Agentic AI for subsurface intelligence: automated SEG-Y to MDIO migration at scale

Long Chen[1], Ben Lasscock[2], Chun-Hao Liu[1], Suren Gunturu[1], Sujitha Martin[1], Brian Michell[2], Arunabha Datta[1], Vidya Sagar Ravipati[1]*

[1]*AWS Generative AI Innovation Center* [2]*TGS ASA*

SUMMARY

Energy companies hold millions of legacy seismic files in SEG-Y format with inconsistent, fragmented metadata that prevents automated processing and AI integration. We present a multi-agent AI system that automates end-to-end metadata reconstruction for large-scale SEG-Y migration to MDIO v1.0, a modern self-describing seismic format. Our system addresses three coupled challenges: (1) seismic product type classification across a canonical taxonomy spanning 2D/3D, pre-stack/post-stack, and marine/land/OBN acquisition modes, (2) header field extraction from free-form text and binary headers, and (3) schema mapping to standardized MDIO coordinates and dimensions. Built on Claude Sonnet 4 with no fine-tuning, our agent-based pipeline achieves 95% accuracy in template classification, 99.17% in field extraction, 98.16% in schema mapping, and 96.83% end-to-end accuracy, surpassing human baseline performance while processing files at 90 seconds per file. Evaluated on expert-labeled samples spanning diverse acquisition types and header conventions, the system demonstrates production-ready performance for migrating TGS's archive of 1.4+ million SEG-Y files. This approach transforms a previously manual, expert-dependent process into a scalable, automated workflow, enabling AI-ready seismic data management at cloud scale.

INTRODUCTION

Energy industry organizations and data providers hold petabytes of seismic data stored predominantly in the SEG-Y format, yet much of this information remains difficult to locate, integrate, and use operationally. The challenge is rarely a lack of data; it is the absence of consistent, machine-readable structure across legacy formats and fragmented metadata sources. In SEG-Y, key concepts such as dimensionality (2D vs. 3D, post-stack vs. gathers), coordinate scalars, and navigation are derived from trace-order and header conventions that vary by project and vendor. Important survey metadata, e.g., geometry, coordinate systems, and processing history, is often embedded in loosely structured textual headers that lack consistent formatting and are not machine-readable. This *metadata gap* creates a significant barrier to automation, large-scale analytics, and the application of AI across entire data portfolios.

At the scale of over 1 million SEG-Y files, manual interpretation of text headers by subject-matter experts (SMEs) is infeasible. The same geophysical concept may be referred to using multiple different aliases (e.g., CDP, CMP, CMP NO.), and custom header overrides (such as non-standard byte locations for inline and crossline coordinates) must be extracted and applied to correctly interpret trace headers. An automated solution guided by domain expertise is required.

This work presents a multi-agent AI system that automates end-to-end metadata reconstruction for large-scale SEG-Y migration into MDIO v1.0 (Sansal et al., 2023a,b; Michell et al., 2025), a modern, open-source, self-describing seismic data format. MDIO replaces the implicit, convention-dependent structure of SEG-Y with explicit dimensions, coordinates, and rich attribute metadata, providing a robust foundation for cloud-native seismic workflows and AI integration. Our agentic solution addresses three coupled tasks: (1) seismic product type classification, (2) header field extraction, and (3) schema mapping to standardized MDIO fields. The system leverages large language models (LLMs) within a structured pipeline that combines free-text reasoning with deterministic domain constraints, achieving production-ready accuracy without model fine-tuning.

METHOD

MDIO v1.0: Target Format and Product Taxonomy

MDIO v1.0 offers a self-describing representation of seismic data with explicit structural metadata rather than inferred conventions. Each dataset specifies its main dimensions (e.g., `cdp`, `angle`, `inline`, `crossline`), along with associated coordinate variables (e.g., `cdp_x`, `cdp_y`, `coordinate_scalar`). The dataset framework is defined as a JSON schema detailing coordinates, dimensions, masks, and key survey metadata as separate arrays, enabling clear interpretation of survey geometry and navigation. MDIO is released as open-source software under the Apache 2.0 license (Sansal and contributors, 2025a), along with cloud-compatible SEG-Y parsing tools (Sansal and contributors, 2025b), and is natively compatible with the Xarray Python library (Hoyer and Hamman, 2017).

To support the diversity of seismic products encountered across acquisition, processing, and migration stages, MDIO v1.0 defines a set of canonical *product templates*. Each template establishes the dataset's dimensional structure and required coordinate variables, providing a standard representation for common seismic products. Table 1 summarizes the templates currently used in operational data management, spanning post-stack sections and volumes, CDP offset and angle gathers, streamer field records and shot gathers, and common-offset/common-azimuth (CoCa) gathers in 2D and 3D configurations.

The canonical seismic product taxonomy used for classification spans multiple axes: survey type (2D, 3D), processing domain (pre-stack, post-stack), gather/product type (seismic section, CDP gathers, shot gathers, receiver gathers, OVT gathers), migration stage (pre-migration, post-migration), and regularization status. It encompasses marine, land, and ocean-bottom node (OBN) acquisition modes (see Table 2).

Multi-Agent AI Pipeline for Metadata Reconstruction

Table 1: Summary of MDIO v1.0 seismic product templates. Each template defines the core dataset dimensions and coordinate variables that structure the seismic data.

Template	Grid Dimensions	Chunk Sizes	Coordinates
PostStack2D	cdp	1024×1024	cdp x/y
PostStack3D	inline, crossline	128×128×128	cdp x/y
CdpOffsetGathers2D	cdp, offset	16×64×1024	cdp x/y
CdpOffsetGathers3D	inline, crossline, offset	8×8×32×512	cdp x/y
CdpAngleGathers2D	cdp, angle	16×64×1024	cdp x/y
CdpAngleGathers3D	inline, crossline, angle	8×8×32×512	cdp x/y
CoCaGathers3D	inline, crossline, offset, azimuth	8×8×32×1×1024	cdp x/y
StreamerFieldRecords3D	sail_line, gun, shot_index, cable, channel	1×1×16×1×32×1024	source x/y, group x/y, shot_point, ffid
StreamerShotGathers2D	shot_point, channel	16×32×2048	source x/y, group x/y
StreamerShotGathers3D	shot_point, cable, channel	8×1×128×2048	source x/y, group x/y, gun

Table 2: Representative entries from the canonical seismic product taxonomy used in MDIO v1.0, classifying products by survey type, processing domain, gather type, migration stage, and regularization status.

Survey	Domain	Gather/Product Type	Migration Stage	Regularization
2D	Post-stack	Seismic section	—	—
2D	Pre-stack	CDP gathers (offset)	Pre-migration	—
3D	Post-stack	Seismic volume	—	—
3D	Pre-stack	CDP gathers	Post-migration	Regularized
3D	Pre-stack	Shot gathers	—	—
OBN	Pre-stack	Receiver gathers	—	Non-regularized
Land	Pre-stack	Shot gathers	—	—
Land	Pre-stack	OVT gathers	Post-migration	—

We aim to ingest a library of more than 1.4 million SEG-Y files into the MDIO v1.0 format. This requires solving three coupled metadata reconstruction tasks: (1) seismic product type identification, (2) header field extraction, and (3) schema mapping to standardized MDIO v1.0 fields. Together, these steps define the minimum requirements for constructing MDIO v1.0 datasets with explicit dimensions, coordinates, and consistent metadata.

To address these challenges, we constructed a labeled training and evaluation dataset derived from 57 manually interpreted SEG-Y files, each containing full text headers, binary header information, and SME-validated ground-truth annotations. Each record included canonical seismic product type labels, standardized header overrides, and unit definitions, providing authoritative reference data for product classification, field extraction, and schema mapping. Although modest in size, this dataset was intentionally curated to span heterogeneous acquisition types, processing workflows, and header conventions, encompassing approximately 800 distinct metadata labels representative of the broader seismic data library.

We developed a multi-agent AI system to perform end-to-end metadata reconstruction. The system leverages LLMs within a structured pipeline that combines free-text reasoning with explicit domain constraints derived from SME guidance. All agents were built using the Strands Agents SDK (AWS, 2025) and deployed on Amazon Bedrock with the Anthropic Claude Sonnet 4 LLM. Each agent follows a ReAct-style reasoning loop (Yao et al., 2023), iteratively observing the input context, reasoning over domain constraints and schema specifications, and selecting actions to the next stage. This design enables dynamic, multi-step reasoning rather than single-pass inference. No model fine-tuning was performed; performance gains were achieved through agent decomposition, prompt refinement, rule encoding, and SME-in-the-loop iteration. The automated pipeline, illustrated in Figure 1, consists of three stages, augmented with verification agents:

1. **Template Classification Agent.** This agent determines the canonical seismic product type consistent with the MDIO v1.0 taxonomy. It evaluates structural cues (e.g., presence of `LINE` versus `INLINE`/`XLINE`), acquisition indicators (e.g., `SOURCE` descriptors such as `VIB`/`DYNAMITE`), and processing descriptions (e.g., migration and regularization statements). Outputs are validated using rule-based consistency checks. A dedicated verification agent cross-checks the predicted template class against the available header evidence.

2. **Field Extraction Agent.** This agent identifies relevant metadata fields from SEG-Y text and binary headers, including field names, byte locations, data types, and semantic intent. It recognizes diverse naming conventions and non-standard abbreviations. Extracted fields are normalized into an intermediate representation independent of the original SEG-Y syntax.

3. **Schema Mapping Agent.** This agent maps extracted fields to standardized MDIO v1.0 metadata namespaces, resolving SEG-Y aliases (e.g., mapping `CDP`, `CMP`, `CMP NO.` to a canonical `cdp` coordinate) and assigning fields to their appropriate roles (dimensions, coordinate variables, or auxiliary metadata). Fields that cannot be mapped unambiguously are explicitly flagged. A verification agent confirms that each mapped alias corresponds to the same geophysical concept as the extracted field. Box 1 illustrates an example of the multi-step reasoning process from the mapping agent, extracted from the Strands Agents SDK.

Box 1: Example for Schema Mapping Agent Reasoning

Input: `{name: 'CMP', byte_location: 21, data_type: 'int32'}`

Agent Reasoning:
Looking at the provided field information, I need to find the corresponding standard field in the schema and apply the mapping rules.

First, let me check the schema for fields with data type `int32` at byte location 21. Byte location 21 corresponds to the `ensemble_num` field which has type `int32`.

Next, I'll apply the rules. Rule 3 specifically states: [`CDP`, `CDP NUMBER`, `CDP Number`, `CMP`] should map to `ensemble_num`. The field name `CMP` is explicitly listed in Rule 3.

This makes sense because in SEG-Y data, CMP (Common Mid Point) is essentially the same concept as CDP (Common Depth Point) or ensemble number - they all refer to the gather of traces sharing the same midpoint location. The data type (`int32`) and the conceptual meaning align perfectly.

Output: `{"mapped_field_name": "ensemble_num"}`

Each stage is followed by automated verification enforcing internal consistency, domain constraints, and schema compatibility. Ambiguous cases and previously unseen patterns are routed through an SME review loop, enabling controlled refinement of prompts, rules, and schema definitions. Table 3 illustrates the mapping of representative SEG-Y header cues to their corresponding MDIO v1.0 interpretations. A notable design finding was that relying on the LLM's pre-trained knowledge of SEG-Y field semantics for schema mapping, rather than providing the full schema descriptions and byte locations, yielded substantially higher accuracy. Providing the complete schema caused the mapping agent to over-rely on byte-location matching, which frequently led to incorrect mappings when non-standard byte locations were used. Besides, enabling verbose reasoning mode in the verification agents (requiring explicit chain-of-thought justification before each decision) improved field mapping accuracy.

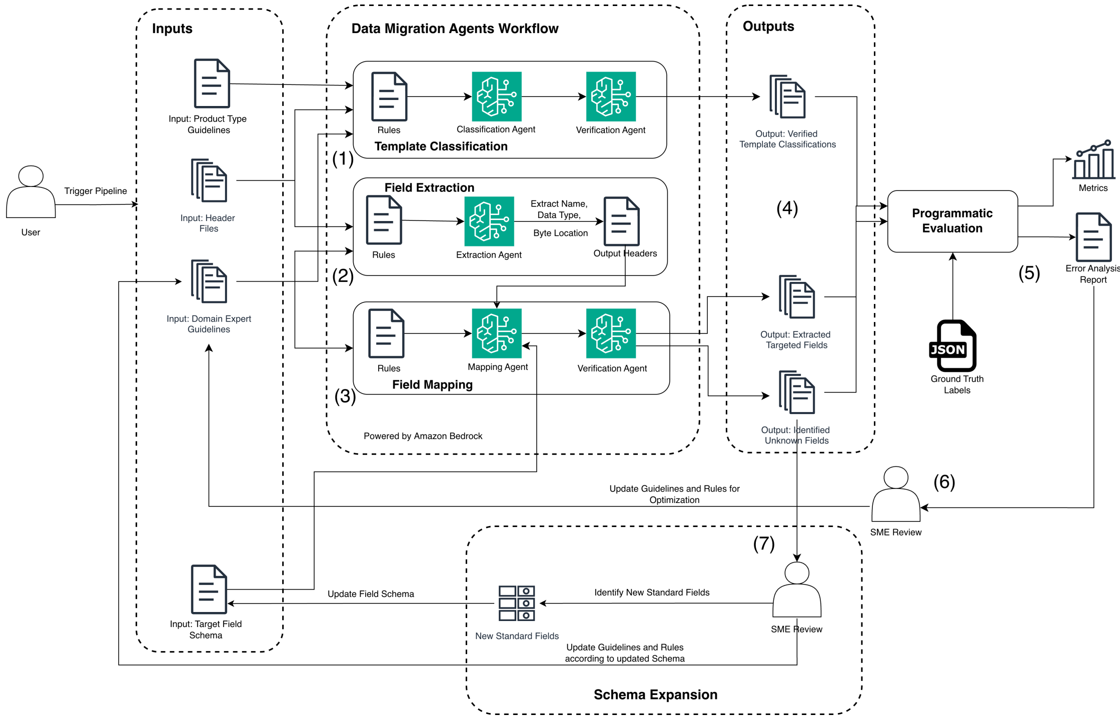


Figure 1: Multi-stage agentic pipeline for reconstructing structured MDIO v1.0 metadata from unstructured SEG-Y headers. The process combines LLM-based semantic parsing with deterministic validation and SME-in-the-loop refinement.

Table 3: Mapping of representative SEG-Y header cues to their corresponding MDIO v1.0 interpretations.

SEG-Y Header Cue	**MDIO Interpretation**
`cdp`, `angle`, `inline`, `crossline`	MDIO dimensions
`cdp_x`, `cdp_y`, `coordinate_scalar`	MDIO coordinate fields
`LINE` = single value	2D survey
`INLINE` or `XLINE` present	3D survey
`SOURCE` = `VIB/DYNAMITE`	Land acquisition
"PSDM", "DEPTH MIGRATION"	Post-migration
"CMP gathers"	Pre-migration
"regularised", "binned", "resampled"	Regularization flag

RESULTS

Evaluation Setup

The system was evaluated end-to-end on the labeled dataset of 57 expert-annotated SEG-Y files across three primary tasks: template classification, field extraction, and field schema mapping. Performance was assessed using exact-match accuracy against SME-validated ground truth. In addition, we report an end-to-end field extraction and mapping metric, in which a field is considered correct if and only if it is both successfully extracted *and* correctly mapped to the MDIO v1.0 schema.

Quantitative Results

Table 4 summarizes the performance of the agentic metadata reconstruction pipeline. The template classification agent achieved 95.00% accuracy in identifying the correct canonical seismic product type from the MDIO v1.0 taxonomy. The field extraction agent achieved 99.17% accuracy with precision and recall of 99.33% and 99.50%, respectively, demonstrating robust extraction across diverse header conventions. The schema mapping agent achieved 98.16% accuracy in mapping extracted fields to their correct MDIO v1.0 canonical names. The end-to-end metric, requiring both correct extraction and correct mapping, reached 96.83% accuracy, with precision of 96.99% and recall of 97.16%.

Table 5 provides a per-stage performance breakdown compared against an independent human baseline from junior SMEs. The AI solution matches human accuracy on template classification (95%) and substantially outperforms it on field extraction, schema mapping, and end-to-end accuracy (96.83% vs. 81.61%), with structured verification mitigating the error compounding observed in manual workflows.

The final performance was achieved through an iterative optimization process. Template classification accuracy improved from 77% (baseline with raw LLM inference) to 95% through progressive incorporation of SME-authored diagnostic rules, LLM-assisted rule discovery, and the addition of a verification agent. Field mapping accuracy improved from 79.43% to 98.16% through optimization rounds that included the introduction of a dedicated verification agent, the finding that relying on the LLM's pre-trained knowledge of SEG-Y field definitions outperformed using the provided schema descriptions, enabling verbose reasoning mode for improved self-consistency, and iterative rule refinement from SME feedback.

Operational Performance and Scalability

Each SEG-Y file is processed in approximately 90 seconds. At this rate, the full conversion of TGS's archive of 1.4+ million SEG-Y files is operationally feasible. The pipeline operates

Table 4: Performance of the agentic AI solution, evaluated against SME-validated ground truths.

Task	Accuracy	Precision	Recall
Template classification	95.00%	N/A	N/A
Field extraction	99.17%	99.33%	99.50%
Field schema mapping	98.16%	98.16%	98.16%
End-to-end (extraction + mapping)	96.83%	96.99%	97.16%

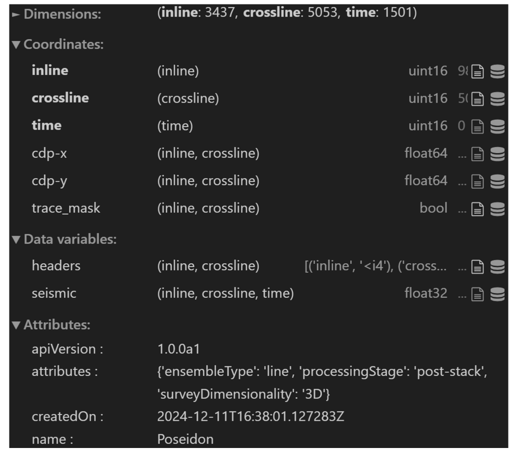
Table 5: Detailed performance comparison between human baseline and the agentic AI solution.

Task	Metric	Human	Solution
Template Classification	Accuracy	95%	95%
	Latency per sample (s)	—	15.2
	Input/output tokens per sample	—	6243/572
Field Extraction	Accuracy – micro	92.98%	99.17%
	Precision – micro	92.98%	99.33%
	Recall – micro	92.98%	99.50%
	Latency per sample (s)	—	8.2
	Input/output tokens per sample	—	1866/528
Field Mapping (GT extractions as input)	Accuracy (All) – micro	88.29%	98.16%
	Precision (All) – micro	88.29%	98.16%
	Recall (All) – micro	88.29%	98.16%
	Precision (Target Fields) – micro	84.04%	97.17%
	Recall (Target Fields) – micro	89.71%	99.36%
	Precision (Unknown Fields) – micro	93.61%	99.29%
	Recall (Unknown Fields) – micro	86.76%	96.86%
	Latency per sample (s)	—	59.84
	Input/output tokens per sample	—	44803/1620
End-to-End (Extraction and Mapping)	Accuracy (All) – micro	81.61%	96.83%
	Precision (All) – micro	81.61%	96.99%
	Recall (All) – micro	81.61%	97.16%
	Precision (Target Fields) – micro	75.30%	95.64%
	Recall (Target Fields) – micro	80.13%	98.40%
	Precision (Unknown Fields) – micro	89.47%	98.56%
	Recall (Unknown Fields) – micro	83.22%	95.80%
	Latency per sample (s)	—	68.04
	Input/output tokens per sample	—	46669/2148

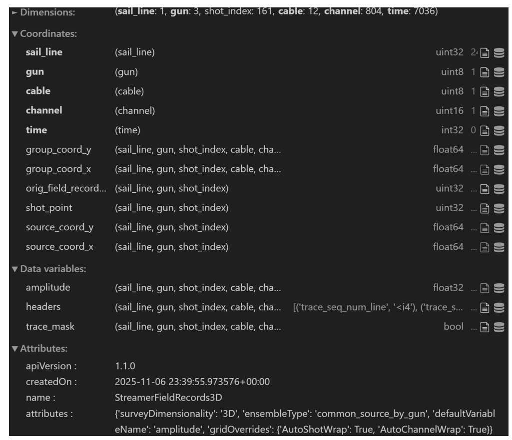
without fine-tuning or retraining; all performance gains derive from agent decomposition, prompt engineering, rule encoding, and iterative SME feedback.

Illustrative Example

Figure 2 shows examples of MDIO v1.0 dataset schemas produced by the pipeline. The top panel displays a 3D post-stack dataset with dimensions (`inline`, `crossline`, `time`), explicit coordinate variables (`cdp_x`, `cdp_y`), and standardized attributes including the processing stage and survey dimensionality. The bottom panel shows a 3D pre-stack streamer field dataset with dimensions (`sail_line`, `gun`, `shot_index`, `cable`, `channel`, `time`), source and receiver coordinates, and pre-stack attributes. In both cases, all metadata that was previously implicit in the SEG-Y text header is now explicit, machine-readable, and conformant with the MDIO v1.0 schema, enabling cloud-native access, AI model training, and automated processing workflows.



(a) post-stack 3D



(b) pre-stack streamer field

Figure 2: Examples of MDIO v1.0 dataset schemas after AI-driven metadata reconstruction, showing a post-stack 3D volume (top) and a pre-stack streamer field dataset (bottom) with defined dimensions, coordinates, and metadata attributes.

CONCLUSIONS

We have presented a multi-agent AI system that automates end-to-end metadata reconstruction for migrating legacy SEG-Y seismic archives to MDIO v1.0, a modern, self-describing, cloud-native format. The system decomposes the migration problem into three coupled tasks (seismic product type classification, header field extraction, and schema mapping), handled by specialized LLM-based agents augmented with deterministic verification and SME-in-the-loop refinement. Evaluated on expert-labeled data spanning diverse acquisition types and header conventions, the pipeline achieves 96.83% end-to-end accuracy, meeting or exceeding human-level performance, while processing each file in approximately 90 seconds. These results demonstrate that production-scale conversion from SEG-Y to MDIO v1.0 for archives exceeding one million files is technically feasible using agentic AI without model fine-tuning.

This work illustrates a broader principle for agentic AI in geoscience: by combining the semantic reasoning capabilities of LLMs with structured domain knowledge, explicit schema constraints, and expert oversight, it is possible to automate complex, knowledge-intensive tasks that previously required extensive manual effort. The resulting MDIO v1.0 datasets, with explicit dimensions, coordinates, and standardized metadata, transform static file archives into structured datasets suitable for scalable analytics and AI workflows.

ACKNOWLEDGMENTS

The authors thank the subject-matter experts at TGS who contributed ground-truth annotations and iterative feedback during system development.

REFERENCES

AWS, 2025, Strands agents: An open-source AI agents framework: https://github.com/strands-agents/sdk-python.

Hoyer, S., and J. Hamman, 2017, xarray: Nd labeled arrays and datasets in python: Journal of Open Research Software.

Michell, B., A. Sansal, B. Lasscock, and M. Roberts, 2025, Mdio v1: Schematizing seismic data for ai and processing: International Meeting for Applied Geoscience & Energy Expanded Abstracts, Society of Exploration Geophysicists, 706–710.

Sansal, A., and contributors, 2025a, mdio-python: Python library for the mdio multidi-mensional energy data format: https://github.com/TGSAI/mdio-python.

———, 2025b, segy: The ultimate python seg-y i/o with cloud support and schemas: https://github.com/TGSAI/segy.

Sansal, A., S. Kainkaryam, B. Lasscock, and A. Valenciano, 2023a, Mdio: Open-source format for multidimensional energy data: The Leading Edge, **42**, 465–473.

Sansal, A., B. Lasscock, and A. Valenciano, 2023b, Integrating energy datasets: the mdio format: First Break, **41**, 69–75.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, 2023, ReAct: Synergizing reasoning and acting in language models: Presented at the International Conference on Learning Representations.