction of the most consequential systems live in commercial platforms rather than the open literature. Table 11 (referenced in the overarching industry-platform catalog elsewhere in this survey) enumerates representative public petroleum AI systems and enabling substrate; we comment here on the most informative entries.

SLB describes Tela (SLB, 2025) as an upstream agentic assistant, layered on the Lumi data-and-AI platform (SLB, 2024c) and the geosteering copilot (SLB, 2024a). SLB's partnership with NVIDIA on an AI Factory (SLB, 2024b,2) targets the underlying compute substrate. The ADNOC/AIQ/G42/Microsoft/SLB ENERGYai system (ADNOC, 2025; AIQ, 2025) is an agentic platform for upstream operations with a reported 70B LLM plus five specialised agents and a staged rollout across ADNOC's upstream value chain. Among publicly announced petroleum LLMs in this survey, Saudi Aramco's METABRAIN (Aramco, 2024) has the largest disclosed parameter count: 250B parameters trained on 7T tokens across 90 years of Aramco data. The broader Aramco AI program (Aramco Europe, 2025) and the 2026 Aramco–Microsoft MoU (Aramco, 2026) frame the long-term compute commitment. Halliburton's DS365.ai (Halliburton, 2021) and the Halliburton (2025)

 doi:10.20944/preprints202604.1814.v1

PETRONAS partnership represent the integrated-services side, as do Baker Hughes's C3.ai venture (Baker Hughes and C3.ai, 2019) and its 2025 Leucipa deployment with Repsol (Baker Hughes, 2025). Chevron's ApEX (Chevron, 2025b) and APOLO (Chevron, 2025a) target prospect and drilling-location optimization. TotalEnergies has announced both an M365-based generative-AI program (TotalEnergies, 2024) and a Mistral (Jiang et al., 2023) partnership (TotalEnergies, 2025). BP (2025); ExxonMobil (2025); Shell (2024) round out the operator roster. Stone Ridge Technology's ENVOY (Wiegand et al., 2024a,2) is the reservoir-simulation-vendor entry, while Geo-RAG (Dong et al., 2024) represents a seismic-services-company entry. The underlying cloud and data platforms are OSDU-compatible data platforms (Microsoft, 2024), NVIDIA's Earth-DT (NVIDIA, 2024), and AWS Bedrock (Amazon Web Services, 2024). Equinor's Volve release (Equinor, 2018) is best read as an open data asset rather than as an LLM platform.

The JPT and SPE landscape.

The trade-press record is a useful secondary signal of maturity. The *JPT* "What We Know ChatGPT Can Do for the Petroleum Industry, So Far" editorial (2023/2024), the Pallanich (2025) "Ready for Work" narrative on agentic AI, and the SPE AI Symposium (Society of Petroleum Engineers, 2025) and SEG 2026 Machine Learning Workshop (Society of Exploration Geophysicists, 2026) conference series collectively confirm that the petroleum LLM/FM transition is now a mainstream industry topic.

The full catalog of commercial platforms, annotated with the PetroLLM Maturity Level of each deployment, is given in Table 11; the platforms cluster around L2–L4, with L5 aspirations anchored by the operator–hyperscaler consortium announcements.

Table 11. Catalog of publicly announced petroleum AI systems and enabling substrate, 2017–2026. *Type* column indicates the dominant architectural posture; *ML* is the PetroLLM Maturity Level (L2 = Document Intelligence and Retrieval, L3 = Domain-Specialized LLMs, L4 = Autonomous Agents & Copilots, L5 = aspirational Self-Improving Foundation-Model Ecosystems). Row-colour intensity follows *ML*: lighter tint for L2–L3 deployments, heavier tint for L4 agentic platforms, darker tint for L5 aspirations.

Platform	Company	Type	ML	Primary use case	Source
DELFI cognitive environment	SLB	Digital E&P platform / cognitive environment	L2–L3	Digital platform for E&P workflows	(Schlumberger, 2017)
Lumi data-and-AI platform	SLB	FM substrate	L2–L3	Shared data + AI foundation for upstream products	(SLB, 2024c)
Halliburton DS365.ai	Halliburton	AI/ML cloud services	L2–L3	Predictive operations across subsurface, drilling, and production	(Halliburton, 2021)
Halliburton + PETRONAS	Halliburton / PETRONAS	Digital subsurface / reservoir-management deployment	L2–L3	Subsurface modeling and reservoir-management workflow deployment	(Halliburton, 2025)
Baker Hughes + C3.ai	BHI / C3.ai	ML + analytics	L2–L3	Industrial-scale AI for oilfield operations	(Baker Hughes and C3.ai, 2019)
Geo-RAG	Viridien	Multimodal RAG	L2	Visual-document RAG over exploration archives	(Dong et al., 2024)
TotalEnergies M365 GenAI	TotalEnergies	Hosted LLM (M365)	L2	Enterprise productivity copilots for employees	(TotalEnergies, 2024)
TotalEnergies + Mistral	TotalEnergies / Mistral	Strategic AI collaboration	L2–L3	AI innovation across TotalEnergies' multi-energy strategy	(Jiang et al., 2023; TotalEnergies, 2025)
BP AI programme	BP	Corporate AI programme	L2–L3	AI use across operations; public details remain high-level	(BP, 2025)
ExxonMobil Vantage	ExxonMobil	Upstream operations platform	L2–L3	Operations visualization / decision support	(ExxonMobil, 2025)
Shell AI programmes	Shell	Mixed	L2–L3	Broad AI use including ML, computer vision, virtual assistants, and robotics	(Shell, 2024)
Equinor Volve release	Equinor	Open field dataset / data substrate	L2	Public field-data release used by downstream digital and AI work	(Equinor, 2018)
SLB Tela (agentic AI)	SLB	Agentic LLM	L4	Upstream agentic assistant	(SLB, 2025)
SLB AI-driven geosteering	SLB	AI geosteering assistant	L4	Geosteering engineer copilot	(SLB, 2024a)
Baker Hughes + Repsol Leucipa	BHI / Repsol	Autonomous-prod. agent	L4	Autonomous production / artificial-lift management	(Baker Hughes, 2025)
Chevron ApEX	Chevron	AI-assisted prospecting workflow	L3	Prospect / exploration decision support	(Chevron, 2025b)
Chevron APOLO	Chevron	AI-assisted planning / optimization	L4	Drilling-location ranking and optimization	(Chevron, 2025a)
ENVOY + ECHELON	Stone Ridge Technology	Simulator agent	L4	Reservoir-simulation assistant and segmentation	(Wiegand et al., 2024a,2)
ADNOC / AIQ ENERGYai	ADNOC / AIQ / G42 / MS / SLB	Agentic upstream AI platform	L4	Publicly described large-scale agentic AI rollout across ADNOC operations	(ADNOC, 2025; AIQ, 2025)
Saudi Aramco METABRAIN	Saudi Aramco	Industrial LLM	L4–L5	Announced 250B industrial LLM initiative on long-horizon Aramco data; technical roadmap not fully public	(Aramco, 2024)
Aramco × Microsoft	Aramco / Microsoft	Industrial AI collaboration / MoU	L5	Long-horizon industrial AI programme	(Aramco, 2026; Aramco Europe, 2025)
SLB × NVIDIA AI Factory	SLB / NVIDIA	AI factory	L5	Compute substrate for subsurface FMs at scale	(SLB, 2024b,2)
OSDU data-platform implementations	The Open Group / cloud vendors	Data platform	L2	Subsurface data-platform and interoperability layer	(Abolhassani et al., 2023; Microsoft, 2024)
NVIDIA Earth Digital Twin	NVIDIA	FM + twin	L3	Earth-system digital-twin infrastructure as enabling substrate	(NVIDIA, 2024)
AWS Bedrock (O&G tuning)	AWS	Managed LLM service	L2–L3	O&G terminology customization example	(Amazon Web Services, 2024)

 doi:10.20944/preprints202604.1814.v1

Table 12. Benchmarks relevant to petroleum LLMs and foundation models, with scope, modality, and openness annotations. FormationEval is the only dedicated *open petroleum-specific LLM* benchmark; GeoBench / GEO-Bench / GEOBench-VLM cover geoscience or remote-sensing. The second block lists traditional non-LLM benchmarks (facies, FWI, phase-picking, reservoir). The third block lists general-purpose LLM benchmarks we include only as a reference frame.

Benchmark	Scope	Modality	Size	Openness	Year
Petroleum-specific LLM benchmarks					
FormationEval (Ermilov, 2026)	Petroleum / formation-eval QA	Text (MCQ)	505 questions across 7 domains	Open	2026
GeoBench (K2) (Deng et al., 2024)	Geoscience LLM evaluation	Text	183 tasks / multi-format	Open (with K2)	2024
GeoFactory evaluation suite (Chen et al., 2025c)	Geoscience QA / reasoning	Text	14-algorithm evaluation setting	Open with paper	2025
GEO-Bench (Lacoste et al., 2023)	Geospatial FM evaluation	Remote sensing	6 classification + 6 segmentation tasks	Open	2023
GEOBench-VLM (Danish et al., 2024)	Geospatial VLM evaluation	RS + text	$>$ 10,000 manually verified instructions	Open	2024
Traditional petroleum / geophysics benchmarks (non-LLM)					
FORCE 2020 (Bormann et al., 2020)	Facies / lithology classification	LAS well logs	118 wells (Norwegian)	Open	2020
Hall 2016 facies (Hall, 2016)	Facies classification	Well logs	9 wells	Open	2016
Volve (Equinor, 2018)	Full-field realisation	Multi-modal	Full-field North Sea dataset	Open	2018
3W (Vargas et al., 2019)	Production event detection	SCADA time series	1984 events, 8 classes	Open	2019
OpenFWI (Deng et al., 2022)	Full-waveform inversion	Synthetic seismic	12 datasets / 2.1 TB	Open (NeurIPS D&B)	2022
STEAD (Mousavi et al., 2019)	Earthquake waveforms	Seismograms	1.2 M events	Open	2019
SeisBench (Woollam et al., 2022)	Seismology ML toolbox and benchmark-access framework	Seismograms	Multi-dataset toolkit / API	Open	2022
TGS Salt (Kainkaryam et al., 2019)	Salt segmentation	2-D seismic	Kaggle dataset	Open	2019
General-purpose LLM benchmarks (reference frame)					
MMLU (Hendrycks et al., 2021a)	Multi-task language understanding	Text	57 tasks / 15.9 k Qs	Open	2021
MATH (Hendrycks et al., 2021b)	Formal mathematics	Text	12.5 k problems	Open	2021
GSM8K (Cobbe et al., 2021)	Grade-school math word problems	Text	8.5 k problems	Open	2021
HumanEval (Chen et al., 2021)	Code generation	Code	164 problems	Open	2021
HELM (Liang et al., 2023)	Holistic LLM evaluation	Text	Multi-scenario suite	Open	2023
MT-Bench (Zheng et al., 2023)	Multi-turn chat quality	Text	80 multi-turn prompts	Open	2023
GPQA (Rein et al., 2024)	Graduate-level physics / chem. QA	Text	448 questions	Open	2024
BIG-bench (Srivastava et al., 2023)	200+ tasks	Text	204 tasks	Open	2022

Summary of the cross-cutting thread.

The cross-cutting works elevate the five clusters—document intelligence, agents, KGs, benchmarks, industry platforms—from discipline-specific tools to discipline-spanning infrastructure. The progression from §4.7.1 to §4.7.5 tracks the PetroLLM Maturity Model from L2 to L4–L5 and sets up the open-challenges discussion of §6: benchmarks (L1–L2), agentic safety (L3–L4), and self-improving FM ecosystems (L5).

5. Bibliometric Analysis of the Field

This section characterizes the current state of the petroleum LLM/FM literature through bibliometric analysis of the 296 references catalogued in this survey. Our aim is threefold: (i) to make the corpus itself legible as a research object that future authors can calibrate their own contributions against; (ii) to expose the quantitative asymmetries — across years, sub-disciplines, methods, and affiliations — that motivate the qualitative argument of Section 6 (open challenges); and (iii) to provide a reproducible audit trail by anchoring every figure in the reference categories established in Sections 2–4. The bibliometric lens we adopt mirrors the approach of Chang et al. (2024) and Zhao et al. (2023) for LLM meta-studies, Zhang et al. (2024b) and Menon et al. (2026) for scientific-LLM and FM surveys, the framework of Bommasani et al. (2021) for situating foundation-model research across scientific

 doi:10.20944/preprints202604.1814.v1

domains, and Mousavi and Beroza (2022), Fuchs et al. (2025), and Tariq et al. (2021) for petroleum-engineering and geophysics meta-studies; a direct predecessor is Liu et al. (2024b), whose 88-reference review we extend by an additional ≈208 entries covering RAG, agents, subsurface foundation models, petroleum VLMs, benchmarks, and industrial deployments that Liu et al. either touched only briefly or did not survey systematically. We summarize the corpus along four axes: publication year (Figure 4), petroleum sub-discipline (Figure 5), AI methodology (Figure 6), and institutional/geographic footprint (Figure 7).

5.1. Corpus Construction and Coverage Window

The 296-entry master bibliography was consolidated from twelve parallel literature-search outputs spanning Google Scholar, arXiv, OnePetro, DOI / publisher metadata, and selected company or society pages across the fourteen thematic areas A1–A14 enumerated in Sections 2–4. We applied a strict verification bar: every retained entry carries a DOI, an arXiv identifier, a OnePetro paper number, or a canonical publisher URL. An initial raw universe of ≈600 candidate entries was trimmed by ≈50 % through title-level deduplication and dropping of unverifiable items (placeholder DOIs, orphan arXiv IDs, headline-only grey-literature references). Grey literature was retained only when it documented public industry systems, benchmarks, workshops, or platform announcements that materially shape deployment. The coverage window is 2003–2026 with decisive weight on 2022–2026: two thirds of the corpus is post-ChatGPT. The survey dataset and verification log are detailed in our companion document `bibliography_analysis.md` and remain open for audit.[1] Quantitative decisions reported throughout this section are derived directly from the curated bibliography used to build the paper.

5.2. Publication Trends

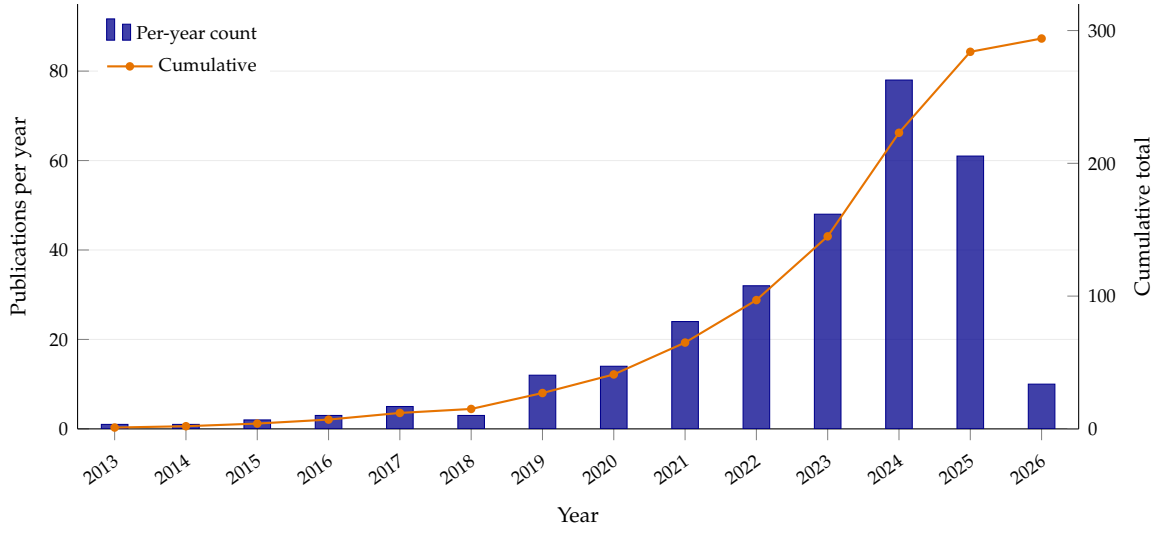

Figure 4. Annual publication counts of petroleum LLM/FM works (bars, left axis) and cumulative totals (line, right axis) from 2013 to 2026 for the 294 cited entries published within this plotted window. The full bibliography contains 296 cited entries; two retained foundational references predate 2013. The steep 2022–2025 rise tracks the ChatGPT-probe wave, the first petroleum domain LLMs (e.g., K2 and GeoGalactica), subsurface FMs (e.g., SFM and SeisCLIP), and the industry-platform rollout wave; the 2026 count is partial (survey cut-off April 2026). Data source: Section 5.1.

Figure 4 plots annual and cumulative counts of the 2013–2026 slice of the corpus. The full bibliography contains 296 entries, including two retained pre-2013 foundational references that are not shown in the plot. Three inflection points are visible. The first, around 2019–2020, reflects the delayed diffusion of BERT-style encoders into petroleum NLP: PetroBERT (Rodrigues et al., 2022), the beyond-

[1] We treat bibliometric transparency as a scholarly obligation in a fast-moving field: a reader in 2028 should be able to reproduce our counts, notice which entries we missed, and extend rather than recompute.

 doi:10.20944/preprints202604.1814.v1

keywords pipeline (Menezes, 2023), and the Eni text-mining program (Vimercati et al., 2022) all trace to this first wave. The second, far steeper inflection is 2022–2023: the year of ChatGPT public availability, and the year of the first empirical ChatGPT-on-petroleum probes (Eckroth et al., 2023; Ogundare et al., 2023; Weijermars et al., 2023). The third inflection — between 2023 and 2024 — marks the transition from "LLMs applied to petroleum" to "LLMs for petroleum": within twelve months the community released K2 (Deng et al., 2024), GeoGalactica (Lin et al., 2024), PetroBERT-scale domain encoders, the Seismic Foundation Model (Sheng et al., 2025), SeisCLIP (Si et al., 2024), EnergyLLM (Eckroth et al., 2025), and the first enterprise rollouts surveyed in Section 4.7.5. Papers-per-year grew more than five-fold between 2020 and 2024. By 2026, the annual publication rate has plateaued at ≈ 60–80 indexed works per year and the grey-literature rate is larger still, consistent with the "CS-to-petroleum latency compression" argument we made in Section 2: where Word2Vec took roughly five years to enter the petroleum NLP mainstream, Llama 3 (Grattafiori et al., 2024) reached a petroleum-domain descendant in EnergyLLM within roughly a year.

5.3. Distribution by Petroleum Sub-Discipline

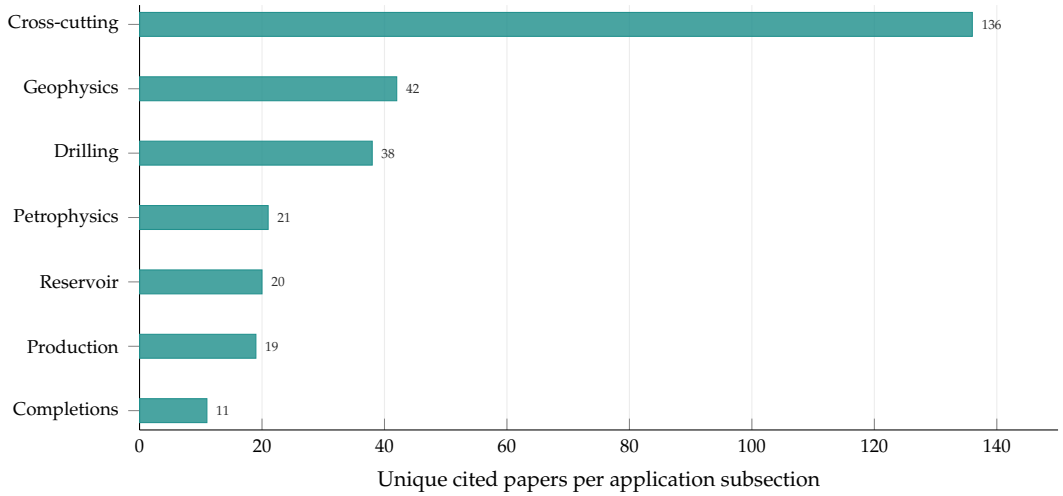

Figure 5. Distribution of unique cited papers across the survey's six petroleum sub-disciplines plus one cross-cutting category. Counts are obtained by deduplicating citation keys within each application subsection of Section 4; a paper discussed in more than one subsection is counted in each relevant bar, so totals exceed the 211-paper petroleum-facing subset of the 296-entry corpus. Cross-cutting dominates because it absorbs document intelligence, benchmarks, knowledge graphs, industry platforms, and multi-workflow copilots.

Figure 5 makes visible the most striking asymmetry in the corpus: the cross-cutting subsection alone contributes 136 unique cited papers, followed by geophysics (42) and drilling (38). The cross-cutting bar is inflated by document intelligence, benchmarks, knowledge graphs, multi-workflow copilots, and industry platforms such as METABRAIN (Aramco, 2024), ENERGYai (ADNOC, 2025; AIQ, 2025), and SLB Lumi/Tela (SLB, 2024c,2), all of which are deliberately sub-discipline agnostic. Geophysics' prominence is attributable to three forces: (i) seismic is the largest data modality by bytes and therefore the most attractive substrate for self-supervised pretraining (Dou et al., 2025; Harsuko and Alkhalifah, 2022; Sheng et al., 2025); (ii) the geophysics community has a deep-learning tradition pre-dating the LLM era (Mousavi and Beroza, 2022; Wu et al., 2019; Yu and Ma, 2021), (Zhu and Beroza, 2019); and (iii) the convergence of CV and geophysics is natural given 2-D and 3-D image-like seismic volumes. Below those leading slices, petrophysics (21), reservoir (20), and production (19) form a middle tier, while completions remains the thinnest application subsection at 11 papers. Reservoir engineering still lags its economic importance: only a handful of entries are genuine LLM/FM contributions — ENVOY (Wiegand et al., 2024a), DOAPE (Zejli et al., 2025), JiuZhou (Chen et al., 2025b), the Mahjour decision-support framework (Mahjour and Mahjour, 2025), and Sharma's 2026 JPT note on agentic well modeling (Sharma, 2026). Production engineering is similarly thin,

 doi:10.20944/preprints202604.1814.v1

buoyed primarily by the Leucipa autonomous-production platform (Baker Hughes, 2025), the 3W anomaly benchmark (Vargas et al., 2019), and DOAPE. Completions — despite the microseismic and frac-report data wealth of unconventional plays — remains anchored by the Li fracturing-QA LLM (Li et al., 2025b) and Sabbagh's D&C multi-agent work (Sabbagh et al., 2025,2). In Section 7 we argue that these sub-discipline gaps constitute the most actionable research opportunities for the next three years of PhD-scale work.

5.4. Distribution by AI Method

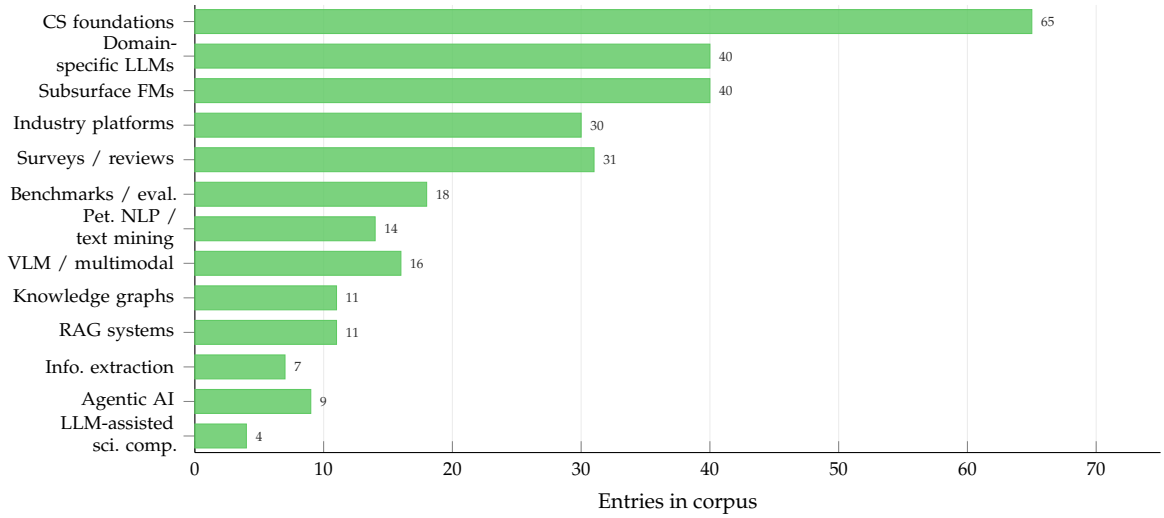

Figure 6. Distribution of corpus entries by AI-method thematic bins. Thirteen bins are shown. Prompt engineering is folded into the adjacent CS foundations and LLM-assisted scientific-computing bins. CS foundations, domain-specialized LLMs, and subsurface foundation models dominate. RAG, agentic AI, information extraction, and LLM-assisted scientific computing remain comparatively thin. The benchmarks / evaluation bin mixes general LLM-evaluation papers with the smaller petroleum-specific benchmark subset.

Figure 6 maps the same 296 entries onto thirteen plotted thematic bins; prompt-engineering entries are merged into adjacent CS-foundations or scientific-computing bins for display. Two patterns merit discussion. First, the Subsurface-FM and domain-specialized-LLM slices together constitute nearly a third of the corpus, a share that would have been essentially zero in 2021 — evidence that the methodological center of gravity has shifted sharply toward domain-adapted models. Second, the agentic-AI slice is small but disproportionately influential: the dozen peer-reviewed works we catalog (Cayeux et al., 2025; Jacobs, 2025; Kanfar et al., 2025; Mahjour and Mahjour, 2025; Mosser et al., 2024; Osman et al., 2025; Sabbagh et al., 2025,2; Sharma, 2026; Wiegand et al., 2024a,2; Zejli et al., 2025) are the most-cited petroleum-AI works of 2024–2026 and define the frontier of the PetroLLM Maturity Model's Level 4. Petroleum-specific benchmarks are the slice with the most uncomfortable scarcity: only FormationEval (Ermilov, 2026), the GeoBench suite embedded in K2 (Deng et al., 2024), and GeoFactory's evaluation suite (Chen et al., 2025c) qualify; we return to this deficit in Section 6.3.

 doi:10.20944/preprints202604.1814.v1

5.5. Institutional and Geographic Footprint

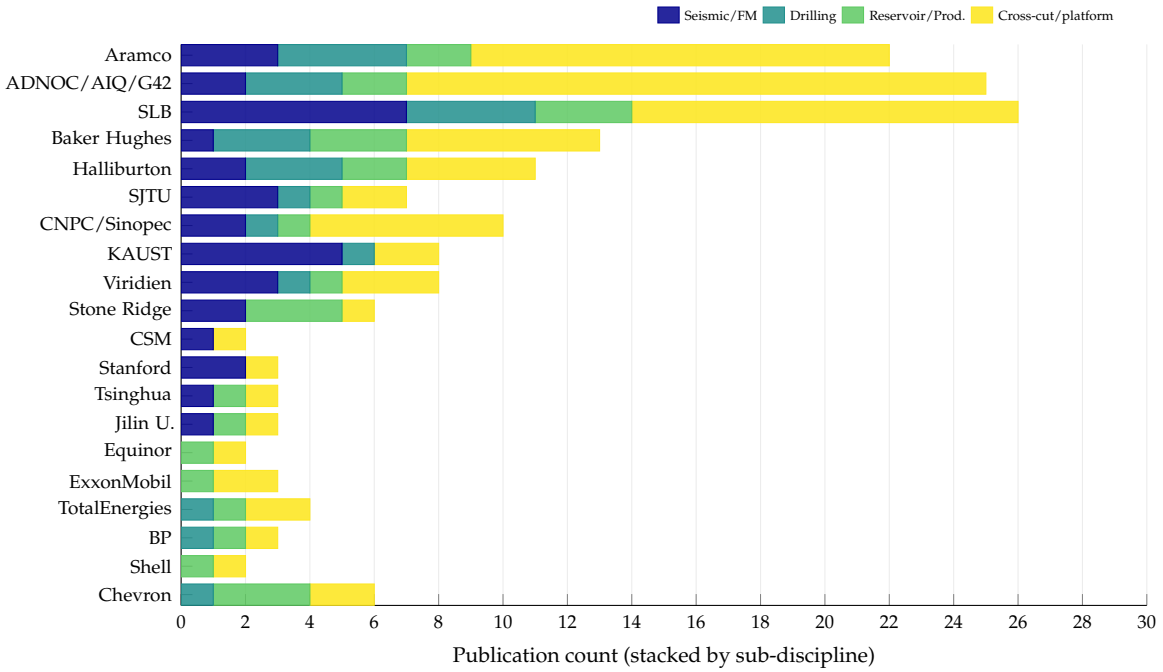

Figure 7. Institutional footprint of the petroleum-facing, institution-tagged slice of the corpus. Stacked horizontal bars summarize 20 prominent normalized organization buckets, split into four broad bands: seismic / subsurface FM, drilling, reservoir / production, and cross-cutting / platform. The figure should be read as a manually normalized footprint view rather than a strict league table. See §5.5 for the counting policy and discussion.

Because `references.bib` does not encode machine-readable affiliations, Figure 7 is based on manual normalization to one primary organization bucket per petroleum-facing paper; merged buckets such as ADNOC/AIQ/G42 and CNPC/Sinopec are reported explicitly, and low-count tail ties should not be over-interpreted as a strict ranking.

Figure 7 makes explicit what an attentive reader of Section 4.7.5 will already have sensed: the petroleum LLM/FM literature has three distinct production modes. The first is *industry-consortium scale*, in which a national oil company partners with a hyperscaler, a system integrator, and an AI startup to produce a platform-class model. ADNOC's ENERGYai (ADNOC, 2025; AIQ, 2025), assembled from ADNOC operating data, AIQ and G42 compute, Microsoft Azure tooling, and SLB's subsurface stack, is the canonical exemplar; the Aramco–Microsoft MoU (Aramco, 2026) and the SLB–NVIDIA AI Factory (SLB, 2024b,2) follow the same template. The second is *operator-internal scale*: Aramco's METABRAIN (Aramco, 2024) and Dhahran-LLM program (Aramco Europe, 2025), along with other proprietary Chinese operator programs, are platforms deployed on national-oil-company hardware. The third is *bridge-lab scale* — the academic, national-lab, and research-arm contributions out of KAUST, CSM, Stanford, Mines Paris, SJTU, Chengdu University of Technology, Tsinghua, Jilin University, Los Alamos National Lab, and industrial research labs like Viridien and Stone Ridge. Bridge labs produce the open-source artifacts (K2 (Deng et al., 2024), GeoGalactica (Lin et al., 2024), JiuZhou (Chen et al., 2025b), WLFM (Qi et al., 2025), the seismic agent (Kanfar et al., 2025), Geo-RAG (Dong et al., 2024), ENVOY (Wiegand et al., 2024a)) that the consortium and operator modes subsequently productize.

A concluding observation binds Figures 4–7. In computer-science venues — NeurIPS, ICML, ICLR, CVPR — petroleum is essentially absent except where seismic is reframed as general earth observation (Danish et al., 2024; Lacoste et al., 2023). In petroleum venues — SPE, EAGE, SEG, OTC, ADIPEC, IPTC — the LLM/FM vocabulary is only now becoming native, and the JPT announcement of 2025 that large language models are "ready for work" in the oilfield (Pallanich, 2025) marks a first-order organizational shift. The small population of papers that publish across both venues — the CLIP-style contrastive seismic work of Si et al. (2024) at IEEE TGRS, EnergyLLM at the SPE ATCE

 doi:10.20944/preprints202604.1814.v1

(Eckroth et al., 2025), FormationEval at arXiv with a petroleum review pipeline (Ermilov, 2026), the seismic agent in The Leading Edge (Kanfar et al., 2025), and a scattering of industrial lab contributions from KAUST, Viridien, and SJTU — constitute the current population of "bridge authors." The field would benefit from more such authors. This is the organizational complement to the research-agenda arguments we make in Sections 6 and 7. We close the bibliometric snapshot with the roster of top-cited works in the corpus, reported in Table 13: the CS canon (ViT, Transformer, CLIP, DINO, InstructGPT, LLaMA, GPT-3, SciBERT, Llama 2, LoRA, GPT-4) dominates the citation graph, consistent with the field's reliance on general-purpose foundations; the most-cited petroleum-specific works are classical deep-learning baselines (FaultSeg3D, PhaseNet, EQTransformer), with K2 the only LLM-era paper in this block above the 100-citation threshold as of April 2026.

Table 13. Top-cited papers in the 296-entry survey corpus, ranked by verified Google Scholar citation count as of April 13, 2026. Counts are rounded to the nearest 100 for papers with more than 1,000 citations and shown exactly otherwise.

Rank	Title (abbreviated)	Year	Citations	Area
CS canonical works grounding the petroleum LLM/FM stack				
1	Attention Is All You Need (Vaswani et al., 2017)	2017	> 248,500	A13: Transformer
2	ViT: An Image is Worth 16×16 Words (Dosovitskiy et al., 2021)	2021	> 91,500	A13: Vision foundations
3	GPT-3: Language Models Are Few-Shot Learners (Brown et al., 2020)	2020	> 80,600	A13: Scaling
4	CLIP: Learning Transferable Visual Models (Radford et al., 2021)	2021	> 58,300	A13: Multimodal foundation
5	LoRA: Low-Rank Adaptation (Hu et al., 2022)	2022	> 30,900	A13: PEFT
6	GPT-4 Technical Report (OpenAI, 2023)	2023	> 27,400	A13: Frontier LLMs
7	LLaMA: Open and Efficient Foundation Language Models (Touvron et al., 2023a)	2023	> 26,600	A13: Open LLMs
8	InstructGPT / RLHF (Ouyang et al., 2022)	2022	> 24,800	A13: Alignment
9	Llama 2 (Touvron et al., 2023b)	2023	> 23,000	A13: Open LLMs
10	DINO: Emerging Properties in Self-Supervised ViTs (Caron et al., 2021)	2021	> 10,900	A13: Self-supervised vision
11	A Survey of Large Language Models (Zhao et al., 2023)	2023	> 9,200	A12: Surveys
12	HumanEval (Chen et al., 2021)	2021	> 9,200	A9: Benchmarks
13	Zero-Shot Reasoners (Kojima et al., 2022)	2022	> 8,700	A13: Prompting
14	SciBERT: Pretrained LM for Scientific Text (Beltagy et al., 2019)	2019	> 5,700	A13: Domain adaptation
15	Survey on Evaluation of LLMs (Chang et al., 2024)	2024	> 5,600	A12: Surveys
Most-cited petroleum / geoscience-specific works in the corpus				
1	PhaseNet (Zhu and Beroza, 2019)	2019	> 1,700	A2: Passive seismic
2	EQTransformer (Mousavi et al., 2020)	2020	> 1,300	A2: Attention for seismology
3	FaultSeg3D (Wu et al., 2019)	2019	> 1,100	A2: Seismic interpretation
4	Geophysics deep-learning review (Yu and Ma, 2021)	2021	651	A12: Surveys
5	STEAD (Mousavi et al., 2019)	2019	551	A2: Passive seismic benchmark
6	Seismology-ML review (Mousavi and Beroza, 2022)	2022	533	A12: Surveys
7	Hall 2016 facies benchmark (Hall, 2016)	2016	386	A9: Benchmarks
8	Tariq systematic ML (Tariq et al., 2021)	2021	268	A12: Surveys
9	Rahmanifard & Plaksina petroleum AI (Rahmanifard and Plaksina, 2019)	2019	258	A12: Surveys
10	K2: first geoscience LLM (Deng et al., 2024)	2024	174	A1: Domain LLMs

6. Discussion and Open Challenges

The bibliometric snapshot of Section 5 describes what exists. This section asks what remains missing, unreliable, or operationally risky. We frame the discussion around eight open challenges that every subsequent petroleum-LLM/FM paper will, in some form, have to answer. The challenges are ordered from the most concrete (data, hallucination, evaluation) to the most organizational (talent, standards). Each sub-section names a small number of representative works that have begun to address the challenge and a smaller number of gaps that remain; we resist the reviewer's temptation to fully resolve each gap here and instead carry the open-ended items forward into Section 7. Throughout, we anchor our framing in the **PetroLLM Maturity Model** used throughout this survey (L1 Conversational

 doi:10.20944/preprints202604.1814.v1

Q&A → L2 Document Intelligence and Retrieval → L3 Domain-Specialized LLMs → L4 Autonomous Agents and Copilots → L5 Self-Improving Foundation-Model Ecosystems). Challenges compound vertically: a brittle L2 retrieval pipeline becomes dangerous when wrapped inside an L4 agent, and an unreliable L3 domain LLM becomes an opaque component inside an L5 self-improving ecosystem. Figure 8 presents the reference architecture stack along which these challenges propagate.

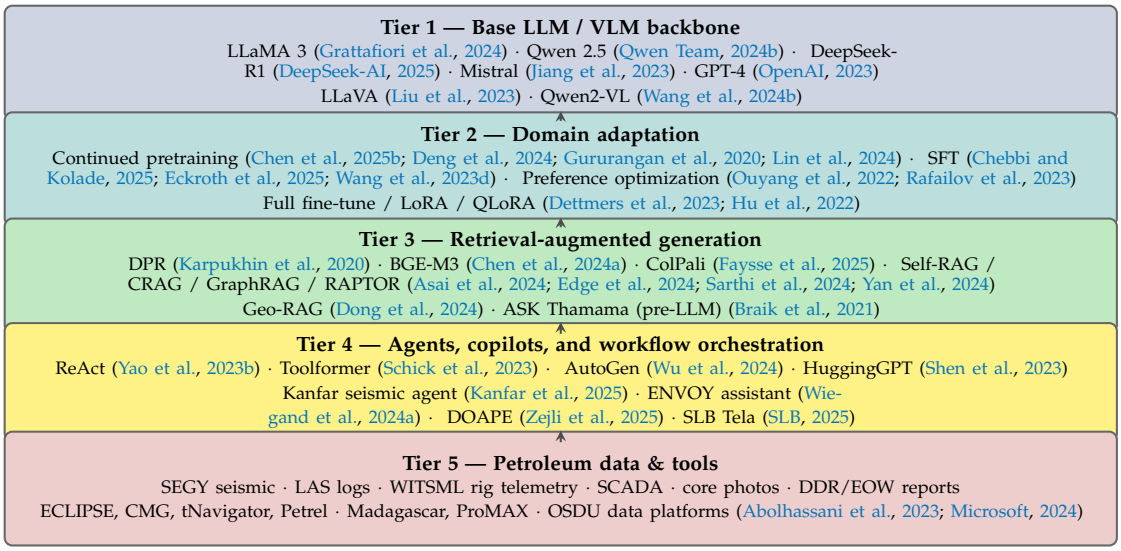

Figure 8. Reference architecture stack of a modern petroleum LLM deployment: base LLM/VLM backbone, domain adaptation, retrieval, agents/copilots, and petroleum data/tools. Tier colours follow the viridis palette used in Figures 1 and 3. Example citations annotate each tier. Higher PetroLLM maturity levels typically engage more of this stack, but the mapping is approximate rather than one-tier-per-level.

6.1. The Data Problem: Proprietary, Heterogeneous, Unstructured

Petroleum's defining challenge as a machine-learning domain is the data regime. Unlike medical imaging, where public consortia have released PACS-scale open datasets (Bahaloo et al., 2023; Tariq et al., 2021) that the survey of Litjens et al. (2017) credits for the entire deep-learning-in-radiology explosion, and unlike environmental science and climate, where community data portals underwrite an entire generation of foundation-model research (Yu et al., 2025), and unlike finance, where exchange and SEC filings are by design public, petroleum data is triply hard. It is *proprietary*: the highest-value training signals — seismic volumes, petrophysical interpretations, well architectures, production histories, drilling NPT root causes, frac-job post-mortems — belong to operators who treat them as competitive assets, and in several jurisdictions (Saudi Arabia, UAE, China, Mexico, Brazil, Russia) they are additionally subject to data-sovereignty regulation that bars cross-border training. The natural consequences are federated and on-premise training regimes (ADNOC, 2025; Aramco, 2024; Gharieb et al., 2024), vanishing open benchmarks (Bormann et al., 2020; Equinor, 2018; Ermilov, 2026; Vargas et al., 2019), and a reproducibility deficit that makes third-party replication nearly impossible for industry-scale claims.

Even within a single operator's vault, petroleum data is *heterogeneous*: daily drilling reports are free-text plus tables plus scanned OCR output (Antoniak et al., 2016; Arumugam et al., 2017; Asif et al., 2024; Hoffimann et al., 2017), well logs are LAS-formatted time-series with header metadata (Koeshidayatullah et al., 2024; Qi et al., 2025), seismic volumes are SEGY with trace-level auxiliary data (Dou et al., 2025; Sheng et al., 2025), production data arrives as SCADA streams (Baker Hughes, 2025; Tharayil et al., 2024), completion reports embed microseismic event catalogs (Chen et al., 2025a), and regulatory filings mix jurisdiction-specific forms. No common ingest pipeline exists; every new LLM project spends the first three months on data plumbing, and interoperability layers remain split across WITSML and related Energistics standards, OSDU-compatible data platforms (Abolhassani et al., 2023; Microsoft, 2024), ISO 15926 (Leal, 2005), and long-standing PPDM data models. Legacy

 doi:10.20944/preprints202604.1814.v1

data adds a third axis: *unstructured*. Ma et al. (2024) showed that extracting location and depth from 160 historical orphan-well documents required a Llama-3.1-405B pipeline and nonetheless achieved only ≈ 70 % accuracy on image-only scans of decades-old reports; Abdelgawad and El Ghattas (2025) report similar challenges for Egyptian legacy well archives.

The tripled data challenge conditions everything else in this survey. Retrieval-augmented systems (Aliyev et al., 2025; Braik et al., 2021; Dong et al., 2024; Mahjour and Mahjour, 2025; Matheus et al., 2025) emerge not because RAG is glamorous but because it is the only architecture that respects operator confidentiality while remaining competitive with fine-tuned baselines. Federated pretraining is the industry default rather than a research novelty. And the ≈ 25 % share that industrial grey-literature papers occupy in our corpus — unusual for a peer-reviewed survey — is a direct consequence: the people with the data do not publish at NeurIPS; they publish at SPE ATCE and ADIPEC.

6.2. Hallucination and Safety in Operational Contexts

LLMs confabulate. In general dialog this is an inconvenience; in petroleum operations it can be life-or-limb. A hallucinated BOP procedure, a wrongly paraphrased H_2S handling rule, or an invented casing-design shortcut exposes crews to incidents whose cost is measured in injuries, lost rigs, and environmental damage. The empirical baseline for this risk is Ogundare et al. (2023), whose Halliburton-affiliated authors showed that both GPT-3.5 and GPT-4 fail, sometimes confidently, on quantitative oil-and-gas PDE problems (Darcy flow, pressure transients, NMO corrections). The Weijermars et al. (2023) probe of ChatGPT on petroleum-engineering problems reported an analogous pattern of plausible but numerically wrong outputs.

Four mitigation patterns have emerged — patterns that the multi-agent survey of Guo et al. (2024) and the knowledge-management LLM work of Zhu et al. (2025) both identify as the main strategies for high-stakes industrial LLM deployment. First, *retrieval grounding* is the most widely adopted: Geo-RAG (Dong et al., 2024), ASK Thamama (Braik et al., 2021), the mud-invasion RAG of Aliyev et al. (2025), the drilling RAG of Matheus et al. (2025), the reservoir framework of Mahjour and Mahjour (2025), and the EAGE synthetic-data RAG of Chang et al. (2025) all constrain generation to retrievable passages. Second, *subject-matter-expert-in-the-loop evaluation* has become a publication requirement for industry-facing systems: EnergyLLM (Eckroth et al., 2025) reports SME blind-evaluation wins over GPT-4o; Mahjour and Mahjour report promising but preprint-only results on a closed 15-case evaluation (Mahjour and Mahjour, 2025); Ma et al. (Ma et al., 2024) report 100 % extraction accuracy on clean documents benchmarked against human annotators. Third, *explicit guardrail layers*: the Kanfar seismic agent (Kanfar et al., 2025) is marketed as the "industry's first guardrails-equipped AI agent" for seismic processing, and the multi-agent architectures of Sabbagh et al. (2025,2) use one agent specifically to critique another. Fourth, *formal verification*: the DOAPE physics-informed agent (Zejli et al., 2025) constrains its outputs by material-balance and flow-rate consistency checks, and the agentic well-modeling argument of Sharma (2026) proposes simulator-grounded sanity checks as a publication requirement.

None of these mitigations resolves the residual risk that no operator will green-light a fully autonomous agent to pull a trip, choke a well, or approve a stimulation plan without a human signing off. The asymptotic safety ceiling of the PetroLLM Maturity Model at its current maturity therefore sits near L4 Agent with supervised autonomy; an unsupervised L5 ecosystem remains categorically aspirational. The Jacobs (2025) review frames autonomous drilling as a ten-year program for precisely this reason.

6.3. The Evaluation Gap

The field currently lacks benchmarks. As of April 2026, Ermilov (2026)'s FormationEval is the only open, petroleum-specific LLM benchmark that has been published with a verifiable arXiv identifier. The GeoBench suite embedded in K2 (Deng et al., 2024) and the GeoFactory evaluation suite (Chen et al., 2025c) cover geoscience broadly but not petroleum-engineering-specifically. Everything else that claims "benchmark" status in the corpus is either (i) a general LLM benchmark repurposed for

 doi:10.20944/preprints202604.1814.v1

petroleum use (Chen et al., 2021; Cobbe et al., 2021; Hendrycks et al., 2021a; Liang et al., 2023; Rein et al., 2024; Srivastava et al., 2023; Zheng et al., 2023), (ii) a non-LLM deep-learning benchmark (Bormann et al., 2020; Deng et al., 2022; Equinor, 2018; Hall, 2016; Kainkaryam et al., 2019; Mousavi et al., 2019; Vargas et al., 2019; Woollam et al., 2022), or (iii) an internal SME evaluation set whose release is blocked by IP concerns (Eckroth et al., 2025; Mahjour and Mahjour, 2025).

The consequence is that claims in the petroleum-LLM literature are comparatively un-falsifiable. When a paper reports that its model "outperforms GPT-4o on petroleum-engineering questions," the reader has no standard test suite to which the improvement can be independently replicated. This is the same bottleneck that ImageNet, GLUE, and MMLU broke in their respective fields. The petroleum community needs — urgently — (i) an open, refereed drilling-LLM benchmark covering DDR summarization, NPT root-cause identification, and well-control procedure QA; (ii) an open reservoir-simulation-copilot benchmark covering ECLIPSE/CMG/tNavigator deck generation and history-matching QA; (iii) an open well-log-interpretation LLM benchmark extending Hall (Hall, 2016) and FORCE 2020 (Bormann et al., 2020) into natural-language tasks; (iv) a petroleum VQA benchmark integrating seismic sections, core photos, log panels, and text; and (v) multilingual petroleum benchmarks in Arabic, Russian, Portuguese, and Chinese — the world's operating languages beyond English. SME blind evaluation, as practiced by the EnergyLLM team and in some closed-set industry or preprint studies (Eckroth et al., 2025; Mahjour and Mahjour, 2025), is the current high-quality fallback but does not scale. Until open benchmarks exist, the PetroLLM Maturity Model cannot graduate from L3 to L4 with scientific rather than anecdotal confidence.

6.4. Physics versus Data: Domain Knowledge Integration

Petroleum engineering is not a pattern-recognition domain. Darcy's law governs flow in porous media; material balance governs reservoir depletion; the pressure-transient equations of Bourdet and Gringarten govern well testing; the elastic wave equation governs seismic imaging. Each of these encodes hard, non-negotiable physical constraints. Pure LLMs hallucinate physics (Ogundare et al., 2023); pure physics-informed neural networks (Kumar et al., 2023; Latrach et al., 2024; Li et al., 2025d) cannot reason about textual reports and unstructured knowledge. The field needs hybrid architectures in which LLMs orchestrate physics-based simulators — calling ECLIPSE, CMG, tNavigator, or Petrel as tools in a ReAct loop (Schick et al., 2023; Wu et al., 2024; Yao et al., 2023b) and reasoning about the returned numerical outputs.

Two early exemplars are ENVOY (Wiegand et al., 2024a,2), Stone Ridge's assistant that wraps the ECHELON reservoir simulator, and DOAPE (Zejli et al., 2025), an explicitly physics-informed agentic framework that validates LLM outputs against mass-balance constraints. The MYCRUNCHGPT system (Kumar et al., 2023) and the CodePDE framework (Li et al., 2025d) prototype LLM-assisted PDE solution generation. None of these yet constitutes a production-grade reservoir copilot; the correct framing is that the LLM should be the *conductor* of physics-based simulators, not their replacement (Samnioti and Gaganis, 2023a,2). An L5 ecosystem (Menon et al., 2026) would be one in which the LLM, the simulator, and the field telemetry close a differentiable feedback loop; this remains aspirational rather than operational.

6.5. Interpretability, Trust, and Engineering Sign-Off

A drilling engineer's signature on an operations plan carries legal weight. The liability regime in petroleum operations demands that recommendations be traceable to verifiable sources; regulatory bodies (BSEE in the US, NOPSEMA in Australia, the UK HSE, Norway's Ptil) expect documented chains of evidence. Black-box LLMs fail this bar. Three solutions are maturing: (i) RAG systems with mandatory citation of retrieved passages (Aliyev et al., 2025; Braik et al., 2021; Dong et al., 2024; Matheus et al., 2025); (ii) structured intermediate outputs or stepwise rationales exposed for human review (Kojima et al., 2022; Wang et al., 2022b; Wei et al., 2022); and (iii) structured outputs with provenance fields, as in the JSON-schema extraction pipeline of Abdelgawad and El Ghattas (2025)

 doi:10.20944/preprints202604.1814.v1

and the knowledge-graph grounding of Garcia et al. (2020); He et al. (2021); Liu et al. (2022); Santos et al. (2024); Tang et al. (2023); Wang et al. (2022a).

The regulatory environment is tightening independently of what petroleum operators do. The EU AI Act Annex III (point 2) lists "safety components in the management and operation of critical digital infrastructure and the supply of water, gas, heating, and electricity" as a high-risk category. Midstream and downstream oil-and-gas operations likely fall in scope under a future implementing act; upstream E&P is a more ambiguous case but seems a likely extension. HSE safety-case discipline — the rigorous, documented demonstration that risks have been identified and mitigated — sits in direct tension with the non-deterministic behavior of a decoder-only LLM. Resolving this tension will require either (i) hard constraints on LLM outputs via guardrails and structured decoding, (ii) formal verification of the narrow slice of tasks for which safety-critical automation is considered, or (iii) deliberate, explicit human-in-the-loop gating at sign-off boundaries. The current literature does not settle the question.

6.6. Reproducibility, Openness, and the Competitive Moat

Most petroleum-LLM papers do not release data, code, or weights. The exceptions are telling: K2 (Deng et al., 2024) released weights, data, and training recipes; GeoGalactica (Lin et al., 2024) released weights; JiuZhou (Chen et al., 2025b) released its corpus and framework; the Seismic Foundation Model (Sheng et al., 2025) released weights; WLFM (Qi et al., 2025) is available on arXiv with partial release; FormationEval (Ermilov, 2026) is open. All industrial platforms — METABRAIN, ENERGYai, SLB Tela and Lumi, Halliburton DS365.ai, Baker Hughes / C3.ai / Leucipa, Chevron ApEX and APOLO, TotalEnergies-Mistral — are closed, and their architectural details are communicated only through press releases, SPE panels, and JPT summaries (AIQ, 2025; Aramco, 2024,2; Baker Hughes, 2025; Baker Hughes and C3.ai, 2019; BP, 2025; Chevron, 2025a,2; ExxonMobil, 2025; Halliburton, 2021; Pallanich, 2025; Shell, 2024; SLB, 2024c,2; TotalEnergies, 2024,2).

This is not a pathology but a structural feature. Competitive moats drive the investment that makes the field commercially viable; open science accelerates the methodological progress that the field rests on. A plausible resolution is the *open-pretrain / closed-finetune* pattern, in which pretrained backbones and pretraining methodology are open while operator-specific fine-tunes and deployments are closed. K2 → EnergyLLM is the archetype: K2's open geoscience pretraining (Deng et al., 2024) provided a substrate that EnergyLLM (Eckroth et al., 2025) closed-finetunes on SPE content. The survey community should actively reward published artifacts in this pattern and treat closed-only claims as provisional.

6.7. Scalability and Deployment: Cloud, Edge, and Rig

Trillion-parameter class models run on cloud hyperscaler GPUs; they do not run on offshore rigs. METABRAIN's 250 B-parameter class roadmap (Aramco, 2024) and similar operator-scale proprietary programs are cloud-side. Offshore installations, MPD geothermal sites, and remote land operations require on-premise or edge deployment: 7 B–13 B models with QLoRA quantization (Dettmers et al., 2023) and heavy parameter-efficient adaptation (Han et al., 2024; Hu et al., 2022) dominate this tier. The control-room language-model screen of Ferrigno et al. (2024) and Tharayil et al. (2024), and the operator-personalized pipeline of Gharieb et al. (2024), are practical demonstrations of the edge-side pattern.

Latency requirements also vary by task type. A seismic agent analyzing ten-gigabyte pre-stack gathers (Kanfar et al., 2025) has minute-scale SLAs; a conversational well-planning assistant has sub-second SLAs; a real-time drilling copilot (Yi et al., 2024) has 100-millisecond SLAs. Economics cuts the same way: deploying fleet-scale predictive maintenance (as in Shell's decade-old collaboration with C3.ai across ≈10,000 equipment instances (Baker Hughes and C3.ai, 2019; Shell, 2024)) has demonstrated ROI, while deploying a bespoke LLM for every asset has not. The resulting architectural pattern across the PetroLLM Maturity Model is a tiered one: an L5-aspirational platform in the cloud, L3 domain LLMs at regional hubs, and small L2/L4 agents at the edge — much like the medical imaging tier of radiology PACS + regional reading centers + bedside devices.

 doi:10.20944/preprints202604.1814.v1

6.8. The Talent and Organizational Challenge

The field demands hybrid talent: petroleum engineering plus machine learning, in one person or one tightly-coupled team. That talent is rare. The Society of Petroleum Engineers now runs an AI Symposium (Society of Petroleum Engineers, 2025), the SEG hosts ML workshops (Society of Exploration Geophysicists, 2026), petroleum MS programs have begun adding ML coursework, and trade-press coverage now treats LLM deployment as a mainstream operational topic (Pallanich, 2025). These are encouraging signals but they do not yet close the gap: there are not enough petroleum engineers who can read a NeurIPS paper, and there are not enough ML researchers who can read a well-control manual.

Organizational structure compounds the talent challenge. AI platforms often sit in a digital or IT organization, separate from the discipline groups (drilling, reservoir, geophysics) whose workflows they are meant to support. The handoff friction — between ML teams who own the models and petroleum teams who own the problems — is the largest operational barrier to the pilot-to-production transitions we describe in Section 7.6. The emerging cohort of petroleum-engineering-trained ML practitioners — the so-called bridge practitioners, each trained in one discipline and deployed in the other — is a precondition for, not an artifact of, field progress.

7. Future Directions

We now extrapolate the current state, the open challenges, and the PetroLLM Maturity Model forward in time. Our predictions are grouped into eight directions. Sections 7.1–7.5 are research-oriented and largely drive L3–L5 of the PetroLLM Maturity Model upward. Sections 7.6–7.8 are deployment-oriented and address how L4 deployments transition into L5 ecosystems. We also interleave a dedicated discussion of the **PetroLLM Maturity Model** itself (Section 7.9) and a closing section on the question the community is most often asked: whether LLMs will replace petroleum engineers (Section 7.10).

7.1. A Petroleum Foundation Model for the Subsurface

One central research objective for 2026–2030 is a petroleum foundation model for the subsurface: a single multimodal backbone pretrained on OnePetro, SPE Books and PetroWiki, the open well-log corpora, the open seismic volumes, and the public core-photo archives; fine-tuned on an operator's proprietary data; and capable of joint reasoning over seismic volumes, well logs, production time-series, and narrative reports. The technical template exists. K2 (Deng et al., 2024) and GeoGalactica (Lin et al., 2024) provide the geoscience-language backbone; SFM (Sheng et al., 2025), SeisCLIP (Si et al., 2024), GEM 3D (Dou et al., 2025), SeisBERT (Pham et al., 2025), and StorSeismic (Harsuko and Alkhalifah, 2022) provide the seismic modality; WLFM (Qi et al., 2025) and the TimeGPT-for-logs work of Koeshidayatullah et al. (2024) provide the well-log modality; the time-series foundation models Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024), Lag-Llama (Rasul et al., 2024), Moirai (Woo et al., 2024), and MOMENT (Goswami et al., 2024) provide the production time-series modality; and PaliGemma 2 (Steiner et al., 2024) and Qwen2-VL (Wang et al., 2024b) show how a modern VLM fuses text and vision.

What is missing is the petroleum-specific instruction-tuning corpus and the fused multimodal backbone. The 2027–2028 paper we predict will be titled something like "PetroLlava" or "SubsurfaceGPT": a 30–70 B-parameter instruction-tuned multimodal model pretrained on $\approx$200 B petroleum tokens plus $\approx$1 PB of seismic and log data, released with weights and benchmark suite. It will likely emerge from a bridge lab rather than an operator: KAUST and SJTU are plausible candidates, with other bridge labs also in contention.

7.2. Autonomous Drilling Copilots

The drilling literature maps to the PetroLLM Maturity Model with unusual cleanliness. L1 is the ChatGPT-on-drilling probes of Ogundare et al. (2023); Weijermars et al. (2023). L2 is the post-

 doi:10.20944/preprints202604.1814.v1

well analysis lineage — Hoffimann (Hoffimann et al., 2017), Arumugam (Arumugam et al., 2017), Antoniak (Antoniak et al., 2016), Kowalchuk (Kowalchuk, 2019), Sidahmed (Sidahmed et al., 2015), the DDR-LLM systems of Asif et al. (2024); Kumar and Kathuria (2023), Ma et al. on orphan wells (Ma et al., 2024), and Wang on mud reports (Wang et al., 2025). L3 is the rock-mechanics LLM of Lin et al. (2025) and the personalized Gharieb approach (Gharieb et al., 2024). L4 is the current frontier: real-time well-construction advisors (Alfarisi et al., 2024; Pacis et al., 2024; Reddicharla et al., 2025; Yi et al., 2024), the Sabbagh multi-agent architecture (Sabbagh et al., 2025,2), the Matheus drilling-RAG pipeline (Matheus et al., 2025), and the agentic drilling frameworks of Cayeux et al. (2025); Jacobs (2025); Osman et al. (2025).

Over the 2026–2028 window we expect routine deployment of L4 post-well analysis copilots at major operators, followed by trial deployment of L4 real-time advisory systems for hole cleaning, mud-weight management, and trip planning. Over the 2028–2030 window we expect safety-envelope-bounded L4 autonomous decision support — explicitly not "autonomous drilling" in the unsupervised sense, but rather "autonomy within a pre-cleared operational corridor." The eDrilling thread (Cayeux et al., 2021,2; Jacobs, 2025) and the digital-rig efforts of the Aker BP / Cognite / NVIDIA stack are the most credible deployment vehicles. True autonomous drilling remains a five-to-ten-year horizon, gated by the safety-case and HSE tensions of Section 6.5.

7.3. Multimodal Petroleum AI

Today, petroleum LLMs (Chen et al., 2025b; Deng et al., 2024; Eckroth et al., 2025; Lin et al., 2024; Ma et al., 2026), subsurface FMs (Dou et al., 2025; Qi et al., 2025; Sheng et al., 2025), VLMs (Liu et al., 2024a; Si et al., 2024), and time-series FMs (Das et al., 2024; Koeshidayatullah et al., 2024) are all separately instantiated. Tomorrow's petroleum AI will fuse them. The analog in general AI is the CLIP $\rightarrow$ LLaVA $\rightarrow$ Qwen-VL $\rightarrow$ PaliGemma 2 trajectory over 2021–2025 (Bai et al., 2023; Beyer et al., 2024; Liu et al., 2023; Radford et al., 2021; Steiner et al., 2024; Wang et al., 2024b): a decade-scale unification of modalities in general-purpose AI is compressing into roughly three years. The missing ingredient in the petroleum case is not methodology but data — a petroleum-specific multimodal instruction corpus, analogous to LLaVA's instruction tuning data but grounded in seismic sections, core photographs, log panels, drilling schematics, and their narrative captions.

Assembling this corpus is, we predict, the largest coordination task now blocking progress in multimodal petroleum AI. It requires roughly one full operator-year of SME time to curate $\approx$10,000 multi-modal examples at quality sufficient for instruction tuning. We expect the SPE AI Symposium (Society of Petroleum Engineers, 2025) and the SEG ML Workshop (Society of Exploration Geophysicists, 2026) to catalyze collaborative datasets; a credible vehicle is the EnergyLLM model of a society-endorsed consortium, expanded to include multimodal capture.

7.4. Digital Twin and LLM Integration

Digital twins in upstream petroleum are today physics-based simulators (ECLIPSE, CMG, tNavigator, the Stone Ridge ECHELON engine wrapped by ENVOY (Wiegand et al., 2024a,2), the NVIDIA Earth-DT stack (NVIDIA, 2024), SLB's DELFI and its subsurface modules (Schlumberger, 2017)). Tomorrow's digital twins will pair these simulators with a natural-language interface: an LLM agent that parses the engineer's question ("What happens to WHP if we increase choke size by 20 % on Well 17-B?"), translates to simulator input deck, queries the twin, and returns a reasoned answer annotated with uncertainty. The Mahjour framework (Mahjour and Mahjour, 2025) prototypes this pattern for reservoir decision-support; ENVOY does it for reservoir simulation; the Sharma (2026) JPT note does it for well modeling; DOAPE (Zejli et al., 2025) adds physics-informed validation; and the fuel-for-work integration of Rodriguez Torrado et al. (2025) hints at field-development planning as the next target workflow. The research agenda is how to handle simulator uncertainty, surrogate coupling, and multi-fidelity ensembles inside an LLM agent's reasoning loop.

 doi:10.20944/preprints202604.1814.v1

7.5. Federated Learning for Petroleum

Operators compete; they cannot pool their proprietary datasets; this makes federated and other privacy-preserving training schemes attractive. The industry pressure for such schemes is already visible in the Microsoft–Aramco MoU (Aramco, 2026), the SLB–ADNOC–AIQ–G42 consortium (ADNOC, 2025; AIQ, 2025), the Baker Hughes–Repsol Leucipa partnership (Baker Hughes, 2025), and the Chevron–Cognite alliances surveyed in Section 4.7.5. These are not demonstrations of federated learning, but they do show the business pressure that motivates it. The academic parametric and privacy-preserving analog lags behind. The open research question — not yet answered in the corpus — is whether federated pretraining and federated continued-pretraining can match centralized training quality at the 7 B and 70 B scales relevant to L3 domain LLMs. Methodologically this is a direct extension of the don't-stop-pretraining agenda (Gururangan et al., 2020) into a distributed-compute setting; the work is yet to be done.

7.6. From Pilot to Production

Operators today count their generative-AI pilots in the dozens. TotalEnergies' Microsoft 365 and Mistral integrations (TotalEnergies, 2024,2), BP's enterprise rollouts (BP, 2025), Shell's decade of predictive-maintenance work (Baker Hughes and C3.ai, 2019; Shell, 2024), Chevron ApEX and APOLO (Chevron, 2025a,2), ExxonMobil Vantage (ExxonMobil, 2025), Devon's ChatDVN (reported in JPT (Pallanich, 2025)), Repsol's Leucipa (Baker Hughes, 2025), and bp's Xpert each represent a pilot that has yet to scale fleet-wide. The digital-transformation wave of 2015–2020 that spawned DELFI (Schlumberger, 2017) and DS365.ai (Halliburton, 2021) produced hundreds of pilots of which fewer than twenty graduated to production; the AI wave is likely to follow a similar funnel shape.

Barriers to pilot-to-production are mostly not technical: change management, data-quality investment, regulatory approval, organizational re-architecture, and cost-model ambiguity dominate. The one technical barrier that matters is the evaluation-gap argument of Section 6.3: without open benchmarks, the internal CFO cannot distinguish a 95 %-accurate pilot from a 70 %-accurate pilot, and the default corporate behavior under uncertainty is "wait another quarter." Closing the evaluation gap is therefore the most actionable research activity in the field today.

7.7. Industry Consolidation Around Platforms

The industry is fragmented today: dozens of bespoke LLM deployments per major operator, each with its own vendor stack, its own data pipeline, its own fine-tune. Within three to five years we expect consolidation around a small number of platform families. The candidate consolidators are (i) AWS Bedrock (Amazon Web Services, 2024), with its petroleum-specific fine-tune guide targeting operator-hosted deployment; (ii) the Microsoft–ADNOC–AIQ–G42–SLB stack (ADNOC, 2025; AIQ, 2025; Aramco, 2026); (iii) a Google-centered pathway leveraging Gemini (Gemini Team, Google, 2023,2) on top of operator-owned OSDU-compatible data platforms (Microsoft, 2024); (iv) the NVIDIA AI Factory with SLB (SLB, 2024b,2); and (v) state-backed Chinese operator-AI platform ecosystems. Each operator will likely pick one primary platform and complement with vendor-supplied modality-specific models — Viridien for seismic workflows (Dong et al., 2024; Kanfar et al., 2025; Sansal et al., 2025), Stone Ridge for reservoir (Wiegand et al., 2024a,2), and specialized vendors for petrophysics and drilling.

7.8. Regulation, Standards, and the Petroleum AI Safety Case

Aviation has DO-178C for safety-critical software; medicine has the FDA's 510(k) pathway for clearance; petroleum-engineering AI has neither. The SPE API standards apparatus and ISO 55000 asset-management framework are the most obvious starting points for a petroleum AI safety case, but no explicit standard exists as of this writing. We predict that within three years the SPE will form an AI Standards Technical Section, and within five years a first-edition petroleum-LLM safety-case document will be circulated for member comment. The analogous precedent is the emergence of the SPE Digital Technical Section and the SEG Digital Transformation Committee in the 2010s.

 doi:10.20944/preprints202604.1814.v1

7.9. The PetroLLM Maturity Model

A recurring navigational device of this survey is the PetroLLM Maturity Model — a five-level scaffold for situating any petroleum LLM or FM deployment on the autonomy spectrum. The model, detailed in Figure 3 and Table 14, is inspired by the Capability Maturity Model Integration (CMMI) for software engineering and by vehicle-automation SAE-J3016's six levels of driving autonomy. It is grounded in petroleum realities.

Table 14. The PetroLLM Maturity Model: five levels of increasing autonomy from conversational question answering to self-improving foundation-model ecosystems, with canonical exemplar systems. Cross-cutting discriminators are shown along the bottom four rows; *Autonomy* here counts tool-using steps per query, not internal reasoning steps.

Level	Name	Capability	Example systems	Representative references
L1	Conversational Q&A	Off-the-shelf general-purpose LLM answering petroleum questions; no retrieval, no tools, no domain fine-tune.	PetroQA; Ogundare ChatGPT baseline; Weijermars probe; early chatbot deployments.	Eckroth et al. (2023); Ogundare et al. (2023); Weijermars et al. (2023); Singh et al. (2023); Pacis et al. (2024).
L2	Document Intelligence & Retrieval	RAG over proprietary corpora; multimodal document ingest; domain embeddings; retrieval grounding without an agent loop.	Geo-RAG, ASK Thamama, Mud-Invasion RAG, Drilling-RAG, Mahjour reservoir DSS, Ma legacy-well pipeline, ColPali-style visual retrieval, RAG-vs-FT study.	Dong et al. (2024); Braik et al. (2021); Aliyev et al. (2025); Matheus et al. (2025); Mahjour and Mahjour (2025); Ma et al. (2024); Faysse et al. (2025); Elyas et al. (2025).
L3	Domain-Specialized LLMs	Continued pretraining / SFT / PEFT on curated petroleum or geoscience corpora; may include domain-adapted embeddings, tokenization, and multilingual variants.	K2, GeoGalactica, EnergyLLM, EnergyGPT, PetroBERT, JiuZhou, QH-GeoGPT, LithoGPT, BB-GeoGPT, GeoFactory, METABRAIN, and related closed industrial domain-LLM programs.	Deng et al. (2024); Lin et al. (2024); Eckroth et al. (2025); Chebbi and Kolade (2025); Rodrigues et al. (2022); Chen et al. (2025b); Ma et al. (2026); Chen et al. (2025c); Li et al. (2025c); Zhang et al. (2024c); Aramco (2024).
L4	Autonomous Agents & Copilots	Tool-using multi-step copilots and agents that call simulators, operator data stores, and telemetry systems; some public examples expose planning or reflection explicitly, others expose tool use only.	Kanfar seismic agent, ENVOY + ECHELON, DOAPE, Sabbagh multi-agent, JPT agentic well modeling, Osman directional-drilling case, SLB Tela, Leucipa.	Kanfar et al. (2025); Wiegand et al. (2024a); Wiegand et al. (2024b); Zejli et al. (2025); Sabbagh et al. (2024); Sabbagh et al. (2025); Sharma (2026); Osman et al. (2025); Jacobs (2025); Cayeux et al. (2025); SLB (2025); Baker Hughes (2025).
L5	Self-Improving FM Ecosystems	Multimodal subsurface FMs continuously ingesting new field data, self-evaluating, and self-updating with human oversight; tightly coupled to physics simulators and operator-scale deployment fabric.	Aspirational operator-scale ecosystems signaled by public roadmaps and partnerships such as Aramco × Microsoft and the SLB × NVIDIA AI Factory; METABRAIN-class industrial programs are relevant but not yet public L5 systems.	Aramco (2024); Aramco (2026); SLB (2026); SLB (2024b); Menon et al. (2026); Pallanich (2025).

Cross-cutting discriminators (L1 → L5):
Grounding: none → retrieval → pretraining → tools + retrieval → tools + retrieval + simulator.
Autonomy: 0 → 0 → 0 → 1 step → multi-step.
Multimodality: text → text + doc → text (mostly) → text + code → seismic + log + text + PVT + SCADA.
Human role: reader → reader → evaluator → supervisor → auditor.

Level 1 — Conversational Q&A. Off-the-shelf general-purpose LLMs (GPT-4, Claude, Gemini) answering petroleum questions with no domain grounding. Representatives: PetroQA (Eckroth et al., 2023), the first systematic ChatGPT-on-O&G probe of Ogundare et al. (2023), Weijermars (Weijermars et al., 2023), the early chatbot deployments of Singh et al. (2023), the zero-shot petroleum classification of Pacis et al. (2024). The definitional discriminator is the absence of retrieval, tools, and domain fine-tuning. The characteristic failure mode is hallucination in quantitative engineering.

Level 2 — Document Intelligence & Retrieval. RAG systems grounded in proprietary corpora: Geo-RAG (Dong et al., 2024), the SLB synthetic-data RAG pipeline (Chang et al., 2025), ASK Thamama (Braik et al., 2021), the mud-invasion RAG of Aliyev et al. (2025), the drilling RAG of Matheus et al. (2025), the reservoir decision framework of Mahjour and Mahjour (2025), the well-data QA pipeline of Reddicharla et al. (2025), the orphan-well extraction of Abdelgawad and El Ghattas (2025); Ma et al. (2024), the ColPali-style visual-document retrieval (Cho et al., 2024; Faysse et al., 2025), the EAGE GraphRAG (Edge et al., 2024; Jiang et al., 2025), plus enterprise retrieval tiers of operator platforms (including ENVOY in its retrieval mode (Wiegand et al., 2024a)). The characteristic discriminator

 doi:10.20944/preprints202604.1814.v1

is retrieval grounding without a tool-use loop; the characteristic failure mode is brittle multimodal retrieval.

Level 3 — Domain-Specialized LLMs. Continued pretraining, SFT, and PEFT on petroleum corpora to produce domain-specific weights. K2 (Deng et al., 2024), GeoGalactica (Lin et al., 2024), EnergyLLM (Eckroth et al., 2025), EnergyGPT (Chebbi and Kolade, 2025), JiuZhou (Chen et al., 2025b), PetroBERT (Rodrigues et al., 2022), QHGeoGPT (Ma et al., 2026), LithoGPT (Li et al., 2025c), BB-GeoGPT (Zhang et al., 2024c), GeoCode-GPT (Hou et al., 2025b), the Zhao knowledge-management work (Zhao et al., 2025; Zhu et al., 2025), the rock-mechanics LLM (Lin et al., 2025), the geotech LLM (Fan et al., 2026), and industrial platforms METABRAIN (Aramco, 2024), Aramco Dhahran (Aramco Europe, 2025), and GeoFactory (Chen et al., 2025c). The characteristic discriminator is modified weights; the characteristic failure mode is closed-weight competitor saturation and benchmark deficit.

Level 4 — Autonomous Agents & Copilots. LLMs wrapped with tools, memory, and plan-act-reflect loops, calling simulators and data stores. The Kanfar seismic agent (Kanfar et al., 2025), ENVOY and ENVOY-SEG (Wiegand et al., 2024a,2), DOAPE (Zejli et al., 2025), Sabbagh's multi-agent well-construction and D&C knowledge-management stacks (Sabbagh et al., 2025,2), Mosser's exploration robot chat (Mosser et al., 2024), the 2026 JPT agentic well-modeling commentary (Sharma, 2026), the autonomous directional-drilling case study of Osman et al. (2025), the Cayeux-line digital-drilling work (Cayeux et al., 2021,2; Jacobs, 2025), the control-room language-model screen of Ferrigno et al. (2024); Tharayil et al. (2024), SLB Tela (SLB, 2025) and the SLB geosteering copilot (SLB, 2024a), and the Repsol Leucipa autonomous-production platform (Baker Hughes, 2025). The characteristic discriminator is the tool-use loop; the characteristic failure mode is the absence of end-to-end closed-loop peer-reviewed validation.

Level 5 — Self-Improving Foundation-Model Ecosystems. Multimodal subsurface FMs that continuously ingest field data, self-evaluate, and self-update with human oversight; tightly coupled to physics simulators as differentiable tools. *There are no mature examples.* The aspirational anchors are the METABRAIN trillion-parameter roadmap (Aramco, 2024), the Microsoft–Aramco MoU (Aramco, 2026), the SLB–NVIDIA AI Factory (SLB, 2024b,2), public roadmap signals from Viridien (Viridien, 2026), and the scientific-foundation-model theoretical framing of Menon et al. (2026). The JPT 2025 declaration of operational readiness (Pallanich, 2025) should still be read as an aspiration rather than as evidence of a mature L5 deployment. Regulatory, safety, carbon, and sovereignty issues all remain open.

7.10. Will LLMs Replace Petroleum Engineers?

LLMs are more likely to augment petroleum engineers than to replace them — not because LLMs are weak, but because petroleum engineering contains an irreducible judgment component that resists full automation. Well control, safety, geologic surprise, and real-time rig-site adaptation require experiential pattern recognition and legal accountability that no current or foreseeable LLM can bear. The oil-and-gas industry's safety-case regime demands a named, qualified, credentialed engineer to sign. A GPT-x cannot sign, cannot be sued, cannot be revoked. The sign-off bottleneck in Section 6.5 is not a temporary technological gap; it is an industrial legal and regulatory invariant.

What *will* compress dramatically is the productivity bar for early- and mid-career engineers. Devon's ChatDVN, reported in JPT, reports a $\approx 7\,\%$ drilling-speed gain and a $\approx 25\,\%$ productivity uplift on early-career engineers' throughput (Pallanich, 2025); the Mahjour preprint reports faster reservoir-characterization workflows across 15 closed field tests (Mahjour and Mahjour, 2025); EnergyLLM (Eckroth et al., 2025) allows a newly onboarded engineer to search decades of SPE literature at interactive latency. New-hire enablement, report generation, report comprehension, regulatory-filing drafting, incident-log summarization, and routine simulator-deck preparation will all be heavily LLM-assisted within three years. Senior engineering judgment — well architecture, field development, reservoir management, operational decision-making under uncertainty — will remain fundamentally human.

 doi:10.20944/preprints202604.1814.v1

The right framing, therefore, is not whether LLMs replace petroleum engineers, but which engineering tasks LLMs allow a given engineer to execute. Young engineers will be more productive; senior engineers will accumulate more experience faster by offloading rote work; the middle cohort — the five- to fifteen-year experience band — will face the sharpest adjustment as their comparative advantage in "rote synthesis of the literature" erodes. The industry's obligation, one that SPE, SEG, and operator HR organizations are only beginning to discharge, is to accelerate the retraining and upskilling of exactly this cohort.

8. Conclusion

This survey has traced the arrival of large language models and foundation models in petroleum engineering from the word-embedding experiments of the late 2010s to the industry-scale platforms and multimodal subsurface foundation models of 2026, analysing 296 verified references across fourteen thematic areas, six petroleum sub-disciplines plus one cross-cutting category, and more than a decade of publication activity. Our unifying contribution is a consistent navigational scaffold — the *PetroLLM Maturity Model* — that spans five capability tiers from conversational question answering through retrieval-augmented document intelligence, domain-specialized language models, autonomous agents and copilots, and the aspirational self-improving foundation-model ecosystems that the field's leading operators are now roadmapping.

Five takeaways emerge from the analysis. First, the petroleum-LLM field has passed the early-adopter inflection: annual publication counts grew more than five-fold between 2020 and 2024, and the lag between a new general-purpose model (LLaMA 3, GPT-4, DeepSeek-R1) and its petroleum descendant (EnergyLLM, QHGeoGPT, the rock-mechanics LLM) has compressed from roughly five years in the Word2Vec era to roughly one year in the fastest recent cases. Second, the methodological distribution is sharply asymmetric: geophysics and cross-cutting document-intelligence dominate the corpus, whereas reservoir simulation, completions, production, and petroleum VQA remain underserved, creating well-defined research openings. Third, petroleum's defining constraints — proprietary data sovereignty, heterogeneous ingest, safety-critical operations, regulatory sign-off — condition every architectural choice we surveyed: retrieval-augmented deployments dominate L2 because confidentiality demands them, closed-weight industrial platforms dominate L3 and L4 because competitive moats fund them, and the scarcity of L5 ecosystems reflects not technical limits but the slow accrual of reliability evidence that safety-case disciplines require. Fourth, the evaluation gap is the most actionable barrier to progress: FormationEval is today the field's only open petroleum-specific LLM benchmark, and closing this gap with open drilling, reservoir, petrophysics, and multimodal VQA benchmarks would unlock the entire pilot-to-production funnel. Fifth, the industry is simultaneously consolidating around a small number of platform-class consortia (Microsoft–ADNOC–AIQ–G42–SLB, Aramco METABRAIN with its Microsoft MoU, SLB-NVIDIA AI Factory, Chinese state-scale operator programs, the AWS Bedrock ecosystem) and fragmenting into dozens of bespoke operator-internal pilots — the tension between these two trajectories will define the deployment landscape of the late 2020s.

We offer the PetroLLM Maturity Model as the survey's central organizing contribution. Where individual datasets, models, and platforms will evolve, the five-level scaffold — Conversational Q&A, Document Intelligence and Retrieval, Domain-Specialized LLMs, Autonomous Agents and Copilots, Self-Improving Foundation-Model Ecosystems — should persist as a common procurement, benchmarking, and research-planning vocabulary. Future operators can locate any vendor proposal on the ladder; future researchers can situate any new model relative to the level-specific gaps we enumerate; future PhD students can select open problems from the level at which they wish to contribute. We have populated each level with canonical exemplars so that the framework is operational rather than merely descriptive.

The paper closes with four practical priorities. The field needs more *open benchmarks* at every sub-discipline (drilling, reservoir, petrophysics, completions, production, and multimodal VQA); more *open*

 doi:10.20944/preprints202604.1814.v1

weights released under the open-pretrain / closed-finetune pattern exemplified by K2, GeoGalactica, JiuZhou, SFM, and WLFM; more *bridge talent* capable of reading both a NeurIPS paper and a well-control manual, supported by SPE AI programs, SEG ML workshops, and the next generation of hybrid petroleum-ML curricula; and more explicit *safety-case standards* for engineers, regulators, and courts. These four agenda items — benchmarks, openness, talent, and standards — are achievable within the next three years if the SPE, SEG, EAGE, OTC, and ADIPEC communities coordinate.

The next decade will not be written by artificial intelligence alone, nor by petroleum engineers alone, but by the growing cohort of hybrid practitioners who speak both languages fluently — who can read an attention-head ablation and a mud-weight log window in the same afternoon, who can critique a physics-informed neural network and a casing-design record on the same whiteboard, and who understand that the subsurface is simultaneously a data problem and an engineering discipline. This survey is offered to that cohort: to the bridge authors already at work, to the graduate students about to begin, and to the senior engineers willing to learn. The field is theirs to build.

Nomenclature

AI	Artificial Intelligence.
AGI	Artificial General Intelligence.
Agentic AI	LLM wrapped with tools, memory, and plan–act–reflect loops; see §2.8.
BERT	Bidirectional Encoder Representations from Transformers.
DPR	Dense Passage Retrieval (bi-encoder architecture for RAG).
CLIP	Contrastive Language–Image Pretraining.
CoT	Chain-of-Thought prompting.
DL	Deep Learning.
DPO	Direct Preference Optimization.
FM	Foundation Model (pretrained, broadly adaptable).
GPT	Generative Pretrained Transformer.
KG	Knowledge Graph.
LLM	Large Language Model.
LoRA / QLoRA	Low-Rank Adaptation; quantised LoRA.
MAE	Masked Autoencoder.
MCP	Model Context Protocol (tool/data-server interface for agents).
MLLM / VLM	Multimodal / Vision–Language LLM.
MLM	Masked Language Modelling pretext task.
OCR	Optical Character Recognition.
PEFT	Parameter-Efficient Fine-Tuning.
RAG	Retrieval-Augmented Generation.
ReAct	Reasoning + Acting prompting pattern.
RLHF	Reinforcement Learning from Human Feedback.
SFT	Supervised Fine-Tuning.
SFM	Subsurface / Seismic Foundation Model.
SSL	Self-Supervised Learning.
TSFM	Time-Series Foundation Model.
ViT	Vision Transformer.
AFE	Authorization For Expenditure.
API	American Petroleum Institute (standards).
BHA	Bottom-Hole Assembly.
BOP	Blowout Preventer.
DAS	Distributed Acoustic Sensing.
DCA	Decline Curve Analysis.
DDR	Daily Drilling Report.
DELFI	SLB cognitive E&P environment.
EOR / IOR	Enhanced / Improved Oil Recovery.

 doi:10.20944/preprints202604.1814.v1

ESP	Electric Submersible Pump (artificial lift).
FDP	Field Development Plan.
FWI	Full-Waveform Inversion.
HAZOP	Hazard and Operability study.
HSE	Health, Safety, and Environment.
LAS	Log ASCII Standard for well logs.
MWD / LWD	Measurement / Logging While Drilling.
NMO	Normal Moveout correction.
NPT	Non-Productive Time (drilling).
IOGP	International Association of Oil & Gas Producers (formerly OGP).
OSDU	Open Subsurface Data Universe data-platform initiative under The Open Group.
P&ID	Piping and Instrumentation Diagram.
PDG	Permanent Downhole Gauge.
PPDM	Professional Petroleum Data Management industry data model.
PVT	Pressure–Volume–Temperature fluid properties.
ROP	Rate of Penetration.
SCADA	Supervisory Control and Data Acquisition.
SEGY	Seismic data exchange format.
SME	Subject-Matter Expert.
SPE / SEG / EAGE	Society of Petroleum Engineers / Exploration Geophysicists / Assoc. of Geoscientists & Engineers.
SRP	Sucker Rod Pump / beam pump (artificial lift).
TLE	*The Leading Edge* (SEG magazine).
WAG	Water-Alternating-Gas injection scheme.
WITSML	Wellsite Information Transfer Standard Markup Language.

References

Abdelgawad, A.E., El Ghattas, A., 2025. Fine-tuned generative AI for automated structured data extraction and insight generation from legacy petroleum well reports: An Egyptian oilfields case study, in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/229443-MS.

Abolhassani, N., Tudor, A., Paul, S., 2023. A data mesh adaptable oil and gas ontology based on open subsurface data universe (OSDU), in: Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KMIS). https://doi.org/10.5220/0012160000003598.

ADNOC, 2025. ADNOC and SLB launch transformative AI-powered solution to boost upstream productivity. Press release, https://www.adnoc.ae/en/news-and-media/press-releases/2025/adnoc-and-slb-launch-transformative-ai-powered-solution-to-boost-upstream-productivity/.

AIQ, 2025. AIQ announces $340 million contract for large-scale deployment of agentic AI across ADNOC operations. Press release, https://aiqintelligence.ae/newsroom/news-and-press-releases/AIQ-announces-340-million-contract-for-large-scale-deployment-of-agentic-AI-across-ADNOC-operations.

Alfarisi, O., Singh, R., Singhal, R., Alzarooni, R.M., Fernandes, S., Ayvaz, Y., Vijayan, M., Mohamed, J., 2024. The first drilling dedicated artificial intelligence ChatGPT pilot, in: GOTECH Conference. https://doi.org/10.2118/219337-MS.

Aliyev, S., Suleymanov, J., Hassan, A., Al Shafloot, T., ElHusseiny, A., 2025. Mud invasion zone detection using retrieval-augmented generation: A generative AI approach, in: SPE Middle East Oil, Gas and Geosciences Show (MEOS GEO). https://doi.org/10.2118/227590-MS.

Alzubaidi, F., Mostaghimi, P., Swietojanski, P., Clark, S.R., Armstrong, R.T., 2021. Automated lithology classification from drill core images using convolutional neural networks. Journal of Petroleum Science and Engineering 197, 107933. https://doi.org/10.1016/j.petrol.2020.107933.

Amazon Web Services, 2024. Customize large language models with oil and gas terminology using Amazon Bedrock. AWS for Industries Blog, https://aws.amazon.com/blogs/industries/customize-large-language-models-with-oil-and-gas-terminology-using-amazon-bedrock/.

 doi:10.20944/preprints202604.1814.v1

Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Pineda Arango, S., Kapoor, S., et al., 2024. Chronos: Learning the language of time series. https://arxiv.org/abs/2403.07815. arXiv:2403.07815.

Antoniak, M., Dalgliesh, J., Verkruyse, M., Lo, J., 2016. Natural language processing techniques on oil and gas drilling data, in: SPE Intelligent Energy International Conference and Exhibition. https://doi.org/10.2118/181015-MS.

Aramco, 2024. Aramco unveils new initiatives to drive digital development (METABRAIN LLM). News item, https://www.aramco.com/en/news-media/news/2024/aramco-unveils-new-initiatives-to-drive-digital-development.

Aramco, 2026. Aramco signs MoU with Microsoft to help advance industrial AI and digital talent transformation. News item, https://www.aramco.com/en/news-media/news/2026/aramco-signs-mou-with-microsoft-to-help-advance-industrial-ai-and-digital-talent-transformation.

Aramco Europe, 2025. Aramco deploys AI and supercomputing to drive digital transformation. News item, https://europe.aramco.com/en/news-media/news/2025/aramco-deploys-ai-and-supercomputing-to-drive-digital-transformation.

Arumugam, S., Rajan, S., Gupta, S., 2017. Augmented text mining for daily drilling reports using topic modeling and ontology, in: SPE Western Regional Meeting. https://doi.org/10.2118/185711-MS.

Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H., 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection, in: International Conference on Learning Representations (ICLR). URL: https://arxiv.org/abs/2310.11511, arXiv:2310.11511.

Asif, W., Al Salt, A.B., Al Sulaimani, T., Al Noufli, N., 2024. Multi-label classification of daily drill reports (DDR) utilizing large language models (LLMs), in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/221870-MS.

Bahaloo, S., Mehrizadeh, M., Najafi-Marghmaleki, A., 2023. Review of application of artificial intelligence techniques in petroleum operations. Petroleum Research https://doi.org/10.1016/j.ptlrs.2022.07.002.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J., 2023. Qwen-VL: A versatile vision-language model. https://arxiv.org/abs/2308.12966. arXiv:2308.12966.

Baker Hughes, 2025. Baker Hughes, Repsol launching new AI-powered functionality in Leucipa. News release, https://www.bakerhughes.com/company/news/baker-hughes-repsol-launching-new-aipowered-functionality-leucipatm.

Baker Hughes, C3.ai, 2019. Baker Hughes and C3.ai announce joint venture to deliver AI solutions across the oil and gas industry. Press release, https://investors.bakerhughes.com/news/press-releases/news-details/2019/Baker-Hughes-a-GE-company-and-C3-ai-Announce-Joint-Venture-to-Deliver-AI-Solutions-Across-the-Oil-and-Gas-Industry-06-24-2019/default.aspx.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: A pretrained language model for scientific text, in: Proceedings of EMNLP-IJCNLP. https://doi.org/10.18653/v1/D19-1371, arXiv:1903.10676.

Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., et al., 2024. PaliGemma: A versatile 3B VLM for transfer. https://arxiv.org/abs/2407.07726. arXiv:2407.07726.

Blecher, L., Cucurull, G., Scialom, T., Stojnic, R., 2023. Nougat: Neural optical understanding for academic documents. https://arxiv.org/abs/2308.13418. arXiv:2308.13418.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Boiger, R., Churakov, S.V., Ballester Llagaria, I., Kosakowski, G., Wust, R., Prasianakis, N.I., 2024. Direct mineral content prediction from drill core images via transfer learning. Swiss Journal of Geosciences 117, 12. https://doi.org/10.1186/s00015-024-00458-3, arXiv:2403.18495.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al., 2021. On the opportunities and risks of foundation models. https://arxiv.org/abs/2108.07258. arXiv:2108.07258.

Bormann, P., Aursand, P., Dilib, F., Manral, S., Dischington, P., 2020. FORCE 2020 well log and lithofacies dataset for machine learning competition. Zenodo https://doi.org/10.5281/zenodo.4351156.

BP, 2025. Driving value: How AI is quietly reshaping bp's operations. Feature, https://www.bp.com/en/global/corporate/news-and-insights/energy-in-focus/how-ai-is-shaping-bp-operations.html.

Braik, M., Al Shehhi, A., Saputelli, L., Mata, C., Badmaev, D., Khan, M., Rahman, A., 2021. Automated subsurface knowledge (ASK) Thamama retrieval engine, in: SPE Annual Technical Conference and Exhibition (ATCE). https://doi.org/10.2118/206372-MS.

 doi:10.20944/preprints202604.1814.v1

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners (GPT-3), in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2005.14165, arXiv:2005.14165.

Cao, Z., Ma, C., Tang, W., Zhou, Y., Zhong, H., Ye, S., Wu, K., Chen, X., Hou, M., 2024. CoreViT: A new vision transformer model for lithofacies identification in cores. Geoenergy Science and Engineering 240, 213012. https://doi.org/10.1016/j.geoen.2024.213012.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers (DINO), in: Proceedings of ICCV. URL: https://arxiv.org/abs/2104.14294, arXiv:2104.14294.

Cayeux, E., Daireaux, B., Ambrus, A., Mihai, R., Carlsen, L., 2021. Autonomous decision-making while drilling. Energies 14, 969. https://doi.org/10.3390/en14040969.

Cayeux, E., Daireaux, B., Pelfrene, G., Mihai, R., 2025. Drilling automation: Revisiting the digital drilling program, in: SPE/IADC International Drilling Conference and Exhibition. https://doi.org/10.2118/223774-MS.

Chang, X., Zhang, Z., Allard, G., Higgins, I.R., 2025. Enhancing oil and gas knowledge retrieval: An LLM-powered pipeline with advanced synthetic data fine-tuning and post-filtering, in: 86th EAGE Annual Conference and Exhibition. https://doi.org/10.3997/2214-4609.202510629.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al., 2024. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology 15. https://doi.org/10.1145/3641289, arXiv:2307.03109.

Chebbi, A., Kolade, B., 2025. Towards EnergyGPT: A large language model specialized for the energy sector. https://arxiv.org/abs/2509.07177. arXiv:2509.07177.

Chen, F., Sun, L., Jiang, B., Huo, X., Pan, X., Feng, C., Zhang, Z., 2025a. A review of AI applications in unconventional oil and gas exploration and development. Energies 18, 391. https://doi.org/10.3390/en18020391.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z., 2024a. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. https://arxiv.org/abs/2402.03216. arXiv:2402.03216.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al., 2021. Evaluating large language models trained on code (HumanEval). https://arxiv.org/abs/2107.03374. arXiv:2107.03374.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations (SimCLR), in: Proceedings of ICML. arXiv:2002.05709.

Chen, Z., Lin, M., Wang, Z., Zang, M., Bai, Y., 2024b. PreparedLLM: Effective pre-pretraining framework for domain-specific large language models. Big Earth Data https://doi.org/10.1080/20964471.2024.2396159.

Chen, Z., Lin, M., Zang, M., Wang, Z., Li, J., Bai, Y., 2025b. JiuZhou: Open foundation language models and effective pre-training framework for geoscience. International Journal of Digital Earth 18. https://doi.org/10.1080/17538947.2025.2449708.

Chen, Z., Wang, X., Zhang, X., Lin, M., Liao, Y., Li, J., Bai, Y., 2025c. GeoFactory: An LLM performance enhancement framework for geoscience factual and inferential tasks. Big Earth Data 9, 225–248. https://doi.org/10.1080/20964471.2025.2506291.

Cheng, S., Harsuko, R., Alkhalifah, T., 2025. A generative foundation model for an all-in-one seismic processing framework. Surveys in Geophysics https://doi.org/10.1007/s10712-025-09912-9.

Chevron, 2025a. How APOLO helps Chevron pinpoint prime drilling locations. Chevron Newsroom, https://www.chevron.com/newsroom/2025/q4/how-apolo-helps-chevron-pinpoint-prime-drilling-locations.

Chevron, 2025b. A smarter way to prospect for oil and gas (ApEX). Chevron Newsroom, https://www.chevron.com/newsroom/2025/q4/a-smarter-way-to-prospect-for-oil-and-gas.

Chikhaoui, K., Alfarraj, M., 2024. Self-supervised learning for efficient seismic facies classification. Geophysics 89, IM61–IM76. https://doi.org/10.1190/geo2023-0508.1.

Cho, J., Mahata, D., İrsoy, O., He, Y., Bansal, M., 2024. M3DocRAG: Multimodal retrieval is what you need for multi-page multi-document understanding. https://arxiv.org/abs/2411.04952. arXiv:2411.04952.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2022. PaLM: Scaling language modeling with pathways. https://arxiv.org/abs/2204.02311. arXiv:2204.02311.

 doi:10.20944/preprints202604.1814.v1

Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al., 2022. Scaling instruction-finetuned language models (FLAN). https://arxiv.org/abs/2210.11416. arXiv:2210.11416.

Cicconeto, F., Vieira, L.V., Abel, M., Alvarenga, R.d.S., Carbonera, J.L., Garcia, L.F., 2022. GeoReservoir: An ontology for deep-marine depositional system geometry description. Computers & Geosciences 159, 105005. https://doi.org/10.1016/j.cageo.2021.105005.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J., 2021. Training verifiers to solve math word problems (GSM8K). https://arxiv.org/abs/2110.14168. arXiv:2110.14168.

Cordeiro, F.C., da Silva, P.F., Tessarollo, A., Freitas, C., de Souza, E., Gomes, D.d.S.M., Souza, R.R., Coelho, F.C., 2024. Petro NLP: Resources for natural language processing and information extraction for the oil and gas industry. Computers & Geosciences 193, 105714. https://doi.org/10.1016/j.cageo.2024.105714.

Danish, M.S., Munir, M.A., Shah, S.R.A., Kuckreja, K., Khan, F.S., Fraccaro, P., Lacoste, A., Khan, S., 2024. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. https://arxiv.org/abs/2411.19325. arXiv:2411.19325.

Das, A., Kong, W., Sen, R., Zhou, Y., 2024. A decoder-only foundation model for time-series forecasting (TimesFM), in: Proceedings of ICML. arXiv:2310.10688.

Dawson, H.L., Dubrule, O., John, C.M., 2023. Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification. Computers & Geosciences 171, 105284. https://doi.org/10.1016/j.cageo.2022.105284.

DeepSeek-AI, 2024a. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. https://arxiv.org/abs/2405.04434. arXiv:2405.04434.

DeepSeek-AI, 2024b. DeepSeek-V3 technical report. https://arxiv.org/abs/2412.19437. arXiv:2412.19437.

DeepSeek-AI, 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

Deng, C., Feng, S., Wang, H., Zhang, X., Jin, P., Feng, Y., Zeng, Q., Chen, Y., Lin, Y., 2022. OpenFWI: Large-scale multi-structural benchmark datasets for seismic full waveform inversion, in: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks. arXiv:2111.02926.

Deng, C., Jia, Y., Xu, H., Zhang, C., Tang, J., Fu, L., Zhang, W., Zhang, H., Wang, X., Zhou, C., 2021. GAKG: A multimodal geoscience academic knowledge graph, in: Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM). https://doi.org/10.1145/3459637.3482003.

Deng, C., Zhang, T., He, Z., Chen, Q., Shi, Y., Xu, Y., Fu, L., Zhang, W., Wang, X., Zhou, C., Lin, Z., He, J., 2024. K2: A foundation language model for geoscience knowledge understanding and utilization, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM), pp. 161–170. URL: https://arxiv.org/abs/2306.05064, https://doi.org/10.1145/3616855.3635772, arXiv:2306.05064.

Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2023. QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2305.14314, arXiv:2305.14314.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/, https://doi.org/10.18653/v1/N19-1423, arXiv:1810.04805.

Dong, T., Subia-Waud, C., Hou, S., 2024. Geo-RAG: Gaining insights from unstructured geological documents with large language models, in: Fourth EAGE Digitalization Conference and Exhibition. https://doi.org/10.3997/2214-4609.202439068.

Dong, X., Yu, W., Lin, J., Guo, Z., Wang, H., Yang, J., 2025. Light-weighted foundation model for seismic data processing based on representative and non-redundant pre-training dataset. https://arxiv.org/abs/2503.10092. arXiv:2503.10092.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale (ViT), in: International Conference on Learning Representations (ICLR). URL: https://arxiv.org/abs/2010.11929, arXiv:2010.11929.

Dou, Y., Wu, X., Bangs, N.L., Sethi, H.S., Li, J., Gao, H., Guo, Z., 2025. Geological everything model 3D (GEM): A promptable foundation model for unified and zero-shot subsurface understanding. https://arxiv.org/abs/2507.00419. arXiv:2507.00419.

 doi:10.20944/preprints202604.1814.v1

Dramsch, J.S., Lüthje, M., 2018. Deep-learning seismic facies on state-of-the-art CNN architectures, in: SEG Technical Program Expanded Abstracts. https://doi.org/10.1190/segam2018-2996783.1.

Eckroth, J., Boden, J., Hough, L., Gatewood, H., Gipson, S., Schoen, E., Gunderson, B., 2025. Building EnergyLLM: A domain-specific large language model trained on SPE content, in: SPE Annual Technical Conference and Exhibition (ATCE). URL: https://doi.org/10.2118/228097-MS, https://doi.org/10.2118/228097-MS.

Eckroth, J., Gipson, M., Boden, J., Hough, L., Elliott, J., Quintana, J., 2023. Answering natural language questions with OpenAI's GPT in the petroleum industry (PetroQA), in: SPE Annual Technical Conference and Exhibition (ATCE). https://doi.org/10.2118/214888-MS.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J., 2024. From local to global: A Graph RAG approach to query-focused summarization. https://arxiv.org/abs/2404.16130. arXiv:2404.16130.

Elyas, O.A., Al Hashim, H.W., Williams, J.R., 2025. Tailoring large language models for drilling applications: A comparative study of retrieval-augmented generation and fine-tuning, in: SPE Western Regional Meeting. https://doi.org/10.2118/224128-MS.

Equinor, 2018. Volve field data sharing. Open data release, https://www.equinor.com/energy/volve-data-sharing.

Ermilov, A., 2026. FormationEval, an open multiple-choice benchmark for petroleum geoscience. https://arxiv.org/abs/2601.02158. https://doi.org/10.2139/ssrn.6074466, arXiv:2601.02158.

ExxonMobil, 2025. Introducing Vantage: A new lens on upstream operations. Corporate page, https://corporate.exxonmobil.com/who-we-are/our-global-organization/business-divisions/upstream/a-new-lens-on-upstream-operations.

Fan, L., Liu, F., Chen, C., 2026. Domain adaptation of large language models for geotechnical applications. Solid Earth Sciences https://doi.org/10.1016/j.sesci.2025.100285.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P., 2025. ColPali: Efficient document retrieval with vision-language models, in: International Conference on Learning Representations (ICLR). URL: https://arxiv.org/abs/2407.01449, arXiv:2407.01449.

Feng, X., Dai, Y., Ji, X., Zhou, L., Dang, Y., 2021. Application of natural language processing in HAZOP reports. Process Safety and Environmental Protection 155, 41–48. https://doi.org/10.1016/j.psep.2021.09.001.

Ferrigno, E., Rodriguez, M., Davidsson, E., 2024. Revolutionizing drilling operations: Next-gen LLM-AI for real-time support in well construction control rooms, in: SPE Annual Technical Conference and Exhibition (ATCE). https://doi.org/10.2118/220798-MS.

Fuchs, F., Fernandez, M.R., Ettrich, N., Keuper, J., 2025. Foundation models for seismic data processing: An extensive review. https://arxiv.org/abs/2503.24166. arXiv:2503.24166.

Gao, H., Wu, X., Liang, L., Sheng, H., Si, X., Gao, H., Li, Y., 2026. A foundation model empowered by a multi-modal prompt engine for universal seismic geobody interpretation across surveys. Information Fusion https://doi.org/10.1016/j.inffus.2025.103437.

Gao, L., Ma, X., Lin, J., Callan, J., 2023. Precise zero-shot dense retrieval without relevance labels (HyDE), in: Proceedings of ACL. arXiv:2212.10496.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H., 2024. Retrieval-augmented generation for large language models: A survey. https://arxiv.org/abs/2312.10997. arXiv:2312.10997.

Garcia, L.F., Abel, M., Perrin, M., dos Santos Alvarenga, R., 2020. The GeoCore ontology: A core ontology for general use in geology. Computers & Geosciences 135, 104387. https://doi.org/10.1016/j.cageo.2019.104387.

Gemini Team, Google, 2023. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805. arXiv:2312.11805.

Gemini Team, Google, 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. https://arxiv.org/abs/2403.05530. arXiv:2403.05530.

Gemma Team, 2024. Gemma: Open models based on Gemini research and technology. https://arxiv.org/abs/2403.08295. arXiv:2403.08295.

Gharieb, A., Gabry, M.A., Soliman, M.Y., 2024. The role of personalized generative AI in advancing petroleum engineering and energy industry: A roadmap to secure and cost-efficient knowledge integration: A case study, in: SPE Annual Technical Conference and Exhibition (ATCE). https://doi.org/10.2118/220716-MS.

Ghorbanfekr, H., Kerstens, P.J., Dirix, K., 2025. Classification of geological borehole descriptions using a domain adapted large language model. Applied Computing and Geosciences https://doi.org/10.1016/j.acags.2025.100229.

 doi:10.20944/preprints202604.1814.v1

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., Dubrawski, A., 2024. MOMENT: A family of open time-series foundation models, in: Proceedings of ICML. arXiv:2402.03885.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al., 2024. The Llama 3 herd of models. https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., et al., 2020. Bootstrap your own latent (BYOL): A new approach to self-supervised learning, in: Advances in Neural Information Processing Systems (NeurIPS). arXiv:2006.07733.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X., 2024. Large language model based multi-agents: A survey of progress and challenges, in: Proceedings of IJCAI. arXiv:2402.01680.

Guo, Z., Wu, X., Liang, L., Sheng, H., Chen, N., Bi, Z., 2025. Cross-domain foundation model adaptation: Pioneering computer vision models for geophysical data analysis. JGR: Machine Learning and Computation 2. https://doi.org/10.1029/2025JH000601, arXiv:2408.12396.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks. Proceedings of ACL. https://doi.org/10.18653/v1/2020.acl-main.740.

Hadid, A., Chakraborty, T., Busby, D., 2024. When geoscience meets generative AI and large language models: Foundations, trends, and future challenges. Expert Systems 41. https://doi.org/10.1111/exsy.13654, arXiv:2402.03349.

Hall, B., 2016. Facies classification using machine learning. The Leading Edge 35, 906–909. https://doi.org/10.1190/tle35100906.1.

Halliburton, 2021. Halliburton launches DS365.ai intelligent automation. Press release, https://www.halliburton.com/en/about-us/press-release/halliburton-launches-ds365ai/.

Halliburton, 2025. Halliburton and PETRONAS to deploy next generation subsurface modeling and reservoir management solutions. Press release, https://www.halliburton.com/en/about-us/press-release/halliburton-petronas-deploy-next-generation-subsurface-modeling-reservoir-management/.

Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S.Q., 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. https://arxiv.org/abs/2403.14608. arXiv:2403.14608.

Harsuko, R., Alkhalifah, T., 2022. StorSeismic: A new paradigm in deep learning for seismic processing. IEEE Transactions on Geoscience and Remote Sensing 60, 1–15. https://doi.org/10.1109/TGRS.2022.3216660, arXiv:2205.00222.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners (MAE), in: Proceedings of CVPR. URL: https://arxiv.org/abs/2111.06377, arXiv:2111.06377.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning (MoCo), in: Proceedings of CVPR. arXiv:1911.05722.

He, Z., Sun, J., Guo, P., Wei, H., Lyu, X., Han, K., 2021. Construction of carbonate reservoir knowledge base and its application in fracture-cavity reservoir geological modeling. Petroleum Exploration and Development https://doi.org/10.1016/S1876-3804(21)60069-1.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J., 2021a. Measuring massive multitask language understanding (MMLU), in: International Conference on Learning Representations (ICLR). arXiv:2009.03300.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J., 2021b. Measuring mathematical problem solving with the MATH dataset, in: Advances in Neural Information Processing Systems (NeurIPS). arXiv:2103.03874.

Hoffimann, J., Mao, Y., Wesley, A., Taylor, A., 2017. Sequence mining and pattern analysis in drilling reports with deep natural language processing. https://arxiv.org/abs/1712.01476. arXiv:1712.01476.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., et al., 2022. Training compute-optimal large language models (Chinchilla), in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2203.15556, arXiv:2203.15556.

Hou, S., Dong, T., Sancheti, O., Liu, H., 2025a. Advancing geologic document digitalization and information retrieval with generative AI. The Leading Edge 44, 108–113. https://doi.org/10.1190/tle44020108.1.

Hou, S., Shen, Z., Zhao, A., Liang, J., Gui, Z., Guan, X., Li, R., Wu, H., 2025b. GeoCode-GPT: A large language model for geospatial code generation. International Journal of Applied Earth Observation and Geoinformation https://doi.org/10.1016/j.jag.2025.104456.

 doi:10.20944/preprints202604.1814.v1

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR). URL: https://arxiv.org/abs/2106.09685, arXiv:2106.09685.

Huang, F., Wu, F., Zhang, Z., Wang, Q., Zhang, L., Boquet, G.M., Chen, H., 2025. GeoGPT-RAG technical report. https://arxiv.org/abs/2509.09686. arXiv:2509.09686.

Huang, Z., Xu, W., Yu, K., 2015. Bidirectional LSTM-CRF models for sequence tagging. https://arxiv.org/abs/1508.01991. arXiv:1508.01991.

Hutahaean, J., Simon, K., 2025. Use of natural language processing and computer vision in deep learning for equipment failure investigation on drilling tools, in: International Petroleum Technology Conference (IPTC). https://doi.org/10.2523/IPTC-24706-MS.

Jacobs, T., 2025. New papers show automated, autonomous drilling systems headed in right direction. Journal of Petroleum Technology (JPT) URL: https://jpt.spe.org/new-papers-show-automated-autonomous-drilling-systems-headed-in-right-direction.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023. Mistral 7B. https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

Jiang, B., Liu, Z., Wang, N., Li, Z., Shi, Y., Lin, B., 2025. Process-oriented dual-layer knowledge GraphRAG for reservoir engineering decision support. Processes 13, 3230. https://doi.org/10.3390/pr13103230.

Johnson, J., Douze, M., Jégou, H., 2021. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7, 535–547. https://doi.org/10.1109/TBDATA.2019.2921572, arXiv:1702.08734.

Kainkaryam, S., Ong, C., Sen, S., Sharma, A., 2019. Crowdsourcing salt model building: Kaggle-TGS salt identification challenge, in: 81st EAGE Conference and Exhibition. https://doi.org/10.3997/2214-4609.201901271.

Kanfar, R., Alali, A., Tonellot, T.L., Salim, H., Ovcharenko, O., 2025. Intelligent seismic workflows: The power of generative AI and language models. The Leading Edge 44, 142–151. https://doi.org/10.1190/tle44020142.1.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t., 2020. Dense passage retrieval for open-domain question answering, in: Proceedings of EMNLP. https://doi.org/10.18653/v1/2020.emnlp-main.550, arXiv:2004.04906.

Khattab, O., Zaharia, M., 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT, in: Proceedings of SIGIR. arXiv:2004.12832.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S., 2022. OCR-free document understanding transformer (Donut), in: Proceedings of ECCV. arXiv:2111.15664.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R., 2023. Segment anything (SAM), in: Proceedings of ICCV. URL: https://arxiv.org/abs/2304.02643, arXiv:2304.02643.

Koeshidayatullah, A., Al-Azani, S., Baraboshkin, E.E., Alfarraj, M., 2022. FaciesViT: Vision transformer for improved core lithofacies prediction. Frontiers in Earth Science 10, 992442. https://doi.org/10.3389/feart.2022.992442.

Koeshidayatullah, A., Al-Fakih, A., Kaka, S., 2024. Leveraging time-series foundation model for subsurface well logs prediction and anomaly detection. https://arxiv.org/abs/2412.05681. arXiv:2412.05681.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y., 2022. Large language models are zero-shot reasoners. https://arxiv.org/abs/2205.11916. arXiv:2205.11916.

Koroteev, D., Tekic, Z., 2021. Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. Energy and AI 3, 100041. https://doi.org/10.1016/j.egyai.2020.100041.

Kowalchuk, P., 2019. Implementing a drilling reporting data mining tool using natural language processing sentiment analysis techniques, in: SPE Middle East Oil and Gas Show and Conference (MEOS). https://doi.org/10.2118/194961-MS.

Kuang, L., Liu, H., Ren, Y., Luo, K., Shi, M., Su, J., Li, X., 2021. Application and development trend of artificial intelligence in petroleum exploration and development. Petroleum Exploration and Development 48, 1–11. https://doi.org/10.1016/S1876-3804(21)60001-0.

Kumar, P., Kathuria, S., 2023. Large language models (LLMs) for natural language processing (NLP) of oil and gas drilling data, in: SPE Annual Technical Conference and Exhibition (ATCE). https://doi.org/10.2118/215167-MS.

Kumar, V., Gleyzer, L., Kahana, A., Shukla, K., Karniadakis, G.E., 2023. MYCRUNCHGPT: A LLM assisted framework for scientific machine learning. Journal of Machine Learning for Modeling and Computing https://doi.org/10.1615/JMachLearnModelComput.2023049518.

 doi:10.20944/preprints202604.1814.v1

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E.D., Kerner, H., Lütjens, B., Irvin, J.A., Dao, D., Alemohammad, H., Drouin, A., et al., 2023. GEO-Bench: Toward foundation models for earth monitoring. https://arxiv.org/abs/2306.03831. arXiv:2306.03831.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C., 2016. Neural architectures for named entity recognition, in: Proceedings of NAACL. arXiv:1603.01360.

Latrach, A., Malki, M.L., Morales, M., Mehana, M., Rabiei, M., 2024. A critical review of physics-informed machine learning applications in subsurface energy systems. Geoenergy Science and Engineering 239, 212938. https://doi.org/10.1016/j.geoen.2024.212938, arXiv:2308.04457.

Lawley, C.J.M., Raimondo, S., Chen, T., Brin, L., Zakharov, A., Kur, D., Hui, J., Newton, G., Burgoyne, S.L., Marquis, G., 2022. Geoscience language models and their intrinsic evaluation. Applied Computing and Geosciences 14, 100084. https://doi.org/10.1016/j.acags.2022.100084.

Leal, D., 2005. ISO 15926 "life cycle data for process plant": An overview. Oil & Gas Science and Technology https://doi.org/10.2516/ogst:2005045.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2005.11401, arXiv:2005.11401.

Li, C., Fomel, S., Chen, Y., Dommisse, R., Savvaidis, A., 2025a. FaultVitNet: A vision transformer assisted network for 3d fault segmentation. JGR: Machine Learning and Computation 2. https://doi.org/10.1029/2024JH000488.

Li, G., Hammoud, H.A.A.K., Itani, H., Khizbullin, D., Ghanem, B., 2023a. CAMEL: Communicative agents for "mind" exploration of large language model society, in: Advances in Neural Information Processing Systems (NeurIPS). arXiv:2303.17760.

Li, J., Pu, J., Zhu, Z., Li, G., Xu, M., He, X., Xu, S., Tian, S., Wang, T., 2025b. Intelligent management of fracturing knowledge in oil and gas fields: A novel transformation approach from document governance to natural language Q&A. Water https://doi.org/10.3390/w17223317.

Li, J., Tang, J., Zhou, J., Wang, D., Zhang, F., Zhao, S., 2025c. LithoGPT-Mini: A practical lightweight language model for fast and accurate complex lithology identification – application to underground gas storage facility, in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/229408-MS.

Li, K., Liu, W., Dou, Y., Xu, Z., Duan, H., Jing, R., 2023b. CONSS: Contrastive learning method for semisupervised seismic facies classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing https://doi.org/10.1109/JSTARS.2023.3308754, arXiv:2210.04776.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F., 2023c. TrOCR: Transformer-based optical character recognition with pre-trained models. https://arxiv.org/abs/2109.10282. arXiv:2109.10282.

Li, S., Marwah, T., Shen, J., Sun, W., Risteski, A., Yang, Y., Talwalkar, A., 2025d. CodePDE: An inference framework for LLM-driven PDE solver generation. https://arxiv.org/abs/2505.08783. arXiv:2505.08783.

Li, Y., Alkhalifah, T., Huang, J., Li, Z., 2023d. Self-supervised pretraining vision transformer with masked autoencoders for building subsurface model. IEEE Transactions on Geoscience and Remote Sensing 61. https://doi.org/10.1109/TGRS.2023.3308999.

Li, Z., Wang, Z., Wei, Z., Zhou, X., Wang, Y., Huai, B., Liu, Q., Yuan, N.J., Gong, R., Chen, E., 2021. Cross-oilfield reservoir classification via multi-scale sensor knowledge transfer. Proceedings of the AAAI Conference on Artificial Intelligence https://doi.org/10.1609/aaai.v35i5.16545.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., et al., 2023. Holistic evaluation of language models (HELM). Transactions on Machine Learning Research (TMLR) arXiv:2211.09110.

Lin, B., Jin, Y., Cao, Q., Meng, H., Pang, H., Wei, S., 2025. Developing a large language model for oil- and gas-related rock mechanics: Progress and challenges. Natural Gas Industry B 12, 110–122. https://doi.org/10.1016/j.ngib.2025.03.007.

Lin, Z., Deng, C., Zhou, L., Zhang, T., Xu, Y., Xu, Y., He, Z., Shi, Y., Dai, B., Song, Y., Zeng, B., Chen, Q., Miao, Y., Xue, B., Wang, S., Fu, L., Zhang, W., He, J., Zhu, Y., Wang, X., Zhou, C., 2024. GeoGalactica: A scientific large language model in geoscience. https://arxiv.org/abs/2401.00434. arXiv:2401.00434.

 doi:10.20944/preprints202604.1814.v1

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J., 2024a. RemoteCLIP: A vision-language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing 62, 1–16. https://doi.org/10.1109/TGRS.2024.3390838, arXiv:2306.11029.

Liu, G., Gong, R., Shi, Y., Wang, Z., Mi, L., Yuan, C., Zhong, J., 2022. Construction of well logging knowledge graph and intelligent identification method of hydrocarbon-bearing formation. Petroleum Exploration and Development https://doi.org/10.1016/S1876-3804(22)60047-8.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2023. Visual instruction tuning (LLaVA), in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2304.08485, arXiv:2304.08485.

Liu, H., Ren, Y., Li, X., Deng, Y., Wang, Y., Cao, Q., Du, J., Lin, Z., Wang, W., 2024b. Research status and application of artificial intelligence large models in the oil and gas industry. Petroleum Exploration and Development 51, 1049–1065. https://doi.org/10.1016/S1876-3804(24)60524-0.

Liu, Q., Ma, J., 2024. Foundation models for geophysics: Review and perspective. https://arxiv.org/abs/2406.03163. arXiv:2406.03163.

Liu, T., Münchmeyer, J., Laurenti, L., Marone, C., de Hoop, M.V., Dokmanić, I., 2024c. SeisLM: A foundation model for seismic waveforms. https://arxiv.org/abs/2410.15765. arXiv:2410.15765.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of ICCV. arXiv:2103.14030.

Lv, X., Xie, Z., Xu, D., Jin, X., Ma, K., Tao, L., Qiu, Q., Pan, Y., 2022. Chinese named entity recognition in the geoscience domain based on BERT. Earth and Space Science 9, e2021EA002166. https://doi.org/10.1029/2021EA002166.

Ma, J., Zhou, Y., Su, L., Yang, H., He, L., 2026. A domain-specific language model for engineering-scale geological reasoning and mineral exploration in the Qin-Hang belt. Engineering Applications of Artificial Intelligence https://doi.org/10.1016/j.engappai.2025.113651.

Ma, Z., Santos, J.E., Lackey, G., Viswanathan, H., O'Malley, D., 2024. Information extraction from historical well records using a large language model. Scientific Reports 14, 31272. https://doi.org/10.1038/s41598-024-81846-5, arXiv:2405.05438.

Macêdo, J.B., das Chagas Moura, M., Aichele, D., Lins, I.D., 2022. Identification of risk features using text mining and BERT-based models: Application to an oil refinery. Process Safety and Environmental Protection 158, 382–399. https://doi.org/10.1016/j.psep.2021.12.025.

Mahjour, S.K., Mahjour, S.S., 2025. Intelligent reservoir decision support: An integrated framework combining LLMs, advanced prompt engineering, and multimodal data fusion for real-time petroleum operations. https://arxiv.org/abs/2509.11376. arXiv:2509.11376.

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., Ermon, S., 2024. GeoLLM: Extracting geospatial knowledge from large language models, in: International Conference on Learning Representations (ICLR). arXiv:2310.06213.

Matheus, J., Tiwari, S., Szemat, W., Amaya, D., 2025. Revolutionizing real-time drilling: Leveraging large language models and retrieval-augmented generation for enhanced operational efficiency, in: Offshore Technology Conference (OTC). https://doi.org/10.4043/35742-MS.

Menezes, R., 2023. Beyond keywords: Comparing insights from unstructured data using generative GPT search and hybrid extractive search in petroleum exploration, in: EAGE Workshop on Data Science - From Fundamentals to Opportunities, pp. 1–8. https://doi.org/10.3997/2214-4609.202377039.

Menon, S.S., Mondal, T., Brahmachary, S., Panda, A., Joshi, S.M., Kalyanaraman, K., Jagtap, A.D., 2026. On scientific foundation models: Rigorous definitions, key applications, and a comprehensive survey. Neural Networks 198. https://doi.org/10.1016/j.neunet.2026.108567.

Microsoft, 2024. Azure Data Manager for Energy – OSDU data platform. Microsoft Azure product page, https://azure.microsoft.com/en-us/products/data-manager-for-energy.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality (word2vec), in: Advances in Neural Information Processing Systems (NeurIPS). arXiv:1310.4546.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J., 2024. Large language models: A survey. https://arxiv.org/abs/2402.06196. arXiv:2402.06196.

 doi:10.20944/preprints202604.1814.v1

Mosser, L., Aursand, P., Brakstad, K.S., Lehre, C., Myhre-Bakkevig, J., 2024. Exploration robot chat: Uncovering decades of exploration knowledge and data with conversational large language models, in: SPE Norway Subsurface Conference. https://doi.org/10.2118/218439-MS.

Mousavi, S.M., Beroza, G.C., 2022. Deep-learning seismology. Science 377. https://doi.org/10.1126/science.abm4470.

Mousavi, S.M., Ellsworth, W.L., Zhu, W., Chuang, L.Y., Beroza, G.C., 2020. Earthquake transformer – an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nature Communications 11, 3952. https://doi.org/10.1038/s41467-020-17591-w.

Mousavi, S.M., Sheng, Y., Zhu, W., Beroza, G.C., 2019. STEAD: A global data set of seismic signals for AI. IEEE Access 7, 179464–179476. https://doi.org/10.1109/ACCESS.2019.2947848.

Nassar, A., Livathinos, N., Lysak, M., Staar, P., 2022. TableFormer: Table structure understanding with transformers, in: Proceedings of CVPR. arXiv:2203.01017.

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., Xiong, C., 2022. CodeGen: An open large language model for code with multi-turn program synthesis. https://arxiv.org/abs/2203.13474. arXiv:2203.13474.

Noshi, C.I., Schubert, J.J., 2018. The role of machine learning in drilling operations; a review, in: SPE/AAPG Eastern Regional Meeting. https://doi.org/10.2118/191823-18ERM-MS.

NVIDIA, 2024. NVIDIA announces Earth climate digital twin. Press release, https://nvidianews.nvidia.com/news/nvidia-announces-earth-climate-digital-twin.

Ogundare, O., Madasu, S., Wiggins, N., 2023. Industrial engineering with large language models: A case study of ChatGPT's performance on oil & gas problems, in: 11th International Conference on Control, Mechatronics and Automation (ICCMA). https://doi.org/10.1109/ICCMA59762.2023.10374622, arXiv:2304.14354.

OpenAI, 2023. GPT-4 technical report. https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al., 2024. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR) URL: https://arxiv.org/abs/2304.07193, arXiv:2304.07193.

Osman, A.M., Ghali, R., Ibrahim, M., Hasan, S., Dewidar, M., 2025. Autonomous directional drilling revolutionizing efficiency and precision on lumpsum turnkey rigs through AI and automation, in: Middle East Oil, Gas and Geosciences Show (MEOS GEO). https://doi.org/10.2118/226893-MS.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback (InstructGPT), in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2203.02155, arXiv:2203.02155.

Pacis, F.J., Alyaev, S., Pelfrene, G., Wiktorski, T., 2024. Enhancing information retrieval in the drilling domain: Zero-shot learning with large language models for question-answering, in: IADC/SPE International Drilling Conference and Exhibition. https://doi.org/10.2118/217671-MS.

Pallanich, J., 2025. Ready for work: Translating AI ambitions into scalable results. Journal of Petroleum Technology (JPT), https://jpt.spe.org/ready-for-work-translating-ai-ambitions-into-scalable-results.

Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S., 2023. Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST). arXiv:2304.03442.

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global vectors for word representation, in: Proceedings of EMNLP. https://doi.org/10.3115/v1/D14-1162.

Pham, N., Di, H., Zhao, T., Abubakar, A., 2025. SeisBERT: A pretrained seismic image representation model for seismic data interpretation. The Leading Edge 44, 96–106. https://doi.org/10.1190/tle44020096.1.

Ponomareva, D., El Droubi, N., El Jundi, O., Assaf, G., Mustapha, H., Al Kindi, Z., 2024. Domain driven methodology adopting generative AI application in oil and gas drilling sector, in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/221883-MS.

Qi, Z., Yu, Q., Wang, J., Zhao, Y.B., Li, Z., Lv, W., 2025. WLFM: A well-logs foundation model for multi-task and cross-well geological interpretation. arXiv preprint arXiv:2509.18152 arXiv:2509.18152. arXiv preprint.

Qiu, Q., Xie, Z., Wu, L., Tao, L., Li, W., 2019. BiLSTM-CRF for geological named entity recognition from the geoscience literature. Earth Science Informatics 12, 565–579. https://doi.org/10.1007/s12145-019-00390-3.

Qwen Team, 2024a. Qwen2 technical report. https://arxiv.org/abs/2407.10671. arXiv:2407.10671.

Qwen Team, 2024b. Qwen2.5 technical report. https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

 doi:10.20944/preprints202604.1814.v1

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision (CLIP), in: Proceedings of ICML. URL: https://arxiv.org/abs/2103.00020, arXiv:2103.00020.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C., 2023. Direct preference optimization: Your language model is secretly a reward model, in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2305.18290, arXiv:2305.18290.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer (T5). Journal of Machine Learning Research 21, 1–67. URL: https://arxiv.org/abs/1910.10683, arXiv:1910.10683.

Rahmanifard, H., Plaksina, T., 2019. Application of artificial intelligence techniques in the petroleum industry: A review. Artificial Intelligence Review 52, 2295–2318. https://doi.org/10.1007/s10462-018-9612-8.

Rasul, K., Ashok, A., Williams, A.R., Ghonia, H., Bhagwatkar, R., Khorasani, A., Bayazi, M.J.D., Adamopoulos, G., Riachi, R., Hassen, N., et al., 2024. Lag-Llama: Towards foundation models for probabilistic time series forecasting. https://arxiv.org/abs/2310.08278. arXiv:2310.08278.

Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., et al., 2024. SAM 2: Segment anything in images and videos. https://arxiv.org/abs/2408.00714. arXiv:2408.00714.

Reddicharla, N., Kumar, A., Vanam, P.R., Sarma, V., Konkati, S., Jain, D., Patil, A.R., 2025. Revolutionizing well data access and analysis with AI: A case study and insights major Middle Eastern oil fields, in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/229370-MS.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of EMNLP-IJCNLP. arXiv:1908.10084.

Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R.Y., Dirani, J., Michael, J., Bowman, S.R., 2024. GPQA: A graduate-level Google-proof Q&A benchmark, in: Conference on Language Modeling (COLM). arXiv:2311.12022.

Rodrigues, R.B.M., Privatto, P.I.M., de Sousa, G.J., Murari, R.P., Afonso, L.C.S., Papa, J.P., Pedronette, D.C.G., Guilherme, I.R., Perrout, S.R., Riente, A.F., 2022. PetroBERT: A domain adaptation language model for oil and gas applications in Portuguese, in: Proceedings of PROPOR 2022: Computational Processing of the Portuguese Language, Springer, LNCS. https://doi.org/10.1007/978-3-030-98305-5_10.

Rodriguez Torrado, R., Pumar Jimenez, A., Shishehbor, M., Badawi, D., 2025. Field development plan optimization under uncertainty using physics-informed neural networks and generative AI: A real-field case study in North America, in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/229333-MS.

Romanenkova, E., Rogulina, A., Shakirov, A., Stulov, N., Zaytsev, A., Ismailova, L., Kovalev, D., Katterbauer, K., AlShehri, A., 2022. Similarity learning for wells based on logging data. Journal of Petroleum Science and Engineering 215. https://doi.org/10.1016/j.petrol.2022.110690.

Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X.E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., et al., 2023. Code Llama: Open foundation models for code. https://arxiv.org/abs/2308.12950. arXiv:2308.12950.

Sabbagh, V.B., Grimberg, M.N., Paes, D.M., 2025. Retrieval-augmented generation for drilling and completion knowledge management, in: OTC Brasil. https://doi.org/10.4043/36203-MS.

Sabbagh, V.B., Lima, C.B.C., Xexéo, G., 2024. Comparative analysis of single and multiagent large language model architectures for domain-specific tasks in well construction. SPE Journal 29, 6869–6882. https://doi.org/10.2118/223612-PA.

Samnioti, A., Gaganis, V., 2023a. Applications of machine learning in subsurface reservoir simulation – a review – part I. Energies 16, 6079. https://doi.org/10.3390/en16166079.

Samnioti, A., Gaganis, V., 2023b. Applications of machine learning in subsurface reservoir simulation – a review – part II. Energies 16, 6727. https://doi.org/10.3390/en16186727.

Sansal, A., Lasscock, B., Valenciano, A., 2025. Scaling seismic foundation models. First Break 43, 73–79. https://doi.org/10.3997/1365-2397.fb2025016.

Santos, N.O., Rodrigues, F.H., Schmidt, D., Romeu, R.K., Nascimento, G., Abel, M., 2024. O3PO: A domain ontology for offshore petroleum production plants. Expert Systems with Applications 238, 122104. https://doi.org/10.1016/j.eswa.2023.122104.

Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., Manning, C.D., 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval, in: International Conference on Learning Representations (ICLR). URL: https://arxiv.org/abs/2401.18059, arXiv:2401.18059.

 doi:10.20944/preprints202604.1814.v1

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T., 2023. Toolformer: Language models can teach themselves to use tools, in: Advances in Neural Information Processing Systems (NeurIPS). arXiv:2302.04761.

Schlumberger, 2017. Schlumberger announces DELFI cognitive E&P environment. News release, https://www.slb.com/newsroom/press-release/2017/pr-2017-0913-delfi.

Sharma, M., 2026. Case study: An agentic AI framework for large-scale well modeling of offshore field developments. Journal of Petroleum Technology (JPT) URL: https://jpt.spe.org/case-study-an-agentic-ai-framework-for-large-scale-well-modeling-of-offshore-field-developments.

Shell, 2024. Artificial intelligence. Shell Global, https://www.shell.com/ai.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y., 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face, in: Advances in Neural Information Processing Systems (NeurIPS). arXiv:2303.17580.

Sheng, H., Wu, X., Si, X., Li, J., Zhang, S., Duan, X., 2025. Seismic foundation model: A next generation deep-learning model in geophysics. Geophysics 90, IM59–IM79. URL: https://arxiv.org/abs/2309.02791, https://doi.org/10.1190/geo2024-0262.1, arXiv:2309.02791.

Shi, P., Wang, N., Wang, W., Yuan, S., Zhu, W., 2025. Denoising offshore distributed acoustic sensing using masked autoencoders to enhance earthquake detection. JGR: Solid Earth 130, e2024JB029728. https://doi.org/10.1029/2024JB029728.

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., Yao, S., 2023. Reflexion: Language agents with verbal reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS). arXiv:2303.11366.

Si, X., Wu, X., Sheng, H., Zhu, J., Li, Z., 2024. SeisCLIP: A seismology foundation model pre-trained by multimodal data for multi-purpose seismic feature extraction. IEEE Transactions on Geoscience and Remote Sensing 62, 1–13. https://doi.org/10.1109/TGRS.2024.3354456, arXiv:2309.02320.

Sidahmed, M., Coley, C.J., Shirzadi, S., 2015. Augmenting operations monitoring by mining unstructured drilling reports, in: SPE Digital Energy Conference and Exhibition. https://doi.org/10.2118/173429-MS.

Singh, A., Ehtesham, A., Kumar, S., Khoei, T.T., Vasilakos, A., 2025. Agentic retrieval-augmented generation: A survey on agentic RAG. https://arxiv.org/abs/2501.09136. arXiv:2501.09136.

Singh, A., Jia, T., Nalagatla, V., 2023. Generative AI enabled conversational chatbot for drilling and production analytics, in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/216267-MS.

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., Oza, H., 2021. Application of machine learning and artificial intelligence in oil and gas industry. Petroleum Research 6, 379–391. https://doi.org/10.1016/j.ptlrs.2021.05.009.

SLB, 2024a. SLB adds AI-driven geosteering to its autonomous drilling solutions to achieve more efficient and productive wells. News release, https://www.slb.com/newsroom/press-release/2024/slb-adds-ai-geosteering.

SLB, 2024b. SLB and NVIDIA collaborate to develop generative AI solutions for the energy sector. Press release, https://www.slb.com/newsroom/press-release/2024/slb-and-nvidia-collaborate-to-develop-generative-ai-solutions-for-the-energy-sector.

SLB, 2024c. SLB launches AI-powered Lumi platform. Press release, https://www.slb.com/newsroom/press-release/2024/slb-launches-ai-powered-lumi-platform.

SLB, 2025. SLB unveils groundbreaking new agentic AI technology for the energy industry (Tela). Press release, https://www.slb.com/news-and-insights/newsroom/press-release/2025/pr-2025-1103-slb-tela-ai.

SLB, 2026. SLB industrializes AI for the energy industry with NVIDIA. News release, https://www.slb.com/newsroom/press-release/2026/pr-2026-0325-slb-nvidia.

Society of Exploration Geophysicists, 2026. Advancing data analytics and machine learning for exploration geophysics. Workshop page, https://seg.org/calendar_events/advancing-data-analytics-machine-learning-for-exploration-geophysics/.

Society of Petroleum Engineers, 2025. SPE AI symposium: Navigating the nexus – where energy meets AI and sustainability. Event page, https://www.spe-events.org/symposium/artificial-intelligence.

 doi:10.20944/preprints202604.1814.v1

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., et al., 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (BIG-bench). Transactions on Machine Learning Research (TMLR) arXiv:2206.04615.

Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., et al., 2024. PaliGemma 2: A family of versatile VLMs for transfer. https://arxiv.org/abs/2412.03555. arXiv:2412.03555.

Tang, X., Feng, Z., Xiao, Y., Wang, M., Ye, T., Zhou, Y., Meng, J., Zhang, B., Zhang, D., 2023. Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development. Geoscience Frontiers 14, 101426. https://doi.org/10.1016/j.gsf.2022.101426.

Tariq, Z., Aljawad, M.S., Hasan, A., Murtaza, M., Mohammed, E., El-Husseiny, A., Alarifi, S.A., Mahmoud, M., Abdulraheem, A., 2021. A systematic review of data science and machine learning applications to the oil and gas industry. Journal of Petroleum Exploration and Production Technology (JPEPT) 11, 4339–4374. https://doi.org/10.1007/s13202-021-01302-2.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., Stojnic, R., 2022. Galactica: A large language model for science. https://arxiv.org/abs/2211.09085. arXiv:2211.09085.

Tharayil, S.M., Aldhalaan, B., Dossary, B., 2024. A language model for natural language interaction with transactional screens in the oil and gas industry, in: GOTECH Conference. https://doi.org/10.2118/219324-MS.

TotalEnergies, 2024. TotalEnergies unlocks the potential of generative artificial intelligence for its employees. Press release, https://totalenergies.com/news/press-releases/totalenergies-unlocks-potential-generative-artificial-intelligence-its.

TotalEnergies, 2025. TotalEnergies to collaborate with Mistral AI to increase the application of artificial intelligence in its multi-energy strategy. Press release, https://totalenergies.com/news/press-releases/totalenergies-collaborate-mistral-ai-increase-application-artificial.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G., 2023a. LLaMA: Open and efficient foundation language models. https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023b. Llama 2: Open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288. arXiv:2307.09288.

Tunkiel, A.T., Sui, D., Wiktorski, T., 2021. Reference dataset for rate of penetration benchmarking. Journal of Petroleum Science and Engineering 196. https://doi.org/10.1016/j.petrol.2020.108069.

Vargas, R.E.V., Munaro, C.J., Ciarelli, P.M., Medeiros, A.G., do Amaral, B.G., Barrionuevo, D.C., de Araujo, J.C.D., Ribeiro, J.L., Magalhaes, L.P., 2019. A realistic and public dataset with rare undesirable real events in oil wells (3W). Journal of Petroleum Science and Engineering 181. https://doi.org/10.1016/j.petrol.2019.106223.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/1706.03762, arXiv:1706.03762. transformer architecture; backbone of every LLM and FM in this survey.

Vimercati, F., Vignati, M., Mariotti, A., Severino, L., Raimondi, A., Onzaca, F., 2022. Application of natural language techniques in reservoir management framework, in: SPE Annual Technical Conference and Exhibition (ATCE). https://doi.org/10.2118/209961-MS.

Viridien, 2026. Viridien AI day: Innovation and insights driving the energy industry. Event page, https://www.viridiengroup.com/resources/media-events/symposia/viridien-ai-day.

Wang, B., Wu, L., Xie, Z., Qiu, Q., Zhou, Y., Ma, K., Tao, L., 2022a. Understanding geological reports based on knowledge graphs using a deep learning approach. Computers & Geosciences 168, 105229. https://doi.org/10.1016/j.cageo.2022.105229.

Wang, C., Li, Y., Chen, J., 2023a. Text mining and knowledge graph construction from geoscience literature legacy: A review, in: Recent Advancement in Geoinformatics and Data Science. Geological Society of America. volume 558, pp. 11–28. https://doi.org/10.1130/2022.2558(02).

Wang, F., Huang, X., Alkhalifah, T., 2024a. Controllable seismic velocity synthesis using generative diffusion models. JGR: Machine Learning and Computation 1, e2024JH000153. https://doi.org/10.1029/2024JH000153, arXiv:2402.06277.

 doi:10.20944/preprints202604.1814.v1

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A., 2023b. Voyager: An open-ended embodied agent with large language models. https://arxiv.org/abs/2305.16291. arXiv:2305.16291.

Wang, H., Chen, S., 2023. Insights into the application of machine learning in reservoir engineering: Current developments and future trends. Energies 16, 1392. https://doi.org/10.3390/en16031392.

Wang, J., Yoon, J., Marzban, A., Castanos, R., Fruge, N., Holman, I., 2025. Automatic daily drilling mud report processing using generative AI to maximize the operational efficiency, in: Offshore Technology Conference (OTC). https://doi.org/10.4043/35625-MS.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.R., 2023c. A survey on large language model based autonomous agents. https://arxiv.org/abs/2308.11432. arXiv:2308.11432.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., et al., 2024b. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. https://arxiv.org/abs/2409.12191. arXiv:2409.12191.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D., 2022b. Self-consistency improves chain of thought reasoning in language models. https://arxiv.org/abs/2203.11171. arXiv:2203.11171.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H., 2023d. Self-Instruct: Aligning language models with self-generated instructions, in: Proceedings of ACL. arXiv:2212.10560.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., et al., 2022c. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks, in: Proceedings of EMNLP. arXiv:2204.07705.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2201.11903, arXiv:2201.11903.

Weijermars, R., Waheed, U.b., Suleymanli, K., 2023. Will ChatGPT and related AI-tools alter the future of the geosciences and petroleum engineering? First Break https://doi.org/10.3997/1365-2397.fb2023043.

Wiegand, K., Bedewi, M., Mukundakrishnan, K., Tishechkin, D., Ananthan, V., Kahn, D., 2024a. Using generative AI to build a reservoir simulation assistant (ENVOY), in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/221987-MS.

Wiegand, K., Mukundakrishnan, K., Bedewi, M., Ananthan, V., Kahn, D., Tishechkin, D., Kajita, M., 2024b. Enhancing reservoir simulation workflows with generative AI for expert model building, quality control, and interpretation, in: Fourth International Meeting for Applied Geoscience & Energy (IMAGE). https://doi.org/10.1190/image2024-4099981.1.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., Sahoo, D., 2024. Unified training of universal time series forecasting transformers (Moirai), in: Proceedings of ICML. arXiv:2402.02592.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., et al., 2022. SEISBENCH: A toolbox for machine learning in seismology. Seismological Research Letters 93, 1695–1709. https://doi.org/10.1785/0220210324, arXiv:2111.00786.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., Wang, C., 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework, in: Conference on Language Modeling (COLM). URL: https://arxiv.org/abs/2308.08155, arXiv:2308.08155.

Wu, X., Liang, L., Shi, Y., Fomel, S., 2019. FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. Geophysics 84, IM35–IM45. https://doi.org/10.1190/geo2018-0646.1.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., et al., 2023. The rise and potential of large language model based agents: A survey. https://arxiv.org/abs/2309.07864. arXiv:2309.07864.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. SimMIM: A simple framework for masked image modeling, in: Proceedings of CVPR. arXiv:2111.09886.

Xu, C., Fu, L., Lin, T., Li, W., Ma, S., 2022. Machine learning in petrophysics: Advantages and limitations. Artificial Intelligence in Geosciences https://doi.org/10.1016/j.aiig.2022.11.004.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., 2020. LayoutLM: Pre-training of text and layout for document image understanding, in: Proceedings of KDD. https://doi.org/10.1145/3394486.3403172, arXiv:1912.13318.

Yan, S.Q., Gu, J.C., Zhu, Y., Ling, Z.H., 2024. Corrective retrieval augmented generation (CRAG). https://arxiv.org/abs/2401.15884. arXiv:2401.15884.

 doi:10.20944/preprints202604.1814.v1

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K., 2023a. Tree of thoughts: Deliberate problem solving with large language models, in: Advances in Neural Information Processing Systems (NeurIPS). URL: https://arxiv.org/abs/2305.10601, arXiv:2305.10601.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y., 2023b. ReAct: Synergizing reasoning and acting in language models, in: International Conference on Learning Representations (ICLR). URL: https://arxiv.org/abs/2210.03629, arXiv:2210.03629.

Yi, M., Ceglinski, K., Ashok, P., Behounek, M., White, S., Peroyea, T., Thetford, T., 2024. Applications of large language models in well construction planning and real-time operation, in: SPE/IADC International Drilling Conference and Exhibition. https://doi.org/10.2118/217700-MS.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E., 2024. A survey on multimodal large language models. National Science Review 11. https://doi.org/10.1093/nsr/nwae403, arXiv:2306.13549.

Yu, R., Chen, S., Xie, Y., Jia, X., 2025. A survey of foundation models for environmental science. https://arxiv.org/abs/2503.03142. arXiv:2503.03142.

Yu, S., Ma, J., 2021. Deep learning for geophysics: Current and future trends. Reviews of Geophysics 59. https://doi.org/10.1029/2021RG000742.

Zejli, A., Lin, A., Calva, B., Noh, M., Miller, P., Ranjith, R., Rai, V., Palanisamy, P., Lin, Y., Partington, B., 2025. Physics-informed agentic AI: An intelligent assistant for production management and optimization (DOAPE), in: SPE Permian Basin Energy Conference and Exhibition. https://doi.org/10.2118/230253-MS.

Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training (SigLIP), in: Proceedings of ICCV. https://doi.org/10.1109/ICCV51070.2023.01100, arXiv:2303.15343.

Zhang, D., Hu, Z., Zhoubian, S., Du, Z., Yang, K., Wang, Z., Yue, Y., Dong, Y., Tang, J., 2024a. SciInstruct: A self-reflective instruction annotated dataset for training scientific language models, in: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks. URL: https://arxiv.org/abs/2401.07950, arXiv:2401.07950.

Zhang, H., Xu, J.J., Cui, H.W., Li, L., Yang, Y., Tang, C.S., Boers, N., 2025. When geoscience meets foundation models: Toward a general geoscience artificial intelligence system. IEEE Geoscience and Remote Sensing Magazine https://doi.org/10.1109/MGRS.2024.3496478, arXiv:2309.06799.

Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., Han, J., 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. https://arxiv.org/abs/2406.10833. https://doi.org/10.18653/v1/2024.emnlp-main.498, arXiv:2406.10833.

Zhang, Y., Wang, Z., He, Z., Li, J., Mai, G., Lin, J., Wei, C., Yu, W., 2024c. BB-GeoGPT: A framework for learning a large language model for geographic information science. Information Processing & Management 61, 103808. https://doi.org/10.1016/j.ipm.2024.103808.

Zhang, Z., Chen, R., Ma, J., 2024d. Improving seismic fault recognition with self-supervised pre-training: A study of 3D transformer-based with multi-scale decoding and fusion. Remote Sensing 16, 922. https://doi.org/10.3390/rs16050922.

Zhang, Z., Hou, T., Kherroubi, J., Khvostichenko, D., 2022. Event detection in drilling remarks using natural language processing, in: IADC/SPE International Drilling Conference and Exhibition. https://doi.org/10.2118/208779-MS.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al., 2023. A survey of large language models. https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

Zhao, X., Wei, S., Wang, L., Deng, R., Zhang, L., Liu, D., Yang, L., Zang, K., Deng, Y., Xu, J., Luo, Y., 2025. Large language model empowered automated reservoir agent: A win-win strategy for reservoir intelligent management and transformation, in: Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC). https://doi.org/10.2118/229646-MS.

Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., et al., 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, in: Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks. arXiv:2306.05685.

Zhong, R., Salehi, C., Johnson, R., 2022. Machine learning for drilling applications: A review. Journal of Natural Gas Science and Engineering https://doi.org/10.1016/j.jngse.2022.104807.

Zhu, J., Du, Y., Zhang, W., Jiang, F., 2025. AI-driven knowledge management in oil and gas: A large language model approach to operational excellence, in: SPE Annual Caspian Technical Conference and Exhibition. https://doi.org/10.2118/230403-MS.

 doi:10.20944/preprints202604.1814.v1

Zhu, W., Beroza, G.C., 2019. PhaseNet: A deep-neural-network-based seismic arrival-time picking method. Geophysical Journal International 216, 261–273. https://doi.org/10.1093/gji/ggy423.

Zwartjes, P., Ovcharenko, O., Yoo, J., Smith, R., Al Ali, A., Rovetta, D., Salim, H., Tonellot, T., 2024. Building a large language model based seismic data processing assistant, in: 85th EAGE Annual Conference & Exhibition, pp. 1–5. https://doi.org/10.3997/2214-4609.202410350.