Large Language Models for Solid-Earth Geoscience: Applications, Modalities, and Memory-Reason-Act Techniques

Yisu Chen[2], Yiran Zhang[1], Yifeng Dai[1], Puzhen Wu[3], Sirui Li[1], Wei Liu[1], Yihao Ding[1]
[1]The University of Western Australia, [2]Wilfrid Laurier University,
[3]The University of Hong Kong
Corresponding author {wei.liu, yihao.ding}@uwa.edu.au

Abstract

Solid Earth geoscience, geology, geochemistry, and geophysics, is increasingly served by large language models (LLMs) and multimodal LLMs (MLLMs) that combine text, geological maps, hyperspectral imagery, 3D structural models, spatial databases, and quantitative measurements into operational decision pipelines. Existing studies are fragmented across tasks, datasets, and system designs. This survey provides a structured review of LLM/MLLM research in solid-Earth geoscience along three axes: applications, modalities, and Memory-Reason-Act techniques. We identify five clusters of application (information extraction, knowledge Q&A and retrieval, multimodal map and 3D interpretation, spatial predictive modelling, and agentic pipelines), characterise four modalities (text, visual, spatial, and quantitative signal) and their encoding gaps, and analyse Memory, Reasoning, and Action design choices with their failure modes. We further examine data resources and bias, and identify the absence of closed-loop drill-outcome validation as the central methodological gap blocking next-generation solid-Earth foundation models.

1 Introduction

Solid Earth geoscience, spanning geology, geochemistry, and geophysics, is an evidence-intensive and decision-critical domain. It supports mineral and energy exploration, subsurface characterisation, geological knowledge services, and geospatial analytics. These workflows rely on heterogeneous evidence, including textual records, geological maps and cross-sections, hyperspectral imagery, geochemical assays, geophysical signals, and spatial databases. Such evidence is multi-scale and high-stakes, linking continental tectonic context to mineral-scale assays, where errors may affect drilling and regulatory decisions.

Recent advances in large language models (LLMs) and multimodal LLMs (Schick et al., 2023; Yao et al., 2023; Zhang et al., 2025b) offer new opportunities for integrating, querying, and reasoning over geoscientific evidence. In solid Earth geoscience, these models have been applied to geological knowledge extraction (Knoblock et al., 2025; Pyo and Chiang, 2024), geoscience question answering and retrieval (Deng et al., 2024; Lin et al., 2024; Chen et al., 2025a; Zhao et al., 2025), multimodal map and 3D interpretation (Huang et al., 2025; Yu et al., 2024; Ma et al., 2026; Anikiev et al., 2026), spatial prediction (Ye et al., 2026; Parsa et al., 2026), and agent-driven geospatial workflow automation (Hou et al., 2024; Zhang et al., 2024b, 2025a; Liu et al., 2025). However, this emerging literature remains fragmented across tasks, modalities, and system designs. Existing work often focuses on isolated capabilities, while solid-Earth applications require systems that can ground multimodal evidence, reason across spatial scales, respect physical constraints, and support decisions that can be validated against field outcomes (Hadid et al., 2024; Zhang and Xu, 2023). These requirements are not well captured by general-purpose LLM benchmarks and call for a domain-specific synthesis.

Related Survey. Prior surveys (Table 6) have examined machine learning, deep learning, foundation models, and generative AI for geoscience from several perspectives. Some provide broad overviews of classical machine learning, deep learning, and scientific machine learning for Earth and environmental sciences (Karpatne et al., 2019; Reichstein et al., 2019; Willard et al., 2022), while others focus on specific modalities or task families, including remote sensing (Zhu et al., 2017), seismic interpretation (Dramsch, 2020), and geoscience language-model evaluation (Lawley et al., 2022). A related line of reviews focuses on mineral exploration, mineral prospectivity mapping,

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4776–4790
November 6-10, 2026 ©2026 Association for Computational Linguistics

and critical-mineral discovery, summarising AI/ML pipelines for exploration targeting, data integration, and predictive modelling (Yang et al., 2024; Sun et al., 2024; Fu et al., 2025b; Davies et al., 2025). More recent work discusses foundation models and LLMs in geoscience more broadly (Zhang and Xu, 2023). Hadid et al. (Hadid et al., 2024) provide the closest high-level reference by surveying generative AI across Earth-system applications, including Earth observation, hazards, and climate.

This paper provides a structured review of LLM and MLLM research in solid Earth geoscience. First, we define the scope of solid-Earth LLM systems and set inclusion criteria for the surveyed literature (§2). Second, we organise studies into five application clusters: information extraction and knowledge-graph construction, geological QA and retrieval, multimodal map and 3D interpretation, spatial predictive modelling, and agentic workflow automation (§3). Third, we examine how systems encode and ground text, visual, spatial, and quantitative signals, identifying modality-specific gaps (§4). Fourth, we introduce a Memory-Reason-Act (MRA) framework to compare how systems store geological knowledge, perform inference, and generate outputs or actions (§5). Finally, we discuss datasets, bias, and evaluation (§6), and highlight open challenges for next-generation systems, especially closed-loop validation against field and drilling outcomes (§7).

2 Survey Methodology

We searched ACL Anthology, arXiv, Springer, Elsevier, Frontiers, MDPI, and SSRN for publications from January 2023 to March 2026. For arXiv, we focused on cs.CL and physics.geo-ph, while also tracking relevant cross-listed papers on solid Earth geoscience. Search queries combined language-model terms, including *large language model*, and *language model*, with geoscience terms such as *geology*, *geochemistry*, *geophysics*, *mineral*, *prospectivity*, *drilling*, and *subsurface*. We further expanded the candidate pool through backward and forward citation tracing. Three inclusion criteria were applied. First, the system must use natural language as an input, output, or intermediate reasoning interface. Second, the target task must lie within solid Earth geoscience, including geological knowledge extraction, mineral exploration, geochemical or geophysical interpretation, subsurface characterisation, and geospatial analysis. Atmospheric and oceanographic studies, as well as general-purpose remote-sensing foundation models without explicit solid-Earth tasks, were excluded. Third, the work had to build on Transformer-based language models, including encoder-only, encoder-decoder, decoder-only, or multimodal variants. Earlier pre-Transformer systems were included only when necessary to clarify the historical development of current LLM-era methods. The final corpus contains 17 representative LLM-era solid-Earth/geoscience systems and related language-interface frameworks, summarised in Table 1.

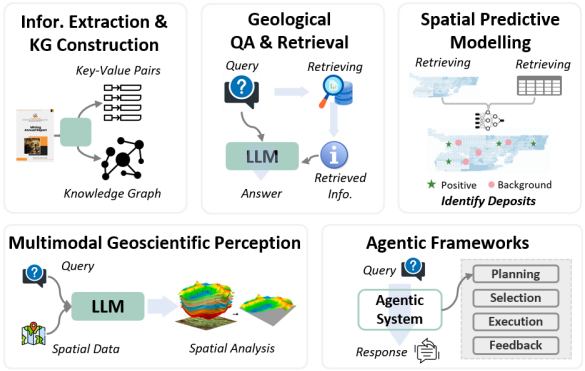


Figure 1: Geo-Related LLM Application Overview

3 Applications

Solid-Earth LLM systems are emerging for structuring, querying, interpreting, predicting from, and operationalising geological evidence. We group them into five application clusters: information extraction and knowledge-graph construction, geological Q&A and retrieval, multimodal map and 3D interpretation, spatial predictive modelling, and agentic workflow automation.

Information Extraction and Knowledge Graph Construction. Many solid-Earth workflows require converting heterogeneous textual sources into structured geological knowledge. Reports, drill logs, and scientific literature contain information on mineral sites, commodities, geological units, observations, and relations that must be extracted for search, comparison, and downstream modelling (Knoblock et al., 2025; Pyo and Chiang, 2024; Ma et al., 2023). Systems (Knoblock et al., 2025; Pyo and Chiang, 2024) use LLMs to extract entities, relations, and mineral-site attributes, while others (Deng et al., 2024; Zhang et al., 2024a) support knowledge-graph alignment through domain adaptation. Related studies on iron-deposit KGs, mineral-exploration semantic search,

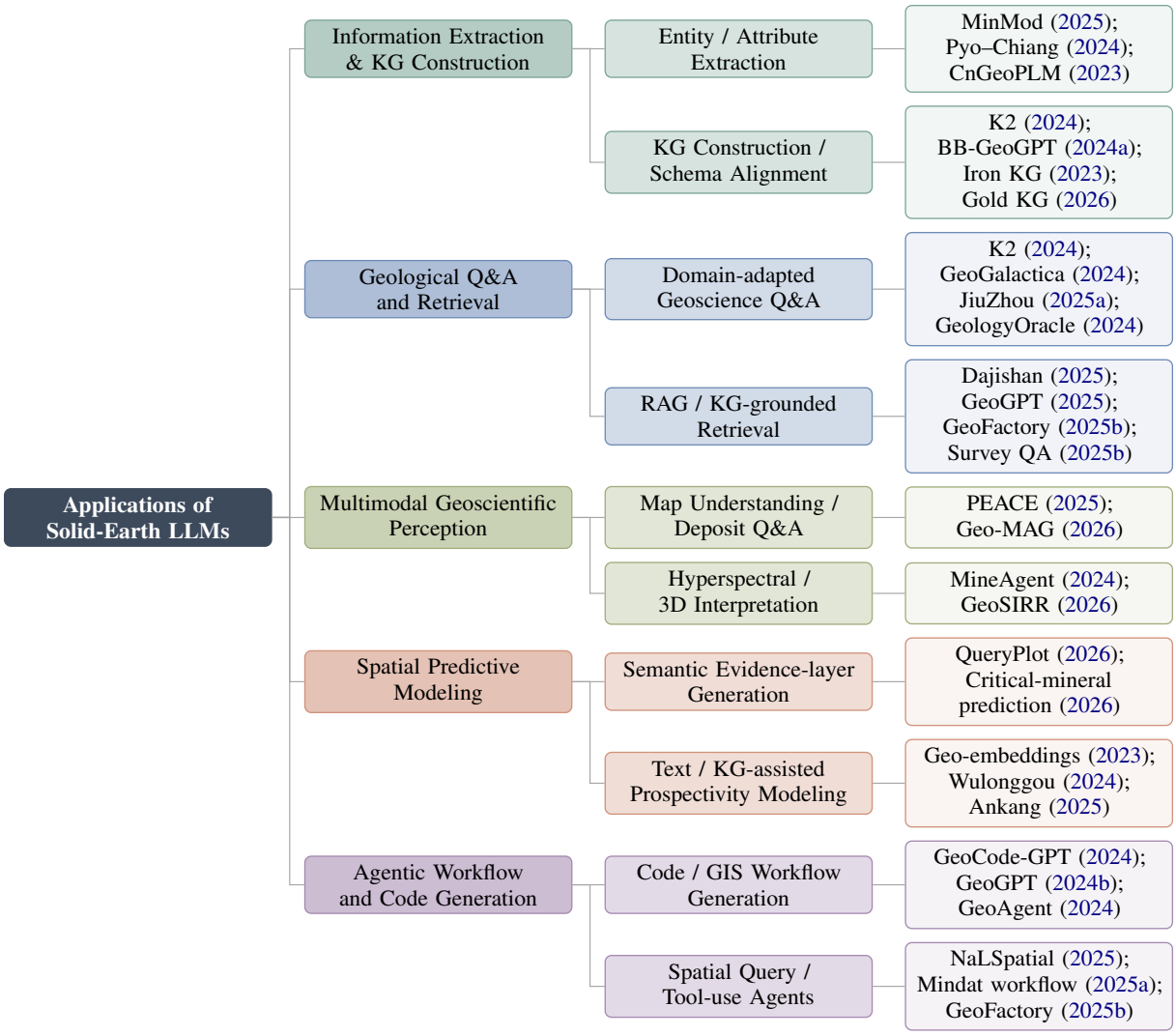


Figure 2: Taxonomy of solid-Earth LLM applications and representative systems. KG = knowledge graph; Q&A = question answering; RAG = retrieval-augmented generation; GIS = geographic information system.

mineral-system NER, and ore-forming condition extraction show that performance depends heavily on schema design, commodity-specific terminology, and geological-attribute granularity (Qiu et al., 2023, 2024; Villacorta Chambi et al., 2025; Enkhsaikhan et al., 2021a). The main challenge is therefore not only extraction accuracy, but also auditability and interoperability, as quantitative attributes, location descriptions, and geological relations remain difficult to standardise across reports, commodities, and ontologies (Wang et al., 2026).

Geological Knowledge Q&A and Retrieval. Once geological knowledge has been collected or indexed, LLMs can provide natural-language access to scientific literature, survey archives, and deposit records. These systems support explanation, fact retrieval, long-document summarisation, and question answering over dispersed geological evidence. Domain models (Deng et al., 2024; Lin et al., 2024; Chen et al., 2025a; Baucon and Neto de Carvalho, 2024) adapt LLMs to geoscience terminology and mineral-exploration knowledge, while retrieval-augmented systems ground answers in literature, survey reports, or deposit KGs (Zhao et al., 2025; Li et al., 2025; Chen et al., 2025b). Field-oriented mineral survey QA systems further show that KG and retrieval grounding work best when questions are tied to explicit survey workflows, mineral-resource entities, and field-report evidence rather than open-ended geological conversation (Wang et al., 2025b). The key limitation is reliability: answers must remain grounded when evidence is incomplete, conflicting, or distributed across documents and modalities.

Multimodal Geoscientific Perception. Beyond text, many geoscience decisions depend on visual

and spatial artefacts such as geological maps, hyperspectral imagery, and 3D structural data (Huang et al., 2025; Yu et al., 2024; Ma et al., 2026). MLLMs are therefore being used to answer map-based questions, relate visual patterns to geological concepts, and assist interpretation of mineralisation environments or structural cross-sections. Representative systems include PEACE for geological-map understanding (Huang et al., 2025), Geo-MAG for map-based deposit Q&A (Ma et al., 2026), MineAgent for hyperspectral mineral-exploration reasoning (Yu et al., 2024), and GeoSIRR for conversational 3D cross-section generation (Anikiev et al., 2026). Across these systems, multimodal gains depend on adding geological structure to MLLMs, including hierarchical map parsing, KG-enhanced retrieval, task-specific map benchmarks, or constrained 3D geological representations.

Spatial Predictive Modeling. The outputs of geological interpretation often feed spatial prediction tasks, particularly mineral prospectivity mapping (MPM) (Ye et al., 2026; Parsa et al., 2026; Zhang et al., 2026b). In this setting, LLMs usually contribute semantic evidence rather than replacing geostatistical or spatial-learning models. They encode geological descriptions, extract text-derived indicators, or link domain knowledge to spatial units used in prediction. Earlier works used geoscience embeddings, entity-attribute representations, or knowledge-graph embeddings as predictive evidence layers (Lawley et al., 2023; Parsa et al., 2025; Yan et al., 2024, 2025), while recent systems use LLM-era models to summarise map-unit descriptions or produce spatial embeddings for downstream prediction (Ye et al., 2026; Parsa et al., 2026).

Agentic Workflow and Code Generation. LLMs are also being used to operationalise geoscience tasks by converting analytical intent into executable workflows. Instead of only returning textual answers, these systems generate GIS scripts, spatial SQL queries, API calls, or multi-step tool-use processes (Hou et al., 2024; Zhang et al., 2024b; Liu et al., 2025; Zhang et al., 2025a). Examples include GeoCode-GPT for geospatial code generation (Hou et al., 2024), GeoGPT and GeoAgent for GIS tool coordination (Chen et al., 2024), NaLSpatial for natural-language spatial database querying (Liu et al., 2025), and Mindat-based agents for mineral co-occurrence analysis (Zhang et al., 2025a). The main application risk is execution reliability: ambiguous user intent, changing APIs, invalid spatial operations, and insufficient verification can turn plausible language outputs into incorrect analytical results, especially when geological plausibility and field validity are not checked.

4 Modality

Solid-Earth LLM systems must integrate text with geological maps, hyperspectral imagery, 3D subsurface models, spatial database entries, and quantitative observations from geochemical assays, geophysical waveforms, and well-log time series (Deng et al., 2024; Huang et al., 2025; Yu et al., 2024; Fu et al., 2025a). These modalities are heterogeneous and span wide spatial and temporal scales, from continental tectonic context to mineral-scale assays (Hadid et al., 2024; Denli et al., 2021). No unified representation has emerged; instead, current systems address each modality through different ad-hoc adaptations, which we describe in turn.

Text. Text, geological reports, scientific publications, drill logs, and domain technical manuals, is the dominant modality across solid-Earth LLM systems (Deng et al., 2024; Lin et al., 2024; Fu et al., 2025a) and the one where encoding strategies are most mature: continued pre-training and instruction tuning on geoscience corpora (§5.1) are the default mechanism for transferring textual knowledge into model parameters. The remaining open issues are data-side rather than method-side: PDF extraction noise, semantic drift across reporting standards, sparse coverage of non-English geoscience archives, and the under-representation of quantitative and spatial reasoning in text corpora limit how much downstream capability text-only adaptation can deliver (Denli et al., 2021; Ma et al., 2023; Zhang et al., 2024a).

Visual. Visual modalities include geological maps, hyperspectral imagery, cross-sections, and 3D structural volumes (Huang et al., 2025; Yu et al., 2024; Ma et al., 2026; Anikiev et al., 2026). Unlike natural images, they encode domain-specific structure: cartographic symbols carry semantic meaning, hyperspectral data extends far beyond RGB channels, and 3D volumes encode topology absent in 2D inputs. Most systems reuse general-purpose MLLM backbones and adapt geoscience inputs via preprocessing rather than training domain-specific encoders. MineAgent (Yu et al., 2024) converts hyperspectral bands into RGB-like deposit signa-

tures, while Geo-MAG (Ma et al., 2026) applies a YOLOv10-based pipeline to extract cartographic features into structured tokens for multimodal RAG. PEACE (Huang et al., 2025) quantifies the limitation, showing that GPT-4o achieves only 36.9

Spatial. Spatial information, coordinates, GIS layers, vector geometries, and spatial database entries, cannot be tokenised as text without losing structural semantics, and LLMs lack native operators for geometric relations such as containment, adjacency, or coordinate-system transformation (Zhang et al., 2024b; Liu et al., 2025). Most systems therefore avoid encoding spatial structure inside the model and instead delegate spatial reasoning to external tools at inference time (§5.2), translating natural-language requests into GIS operations or spatial SQL (Liu et al., 2025; Chen et al., 2024; Zhang et al., 2024b). The delegation pattern bypasses the context-window cost of representing large or high-resolution spatial data but transfers the burden of spatial correctness to the underlying tools and to the quality of generated code.

Quantitative signals. Geochemical assays, geophysical waveforms, well-log time series, and gravity-magnetic measurements are central to subsurface characterisation but remain the least developed modality. Existing approaches either serialise numerical sequences into text or bypass them entirely by routing analysis to external tools. The pre-LLM lineage, including GloVe embeddings (Lawley et al., 2023), BERT entity-attribute embeddings for the Wulonggou region (Yan et al., 2024), and semantic knowledge-graph embeddings feeding Conformer architectures for Ankang gold prospecting (Yan et al., 2025), shows that domain-specific dense representations of quantitative observations are feasible, but no published LLM-era system uses a generative language model as a predictive backbone over assay or geophysical signal data (Ye et al., 2026; Parsa et al., 2026). Closing this gap will likely require foundation models for numerical geoscience signals, parallel to recent seismic foundation models (Sheng et al., 2025; Dou et al., 2025), or principled interfaces between language models and signal-specific encoders.

5 MRA-Oriented Techniques

We organise the techniques used by solid-Earth LLM systems along three categories that describe how a system uses its modalities: *memory*, where geological knowledge is stored; *reasoning*, how inference proceeds over that knowledge; and *action*, what form the output takes. Every system combines choices along all three axes, and these choices jointly determine its hallucination risk, evaluation surface, and deployment shape.

5.1 Memory: Where Knowledge Lives

Continued pre-training and model alignment. Continued pre-training (CPT) on large geoscience corpora is the most common adaptation strategy for parametric memory. Parameters from some language model pre-trained on general corpora are first initialised. The same language-modelling objective is then used for geoscience domain corpora with the aim to produce richer internal representations for geoscientific vocabulary and factual associations. K2 (Deng et al., 2024), GeoGalactica (Lin et al., 2024), GeoProspect (Wu et al., 2025), and JiuZhou (Chen et al., 2025a) all follow this pattern.

Instruction tuning then aligns model behaviour with the input-output format desired for geoscience tasks. This typically takes the form of QA rather than strictly the sequence continuation task in pre-training. The instruct datasets encompass general instruction-following data and domain-specific QA or geospatial code generation. The method used for alignment amongst the models in Table 1 is supervised fine-tuning with teacher forcing.

Symbolic and retrieval-augmented memory. Symbolic knowledge graphs encode deposit ontologies, stratigraphic relations, and mineralogical associations as explicit triples (Knoblock et al., 2025; Zhao et al., 2025; Li et al., 2025), while vector indexes retrieve passages from literature and survey documents through dense similarity search (Ye et al., 2026; Baucon and Neto de Carvalho, 2024; Chen et al., 2025b). Hybrid architectures combine both: Geo-MAG (Ma et al., 2026) reports a 23.5% $\rightarrow$ 77.2% accuracy progression on polymetallic map QA by adding a Neo4j knowledge graph atop multimodal RAG over GPT-4o. KG branches encode long-shelf-life deposit-type theory under schema-checked constraints; vector branches absorb evolving regional reports without forcing schema migration. The cost is engineering complexity and the absence of a shared ontology across systems.

System	# Params	Task	Input Modality				Base Model	Training			Infer.	Avail.
			Text	Image	Spatial	Temporal		CPT	SFT	RL		
Domain-specific Geoscience LLM												
K2, 2024	7B	GQA	✓	✗	✗	✗	LLaMA-7B	✓	✓	✗	P	✓
GeoGalactica, 2024	30B	GQA, KCE	✓	✗	✗	✗	Galactica	✓	✓	✗	P	✓
BB-GeoGPT, 2024a	7B	GQA, GDA	✓	✗	✗	✗	LLaMA-2-7B	✓	✓	✗	P	✗
GeoMinLM, 2025a	7B/13B	GQA, DSF	✓	✗	✗	✗	Baichuan-2	✗	✓	✗	R	✓
Jiuzhou, 2025a	7B	GQA	✓	✗	✗	✗	Mistral-7B	✓	✓	✗	P	✓
GeoProspect, 2025	7B	GQA, DSF	✓	✗	✗	✗	Qwen2-7B	✓	✓	✗	P	✗
LLM Frameworks & Agents for Geoscience												
GeoWorkflow-LLM, 2025a	✗	GWA, GDA	✓	✓	✓	✗	GPT-4o	✗	✗	✗	T	✗
MineAgent, 2024	✗	RS, GWA	✓	✗	✓	✗	MLLMs	✗	✗	✗	A	✗
Geo-MAG, 2026	✗	GQA, GDA	✓	✓	✓	✗	GPT-4o	✗	✗	✗	R	✓
GeoSIRR, 2026	✗	GDA, GWA	✓	✗	✓	✗	LLMs	✗	✗	✗	T	✗
QueryPlot, 2026	✗	GDA, DSF	✓	✗	✓	✗	Text encoder	✗	✗	✗	R	✓
GeoFactory, 2025b	✗	GQA, GDA	✓	✗	✗	✗	Mistral-7B	✗	✗	✗	T	✓

Table 1: Summary of representative solid-Earth/geoscience LLM systems and closely related language-interface frameworks. ✓ = supported; ✗ = not applicable or not reported. CPT = continued pre-training; SFT = supervised fine-tuning; RL = reinforcement learning; Infer. = inference strategy; P = prompting; R = retrieval-augmented generation; T = tools-augmented generation; A = agentic system; Avail. = availability; Task category explanations are provided in Table 2.

5.2 Reasoning: How Inference Proceeds

Prompting and chain-of-thought. Prompt engineering and chain-of-thought (CoT) reasoning (Wei et al., 2022; Zhang et al., 2026a) require no parameter updates and remain the most widely used inference paradigm. In solid-Earth applications, prompts typically constrain output formats and elicit multi-step reasoning over GIS operations (Zhang et al., 2024b) or geological concept descriptions (Huang et al., 2025). CoT carries a domain-specific risk: a fabricated ore-genesis chain is geologically *coherent* because the model has absorbed ore-genesis textbook material to produce plausible fluid-pathway sequences, yet may be attached to the wrong deposit. Unlike mathematical or program CoT, ore-genesis chains lack automatic verifiers, and their plausibility is itself the failure mode (Li et al., 2025).

Tool-augmented and KG-guided reasoning. Tool augmentation delegates computation, spatial operations, and data-intensive tasks to external software whose outputs ground model generation in deterministic results (Zhang et al., 2025a; Chen et al., 2025b; Liu et al., 2025). KG-guided variants constrain reasoning to traverse triples retrieved from a domain knowledge graph, lowering hallucination risk at the cost of restricted expressiveness (Zhao et al., 2025; Li et al., 2025; Ma et al., 2026). The failure mode across both is interface translation: ambiguous requests must be mapped onto tool calls or KG paths, and reasoning-tuned models have been observed to *hurt* extraction quality by inducing hallucinated relations (Li et al., 2025).

Agentic reasoning. Agentic inference combines prompting, retrieval, and tool use within iterative decision loops. Rather than producing a single response, the system decomposes the task, executes a sequence of actions, observes results, and updates its plan accordingly. Existing geoscience agents adopt ReAct-style architectures (Yao et al., 2023), MineAgent's four-module hierarchy (Yu et al., 2024), Zhang et al.'s supervisor, worker pattern over Mindat (Zhang et al., 2025a), and GeoAgent's code-execution loop (Chen et al., 2024), solving multi-step problems single-pass generation cannot, particularly when geospatial data, code execution, or cross-document evidence are required.

5.3 Action: Output Forms and Validation Deficit

Output artefacts. Solid-Earth LLM systems produce four kinds of output, each with distinct verifiability and trust thresholds: *knowledge graph triples* (highly verifiable but indirect for decision support, as in MinMod (Knoblock et al., 2025) and Dajishan (Zhao et al., 2025)); *spatial prediction maps* (subject to holdout cross-validation and directly tied to drill targeting, as in QueryPlot (Ye et al., 2026) and the Canadian critical-mineral system (Parsa et al., 2026)); *natural-language recommendations* (low verifiability but high expressive reach, as in

GeoMinLM (Fu et al., 2025a) and GeoProspect (Wu et al., 2025)); and *agent actions or API calls* (fully logged but bounded by the underlying tool surface, as in Zhang et al. (Zhang et al., 2025a) and MineAgent (Yu et al., 2024)).

Closed-loop validation and verifiable rewards. None of the LLM-era systems reviewed close the loop from model output to a physical geological outcome. MPM systems are evaluated against held-out known deposits, which is the standard MPM protocol but does not track new recommendations through to drilling (Ye et al., 2026; Parsa et al., 2026); decision-support QA systems are scored on F1, BLEU, or entity-level extraction accuracy, none of which correlate with whether a recommended target hosts ore (Fu et al., 2025a; Wu et al., 2025; Zhao et al., 2025); KG-construction systems are scored on triple-level precision and recall, measuring faithfulness to the source report rather than to physical geology (Knoblock et al., 2025). The absence of closed-loop ground truth also blocks reinforcement learning with verifiable rewards (Guo et al., 2025), currently feasible only when generated geospatial code can be executed against a reference system (Chen et al., 2025b). Closing this gap is the central methodological challenge for solid-Earth LLM evaluation (§7).

6 Datasets

Solid-Earth LLM datasets fall into three stages: pre-training for domain knowledge, instruction tuning for prompt-following behaviour, and supervised fine-tuning (SFT) for task-level capability, each with distinct scale, supervision cost, and quality risks. A fourth concern cutting across all three is dataset bias and evaluation validity, which remain under-addressed in the literature.

Pre-training corpora. Pre-training shifts model statistics toward geoscientific terminology through large unsupervised corpora. K2 (Deng et al., 2024) draws on Earth-science Wikipedia and the GAKG knowledge graph (∼1.1M geoscience papers and 4.3M metadata records from DDE Scholar), GeoGalactica (Lin et al., 2024) extends Galactica with 52.7B tokens of geological literature and arXiv content, JiuZhou (Chen et al., 2025a) interleaves a 6.95B-token geoscience subcorpus with general text, and GeoProspect (Wu et al., 2025) curates 181M tokens from ChinaXiv, CNKI, and public geological repositories. Chinese systems CnGeoPLM (Ma et al., 2023) and GeoMinLM (Fu et al., 2025a) target under-resourced Mandarin reports, while GeoBERT (Denli et al., 2021) remains the only public account of an industrial corpus (20M proprietary records from ExxonMobil). PDF extraction noise, diagram and cross-section loss, and English/Chinese over-representation relative to other geoscience-producing languages remain the dominant quality bottlenecks.

Supervised fine-tuning resources. SFT data turns domain knowledge into task-level capability under explicit supervision to align LLM models with the desired input-output format. Five alignment categories dominate solid-Earth work: (i) domain QA pairs from geological textbooks and survey reports (Fu et al., 2025a; Wu et al., 2025); (ii) KG-aligned extraction triples, exemplified by the MinMod NI 43-101 corpus of 680,000 mineral sites across 190 commodities (Knoblock et al., 2025) and Pyo and Chiang's silver-standard mineral-site linkage pairs (Pyo and Chiang, 2024); (iii) geospatial code generation corpora such as GeoCode-GPT's 275K snippets and 10K operator examples mined from ArcGIS, MATLAB, R, and Python repositories (Hou et al., 2024); (iv) tool-use supervision derived from API-call traces over domain-specific platforms such as the Mindat mineralogical database (Zhang et al., 2025a); and (v) map and 3D visual QA benchmarks built from polymetallic geological maps and SGMC polygon descriptions (Huang et al., 2025; Ma et al., 2026; Ye et al., 2026). Most resources remain text-centred, fragmented across study areas and commodities, and weakly standardised, limiting cross-system comparison.

Quality, bias, and evaluation validity. Three under-discussed concerns cut across all three stages. First, *geographic and commodity bias*: existing corpora cluster around Chinese, Canadian, and Australian deposits and over-represent critical-mineral verticals (gold, tungsten, MVT Zn–Pb, fluorite–barite), while Latin American, African, and industrial-mineral contexts are sparsely covered (Fu et al., 2025a; Parsa et al., 2026; Zhao et al., 2025). Second, *annotation-bias propagation*: synthetic supervision generated through LLM teachers risks encoding teacher-model errors as gold labels, and few datasets undergo expert validation (Deng et al., 2024; Zhang et al., 2024a). Third, *evaluation validity*: F1, BLEU, and entity-level precision currently used for benchmarking do not correlate with field-level success criteria such as drill-outcome hit

rate (Ye et al., 2026; Parsa et al., 2026), leaving the most operationally meaningful evaluation signal unmeasured. These gaps directly motivate the closed-loop validation challenge in §7.

7 Challenges and Emerging Trends

KG Interoperability. KGs are now the dominant external memory in solid-Earth LLM systems, but most systems adopt incompatible deposit ontologies, mineral schemas, or geological classifications. MinMod (Knoblock et al., 2025) normalises 680,000 mineral sites under a 190-commodity ontology, while Dajishan (Zhao et al., 2025), LightRAG GeoGPT (Li et al., 2025), and Geo-MAG (Ma et al., 2026) construct project-specific KGs using LLM-driven or pipeline-based extraction without shared vocabularies. This lack of standardisation makes cross-corpus integration and downstream prospectivity modelling dependent on manual schema alignment, despite evidence from non-LLM gold-deposit community detection (Wang et al., 2026) that a unified KG could support strong performance without LLM inference. Therefore, interoperable solid-Earth ontologies—or alignment with existing schemas such as MinMod—are essential for scalable geological memory.

Closed-Loop Validation and Field-Outcome Benchmarks. As discussed in §5.3, none of the LLM-era solid-Earth systems reviewed close the loop from model output to a geological outcome, drill-hole intersection, assay confirmation, or production-scale validation. Current benchmarks measure proxy quantities: QA F1, BLEU, triple-level extraction precision, or holdout known-deposit recall. These proxies do not correlate with field-level hit rate (Ye et al., 2026; Parsa et al., 2026; Zhao et al., 2025), and their use as evaluation signals risks selecting systems that are persuasive to readers but operationally unhelpful. Closing this gap requires a community benchmark that tracks model recommendations through to drilling outcomes, potentially built by composing existing assets (OzROCK (Enkhsaikhan et al., 2021b), MinMod (Knoblock et al., 2025), geological-map QA corpora (Huang et al., 2025; Ma et al., 2026), and time-stamped exploration project records) into a forward-evaluable evaluation pipeline.

Physical Grounding. Solid-Earth phenomena obey conservation laws, geochronological constraints, and structural-geometric consistency rules that current LLM outputs routinely violate without detection. An ore-genesis chain may invoke retrograde alteration in a province where the relevant fluid pathway is unsupported by regional tectonics; a generated cross-section (Anikiev et al., 2026) may produce stratigraphically impossible layer sequences; an MPM score may rely on indicators that contradict the host deposit type. None of the output forms surveyed in §5.3 is checked against physics-based simulation tools, geomechanical solvers, or formal geochronological constraints. Integrating physics-aware verification as post hoc consistency checks, or as differentiable training objectives, remains an open methodological direction that current tool-augmented pipelines do not yet support.

Beyond Text: Modality Coverage Gaps. §4 identified four modalities relevant to solid Earth, of which only text is well handled by current LLM systems. Three major gaps remain. First, *quantitative geochemical and geophysical signals*, including assays, seismic waveforms, well-logs, and gravity–magnetic measurements, lack LLM-era generative-backbone treatment despite a mature pre-LLM lineage (Lawley et al., 2023; Yan et al., 2024, 2025). Second, *3D structural and subsurface volumes* remain thinly explored: recent seismic foundation models (Sheng et al., 2025; Dou et al., 2025; Gao et al., 2025) and 3D-VQA benchmarks (Tsujimoto et al., 2026; Wang et al., 2025a) develop separately from language-model pipelines, with no unified framework linking them to text-grounded LLMs. Third, *hyperspectral imagery* is currently absorbed only through RGB preprocessing (Yu et al., 2024), leaving spectral specificity unused. A unified multimodal pre-training target spanning text, maps, 3D structures, hyperspectral channels, and quantitative signals remains the key gap for next-generation solid-Earth foundation models.

Reinforcement Learning with Verifiable Rewards. RLVR-style training (Guo et al., 2025) remains largely unexplored in solid-Earth LLM systems because reward signals are scarce. The natural candidates are precisely those signals identified above as missing: drill-outcome validation, physics-consistency checks, or executable verification against tool outputs. The one feasible niche today is geospatial-code execution, where GeoFactory (Chen et al., 2025b) identifies RL as a key direction for improving agent decision-making over executable workflows. Broader adoption of RL in solid-Earth geoscience therefore depends on the

same infrastructure investments required for closed-loop validation and physics-aware verification.

Generalisable Training Pipeline and Model Selection. A stabilised pipeline for geo-LLM training is emerging (Table 1), typically involving CPT on geoscience corpora followed by supervised fine-tuning on in-domain instruction data for alignment. This recipe is shared by models such as JiuZhou (Chen et al., 2025a), GeoProspect (Wu et al., 2025), and K2 (Deng et al., 2024), and is applicable across decoder-only transformer architectures. However, model checkpoint selection is often ad hoc or based on proxy indicators such as benchmark performance, which may not reflect true domain adaptation potential.

Systematic studies on checkpoint adaptation ability remain limited and are needed to enable quantitative selection for CPT and alignment, though this is complicated by the lack of a canonical geoscience corpus. Moreover, geo-LLMs show substantial performance variance on shared benchmarks (e.g., NPEE and APTest (Deng et al., 2024; Chen et al., 2025a)) despite similar pipelines, suggesting a strong corpus-dependent effect.

Evaluation Methodology. Subjective evaluation is heavily utilised in scoring the performance of geo-LLM models. Fu et al. (2025a) exclusively uses domain experts to score their model, while Deng et al. (2024) uses subjective evaluation as part of their benchmark. Chen et al. (2025a) uses the same subjective evaluation benchmark as Deng et al. (2024) but instead uses GPT-4 as the evaluator. The subjective scores often do not align with objective benchmarks in ranking these models. This may point to specific features that are difficult to measure using objective benchmarks such as the ability for coherent reasoning from in-domain first principles (wherein the answer may be wrong but reasoning is valuable). In essence, the appearance of understanding of the domain is in question.

Limitations

Evaluation Focus: We primarily summarize reported capabilities, datasets, and system designs. Quantitative benchmarking or empirical comparisons between systems were beyond the scope, limiting direct performance assessment.

Responsible AI and Governance: While we highlight Responsible AI considerations, ethical, regulatory, and operational aspects in real-world deployment are complex and underexplored in this survey.

Language and Regional Coverage: Most reviewed systems and datasets are English-centric. Non-English geoscience literature, local reports, and regional workflows are underrepresented.

Acknowledgements

This research was supported by the Australian Research Council (ARC) Training Centre for Critical Resources for the Future (CCRF) under grant number IC230100035.

References

Denis Anikiev, Juan Esteban Mosquera, Korhan Ayranci, Judith Bott, and Umair Bin Waheed. 2026. Geosirr 1.0: Conversational geological cross-section modeling using large language models. *EGUsphere*. Preprint; discussion started 10 March 2026.

Andrea Baucon and Carlos Neto de Carvalho. 2024. Can ai get a degree in geoscience? performance analysis of a gpt-based artificial intelligence system trained for earth science (geologyoracle). *Geoheritage*, 16:121.

Yuxing Chen, Weijie Wang, Sylvain Lobry, and Camille Kurtz. 2024. An llm agent for automatic geospatial data analysis. *arXiv preprint arXiv:2410.18792*.

Zhou Chen, Ming Lin, Mingrun Zang, and 1 others. 2025a. Jiuzhou: Open foundation language models and effective pre-training framework for geoscience. *International Journal of Digital Earth*, 18(1):2449708.

Zhou Chen, Xiao Wang, Xinan Zhang, Ming Lin, Yuanhong Liao, Juanzi Li, and Yuqi Bai. 2025b. Geofactory: An llm performance enhancement framework for geoscience factual and inferential tasks. *Big Earth Data*, 9(4):714–746.

Rhys S. Davies, McLean Trott, Jaakko Georgi, and Alexander Farrar. 2025. Artificial intelligence and machine learning to enhance critical mineral deposit discovery. *Geosystems and Geoenvironment*, 4(2):100361.

Cheng Deng, Tianhang Zhang, Zhongmou He, and 1 others. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the Seventeenth ACM International Conference on Web Search and Data Mining*, pages 161–170.

Huseyin Denli, Hassan A. Chughtai, Brian Hughes, Robert Gistri, and Peng Xu. 2021. Geoscience language processing for exploration. In *Abu Dhabi International Petroleum Exhibition & Conference*. Society of Petroleum Engineers.

Yimin Dou, Xinming Wu, Nathan L. Bangs, Harpreet Singh Sethi, Jintao Li, Hang Gao, and Zhixiang Guo. 2025. Geological everything model 3d: A physics-informed promptable foundation model for unified and zero-shot subsurface understanding. *arXiv preprint arXiv:2507.00419*.

Jesper Sören Dramsch. 2020. 70 years of machine learning in geoscience in review. *Advances in Geophysics*, 61:1–55.

Majigsuren Enkhsaikhan, Eun Jung Holden, Paul Duuring, and Wei Liu. 2021a. Understanding ore-forming conditions using machine reading of text. *Ore Geology Reviews*, 135:104200.

Majigsuren Enkhsaikhan, Wei Liu, Eun-Jung Holden, and Paul Duuring. 2021b. Auto-labelling entities in low-resource text: A geological case study. *Knowledge and Information Systems*, 63:695–715.

Yu Fu, Mingguo Wang, Chengbin Wang, Shuaixian Dong, Jianguo Chen, Jiyuan Wang, Hongping Yu, Jing Huang, Liheng Chang, and Bo Wang. 2025a. Geominlm: A large language model in geology and mineral survey in yunnan province. *Ore Geology Reviews*, 182:106638.

Zekang Fu, Xiaojun Zheng, Yongfeng Yan, Xiaofei Xu, Fanchao Zhou, Xiao Li, Quantong Zhou, and Weikun Mai. 2025b. The evolution of machine learning in large-scale mineral prospectivity prediction: A decade of innovation (2016–2025). *Minerals*, 15(10):1042.

Hang Gao, Xinming Wu, Luming Liang, Hanlin Sheng, Xu Si, Hui Gao, and Yaxing Li. 2025. A foundation model enpowered by a multi-modal prompt engine for universal seismic geobody interpretation across surveys. *Information Fusion*, page 103437.

Daya Guo, Dejian Yang, Haowei Zhang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Shuyang Hou, Zhangxiao Shen, Anqi Zhao, and 1 others. 2024. Geocode-gpt: A large language model for geospatial code generation tasks. *arXiv preprint arXiv:2410.17031*.

Yangyu Huang, Tianyi Gao, Haoran Xu, Qihao Zhao, Yang Song, Zhipeng Gui, Tengchao Lv, Hao Chen, Lei Cui, Scarlett Li, and Furu Wei. 2025. Peace: Empowering geologic map holistic understanding with mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3899–3908.

Anuj Karpatne, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. 2019. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1544–1554.

Craig A. Knoblock, Binh Vu, Basel Shbita, Yao-Yi Chiang, Pothula Punith Krishna, Xiao Lin, Goran Muric, Jiyoon Pyo, Adriana Trejo-Sheu, and Meng Ye. 2025. Exploiting llms and semantic technologies to build a knowledge graph of historical mining data. In *The Semantic Web – ISWC 2025*, Lecture Notes in Computer Science, pages 451–471. Springer.

Christopher J. M. Lawley, Michael G. Gadd, Mohammad Parsa, Graham W. Lederer, Garth E. Graham, and Arianne Ford. 2023. Applications of natural language processing to geoscience text data and prospectivity modeling. *Natural Resources Research*, 32(4):1503–1527.

Christopher J. M. Lawley, Stefania Raimondo, Tianyi Chen, and 1 others. 2022. Geoscience language models and their intrinsic evaluation. *Applied Computing and Geosciences*, 14:100084.

Ke Li, Sheng-Bao Guo, Yan-Lei Wang, Yan-Wu Pan, and Xiao Cheng. 2025. Building a geological knowledge question-answering large language model with lightrag technology based on geogpt: A method study. *SSRN preprint*.

Zhouhan Lin, Cheng Deng, Le Zhou, and 1 others. 2024. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

Mengyi Liu, Xieyang Wang, Jianqiu Xu, Hua Lu, and Yongxin Tong. 2025. Nalspatial: A natural language interface for spatial databases. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):2056–2070.

Kai Ma, Qingzhong Zhan, Tongyun Zhang, Yuzheng Shi, Zhenxi Fang, Syed Yasir Ali Shah, Or Aimon Brou Koffi Kablan, and Baoyi Zhang. 2026. Geomag: A knowledge graph (kg)-enhanced multimodal retrieval-augmented generation (rag) framework for geological map understanding. *Transactions in GIS*, 30:e70226.

Kai Ma, Shuai Zheng, Miao Tian, Qinjun Qiu, Yanlin Tan, Xing Hu, Haoran Li, and Zhong Xie. 2023. Cngeoplm: Contextual knowledge selection and embedding with pretrained language representation model for the geoscience domain. *Earth Science Informatics*, 16:3629–3646.

Mohammad Parsa, Renato Cumani, Hossein Jodeiri Akbari Fam, and Bilal Tawbe. 2026. Large language models and geoscience transformers for predictive mapping of canadian critical minerals. *Natural Resources Research*, 35:207–225.

Mohammad Parsa, Christopher J. M. Lawley, Tarryn Cawood, Tania Martins, Renato Cumani, Steven E. Zhang, Aaron Thompson, Ernst Schetselaar, Steve Beyer, David R. Lentz, Jeff Harris, Hossein Jodeiri

Akbari Fam, and Alexandre Voinot. 2025. Pan-canadian predictive modeling of lithium–cesium–tantalum pegmatites with deep learning and natural language processing. *Natural Resources Research*, 34(2):639–668.

Jiyoon Pyo and Yao-Yi Chiang. 2024. Leveraging large language models for generating labeled mineral site record linkage data. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '24, pages 86–98. Association for Computing Machinery.

Qinjun Qiu, Kai Ma, Hairong Lv, Liufeng Tao, and Zhong Xie. 2023. Construction and application of a knowledge graph for iron deposits using text mining analytics and a deep learning algorithm. *Mathematical Geosciences*, 55(3):423–456.

Qinjun Qiu, Miao Tian, Liufeng Tao, Zhong Xie, and Kai Ma. 2024. Semantic information extraction and search of mineral exploration data using text mining and deep learning methods. *Ore Geology Reviews*, 165:105863.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. 2019. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Hanlin Sheng, Xinming Wu, Xu Si, Jintao Li, Sibo Zhang, and Xudong Duan. 2025. Seismic foundation model: A next generation deep-learning model in geophysics. *Geophysics*, 90(2):IM59–IM79.

Kang Sun, Yansi Chen, Guoshuai Geng, Zongyue Lu, Wei Zhang, Zhihong Song, Jiyun Guan, Yang Zhao, and Zhaonian Zhang. 2024. A review of mineral prospectivity mapping using deep learning. *Minerals*, 14(10):1021.

Mai Tsujimoto, Junjue Wang, Weihao Xuan, and Naoto Yokoya. 2026. Geo3dvqa: Evaluating vision-language models for 3d geospatial reasoning from aerial imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4994–5003.

Sandra Paula Villacorta Chambi, Mark Lindsay, Jens Klump, Klaus Gessner, Erin Gray, and Helen McFarlane. 2025. Assessing named entity recognition by using geoscience domain schemas: The case of mineral systems. *Frontiers in Earth Science*, 13:1530004.

Bo Wang, Mingguo Wang, Chengbin Wang, Jianguo Chen, Liheng Chang, and Nan Li. 2026. Gold deposit knowledge graph mining based on network analysis: Implications for the gold deposit understanding and classification. *Natural Resources Research*.

Jiayu Wang, Ruizhi Wang, Jie Song, and 1 others. 2025a. Rs3dbench: A comprehensive benchmark for 3d spatial perception in remote sensing. *arXiv preprint arXiv:2509.18897*.

Mingguo Wang, Chengbin Wang, Jianguo Chen, Bo Wang, Wei Wang, Xiaogang Ma, Jiangtao Ren, Zichen Li, Yicai Ye, Jiakai Zhang, and Yue Wang. 2025b. A lightweight knowledge graph-driven question answering system for field-based mineral resource survey. *Applied Computing and Geosciences*, 27:100268.

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

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):1–37.

Gang Wu, Haitao Wang, Kunyan Zhang, Shengguang Zhu, Jian Cui, Mengzhe Fan, Hailiang Wang, Hengliang Guo, Gubin Zhang, Dujuan Zhang, Haitao Wei, and Shan Zhao. 2025. Geoprospect: A domain-specific geological large language model with enhanced continual learning. *Neurocomputing*, 650:130801.

Qun Yan, Linfu Xue, Yongsheng Li, Rui Wang, Ke Ding, and Zhenglin Xu. 2025. Mineral prospectivity mapping using geological map semantic knowledge graph embedding: A case study of gold prospecting in ankang, shaanxi province, china. *International Journal of Digital Earth*, 18(1):2517827.

Qun Yan, Juan Zhao, Linfu Xue, Liqiong Wei, Mingjia Ji, Xiangjin Ran, and Junhao Dai. 2024. Mineral prospectivity mapping based on spatial feature classification with geological map knowledge graph embedding: Case study of gold ore prediction at wulonggou, qinghai province (western china). *Natural Resources Research*, 33(6):2385–2406.

Fanfan Yang, Renguang Zuo, and Oliver P. Kreuzer. 2024. Artificial intelligence for mineral exploration: A review and perspectives on future directions from data science. *Earth-Science Reviews*, 258:104941.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Meng Ye, Xiao Lin, Georgina Lukoczki, Graham W. Lederer, and Yi Yao. 2026. Queryplot: Generating geological evidence layers using natural language queries for mineral exploration. *arXiv preprint arXiv:2602.17784*. Manuscript submitted to ACM.

Beibei Yu, Tao Shen, Hongbin Na, Ling Chen, and Denqi Li. 2024. Mineagent: Towards remote-sensing mineral exploration with multimodal large language models. *arXiv preprint arXiv:2412.17339*.

Hao Zhang and Jin-Jian Xu. 2023. When geoscience meets foundation models: Towards general geoscience artificial intelligence system. *arXiv preprint arXiv:2309.06799*.

Jiyin Zhang, Cory Clairmont, Xiang Que, Wenjia Li, Weilin Chen, Chenhao Li, and Xiaogang Ma. 2025a. Streamlining geoscience data analysis with an llm-driven workflow. *Applied Computing and Geosciences*, 25:100218.

Yifan Zhang, Zhiyun Wang, Zhengting He, and 1 others. 2024a. Bb-geogpt: A framework for learning a large language model for geographic information science. *Information Processing & Management*, 61(5):103808.

Yifan Zhang, Cheng Wei, Zhengting He, and Wenhao Yu. 2024b. Geogpt: An assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131:103976.

Yiran Zhang, Jincheng Hu, Mark Dras, and Usman Naseem. 2025b. Cogmem: A cognitive memory architecture for sustained multi-turn reasoning in large language models. *arXiv preprint arXiv:2512.14118*.

Yiran Zhang, Mingyang Lin, Mark Dras, and Usman Naseem. 2026a. Beyond the black box: Demystifying multi-turn llm reasoning with vista. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 41745–41747.

Yiran Zhang, Jinwen Liu, Daniel Su, Yisu Chen, Qiang Sun, Chris Gonzalez, Eun-Jung Holden, Marco Fiorentini, Wei Liu, and Yihao Ding. 2026b. Minetrace: An evidence-grounded interactive reasoning system for mineral prospectivity. *arXiv preprint arXiv:2609.02060*.

Yu Zhao, Kun Wang, Qishen Chen, Yanfei Zhang, Qing Guan, Jiayun Xing, Xin Ren, Chenghong Shang, and Hang He. 2025. Llm-driven rapid construction of knowledge graph for mineral resources: A case study of the dajishan hydrothermal tungsten deposit. *Frontiers in Earth Science*, 13:1738096.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36.

A Detailed Tables for Geoscience LLM Data Resources

Abbrev	Full Task Name	Explanation
GQA	Geoscience Question Answering	LLMs answering domain-specific geoscience questions, including geological reasoning and knowledge retrieval.
GDA	Geoscience Data Analysis	LLM-assisted exploration and analysis of geoscience datasets via natural language or interactive interfaces.
GWA	Geospatial Workflow Automation	LLMs generating GIS workflows, geospatial code, or automating spatial analysis pipelines.
RS	Remote Sensing & Earth Observation	AI models applied to satellite imagery, mineral exploration, or Earth observation tasks.
DSF	Decision Support & Forecasting	AI models supporting prediction or decision-making in Earth systems such as disasters, weather, or environmental forecasting.
KCE	Knowledge Communication & Education	LLM systems for geoscience education, science communication, or geotourism content generation.

Table 2: Overview of task categories used in geoscience MLLM studies.

Paper	Dataset	Data size	Source	Purpose
K2	Geoscience papers	1,122,094 docs	GAKG	CPT; domain knowledge injection
	Geoscience paper metadata	4,274,716 docs	DDE Scholar	Metadata-enriched corpus
	Earth science Wikipedia	767,341 docs	Wikipedia (Earth science)	General geoscience grounding
BB-GeoGPT	GIS papers + GIS-related Wikipedia	2,499 + 24,408 docs	GIS literature; Wikipedia	Domain knowledge injection
GeoGalactica	GeoCorpus	52.7B tokens	LetPub, Acemap	Large-scale geoscience CPT
	ArXiv preprints	13.7B tokens	ArXiv	Scientific reasoning coverage
	Code data	12.4B tokens	Not reported	Programming & structured reasoning
GeoCode-GPT	GeoCode-PT: code snippets + operator examples	275,374 snippets; 10,190 operators	GitHub; GEE/PIE; ArcGIS, MATLAB, R, Python	Geospatial code pre-training
GeoProspect	Geological corpus: papers, books, dictionaries	181.66M tokens	ChinaXiv, CNKI, public geological repositories	Domain knowledge injection
CnGeoPLM	GeoCorpus-CN: reports, abstracts, dissertations	∼232 MB	National Geological Archives of China; journals; dissertations	Contextual embedding (Chinese)
JiuZhou	JiuZhou-Corpus: general + geoscience corpus	6.95B tokens	Encyclopedias, books, QA forums, Wikipedia	Geoscience & general knowledge injection
Geoscience GloVe/BERT	Public geoscience publication corpus	∼45,000 docs	NRCan GEOSCAN; provincial surveys; DOAJ	Geochemical & mineral representation
GeoBERT	Internal geoscientific records	20M records	ExxonMobil internal database	Oil & gas industry language representation

Table 3: Summary of reported pre-training or domain-retraining data resources in geoscience language-model research. **Abbr.**: GAKG = Geoscience Academic Knowledge Graph; DDE = Deep-time Digital Earth; CNKI = China National Knowledge Infrastructure; NRCan = Natural Resources Canada; DOAJ = Directory of Open Access Journals; CPT = continued pre-training. Data sizes follow the units reported in the original papers.

Paper	Dataset	Data size	Source	Purpose
K2	Alpaca-GPT4	52,002 pairs	Stanford Alpaca	General instruction following
	Dolly-15K	15,011 pairs	Databricks	Human-written instruction data
	Natural Instructions	2,446 pairs	AllenAI	Objective task coverage
	AI2 Reasoning Challenge	7,787 pairs	AI2	Science reasoning QA
BB-GeoGPT	GPT4-Alpaca / instruction alignment data	Not reported	Not reported	Human-intent alignment
GeoGalactica	Alpaca	52,000 pairs	Stanford Alpaca	Human-intent alignment
GeoCode-GPT	Alpaca-GPT-4 multilingual subset	364,000 pairs	HuggingFace; Alpaca-GPT-4	Alignment; catastrophic forgetting prevention
GeoProspect	Alpaca-GPT4-zh + Belle	Not reported	HuggingFace open-source sets	Chinese conversational alignment
JiuZhou	Alpaca-GPT4 + Super-NaturalInstructions	Not reported	HuggingFace; AllenAI	Human alignment; instruction following

Table 4: Summary of reported instruction-tuning and alignment data resources in geoscience LLM research. **Abbr.**: IFT = instruction fine-tuning. "Not reported" indicates that the data size or source is not explicitly specified in the reviewed paper. Most listed resources are general-purpose instruction-following datasets rather than expert-authored solid-Earth instruction corpora.

Paper / Resource	Dataset / Resource	Source	Data size	Purpose
Domain LLM SFT and QA resources				
K2 (Deng et al., 2024)	GeoSignal	GAKG, DDE Scholar, DataExpo, GSO, geoscience QA, self-instruct data	Not reported	Domain SFT for geoscience QA, reasoning, fact verification, summarisation, classification, and NER
	GeoTool	Manual GeoSearch examples and ToolBench-style tool-use data	Not reported	Tool invocation, decision-making, execution planning, and parameter extension
BB-GeoGPT (Zhang et al., 2024a)	BB-GeoSFT and auxiliary geospatial NLP/chat resources	LLaMA-2-derived QA data, GROBID, Broad Twitter Corpus, LNEX, SemEval-2015 Task 8, UltraChat	35,876 instructions	GIS knowledge QA, word semantics, summarisation, toponym extraction, spatial concept identification, and GIS-topic adaptation
GeoGalactica (Lin et al.,	Entity, citation, and academic-graph data	Wikipedia, Deep Literature, DataExpo	7.4M + 45M records	NER, reference resolution, citation understanding, and scientific summarisation
	Ontology, dictionary, figure, and table resources	GSO, GAKG	360K + 24M records	Terminology explanation, semantic alignment, figure-to-text, and table-to-text generation
	Mineral and resource QA subsets	RRUFF mineral database; USGS MRData	84,069 + 148,212 records	Mineral identification QA and mineral-resource distribution QA
	Broader geoscience factual QA subsets	Dinosaur, fossil, earthquake, and hazard datasets	11,348 + 4,959 + 37,284 + 37,823 records	Broader geoscience factual QA used in GeoGalactica task coverage
GeoCode-GPT (Hou et al., 2024)	GeoCode-SFT: code summarisation, code generation, and toolkit knowledge	GeoCode-GPT project resources	502,047 pairs	Spatial code generation, geospatial programming, and toolkit-aware code reasoning
GeoProspect (Wu et al., 2025)	Professional geology instruction dataset	Geological textbooks, Chinese geological exams, and professional geological reports	Not reported	Geological-domain instruction tuning and mineral prospecting QA
GeoMinLM (Fu et al., 2025a)	Hybrid geological corpus and KG-based QA corpus	Yunnan Gold & Mining Group reports, textbooks, monographs, and domain KG resources	∼5.16M words	Geological and mineral-survey QA, exploration-domain reasoning, and KG-supported inference
JiuZhou (Chen et al., 2025a)	GeoSignal geoscience QA and structural pairs	GAKG, DDE Scholar, geoscience QA platforms, and structured geoscience resources	39,749 pairs	Task specialisation for GeoBench-style geoscience QA and structural reasoning
KG, extraction, and mineral-resource QA resources				
MinMod (Knoblock et al., 2025)	Historical mining-data knowledge graph	NI 43-101 reports, USGS resources, mining databases, and tabular mineral records	∼680,000 mineral sites; 190 commodities	Mineral-site KG construction, mineral-resource integration, record harmonisation, and downstream prospectivity analysis support
Pyo & Chiang (Pyo and Chiang, 2024)	LLM-generated labelled mineral-site record-linkage data	Mineral-site records with spatial and non-spatial tabular attributes	Not reported	Training-data generation and PLM fine-tuning for mineral-site record linkage
Wang et al. (Wang et al., 2025b)	Field-based mineral-resource survey KGQA resource	Mineral-resource survey data, KG techniques, question templates, and matching algorithms	Not reported	Lightweight question answering for field-based geologists and mineral-resource survey workflows
Dajishan KG (Zhao et al., 2025)	Dajishan hydrothermal tungsten-deposit KG	Mineral-resource knowledge base and unstructured geological texts for Dajishan tungsten mine	Not reported	LLM-driven mineral-resource KG construction and prospecting-indicator analysis
GeoWorkflow-LLM (Zhang et al., 2025a)	Mindat API-call and workflow resources	Mindat open data API and mineralogical database	Not reported	Tool-grounded workflow automation, Mindat-based mineral-data querying, and reproducible geoscience data-analysis execution
Multimodal map, hyperspectral, and spatial task resources				
PEACE / GeoMap-Agent (Huang et al., 2025)	GeoMap-Bench	Geologic maps with multimodal multiple-choice annotations	124 maps; 3,864 questions	Geologic-map extraction, referring, grounding, reasoning, and analysis with MLLMs
Geo-MAG (Ma et al., 2026)	Geological-map KG and multimodal RAG resources	Geological reports and geological maps	Not reported	KG-enhanced multimodal RAG for geological-map interpretation and semantic reasoning
MineAgent (Yu et al., 2024)	MineBench	Geological and hyperspectral remote-sensing data	Not reported	Multi-image reasoning, spatial–spectral integration, and mineral-exploration MLLM evaluation
QueryPlot (Ye et al., 2026)	Deposit-model descriptions and SGMC structured polygon texts	Descriptive deposit models and State Geologic Map Compilation polygons	>120 deposit types; SGMC polygons	Natural-language query translation, evidence-layer generation, and mineral-prospectivity mapping support
Pre-LLM and non-generative IE / schema resources				
CnGeoPLM (Ma et al., 2023)	Task-specific Chinese geological annotations	Manually annotated Chinese geological reports and GeoCorpus-CN-derived resources	Small-scale	Geological NER, mineral-domain text classification, and encoder-based downstream evaluation
OzROCK (Enkhsaikhan et al., 2021b)	BIO-annotated geological entity dataset	Mineral-exploration reports with distant supervision and iterative auto-labelling	Not reported	Low-resource geological NER across six geological entity categories
Ore-forming MR (Enkhsaikhan et al., 2021a)	Ore-forming condition extraction resource	Under-utilised mineral-exploration reports	Not reported	Geological term labelling, entity typing, relation extraction, and ore-forming condition KG construction
IronKG (Qiu et al., 2023)	Iron-deposit KG construction resource	Exploration reports and text-mined ore-forming circumstance descriptions	Not reported	Iron-deposit KG construction, entity–relation extraction, and data-driven mineral discovery support
Mineral semantic IE (Qiu et al., 2024)	Mineral-exploration semantic extraction and search resource	Mineral-exploration reports	Not reported	Keyword/keyphrase extraction, topic graph construction, geological entity recognition, relation extraction, and semantic search
SMS NER schema (Villacorta Chambi et al., 2025)	Schema for Mineral Systems	Domain-characterised mineral-system texts and expert-designed entity taxonomy	9 geological + 5 general entity classes	Mineral-system NER schema design, annotation consistency, and evaluation validity

Table 5: Summary of supervised fine-tuning and task-level data resources used in solid-Earth geoscience LLM, MLLM, KG, and language-interface research. The table includes resources for domain SFT, QA, tool use, KG construction, record linkage, geologic-map understanding, hyperspectral mineral exploration, evidence-layer generation, and pre-LLM information extraction. **Abbr.**: SFT = supervised fine-tuning; KG = knowledge graph; QA = question answering; IE = information extraction; NER = named entity recognition; MLLM = multimodal large language model; GAKG = Geoscience Academic Knowledge Graph; DDE = Deep-time Digital Earth; GSO = GeoScience Ontology; RRUFF = RRUFF mineral database; USGS = United States Geological Survey; SGMC = State Geologic Map Compilation. "Not reported" indicates that the data size is not explicitly specified in the reviewed paper or public project page.

Survey / Review	Year	LLM/ MLLM	Solid Earth	Mineral/ Resource	Multi modal	Agent/ Tool	MRA	Data Bias	Field Valid.
Other surveys									
Zhu et al.	2017	✗	✗	✗	✓	✗	✗	**Partly**	✗
Karpatne et al.	2019	✗	**Partly**	✗	**Partly**	✗	✗	**Partly**	✗
Reichstein et al.	2019	✗	✗	✗	**Partly**	✗	✗	**Partly**	✗
Dramsch	2020	✗	✓	✗	**Partly**	✗	✗	**Partly**	✗
Willard et al.	2022	✗	✗	✗	**Partly**	✗	✗	✗	✗
Lawley et al.	2022	✗	✓	✓	✗	✗	✗	**Partly**	✗
Zhang and Xu	2023	✓	**Partly**	✗	✓	**Partly**	✗	**Partly**	✗
Yang et al.	2024	✗	✓	✓	**Partly**	✗	✗	**Partly**	**Partly**
Sun et al.	2024	✗	✓	✓	**Partly**	✗	✗	**Partly**	**Partly**
Hadid et al.	2024	✓	**Partly**	✗	**Partly**	**Partly**	✗	**Partly**	✗
Fu et al.	2025	✗	✓	✓	**Partly**	✗	✗	**Partly**	**Partly**
Davies et al.	2025	✗	✓	✓	**Partly**	✗	✗	**Partly**	**Partly**
This survey									
Ours	**2026**	✓	✓	✓	✓		✓	✓	✓

Table 6: Coverage comparison between this survey and other reviews on geoscience, mineral exploration, foundation models, and generative AI. LLM/MLLM = large language model / multimodal large language model; MRA = Memory–Reason–Act; Data Bias = dataset bias and representativeness; Field Valid. = field or drilling validation; ✓, ✗, and **Partly** denote full, absent, and partial coverage.