AUTOREGRESSIVE DRILLHOLE MODELLING UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive modelling has achieved remarkable success in language and sequence tasks by learning to predict future states from previous observation. Mineral-exploration drillholes provide a natural but largely unexplored setting for this paradigm: as drilling proceeds, lithology is revealed sequentially from shallow to deep, making prediction of deeper strata inherently autoregressive. Existing drillhole modelling, however, is dominated by spatial interpolation and reconstruction, or largely rely on masked modelling, leaving strictly autoregressive prediction largely underexplored. We introduce DrillBench, a benchmark of 49,671 Western Australian drillholes for next-layer prediction and autoregressive stratigraphic generation across a graded transfer spectrum, from local prediction through spatial shift to cross geological province transfer. Benchmarking classical, geostatistical, and neural models reveals a clear *transfer boundary*: spatial and geochemical conditioning provides large local gains but deteriorates sharply under stronger shift, whereas lithology-sequence autoregressive models transfer more robustly. Guided by this finding, we develop a backbone-agnostic recipe combining large-scale pretraining on historical drillholes with spatial retrieval of neighbouring lithology. Retrieval is most effective in weathered cover, when local spatial continuity remains informative, whereas pretraining contributes more strongly in bedrock and under broader geological shift. Together, they retain strong local performance while improving generalisation under spatial and cross-province shift, most markedly on the most distant splits.

1 INTRODUCTION

Autoregressive modelling has expanded beyond language (Achiam et al., 2023; Brown et al., 2020) to scientific sequences such as proteins (Madani et al., 2023), genomes (Nguyen et al., 2024), and general time series (Garza et al., 2023). By repeatedly predicting the next state from the observed sequence, autoregressive models can learn recurring structural dependencies that support both next-step prediction and free-running generation. Such dependencies may also transfer across settings when similar underlying structure persists. Meanwhile, the machine-learning community has shown growing interest in geoscience, particularly through remote sensing (Bai et al., 2026; Ji et al., 2026) and Earth observation (Danish et al., 2026). However, most of this progress focuses on surface and near-surface observations, while subsurface geological sequences, including large drillhole archives, remain comparatively underexplored by modern sequence modelling.

The drilling process naturally induces an autoregressive prediction setting (Figure 1). As drilling progresses downwards, lithology is revealed layer by layer, so at any stage only the strata already encountered above the drill bit are available, while deeper formations remain unknown. This makes it natural to predict the next stratum from the sequence observed so far and recursively extend the hole to deeper. More importantly, autoregressive modelling can learn recurring stratigraphic dependencies within a hole, for example, a weathering profile progressing from laterite to saprolite and then to fresher bedrock, rather than relying entirely on nearby observations. Existing mineral-drillhole modelling, however, is dominated by interpolation, simulation, and 3D reconstruction (Strebelle, 2002; Ren et al., 2025; Li et al., 2026a), which primarily exploit spatial relationships between observations. Petroleum well-log studies retain the depth sequence but largely focus on masked representation learning or point-wise classification (Qi et al., 2025; Sun et al., 2023; Carvalho et al., 2026). Neither setting directly evaluates whether a model can use the observed stratigraphic sequence to roll a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

partially drilled hole forward from its own predictions, leaving autoregressive subsurface generation without a public benchmark or shared evaluation protocol.

Thus, we introduce **DrillBench**, a public benchmark of 49,671 mineral-exploration drillholes, together with about 112k-hole pretraining archive (WA-DASC-full) spanning several tectonic provinces in Western Australia. DrillBench evaluates two tasks: *next-layer prediction*, which predicts the next lithology from the ground-truth history above it, and *free-running generation*, which recursively generates the remaining lithology from an observed prefix. Both tasks are evaluated across progressively larger spatial and geological shifts, from in-distribution prediction through increasing spatial shift to cross-province transfer. Benchmarking a broad range of classical, geostatistical, and neural methods reveals a clear advantage of sequence-only autoregressive modelling: it provides a better balance between in-distribution accuracy and transfer. By relying on the observed sequence rather than location-specific signals, these models degrade more gradually as spatial and geological shift increases. In contrast, direct spatial and geochemical conditioning improves local accuracy but can become harmful under stronger shift, exposing a clear ***transfer boundary***.

Building on this transfer-safe autoregressive foundation, we further develop a backbone-agnostic training-and-inference recipe that introduces additional geological information. **(i) Large-scale pretraining** strengthens transferable stratigraphic structure using historical drillholes, with sparse geochemistry used only as auxiliary supervision. **(ii) Spatial retrieval** then introduces neighbouring lithology at inference without directly exposing absolute coordinates at test time. The two mechanisms complement the autoregressive model in different geological regimes. Retrieval is most effective in weathered cover within the source province but adds little across provinces. Pretraining shows the opposite trend: it contributes little in-distribution but becomes more valuable in bedrock and under broader geological transfer. Together, these results show how the three components complement each other: autoregressive sequence modelling provides the transferable foundation, retrieval strengthens prediction where local continuity is reliable, particularly in weathered cover, and pretraining becomes more important in bedrock and under broader geological transfer.

Our contributions are: **i) Formulation.** We formulate drillhole modelling as generative next-stratum prediction: the model uses only the strata already observed to predict the unseen geology below. We evaluate this through next-stratum prediction and free-running generation with a unified protocol. **ii) The transfer boundary.** Coordinates and assays improve local accuracy but degrade sharply under spatial and geological shift, whereas sequence-based models transfer more robustly. **iii) Pretraining and retrieval.** We extend autoregressive models with large-scale pretraining on large-scale drillholes and retrieval of neighbouring lithology, without directly conditioning on absolute coordinates or assays. Retrieval is most useful for local continuity in weathered cover, while pretraining contributes more in bedrock and cross-province transfer. **iv) Resources.** We release DrillBench with 49,671 drillholes, the WA-DASC-full pretraining corpus, pretrained checkpoints, and evaluation code.

2 Related Work

Recent machine-learning research has increasingly explored geoscience through remote sensing (Xiao et al., 2026) and Earth observation (Li et al., 2026b), including large-scale and multimodal foundation models trained on satellite imagery (Bai et al., 2026; Danish et al., 2026; Perron et al., 2026). These advances primarily model the Earth's surface and near-surface environment, whereas large-scale learning of subsurface geological structure has received substantially less attention. Mineral-drillhole modelling is dominated by *spatial interpolation and geostatistical simulation*, (Carle & Fogg, 1996; Strebelle, 2002; Caers, 2011; Ren et al., 2025; Li et al., 2026a). These approaches rely heavily on coordinates and neighbouring observations, making them effective for local reconstruction but less suited to learning stratigraphic relationships that may recur across locations. Petroleum well-log research instead models observations along depth, but most methods use masked modelling and bidirectional encoders (Qi et al., 2025; Liu et al., 2020; Sun et al., 2023; Carvalho et al., 2026). Because they can access information from deeper positions or predict individual depths independently, these formulations do not match the one-directional drilling process in which only the already drilled sequence is available. In addition, existing public benchmarks are generally small and geographically restricted (Hall, 2016; Bormann et al., 2020; Carvalho et al., 2026), limiting the study of graded transfer, large-scale pretraining, and retrieval-based use of neighbouring drillholes. This paper addresses these gaps by formulating drillhole modelling autoregressively, evaluating predic-

tion and generation across progressively stronger spatial and geological shifts, and using a historical archive to study how pretraining and retrieval contribute to transfer.

3 BENCHMARK PROTOCOL

3.1 DRILLHOLE REPRESENTATION

A single drillhole representation format does not suit all modelling approaches. Autoregressive models naturally operate on geological layers, while spatial and point-wise methods make predictions at specific depths. We therefore provide two representations of each drillhole: a compact *layer-level* sequence and a *cell-level* sequence defined over depth.

a) Layer-level representation. A drillhole is represented as $\mathcal{H}^{\text{layer}} = \big((\ell_1, D_1), \ldots, (\ell_L, D_L)\big)$, where $\ell_i \in \mathcal{L}$ is the lithology of stratum i, $D_i > 0$ is its thickness, and $\mathcal{L}$ contains the 37 standardised lithology codes. This representation directly models *which lithology comes next* and *how thick it is*, without repeating the same lithology. It is used by layer-transition baselines such as Markov models and by layer-level neural models such as LSTM and Transformer. **b) Cell-level representation.** The drillhole is represented as a depth-indexed sequence $\mathcal{H}^{\text{cell}} = (x_1, \ldots, x_T), \quad x_t \in \mathcal{L}'$, where $\mathcal{L}' = \mathcal{L} \cup \{\texttt{CONT}, \texttt{UNK}\}$ and UNK denotes missing or unmappable lithology. Given a depth spacing Δd, the first cell of each stratum carries its lithology code and the remaining cells are encoded as CONT. The number of CONT cells therefore preserves stratum thickness without revealing where the stratum ends in advance. Cell-level sequence models such as LSTM, and Mamba use this sequence directly. For spatial methods such as k-NN, and feature-based models such as XGBoost, CONT cells are replaced by the corresponding stratum lithology so that each depth cell has an explicit label.

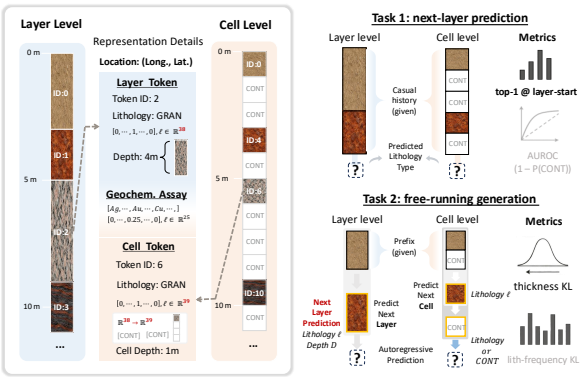

Figure 1: Two granularities of drillhole representation and DrillBench task definitions.

3.2 TASK DEFINITION

3.2.1 TASK 1: NEXT-LAYER PREDICTION

Definition. At each true stratum boundary, the model predicts the lithology of the next stratum using only information available up to that boundary. Ground-truth boundaries are provided so that the task isolates *which lithology comes next*, rather than jointly evaluating boundary or thickness prediction. Layer-level models use the preceding strata, while cell-level models use the corresponding cell sequence. **Evaluation.** All methods are evaluated over the same 37 lithology classes at the ground-truth stratum boundaries. Layer-level models directly predict the 37 lithology classes. Cell-level models also predict CONT; at the true stratum boundaries, we remove CONT and renormalise the lithology probabilities. We report macro-F1 over the lithology classes with at least 20 events in each split as the main metric (Appendix C.2). **Baselines.** We include methods operating at both levels. Cell-level baselines include frequency and Markov models, geostatistical methods such as spatial k-NN and TPROGS (Caers, 2011; Strebelle, 2002; Carle & Fogg, 1996), XGBoost (Chen & Guestrin, 2016), and neural sequence models including GRU, LSTM, TCN, Transformer, and Mamba. Layer-level baselines include corresponding transition models and neural sequence architectures applied directly to the stratum sequence.

3.2.2 TASK 2: FREE-RUNNING GENERATION

Definition. Given the true first $p\%$ of a drillhole, the model generates the remainder autoregressively. At each step, the model decides whether a new layer begins and, if so, predicts its lithology. Each prediction becomes the next input until the known hole depth is reached. No ground-truth

boundaries are provided, so errors can accumulate. **Evaluation.** Our primary metric is the *joint KL*, $\mathrm{KL}(\text{true}|\text{generated})$, between the empirical joint distributions of lithology and layer thickness over each split. Lithology uses the 37 DrillBench classes and thickness seven log-spaced bins, so the metric penalises mismatches in either rock composition or thickness. We average over $M{=}5$ rollouts per hole; lower is better. (Details and additional metrics are reported in Appendix D.2). **Baselines.** Markov, frequency, and neural sequence models from Task 1 can use their own predictions as input and therefore support free-running generation. Spatial methods such as kriging, spatial k-NN, and simulation cannot recursively extend a hole and are reported for reference only.

4 DrillBench Dataset

4.1 Source Corpus

DrillBench is derived from the Western Australia[1], a public archive of mineral exploration submissions released under CC-BY-4.0. We focus on two geologically distinct areas: the *Eastern Goldfields* (EG; Archean greenstone–granite, gold-dominated), with 46,393 drillholes, and the *Pilbara* (Archean BIF iron-formation, iron-dominated), with 3,278 drillholes. After curation (§4.2), DrillBench contains **49,671 drillholes and 1,499,181 depth intervals** collected over 30 years data, stored as nested Apache Parquet files. To our knowledge, this is the first public drillhole corpus formatted explicitly for 1D autoregressive sequence modelling.

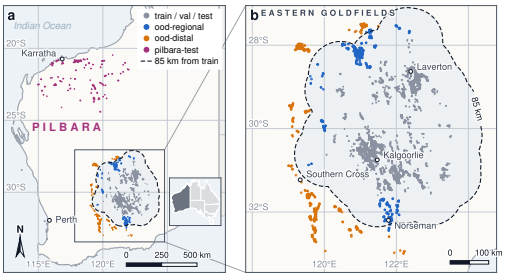

Figure 2: DrillBench six splits spatial layout.

4.2 Dataset Curation

To construct a clean, standardised, and leakage-resistant benchmark, we apply six preprocessing steps (details and per-step counts in Appendix B.1). **(1) Invalid-record removal.** We remove 26 corrupted records with implausible drill depths ($> 10\,\mathrm{km}$) or malformed `hole_id` fields. **(2) Hole-level quality filtering.** We retain drillholes that are at least $30\,\mathrm{m}$ deep, containing at least five logged intervals, and have substantive lithology text for at least half of them, ensuring sufficient context for autoregressive evaluation. **(3) Lithology vocabulary normalisation.** Free-text descriptions are mapped to a canonical 37-code vocabulary shared across data providers. About 68% of intervals are mapped; unmapped intervals retain their raw text and are treated as unobserved. **(4) Physical-bounds masking.** For each of the 25 elements, values above a stoichiometric upper bound (e.g., $\mathrm{Fe} \geq 72\%$) are set to `null`. Negative values within the valid range are retained because they encode below-detection-limit measurements. **(5) Coordinate harmonisation.** Collar coordinates reported in different zones and datums, including GDA94 and AGD84, are transformed into a common regional reference frame so that splitting, de-duplication, and retrieval use consistent ground distances. We exclude 27 drillholes without valid coordinates. **(6) Twin-hole de-duplication.** The same physical drillhole may appear in multiple statutory reports under different identifiers, creating evaluation leakage. We remove evaluation holes with a training-set twin within $25\,\mathrm{m}$ whose lithology sequence is identical or whose interval grid matches within $0.11\,\mathrm{m}$. This removes 1,051 duplicate holes from VAL and TEST, leaving 49,671 drillholes and 1,499,181 intervals.

4.3 Drillhole Data Modalities

Each drillhole comprises hole-level metadata and a sequence of depth intervals. (1) *Collar metadata* provides drillhole-level context, including location, elevation, region and planned maximum depth. (2) *Lithology* records the rock type at each depth interval and serves as the main prediction target. Each mapped interval is assigned one of 37 predefined lithology labels (s ee Appendix B.2), such as banded iron formation (BIF). (3) *Geochemistry* records the concentrations of 25 elements

[1] `https://dasc.dmirs.wa.gov.au/`

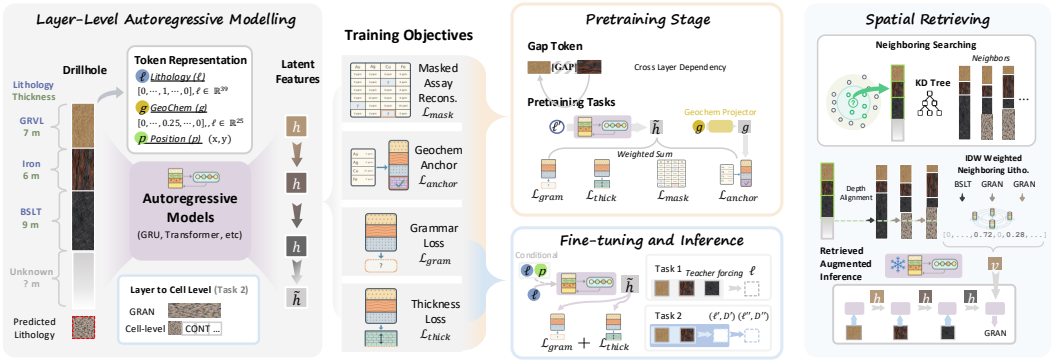

Figure 3: A Transfer-Safe Recipe. Pretraining, fine-tuning, and retrieving augmented pipeline.

at each interval. Because measurements are incomplete across depths, each geochemical vector is accompanied by a binary mask ($m_t \in \{0,1\}^{25}$) indicating the available elements.

4.4 DrillBench Splits

Each drillhole in DrillBench is assigned to one of six subsets (Table 1), defining three evaluation settings with progressively stronger distribution shifts: *i)* ***In-distribution***, using spatially clustered Eastern Goldfields (EG) drillholes for training, validation, and testing; *ii)* ***Graded spatial shift***, using two groups of EG drillholes increasingly separated from the training data (with same geological province but different spatial distance), with OOD-REGIONAL located 15–98km from the nearest TRAIN collar (median 58km) and OOD-DISTAL 86–205km away (median 129km); and *iii)* ***Cross-province transfer***, using PILBARA-TEST to evaluate generalisation to a geologically distinct province. Together, these settings form a controlled transfer spectrum from local prediction, through within-province spatial separation, to cross-province geological generalisation (Figure 2).

Table 1: DrillBench split summary across splits.

Split	Region	Holes n	Intervals n	1m tokens n	Depth $\bar{d}$ (m)	Lith. (%)
In-distribution: Eastern Goldfields (EG) core						
TRAIN	EG	27,435	818,350	2,076,335	75.9	70.1
VAL	EG	1,951	60,998	152,817	78.5	69.4
TEST	EG	3,856	120,196	301,531	78.3	70.0
Graded spatial shift: EG outer rings						
OOD-REGIONAL	EG	8,874	201,110	743,314	83.7	62.7
OOD-DISTAL	EG	4,277	118,908	251,170	58.8	59.2
Cross-province transfer: Pilbara						
PILBARA-TEST	Pilbara	3,278	179,619	239,708	73.6	70.5
Total		**49,671**	**1,499,181**	**3,764,875**	**76.0**	**68.3**

WA-DASC-full. Alongside DrillBench, we release WA-DASC-full, containing 498,713 drillholes across 30 DASC tectonic-province labels, 13 of which have more than 5,000 holes. About 112k holes with at least two logged lithology layers are used for pretraining. Unlike DrillBench, WA-DASC-full is only lightly standardised and is used for pretraining and retrieval rather than benchmark evaluation. To prevent leakage, all DrillBench evaluation holes, and their archive twins within 500 m, are excluded from pretraining, fine-tuning, and retrieval.

5 A Transfer-Safe Training-and-Inference Recipe

We introduce a backbone-agnostic training-and-inference recipe, which aims to bring in broader geological knowledge and local evidence without providing absolute coordinates and assay values out of the prediction input. The recipe has three components: a *layer-level autoregressive* backbone over strata and thicknesses (§5.1), large-scale *pretraining on the uncurated* WA-DASC-full archive with geochemistry as auxiliary supervision (§5.2), and *spatial retrieval* that uses coordinates only to look up neighbouring lithology (§5.3). *Fine-tuning, decoding, and inference* follow in §5.4.

5.1 Layer-Level Autoregressive Model

To directly model the dependencies between successive lithological strata, we represent each drillhole as a sequence of layers, $(\ell_1, D_1), \ldots, (\ell_L, D_L)$, where ℓ_i denotes the lithology of layer i and $D_i > 0$ is its thickness. It follows a layer-level autoregressive architecture: given the first i layers,

it predicts the lithology and thickness of layer $i{+}1$. A sequence backbone f_θ maps the embedded lithologies to $h_i = f_\theta(\mathbf{e}(\ell_1), \ldots, \mathbf{e}(\ell_i))$, where $\mathbf{e}$ is a learned lithology embedding. The hidden state is first passed to a **lithology head**, which predicts a categorical distribution $q_i(k)$ over the lithology classes of the next layer. In addition, a separate **thickness head** predicts the corresponding layer thickness using a log-normal distribution. The predicted lithology and thickness together form the next layer (ℓ_{i+1}, D_{i+1}), which is appended to the generated sequence and the process repeats autoregressively until the drillhole is completed.

5.2 Pretraining on Uncurated Archives

Large-scale pretraining allows model to learn lithological dependencies that generalise beyond individual regions. Geological sequences contain recurring patterns across regions, such as weathering profiles from laterite through saprolite to fresher bedrock. Pretraining on a much larger archive can therefore provide more robust priors for cross-region generalisation. However, WA-DASC-full is highly heterogeneous, with only $\sim$15% of intervals mapping to the target lithology vocabulary. We address this heterogeneity through gap tokenisation and multi-task supervision.

Gap tokenisation. To preserve stratigraphic continuity when lithology records are incomplete, we introduce an explicit GAP token for unlogged intervals rather than breaking a drillhole into disconnected fragments. For example, $A \rightarrow [\text{GAP}] \rightarrow B$ preserves the relative ordering of the observed strata while marking the missing section. This allows models to learn from longer layer sequences without treating missing records as observed lithologies.

Multi-task pretraining. To learn transferable geological representations from the large-scale archive, we introduce four training objectives. The two primary tasks predict the next lithology and its thickness through $\mathcal{L}_{\text{gram}}$ and $\mathcal{L}_{\text{thick}}$. We further use geochemical assays as auxiliary supervision where available: $\mathcal{L}_{\text{mask}}$ reconstructs randomly masked element values to capture compositional relationships, while $\mathcal{L}_{\text{anchor}}$ predicts the corresponding lithology from chemistry when a mapped label is available. These auxiliary tasks allow assayed intervals, including those with unreliable or unmappable lithology descriptions, to provide additional training signals. Geochemistry *is used only during pretraining* and removed during fine-tuning and inference, so the downstream model does not depend on assay availability in unseen regions. The overall objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{gram}} + \lambda_{\text{thick}}\mathcal{L}_{\text{thick}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}}, \tag{1}$$

where λ balances the pretraining objectives.

5.3 Spatial retrieval without explicit coordinate inputs

We use neighbouring drillholes as local geological evidence without exposing absolute coordinates to the model. Coordinates are used only to retrieve the $K{=}10$ nearest logged holes in the same region from WA-DASC-full and DrillBench TRAIN. Evaluation holes are excluded from the retrieval pool, and neighbours within $25\,\text{m}$ of the query collar are skipped to reduce the risk of retrieving duplicate holes. To predict layer $i{+}1$, each retrieved neighbour is queried at the corresponding elevation when valid elevation information is available; otherwise, the same down-hole depth is used. A neighbour contributes a vote only if one of its beds starts within the corresponding $1\,\text{m}$ depth cell. Its lithology is mapped to the 37-class vocabulary, and votes are combined using inverse-distance weighting:

$$r_i = \sum_{j \in \mathcal{N}_i} \alpha_{ij}\, \mathbf{1}(c_{ij}), \qquad \alpha_{ij} = \frac{1/d_j}{\sum_{m \in \mathcal{N}_i} 1/d_m}, \tag{2}$$

where $\mathcal{N}_i$ is the set of voting neighbours, $\mathbf{1}(c_{ij})$ is the one-hot lithology vector, and d_j is the collar distance. Thus, r_i captures local lithological context without directly encoding absolute position. We fuse the retrieval signal with the current layer representation as $x_i = \mathbf{e}(\ell_i) + \phi(r_i)$, where ϕ is a small MLP. If no neighbour provides a valid vote, we set $r_i = \mathbf{0}$. Model parameters are fixed across all test sets, so at test time target-region information enters only through retrieved neighbours.

5.4 Fine-Tuning, Decoding, and Evaluation

Fine-tuning. The model f_θ is fine-tuned on TRAIN for the 37 lithology classes and thickness (chemistry pathway removed). The retrieval encoder ϕ, absent from pretraining, is learned jointly under

label smoothing ($\varepsilon = 0.3$). The layer input is $x_i = \mathbf{e}(\ell_i) + \phi(r_i)$, and $h_i = f_\theta(x_1, \ldots, x_i)$. Retrieval is zeroed for 30% of training holes, so the model can fall back to the sequence alone. The heads give

$$q_i(k) = P(\ell_{i+1} = k \mid h_i), \qquad p(D_{i+1} \mid h_i, \ell_{i+1}), \tag{3}$$

where $p(D)$ is a discretised log-normal over thickness in cells.

Decoding and evaluation. Every baseline answers the same question on the 1 m grid: whether the next cell is CONT, meaning that the current stratum continues, or a lithology k, meaning that a new stratum begins. Our model instead predicts the next stratum's lithology distribution q and thickness distribution. We convert these predictions to the same cell-level form. If λ is the probability that the current stratum ends at the present cell, then

$$P(\texttt{CONT}) = 1 - s\lambda, \qquad P(k) = s\lambda\, q(k), \tag{4}$$

where s is a single calibration factor for the predicted boundary probability. A larger s produces more stratum boundaries, while a smaller s produces fewer. From this point, all methods have the same output space and are evaluated by the same code. In *Task 1*, each method is given the true history above a stratum boundary and is scored by whether its most probable lithology matches the stratum that actually begins there. In *Task 2*, each method is given the true upper half of a hole and generates the remainder autoregressively until the logged depth. The generated cells are then grouped into strata and compared with the true lower half using their joint lithology–thickness distribution. For each generator, s is fitted only on TRAIN holes to match the observed boundary rate, and is then kept fixed for all test sets (Appendix C.2).

6 Results and Discussion

We present the results in three stages. We first benchmark autoregressive models against classical, geostatistical, and learning-based alternatives on next-layer prediction and free-running generation. We then study how performance changes with increasing spatial and geological shift, using this analysis to expose the *transfer boundary* and assess how pretraining and retrieval mitigate it. Finally, case studies and error analysis highlight the remaining limitations and point to future directions. Additional results and analyses are reported in the Appendix D.

Table 2: **Tasks 1 and 2 performance.** Task 1 reports macro-F1 (%, ↑); Task 2 reports joint KL (↓). L: lithology; S: location; R: retrieval.

		Inputs			Increasing spatial shift →			
Family	**Method**	L	S	R	Test	OOD-R	OOD-D	Pil-T
Task 1: next-layer macro-F1 (%, ↑)								
Cell level (1 meter token)								
Freq.	Unigram				2.0	0.8	1.1	1.7
	3-gram	✓			9.4	3.2	4.0	3.7
Geostat.	TPROGS	✓	✓		4.5	1.1	1.2	1.7
	Spatial k-NN			✓	68.9	11.5	9.6	8.0
ML	XGBoost	✓	✓		64.8	2.4	2.5	2.2
DL	GRU	✓			58.8	21.4	23.0	23.2
	Transformer	✓			60.3	26.5	23.6	22.5
Layer level (stratum token)								
Markov	Layer Markov	✓			56.7	36.3	26.1	22.0
DL	GRU	✓			59.4	37.4	28.3	28.2
	Transformer	✓			58.9	36.6	29.1	26.0
Ours	GRU + R	✓		✓	**70.0**	37.4	33.9	26.9
	GRU + PT	✓			58.5	**39.2**	33.0	**31.7**
	GRU + PT + R	✓		✓	69.2	38.8	**37.2**	30.6
Task 2: joint KL (↓)								
Cell level (1 meter token)								
Freq.	Unigram				0.226	0.599	0.872	1.111
	3-gram	✓			0.186	0.533	0.781	0.960
Geostat.	TPROGS	✓	✓		0.183	0.492	0.716	0.885
	Spatial k-NN*			✓	0.553	0.950	0.852	0.729
ML	XGBoost	✓	✓		0.206	0.518	0.950	1.776
DL	GRU	✓			0.340	0.612	0.559	1.003
	Transformer	✓			0.356	0.585	0.533	0.995
Layer level (stratum token)								
Markov	Layer Markov	✓			0.195	0.385	0.509	0.849
DL	GRU	✓			0.183	0.356	0.464	0.550
	Transformer	✓			**0.172**	0.320	0.495	0.544
Ours	GRU + R	✓		✓	0.192	0.341	0.479	0.561
	GRU + PT	✓			0.185	**0.303**	0.488	0.551
	GRU + PT + R	✓		✓	0.186	0.322	**0.408**	**0.458**

6.1 Benchmark Performance

Task 1: Next-layer prediction. Frequency-based methods remain weak across all settings. Spatial k-NN and XGBoost perform well on the in-distribution test set, but their performance drops sharply on more distant regions. In contrast, autoregressive neural models degrade more gradually and maintain much stronger performance under spatial shift. Cell- and layer-level neural baselines perform similarly in-distribution, but layer-level models are generally stronger under spatial shift. The benefit becomes clearer when pretraining and retrieval are added. Retrieval gives the largest improvement in-distribution, with GRU+R reaching 70.0% macro-F1. Pretraining is more helpful for transfer, giving the best OOD-R and Pilbara-test results (39.2% and 31.7%). Combining pretraining and retrieval achieves the best OOD-D result at 37.2% and remains competitive in the other settings. Overall, these results show that retrieval mainly strengthens local prediction, while pretraining improves transfer to more distant regions. We therefore focus the following analysis on layer-level models.

Task 2: Free-running generation. This task is more challenging because errors accumulate over the rollout. Frequency, geostatistical, and ML methods are competitive in some settings but degrade under larger spatial shifts. Neural autoregressive models are more stable, with layer-level models generally outperforming cell-level models on the harder transfer splits. At the layer level, retrieval and pretraining provide complementary gains. GRU+PT+R achieves the best OOD-D and Pilbara-test results, with joint KL of 0.408 and 0.458, while GRU+PT is strongest among our models on OOD-R. Together with Task 1, these results support modelling lithology directly at the stratum level for more transferable autoregressive generation.

6.2 TRANSFER BOUNDARY AND MITIGATION STRATEGIES - OUR TRANSFER SAFE RECIPE

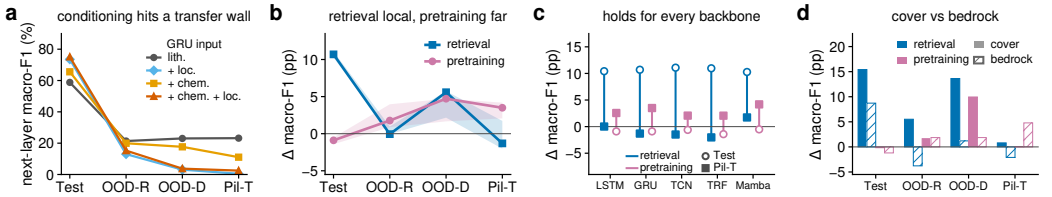

Figure 4: **What transfers.** Next-layer prediction performance under increasing shift. **(a)** Direct conditioning and the transfer boundary (lith only). **(b)** Local retrieval vs. broad pretraining. **(c)** Consistency across five backbones. **(d)** Regolith cover vs. bedrock.

i) Direct conditioning causes the transfer boundary. Figure 4(a) shows a clear trade-off between local accuracy and transfer. Adding location and chemistry improves TEST performance, but these gains collapse under distribution shift: the fully conditioned model drops from 75% on TEST to below 3% on PILBARA-TEST, while the lithology-only model remains above 20% on all shifted splits. We call this reversal the *transfer boundary*: signals that are highly predictive locally can become harmful when their regional associations no longer hold. This is consistent with known limitations of coordinates as spatial predictors (Meyer et al., 2019; Ploton et al., 2020); our receip therefore avoids using raw coordinates and assays as model inputs. **ii) Pretraining supports broad geological transfer.** Figure 4(b) shows that pretraining provides little gain in-distribution, but becomes positive once the data shift and stays positive on every shifted split, indicating that it mainly contributes transferable geological knowledge rather than local fit. This trend is consistent across multiple deep-learning backbones (Figure 4(c)): pretraining has almost no effect on TEST, but consistently improves PILBARA-TEST performance. **iii) Retrieval supports local spatial adaptation.** Retrieval shows the opposite pattern. It provides the largest gains in distribution, where nearby drillholes supply directly relevant geological evidence, and its benefit is smallest across provinces (Figure 4(b)). The same behaviour appears across multiple backbones (Figure 4(c)). This suggests that retrieval is most useful for local adaptation rather than broad geological transfer. **iv) Pretraining and retrieval are complementary across geological regimes.** Figure 4(d) separates their effects between regolith cover and bedrock. Cover comprises weathered or transported material above the bedrock and is strongly shaped by local regolith–landform processes, whereas bedrock more directly records the geological units relevant to primary mineralisation (Anand & Butt, 2010). Consistent with this distinction, retrieval is most effective in Eastern Goldfields cover, where nearby holes provide useful local context, while pretraining is more valuable in bedrock on OOD-R and across provinces. OOD-D is the exception because most scored layers are regolith (Table 12). Overall, retrieval captures local near-surface continuity, whereas pretraining better supports transfer of deeper geological structure.

6.3 FAILURE MODES AND CASE STUDIES.

Failures Concentrate on Poorly Supported Lithologies. Figure 5(a) compares per-class recall in-distribution and on PILBARA-TEST. In-distribution, the model performs well on most common cover and bedrock classes. Under cross-province transfer, however, errors are concentrated in lithologies with little source-region support. Well-represented classes such as CLAY, SAPL, and VEIN remain reasonably predictable, whereas every zero-recall class on PILBARA-TEST has at most 71 training layers. This indicates a clear class-support shift: stratigraphic patterns may transfer, but lithologies that are rarely observed in the source region remain difficult to recover from sequence context alone. Future gains may therefore come more from larger and geologically broader training corpora, or limited target-region supervision, than from simply increasing model capacity.

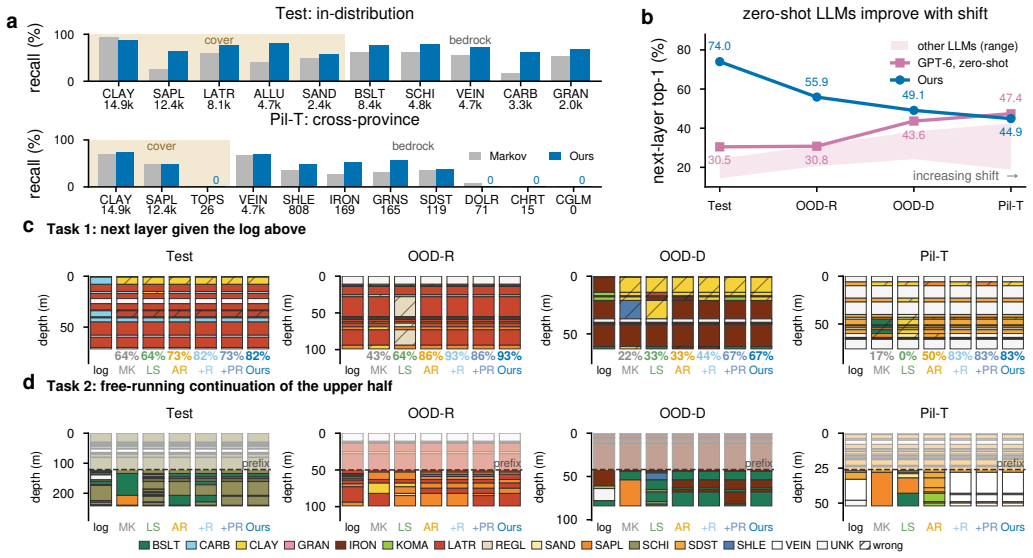

Figure 5: **Failure analysis under geological shift.** (a) Cross-province errors concentrate on lithologies with limited training support. (b) Without retrieval, zero-shot LLMs degrade less under shift (top-1 accuracy on 150 randomly selected holes per split; our model is scored on the same holes). (c–d) Case studies show the complementary effects of retrieval and pretraining, together with remaining lithology and boundary errors.

LLMs Perform Better OOD than In-Domain. We additionally evaluate several recent LLMs under zero-shot prompting, with and without retrieved geological context. Surprisingly, without retrieved context their performance improves with increasing shift: GPT-6 rises from 30.5% to 47.4% under cross-province shift, slightly above our model (44.9%), while our trained model shows the opposite trend. This suggests that LLMs may draw on broad geological knowledge that is less tied to the source region, although part of it may reflect exposure to the public DASC records during pretraining (Appendix D.3). Although weaker than specialised models in-distribution, they may be particularly useful when target-region labels are scarce. A promising direction is to combine LLMs with our retrieval and autoregressive components in an agentic system, using the LLM for broad geological reasoning and the specialised model for local prediction.

Case Studies Reveal Complementary Strengths and Remaining Failures. Figure 5(c–d) shows how these trends appear at the drillhole level. The autoregressive model already captures major stratigraphic transitions better than Markov and cell-level baselines, while retrieval helps when nearby holes provide relevant local structure and pretraining becomes more useful as the geological shift increases. Combining both gives the most consistent predictions across the four spatial settings, including substantial gains on the OOD-D and PILBARA-TEST examples. The free-running cases show similar progress, but also expose remaining errors in lithology identity and layer boundaries as predictions accumulate. Together, these examples reinforce the main result: autoregressive modelling provides the transferable foundation, retrieval contributes local evidence, and pretraining supplies broader geological priors. Further gains are likely to require broader geological coverage and better use of sparse target-region evidence (e.g. LLMs), rather than simply larger backbones.

7 CONCLUSION

We study autoregressive modelling for mineral-exploration drillholes, a domain where predictions must remain useful across strong spatial and geological shifts. We introduce DrillBench to evaluate next-layer prediction and free-running generation, and show that location- and assay-based conditioning can create a clear transfer boundary. To address this domain-specific challenge, we develop a transfer-safe recipe that combines layer-level autoregression, large-scale geological pretraining, and spatial retrieval: retrieval exploits local continuity, while pretraining supports broader geological transfer. Our results show that effective subsurface modelling requires not only stronger sequence models, but mechanisms designed around geological structure, data sparsity, and regional shift.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, translation, manuscript organisation, literature search, and the refinement of methodological descriptions. They were also used to discuss experimental design and interpretation of results. All AI-assisted suggestions were critically reviewed by the authors and checked against the underlying data, experiments, code, and cited literature. The authors made all final methodological and scientific decisions and take full responsibility for the content, results, and claims of this work.

ETHICS STATEMENT

This work uses only public mineral-exploration records from the DEMIRS DASC archive (CC-BY-4.0), which contain no personal data and involve no human subjects. We foresee no ethical concerns beyond standard caution in using model predictions for exploration decisions.

REPRODUCIBILITY STATEMENT

All implementation details needed to reproduce our results are provided in the appendix, including the full data curation pipeline and lithology vocabulary (Appendices B.1 and B.2) together with the split construction (§4.4), architectures, hyperparameters, and training schedules for every baseline and for our model (Appendix C.1), the evaluation protocol and metrics (Appendix C.2), and the exact prompts and inference settings for the LLM baselines (Appendix D.3). We further report extensive ablations on every component of the recipe (Appendix D.4; all settings are summarised in Table 19), together with seeds and hole-cluster bootstrap confidence intervals (Appendices D.1 and D.2). The DrillBench dataset, the WA-DASC-full pretraining and retrieving corpus, all code, pretrained checkpoints, and raw evaluation outputs will be publicly released upon acceptance.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

RR Anand and CRM Butt. A guide for mineral exploration through the regolith in the yilgarn craton, western australia. *Australian Journal of Earth Sciences*, 57(8):1015–1114, 2010.

Lubin Bai, Xiuyuan Zhang, Siqi Zhang, Zepeng Zhang, Haoyu Wang, Wei Qin, and Shihong Du. Geolink: Empowering remote sensing foundation model with openstreetmap data. *Advances in Neural Information Processing Systems*, 38:168557–168587, 2026.

Peter Bormann, Peder Aursand, Fahad Dilib, Surrender Manral, and Peter Dischington. Force 2020 well well log and lithofacies dataset for machine learning competition. *Zenodo: Geneva, Switzerland*, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jef Caers. *Modeling uncertainty in the earth sciences*. John Wiley & Sons, 2011.

Steven F Carle and Graham E Fogg. Transition probability-based indicator geostatistics. *Mathematical geology*, 28(4):453–476, 1996.

Henrique MO Carvalho, Henrique Daniel, André Korenchendler, Matheus CA Sobreira, and Alexei MC Machado. Lithology classification based on well log data: A benchmark for machine learning models. *Mathematical Geosciences*, pp. 1–23, 2026.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Shah, Muhammad Haris Khan, Rao Anwer, Jorma Laaksonen, Fahad Khan, and Salman Khan. Terrafm: A scalable foundation model for unified multisensor earth observation. In *International Conference on Learning Representations*, volume 2026, pp. 12787–12809, 2026.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1. *arXiv preprint arXiv:2310.03589*, 2023.

Hossein Ghorbanfekr, Pieter Jan Kerstens, and Katrijn Dirix. Classification of geological borehole descriptions using a domain adapted large language model. *Applied Computing and Geosciences*, 25:100229, 2025.

Brendon Hall. Facies classification using machine learning. *The Leading Edge*, 35(10):906–909, 2016.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7038–7051, 2021.

Jinheng Ji, Jiahui Qu, Wenqian Dong, and Yunsong Li. Cf-ipt: Cross-modal fusion interactive prompt tuning of vision-language pre-trained model for multisource remote sensing data classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23021–23030, 2026.

Hao Li, Chi Zhang, and Zhenwen He. A generalized borehole autoregressive neural network for 3d geological modeling. *Algorithms*, 19(2):128, 2026a.

Xiang Li, Yong Tao, Siyuan Zhang, Siwei Liu, Zhitong Xiong, Chunbo Luo, Lu Liu, Mykola Pechenizkiy, Xiaoxiang Zhu, and Tianjin Huang. Reobench: Benchmarking robustness of earth observation foundation models. *Advances in Neural Information Processing Systems*, 38, 2026b.

Haining Liu, Yuping Wu, Yingchang Cao, Wenjun Lv, Hongwei Han, Zerui Li, and Ji Chang. Well logging based lithology identification model establishment under data drift: A transfer learning method. *Sensors*, 20(13):3643, 2020.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41 (8):1099–1106, 2023.

Hanna Meyer, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. *Ecological Modelling*, 411:108815, 2019.

Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. Sequence modeling and design from molecular to genome scale with evo. biorxiv. *Published online March*, 6:2024–02, 2024.

Yohann Perron, Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Universat: Resolution-and modality-agnostic transformers for earth observation. *arXiv preprint arXiv:2606.23503*, 2026.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 11(1):4540, 2020.

Zhenyu Qi, Qing Yu, Jichen Wang, Yun-Bo Zhao, Zerui Li, and Wenjun Lv. Wlfm: A well-logs foundation model for multi-task and cross-well geological interpretation. *arXiv preprint arXiv:2509.18152*, 2025.

Anjing Ren, Liang Wu, Jianglong Xu, Yanjie Xing, Qinjun Qiu, and Zhong Xie. A deep learning method for 3d geological modeling using et4dd with offset-attention mechanism. *Computers & Geosciences*, 200:105929, 2025.

Sebastien Strebelle. Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical geology*, 34(1):1–21, 2002.

Longxiang Sun, Zerui Li, Kun Li, Haining Liu, Ge Liu, and Wenjun Lv. Cross-well lithology identification based on wavelet transform and adversarial learning. *Energies*, 16(3):1475, 2023.

Daniel Wedge, Owen Hartley, Andrew McMickan, Thomas Green, and Eun-Jung Holden. Machine learning assisted geological interpretation of drillhole data: Examples from the pilbara region, western australia. *Ore Geology Reviews*, 114:103118, 2019.

Aoran Xiao, Shihao Cheng, Yonghao Xu, Yexian Ren, Hongruixuan Chen, and Naoto Yokoya. Geommbench and geommagent: Toward expert-level multimodal intelligence in geoscience and remote sensing. *arXiv preprint arXiv:2604.08896*, 2026.

Table 3: Comparison with representative studies on subsurface modelling.

Work	Object	Input	Task formulation	Evaluation setting	Scale	Public data
Petroleum well logs						
Lithology benchmark Carvalho et al. (2026)	well log	curves and facies	point-wise classification	within-basin CV	341 wells	yes
SEG'16 contest Hall (2016)	well log	curves and facies	point-wise classification	within-basin	10 wells	yes
TL under data drift Liu et al. (2020)	well log	continuous curves	classification and adaptation	labelled target wells	7 wells	no
FORCE'20 contest Bormann et al. (2020)	well log	curves and facies	point-wise classification	within-basin	118 wells	yes
Adversarial cross-well Sun et al. (2023)	well log	continuous curves	classification and adaptation	labelled target wells	∼32k pts	no
TimeGPT Garza et al. (2023)	well log	single curve	zero-shot forecasting	cross-basin	100B pts[‡]	no
WLFM Qi et al. (2025)	well log	multiple curves	masked pretraining	cross-well, same basin	1.2k wells	no
Mineral drillholes						
ML-assisted logging Wedge et al. (2019)	drillhole	lithology, assays	stratigraphic interpretation	single province	396 holes	no
ET4DD Ren et al. (2025)	drillhole	coordinates	3D point classification	single area	719 holes	no
GBARNN Li et al. (2026a)	drillhole	coordinates	spatial autoregression	single area	64 holes	no
GEOBERTje Ghorbanfekr et al. (2025)	drillhole text	free text	lithology classification	not evaluated	283k desc.	partial
Ours	**drillhole**	**lithology**	**causal prediction and generation**	**cross-province zero-shot**	**50k holes**	**yes**

Table 4: Curation funnel. Steps 4–5 change values, not rows; step 7 moves holes out of the benchmark, it does not delete them.

Step	Removed	Holes	Intervals
(0) Logged holes in region	–	706,198	16,008,692
(1) Corrupt records	26 records	–	–
(2) Quality filter	655,085 holes	51,113	1,533,265
(3) No valid coordinates	27 holes	51,086	1,533,004
(4) Lithology mapping	0	68.2% intervals mapped	
(5) Bounds masking	0	409,821 values nulled	
(6) Twin de-duplication (VAL, TEST)	1,051 holes	50,035	1,511,299
(7) Pilbara context holes → retrieval archive	364 holes	**49,671**	**1,499,181**

A Scope and Limitations

The aim of this paper is to establish autoregressive next-stratum modelling as a paradigm for drillholes: we formulate the task, build a benchmark and protocol that test it under graded spatial and cross-province shift, and show with a transfer-safe recipe that sequence structure learned from historical archives transfers where location-based conditioning does not. Several scope choices serve this aim. We use company logs as they are: depths and thicknesses are measured along the hole without correcting for its orientation (dip and azimuth), so thicknesses are apparent rather than true, and lithology comes from a rule-based harmonisation of free-text descriptions (App. B.1); the formulation must work on the records that exploration actually produces. We evaluate cross-province transfer on the Pilbara, which we harmonised to benchmark quality alongside the Eastern Goldfields; extending the benchmark to the other provinces of WA-DASC-full, where only about 15% of intervals map to our vocabulary (§5.2), requires reconciling each company's logging vocabulary, a curation effort that we leave to future releases. We keep the models small (Table 8) and their inputs to lithology (App. D.4.1), so that the gains can be attributed to the formulation rather than to scale or to region-specific signals. Orientation-corrected depths, larger backbones, per-hole likelihood evaluation of generation (App. D.2), and transferable encodings of location and chemistry are natural next steps, and the benchmark is built to support them.

B Dataset Details

B.1 Dataset Curation Details

Table 4 reports how many holes and intervals each step removes. Steps are grouped by what they act on rather than listed in execution order. The vocabulary mapping (4) is computed before the quality filter (2), whose condition (iv) depends on it, and value-level masking (5) removes no rows, so the counts are unaffected by ordering.

(1) Corrupt-record removal. The source is distributed as native BCP exports, and a small fraction of rows are byte-desynchronised. We detect these by values no drillhole can take: a maximum depth above $10\,\mathrm{km}$, or a `hole_id` longer than 500 characters. The 26 affected collar records (941,143

$\rightarrow$ 941,117 in the two study regions) are dropped, not repaired, because the remaining fields of a desynchronised row cannot be trusted.

(2) Hole-level quality filter. Missing `to_depth` values are first filled from the next interval's `from_depth`. Intervals that still lack a base or have non-positive thickness are then discarded. A hole is admitted only if it satisfies all four conditions: (i) maximum depth $\geq 30\,\mathrm{m}$; (ii) at least five logged intervals; (iii) substantive lithology text on at least 50% of intervals; (iv) at least 30% of intervals mapped to the vocabulary. Conditions (i)–(ii) exclude shallow reconnaissance drilling that is too short to form a stratigraphic sequence and gives an autoregressive model insufficient context. For (iii), text is non-substantive if it is empty, purely numeric (e.g. recovery percentages entered in the lithology column), a single character, or a bare modifier such as `Fg`. Such text passes a naive non-null check but carries no lithology. Condition (iv) keeps logs that are mostly unreadable by the vocabulary out of the benchmark. The filter is the dominant reduction (Table 4) and is not uniform across regions: Pilbara iron-ore exploration relies heavily on shallow rotary-air-blast drilling, so a larger share of its holes fails (i)–(ii).

(3) Coordinate harmonisation and unlocatable holes. Every DrillBench split is defined by spatial distance, so a hole without valid collar coordinates cannot be assigned to a split; we exclude 27 such holes (261 intervals), all in the Eastern Goldfields. For the remaining holes, the coordinates as reported are not mutually comparable: companies file eastings and northings in MGA zone 50 or 51 and in either GDA94 or the older AGD84 datum, whose grids differ by about $211\,\mathrm{m}$ (24,802 holes, 49%, are filed in AGD84). We therefore re-derive eastings and northings from the GDA94 longitude and latitude held by the source database, in one zone per region (Eastern Goldfields EPSG:28351, Pilbara EPSG:28350), and keep the reported values in separate columns. All distances used for splitting, de-duplication and retrieval are computed in this frame.

(4) Lithology vocabulary. Free-text descriptions are mapped to the 37-code vocabulary (Table 7) by exact lookup followed by ordered regular-expression rules. On the 1,533,265 retained intervals, 68.2% receive a code. The remaining 31.8% break down as follows: unmapped raw text 28.5%, spread over 9,988 distinct strings, dominated by project-specific shorthand such as `TAC`, `OTA`, `CPZ`; modifier-only text 2.2%; blank 0.9%; explicit "unknown" 0.3%. Unmapped intervals keep their raw text in `lith_raw` with a null code and are treated as unobserved in every model input and metric. The per-split lithology coverage in Table 1 therefore measures available supervision, not logging completeness.

(5) Physical-bounds masking. Assays are harmonised to 25 elements: Au in ppb, trace elements in ppm, major oxides in wt%, with Al reported as $0.5293 \times \mathrm{Al_2O_3}$. A value whose magnitude exceeds its element's stoichiometric bound (Table 5) is set to `null`; the rest of the record is kept, so masking removes no interval. Over the 70.96M source assay rows, 409,821 values are nulled. Negative values within bounds are kept. By Western Australian reporting convention a negative encodes a below-detection-limit result: `As_ppm`$= -3000$ means $< 3000\,\mathrm{ppm}$. Downstream code decodes this convention at use time, and users of the geochemistry channel must do the same.

(6) Twin-hole de-duplication. Statutory reporting means the same physical hole is often filed more than once, each time under a different `hole_id` and project name. A hole-level split would then place near-copies on both sides and flatter any method that predicts from neighbouring holes. From each evaluation split we remove every hole that has a *twin* in its reference pool: a hole within $25\,\mathrm{m}$ collar distance whose full lithology sequence is identical, or whose interval count matches and whose depth grid agrees within $0.11\,\mathrm{m}$. Table 6 breaks the removals down by split. TRAIN, both spatial OOD splits, and PILBARA-TEST are unchanged; a separate scan against TRAIN finds no twins in the OOD splits. Distinct holes within $25\,\mathrm{m}$ of each other are deliberately kept, since they are legitimate evidence for interpolating methods. The twin search is run on the harmonised coordinates: in the coordinates as reported, a hole filed once in AGD84 and once in GDA94 lies about $211\,\mathrm{m}$ from its own copy and escapes a $25\,\mathrm{m}$ radius, which hid 198 verbatim copies (69 in VAL, 129 in TEST). Three further TEST holes are removed because an identical log of at least three layers, on the same depth grid, is filed under a *different* report within $500\,\mathrm{m}$ of a training hole (re-surveyed collars); they are counted under "same lith." in Table 6.

Table 5: Per-element physical upper bounds and number of values nulled.

Element	Bound	Nulled	Element	Bound	Nulled
Au (ppb)	10^6	22,090	Fe (%)	72	16,663
Ag (ppm)	50,000	11,107	Al (%)	30	29,843
As (ppm)	200,000	643	SiO_2 (%)	100	53,650
Sb (ppm)	50,000	546	MgO (%)	60	32,736
Bi (ppm)	50,000	4,146	CaO (%)	60	28,301
Cu (ppm)	300,000	1,093	Na_2O (%)	20	30,536
Pb (ppm)	300,000	712	K_2O (%)	20	29,588
Zn (ppm)	300,000	2,845	TiO_2 (%)	30	28,153
Ni (ppm)	100,000	19,058	P_2O_5 (%)	50	24,359
Co (ppm)	50,000	4,275	LOI (%)	50	30,745
Cr (ppm)	200,000	1,678	S (%)	60	18,395
V (ppm)	10,000	782			
Ba (ppm)	100,000	3,341			
Sr (ppm)	10,000	14,536		**Total**	**409,821**

Twins in the archive. Exclusion by `hole_id` does not catch copies filed under a different identifier, and the 25 m retrieval guard does not catch re-surveyed collars. We therefore scan every evaluation split against the full WA-DASC-full archive, whose collars are re-derived from GDA94 longitude and latitude in the same frame as DrillBench. An archive hole within 500 m is a twin if its raw lithology text sequence or its mapped sequence is identical over at least three layers, or if its interval count matches and its depth grid agrees within 0.11 m; raw text is compared because only about 15% of archive intervals map to our vocabulary. Table **??** reports the matches. All matched archive holes are removed from both the pretraining corpus and the retrieval pool before any result in this paper is computed; the benchmark splits are unchanged.

Table 6: Twin-hole de-duplication by evaluation split.

		Holes		Removed by rule	
Split	Reference pool	before → after	removed	same lith.	same grid
VAL	TRAIN	2,286 → 1,951	335	292	43
TEST	TRAIN	4,572 → 3,856	716	654	62
Total			1,051	946	105

B.2 Lithology Vocabulary Details

Lithology labels are standardised to a single vocabulary of 37 codes (Table 7), which is used for training, prediction, and evaluation throughout. Free-text descriptions from all data providers are mapped into these codes by exact matching followed by regular-expression rules (§B.1), so that holes logged by different companies and geologists become directly comparable. Intervals whose text cannot be mapped keep their original wording in `lith_raw` but carry no code, and are treated as unobserved rather than assigned a fallback class. `<UNK>` denotes missing or unmapped lithology and is excluded from training and evaluation; it is distinct from `UNKN`, which is a valid code meaning "undifferentiated." Cell-level tokenisation additionally uses `<CONT>` for layer continuation (§3.1), and layer-level pretraining uses `GAP` for unlogged runs (§5.2). The COVER/BEDROCK assignment is used only for analysis, not for training or prediction. It follows the regolith–bedrock distinction of Anand & Butt (2010): COVER spans transported material (alluvium, colluvium, gravel, sand) and the in-situ weathering profile (laterite, saprolite, clay), whereas BEDROCK covers fresh lithologies. Saprolite and clay sit at the boundary, since both are weathering products of the underlying rock; we place them in COVER because they are logged by weathering state rather than protolith. `UNKN` carries no zone and is excluded from the zone analysis.

Table 7: **Shared lithology vocabulary.** The 37 fine-grained codes (ids 0–36) form the prediction target for both tasks and are shared across data providers. For the cover/bedrock analysis (§6, Table 12) each code is additionally assigned to one of two zones: COVER groups transported and weathered material above fresh rock, while BEDROCK groups the in-situ units relevant to primary mineralisation. **Fill** denotes the colour used in the case-study lithology strips; hatching (/ / /) marks incorrect predictions. Two ids are reserved: <UNK> (37) for missing or out-of-vocabulary lithology and <CONT> (38) for layer continuation. UNKN (35) is a valid fine code for "undifferentiated" and is distinct from <UNK>; both are shown in grey and excluded from scoring.

id	Fill	Code	Description	id	Fill	Code	Description
COVER — transported and weathered material (9 codes)							
0		ALLU	Alluvium	25		REGL	Regolith / overburden
7		CLAY	Clay	27		SAND	Sand
8		COLL	Colluvium	28		SAPL	Saprolite / saprock
17		GRVL	Gravel	34		TOPS	Transported overburden
20		LATR	Laterite / ferricrete				
BEDROCK — in-situ rock (27 codes)							
1		AMPH	Amphibolite	18		IRON	Ironstone / BIF
2		ANDS	Andesite	19		KOMA	Komatiite
3		BSLT	Basalt	21		PEGM	Pegmatite
4		CARB	Limestone / dolomite	22		PHYL	Phyllite
5		CGLM	Conglomerate / breccia	23		PORP	Porphyry
6		CHRT	Chert	24		QRTZ	Quartzite
9		DACI	Dacite	26		RHYO	Rhyolite
10		DIOR	Diorite	29		SCHI	Schist
11		DOLR	Dolerite / dyke	30		SDST	Sandstone
12		GABR	Gabbro	31		SERP	Serpentinite / talc
13		GNSS	Gneiss	32		SHLE	Shale / siltstone
14		GRAN	Granite	33		SKAR	Skarn
15		GRND	Granodiorite	36		VEIN	Quartz vein
16		GRNS	Greenstone				
Unzoned and reserved							
35		UNKN	Unknown / undifferentiated	38	–	<CONT>	layer continuation
37		<UNK>	out-of-vocab / missing				

C Implementation Details

C.1 Baseline Implementation Details

We include both cell-level and layer-level statistical baselines. Cell-level methods are shown in *light blue*, while layer-level methods are shown in *light orange*. The first five methods operate on the 1 m cell sequence, while the final two operate directly on stratum-level sequences.

C.1.1 Frequency and Markov Models

We include both cell-level and layer-level frequency- and Markov-based baselines. Cell-level methods operate on the 1 m sequence and are marked in *light blue*, while layer-level methods operate on stratum sequences and are marked in *light orange*. Unless stated otherwise, count tables use add-one smoothing. UNK is excluded from count estimation and receives only smoothing mass.

Unigram predicts the marginal distribution of the 37 lithology classes in TRAIN, without using sequence context.

n-gram uses an order-3 model,

$$P(\ell_t \mid \ell_{t-1}, \ell_{t-2}),$$

with interpolated back-off. For a context observed N_{ctx} times, the higher-order estimate receives weight

$$\lambda_{\text{ctx}} = \frac{N_{\text{ctx}}}{N_{\text{ctx}} + \gamma}, \qquad \gamma = 8,$$

and the remaining weight is assigned to the next lower-order model. This is applied recursively from trigram to bigram to unigram, so rare contexts rely more on shorter histories.

Depth-Markov uses a bigram additionally conditioned on depth,

$$P(\ell_t \mid d_t, \ell_{t-1}),$$

where d_t is the 10 m depth band containing the interval top. We use 20 bands from 0–200 m and one overflow band for deeper intervals. If a (d_t, ℓ_{t-1}) context is unseen, the model backs off as

$$P(\ell_t \mid d_t, \ell_{t-1}) \to P(\ell_t \mid d_t) \to P(\ell_t).$$

Persistence assigns probability $1 - \epsilon$ to the previous cell's lithology and distributes the remaining probability ϵ over the other classes, with $\epsilon = 0.01$. At the first cell of a hole, or when the previous cell is UNK, it falls back to the training marginal. Since only about 1.5–2.4% of cells correspond to stratum boundaries, this provides a simple reference for prediction on the 1 m grid.

HMM is a discrete-emission hidden Markov model with $K = 24$ latent states. It is trained with Baum–Welch on a fixed-seed subsample of 8,000 training holes for at most 25 EM iterations, with convergence tolerance 10^{-3} and a self-transition bias of 5.0. UNK observations are treated as missing by assigning them uniform emission likelihood across states, so they do not contribute to emission-count updates. Evaluation uses all holes in each split.

Layer Markov operates directly on stratum lithologies. It uses an order-2 transition model with hierarchical $2 \to 1 \to 0$ back-off. Each higher-order distribution is smoothed toward the next lower-order distribution using a pseudo-count prior of weight $\kappa = 5$, while the context-free marginal uses add-one smoothing. To convert layer predictions to the common 1 m grid, the model also estimates a discrete thickness hazard,

$$P(T = r \mid T \geq r),$$

where T is stratum thickness in cells and r is the current run length. A separate hazard is estimated for each lithology and shrunk toward the global hazard with weight $\kappa = 5$.

Layer marginal removes sequence dependence. Each new stratum lithology is sampled from the training marginal, while its thickness is generated from the global thickness hazard. It serves as a Task 2 reference for distinguishing sequence generation quality from simply matching overall lithology frequencies.

C.1.2 GEOSTATISTICAL MODELS

The geostatistical baselines fall into two groups according to where their information comes from. *Neighbour-based methods* retrieve nearby logged drillholes at prediction time, whereas *region-level methods* estimate transition or pattern statistics from TRAIN and use no query-side neighbours.

Retrieval pool. To make the fair comparison, all neighbour-based baselines use the same retrieval pool as our model (§5.3): DrillBench TRAIN and WA-DASC-full, with the archive twins of every evaluation hole removed. Retrieval is restricted to the same geological region: Eastern Goldfields queries use TRAIN plus archive holes (pretraining corpus). Thus, these methods differ from our retriving approach in how they use retrieved evidence, not in which drillholes are available.

Coordinate frame. DrillBench collars are stored in local MGA zones: TRAIN, VAL, TEST, and OOD-REGIONAL use zone 51, whereas OOD-DISTAL and the Pilbara splits use zone 50. All collars are converted to a common metric coordinate system before computing distances.[2]

Duplicate guard and missing evidence. As in our receip, any pool hole within $25\,\mathrm{m}$ of the query collar is excluded to avoid re-reported copies of the same physical drillhole. At a queried depth, neighbours that do not reach that depth provide no vote. If no retrieved neighbour covers the queried position, the prediction falls back to the TRAIN marginal.

Neighbour-based methods.

[2]The zone assignment was verified using 5,038 TRAIN holes shared with the archive and by checking nearest-archive distances for the remaining splits. The correct transformation places holes in their expected provinces, whereas using the wrong zone produces displacements of tens to hundreds of kilometres.

Spatial k-NN retrieves the $k = 10$ nearest eligible collars in the same region. At each query depth, the lithology of every available neighbour is obtained by interval lookup, and class probabilities are formed using inverse-distance weighting.

Indicator kriging uses the same depth-wise lookup with $k = 20$ neighbours, but replaces inverse-distance weighting with simple indicator kriging. We use a spherical variogram with range $1\,\mathrm{km}$, nugget 0.05, and sill 0.95, and blend the resulting estimate with the TRAIN marginal using $\alpha = 0.05$. As neighbours move beyond the variogram range, their contribution decreases and the prediction approaches the marginal.

Sequential indicator simulation (SIS) uses the same kriging conditional distribution but draws M realisations and reports their empirical class distribution. It therefore provides a stochastic counterpart to indicator kriging, suitable for Task 2 where complete stratigraphic continuations must be generated.

Retrieval-only readout isolates the information contained in our retrieval vector r_i by removing the autoregressive sequence model. It uses the same retrieval pool, k, duplicate guard, and depth alignment as ours. We evaluate both a parameter-free readout, which predicts the arg-max of r_i, and a readout learned on TRAIN layer-start cells. Their accuracies differ by less than one percentage point, indicating that performance is mainly limited by the retrieved evidence rather than the readout.

When collar elevation is available, neighbour lithology is aligned by elevation rather than down-hole depth. Because the public archive does not provide collar elevation, this adjustment has no effect when all retrieved evidence comes from archive holes; the same limitation applies to our receip.

Region-level methods.

These methods use only TRAIN sequences when fitting their statistics and do not access the retrieval pool at prediction time.

TPROGS separates persistence within a lithology from transitions between lithologies. For each region, it estimates

$$p_{\mathrm{self}}(c),$$

the probability that lithology c continues into the next cell, and an embedded transition matrix

$$T_e(c \to c'),$$

which gives the probability of transitioning from c to a different lithology c' once a boundary occurs. This separates stratum run length from lithology-transition structure.

MPS-lite is a vertical approximation of multi-point statistics. The regional TRAIN sequences form the training patterns. At position t, the model matches the preceding $W = 5$ tokens and uses the empirical distribution of the following token as its prediction. Sparse patterns progressively back off to shorter contexts using

$$\lambda_{\mathrm{ctx}} = \frac{N_{\mathrm{ctx}}}{N_{\mathrm{ctx}} + \gamma}, \qquad \gamma = 8,$$

where N_{ctx} is the number of training occurrences of the current pattern. This tests whether a longer local pattern provides useful information beyond the shorter contexts captured by the n-gram baseline.

C.1.3 Feature-based Machine Learning

XGBoost represents the feature-based machine-learning baseline. It uses gradient-boosted decision trees with only information available at prediction time: z-scored collar coordinates, region one-hot features, depth, current run length, and previous lithologies. We train 200 boosting rounds with maximum tree depth 6 and learning rate 0.1. Training uses early stopping if validation performance does not improve for 20 consecutive rounds.

C.1.4 Cell-level autoregressive models.

All five neural baselines operate directly on the 1 meter token sequence and use only lithology history as input; collar information, geochemistry, and auxiliary objectives are excluded. They are

trained with teacher forcing and cross-entropy, so the main difference between them is how past tokens are represented and combined.

LSTM uses a 64-dimensional token embedding, 256 hidden units, 2 recurrent layers, and dropout 0.1. It is trained for 20 epochs with AdamW using learning rate 2×10^{-3}, weight decay 10^{-4}, and batch size 64. The recurrent hidden state summarises the sequence observed above the current cell.

GRU uses the same embedding size, hidden size, number of layers, dropout, optimiser, and training schedule as the LSTM, replacing only the LSTM recurrent unit with a GRU. This provides a controlled comparison between the two recurrent architectures.

Transformer replaces recurrence with causal self-attention. It uses $d_{\text{model}} = 192$, 6 attention heads, 6 layers, feed-forward width 768, sinusoidal positional encodings, context length 2048, and dropout 0.1. It is trained for 30 epochs with AdamW using learning rate 3×10^{-4}, batch size 32, and a warm-up followed by inverse-square-root decay.

TCN uses six residual blocks, each containing two dilated causal convolutions with kernel size 3. The dilation factors are $1, 2, 4, 8, 16, 32$, giving a receptive field of 253 cells. Each block uses GELU, dropout, a residual connection, and LayerNorm. The hidden width and optimisation schedule match the Transformer.

Mamba uses $d_{\text{model}} = 192$ and 8 selective state-space blocks, with state size 16, convolution width 4, and expansion factor 2. The blocks are arranged in a pre-norm residual stack and trained with the same optimisation schedule as the Transformer.

For all five models, we retain the checkpoint with the lowest negative log-likelihood on VAL.

C.1.5 Our layer-level model.

The body is deliberately small: $d_{\text{model}} = 192$ with a 2-layer GRU or a 3-layer, 4-head Transformer, at most 128 strata per hole, 0.65–1.6 M parameters. Thickness and depth are not fed back as inputs, since both are province fingerprints; the inputs are the lithology embedding and, when retrieval is enabled, $\phi(r_i)$. Training runs for a fixed 20 epochs with AdamW (learning rate 3×10^{-4}, weight decay 0.01, batch size 256) and label smoothing 0.3 on the lithology head; the thickness head is trained by Gaussian negative log-likelihood on $\log D$ with weight 0.3. No checkpoint selection is performed, so VAL is never used for model choice. Retrieval draws $k = 10$ neighbours from the pool, excludes neighbours closer than 25 m to the query collar (a duplicate-report guard), aligns the look-up by elevation rather than hole depth, and summarises the neighbours into a 37-class inverse-distance-weighted lithology distribution; the whole retrieval vector is dropped for 30% of training holes. Pretraining runs on the deduplicated multi-province corpus, layerised with an explicit `GAP` token for unlogged runs, over a lithology vocabulary plus `GAP` and `UNK`; batch size 512, learning rate 3×10^{-4}, label smoothing 0.3, 24 epochs for the checkpoint used in the main tables, with auxiliary thickness, masked-token, and chemistry-anchor losses weighted 0.3, 0.5 and 0.5.

Evaluation and reporting. Layer-level rows of Table 2 are run with ten seeds, other trained arms with three, and single-run baselines once; we report the mean. Task 1 reports macro-F1 at layer-start cells, with top-1 as a secondary metric, and 95% confidence intervals from a hole-level cluster bootstrap; Task 2 reports the joint KL of free-running rollouts under the protocol of §C.2.

Compute. All runs are single-GPU on a SLURM cluster (V100, A6000 or H100 depending on availability). A layer-level fine-tuning run takes 3–4 minutes, a cell-level run with retrieval about 17 minutes, and corpus pretraining 20 minutes (GRU) to 7 hours (12-layer Transformer). The complete set of results in this paper is roughly 400 GPU-hours, dominated by the pretraining checkpoints and the data-size sweep. Cross-province numbers are sensitive to the CPU/GPU split on OOD-DISTAL, so every paired comparison is run on the same device type.

C.2 Evaluation Protocol Details

Unified output space. The evaluated methods predict different objects. Cell-level methods directly produce a distribution over the next 1 m token, whereas layer-level methods predict the lithol-

ogy and thickness of the next stratum. To make them directly comparable, every method is converted to the same per-cell distribution over $\{\texttt{CONT}\} \cup \{1, \ldots, 37\}$.

For a layer-level model, let $q(k)$ denote the probability that the next stratum has lithology k, and let λ denote the probability that the current stratum ends at the current cell. Its equivalent cell-level distribution is

$$P(\texttt{CONT}) = 1 - \lambda, \qquad P(k) = \lambda q(k).$$

Cell-level methods already produce this distribution directly. All methods are therefore scored by the same evaluation code on the same 1 m grid.

Task 1: next-layer lithology prediction. Task 1 asks which lithology begins at the next true stratum boundary. Given the true history above a boundary, we evaluate the predicted distribution at the first cell of the new stratum and report macro-F1, averaged with equal weight over the lithology classes that have at least 20 true next-stratum events in the split. CONT is excluded from the argmax because the boundary location is already given in this task.

We do not use ordinary per-cell accuracy as the primary metric because about 96% of logged cells are continuations and new strata start at only 1.2–2.6% of cells. A model that always predicts CONT, i.e. that the current stratum simply continues, therefore reaches about 96% per-cell accuracy without learning any geological transition. Evaluating only true stratum starts instead isolates the intended question: *which rock comes next?*

We additionally report top-1 accuracy, balanced accuracy, support-weighted F1 and Cohen's κ as secondary metrics (App. D.1); on OOD-R and OOD-D, where single exploration reports hold hundreds of holes, top-1 is macro-averaged over reports. Confidence intervals are 95% bootstrap intervals (B=1,000) that resample drillholes rather than individual cells, because observations within a hole are strongly correlated; trained models additionally resample seeds.

Task 2: free-running generation. Task 2 evaluates whether a model can continue a partially observed drillhole using its own predictions. Each model receives the true upper 50% of a hole, then samples one cell at a time until the known total depth is reached.

A generated sequence should not be compared with the true suffix cell by cell: two plausible stratigraphic columns may contain similar lithologies and thicknesses but place a boundary a few metres apart. We therefore evaluate the distribution of generated strata rather than exact cell alignment.

Our primary metric is the KL divergence $\mathrm{KL}(P_{\text{true}} \,\|\, P_{\text{gen}})$ between the true and generated joint distributions of

$$(\text{lithology}, \text{thickness bin}),$$

pooled over all evaluated holes and rollouts. Lithology uses the 37 DrillBench classes. Thickness is discretised into seven log-spaced bins with edges

$$1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 31 \text{ m},$$

with the final bin absorbing thicker strata. Before counting, both true and generated suffixes are converted from cells back into strata by run-length encoding, and strata of unknown lithology are dropped. KL divergence uses $\varepsilon = 10^{-6}$ numerical smoothing.

The joint metric is used because lithology and thickness marginals can each be matched independently while still producing unrealistic stratigraphy. For example, a model may reproduce the correct lithology frequencies while fragmenting units into unrealistically thin layers, or match the thickness distribution while assigning it to the wrong rocks.

Generation protocol and calibration. All generators use the same true 50% prefix, sample autoregressively from their full predicted distribution, and stop at the logged hole depth. We draw five rollouts per hole on a fixed, report-stratified sample: each exploration report contributes at most c randomly drawn holes, with c the smallest cap that yields at least 600 holes, giving 610–827 holes per split. The sample is drawn once, independently of any model, so every method is scored on the same holes. Confidence intervals resample whole reports (B=1,000), because holes of one report share a project and a logging vocabulary.

Because some generators naturally over- or under-segment, we fit one scalar calibration factor s on TRAIN rollouts for every method:

$$P(\texttt{CONT}) = 1 - s\lambda, \qquad P(k) = s\lambda q(k),$$

with $s\lambda$ clipped to $[10^{-4}, 0.99]$. The factor rescales the layer-start probability so that the generated boundary rate matches the training data. The same procedure is applied to all generators, and s is never fitted on validation or test data. Calibration is used only for Task 2.

Fair comparison. All methods are converted to the same per-cell output space and scored by the same implementation. All generators receive the same prefix, stopping depth and hole sample, and the same train-only boundary calibration is applied to every generator. The layer-level neural models, whose lithology heads are trained with label smoothing, additionally sharpen q with a temperature selected on TRAIN rollouts; the other methods receive no such temperature. Spatial k-NN is marked with * in Task 2 because it is a depth-wise lookup rather than an autoregressive generator.

D Additional Results

D.1 Additional Results: Next-Layer Prediction

D.1.1 Protocol and uncertainty

All numbers are computed from per-event logs (hole, true next lithology, predicted next lithology) written by the shared evaluation harness, so every method is scored on exactly the same layer starts. Macro-F1 is averaged over the lithology classes with at least 20 events in the split. Top-1 is pooled on TEST and PILBARA-TEST and macro-averaged over exploration reports on the two OOD rings, where single reports hold up to several hundred holes. Intervals are 95% hole-cluster bootstrap intervals (B=1000): all layer starts of a hole are resampled together, and rows trained with several seeds (10 for the five layer-level rows of Table 2, 3 for every other trained row) additionally resample seeds. One resample per split is shared by all rows, so the difference between any two rows is a *paired* bootstrap. The unpaired intervals are wide on macro-F1 ($\pm$3–5 points, because rare classes carry equal weight); the evidence for every claim in §6 is therefore the paired difference (Table 10).

D.1.2 All baselines

Table 9 extends Table 2 to all 37 methods. *Frequency and Markov models at the cell level* never exceed 10% macro-F1: under the CONT tokenisation they cannot copy the previous metre, and a first-order chain carries no stratigraphic memory. *Spatial methods* are strong only where training holes are within reach: k-NN, indicator kriging and SIS reach 66–72% macro-F1 on TEST and fall to 1–5% on every shifted split when restricted to the training pool; giving them the same archive pool that our retrieval uses lifts them only to 6–12%, so the advantage of the proposed recipe is not access to a larger pool. *Directly conditioned neural baselines* expose the transfer boundary: adding location to an LSTM raises TEST top-1 from 60.1 to 80.5 and lowers PILBARA-TEST from 38.5 to 22.6; for a GRU the same input gives 79.6 in distribution and 3.0 across provinces. Geochemistry alone is milder (67.7 and 23.2 for the GRU), and the two together are the strongest in-distribution models of the benchmark (up to 81.8 top-1) and among the weakest under cross-province transfer. Spatial baselines and the collar-conditioned networks were re-fitted on the coordinates; sequence-only baselines are unaffected by the coordinate revision because the training split is unchanged.

D.1.3 Paired differences

Table 10 scores both methods of each row on the same bootstrap resample of holes. *Components.* Relative to the plain layer-level GRU, which is trained with label smoothing like every arm, retrieval adds +10.7 macro-F1 points on TEST and +5.6 on OOD-D, and nothing measurable on OOD-R (-0.1) or PILBARA-TEST (-1.3); both intervals include zero. Pretraining shows the complementary profile: no effect in distribution (-0.9, not significant) and +1.8, +4.7 and +3.5 on the three shifted splits. The PILBARA-TEST gain from pretraining does not come from Pilbara holes in the archive: it persists when the whole province is removed from the pretraining corpus (App. D.4.2). Combined, the full recipe improves on the plain GRU on all four splits, on layer Markov, on the layer-level Transformer and on the cell-level Transformer, with every interval excluding zero; against

the strongest spatial baselines it is on par in distribution and $+23$ to $+36$ points ahead under shift. On OOD-R and PILBARA-TEST the full recipe is numerically below GRU + PT in macro-F1 (38.8 vs. 39.2 and 30.6 vs. 31.7), but the paired differences (-0.4 and -1.1) are not significant: once the model is pretrained, retrieval neither helps nor hurts there. Under top-1, balanced accuracy, support-weighted F1 and Cohen's κ the full recipe is the best row on all three shifted splits (e.g. $\kappa\times100 = 45.8/40.3/37.6$ vs. $32.5/29.9/34.6$ for GRU + PT); retrieval thus helps mostly on frequent lithologies, and pretraining is what moves the rare ones. *The transfer boundary.* The last rows make the wall a paired statement on identical holes: adding location to the cell-level GRU changes macro-F1 by $+14.7$ on TEST and by -8.4, -19.9 and -22.7 on OOD-R, OOD-D and PILBARA-TEST; geochemistry by $+6.8$ and -1.4, -5.3, -12.2; an LSTM shows the same sign pattern. Every interval except geochemistry on OOD-R excludes zero, so the reversal is not a property of one backbone or of seed noise.

D.1.4 BACKBONES AND TOKEN GRANULARITY

Table 11 repeats the four arms for five layer-level backbones, each with a label-smoothed plain arm. The pattern of Figure 4(c) holds for each of them: in macro-F1, retrieval adds 10–11 points on TEST and nothing measurable on PILBARA-TEST (-2.0 to $+1.7$), while pretraining adds nothing on TEST and $+2.1$ to $+4.2$ on PILBARA-TEST. With the full recipe GRU, LSTM, TCN and Transformer are within a few points of each other; Mamba is the weakest. We use the GRU in the main text because it is the most even across the two tasks; the conclusions do not depend on this choice. All backbones share one learning rate (3×10^{-4}). The same recipe applied to a *cell-level* Transformer behaves differently: retrieval transfers ($+10.3$ macro-F1 on TEST), but pretraining lowers macro-F1 on three of the four splits ($60.3 \to 53.3$ on TEST, $23.6 \to 19.4$ on OOD-D, $22.5 \to 19.1$ on PILBARA-TEST), and the full cell-level recipe is 10–14 points below the layer-level one on the three shifted splits. At one-metre resolution most targets are `CONT`, so the pretraining signal is dominated by thickness rather than by lithological succession; the layer-level formulation removes that imbalance, which is why the recipe is built on it.

D.1.5 COVER AND BEDROCK

Table 12 gives the numbers behind Figure 4(d). Retrieval is a regolith mechanism: within the Eastern Goldfields its gain in cover is $+16.0$, $+15.3$ and $+10.9$ points, against $+13.6$, $+0.7$ and $+4.9$ in bedrock, and it all but vanishes across provinces ($+0.8$ in both zones). Pretraining is the mirror image on the two splits where bedrock differs most from training: $+1.4$ vs. $+0.7$ on OOD-R and $+2.7$ vs. $+1.2$ on PILBARA-TEST, with no effect in distribution. OOD-D is the exception and we report it as such: there the pretraining gain is larger in cover ($+4.3$) than in bedrock ($+0.5$); the distal ring is 77% regolith by scored layer starts, and a plausible reading is that its weathering profiles differ from those of the training camps, so that the archive's regolith sequences are what transfers; we have not tested this directly.

D.1.6 ERROR ANALYSIS

Figure 6(a) shows that the PILBARA-TEST margin over layer Markov is carried by bedrock classes that are rare in the Eastern Goldfields training region: ironstone/BIF ($+31$ recall points), greenstone ($+26$), shale ($+16$) and conglomerate ($0 \to 15$), together with saprolite ($+14$). Two classes remain unsolved by every method: transported overburden (TOPS) is 12% of Pilbara layer starts and essentially absent from the training region, and chert is absorbed by clay and the residual classes. On OOD-D (panel c) the dominant errors are within the regolith profile (alluvium $\to$ laterite 48%, regolith $\to$ clay 26%) and quartz vein $\to$ clay (82%), i.e. confusions between units that loggers themselves separate inconsistently, rather than failures of stratigraphic order. Panel (d) shows that the gain is broad: on OOD-D 54% of the scored layers sit in holes where the recipe beats layer Markov and 14% in holes where it loses; on PILBARA-TEST the split is 41% to 13%. By exploration report, the natural unit of a project, a logging geologist and a local vocabulary, the recipe beats layer Markov in 39 of 44 OOD-D reports (event-weighted margin $+17.3$ points); on PILBARA-TEST it wins 20 and loses 18 of 43 reports, and the $+7.9$-point margin is carried by a few large reports (up to $+41$ points; $+23$ on the largest, with 707 layer starts), while most losses are within 5 points. Cross-province transfer therefore improves the typical large project but is not yet uniform across reports.

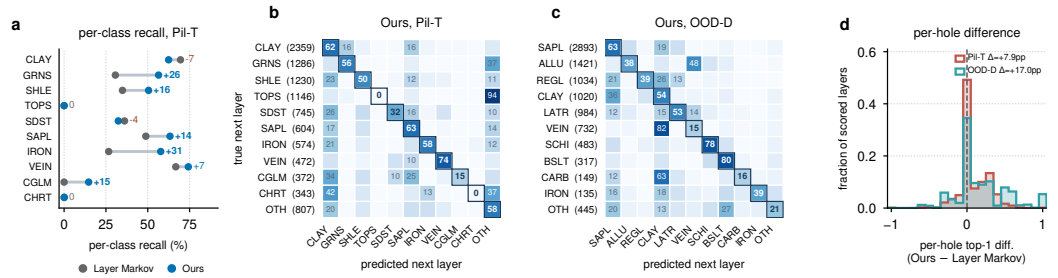

Figure 6: **Error analysis** (full recipe, 10 seeds pooled; row counts are per run). **(a)** Per-class recall on PILBARA-TEST for the ten most frequent true classes, layer Markov → ours. **(b, c)** Row-normalised confusion on PILBARA-TEST and OOD-D. **(d)** Distribution of the per-hole top-1 difference, ours minus layer Markov, weighted by scored layers.

D.2 Additional Results: Free-Running Generation

D.2.1 Protocol and uncertainty

Every method generates the lower half of the same report-stratified sample of holes (≈600 per split, drawn so that every exploration report is represented in proportion to its size rather than by file order), from the true upper half, with $M{=}5$ rollouts per hole. Generated and true suffixes are cut into layers and compared through their joint (lithology, thickness) law over the 37 codes and seven log-spaced thickness bins; the pooled forward KL of this law is the main-text metric (App. C.2 defines it and the secondary metrics). Because a divergence is a function of sums of per-hole histograms, it can be bootstrapped: intervals are 95% report-cluster bootstrap intervals ($B{=}1000$; all holes of a WAMEX report are resampled together, since they share a camp, a logger and a vocabulary), and trained rows additionally resample seeds. Plug-in divergences are biased upward under resampling, so intervals are re-centred on the full-sample estimate. One report resample per split is shared by all rows, so every difference in Table 14 is paired. The intervals are wide on OOD-D, where one camp holds more than half of the holes and the resample often drops or doubles it.

D.2.2 All baselines

Table 13 extends Table 2 to all 34 fitted baselines and reports, next to the joint KL, the bounded JSD of the same law and the per-metre agreement of the generated suffix with the hole's own log. Three observations. *Per-metre agreement rewards copying, not generation*: persistence (repeat the last lithology to the bottom) has the highest Acc/m on every shifted split (64.8/66.3/57.3) and the worst joint KL of the table (1.3–3.0), because it produces a single layer of the wrong thickness; indicator kriging on the archive pool likewise tops Acc/m on TEST (77.8) while its KL is 0.437. This is why the joint law, not agreement, is the headline metric. *Spatial lookups do not roll a hole forward*: k-NN and kriging read the neighbours' lithology at each depth and reach KL 1.4–2.0 on the two rings, above every sequence model; they are marked * and never highlighted. *The transfer boundary appears in generation as it does in prediction*: cell-level networks conditioned on location produce the most implausible rollouts of the benchmark once the coordinates are far from training (GRU + location KL 2.64 on OOD-D and 6.02 on PILBARA-TEST, against 0.56 and 1.00 for the same network without coordinates), whereas the unconditioned layer-level models degrade gradually (0.18 → 0.55 for the GRU). Among fitted baselines the best rows are layer Markov on TEST (0.150) and the 3-gram / TPROGS pair on the cell level; on the shifted splits the layer-level neural rows lead.

D.2.3 Paired differences

The paired differences separate what the recipe changes from sampling noise. The plain layer-level GRU is trained with label smoothing like every arm. Relative to it, neither component moves the joint KL in distribution ($|\Delta| \le 0.01$, intervals include zero), and on OOD-R only pretraining does (-0.053): the generated law of the Eastern Goldfields core is already matched by a sequence model

trained on it. On the two hardest splits neither component alone is distinguishable from the plain GRU. The full recipe is, on PILBARA-TEST (-0.092), but not on OOD-D (-0.057, interval includes zero). On PILBARA-TEST adding pretraining to the retrieval model (-0.103) and adding retrieval to the pretrained model (-0.093) are each significant, so the two components are complementary there: each helps only in the presence of the other. Against the layer-level Transformer the recipe is better on OOD-D (-0.087) and PILBARA-TEST (-0.086). Against layer Markov the recipe is 0.39 better on PILBARA-TEST and not distinguishable on the three Eastern Goldfields splits ($+0.036$, -0.063, -0.102, all intervals include zero; on JSD layer Markov is in fact significantly better in distribution, $+0.016$). Against the cell-level Transformer, XGBoost and the 3-gram the recipe is ahead on every shifted split with intervals that exclude zero. Two entries of Table 2 should be read with this: the best TEST row (layer Markov, 0.150) and the best OOD-R row (GRU + PT, 0.303) are within noise of the full recipe (0.186 and 0.322). The recipe's advantage is confined to OOD-D and PILBARA-TEST. There it is large against the fitted and cell-level baselines and smaller, but significant on PILBARA-TEST, against the label-smoothed layer-level networks.

D.2.4 BACKBONES

Table 15 repeats the four arms for five layer-level backbones (GRU and plain Transformer with 10 seeds, the rest with 3, so their intervals are wider). Every plain arm is trained with label smoothing. On TEST and OOD-R the four arms of every backbone lie within 0.16–0.36 and nothing separates them. On PILBARA-TEST the best augmented arm of every backbone is below its plain arm (plain 0.54–0.60, best arm 0.46–0.54), but which arm is best varies, and retrieval alone does not lower the KL for every backbone; against the label-smoothed plain arm most of these differences are within noise (Table **??**). With three seeds the ordering of the two augmented arms is not stable: LSTM + PT ($0.403/0.494$ on OOD-D / PILBARA-TEST) and TCN + PT (0.345 on OOD-D) beat their own full recipe, and Transformer + PT + R is the best OOD-R row of the table (0.275). For generation we therefore claim no per-component gain that holds for every backbone. We use the GRU in the main text because it is the most even across the two tasks. All backbones share one learning rate (3×10^{-4}); Mamba at its own larger rate diverged for seed-determined runs, and the shared rate removes the divergence without changing its scores.

D.2.5 OTHER METRICS

Re-scoring the main-table methods under the joint law at 13 coarse classes, under JSD, under a report-balanced KL in which every report contributes one normalised histogram, and under three per-hole metrics (Acc/m, normalised layer-sequence edit distance, boundary F1 at ±1 m) changes no conclusion. The full recipe is the best row on OOD-D and PILBARA-TEST under KL, KL-13, JSD and the balanced KL (e.g. KL-13 $0.225/0.196$ vs. $0.267/0.385$ for layer Markov); layer Markov keeps TEST under every divergence, and OOD-R goes to GRU + PT under KL, JSD and the balanced KL and to the full recipe under KL-13. The per-hole metrics are flat across all layer-level rows (Acc/m 56–65, edit distance 0.58–0.63, boundary F1 10–12%) and are not what separates methods: a free-running rollout is not expected to reproduce the exact boundaries of one hole, only the layer statistics of the split, which is what the divergences measure. The price is that the pooled law does not score the fit of each continuation to its own prefix: in distribution a prefix-agnostic generator comes close (layer marginal 0.211 vs. 0.186), so the metric discriminates mainly under shift, and all generators are also given the logged stopping depth; a per-hole likelihood of the true suffix would complement it.

D.3 ZERO-SHOT LLM BASELINES: IMPLEMENTATION AND PROMPTS

This appendix documents how the zero-shot large-language-model (LLM) rows are produced: which points are scored, the exact text every model receives, how the answer is read out, and how each model was run. All providers import one prompt builder and one scoring routine, so every model sees byte-identical text for the same prediction point and is scored by the same rule as every trained method.

D.3.1 PIPELINE

1. **Hole sampling.** For each split we draw 150 holes with at least two intervals by a seeded random permutation (seed 0). Holes are stored project by project, so a prefix slice would cover only a few prospects.
2. **Prediction points.** A point is a cell that starts a layer and whose own token is a real lithology, i.e. neither `CONT` nor unmapped. This is the Task 1 mask used for every other method. The 750 sampled holes yield 2,161 points (Table 16).
3. **Context.** The model sees only the layers that start strictly above the point, top to bottom. Unmapped layers are shown as `UNLOGGED`; 46.8% of context layers carry this placeholder. Points with no layer above them (215 of 2,161) are cold starts and receive a dedicated sentence.
4. **One request per point.** Showing a whole hole and asking for every layer would leak labels, because the later layers of a hole are the targets of its earlier points. Each point is therefore an independent request.
5. **Read-out.** Open-weight models are scored by candidate likelihood; hosted models return a ranked top-3 in JSON.
6. **Scoring.** Top-1 and top-3 over the 37 lithologies, with per-hole hit/point tallies so that any LLM row can be paired with any other method on the same holes by the hole-cluster bootstrap. Layer-start AUROC and per-cell top-1 are left empty: the models are never asked where a layer ends, and deriving a boundary probability from the true layer starts would score them on the answer key.

D.3.2 PROMPTS

System prompt. Shared by every point of an arm. The full text of the **strict** arm is reproduced below; the code list is one code per line in the actual prompt and is abbreviated here.

System prompt (strict arm)

```
You are logging a mineral exploration drill hole. Given the
lithological units already logged from the surface downward, name the
single most likely lithology of the NEXT unit below.

Answer with codes from this list only (37 options):
  ALLU alluvium
  AMPH amphibolite
  ANDS andesite
  ...  (one line per code, 37 lines)
  UNKN unknown / unlogged
  VEIN vein

Units shown as UNLOGGED are intervals whose driller's code could not
be mapped to this list. They are gaps in the record, not a rock type:
reason across them to the rock that most likely lies below.

Return your three best guesses, most likely first.
```

The 37 codes and their glosses are: ALLU alluvium, AMPH amphibolite, ANDS andesite, BSLT basalt, CARB carbonate, CGLM conglomerate, CHRT chert, CLAY clay, COLL colluvium, DACI dacite, DIOR diorite, DOLR dolerite, GABR gabbro, GNSS gneiss, GRAN granite, GRND granodiorite, GRNS greenstone, GRVL gravel, IRON iron formation / BIF, KOMA komatiite, LATR laterite, PEGM pegmatite, PHYL phyllite, PORP porphyry, QRTZ quartz, REGL regolith, RHYO rhyolite, SAND sand, SAPL saprolite, SCHI schist, SDST sandstone, SERP serpentinite, SHLE shale, SKAR skarn, TOPS topsoil, UNKN unknown / unlogged, VEIN vein. Spelling out each code removes the possibility that a model fails only because it does not recognise an abbreviation. The `UNLOGGED` clause separates a gap in the record from `UNKN`, which is a logged category in the schema.

The other two arms insert one paragraph after the first sentence of the system prompt:

Inserted paragraph (geo and retrieval arms)

```
[geo] Each layer also carries its top depth and thickness in metres,
and you are told which province the hole is in.

[retrieval] You are also given what NEARBY drill holes logged at the
same depth. These are real observations from the same area,
inverse-distance weighted, grouped into broad rock classes. They are
strong evidence but not ground truth: when few neighbours reach the
depth, or they sit far away, rely more on the sequence itself.
```

User turn. The three arms differ only here. Below is the same `test` point under each arm: the fifth layer of a hole in the Eastern Goldfields, whose true lithology is `ALLU`.

strict

```
Units logged so far,
top to bottom:
  1. CLAY
  2. UNLOGGED
  3. CLAY
  4. UNLOGGED

What is the next
unit's lithology?
```

geo

```
Province: the Eastern Goldfields,
Yilgarn Craton, Western Australia.

Units logged so far, top to bottom:
  1. CLAY (top 0 m, 1 m thick)
  2. UNLOGGED (top 1 m, 2 m thick)
  3. CLAY (top 3 m, 45 m thick)
  4. UNLOGGED (top 48 m, 10 m
     thick)

The next unit starts at 58 m.

What is the next unit's lithology?
```

retrieval

```
Units logged so far, top
to bottom:
  1. CLAY
  2. UNLOGGED
  3. CLAY
  4. UNLOGGED

Nearby holes logged at
this depth:
cover/alluvium 100%.
(10% of the retrieved
holes reach this depth;
mean distance 100 m.)

What is the next unit's
lithology?
```

Cold starts replace the layer list with `No units have been logged yet --- this is the first unit, at the surface.` (**geo**: `No units logged yet; this is the first unit, at 0 m.`, after the province line). The **strict** arm matches the inputs of our model's lithology head, which also sees neither depth nor thickness. The **geo** arm measures how much of a score comes from a regional prior. The Pilbara province line reads `the Pilbara Craton, Western Australia`.

Retrieval evidence. The neighbour sentence is rendered from the same retrieval context as our model (§5.3): TRAIN and WA-DASC-full, k=10, elevation-aligned where possible, a 25 m minimum distance against re-reported twin holes, and a query hole never retrieves itself. For readability, the 37-class inverse-distance-weighted distribution at the query depth is grouped into 13 broad classes, and the five largest classes above 1% are shown, together with the share of the k neighbours that reach this depth and their mean distance, so that the model can judge how much evidence is available. When no neighbour reaches the depth, the sentence is `No nearby hole reaches this depth`. Neighbour bed thickness is not shown, so the retrieval arm adds only neighbour lithology, coverage and distance to the strict arm.

D.3.3 OPEN-WEIGHT MODELS: LIKELIHOOD SCORING

Every candidate code is appended to the chat-templated prompt and scored by teacher-forced joint log-likelihood; the prediction is the ranking of the candidates. The user turn gets one extra line, `Answer with the single code only.`, so that the answer begins at the next token. Qwen3 is run with `enable_thinking=False` for the same reason. The model never has to emit a parseable string, so there are no parse failures by construction.

Raw likelihood is biased here. On the Qwen3 tokeniser the 36 candidates are one token (1 code), two tokens (27) or three tokens (8), and the model places almost all next-token mass on `UN`, echoing the `UNLOGGED` placeholders. Two corrections are applied by default:

- **Domain-conditional PMI** (Holtzman et al., 2021): $s(c) = \log P(c \mid \text{prompt}) - \log P(c \mid \text{neutral})$. The neutral prompt keeps the system prompt and the question but removes all hole evidence, so the correction cancels each code's surface preference.
- **`UNKN` is not a candidate.** It is the target on 2 of 4,258 points in the scored splits of the pilot, all in `ood_far`; those points count as misses.

Candidates are appended with no leading space, because the chat template ends on a newline: `" GRAN"` tokenises as `[' GR','AN']` and scored 16.5 nats below `"GRAN"` on a diagnostic point. The implementation repeats the prompt once per candidate in chunks of 8 and requests logits for the last $L+1$ positions only, which keeps a 32B model inside one GPU.

Generative bridge. Because hosted models cannot be likelihood-scored, Qwen3-8B and Qwen3-32B are also run in generation mode: the exact hosted prompt with no extra line, greedy decoding with at most 96 new tokens, thinking disabled, and the same parser as the hosted models. These rows isolate the effect of the read-out instrument from the effect of the model.

D.3.4 HOSTED MODELS: GENERATIVE SCORING

Hosted models return {`"top3": [c1, c2, c3]`} under a strict JSON schema with exactly three strings. The parser reads the JSON and falls back to any four-capital-letter tokens in the text, truncates each entry to four characters, keeps only valid codes, and removes duplicates. A point whose response cannot be retrieved or parsed is excluded from that row's denominator and counted in the parse-failure column of Table 18. Such rows are compared with our full recipe (GRU + PT + R) on the same quantity, top-1 under the shared mask, but not directly with likelihood-scored rows.

- **GPT-6-astra**: OpenAI Batch API, `/v1/chat/completions` with a 24-hour completion window; system and user messages as above; `reasoning_effort=medium`.
- **Gemini-pro-latest**: Gemini batch API with `response_mime_type=application/json`, the same schema, and `thinking_level=medium`. The batch payload has no system field, so the system prompt is placed in the user turn followed by a blank line; the text the model reads is otherwise identical.

D.3.5 MODELS AND CONFIGURATION

D.3.6 LLM LIMITATIONS

Training-data contamination. The WA DASC drill-hole records are public. A hosted model may have seen them in pre-training, in which case part of its cross-province score reflects recall rather than inference. We cannot audit closed training corpora, and this caveat applies to every hosted row.

Structured-output reliability. GPT-6-astra and all open-weight runs parsed every point. Gemini-pro-latest failed on 1,409 of 2,161 points (65.2%) in the **retrieval** arm, whose prompts are longest, and on up to 97% of `pilbara_test`. The surviving points are a biased subset.

Sampling. Each cell is one run over 150 holes per split. The read-out is deterministic given the prompt for the likelihood and greedy rows, so there is no seed variance, but hole-sampling variance remains: the hole-cluster bootstrap 95% interval is about $\pm$3–4 points, and smaller differences should not be read as an ordering.

Cost. The three GPT-6-astra arms consumed 3.36 M input and 0.53 M output tokens, 65% of them reasoning tokens, for one evaluation pass. A likelihood-scored open-weight arm takes about 2 h on one GPU. Our full recipe (GRU + PT + R) has 0.68 M parameters and runs locally.

D.4 PRETRAINING, RETRIEVAL AND HYPERPARAMETERS

This section collects every experiment on the two components of the recipe and on the settings around them. Unless stated otherwise, experiments use the layer-level GRU of the main text on Task 1 and are scored as in App. D.1: macro-F1 over classes with at least 20 events, and *paired* differences against a stated reference, with both arms scored on the same hole-cluster bootstrap resample and seeds resampled. Arms use three fine-tuning seeds, and the deployed configuration uses ten. Each pretraining variant is a single checkpoint. "Full recipe" means pretraining plus retrieval. Table 19 lists every setting, its deployed value and where it is varied, and Figure 7 summarises the quantitative axes.

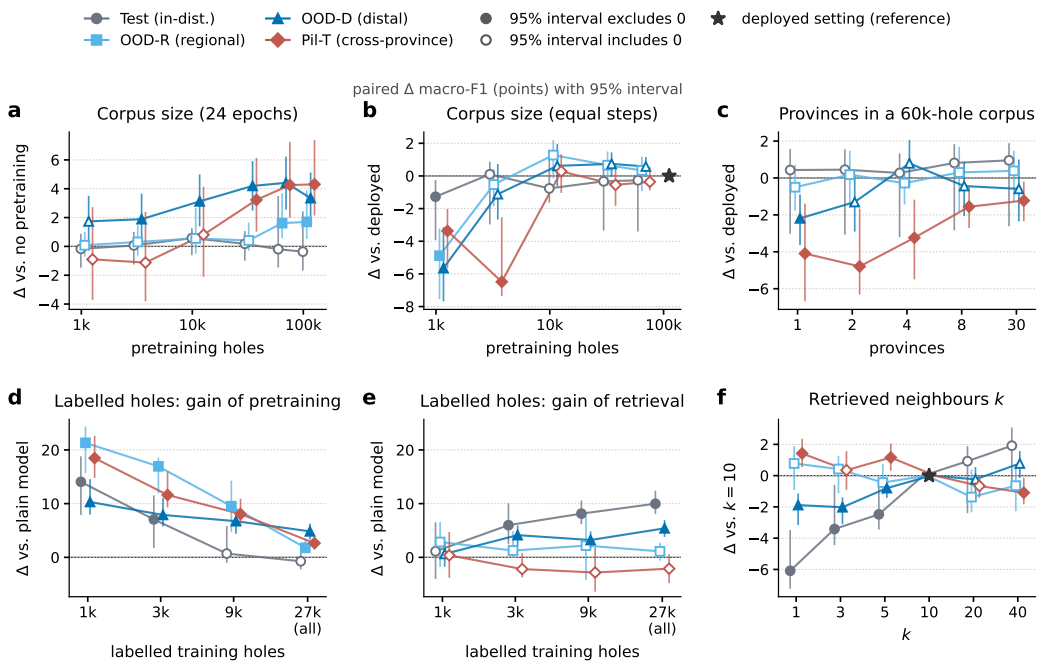

Figure 7: **Pretraining and retrieval ablations** on Task 1: paired Δ macro-F1 (points) with 95% intervals; filled markers are differences whose interval excludes zero, the star marks the deployed reference. Panels a, b, c and f use the full recipe. **(a)** Pretraining-corpus size at a fixed 24 epochs, against no pretraining. **(b)** The same corpus sizes with the number of optimisation steps held at that of the full corpus, against the deployed checkpoint. **(c)** Number of provinces in a 60k-hole corpus, against the deployed checkpoint. **(d, e)** Labelled training holes: gain of pretraining (d) and of retrieval (e) over the plain model trained on the same holes. **(f)** Number of retrieved neighbours k, against the deployed k=10.

D.4.1 What each component contributes

Table 20 adds the components one at a time. Label smoothing, applied to the lithology-only model, helps on all three shifted splits at no cost in distribution. Retrieval then gives the largest single gain in distribution and a clear gain on the distal ring, and nothing measurable across provinces. Pretraining shows the opposite profile: nothing in distribution and a significant gain on every shifted split. Feeding thickness and depth back as inputs helps slightly in distribution, not significantly, and costs a significant amount on PILBARA-TEST. This is the transfer boundary of §6 at a smaller scale, and the reason the deployed model reads lithology only.

The same division of labour holds when each component is measured on its own against the plain model (Table 10). Retrieval buys in-distribution and ring accuracy, pretraining buys the shifted splits, and the full recipe improves on the plain model on all four splits. On Task 2 (Table 14) neither component changes the generated law in distribution or on the regional ring, where a sequence model trained on the Eastern Goldfields already matches it. On the distal ring and across provinces both help. There the full recipe is the only arm whose gain is significant on both splits, and on PILBARA-TEST the two components are additive.

D.4.2 Pretraining

Corpus size. The pretraining corpus consists of the roughly 112k archive holes with at least two logged lithology layers. Table 21 pretrains on nested random subsets of it for a fixed 24 epochs. In distribution the corpus size never matters. Under shift the benefit grows with the corpus, but the splits differ in how much they need. The distal ring gains from a few thousand holes on, whereas PILBARA-TEST gains only from tens of thousands of holes on. Below that, pretraining without re-

trieval is harmful across provinces, though not significantly: a small corpus mostly repeats the Eastern Goldfields and teaches nothing that transfers.

Data or optimisation? A fixed epoch count also gives a small corpus fewer optimisation steps. Table 22 holds the number of steps at that of the full corpus. With equal compute the curve saturates at about a tenth of the corpus. Apart from one significant dip across provinces at 30% without retrieval, only the two smallest fractions remain clearly worse. Much of the climb in Table 21 beyond that point is therefore optimisation rather than data. What the data itself buys is concentrated in the smallest corpora.

Province diversity. Subsampling also removes provinces. Table 23 fixes the corpus at 60k holes and draws it from one to thirty provinces. PILBARA-TEST responds to diversity. Its deficit against the full archive is largest for a one- or two-province corpus and at least halves at thirty provinces, while the in-distribution score does not move. The distal ring, which lies in the training province, responds much less. Volume alone does not transfer across provinces; geological breadth does.

Held-out provinces. The archive covers all of Western Australia and therefore contains unlabelled Pilbara holes. A gain on PILBARA-TEST could reflect a transferable prior or mere exposure to the target province. We repeat pretraining, with an identical recipe, on two reduced archives. *Pilbara held out* removes every archive hole inside the bounding box of the PILBARA-TEST collars dilated by 100 km, about a quarter of the archive; the nearest remaining hole is more than 100 km from any Pilbara collar. *Pilbara and EG held out* also removes the bounding box of the two OOD rings dilated by 100 km, which covers the training footprint as well, so that pretraining sees neither province evaluated in this paper. This removes about half of the archive. Fine-tuning, seeds and the retrieval pool are unchanged. Most of the cross-province gain survives. With retrieval, four fifths or more of the PILBARA-TEST gain remains (82% with Pilbara held out, 96% with the Eastern Goldfields held out as well). Without retrieval the gain over the label-smoothed GRU is smaller to begin with, and holding Pilbara out keeps about half of it (45%; 74% with the Eastern Goldfields also held out). Top-1 accuracy on PILBARA-TEST does not change beyond seed noise (Table 24). Removing the Eastern Goldfields as well costs nothing further, and the rings lose at most about a point. The gain therefore largely reflects stratigraphic structure learned from other provinces. The residual difference is the part that does come from having read the target province, and it is what a practitioner gains by pretraining on the full archive.

Objective. Table 25 varies the pretraining targets on the full corpus. The fine-grained lithology objective is the component that matters. Replacing it by the 13-class coarse objective costs a significant amount on all three shifted splits, with or without retrieval and with or without chemistry. The chemistry losses themselves are not measurable at this scale: switching them off makes no split worse, and the one significant difference, on the regional ring without retrieval, favours switching them off. We keep them in the released recipe because they do no harm and make the geochemistry channel of the archive available, but no transfer claim rests on them.

Labelled data. Table 26 fine-tunes on random subsets of TRAIN. Pretraining is worth most when labels are scarce. At a thousand labelled holes it lifts every split, PILBARA-TEST included, by a wide margin, and a pretrained model fine-tuned on a thousand holes already beats a model trained from scratch on all of TRAIN across provinces. As labels grow, its in-distribution gain vanishes and its shifted-split gain settles to the values of Table 20. Retrieval behaves the other way. Its in-distribution gain grows with the labelled set, because the labelled part of the pool grows with it, and it adds nothing on PILBARA-TEST at any size.

D.4.3 RETRIEVAL

Number of neighbours. Table 27 varies k. In distribution accuracy keeps rising up to the largest k tried. Across provinces it moves the other way, significantly at the largest k: the farther neighbours of a Pilbara hole resemble it less than the nearest few, and averaging them dilutes the evidence. A single neighbour already gives about half of the gain on the distal ring and is slightly better than $k{=}10$ across provinces. The deployed $k{=}10$ sits on the plateau of the shifted splits; a user who cares only about in-distribution accuracy should raise it. Restricting neighbours to a 1 km radius hurts on

the distal ring, where the nearest evidence lies farther away, and re-ranking candidates by mutual agreement does not help anywhere. Distance is the better selection rule.

Pool composition. Table 28 changes what the model can retrieve from. Searching only the training split leaves in-distribution accuracy unchanged but costs both rings heavily. Their nearest training holes are kilometres away, and the archive supplies the neighbours they use. Across provinces the training-only pool is no worse, so the archive's Pilbara holes are not what carries the cross-province result. Table 29 makes this explicit by thinning the 357 Pilbara holes that an earlier release set aside as a target-region subset. Removing any number of them leaves PILBARA-TEST unchanged within noise in either pool, which is why the subset was folded back into the archive.

The same pool for the spatial baselines. A larger pool could favour our model simply by showing it more holes. We therefore refit spatial k-NN, indicator kriging and SIS on the same train + archive pool (Table 9). The archive lifts them under every shift but leaves them far below the recipe on every shifted split (Table 10). What the recipe adds is not access to the pool but a sequence model that reads the neighbours as context.

Design choices. Table 30 changes the retrieval module one choice at a time. Neighbour dropout, elevation alignment, the self-exclusion radius and a learned gate on the retrieval term all stay within about a point of the deployed setting. The gate, initialised open, learns nothing that the additive fusion does not already do. The one change that matters is how neighbour evidence is read. The deployed module lets a neighbour vote only at depths where one of its beds starts, so across provinces only about a third of the cells receive any evidence. Letting every bed vote over its whole thickness gives the largest in-distribution gain of any change we tried, but costs significantly on PILBARA-TEST: denser evidence from the same neighbours helps where they are right and hurts where they are not. We report it as the main untapped in-distribution gain and leave it out of the deployed recipe, whose target is transfer.

D.4.4 WHERE EACH COMPONENT ACTS

Two analyses of where the components act are reported with the Task 1 and Task 2 results and are not repeated here. Cover versus bedrock (Table 12) shows that retrieval is mainly a cover mechanism and pretraining mainly a bedrock one, with the distal ring as the exception. The five-backbone comparison (Tables 11 and 15) shows that the division of labour of Table 20 holds for every backbone on both tasks.

Token unit. Table 31 keeps the Transformer backbone and the recipe fixed and changes only the token: a 1 m cell or a whole stratum. Retrieval helps in distribution at both levels. Pretraining does not: at the cell level it lowers macro-F1 on most splits, whereas at the stratum level it raises it on all three shifted splits. The full recipe at the cell level is therefore well below the stratum-level one under every shift. At one-metre resolution most targets are `CONT`, so the pretraining signal is dominated by bed thickness rather than by lithological succession. The stratum token removes that imbalance, which is why the recipe is built on it.

D.4.5 TOKENISATION GRID

Changing the cell spacing changes the task, so the benchmark is frozen at 1 m. Table 32 retrains the plain model and the full recipe at 0.5 m and 2 m as a sensitivity probe. The plain model is indifferent to the finer grid and loses at the coarser one, where coarsening erases layer starts. The full recipe at 0.5 m is slightly worse in distribution and unchanged under shift. A likely reason is that a neighbour votes only in the cell where one of its beds starts, and a half-metre cell is missed more often once the look-up is shifted by elevation. At 2 m the full recipe is unchanged in distribution and loses across provinces. Neither alternative improves any split significantly, and each costs on at least one, so we keep the 1 m grid, which also matches the logging resolution of the source data.

Table 8: **Model sizes and compute.** Wall-clock is one run on one GPU (V100, A6000 or H100, whichever was free); classical baselines are fitted on CPU. Deep arms are trained with three seeds, so the cost of a reported row is three times the figure shown. Pretraining is paid once per checkpoint and is reused by every arm that starts from it.

Method	Inputs	Params	Wall-clock
Classical and geostatistical (fit + score, CPU)			
Unigram / per-region unigram	lith.	–	$<$ 1 min
Bigram / 3-gram / depth-Markov	lith.	–	1 min
HMM (K=24)	lith.	–	1 min
Layer Markov (order 2)	lith.	–	1 min
TPROGS	lith. + loc.	–	1 min
Spatial k-NN / kriging / SIS	loc.	–	1–2 min
MPS-lite	lith. + loc.	–	2 min
XGBoost	lith. + loc.	–	47 min
Cell-level autoregressive (1 m token, 20–30 epochs)			
GRU	lith.	0.65 M	4 min
LSTM	lith.	0.87 M	7 min
TCN	lith.	1.35 M	11 min
Transformer	lith.	2.68 M	16 min
Mamba	lith.	2.03 M	4.7 h
GRU + loc. / + chem.	–	0.73 M	5–7 min
LSTM + chem. / + chem. + loc.	–	0.88–1.01 M	10 min
TCN + chem. + loc.	–	1.40 M	11 min
Transformer + chem. + loc.	–	2.74 M	23 min
Mamba + chem. + loc.	–	2.08 M	4.7 h
Layer level (stratum token, 20 epochs)			
GRU	lith.	0.65 M	4 min
Transformer	lith.	1.6 M	6 min
GRU + retrieval ($\pm$ pretraining)	lith. + retr.	0.68 M	4 min
+ Task 2 rollouts	–	–	17 min
Cell level, ours (10 epochs, archive retrieval)			
GRU + retrieval ($\pm$ pretraining)	lith. + retr.	3.4 M	10 min
Transformer + retrieval ($\pm$ pretr.)	lith. + retr.	38.8 M	27 min
+ Task 2 rollouts	–	–	53 min
Corpus pretraining			
Layer-level GRU, 24 epochs	lith. (+ chem.)	0.65 M	20 min
Cell-level GRU, d=512, 2 layers	lith.	3.4 M	22 min
Cell-level Transformer, d=512, 12 layers	lith.	38.8 M	7 h
Zero-shot LLMs (150 holes per split, no training)			
Qwen3-8B / 14B / 32B, phi-4	prompt	8–32 B	2.0 h per arm
GPT-6, Gemini-pro	prompt	–	batch API

Table 9: **Task 1, all baselines** on the four evaluation splits. Layer-level neural rows use label smoothing (ε=0.3). $\pm$: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with $\geq$20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports.

		macro-F1 (%)				top-1 accuracy (%)			
Method	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
Frequency / Markov (cell level)									
Unigram	1	2.0±0.2	0.8±0.1	1.1±0.2	1.7±0.2	20.0±1.0	12.0±1.3	13.2±2.0	23.7±1.2
Per-region unigram	1	2.0±0.2	0.8±0.1	1.1±0.2	1.7±0.2	20.0±1.0	12.0±1.3	13.2±2.0	23.7±1.2
Persistence	1	2.5±0.3	0.6±0.1	0.7±0.1	1.6±0.2	11.6±0.8	3.4±1.3	4.8±1.7	10.5±1.0
Bigram	1	4.5±0.7	1.1±0.2	1.2±0.2	1.7±0.2	21.3±1.0	12.2±1.3	13.2±2.1	22.7±1.2
3-gram	1	9.4±1.3	3.2±1.0	4.0±0.9	3.7±0.8	23.3±0.9	13.6±1.4	15.9±2.4	21.0±1.0
Depth-Markov	1	9.0±1.2	3.7±0.5	3.4±0.6	1.5±0.2	28.0±1.2	18.0±1.6	15.6±2.4	10.4±0.7
HMM (K=24)	1	2.0±0.2	0.8±0.1	1.1±0.2	1.7±0.2	20.0±1.0	12.0±1.3	13.2±2.0	23.7±1.2
Geostatistical (cell level)									
Spatial k-NN, train pool	1	71.5±3.7	0.9±0.1	2.0±0.3	2.1±0.2	78.2±1.1	12.5±1.3	14.0±2.1	24.2±1.2
Spatial k-NN, archive pool	1	68.9±3.8	11.5±1.4	9.6±1.3	8.0±1.0	75.7±1.1	24.9±1.9	23.5±2.6	19.8±1.2
Indicator kriging, train pool	1	71.6±3.8	0.9±0.1	2.0±0.3	2.1±0.2	76.1±1.1	12.5±1.3	14.0±2.1	24.2±1.2
Indicator kriging, archive pool	1	69.5±3.8	6.4±0.8	7.6±1.7	6.8±1.1	74.5±1.1	22.8±1.8	24.3±2.3	23.8±1.2
SIS, train pool	1	68.3±3.8	4.0±0.5	4.6±0.7	2.4±0.3	74.8±1.2	11.2±1.5	8.5±1.5	6.6±0.5
SIS, archive pool	1	66.2±3.7	8.3±0.9	8.5±1.1	5.6±0.7	73.2±1.2	19.6±1.5	19.9±2.6	14.0±1.0
MPS-lite	1	18.8±2.4	5.6±1.1	6.4±1.3	7.4±1.4	28.7±1.1	16.8±1.3	16.2±2.5	22.7±1.1
TPROGS	1	4.5±0.7	1.1±0.2	1.2±0.2	1.7±0.2	21.3±1.0	12.2±1.3	13.2±2.1	22.7±1.2
Machine learning (cell level)									
XGBoost	1	64.8±4.9	2.4±0.3	2.5±0.4	2.2±0.3	76.1±1.1	19.3±1.7	11.5±2.2	11.4±0.8
Deep learning, cell level, lithology only									
LSTM	1	58.9±5.1	25.8±3.1	22.2±3.4	22.3±2.5	60.1±1.2	39.1±2.0	31.7±2.7	38.5±1.3
GRU	1	58.8±5.3	21.4±2.6	23.0±3.4	23.2±2.2	60.3±1.1	37.1±2.0	35.5±3.0	36.3±1.4
Mamba	1	59.2±5.0	24.8±2.6	22.2±3.2	19.4±1.6	60.1±1.2	38.5±1.9	34.9±2.9	35.9±1.2
TCN	1	51.6±5.3	17.9±2.2	17.5±2.5	16.3±1.7	58.6±1.1	34.9±2.1	32.4±2.7	31.7±1.3
Transformer	1	60.3±5.1	26.5±3.0	23.6±3.4	22.5±2.2	60.6±1.2	39.7±2.0	34.8±3.0	38.2±1.3
Deep learning, cell level, conditioned on location / geochemistry									
LSTM + location	1	73.4±4.0	11.5±1.5	10.5±1.8	7.5±1.2	80.5±1.0	23.1±1.7	20.4±2.0	22.6±1.1
LSTM + geochem	1	65.0±5.3	17.8±2.2	17.2±2.5	9.2±1.3	68.1±1.0	34.0±2.3	27.2±2.7	19.9±1.2
LSTM + geochem + location	1	75.6±4.4	10.4±1.4	7.8±1.4	3.4±0.5	81.1±0.9	30.3±2.0	18.0±2.3	8.0±0.8
GRU + location	1	73.6±5.5	13.0±1.6	3.1±0.6	0.5±0.1	79.6±0.9	34.4±2.4	12.1±1.9	3.0±0.4
GRU + geochem	1	65.6±5.3	19.9±2.2	17.7±2.5	11.0±1.5	67.7±1.0	38.1±2.3	30.8±2.9	23.2±1.2
GRU + geochem + location	1	75.1±3.8	15.2±2.0	3.7±0.8	2.5±0.4	80.4±0.9	32.2±2.2	12.4±2.1	11.3±1.2
Transformer + geochem + location	1	**77.2**±4.9	18.1±2.2	11.0±1.7	7.2±1.0	**81.8**±0.9	32.8±2.1	21.8±2.6	22.5±1.1
Mamba + geochem + location	1	74.2±5.2	18.7±2.3	10.6±1.7	9.4±1.2	80.8±0.9	33.9±2.2	24.6±2.7	26.4±1.2
TCN + geochem + location	1	66.5±5.6	15.4±1.8	13.0±1.9	9.2±1.5	76.0±1.0	35.3±2.3	20.6±2.3	21.3±1.1
Layer level									
Layer marginal	1	2.0±0.2	0.8±0.1	1.1±0.2	1.7±0.2	20.0±1.0	12.0±1.3	13.2±2.0	23.7±1.2
Layer Markov	1	56.7±5.0	36.3±4.0	26.1±3.7	22.0±3.3	57.0±1.2	37.4±1.8	32.8±2.6	36.5±1.4
GRU	10	59.4±5.0	37.4±3.5	28.3±4.0	28.2±3.1	58.6±1.0	40.5±1.8	34.7±3.2	42.0±1.3
Transformer	10	58.9±4.9	36.6±3.5	29.1±3.9	26.0±2.9	58.3±1.1	40.2±1.8	35.1±3.0	39.7±1.3
GRU + R	10	70.0±5.0	37.4±3.3	33.9±4.5	26.9±2.9	73.5±0.9	52.5±2.0	46.9±3.2	42.8±1.2
GRU + PT	10	58.5±5.1	**39.2**±3.3	33.0±4.2	**31.7**±2.8	58.6±1.1	40.8±1.8	35.4±3.0	44.0±1.2
GRU + PT + R (ours)	10	69.2±5.0	38.8±3.5	**37.2**±4.6	30.6±2.9	73.7±0.9	**52.9**±2.1	**48.5**±3.3	**44.3**±1.2

Table 10: **Task 1 paired differences** (percentage points) with 95% intervals. Both methods are scored on the same hole-cluster bootstrap resample (B=1000), seeds resampled independently. Intervals that exclude zero are bold. GRU and layer Transformer: the plain models with label smoothing (ε=0.3).

	Δ macro-F1				Δ top-1			
Comparison	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
GRU + R − GRU	**+10.7** [+8.9, +12.7]	-0.1 [-1.0, +2.2]	**+5.6** [+4.3, +6.9]	-1.3 [-3.1, +0.6]	**+15.0** [+14.2, +15.8]	**+12.0** [+10.9, +14.1]	**+12.2** [+8.4, +16.1]	+0.8 [-0.0, +1.7]
GRU + PT − GRU	-0.9 [-2.0, +0.3]	**+1.8** [+1.1, +3.2]	**+4.7** [+3.4, +5.8]	**+3.5** [+2.2, +4.8]	+0.0 [-0.2, +0.3]	+0.3 [-0.7, +1.5]	+0.7 [-2.7, +3.6]	**+2.0** [+1.0, +3.0]
Ours − GRU	**+9.9** [+7.8, +13.1]	**+1.4** [+0.4, +4.0]	**+8.9** [+7.5, +10.6]	**+2.4** [+1.0, +4.3]	**+15.1** [+14.3, +16.0]	**+12.4** [+11.2, +14.4]	**+13.8** [+10.4, +17.6]	**+2.4** [+1.5, +3.3]
Ours − GRU + R (adds PT)	-0.8 [-1.8, +2.9]	**+1.4** [+0.4, +2.6]	**+3.4** [+2.1, +4.6]	**+3.7** [+2.2, +5.4]	+0.2 [-0.2, +0.5]	+0.4 [-0.8, +1.1]	+1.6 [-0.9, +4.9]	**+1.5** [+1.0, +2.1]
Ours − GRU + PT (adds R)	**+10.8** [+9.2, +13.1]	-0.4 [-1.8, +1.6]	**+4.2** [+3.3, +5.5]	-1.1 [-2.2, +0.4]	**+15.1** [+14.2, +16.0]	**+12.1** [+10.6, +14.1]	**+13.1** [+10.0, +17.1]	+0.3 [-0.5, +1.0]
Ours − Layer Transformer	**+10.4** [+8.4, +13.5]	**+2.2** [+0.9, +5.3]	**+8.1** [+6.5, +9.7]	**+4.6** [+3.4, +6.6]	**+15.4** [+14.5, +16.3]	**+12.7** [+11.6, +14.8]	**+13.5** [+10.1, +17.4]	**+4.6** [+3.7, +5.5]
Ours − Layer Markov	**+12.5** [+10.1, +15.7]	**+2.5** [+1.8, +5.7]	**+11.1** [+9.0, +13.2]	**+8.7** [+4.9, +11.5]	**+16.7** [+15.7, +17.8]	**+15.5** [+13.8, +17.8]	**+15.8** [+13.8, +20.1]	**+7.9** [+6.7, +8.9]
Ours − cell Transformer	**+8.9** [+6.9, +12.1]	**+12.3** [+8.9, +14.3]	**+13.6** [+11.0, +15.8]	**+8.1** [+6.1, +12.4]	**+13.0** [+12.1, +14.1]	**+13.3** [+11.3, +15.6]	**+13.7** [+11.1, +18.1]	**+6.2** [+5.2, +7.2]
Ours − Spatial k-NN (archive)	+0.3 [-7.1, +4.0]	**+27.3** [+24.4, +30.3]	**+27.6** [+23.2, +30.9]	**+22.6** [+20.7, +27.2]	**-2.0** [-3.1, -0.8]	**+28.0** [+25.1, +29.4]	**+25.1** [+22.0, +29.1]	**+24.5** [+23.2, +25.9]
Ours − XGBoost	+4.4 [-7.2, +8.2]	**+36.4** [+33.4, +40.2]	**+34.8** [+29.9, +38.7]	**+28.4** [+27.0, +32.9]	**-2.4** [-3.5, -1.3]	**+33.7** [+31.1, +35.9]	**+37.0** [+33.4, +42.1]	**+32.9** [+31.5, +34.0]
GRU + location − GRU (cell)	**+14.7** [+12.3, +17.7]	**-8.4** [-10.7, -7.2]	**-19.9** [-22.7, -16.4]	**-22.7** [-24.5, -20.1]	**+19.3** [+18.3, +20.5]	-2.7 [-5.0, +1.0]	**-23.4** [-26.0, -20.0]	**-33.3** [-34.8, -31.8]
GRU + geochem. − GRU (cell)	**+6.8** [+4.2, +9.1]	-1.4 [-3.1, +0.0]	**-5.3** [-6.9, -3.4]	**-12.2** [-13.5, -8.6]	**+7.4** [+6.5, +8.2]	+1.1 [-1.1, +3.8]	**-4.7** [-7.3, -1.9]	**-13.1** [-14.4, -11.9]
GRU + both − GRU (cell)	**+16.3** [+13.3, +21.5]	**-6.1** [-8.4, -4.8]	**-19.3** [-22.0, -15.8]	**-20.7** [-22.3, -18.1]	**+20.1** [+19.0, +21.3]	**-4.9** [-7.6, -2.2]	**-23.1** [-25.6, -20.4]	**-25.0** [-26.7, -23.2]
LSTM + location − LSTM (cell)	**+14.5** [+12.9, +20.8]	**-14.2** [-17.3, -13.1]	**-11.7** [-13.6, -9.0]	**-14.8** [-18.6, -13.3]	**+20.4** [+19.3, +21.5]	**-16.0** [-17.7, -14.2]	**-11.3** [-12.5, -7.6]	**-16.0** [-17.3, -14.6]

Table 11: **Task 1 across layer-level backbones.** Each backbone is trained plain (lithology only), with retrieval (R), with fine-objective pretraining (PT), and with both; all four arms use label smoothing (ε=0.3, App. **??**). ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports.

		macro-F1 (%)				top-1 accuracy (%)			
Method	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
GRU									
GRU	10	59.4±4.9	37.4±3.5	28.3±3.9	28.2±3.0	58.6±1.1	40.5±1.9	34.7±3.1	42.0±1.4
+ R	10	**70.0**±5.0	37.4±3.3	33.9±4.5	26.9±2.8	73.5±1.0	52.5±1.9	46.9±3.4	42.8±1.2
+ PT	10	58.5±5.1	39.2±3.2	33.0±4.2	**31.7**±2.8	58.6±1.1	40.8±1.9	35.4±3.0	44.0±1.3
+ PT + R	10	69.2±4.9	38.8±3.5	37.2±4.7	30.6±2.9	73.7±0.9	**52.9**±2.1	48.5±3.3	44.3±1.2
LSTM									
LSTM	3	59.1±5.1	37.4±4.0	28.1±4.1	25.0±3.7	58.5±1.1	41.2±2.2	36.0±4.6	39.8±1.8
+ R	3	69.5±5.3	37.3±3.9	33.8±4.9	25.0±3.8	73.6±0.9	52.3±2.0	48.1±4.0	41.5±1.8
+ PT	3	58.2±5.3	38.9±3.4	30.5±4.2	27.6±3.1	58.5±1.1	41.5±2.3	36.4±4.3	43.2±1.7
+ PT + R	3	67.9±5.6	38.5±3.2	34.6±4.6	29.1±3.5	73.6±1.0	52.6±2.1	47.7±4.0	**45.5**±1.4
Transformer									
Transformer	10	58.9±5.0	36.6±3.6	29.1±3.9	26.0±3.0	58.3±1.1	40.2±1.8	35.1±3.1	39.7±1.3
+ R	3	69.8±4.8	36.6±3.8	34.8±4.6	24.0±3.5	**73.7**±1.0	52.0±2.2	48.4±4.5	39.3±1.4
+ PT	3	57.5±5.2	38.8±3.3	30.8±3.8	28.1±2.9	58.1±1.1	40.9±2.1	36.4±4.2	40.7±1.3
+ PT + R	3	68.2±5.3	**39.6**±3.3	34.9±4.1	26.0±3.8	73.4±1.0	52.5±2.3	48.3±4.9	41.8±1.3
TCN									
TCN	3	58.4±4.9	32.0±2.7	30.1±4.1	27.9±2.7	58.1±1.1	40.3±2.1	35.0±4.6	43.5±1.3
+ R	3	69.4±4.9	33.0±3.2	32.3±4.3	26.4±2.8	73.5±1.0	52.6±2.1	47.2±4.4	42.8±1.3
+ PT	3	57.7±4.9	36.0±3.5	33.2±4.0	29.9±2.5	58.0±1.1	41.7±2.2	37.7±4.7	42.1±1.4
+ PT + R	3	68.7±4.9	37.7±3.5	**37.9**±4.7	29.2±2.3	73.3±1.0	52.7±2.2	**48.7**±4.3	43.7±1.3
Mamba									
Mamba	3	58.0±4.9	28.4±3.5	25.5±3.6	19.5±3.2	57.9±1.1	41.2±2.2	36.9±2.6	33.6±2.4
+ R	3	68.2±5.5	28.6±3.7	29.6±4.5	21.2±2.1	73.2±1.0	51.3±2.2	46.4±4.5	37.0±3.0
+ PT	3	57.5±4.9	29.3±3.2	27.3±3.8	23.7±2.8	57.8±1.1	40.3±2.4	35.2±4.6	38.1±1.2
+ PT + R	3	67.6±5.2	30.9±3.3	30.5±4.1	22.6±2.7	73.1±1.0	51.1±2.3	46.7±4.1	39.6±1.4

Table 12: **Cover vs. bedrock** (layer-level GRU, top-1 %, pooled inside each zone by the *true* next lithology; 10 seeds). ΔR: retrieval gain over the plain model (with label smoothing, as every arm); ΔPT: pretraining gain without retrieval. n: scored layer starts.

Split	Zone	n	plain	+R	+PT	+PT+R	ΔR	ΔPT
TEST	cover	5,870	59.7	75.7	59.7	76.0	+16.0	+0.0
	bedrock	4,539	57.1	70.7	57.2	70.7	+13.6	+0.1
OOD-R	cover	26,510	43.2	58.6	43.9	59.3	+15.3	+0.7
	bedrock	4,100	53.3	54.0	54.7	56.6	+0.7	+1.4
OOD-D	cover	7,380	36.7	47.6	41.0	52.3	+10.9	+4.3
	bedrock	2,233	34.1	39.1	34.6	39.7	+4.9	+0.5
PILBARA-TEST	cover	4,214	43.6	44.4	44.8	44.8	+0.8	+1.2
	bedrock	5,724	40.7	41.6	43.4	44.0	+0.8	+2.7

Table 13: **Task 2, all baselines.** Pooled joint KL, JSD and per-metre agreement. *: lookup, not free-running; not highlighted. ±: half-width of the 95% report-cluster bootstrap interval (B=1000; holes of one exploration report are resampled together); trained rows also resample seeds. Plug-in divergences are biased upward under resampling, so intervals are re-centred on the full-sample estimate. Layer-level neural rows (and the plain arms of the backbone table) use label smoothing (ε=0.3).

		KL ↓				JSD ↓				Acc/m ↑			
Method	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
Frequency / Markov (cell level)													
Unigram	1	0.226±0.077	0.599±0.149	0.872±0.476	1.111±0.316	0.047±0.013	0.130±0.024	0.133±0.050	0.203±0.041	50.7±2.6	48.2±4.3	51.5±5.3	41.9±4.9
Per-region unigram	1	0.226±0.077	0.599±0.149	0.872±0.476	1.111±0.316	0.047±0.013	0.130±0.024	0.133±0.050	0.203±0.041	50.7±2.6	48.2±4.3	51.5±5.3	41.9±4.9
Persistence	1	1.318±0.303	1.848±0.487	2.959±1.397	2.112±0.551	0.213±0.020	0.286±0.034	0.301±0.075	0.285±0.032	67.9±3.3	**64.8**±5.4	**66.3**±6.3	**57.3**±6.4
Bigram	1	0.184±0.052	0.489±0.149	0.639±0.411	0.949±0.298	0.042±0.011	0.104±0.022	0.111±0.048	0.179±0.037	46.2±2.6	43.7±4.3	47.5±5.1	37.4±5.1
3-gram	1	0.186±0.067	0.533±0.190	0.781±0.498	0.960±0.285	0.040±0.012	0.110±0.021	0.110±0.046	0.175±0.036	48.9±2.6	47.5±4.3	50.9±5.1	38.8±5.1
Depth-Markov	1	0.220±0.062	0.518±0.125	0.564±0.335	0.797±0.223	0.058±0.012	0.122±0.021	0.116±0.045	0.176±0.032	48.7±2.6	46.4±4.3	49.1±5.0	38.2±5.3
HMM (K=24)	1	0.226±0.073	0.584±0.143	0.774±0.456	1.059±0.299	0.047±0.012	0.126±0.025	0.130±0.049	0.203±0.041	49.9±2.6	47.5±4.4	50.5±5.2	41.0±5.0
Geostatistical (cell level)													
Spatial k-NN, train pool*	1	0.546±0.052	1.445±0.394	2.022±0.942	2.028±0.679	0.161±0.013	0.226±0.031	0.254±0.068	0.260±0.037	52.4±2.3	63.0±5.4	64.3±6.3	55.9±6.3
Spatial k-NN, archive pool*	1	0.553±0.053	0.950±0.105	0.852±0.352	0.729±0.124	0.162±0.013	0.262±0.022	0.219±0.047	0.199±0.028	54.6±2.3	39.3±4.5	39.8±4.3	33.7±4.8
Indicator kriging, train pool	1	0.192±0.066	1.445±0.394	2.022±0.942	2.028±0.679	0.039±0.010	0.226±0.031	0.254±0.068	0.260±0.037	70.7±2.6	63.0±5.4	64.3±6.3	55.9±6.3
Indicator kriging, archive pool	1	0.437±0.119	1.205±0.402	1.991±0.892	1.475±0.633	0.089±0.016	0.205±0.034	0.249±0.067	0.218±0.038	**77.8**±2.4	62.9±5.3	64.9±6.0	55.9±5.9
SIS, train pool	1	0.387±0.038	0.844±0.125	0.902±0.440	0.806±0.169	0.117±0.009	0.220±0.023	0.205±0.061	0.197±0.026	58.5±2.4	47.4±4.5	52.0±5.2	41.4±5.1
SIS, archive pool	1	0.424±0.049	0.811±0.126	0.890±0.431	0.742±0.150	0.124±0.010	0.212±0.025	0.203±0.061	0.186±0.026	62.0±2.4	47.8±4.5	52.2±5.1	42.5±5.1
MPS-lite	1	0.206±0.083	0.526±0.188	0.641±0.389	0.968±0.292	0.037±0.011	0.106±0.019	0.108±0.043	0.172±0.035	51.0±2.7	50.5±4.5	51.9±5.4	40.4±5.1
TPROGS	1	0.183±0.051	0.492±0.151	0.716±0.380	0.885±0.278	0.042±0.011	0.103±0.022	0.112±0.048	0.177±0.037	46.2±2.6	43.7±4.3	47.5±5.1	37.4±5.1
Machine learning (cell level)													
XGBoost	1	0.206±0.076	0.518±0.164	0.950±0.475	1.776±0.320	0.045±0.011	0.119±0.037	0.185±0.059	0.343±0.043	62.7±2.6	43.2±4.6	33.7±4.1	27.0±4.8
Deep learning, cell level, lithology only													
LSTM	1	0.341±0.089	0.547±0.127	0.529±0.341	1.054±0.398	0.078±0.015	0.123±0.027	0.096±0.047	0.191±0.045	69.1±2.7	62.4±5.1	59.0±5.8	51.7±6.1
GRU	1	0.340±0.081	0.612±0.136	0.559±0.342	1.003±0.449	0.078±0.016	0.132±0.025	0.103±0.030	0.173±0.043	69.4±2.7	61.8±5.2	60.6±5.6	49.9±6.3
Mamba	1	0.345±0.077	0.580±0.118	0.536±0.358	0.900±0.245	0.083±0.014	0.132±0.025	0.098±0.038	0.185±0.035	69.4±2.7	64.2±4.7	61.1±5.5	51.7±6.0
TCN	1	0.348±0.099	0.683±0.193	0.668±0.397	0.899±0.294	0.071±0.013	0.143±0.025	0.114±0.036	0.160±0.029	70.7±2.7	64.5±4.7	62.1±5.7	54.9±6.5
Transformer	1	0.356±0.087	0.585±0.114	0.533±0.332	0.995±0.373	0.079±0.016	0.131±0.022	0.103±0.029	0.178±0.033	69.7±2.8	62.8±5.1	59.5±5.6	52.6±6.1
Deep learning, cell level, conditioned on location / geochemistry													
LSTM + location	1	0.283±0.083	0.801±0.237	1.250±0.613	1.511±0.439	0.064±0.015	0.166±0.047	0.240±0.084	0.252±0.040	76.2±2.4	55.1±6.1	50.3±5.7	42.8±5.9
LSTM + geochem	1	0.595±0.121	0.703±0.159	0.783±0.420	1.519±0.469	0.131±0.018	0.152±0.027	0.136±0.054	0.259±0.038	65.4±2.9	60.4±5.5	56.6±6.0	43.1±6.3
LSTM + geochem + location	1	0.379±0.095	0.977±0.276	1.434±0.631	3.082±0.639	0.086±0.016	0.208±0.058	0.253±0.075	0.442±0.037	68.6±2.9	51.3±5.6	43.2±6.4	27.9±5.1
GRU + location	1	0.283±0.075	0.540±0.232	2.638±0.649	6.016±0.978	0.065±0.016	0.116±0.027	0.389±0.067	0.499±0.061	75.4±2.6	56.0±6.1	24.5±6.2	5.9±3.2
GRU + geochem	1	0.384±0.102	0.772±0.174	0.689±0.424	1.081±0.417	0.078±0.014	0.161±0.028	0.112±0.046	0.202±0.051	69.0±3.0	61.4±5.5	59.6±6.0	50.9±6.1
GRU + geochem + location	1	0.448±0.104	1.350±0.403	3.882±0.904	4.537±1.022	0.104±0.017	0.274±0.069	0.485±0.049	0.398±0.056	73.5±2.6	54.5±6.0	10.7±4.4	18.9±6.7
Transformer + geochem + location	1	0.472±0.130	0.664±0.159	1.360±0.448	2.110±0.396	0.104±0.024	0.137±0.027	0.247±0.040	0.340±0.034	72.3±2.8	58.0±6.4	45.3±5.4	31.1±5.0
Mamba + geochem + location	1	0.384±0.099	1.046±0.276	1.157±0.490	1.894±0.419	0.086±0.016	0.204±0.045	0.225±0.065	0.324±0.045	70.5±2.9	61.5±5.6	52.7±5.5	39.8±5.5
TCN + geochem + location	1	0.285±0.105	0.621±0.251	0.978±0.632	1.084±0.313	0.053±0.011	0.117±0.030	0.133±0.076	0.196±0.038	71.5±2.7	59.6±5.1	59.3±5.4	47.6±5.9
Layer level													
Layer marginal	1	0.211±0.086	0.563±0.146	0.653±0.341	1.072±0.303	0.043±0.013	0.118±0.024	0.113±0.046	0.202±0.041	51.5±2.6	49.2±4.3	51.1±5.2	43.0±5.0
Layer Markov	1	0.195±0.063	0.385±0.089	0.509±0.330	0.849±0.301	**0.035**±0.011	0.096±0.021	0.093±0.033	0.152±0.034	56.9±2.4	53.4±3.9	54.6±5.1	47.0±4.9
GRU	10	0.183±0.047	0.356±0.076	0.464±0.288	0.550±0.212	0.047±0.012	0.085±0.017	0.093±0.028	0.118±0.029	63.1±2.5	57.3±4.3	56.7±5.3	51.3±4.6
Transformer	10	**0.172**±0.043	0.320±0.073	0.495±0.327	0.544±0.212	0.045±0.012	0.077±0.017	0.095±0.030	0.114±0.029	62.9±2.5	57.4±4.4	56.4±5.3	50.5±4.7
GRU + R	10	0.192±0.058	0.341±0.076	0.479±0.311	0.561±0.211	0.048±0.014	0.083±0.020	0.094±0.029	0.118±0.029	64.7±2.5	57.4±4.3	56.8±5.2	51.5±4.6
GRU + PT	10	0.185±0.048	**0.303**±0.077	0.488±0.347	0.551±0.223	0.046±0.010	**0.072**±0.016	**0.090**±0.035	0.113±0.031	63.3±2.5	57.6±4.1	56.2±5.3	52.0±4.6
GRU + PT + R (ours)	10	0.186±0.046	0.322±0.076	**0.408**±0.259	**0.458**±0.172	0.052±0.013	0.081±0.020	0.090±0.029	**0.106**±0.027	63.8±2.3	56.3±4.2	56.1±5.2	51.1±4.8

Table 14: **Task 2 paired differences** in pooled joint KL and JSD (negative = first method better). Both methods are scored on the same report-cluster bootstrap resample (B=1000), seeds resampled independently. Intervals that exclude zero are bold. GRU and layer Transformer: the plain models with label smoothing (ε=0.3).

	Δ KL				Δ JSD			
Comparison	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
GRU + R − GRU	+0.008 [-0.023, +0.045]	-0.015 [-0.055, +0.024]	+0.014 [-0.088, +0.157]	+0.011 [-0.057, +0.093]	+0.001 [-0.010, +0.014]	-0.002 [-0.015, +0.013]	+0.001 [-0.006, +0.010]	-0.001 [-0.010, +0.009]
GRU + PT − GRU	+0.001 [-0.022, +0.026]	**-0.053 [-0.088, -0.019]**	+0.023 [-0.065, +0.211]	+0.000 [-0.085, +0.100]	-0.001 [-0.009, +0.006]	**-0.013 [-0.022, -0.004]**	-0.003 [-0.014, +0.011]	-0.005 [-0.016, +0.005]
Ours − GRU	+0.003 [-0.028, +0.042]	-0.034 [-0.072, +0.013]	-0.057 [-0.157, +0.038]	**-0.092 [-0.202, -0.016]**	+0.005 [-0.005, +0.017]	-0.004 [-0.015, +0.011]	-0.003 [-0.013, +0.010]	**-0.013 [-0.023, -0.001]**
Ours − GRU + R (adds PT)	-0.005 [-0.050, +0.040]	-0.019 [-0.067, +0.032]	-0.071 [-0.225, +0.022]	**-0.103 [-0.207, -0.023]**	+0.004 [-0.011, +0.017]	-0.002 [-0.019, +0.015]	-0.004 [-0.014, +0.009]	-0.012 [-0.023, +0.001]
Ours − GRU + PT (adds R)	+0.002 [-0.030, +0.039]	+0.019 [-0.025, +0.072]	-0.080 [-0.253, +0.018]	**-0.093 [-0.191, -0.019]**	+0.006 [-0.003, +0.017]	+0.009 [-0.002, +0.023]	+0.000 [-0.012, +0.016]	-0.007 [-0.019, +0.005]
Ours − Layer Transformer	+0.015 [-0.017, +0.049]	+0.003 [-0.037, +0.052]	**-0.087 [-0.254, -0.002]**	**-0.086 [-0.182, -0.016]**	+0.007 [-0.003, +0.019]	+0.004 [-0.007, +0.019]	-0.004 [-0.015, +0.008]	-0.008 [-0.022, +0.005]
Ours − Layer Markov	+0.036 [-0.014, +0.085]	-0.063 [-0.123, +0.001]	-0.102 [-0.312, +0.058]	**-0.391 [-0.586, -0.205]**	**+0.016 [+0.006, +0.028]**	-0.015 [-0.029, +0.002]	-0.003 [-0.051, +0.020]	**-0.046 [-0.070, -0.022]**
Ours − cell Transformer	**-0.170 [-0.255, -0.093]**	**-0.263 [-0.371, -0.152]**	**-0.125 [-0.294, -0.004]**	**-0.537 [-0.821, -0.310]**	**-0.028 [-0.045, -0.011]**	**-0.051 [-0.072, -0.027]**	-0.012 [-0.035, +0.005]	**-0.073 [-0.093, -0.048]**
Ours − XGBoost	-0.019 [-0.098, +0.042]	**-0.195 [-0.390, -0.047]**	**-0.542 [-0.874, -0.313]**	**-1.318 [-1.612, -1.089]**	+0.007 [-0.006, +0.022]	**-0.038 [-0.083, -0.002]**	**-0.095 [-0.158, -0.057]**	**-0.238 [-0.273, -0.202]**
Ours − 3-gram	+0.000 [-0.060, +0.054]	**-0.211 [-0.449, -0.113]**	**-0.373 [-0.729, -0.064]**	**-0.502 [-0.681, -0.311]**	+0.011 [-0.000, +0.024]	**-0.029 [-0.047, -0.011]**	-0.019 [-0.082, +0.015]	**-0.069 [-0.097, -0.042]**

Table 15: **Task 2 across layer-level backbones.** Each backbone is trained plain, with retrieval (R), with fine-objective pretraining (PT), and with both. Pooled joint KL, JSD and per-metre agreement. ±: half-width of the 95% report-cluster bootstrap interval (B=1000; holes of one exploration report are resampled together); trained rows also resample seeds. Plug-in divergences are biased upward under resampling, so intervals are re-centred on the full-sample estimate. Layer-level neural rows (and the plain arms of the backbone table) use label smoothing (ε=0.3).

		KL ↓				JSD ↓				Acc/m ↑			
Method	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
GRU													
GRU	10	0.183±0.050	0.356±0.076	0.464±0.278	0.550±0.207	0.047±0.011	0.085±0.017	0.093±0.028	0.118±0.029	63.1±2.4	57.3±4.4	56.7±5.2	51.3±4.6
+ R	10	0.192±0.059	0.341±0.081	0.479±0.306	0.561±0.212	0.048±0.014	0.083±0.020	0.094±0.029	0.118±0.029	**64.7**±2.5	57.4±4.4	56.8±5.3	51.5±4.7
+ PT	10	0.185±0.050	0.303±0.078	0.488±0.349	0.551±0.224	0.046±0.010	0.072±0.017	0.090±0.035	0.113±0.030	63.3±2.4	57.6±4.2	56.2±5.2	**52.0**±4.6
+ PT + R	10	0.186±0.047	0.322±0.074	0.408±0.264	**0.458**±0.169	0.052±0.013	0.081±0.020	0.090±0.030	**0.106**±0.027	63.8±2.4	56.3±4.4	56.1±5.2	51.1±4.5
LSTM													
LSTM	3	0.164±0.044	0.338±0.079	0.439±0.275	0.549±0.217	0.045±0.011	0.081±0.018	0.092±0.030	0.121±0.030	62.7±2.3	57.8±4.3	55.9±5.4	51.0±4.6
+ R	3	0.170±0.048	0.328±0.080	0.486±0.306	0.571±0.224	0.044±0.013	0.078±0.021	0.093±0.028	0.119±0.030	64.5±2.6	58.0±4.8	56.7±5.6	51.2±4.7
+ PT	3	0.173±0.041	0.321±0.072	0.403±0.236	0.494±0.172	0.049±0.012	0.083±0.019	0.092±0.027	0.118±0.029	62.3±2.5	57.2±4.4	55.6±5.2	50.1±4.5
+ PT + R	3	0.178±0.050	0.303±0.076	0.459±0.309	0.543±0.204	0.048±0.012	0.075±0.021	0.092±0.028	0.119±0.031	63.6±2.5	56.8±4.4	55.6±5.2	50.5±4.9
Transformer													
Transformer	10	0.172±0.045	0.320±0.072	0.495±0.333	0.544±0.210	0.045±0.012	0.077±0.017	0.095±0.030	0.114±0.027	62.9±2.6	57.4±4.2	56.4±5.3	50.5±4.6
+ R	3	0.182±0.045	0.321±0.070	0.441±0.263	0.509±0.173	0.049±0.014	0.080±0.019	0.092±0.026	0.118±0.029	63.9±2.4	57.6±4.6	56.4±5.5	49.8±4.9
+ PT	3	0.168±0.045	0.285±0.085	0.473±0.368	0.541±0.219	0.045±0.011	0.072±0.020	0.088±0.030	0.120±0.032	63.0±2.4	56.9±4.4	55.7±5.2	51.0±4.7
+ PT + R	3	0.175±0.047	**0.275**±0.074	0.451±0.308	0.521±0.189	0.047±0.015	0.071±0.021	0.090±0.030	0.113±0.030	64.1±2.5	57.2±4.3	56.2±5.6	49.9±5.0
TCN													
TCN	3	**0.161**±0.042	0.328±0.081	0.450±0.317	0.596±0.228	0.042±0.010	0.077±0.017	0.089±0.031	0.131±0.033	63.7±2.3	**58.5**±4.3	**56.9**±5.6	51.2±4.6
+ R	3	0.171±0.047	0.324±0.073	0.409±0.240	0.574±0.205	0.048±0.014	0.080±0.020	0.091±0.025	0.134±0.034	64.0±2.6	58.3±4.7	56.7±5.4	49.9±5.2
+ PT	3	0.204±0.049	0.349±0.078	**0.345**±0.163	0.531±0.163	0.059±0.015	0.092±0.020	0.089±0.026	0.131±0.030	62.0±2.4	57.9±4.4	55.3±5.2	48.8±4.6
+ PT + R	3	0.169±0.044	0.282±0.073	0.426±0.258	0.580±0.221	0.045±0.011	**0.068**±0.016	0.090±0.030	0.124±0.032	64.6±2.6	57.6±4.4	56.4±5.4	50.3±4.8
Mamba													
Mamba	3	0.175±0.043	0.346±0.074	0.451±0.257	0.587±0.195	0.048±0.012	0.088±0.019	0.096±0.027	0.133±0.034	62.3±2.4	57.1±4.5	55.5±5.3	48.7±5.3
+ R	3	0.167±0.042	0.316±0.073	0.467±0.283	0.599±0.216	0.045±0.011	0.076±0.017	0.095±0.028	0.128±0.032	63.7±2.4	57.6±4.5	55.7±5.4	49.3±5.2
+ PT	3	0.193±0.043	0.345±0.078	0.390±0.220	0.544±0.160	0.054±0.014	0.089±0.023	0.088±0.029	0.130±0.028	62.0±2.4	56.4±4.5	55.2±5.3	49.0±5.1
+ PT + R	3	0.162±0.051	0.297±0.084	0.456±0.326	0.597±0.219	**0.042**±0.010	0.069±0.017	**0.086**±0.029	0.126±0.033	64.3±2.4	57.4±4.2	56.6±5.3	49.7±4.7

Split	**Points**	**w/ retrieval**	**median dist.**
`val`	466	46.6%	115 m
`test`	361	43.8%	137 m
`ood_near`	471	15.7%	1,503 m
`ood_far`	401	18.7%	498 m
`pilbara_test`	462	32.0%	1,202 m
Total	2,161		

Table 16: Prediction points per split (150 holes each) and, for the **retrieval** arm, the share of points with any neighbour evidence at that depth and the median weighted neighbour distance among them.

Model	Access	Read-out	Precision	Hardware / setting
Qwen3-8B	open	likelihood+PMI	fp16	V100 32 GB
Qwen3-14B	open	likelihood+PMI	bf16	H100
Qwen3-32B	open	likelihood+PMI	bf16	H100
phi-4 (14B)	open	likelihood+PMI	bf16	H100
Qwen3-8B, 32B	open	generative bridge	bf16	A6000, greedy
GPT-6-astra	hosted	generative	–	batch, effort medium
Gemini-pro-latest	hosted	generative	–	batch, thinking medium

Table 17: Zero-shot configurations. Llama and Gemma were not run because their weights require manual approval.

Model	Arm	val	test	ood_r	ood_d	pilbara_test	parse fail.
Likelihood + PMI							
Qwen3-8B	strict	14.6 / 36.9	14.4 / 32.7	20.6 / 41.4	24.2 / 50.1	18.8 / 32.7	0
	geo	15.4 / 30.3	17.2 / 26.6	21.7 / 25.5	29.7 / 31.9	8.2 / 26.6	0
	retrieval	29.2 / 54.5	33.0 / 52.1	25.5 / 52.2	37.7 / 51.1	23.4 / 44.6	0
Qwen3-14B	strict	29.0 / 44.2	24.4 / 33.2	31.2 / 39.5	38.4 / 51.4	42.4 / 60.2	0
	geo	21.0 / 35.4	16.6 / 25.2	22.9 / 35.2	29.7 / 43.1	14.5 / 38.5	0
	retrieval	48.7 / 63.7	51.5 / 58.7	36.1 / 46.7	49.1 / 56.6	40.5 / 59.7	0
Qwen3-32B	strict	22.8 / 46.6	20.8 / 37.4	29.3 / 52.6	32.7 / 41.1	35.9 / 50.4	0
	geo	14.4 / 34.3	10.8 / 26.9	21.0 / 27.0	12.2 / 29.9	28.1 / 43.7	0
	retrieval	44.6 / 67.0	48.5 / 64.0	36.5 / 43.7	39.6 / 49.1	32.9 / 55.4	0
phi-4	strict	26.0 / 38.4	21.1 / 31.3	29.3 / 48.0	36.2 / 48.9	30.3 / 41.3	0
	geo	6.2 / 17.4	2.2 / 14.1	5.1 / 14.2	12.5 / 33.4	18.2 / 23.4	0
	retrieval	44.2 / 54.5	47.4 / 55.4	29.5 / 51.2	37.4 / 54.6	34.2 / 44.6	0
Generative (bridge)							
Qwen3-8B	strict	29.0 / 49.6	23.3 / 44.6	30.1 / 48.8	35.4 / 53.9	45.5 / 63.4	0
	geo	25.1 / 34.3	20.2 / 29.1	24.6 / 31.4	29.7 / 35.4	32.7 / 44.2	0
	retrieval	39.7 / 54.3	35.2 / 44.9	32.1 / 40.8	41.9 / 54.1	44.4 / 59.1	0
Qwen3-32B	strict	23.4 / 33.9	16.1 / 25.8	7.9 / 10.2	17.2 / 20.2	26.8 / 40.7	0
	geo	17.8 / 28.5	9.7 / 14.7	14.9 / 20.0	15.5 / 23.9	27.1 / 43.1	0
	retrieval	42.9 / 56.4	47.4 / 55.1	24.6 / 32.7	30.4 / 37.9	29.4 / 47.0	0
Generative (hosted)							
GPT-6-astra	strict	38.6 / 56.9	30.5 / 46.5	30.8 / 46.3	43.6 / 61.1	47.4 / 59.3	0
	geo	41.6 / 60.9	35.2 / 56.2	27.2 / 53.5	44.4 / 70.1	40.3 / 53.7	0
	retrieval	55.4 / 71.9	53.2 / 67.0	38.0 / 56.9	51.1 / 68.1	45.9 / 64.1	0
Gemini-pro-latest	strict	9.7 / 21.5	8.9 / 21.6	11.2 / 22.3	20.7 / 31.9	5.4 / 14.9	0
	geo	13.7 / 30.0	14.7 / 30.5	15.7 / 25.3	16.3 / 31.0	4.1 / 14.1	26
	retrieval†	55.1 / 74.5	59.7 / 73.7	32.0 / 51.8	37.2 / 55.5	50.0 / 58.3	1,409

Table 18: Zero-shot LLM next-layer top-1 / top-3 (%) on the 150-hole samples of the original five splits; 2,161 points per row. Parse failures are excluded from the denominator. †Scored on the 752 points that parsed; not a capability estimate.

Table 19: **Settings of the recipe**: deployed value, the alternatives tested, and where. Settings without a table entry were fixed a priori and not tuned.

Setting	Deployed	Tested	Where
Pretraining			
Corpus	all 112k holes	1.1k–67k holes; 1–30 provinces; region held out	Tab. 21, 23, 24
Optimisation length	24 epochs	compute-matched fractions	Tab. 22
Objective	fine + chemistry	fine only; coarse $\pm$ chemistry	Tab. 25
Retrieval			
Neighbours k	10	1–40; 1 km radius; consensus	Tab. 27
Pool	train + archive	none; train only; + target context	Tab. 28, 29
Neighbour dropout	0.3	0, 0.6	Tab. 30
Self-exclusion radius	25 m	none	Tab. 30
Depth alignment	elevation	hole depth	Tab. 30
Fusion	additive	learned gate	Tab. 30
Neighbour votes	at bed starts	over whole beds	Tab. 30
Model and data			
Token	stratum (layer)	1 m cell	Tab. 31
Backbone	GRU	LSTM, Transformer, TCN, Mamba	Tab. 11, 15
Inputs	lithology	+ thickness, depth	Tab. 20
Label smoothing	0.3	none	Tab. 20
Grid	1 m	0.5 m, 2 m	Tab. 32
Labelled holes	all of TRAIN	1k, 3k, 9k	Tab. 26
Learning rate	3×10^{-4}	shared by all backbones	–
Fine-tuning length	20 epochs	no checkpoint selection	–
Generation (Task 2 only)			
Boundary-rate scale s	fitted on TRAIN	every generator	App. C.2
Lithology temperature τ	0.7, set on TRAIN	–	App. C.2

Table 20: **Cumulative components.** Layer-level GRU, 10 seeds per row; Δ vs. the previous row. $\pm$: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with $\geq$20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
A0: layer GRU, lithology only	10	59.5$\pm$4.8	34.9$\pm$3.3	26.1$\pm$3.8	26.5$\pm$2.6	58.6$\pm$1.1	40.2$\pm$2.1	34.5$\pm$3.1	40.2$\pm$1.4	–	–	–	–
A1: + label smoothing	10	59.4$\pm$5.0	37.4$\pm$3.5	28.3$\pm$4.0	28.2$\pm$3.2	58.6$\pm$1.1	40.5$\pm$2.0	34.7$\pm$3.2	42.0$\pm$1.3	-0.1 [-0.8, +0.5]	**+2.6 [+1.3, +3.6]**	**+2.2 [+1.2, +3.1]**	**+1.7 [+0.2, +4.4]**
A2: + retrieval	10	70.0$\pm$5.0	37.4$\pm$3.3	33.9$\pm$4.6	26.9$\pm$3.0	73.5$\pm$0.9	52.5$\pm$2.2	46.9$\pm$3.3	42.8$\pm$1.1	**+10.7 [+8.9, +12.6]**	-0.1 [-1.1, +2.2]	**+5.6 [+4.4, +7.0]**	-1.3 [-3.0, +0.6]
A3: + pretraining (full recipe)	10	69.2$\pm$4.9	38.8$\pm$3.5	37.2$\pm$4.8	30.6$\pm$2.9	73.7$\pm$0.9	52.9$\pm$2.2	48.5$\pm$3.4	44.3$\pm$1.1	-0.8 [-1.8, +2.8]	**+1.4 [+0.5, +2.5]**	**+3.4 [+2.1, +4.6]**	**+3.7 [+2.3, +5.3]**
A4: + thickness / depth inputs	10	71.3$\pm$4.9	39.6$\pm$4.2	37.3$\pm$5.0	29.0$\pm$3.0	75.1$\pm$0.9	53.1$\pm$2.4	47.9$\pm$3.0	44.2$\pm$1.1	+2.1 [-1.3, +3.1]	+0.8 [-0.6, +4.1]	+0.1 [-1.9, +1.3]	**-1.6 [-3.2, -0.5]**

Table 21: **Pretraining-corpus size.** Nested random subsets of the 112k-hole pretraining corpus, fine objective, 24 pretraining epochs each; Δ vs. no pretraining in the same retrieval setting. $\pm$: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with $\geq$20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
no PT, no R	3	59.5±4.9	37.2±3.6	28.5±4.2	29.0±3.0	58.5±1.1	41.4±2.4	36.3±4.3	42.0±1.4	–	–	–	–
1.1k, no R	3	59.4±4.9	38.2±3.3	29.9±4.3	26.5±3.7	58.5±1.1	41.6±2.3	36.3±4.6	41.0±1.4	-0.1 [-1.4, +1.2]	**+1.0 [+0.1, +2.2]**	+1.3 [-0.3, +2.9]	-2.5 [-4.6, +0.4]
3.4k, no R	3	59.2±5.0	38.0±3.4	30.6±4.5	25.2±3.2	58.5±1.1	41.6±2.4	36.3±4.3	41.2±1.5	-0.3 [-1.6, +1.0]	+0.8 [-0.4, +2.0]	**+2.1 [+0.5, +3.9]**	-3.8 [-5.0, +0.7]
11k, no R	3	59.4±4.9	38.7±3.6	30.4±4.4	26.8±3.2	58.5±1.1	41.7±2.3	36.4±4.3	42.7±1.4	-0.1 [-1.2, +1.0]	**+1.5 [+0.3, +2.8]**	**+1.8 [+0.5, +3.3]**	-2.2 [-3.6, +2.5]
34k, no R	3	59.1±5.0	39.1±3.5	32.1±4.6	29.5±3.6	58.5±1.1	41.7±2.4	36.7±4.2	43.0±1.3	-0.4 [-1.7, +1.0]	**+1.9 [+1.0, +3.9]**	**+3.5 [+2.0, +5.0]**	+0.6 [-1.8, +2.6]
67k, no R	3	59.0±5.0	39.0±3.3	34.1±4.7	31.7±3.2	58.5±1.1	42.0±2.5	37.3±4.4	43.5±1.5	-0.5 [-1.9, +1.1]	**+1.8 [+1.1, +4.1]**	**+5.5 [+3.7, +7.1]**	**+2.7 [+1.3, +4.6]**
112k, no R	3	58.7±5.1	39.1±3.3	32.9±4.7	31.8±3.1	58.5±1.1	42.1±2.3	37.1±4.5	43.5±1.6	-0.8 [-2.2, +0.8]	**+1.9 [+1.2, +3.5]**	**+4.4 [+2.8, +6.0]**	**+2.8 [+1.5, +4.6]**
no PT, + R	3	69.5±5.1	38.3±3.7	33.9±4.6	26.9±3.4	73.4±1.0	52.4±2.2	46.2±4.4	42.8±1.2	–	–	–	–
1.1k, + R	3	69.3±5.1	38.4±3.6	35.7±4.8	26.0±3.6	73.4±0.9	52.3±2.3	46.1±4.4	42.8±1.3	-0.2 [-1.5, +0.9]	+0.1 [-0.7, +1.0]	+1.7 [-0.1, +3.5]	-0.9 [-3.7, +2.7]
3.4k, + R	3	69.5±5.0	38.6±3.7	35.8±4.8	25.7±3.5	73.4±1.0	52.5±2.2	46.2±4.5	42.7±1.4	+0.1 [-1.3, +1.0]	+0.3 [-0.6, +1.5]	**+1.9 [+0.1, +3.7]**	-1.1 [-3.8, +2.4]
11k, + R	3	70.1±5.0	38.9±3.7	37.1±5.1	27.7±3.5	73.6±0.9	53.0±2.2	48.6±4.0	43.4±1.2	+0.6 [-0.6, +1.3]	+0.6 [-0.5, +2.7]	**+3.1 [+1.4, +5.0]**	+0.8 [-2.1, +4.1]
34k, + R	3	69.7±5.1	38.7±3.5	38.1±5.0	30.1±2.9	73.5±1.0	53.0±2.3	48.8±4.0	44.2±1.1	+0.2 [-1.0, +1.1]	+0.4 [-0.3, +1.6]	**+4.2 [+2.5, +5.9]**	**+3.2 [+1.0, +6.1]**
67k, + R	3	69.3±4.8	39.9±3.9	38.4±5.1	31.1±2.9	73.6±0.9	53.1±2.2	49.1±4.1	44.7±1.2	-0.2 [-1.5, +3.5]	**+1.6 [+0.6, +3.7]**	**+4.4 [+2.5, +6.2]**	**+4.3 [+2.0, +7.2]**
112k, + R	3	69.1±5.2	40.0±4.0	37.3±4.9	31.2±2.9	73.7±0.9	53.1±2.3	49.2±4.0	44.6±1.2	-0.4 [-1.7, +2.4]	**+1.7 [+0.5, +3.9]**	**+3.3 [+1.8, +5.1]**	**+4.3 [+2.1, +7.4]**

Table 22: **Compute-matched corpus size.** Corpus fractions pretrained for $24/\text{fraction}$ epochs, so that every checkpoint takes the same number of optimisation steps; Δ vs. the deployed full-corpus checkpoint in the same retrieval setting. $\pm$: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with $\geq$20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
fine + chem (deployed), no R	10	58.5±5.1	39.2±3.3	33.0±4.5	31.7±2.9	58.6±1.1	40.8±2.1	35.4±2.9	44.0±1.2	–	–	–	–
1% (2400 ep), no R	3	58.6±4.9	35.6±3.7	27.5±3.8	28.1±2.9	58.2±1.1	41.1±2.4	34.9±4.4	41.8±1.5	+0.1 [-1.1, +1.2]	**-3.6 [-6.5, -2.5]**	**-5.5 [-7.4, -2.7]**	**-3.6 [-5.6, -2.1]**
3% (800 ep), no R	3	58.7±5.1	39.0±3.6	29.4±4.4	25.4±3.1	58.5±1.1	41.4±2.4	35.8±4.3	41.5±1.4	+0.2 [-0.6, +1.2]	-0.2 [-1.9, +0.5]	**-3.6 [-5.0, -2.1]**	**-6.3 [-7.7, -2.7]**
10% (240 ep), no R	3	57.9±5.1	39.2±3.8	33.2±4.7	31.6±2.9	58.5±1.1	40.7±2.2	35.5±4.6	43.6±1.6	-0.5 [-1.3, +1.2]	+0.0 [-1.8, +1.0]	+0.2 [-1.1, +1.5]	-0.1 [-1.6, +1.1]
30% (80 ep), no R	3	58.7±5.1	39.9±3.6	33.5±4.6	29.7±3.0	58.5±1.1	42.2±2.4	37.3±4.5	42.6±1.5	+0.3 [-0.3, +0.9]	+0.7 [-0.5, +1.3]	+0.5 [-0.8, +1.7]	**-2.0 [-2.9, -1.0]**
60% (40 ep), no R	3	58.7±5.1	39.3±3.6	33.4±4.5	31.8±2.9	58.6±1.1	42.1±2.4	37.2±4.1	43.7±1.4	+0.3 [-0.7, +1.1]	+0.1 [-0.8, +0.8]	+0.4 [-0.8, +1.5]	+0.0 [-0.8, +1.0]
fine + chem (deployed), + R	10	69.2±5.0	38.8±3.5	37.2±4.9	30.6±3.0	73.7±0.9	52.9±2.2	48.5±3.3	44.3±1.1	–	–	–	–
1% (2400 ep), + R	3	68.0±5.3	33.9±3.4	31.6±4.3	27.3±2.7	73.2±1.0	51.2±2.5	48.1±2.7	42.7±1.1	**-1.3 [-3.9, -0.3]**	**-4.9 [-7.5, -3.2]**	**-5.6 [-7.7, -3.1]**	**-3.4 [-5.6, -2.0]**
3% (800 ep), + R	3	69.3±5.3	38.3±3.4	36.1±5.1	24.2±2.7	73.7±1.0	52.0±2.5	50.1±2.9	41.8±1.2	+0.1 [-3.0, +0.9]	-0.5 [-1.9, +0.4]	-1.1 [-2.7, +0.7]	**-6.5 [-7.4, -2.5]**
10% (240 ep), + R	3	68.5±5.2	40.1±3.7	37.9±5.1	30.9±2.8	73.2±1.0	51.8±2.4	47.3±4.3	44.4±1.2	-0.8 [-1.6, +1.5]	+1.3 [-0.1, +2.2]	+0.6 [-0.6, +2.0]	+0.3 [-1.0, +1.3]
30% (80 ep), + R	3	68.9±5.4	39.4±3.7	38.0±4.9	30.1±3.0	73.6±0.9	53.0±2.3	49.4±4.2	44.1±1.2	-0.3 [-3.3, +0.4]	+0.7 [-0.6, +1.5]	+0.7 [-0.0, +1.4]	-0.5 [-1.8, +0.6]
60% (40 ep), + R	3	69.0±5.2	39.1±3.7	37.8±4.9	30.3±3.1	73.6±1.0	52.6±2.4	49.2±4.3	44.6±1.2	-0.3 [-3.4, +0.2]	+0.3 [-0.9, +1.2]	+0.6 [-0.0, +1.2]	-0.4 [-0.9, +0.5]

Table 23: **Province diversity of the pretraining corpus.** 60k-hole corpora drawn from the n largest provinces; Δ vs. the deployed full-corpus checkpoint in the same retrieval setting. $\pm$: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with $\geq$20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
fine + chem (deployed), no R	10	58.5±5.1	39.2±3.3	33.0±4.5	31.7±2.9	58.6±1.1	40.8±2.1	35.4±2.9	44.0±1.2	–	–	–	–
1 province, no R	3	59.6±4.9	38.4±3.5	30.3±4.5	25.3±3.1	58.5±1.1	41.7±2.3	36.4±4.5	42.0±1.4	+1.2 [-0.8, +2.3]	**-0.8 [-2.1, -0.1]**	**-2.7 [-4.2, -1.3]**	**-6.4 [-7.5, -2.4]**
2 provinces, no R	3	59.2±4.9	38.7±3.8	30.6±4.3	26.4±3.2	58.5±1.1	41.6±2.4	36.5±4.1	42.2±1.3	+0.7 [-0.8, +2.1]	-0.4 [-2.3, +0.1]	**-2.4 [-3.7, -1.2]**	**-5.4 [-6.7, -1.7]**
4 provinces, no R	3	59.0±5.1	38.8±3.6	31.9±4.5	26.7±3.1	58.5±1.1	41.6±2.4	36.6±4.2	42.7±1.5	+0.6 [-0.9, +1.8]	-0.4 [-1.6, +0.2]	-1.2 [-2.5, +0.0]	**-5.0 [-6.4, -1.5]**
8 provinces, no R	3	59.4±5.0	39.0±3.6	30.6±4.4	26.4±3.2	58.5±1.1	41.7±2.3	36.4±4.2	42.3±1.4	+0.9 [-0.5, +2.0]	-0.2 [-1.5, +0.4]	**-2.4 [-3.7, -1.2]**	**-5.4 [-6.5, -1.4]**
30 provinces, no R	3	59.2±4.9	38.2±3.3	30.0±4.4	28.5±3.5	58.5±1.1	41.5±2.5	36.1±4.6	42.5±1.4	+0.8 [-0.8, +2.0]	**-1.0 [-2.0, -0.3]**	**-3.0 [-4.6, -1.5]**	**-3.2 [-5.3, -1.4]**
fine + chem (deployed), + R	10	69.2±5.0	38.8±3.5	37.2±4.9	30.6±3.0	73.7±0.9	52.9±2.2	48.5±3.3	44.3±1.1	–	–	–	–
1 province, + R	3	69.7±5.0	38.3±3.6	35.1±4.6	26.6±3.6	73.4±0.9	52.3±2.2	46.3±4.3	43.5±1.4	+0.4 [-3.0, +1.6]	-0.5 [-1.8, +0.7]	**-2.2 [-3.6, -0.9]**	**-4.1 [-6.7, -1.4]**
2 provinces, + R	3	69.7±4.9	39.0±3.7	35.9±4.9	25.9±3.0	73.5±1.0	53.0±2.1	48.1±3.9	43.1±1.4	+0.4 [-3.1, +1.6]	+0.2 [-1.6, +1.5]	-1.3 [-2.9, +0.1]	**-4.8 [-6.3, -1.7]**
4 provinces, + R	3	69.5±5.0	38.5±3.5	38.0±5.0	27.4±3.3	73.5±1.0	52.9±2.2	48.8±4.1	43.8±1.2	+0.3 [-3.2, +1.3]	-0.3 [-1.4, +0.4]	+0.8 [-0.5, +2.1]	**-3.2 [-5.5, -1.2]**
8 provinces, + R	3	70.0±5.0	39.1±3.8	36.8±4.9	29.1±3.0	73.6±0.9	52.8±2.2	48.6±4.0	44.1±1.2	+0.8 [-2.8, +1.8]	+0.3 [-1.8, +1.7]	-0.4 [-2.1, +0.9]	**-1.6 [-2.7, -0.4]**
30 provinces, + R	3	70.2±5.0	39.2±3.7	36.7±5.0	29.4±3.0	73.6±1.0	53.0±2.2	48.3±4.2	43.8±1.2	+1.0 [-2.6, +1.9]	+0.4 [-1.3, +1.5]	-0.6 [-2.3, +1.0]	**-1.2 [-2.3, -0.2]**

Table 24: **Province-held-out pretraining**, Task 1 on DrillBench, layer-level GRU, 10 seeds. Same convention as Table 11: macro-F1 over classes with ≥20 events, top-1 macro-averaged over reports on the two rings; ±: half-width of the 95% hole-cluster bootstrap interval. Δ rows: paired difference to the full-archive row above (same resample); intervals that exclude zero are bold. All rows use label smoothing (ε=0.3).

	macro-F1 (%)				top-1 (%)			
Pretraining corpus	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
Without retrieval								
None (GRU)	59.4±4.9	37.4±3.5	28.3±3.9	28.2±3.0	58.6±1.1	40.5±1.9	34.7±3.1	42.0±1.4
Full archive (GRU + PT)	58.5±5.1	39.2±3.2	33.0±4.3	31.7±2.9	58.6±1.1	40.8±1.8	35.4±3.0	44.0±1.3
Pilbara held out	58.1±5.1	38.3±3.4	32.1±4.2	29.8±2.9	58.5±1.1	40.7±1.9	35.2±3.0	43.8±1.3
Δ vs. full archive	-0.3 [-1.0, +0.4]	**-0.9 [-1.9, -0.1]**	-0.9 [-1.9, +0.0]	**-1.9 [-2.6, -0.9]**	-0.1 [-0.3, +0.2]	-0.1 [-1.1, +0.9]	-0.2 [-3.4, +2.7]	-0.2 [-1.2, +0.7]
Pilbara and EG held out	58.5±5.2	38.5±3.3	31.9±4.1	30.8±2.7	58.6±1.1	40.9±1.9	35.6±3.1	43.4±1.3
Δ vs. full archive	-0.0 [-0.9, +0.8]	-0.7 [-1.6, +0.2]	**-1.1 [-1.9, -0.3]**	-0.9 [-3.0, +0.2]	-0.0 [-0.3, +0.2]	+0.1 [-1.0, +1.0]	+0.2 [-2.6, +3.8]	-0.6 [-1.6, +0.4]
With retrieval								
None (GRU + R)	70.0±5.0	37.4±3.4	33.9±4.5	26.9±3.1	73.5±0.9	52.5±2.0	46.9±3.3	42.8±1.2
Full archive (GRU + PT + R)	69.2±4.9	38.8±3.5	37.2±4.6	30.6±3.0	73.7±0.9	52.9±2.1	48.5±3.3	44.3±1.2
Pilbara held out	69.5±5.1	38.6±3.3	37.5±4.8	30.0±2.8	73.6±0.9	52.5±2.1	48.5±3.2	44.4±1.2
Δ vs. full archive	+0.3 [-3.0, +0.8]	-0.2 [-0.9, +0.7]	+0.3 [-0.7, +1.2]	-0.7 [-1.8, +0.2]	-0.1 [-0.5, +0.3]	-0.4 [-1.2, +0.3]	-0.0 [-2.7, +2.6]	+0.1 [-0.5, +0.6]
Pilbara and EG held out	69.5±5.3	39.3±3.5	37.1±4.7	30.5±2.5	73.6±0.9	52.4±2.2	48.3±3.2	44.9±1.2
Δ vs. full archive	+0.2 [-2.6, +0.8]	+0.5 [-0.6, +1.3]	-0.2 [-1.2, +0.8]	-0.2 [-2.1, +0.9]	-0.1 [-0.5, +0.3]	-0.5 [-1.4, +0.1]	-0.2 [-2.8, +2.2]	+0.6 [-0.1, +1.2]

Table 25: **Pretraining objective.** Full corpus; fine: 37-code targets with coarse fallback; coarse: 13-class targets; chem: auxiliary geochemistry losses. Δ vs. the deployed checkpoint (fine + chem) in the same retrieval setting. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
fine + chem (deployed), no R	10	58.5±5.1	39.2±3.3	33.0±4.5	31.7±2.9	58.6±1.1	40.8±2.1	35.4±2.9	44.0±1.2	–	–	–	–
fine, chem losses off, no R	3	58.8±5.1	40.6±3.3	33.7±4.7	31.1±3.1	58.6±1.1	42.0±2.2	37.5±4.3	44.7±1.7	+0.3 [-1.1, +1.6]	**+1.4 [+0.8, +3.0]**	+0.7 [-0.4, +1.7]	-0.7 [-1.5, +0.8]
coarse + chem, no R	3	58.0±5.6	36.2±3.6	26.6±4.2	27.6±2.5	58.5±1.1	41.6±2.3	35.7±4.3	43.5±1.6	-0.4 [-2.6, +1.0]	**-3.0 [-4.8, -1.2]**	**-6.4 [-7.9, -4.4]**	**-4.1 [-7.8, -2.5]**
coarse, no chemistry, no R	3	57.4±5.7	35.7±3.2	26.2±4.1	27.4±2.8	58.4±1.2	40.4±2.2	34.6±4.3	42.8±1.8	-1.0 [-3.1, +0.6]	**-3.4 [-4.9, -1.3]**	**-6.8 [-8.1, -4.8]**	**-4.4 [-8.3, -2.7]**
fine + chem (deployed), + R	10	69.2±5.0	38.8±3.5	37.2±4.9	30.6±3.0	73.7±0.9	52.9±2.2	48.5±3.3	44.3±1.1	–	–	–	–
fine, chem losses off, + R	3	69.3±5.2	39.2±3.5	38.2±5.0	31.6±3.1	73.6±1.0	52.7±2.3	49.0±4.2	45.4±1.4	+0.0 [-3.4, +0.7]	+0.4 [-0.6, +1.7]	+1.0 [-0.3, +2.0]	+1.0 [-0.0, +2.2]
coarse + chem, + R	3	68.1±6.0	36.7±3.5	31.2±4.8	27.4±2.5	73.5±1.0	52.2±2.2	45.6±4.5	45.3±1.5	-1.1 [-4.2, +0.9]	**-2.1 [-3.7, -1.0]**	**-6.0 [-8.0, -3.2]**	**-3.3 [-7.2, -1.7]**
coarse, no chemistry, + R	3	67.9±5.9	36.4±3.6	30.6±4.5	26.3±2.3	73.3±1.0	51.8±2.3	45.8±4.5	43.5±1.4	-1.3 [-4.4, +1.0]	**-2.4 [-3.7, -1.3]**	**-6.6 [-8.0, -4.6]**	**-4.3 [-8.5, -2.8]**

Table 26: **Labelled training holes.** Random subsets of TRAIN; the labelled part of the retrieval pool is restricted to the same subset and the archive part is unchanged; Δ vs. the plain model at the same size. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
1k, plain	3	42.5±5.5	17.0±2.9	19.4±2.8	17.4±3.6	53.0±1.2	35.0±2.1	31.0±2.8	34.0±1.9	–	–	–	–
1k, + PT	3	56.5±4.9	38.3±3.7	29.7±3.8	35.9±2.8	57.4±1.2	39.3±2.1	32.5±2.6	45.7±1.6	**+14.1 [+7.9, +18.8]**	**+21.3 [+15.7, +24.4]**	**+10.3 [+8.0, +14.6]**	**+18.5 [+14.8, +22.6]**
1k, + R	3	43.6±5.8	19.9±3.9	20.0±3.4	17.8±3.2	56.7±1.1	39.2±2.2	35.8±3.9	34.3±2.3	+1.1 [-4.0, +6.5]	+2.9 [-1.8, +6.6]	+0.6 [-1.7, +3.0]	+0.4 [-3.8, +4.7]
1k, + PT + R	3	58.7±4.8	36.5±3.1	31.9±4.0	36.7±2.9	61.4±1.5	45.8±2.6	43.1±2.8	45.5±2.0	**+16.2 [+10.2, +20.7]**	**+19.5 [+15.9, +22.9]**	**+12.5 [+9.9, +16.7]**	**+19.4 [+15.5, +23.5]**
3k, plain	3	50.2±5.9	23.6±2.8	23.4±3.7	24.0±2.3	56.9±1.2	37.1±2.1	31.3±2.6	38.9±1.6	–	–	–	–
3k, + PT	3	57.3±5.0	40.5±3.4	31.3±3.8	35.6±3.2	58.0±1.2	40.2±2.1	33.2±2.8	45.4±1.5	**+7.1 [+1.8, +11.5]**	**+16.9 [+12.0, +18.6]**	**+7.9 [+5.8, +11.0]**	**+11.6 [+9.3, +15.9]**
3k, + R	3	56.2±6.3	24.8±2.8	27.6±3.9	21.8±2.3	66.0±1.1	48.9±2.2	44.5±4.8	38.2±1.6	**+6.0 [+2.4, +10.1]**	+1.3 [-0.1, +4.8]	**+4.1 [+2.1, +5.9]**	-2.2 [-3.7, +0.8]
3k, + PT + R	3	62.5±5.2	41.3±3.3	33.5±4.2	36.4±3.0	66.6±1.1	48.4±2.7	40.6±3.1	46.0±1.3	**+12.3 [+7.6, +16.5]**	**+17.7 [+13.4, +19.6]**	**+10.1 [+7.7, +12.5]**	**+12.4 [+10.3, +16.5]**
9k, plain	3	57.4±5.1	30.8±5.5	26.4±3.9	26.3±2.9	58.2±1.1	39.0±1.9	33.6±4.2	40.4±1.7	–	–	–	–
9k, + PT	3	58.1±4.8	40.4±3.5	33.1±4.7	34.4±2.8	58.5±1.1	39.9±2.1	33.8±2.7	44.6±1.3	+0.7 [-1.0, +5.9]	**+9.5 [+4.8, +14.3]**	**+6.7 [+4.4, +8.8]**	**+8.1 [+5.9, +10.9]**
9k, + R	3	65.5±5.5	33.0±5.0	29.6±4.2	23.5±4.6	70.7±1.0	50.0±2.4	44.0±4.6	40.1±1.6	**+8.1 [+6.2, +10.6]**	+2.2 [-4.2, +7.7]	**+3.2 [+1.7, +4.8]**	-2.8 [-6.4, +1.9]
9k, + PT + R	3	66.0±5.1	39.1±3.9	37.4±4.6	33.9±3.1	70.8±1.0	51.3±2.4	46.7±4.5	45.0±1.5	**+8.6 [+6.7, +14.3]**	**+8.3 [+3.2, +14.2]**	**+11.0 [+8.8, +13.4]**	**+7.6 [+5.3, +10.8]**
27k (all), plain	3	59.5±4.8	37.2±3.5	28.5±4.3	29.0±3.0	58.5±1.1	41.4±2.4	36.3±4.2	42.0±1.3	–	–	–	–
27k (all), + PT	3	58.8±5.0	39.0±3.2	33.4±4.7	31.6±2.9	58.5±1.1	42.0±2.4	37.3±4.4	43.4±1.5	-0.7 [-2.3, +1.1]	**+1.8 [+0.8, +3.4]**	**+4.8 [+3.3, +6.3]**	**+2.6 [+1.5, +4.2]**
27k (all), + R	3	69.5±4.9	38.3±3.6	33.9±4.8	26.9±3.5	73.4±0.9	52.4±2.3	46.2±4.5	42.8±1.2	**+10.0 [+8.1, +12.4]**	+1.1 [-0.6, +2.7]	**+5.4 [+3.8, +6.9]**	-2.1 [-4.8, +0.6]
27k (all), + PT + R	3	69.2±5.1	39.1±3.6	37.5±4.9	31.1±3.0	73.7±1.0	53.0±2.4	49.1±4.1	44.4±1.3	**+9.7 [+7.7, +12.8]**	**+1.9 [+0.2, +4.2]**	**+9.0 [+7.3, +10.7]**	**+2.1 [+0.7, +4.4]**

Table 27: **Number of retrieved neighbours.** Full recipe; radius: neighbours restricted to 1 km; consensus: candidates re-ranked by agreement with each other. Δ vs. the deployed k=10. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
no retrieval	3	58.8±5.1	39.0±3.2	33.4±4.6	31.6±2.9	58.5±1.1	42.0±2.4	37.3±4.3	43.4±1.4	**-10.4 [-13.3, -8.9]**	-0.1 [-1.6, +1.3]	**-4.1 [-5.5, -2.9]**	+0.5 [-1.3, +1.9]
k=1	3	63.1±5.0	39.9±3.5	35.6±4.9	32.5±3.0	64.6±1.0	47.1±2.4	42.8±4.5	44.7±1.4	**-6.1 [-7.2, -3.5]**	+0.8 [-0.9, +1.9]	**-1.9 [-3.2, -1.1]**	**+1.4 [+0.5, +2.4]**
k=3	3	65.7±5.1	39.6±3.7	35.5±4.9	31.5±3.1	68.9±1.0	48.7±2.3	44.6±4.1	44.6±1.3	**-3.4 [-4.4, -0.6]**	+0.4 [-1.1, +1.3]	**-2.0 [-3.1, -1.4]**	+0.3 [-0.9, +1.6]
k=5	3	66.7±5.1	38.7±3.5	36.7±4.9	32.3±3.0	70.9±0.9	50.3±2.5	46.9±4.2	44.4±1.3	**-2.5 [-3.4, -0.6]**	-0.4 [-1.2, +0.7]	**-0.8 [-1.4, -0.2]**	**+1.2 [+0.3, +2.1]**
k=10	3	69.2±5.1	39.1±3.6	37.5±5.0	31.1±2.9	73.7±0.9	53.0±2.3	49.1±4.0	44.4±1.3	–	–	–	–
k=20	3	70.1±5.4	37.8±3.5	37.3±5.0	30.4±2.9	76.1±1.0	54.1±2.6	49.2±3.6	44.2±1.2	+0.9 [-2.4, +1.9]	-1.4 [-2.3, +0.4]	-0.2 [-1.3, +0.5]	-0.7 [-1.4, +0.0]
k=40	3	71.1±5.2	38.5±3.8	38.3±5.2	30.0±3.1	77.5±0.9	53.9±2.6	50.2±3.3	44.4±1.3	+1.9 [-0.4, +3.1]	-0.6 [-2.3, +0.5]	+0.8 [-0.1, +1.6]	**-1.1 [-1.8, -0.1]**
k=10, radius 1 km	3	69.3±4.9	39.3±3.7	36.1±4.8	31.3±3.1	73.6±0.9	48.3±2.3	46.6±4.8	44.0±1.2	+0.2 [-0.7, +2.0]	+0.2 [-1.2, +1.1]	**-1.4 [-2.0, -0.9]**	+0.2 [-0.5, +1.2]
k=10, consensus	3	68.2±5.3	36.9±3.2	37.2±5.0	31.7±2.9	72.1±1.0	53.3±2.4	45.4±3.2	44.5±1.2	-1.0 [-4.1, +0.2]	-2.2 [-3.2, +0.6]	-0.4 [-1.8, +0.5]	+0.6 [-0.1, +1.7]

Table 28: **Retrieval pool.** Full recipe; target context: the 357 Pilbara holes formerly released as a fine-tuning subset. Δ vs. the deployed train + archive pool. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
none	10	58.5±5.1	39.2±3.2	33.0±4.2	31.7±2.8	58.6±1.1	40.8±1.8	35.4±3.1	44.0±1.3	**-10.8 [-13.2, -9.1]**	+0.4 [-1.6, +1.8]	**-4.2 [-5.5, -3.3]**	+1.1 [-0.4, +2.3]
train only	3	69.5±5.0	30.8±2.5	29.7±4.1	32.3±3.2	73.6±1.0	37.9±2.1	33.8±3.8	43.4±1.6	+0.3 [-1.3, +0.9]	**-8.0 [-9.2, -4.7]**	**-7.5 [-9.4, -5.9]**	**+1.7 [+0.3, +3.0]**
train + target context	3	69.5±4.9	30.8±2.5	29.7±4.0	31.5±3.0	73.6±1.0	37.9±2.1	33.8±3.9	43.3±1.7	+0.3 [-1.3, +0.9]	**-8.0 [-9.2, -4.7]**	**-7.5 [-9.4, -5.9]**	+0.8 [-0.3, +1.8]
train + archive	10	69.2±5.0	38.8±3.4	37.2±4.6	30.6±2.9	73.7±0.9	52.9±2.1	48.5±3.2	44.3±1.1	–	–	–	–

Table 29: **Target-region holes in the retrieval pool.** Full recipe; n of the 357 Pilbara context holes kept in the pool. Δ vs. n=357 in the same pool. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
train pool, n=0	3	69.5±5.0	30.8±2.5	29.7±4.3	32.3±3.2	73.6±0.9	37.9±2.2	33.8±3.7	43.4±1.6	+0.0 [-0.9, +0.9]	+0.0 [-0.5, +0.4]	+0.0 [-1.1, +1.1]	+0.8 [-0.4, +1.9]
train pool, n=10	3	69.5±5.0	30.8±2.5	29.7±4.3	31.9±3.1	73.6±1.0	37.9±2.2	33.8±4.0	43.1±1.7	+0.0 [-0.8, +1.6]	+0.0 [-0.5, +0.4]	+0.0 [-1.2, +1.2]	+0.4 [-0.6, +1.4]
train pool, n=50	3	69.5±4.9	30.8±2.5	29.7±4.4	30.6±3.4	73.6±1.0	37.9±2.3	33.8±3.9	43.4±1.5	+0.0 [-0.8, +1.3]	+0.0 [-0.5, +0.4]	+0.0 [-1.1, +1.1]	-0.9 [-2.5, +0.6]
train pool, n=100	3	69.5±4.9	30.8±2.5	29.7±4.3	30.8±3.2	73.6±1.0	37.9±2.2	33.8±3.7	43.5±1.6	+0.0 [-0.7, +0.9]	+0.0 [-0.5, +0.4]	+0.0 [-1.2, +1.1]	-0.7 [-1.8, +0.4]
train pool, n=200	3	69.5±5.1	30.8±2.5	29.7±4.3	31.4±2.9	73.6±1.0	37.9±2.2	33.8±3.8	43.6±1.5	+0.0 [-0.7, +0.9]	+0.0 [-0.5, +0.4]	+0.0 [-1.1, +1.2]	-0.1 [-0.9, +0.8]
train pool, n=357	3	69.5±5.0	30.8±2.5	29.7±4.4	31.5±3.1	73.6±1.0	37.9±2.2	33.8±3.8	43.3±1.6	–	–	–	–
train + archive, n=0	3	69.2±5.0	39.1±3.6	37.5±5.0	31.2±3.0	73.7±0.9	53.0±2.3	49.1±4.2	43.7±1.1	+0.0 [-1.1, +0.8]	-0.0 [-0.9, +1.0]	+0.0 [-0.3, +0.3]	+0.1 [-0.4, +1.0]
train + archive, n=10	3	69.2±5.0	39.1±3.7	37.5±4.9	31.1±3.1	73.7±0.9	53.0±2.3	49.1±4.2	44.0±1.2	+0.0 [-0.9, +1.2]	+0.0 [-0.9, +1.0]	+0.0 [-0.3, +0.3]	-0.1 [-0.8, +0.8]
train + archive, n=50	3	69.2±5.0	39.1±3.6	37.5±5.0	30.7±3.1	73.7±0.9	53.0±2.3	49.1±4.3	44.4±1.2	+0.0 [-0.8, +0.7]	+0.0 [-0.9, +0.9]	+0.0 [-0.3, +0.3]	-0.4 [-1.3, +0.6]
train + archive, n=100	3	69.2±5.0	39.1±3.6	37.5±5.0	31.0±3.0	73.7±0.9	53.0±2.4	49.1±4.2	44.5±1.2	+0.0 [-0.8, +1.0]	-0.0 [-0.9, +0.9]	+0.0 [-0.3, +0.3]	-0.1 [-0.8, +0.7]
train + archive, n=200	3	69.2±5.0	39.1±3.7	37.5±5.0	31.2±3.0	73.7±0.9	53.0±2.3	49.1±4.3	44.5±1.2	+0.0 [-0.8, +0.7]	-0.0 [-0.9, +0.9]	+0.0 [-0.3, +0.3]	+0.1 [-0.5, +0.9]
train + archive, n=357	3	69.2±5.0	39.1±3.6	37.5±5.0	31.1±3.0	73.7±0.9	53.0±2.4	49.1±4.1	44.4±1.3	–	–	–	–

Table 30: **Retrieval design choices.** Full recipe, one change at a time. Δ vs. the deployed setting. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
deployed	10	69.2±4.9	38.8±3.5	37.2±4.9	30.6±2.9	73.7±0.9	52.9±2.2	48.5±3.2	44.3±1.1	–	–	–	–
neighbour dropout 0	3	69.7±5.0	39.4±3.8	37.8±5.0	30.9±2.9	73.8±1.0	53.4±2.2	49.4±4.2	44.2±1.3	+0.4 [-0.5, +1.7]	+0.6 [-0.9, +2.4]	+0.6 [-0.7, +1.7]	+0.2 [-0.6, +0.9]
neighbour dropout 0.6	3	68.9±5.3	39.0±3.6	37.3±5.0	31.9±3.0	73.6±1.0	52.6±2.4	49.2±4.2	44.9±1.3	-0.3 [-3.3, +0.3]	+0.2 [-0.9, +1.2]	+0.0 [-0.6, +0.5]	**+1.2 [+0.6, +2.1]**
no elevation alignment	3	70.2±4.9	39.5±3.7	37.6±4.9	30.9±3.1	74.8±1.0	53.6±2.3	49.6±4.1	44.3±1.4	**+0.9 [+0.1, +3.0]**	+0.7 [-0.7, +1.6]	+0.4 [-0.2, +0.9]	+0.2 [-0.6, +1.1]
no 25 m self-exclusion	3	70.3±4.9	39.0±3.6	37.5±5.0	30.4±2.9	74.3±1.0	52.5±2.4	49.5±4.0	44.2±1.2	**+1.0 [+0.2, +3.5]**	+0.2 [-1.1, +1.0]	+0.2 [-0.4, +0.8]	-0.2 [-1.2, +0.8]
learned gate	3	69.1±5.1	39.1±3.6	37.5±5.0	31.1±2.9	73.7±0.9	53.0±2.3	49.1±4.3	44.4±1.3	-0.1 [-0.9, +0.5]	+0.4 [-0.8, +1.2]	+0.3 [-0.2, +0.7]	+0.4 [-0.3, +1.2]
learned gate + base-of-cover	3	69.1±5.1	39.1±3.6	37.5±4.9	31.1±3.0	73.7±0.9	53.0±2.3	49.1±4.3	44.4±1.2	-0.1 [-1.0, +0.7]	+0.4 [-0.8, +1.2]	+0.3 [-0.2, +0.7]	+0.4 [-0.3, +1.2]
bed-filled votes	3	73.6±5.6	38.5±3.7	38.0±5.1	29.4±2.7	81.1±0.9	55.2±2.4	51.8±3.9	44.0±1.1	**+4.4 [+1.6, +6.0]**	-0.3 [-2.3, +0.7]	+0.7 [-0.7, +2.1]	**-1.2 [-2.5, -0.3]**

Table 31: **Token unit: 1 m cell vs. stratum.** The same Transformer backbone and the same recipe, retrieval (R) and corpus pretraining (PT), with a 1 m cell or a whole stratum as the token. AR: plain autoregressive model. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports.

		macro-F1 (%)				top-1 accuracy (%)			
Method	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
Cell level (1 m token, Transformer)									
AR	1	60.3±5.1	26.5±3.1	23.6±3.3	22.5±2.5	60.6±1.2	39.7±2.1	34.8±3.0	38.2±1.4
+ R	3	**70.6**±5.5	28.2±2.8	25.0±3.6	19.9±1.9	**80.6**±0.9	51.4±2.5	46.0±3.5	39.0±1.5
+ PT	3	53.3±5.1	26.7±2.7	19.4±2.8	19.1±2.3	57.9±1.1	37.9±2.0	33.5±2.6	31.0±1.6
+ PT + R	3	64.4±5.6	28.2±3.6	23.5±3.5	18.5±2.4	76.5±1.1	49.9±2.6	44.3±4.2	30.7±2.5
Layer level (stratum token, Transformer)									
AR	10	58.2±5.1	34.4±3.8	26.7±3.9	24.4±2.8	58.2±1.1	40.2±2.0	35.3±3.1	38.1±1.3
+ R	3	69.8±4.9	36.6±3.6	34.8±4.6	24.0±3.7	73.7±1.0	52.0±2.2	**48.4**±4.3	39.3±1.4
+ PT	3	57.5±5.2	38.8±3.3	30.8±3.9	**28.1**±3.1	58.1±1.1	40.9±2.3	36.4±4.2	40.7±1.3
+ PT + R	3	68.2±5.3	**39.6**±3.2	**34.9**±4.2	26.0±3.6	73.4±1.0	**52.5**±2.4	48.3±4.9	**41.8**±1.3

Table 32: **Tokenisation grid (without label smoothing).** Plain GRU and full recipe retrained at each spacing; Δ vs. 1 m in the same arm. ±: half-width of the 95% hole-cluster bootstrap interval (B=1000); trained rows also resample seeds. macro-F1 over classes with ≥20 events; top-1 on OOD-R / OOD-D is macro-averaged over exploration reports. Δ: paired difference in macro-F1 (points) with its 95% interval; bold when the interval excludes zero.

		macro-F1 (%)				top-1 (%)				Δ macro-F1 vs. reference			
Arm	**runs**	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T	Test	OOD-R	OOD-D	Pil-T
0.5 m, plain	3	60.2±5.4	36.0±3.6	27.2±4.1	27.8±2.5	59.5±1.1	41.4±2.4	36.4±4.5	41.1±1.2	+0.7 [-3.0, +1.6]	+1.1 [-0.3, +2.8]	+1.1 [-0.5, +2.4]	+1.3 [-0.4, +2.9]
1 m, plain	10	59.5±4.8	34.9±3.2	26.1±3.8	26.5±2.7	58.6±1.1	40.2±2.0	34.5±3.2	40.2±1.4	–	–	–	–
2 m, plain	3	55.5±4.8	30.1±3.1	23.6±3.5	25.2±2.8	55.4±1.0	37.4±2.2	31.8±4.6	36.7±2.2	**-4.0 [-7.0, -1.0]**	**-4.7 [-9.0, -1.5]**	-2.5 [-5.4, +0.2]	-1.4 [-4.5, +0.3]
0.5 m, full recipe	3	66.4±5.7	39.8±3.7	37.8±5.0	31.6±3.0	70.0±1.0	52.4±2.2	44.8±5.0	45.2±1.4	**-2.8 [-6.7, -1.3]**	+1.0 [-0.6, +2.1]	+0.6 [-1.2, +1.6]	+1.0 [-0.0, +2.0]
1 m, full recipe	10	69.2±4.9	38.8±3.6	37.2±4.9	30.6±3.1	73.7±0.9	52.9±2.2	48.5±3.2	44.3±1.1	–	–	–	–
2 m, full recipe	3	68.2±5.2	41.0±3.4	36.0±4.7	27.7±3.4	72.7±1.0	53.5±2.5	46.6±3.1	40.9±1.2	-1.1 [-6.0, +1.3]	+2.2 [-2.0, +3.4]	-1.2 [-4.7, +2.3]	**-3.0 [-6.6, -0.9]**