Domain-Aware Tabular Data Augmentation Using Large Language Models

Suraj Neelakantan Martin Längkvist Amy Loutfi
Machine Perception and Interaction Lab (MPI), Örebro University
Fakultetsgatan 1, 701 82 Örebro, Sweden
{suraj.neelakantan, martin.langkvist, amy.loutfi}@oru.se

Abstract

Traditional tabular augmentation methods, such as SMOTE and Gaussian sampling, treat features as generic vectors, disregarding the domain-specific constraints often present in scientific tabular data. This work introduces a domain-aware augmentation approach that leverages Large Language Models (LLMs) to encode scientific knowledge through policy generation. The effectiveness of this approach is demonstrated using a case study on geochemical compositions, where data must satisfy closure constraints and exhibit intrinsic correlations that geometric interpolation methods fail to preserve. Evaluated on an imbalanced geochemical rock classification dataset, the LLM-based augmentation achieves 95.74% accuracy and a 0.9544 macro-F1 score, outperforming SMOTE, Gaussian sampling, and no-augmentation baselines while requiring fewer synthetic samples.

1 Introduction

Severe class imbalance challenges machine learning (ML) on real-world tabular datasets across domains: predictive maintenance (rare fault classes), medical diagnosis (uncommon diseases), fraud detection (sparse anomalies), and compositional data analysis (rare classes reflecting natural distributions)(1). Traditional augmentation methods like Synthetic Minority Over-sampling Technique (SMOTE) (3) and Gaussian sampling address imbalance through geometric interpolation or sampling in feature space, treating attributes as generic vectors. These approaches can hinder the performance of ML models on tabular data with domain-specific physical constraints, such as compositional data common in geochemistry, metabolomics, and microbiome analysis.

1.1 Related Work and Limitations

SMOTE (3) generates synthetic samples via linear interpolation between minority class neighbors in feature space. SMOTE and its variants (4; 5) treat features as unconstrained vectors with fixed augmentation ratios (k-nearest neighbors), failing to preserve domain-specific relationships or adapt to natural abundance patterns. Aitchison's centered log-ratio (CLR) transformation (2) maps compositional data from the simplex to Euclidean space, enabling standard statistical methods. Applying CLR preprocessing before training ML models remains geometry-based, while these approaches respect closure via inverse transforms but lack domain awareness, ignoring correlations driven by underlying physical processes.

Deep generative models for tabular data such as Variational Autoencoders (TVAE) (7) and Generative Adversarial Networks (CTGAN, CTAB-GAN) (8; 9) learn complex data distributions for synthetic data generation. Graph-based approaches like the Causal-Graph Lithology Classifier (15) apply spatial relationship modeling for lithology classification but require sequential data that can be unavailable in areas like compositional geochemistry. However, these methods require large training sets (thousands

AI for Tabular Data workshop at EurIPS 2025

of samples), struggle with extreme imbalance, and cannot encode known domain constraints beyond patterns present in limited training data. They also require large synthetic sample counts to improve minority class performance, risking overfitting to unrealistic data. Recent tabular foundation models (10) show promise but remain data-hungry and domain-agnostic, lacking mechanisms to concentrate augmentation where most beneficial while minimizing majority class noise.

Large Language Models (LLMs), when combined with carefully chosen inputs, have shown improvements in LLM performance on a variety of tabular tasks (11). TabLLM (13) and GReaT (14) serialize tabular rows as text and fine-tune language models for generation, achieving strong performance on benchmarks. However, these approaches require task-specific fine-tuning and do not explicitly leverage domain knowledge available in LLM pre-training. Most critically, existing methods are not policy-driven frameworks, where LLMs design augmentation strategies rather than directly generating samples.

This work proposes tabular data augmentation using LLMs to encode domain specific constraints through policy generation. The approach generalizes to standard and compositional tabular settings with domain specific constraints. Rock classification using geochemical tabular data is presented as an illustrative application to demonstrate this approach.The remainder of the paper describes the dataset and methodology (Section 2), presents experimental results comparing four augmentation strategies on two classification models (Section 3), and finally the conclusion.

2 Dataset and Methodology

2.1 Dataset

The dataset comprises 752 training and 188 test samples from the Aleutian Arc geochemical database (16), containing volcanic rock compositions from convergent margin settings. Each sample is characterized by 9 major element oxide compositions (SiO_2, TiO_2, Al_2O_3, FeO_t, MgO, CaO, MnO, Na_2O, K_2O) measured by X-ray fluorescence in weight percent. More deatils on the dataset can be found in Appendix A.

2.2 LLM-Based Augmentation

Policy Generation by prompting a LLM. Claude Sonnet 3.5 was given a comprehensive prompt containing petrological descriptions of the 5 rock types and their positions in the calc-alkaline series, dataset statistics including per-oxide distributions and the strong Al_2O_3-MgO anticorrelation (r=-0.906), domain constraints such as compositional closure, and a task specification requesting a JSON policy for class-adaptive augmentation addressing severe imbalance while respecting geological constraints.

The prompt explicitly encodes domain knowledge by requesting heavy augmentation for rare evolved compositions (rhyolite), minimal augmentation for abundant intermediate compositions (andesite), and preservation of diagnostic element correlations from fractional crystallization. The complete prompt appears in Appendix B.

The LLM generates a JSON policy that defines the augmentation strategy. With class-adaptive ratios (minority: 3.5x, mid: 0.65x, major: 0.12x), the policy uses a logistic-normal family in CLR space to handle compositional geometry and employs Ledoit-Wolf covariance estimation with ridge regularization for numerical stability. Element-specific analytical uncertainties reflect XRF measurement precision. Mean shift augmentation (60% probability) introduces within-class heterogeneity, while sample bounds (per-class max 3,500, global max 18,000) prevent excessive generation. The complete prompt text is shown in Appendix C.

Synthetic Sample Generation. Algorithm 1 implements policy-driven generation of synthetic samples. For each class, real samples are transformed to CLR space, class mean and covariance are estimated via Ledoit-Wolf with ridge regularization, synthetic samples are drawn from the fitted Gaussian with clipping to prevent extreme outliers, and inverse CLR transform with closure normalization returns valid oxide compositions.

Algorithm 1 LLM Policy-Driven Synthetic Data Generation

Require: Training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, policy $\mathcal{P}$ (JSON)
Ensure: Synthetic dataset $\mathcal{D}_{\text{syn}}$
Parse $\mathcal{P}$: ratios $\{r_c\}$, ridge λ, clip threshold τ
for each class c **do**
  $\mathcal{D}_c \leftarrow \{(\mathbf{x}_i, y_i) : y_i = c\}$, $n_c \leftarrow |\mathcal{D}_c|$
  Transform to CLR: $\mathbf{z}_i \leftarrow \log(\mathbf{x}_i) - \frac{1}{d}\sum_j \log(x_{ij})$
  Estimate: $\boldsymbol{\mu}_c \leftarrow \frac{1}{n_c}\sum_i \mathbf{z}_i$, $\boldsymbol{\Sigma}_c \leftarrow \text{LedoitWolf}(\{\mathbf{z}_i\}) + \lambda\mathbf{I}$
  for $j = 1$ to $\lfloor r_c \cdot n_c \rfloor$ **do**
    Sample: $\mathbf{z}_{\text{new}} \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$
    Clip: $\mathbf{z}_{\text{new}} \leftarrow \max(-\tau, \min(\tau, \mathbf{z}_{\text{new}}))$
    Inverse CLR: $\mathbf{x}_{\text{new}} \leftarrow \exp(\mathbf{z}_{\text{new}})/\|\exp(\mathbf{z}_{\text{new}})\|_1 \times 100$
    Add to $\mathcal{D}_{\text{syn}}$
  end for
end for
return $\mathcal{D}_{\text{syn}}$

2.3 Baseline Methods

SMOTE-CLR augments compositional tabular data in CLR space with class-adaptive for fair comparison and generates 863 synthetic samples. **Gaussian-CLR** samples per-class Gaussians in CLR space using Ledoit-Wolf covariance with ridge generates 376 synthetic samples. **NoAug** trains on 752 real samples only with tempered class reweighting to address imbalance. All methods apply CLR transformation before augmentation, inverse transform to compositional space, and validate oxide closure (sum $\approx$ 100 wt%).

2.4 Experimental Protocol

Two classifiers are used in this work: Random Forest (RF) (600 trees, max_features='sqrt', min_samples_leaf=1) and XGBoost (XGB)(500 trees, lr=0.08, depth=6, subsample=0.9, colsample=0.9, L2=1.0). Metrics include macro-F1 (equal class weight, emphasizes minorities) and accuracy. Experiments were repeated over 5 seeds. Implementation uses imbalanced-learn v0.11 (SMOTE), XGBoost v2.0 with GPU acceleration, and Anthropic's Claude LLM.

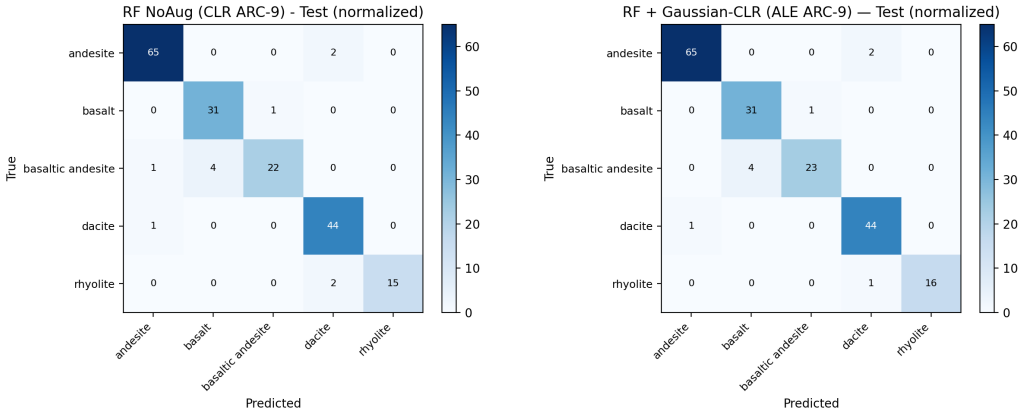
3 Results and Discussion

Table 1: Classification performance on Aleutian Arc test set (5 seeds)

Method	Random Forest		XGBoost	
	Macro-F1	Accuracy	Macro-F1	Accuracy
NoAug	0.9289 ± 0.0066	0.9394 ± 0.0048	0.9377 ± 0.0021	0.9468 ± 0.0013
SMOTE-CLR	0.9474 ± 0.0064	0.9532 ± 0.0045	0.9563 ± 0.0050	0.9596 ± 0.0029
Gauss-CLR	0.9450 ± 0.0026	0.9511 ± 0.0024	0.9671 ± 0.0030	0.9670 ± 0.0024
LLM-CLR	**0.9544 ± 0.0012**	**0.9574 ± 0.0011**	**0.9675 ± 0.0025**	**0.9678 ± 0.0024**

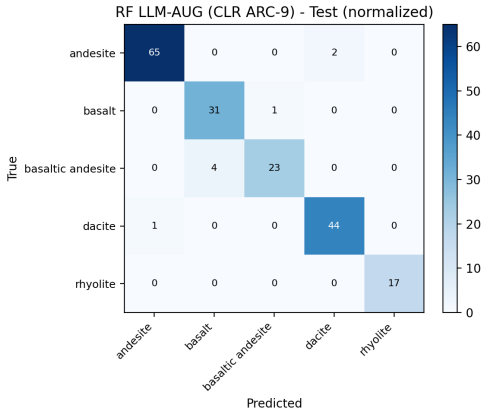
Table 1 presents classification performance across four augmentation strategies and two classification models. LLM-based augmentation achieves the highest macro-F1 scores for both RF (0.9544) and XGB (0.9675) classifiers, outperforming SMOTE-CLR by +0.70 and +1.12 percentage points respectively. Gaussian-CLR shows strong performance with XGBoost (0.9671) but underperforms LLM by +0.94 percentage points (RF) and +0.04 percentage points (XGB) but slightly underperforms SMOTE-CLR on RF (0.9450 vs. 0.9474) but substantially outperforms SMOTE on XGB (+1.08 percentage points, 0.9671 vs. 0.9563). The no-augmentation baseline achieves 0.9289 macro-F1 (RF) and 0.9377 (XGBoost), the lowest of all the methods.

Figure 1 presents confusion matrices for RF classifier across three augmentation approaches. Rhyolite (minority class) shows progressive improvement: NoAug correctly classifies 15 of 17 samples (88%), Gaussian-CLR improves to 16 of 17 (94%), and LLM-CLR achieves a perfect 17 of 17. This pattern



(a) No Aug

(b) Gaussian-CLR



(c) LLM-CLR

Figure 1: RF confusion matrices (row-normalized). Both augmentations improve the minority class; LLM-CLR removes basaltic–andesite/andesite confusion while preserving majority-class accuracy and achieves perfect minority classification.

suggests that while CLR-space sampling helps minority classes, domain-aware policy generation fully resolves errors by encoding petrological constraints that distinguish highly evolved from intermediate compositions.

Basaltic andesite improves from 22 of 27 (81%) with NoAug to 23 of 27 (85%) for both Gaussian and LLM. Both eliminate andesite confusion, but 4 samples are still misclassified as basalt, which is geologically reasonable given class overlap at 52-53 wt% SiO_2 (17) where discrimination requires trace elements absent from this dataset. Identical performance suggests CLR handles compositional geometry effectively for moderate imbalance, while LLM's advantage emerges for extreme minorities (rhyolite). Majority classes maintain 97-98% recall across methods, confirming that augmentation improves underrepresented classes without degrading the well-represented classes. LLM-CLR achieves perfect minority recall using only 181 synthetic samples, compared to 372 for Gaussian-CLR.

4 Conclusion

This work proposes domain-aware augmentation of compositional tabular data using LLMs to encode scientific constraints through policy generation. Evaluated on imbalanced classification of rocks using geochemical data, the approach outperforms SMOTE and Gaussian augmentation methods while using fewer synthetic samples. The framework preserves some domain specific correlations (SiO_2-MgO: r=-0.847 vs. real -0.850) and achieves very high minority class recall (rhyolite F1=1.00). Future work involves validating the framework across diverse compositional

domains beyond geochemistry and exploring automated policy optimization through reinforcement learning on downstream performance.

Acknowledgements

This work was supported by the Industrial Graduate School Collaborative AI and Robotics (Swedish Knowledge Foundation, Dnr:20190128). Computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS).

References

[1] Chen, Wuxing, et al. "A survey on imbalanced learning: latest research, applications and future directions." Artificial Intelligence Review 57.6 (2024): 137.

[2] Aitchison, John. "The statistical analysis of compositional data." Journal of the Royal Statistical Society: Series B (Methodological) 44.2 (1982): 139-160.dataset

[3] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[4] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." International conference on intelligent computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[5] He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). Ieee, 2008.

[6] Lachaud, Alix, Marcus Adam, and Ilija Mišković. "Comparative study of random forest and support vector machine algorithms in mineral prospectivity mapping with limited training data." Minerals 13.8 (2023): 1073.

[7] Tazwar, Syed Mahir, et al. "Tab-VAE: A Novel VAE for Generating Synthetic Tabular Data." ICPRAM. 2024.

[8] Xu, Lei, et al. "Modeling tabular data using conditional gan." Advances in neural information processing systems 32 (2019).

[9] Zhao, Zilong, et al. "Ctab-gan: Effective table data synthesizing." Asian conference on machine learning. PMLR, 2021.

[10] Zhang, Han, et al. "Towards foundation models for learning on tabular data." (2023).

[11] Sui, Yuan, et al. "Table meets llm: Can large language models understand structured table data? a benchmark and empirical study." Proceedings of the 17th ACM International Conference on Web Search and Data Mining. 2024.

[12] Perfit, Mr R., et al. "Chemical characteristics of island-arc basalts: implications for mantle sources." Chemical Geology 30.3 (1980): 227-256.

[13] Hegselmann, Stefan, et al. "Tabllm: Few-shot classification of tabular data with large language models." International conference on artificial intelligence and statistics. PMLR, 2023.

[14] Borisov, Vadim, et al. "Language models are realistic tabular data generators." arXiv preprint arXiv:2210.06280 (2022).

[15] Sun, Youzhuang, et al. "Causal-Graph Lithology Classifier: Synergizing Causal Inference with Graph Neural Networks for High-Accuracy Rock Classification in Well Logging." Marine and Petroleum Geology (2025): 107452.

[16] PetDB Team, T., 2019. EarthChem Data-To-Go: Geochemical Data for the Aleutian Arc, version February 2019, Version 1.0. Interdisciplinary Earth Data Alliance (IEDA). https://doi.org/10.1594/IEDA/111305. Accessed 2025-10-16

[17] M. J. LE BAS, R. W. LE MAITRE, A. STRECKEISEN, B. ZANETTIN, IUGS Subcommission on the Systematics of Igneous Rocks, A Chemical Classification of Volcanic Rocks Based on the Total Alkali-Silica Diagram, Journal of Petrology, Volume 27, Issue 3, June 1986, Pages 745–750

Appendix

A Dataset Statistics

The dataset was cleaned first by filtering the original 1,305 samples to ensure geological validity. Then it was further filtered to retain only samples with complete oxide measurements, then normalizes compositions to 100 wt% closure. Unlabeled entries, xenoliths, and non-igneous lithologies are removed. Rock-type-specific compositional bounds enforce petrological constraints, removing samples that violate known chemical ranges like basalts must have SiO_2 between 44-53 wt% with total alkalies below 5.5 wt%, while rhyolite requires SiO_2 between 70-78 wt% and alkalies below 9.5 wt%. Classes with insufficient support (<40 training or <20 test samples) were dropped. The final dataset contains 5 volcanic rock types reflecting calc-alkaline differentiation: andesite (n=274, 36%), basalt (n=257, 34%), basaltic andesite (n=133, 18%), dacite (n=57, 8%), and rhyolite (n=31, 4%).

This geological compositional data impose three constraints on augmentation methods. First, oxides must satisfy closure, i.e. their sum equal to 100, inducing spurious negative correlations (2). Second, element ratios must respect thermodynamic equilibria (e.g., Mg# = Mg/(Mg+Fe) should be between [30,80] for natural arc magmas (12)). Third, diagnostic correlations like SiO_2-MgO anti-correlation (r = -0.85) arise from fractional crystallization and must be preserved. Traditional geometry-based augmentation often violates these constraints. All methods apply centered log-ratio (CLR) transformation $\text{CLR}(\mathbf{x}) = \log(\mathbf{x}) - \frac{1}{d}\sum_{j=1}^{d}\log(x_j)$ before augmentation to handle compositional geometry.

B LLM Prompt for Policy Generation

The complete prompt provided to Claude Sonnet 3.5 for generating the class-adaptive augmentation policy. The prompt demonstrates how geochemical domain knowledge is encoded through natural language to guide policy design.

B.1 Complete Prompt Text

```
You are an expert in geochemical data augmentation and imbalanced
learning. Generate a JSON policy for creating synthetic rock
composition samples using centered log-ratio (CLR) space.

DATASET SUMMARY:
Dataset: Aleutian Arc volcanic rocks
Total: 752 samples, 5 classes
Features: 9 major element oxides (SiO2, TiO2, Al2O3, FeOt, MgO,
          CaO, MnO, Na2O, K2O)

CLASS DISTRIBUTION:
  andesite          : 274 (36.4%) - MAJOR
  basalt            : 257 (34.2%) - MAJOR
  basaltic andesite : 133 (17.7%) - MID
  dacite            :  57 ( 7.6%) - MINORITY
  rhyolite          :  31 ( 4.1%) - MINORITY

Imbalance ratio: 8.84:1

GEOCHEMICAL CONTEXT:
Calc-alkaline differentiation series (basalt → andesite → dacite
→ rhyolite):
- Progressive SiO2 increase (50% → 72%), MgO decrease (7% → 0.8%)
- Key correlations: SiO2-MgO (-0.850), Al2O3-MgO (-0.906),
```

```
  SiO2-CaO (-0.723)
- Natural pattern: Intermediate compositions dominate, evolved
  rhyolites rare
- Compositional constraint: All oxides must sum to 100 wt%

TASK:
Generate class-adaptive policy to reduce imbalance from 8.84:1 to
~3-4:1 while preserving fractional crystallization trends and
diagnostic correlations.

REQUIREMENTS:
1. **Strategy**: "class_adaptive_legacy"

2. **Bucket Ratios**:
   - minority (rhyolite, dacite): 3.0-4.0× augmentation
   - mid (basaltic andesite): 0.5-0.8× augmentation
   - major (andesite, basalt): 0.1-0.2× augmentation

3. **Quantile Breaks**: [0.15-0.25, 0.85-0.95]

4. **Covariance**:
   - "ledoit_wolf" estimator
   - ridge: 0.0005-0.001

5. **Quality Controls**:
   - clip_z: 7.5-8.5
   - max_resample: 1500-2000
   - per_class_max: 3500-4000
   - per_class_min: 4-8
   - global_max: 18000-20000

6. **Mean Shift**:
   - enable: true
   - frac: 0.55-0.65
   - apply_prob: 0.9-1.0

7. **Instrument Noise** (relative std by oxide):
   - Major (SiO2, Al2O3): 0.025-0.035
   - Intermediate (FeOt, MgO, CaO): 0.10-0.15
   - Minor (TiO2, Na2O, K2O): 0.15-0.25
   - Trace (MnO): 0.20-0.30

8. **Seed**: 42

EXPECTED OUTCOMES:
- Total synthetic: ~300-450 samples
- Final imbalance: ~3-4:1 (from 8.84:1)
- Preserve key correlations (SiO2-MgO within ±0.01)
- Target metrics: >95% accuracy, >0.93 macro-F1

OUTPUT:
Return valid JSON with "logistic_normal_clr" family and
"class_adaptive_legacy" strategy. Include all required parameter
fields.
```

B.2 Domain Knowledge Encoding

The prompt encodes geochemical expertise through three mechanisms:

Statistical Context. Complete class distribution (752 samples across 5 classes with 8.84:1 imbalance), compositional ranges for the calc-alkaline differentiation series, and diagnostic element correlations (SiO_2-MgO: r=-0.850) provide quantitative grounding. These statistics reflect 50+ years of igneous petrology research on arc volcanism.

Petrological Constraints. The prompt describes fractional crystallization trends (progressive SiO_2 increase, MgO depletion) and natural abundance patterns (intermediate compositions dominate, evolved rhyolites rare due to thermodynamic barriers). Compositional closure (oxides sum to 100 wt%) and correlation preservation requirements ensure synthetic samples respect physical laws.

Task-Specific Guidance. Explicit augmentation ratio ranges for each class tier (minority 3-4×, mid 0.5-0.8×, major 0.1-0.2×) and quality control parameters (ridge regularization, CLR clipping thresholds) translate domain knowledge into actionable policy specifications. Element-specific analytical uncertainties reflect X-ray fluorescence measurement precision.

C Generated JSON Policy

Claude Sonnet 3.5 generated the following policy, which was used in the experiments reported in this paper. Note that prompt engineering and parameter tuning can yield alternative policies; the policy below represents one effective configuration among several explored during method development.


```
{
  "family": "logistic_normal_clr",
  "strategy": "class_adaptive_legacy",
  "bucket_ratio": {
    "minority": 3.50,
    "mid": 0.65,
    "major": 0.12
  },
  "quantile_breaks": [0.18, 0.88],
  "ridge": 0.0008,
  "estimator": "ledoit_wolf",
  "clip_z": 8.5,
  "max_resample": 1500,
  "per_class_max": 3500,
  "per_class_min": 4,
  "global_max": 18000,
  "mean_shift": {
    "enable": true,
    "frac": 0.60,
    "apply_prob": 1.0
  },
  "instrument_sigma_rel": {
    "sio2": 0.028,
    "tio2": 0.165,
    "al2o3": 0.032,
    "feot": 0.125,
    "mgo": 0.155,
    "cao": 0.135,
    "mno": 0.245,
    "na2o": 0.145,
    "k2o": 0.220
  },
  "seed": 42
}
```


Policy Interpretation. The generated policy reflects domain-informed decisions: minority classes (rhyolite n=31, dacite n=57) receive 3.50× augmentation, mid-frequency class (basaltic andesite n=133) receives 0.65× augmentation, and majority classes (andesite n=274, basalt n=257) receive minimal 0.12× augmentation. Quantile breaks [0.18, 0.88] partition classes into three tiers based on natural abundance.

Ledoit-Wolf covariance estimation with ridge regularization ($\lambda = 8 \times 10^{-4}$) stabilizes synthetic generation for small minority classes. CLR clipping threshold (8.5) balances diversity and realism, preventing extreme outliers while allowing sufficient within-class variance. Element-specific analytical uncertainties mirror X-ray fluorescence precision: major oxides (SiO_2 2.8%, Al_2O_3 3.2%)

exhibit low relative error, while minor oxides (MnO 24.5%, K_2O 22%) show higher measurement uncertainty typical of low concentrations.

Mean shift augmentation (60% fraction, 100% application probability) adds controlled diversity beyond covariance-based sampling, particularly beneficial for minority classes with limited real samples. This policy generated 181 synthetic samples distributed as: rhyolite (6), dacite (60), basaltic andesite (50), basalt (38), andesite (27), achieving the target class balance while preserving diagnostic geochemical correlations (SiO_2-MgO: r=-0.847 in synthetic vs. r=-0.850 in real data).

The policy shown above represents one successful configuration among several explored during development. We emphasize that the core contribution is not this specific JSON, but rather the *paradigm*: LLMs can translate domain expertise expressed in natural language into executable augmentation policies. This approach achieved superior classification performance (0.9675 macro-F1) using only 181 synthetic samples—50% fewer than Gaussian-CLR—by encoding petrological constraints that purely statistical methods cannot capture.