A General Pipeline for LLM Finetune Ingestion of Scientific Tabular Data

Victor Y. Shirasuna[1], **Enzo Reis de Oliveira**[1], **Caio Rodrigues Gama**[1], **Daniel Djinishian de Briquez**[1], **Eduardo Soares**[1], **Sandro Rama Fiorini**[1], **Dmitry Zubarev**[2], **Nathaniel H. Park**[2], **Rodrigo Neumann Barros Ferreira**[1], **Emilio Vital Brazil**[1]

[1]IBM Research Brazil
[2]IBM Research Almaden
vshirasuna@ibm.com, Enzo.Reis@ibm.com, caiogama@ibm.com, daniel.briquez@ibm.com, Eduardo.Soares@ibm.com, srfiorini@ibm.com, Dmitry.Zubarev@ibm.com, npark@us.ibm.com, rneumann@br.ibm.com, evital@br.ibm.com

Abstract

Large Language Models (LLMs) excel in linguistic reasoning but remain limited in processing structured scientific data such as tables and measurement datasets. We introduce a scalable framework that converts tabular scientific data into validated natural-language question–answer (Q&A) corpora for LLM fine-tuning. The pipeline integrates statistical quantization, automated Q&A generation, linguistic refinement, and *LLM-as-a-judge* evaluation to ensure factual and linguistic quality. Applied to the QM9, QMOF, and PubChem datasets, it produced over 1.3 billion tokens across 12.5 million samples with high fluency and grammatical accuracy. This data-to-text paradigm bridges numerical and linguistic modalities, enabling LLMs to reason over empirical data and advancing the development of scientifically grounded, multimodal language models. All resulting corpora will be open-sourced.

Introduction

The advent of large language models (LLMs) has transformed the landscape of scientific computing and data-driven discovery (Zhang et al. 2025; Guo et al. 2024). LLMs have shown remarkable capabilities in reasoning, summarization, and contextual understanding across a wide range of domains (Matarazzo and Torlone 2025; Huang and Chang 2022). However, most LLMs are primarily trained on unstructured textual data—scientific articles, technical manuals, and web text—whereas the majority of real-world scientific information exists in structured formats such as tables, spreadsheets, or multidimensional measurement datasets (Van Breugel and Van Der Schaar 2024; Hollmann et al. 2025). These tabular datasets capture rich quantitative relationships among physical, chemical, or biological variables, but they remain largely inaccessible to LLMs in their native form (Medupin, Bannister, and Schwartz 2020). Consequently, while current models excel at linguistic reasoning, they often lack the ability to interpret, analyze, or explain patterns that emerge from numerical data, thereby limiting their potential for true scientific understanding (Akhtar et al. 2023; Huang et al. 2024).

Bridging this gap requires new methodologies that can translate structured numerical data into rich textual representations suitable for natural-language processing (Suadaa et al. 2021; Ruan et al. 2024). Yet, this transformation is nontrivial: tabular data encapsulates implicit relationships between features that must be rendered explicitly through linguistic abstraction (Wang et al. 2022). For example, a simple correlation between temperature and reaction yield in a chemistry dataset might correspond to a verbal statement such as “Higher reaction temperatures tend to increase yield up to an optimal point.” Such textualization demands not only statistical summarization but also contextual framing that mirrors how scientists interpret and communicate findings (Saebi et al. 2023; Raghavan et al. 2023). Without this bridge, LLMs remain disconnected from the quantitative backbone of empirical science.

To address this challenge, we introduce a systematic framework for converting tabular scientific datasets into high-quality textual question–answer (Q&A) pairs suitable for LLM fine-tuning (Fig. 1). The process begins with statistical analysis of the numerical and categorical features to identify key patterns—such as correlations, trends, and outliers—within the dataset. These insights inform a Q&A generation module that formulates natural-language questions and corresponding factual answers derived from the data. A subsequent text humanization stage enhances linguistic naturalness and domain relevance, producing coherent, human-readable prompts. To ensure factual and interpretive accuracy, the generated Q&A pairs undergo expert validation and LLM-as-a-judge refinement, where human reviewers and secondary models assess clarity, correctness, and explanatory depth. The resulting collection of validated prompts forms a textualized corpus that encapsulates the original tabular knowledge, enabling efficient fine-tuning of domain-specific LLMs.

This data-to-text paradigm unlocks new possibilities for integrating structured and unstructured scientific information within a unified language model. Fine-tuned models trained on the generated Q&A corpus can reason over both textual and numerical content—explaining correlations, summarizing dataset properties, and generating hypotheses in natural language. More broadly, the proposed framework offers a scalable pathway for scientific knowledge grounding, allowing LLMs to assimilate insights from large experimental databases without manual annotation. By transforming tabular data into interpretable linguistic form, our approach paves the way for data-aware scientific assistants capable of interactive analysis, automated report generation, and context-sensitive discovery across disciplines such as materials science, chemistry, biology, and engineering.

In summary, the main contributions of this work are:

1. A novel pipeline for converting tabular scientific data into validated natural-language Q&A prompts suitable for LLM fine-tuning.
2. An automated statistical-to-textual translation method that integrates quantitative analysis, linguistic transformation, and expert or model-based validation.
3. A framework for grounding LLMs in structured scientific knowledge, improving their ability to reason about empirical data, generate insights, and assist in data interpretation tasks.
4. A demonstration of scalability and adaptability, showing how the approach can generalize across diverse domains and datasets with minimal manual intervention.

Proposed Pipeline

The objective of this work is to design a reproducible and extensible data pipeline that converts tabular scientific datasets into formats compatible with large language models (LLMs), including support for multimodal inputs. The proposed pipeline restructures structured numerical data into question–answer (Q&A) pairs suitable for fine-tuning language models while preserving traceability across all transformation stages. Figure 1 illustrates the complete data flow, exemplified for the QM9 dataset, though the framework is designed to be dataset-agnostic and easily adaptable to other scientific domains.

The pipeline begins by parsing tabular datasets containing numerical and categorical features. For each dataset, individual properties are statistically analyzed to determine their underlying distribution and modality. Continuous numerical variables are discretized into categorical intervals to enhance interpretability and improve alignment with the token space of language models. This quantization step allows the model to operate on linguistically meaningful labels (*very low, low, medium, high, very high*) instead of raw numerical values. Each transformation step—including statistical distribution type, thresholds used for binning, and property metadata—is stored to ensure full traceability of the data generation process.

Q&A Generation and Linguistic Refinement

Following discretization, the framework automatically generates Q&A pairs derived from the feature columns of the dataset. Each Q&A item associates a scientific observation (e.g., a molecular property) with its corresponding value, expressed in natural language. Three complementary Q&A formats are supported:

- Single-property pairs: Each entry corresponds to one property and its associated textual description.
- Property-category pairs: Each entry summarizes a subset of related properties (e.g., electronic or thermodynamic properties) into categorical groups.
- All-properties pairs: Each entry contains questions and answers covering all available properties in a given sample.

Initially, Q&A pairs are generated with fixed syntactic templates to ensure structural consistency. These structured templates are then refined using an instruction-tuned language models, which enhances lexical diversity and improves the fluency of the generated text. During this stage, the model paraphrases and reformulates the answers while maintaining their factual content. The refinement process can be orchestrated by a LLM platform, such as `Ollama`, that programmatically interfaces with the LLM and stores the outputs in a relational database, preserving all intermediate versions of the text.

LLM-as-a-Judge Evaluation

To ensure the linguistic and factual quality of the generated Q&A pairs, a secondary model acts as an evaluator in an **LLM-as-a-judge** configuration. This evaluation step assesses three main dimensions: (i) semantic coherence between question and answer, (ii) factual correctness with respect to the original tabular data, and (iii) linguistic relevance and clarity. The resulting evaluation reports are saved alongside the Q&A corpus, providing quantitative and qualitative diagnostics for each dataset.

Among the three main dimensions, we define four evaluation criteria for the LLM-as-a-judge framework for each generated Q&A sample:

- **C1 – Naturalness:** Judges linguistic fluency and human-likeness, focusing solely on writing flow, tone, and variation. Ignores factual correctness and minor grammatical errors.
- **C2 – Faithfulness:** Checks factual alignment between the generated text and the original tabular record.
- **C3 – Clarity and Scope:** Evaluates whether the answer explicitly covers all six schema properties, stays within scope (no applications, comparisons, or unrelated methods), and maintains readability and cohesion.
- **C4 – Surface Form:** Detects spelling, grammar, punctuation, capitalization, and formatting inconsistencies.

These criteria are applied automatically by the evaluator model and later aggregated into dataset-level pass rates.

Multimodal Integration

Although the framework is primarily designed for tabular data, it also supports multimodal extensions. In the case of the QM9 dataset, each molecule is represented not only by its SMILES string but also by a corresponding 3D electron-density grid. These grids, stored as `.npy` tensors of dimension $128 \times 128 \times 128$, encode spatial electron-density distributions and are referenced within the Q&A structure using unique identifiers. This design allows the resulting dataset to serve both unimodal (text-only) and multimodal (text + 3D grid) fine-tuning pipelines, broadening its applicability to generative and retrieval-based multimodal tasks.

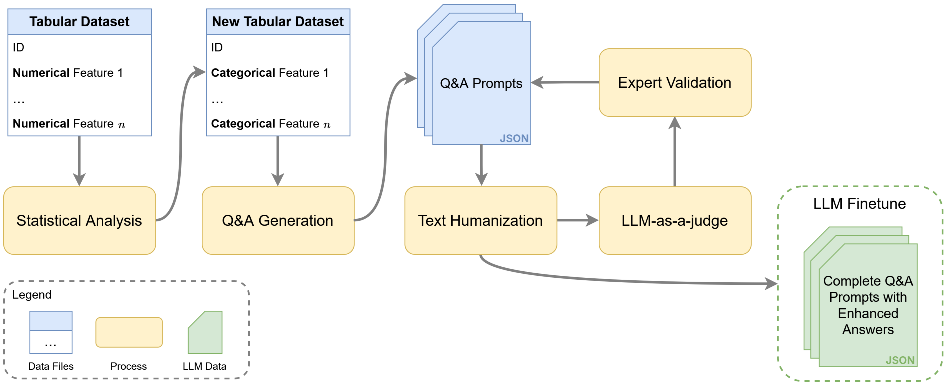


Figure 1: LLM Finetune Dataset Pipeline Generation.

Dataset	#Samples	#Tasks	Description
QM9	133,885	19	Quantum-mechanical properties of small organic molecules.
QMOF	20,373	6	Quantum-chemical and structural descriptors for metal–organic frameworks.
PubChem	55,433	7	Large-scale database of molecular structures and physicochemical properties.

Table 1: Tabular datasets used for pipeline validation. Each dataset was processed through the proposed framework to generate structured and linguistically refined Q&A corpora.

Categorization and Statistical Quantization

To convert continuous molecular properties into interpretable classes, a statistical analysis is performed to determine each property's distributional characteristics (Gaussian, symmetric, or skewed). Outliers are removed to avoid distortion, and percentile-based thresholds are computed according to predefined intervals (0–5–30–70–95–100). Each interval is assigned a categorical label corresponding to qualitative bins such as *very low*, *low*, *medium*, *high*, and *very high*. The quantization process is fully documented in a JSON descriptor , which records thresholds, units, and statistical summaries for each property.

Large-Scale Processing and Parallelization

Given that datasets like QM9 and PubChem contain millions of molecular samples, the pipeline supports distributed execution on high-performance clusters (e.g., AWS). Lightweight instruction-tuned LLMs are employed to maximize throughput while maintaining linguistic quality. Multithreaded inference is implemented using the LiteLLM library, which manages concurrent API requests and fault recovery. The subsequent merging of LLM-enhanced responses with their structured counterparts is conducted using PySpark, ensuring scalability and fault tolerance. Final outputs are stored in both `JSONL` and `Parquet` formats to facilitate downstream processing and model training.

Traceability and Output Artifacts

A key design principle of the pipeline is end-to-end traceability. For each dataset processed, the framework outputs:

1. A mapping table linking original and transformed features.
2. Statistical metadata and quantization thresholds.
3. The prompts used for linguistic refinement.
4. The full Q&A dataset in multiple formats.
5. Evaluation reports from the LLM-as-a-judge step.

These artifacts ensure reproducibility and transparency, allowing other researchers to audit or extend the data transformation pipeline.

In summary, the proposed pipeline generalizes the process of converting numerical and structured data into textual representations suitable for LLM training and evaluation. By integrating statistical quantization, automated Q&A generation, multimodal support, and LLM-based refinement and evaluation, the framework provides a robust and extensible foundation for creating high-quality scientific corpora that bridge the gap between numerical data and natural language understanding.

Datasets

This study leverages several well-established datasets in computational chemistry and materials science to evaluate the proposed pipeline for transforming tabular data into question–answer (Q&A) pairs suitable for large language model (LLM) fine-tuning and multi-modality. Each dataset represents a distinct class of scientific tabular information, ranging from quantum-mechanical molecular properties to high-throughput materials descriptors. Table 1 summarizes their key characteristics.

The QM9 dataset (Ramakrishnan et al. 2014) comprises 133,885 small organic molecules represented by their SMILES strings and associated with 19 computed quantum-mechanical properties, including HOMO/LUMO energies, dipole moment, and heat capacity. Each molecule is also paired with a corresponding 3D electron-density grid, which enables multimodal extensions by combining textual and volumetric representations.

The QMOF dataset (Rosen et al. 2021) contains over 20,000 metal–organic frameworks annotated with computed quantum-chemical descriptors such as band gap, total energy, and formation enthalpy. Its diverse composition and large property space make it an excellent benchmark for evaluating the generalization capacity of tabular-to-text conversion pipelines.

PubChem (Kim et al. 2021) is a large public repository containing millions of molecules with annotated biological and physicochemical properties. In this work, a representative subset of 55,433 samples was used to assess scalability and robustness when applying the proposed Q&A generation and linguistic refinement pipeline to massive datasets.

All datasets were standardized and processed through the same pipeline stages: feature analysis, statistical quantization, Q&A generation, linguistic refinement, and evaluation via an LLM-as-a-judge framework. The resulting corpora enable controlled studies of how textualization quality and lexical diversity vary across scientific domains and dataset sizes.

Results

This section presents a comprehensive evaluation of our LLM-based data generation pipeline. We begin by assessing the quality of the generated text across three distinct datasets: QM9, QMOF, and PubChem using an LLM-as-a-judge framework to measure fluency and factual fidelity. We then detail an analysis of the factual fidelity over the linguistic refinement. Finally, we present a summary about the total tokens and samples generated for this study.

LLM-as-a-Judge Evaluation

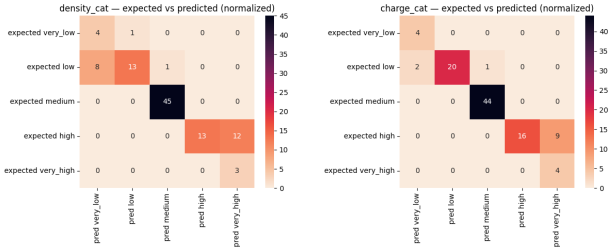
A baseline evaluation using raw, non-humanized templates confirmed the reliability of the LLM-as-a-judge framework. As expected, baseline texts scored 0% in Naturalness (C1) but 100% in Fidelity (C2), Clarity (C3), and Grammar (C4), validating that the evaluator distinguishes fluency from factual accuracy. After refinement, QM9 achieved near-perfect fluency (100% C1/C4, 97% C3) but lower Fidelity (58%). QMOF followed the same trend with strong fluency (100% C1, 90% C4) yet reduced Fidelity (68%) and Clarity (42%), likely due to higher structural complexity. PubChem showed similar behavior (97% C1, 90% C4, 52% C2), reinforcing that increased linguistic variation often lowers factual precision across domains.

Criterion	Baseline	QM9	QMOF	PubChem
C1. Naturalness	0%	100%	100%	97%
C2. Fidelity	100%	58%	68%	52%
C3. Clarity	100%	97%	42%	82%
C4. Grammar	100%	100%	90%	90%

Table 2: LLM-as-a-judge evaluation pass rates (%) based on 100 random samples per dataset.

Factual Correctness Analysis

We observed that most of the C2 failures comes from intensity swaps in `denisty_cat` and `charge_cat` (e.g., `low` ↔ `very_low`, `high` ↔ `very_high`). Less frequent issues include `band_gap_mismatch`, occasional `pld_cat_mismatch`/`lcd_cat_mismatch`, and a few rule/omission cases for band gap. Overall, pushing fluency without extra anchoring increases these label-intensity flips; stricter surface rules help C4 but do not by themselves recover C2/C3.



(a) density_cat — expected vs predicted

(b) charge_cat — expected vs predicted

Figure 2: Confusion matrices for density_cat and charge_cat categories illustrating the main C2 (factual) swap patterns.

Generated Tokens

At the end of the pipeline process, the resulting corpora is ready available for LLM finetune ingestion. For each dataset we determine the number of generated tokens and the resulting amount of training samples, which are shown in Table 3.

Dataset	Tokens	#LLM Samples
QM9	836M	2.5M
QMOF	9.5M	179.2K
PubChem	454M	9.87M
Total	**1.30B**	**12.5M**

Table 3: Token statistics of the generated LLM data for the QM9, QMOF, and PubChem.

Conclusion and Future Work

In this work, we present a general pipeline that helps transform scientific tabular data into meaningful Q&A pairs for LLM ingestion. The proposed pipeline enables reproducibility and traceability of the entire process generation, ensuring standardization of the resulting corpora. Our experiments with three distinct scientific tabular datasets demonstrate the feasibility of the pipeline in providing validated content to LLM finetuning. However, we recognize some limitations that need to be addressed in future work. For example, we aim to improve Q&A generation and linguistic refinement within the evaluated LLM-as-a-judge dimensions, such as the fidelity and clarity criteria. Similarly, the limitation of the current approach for factual fidelity should be addressed to reflect the categorical assertiveness, avoiding information erroneous with the original data. Finally, all generated LLM data will be open-sourced and publicly available for the community.

References

Akhtar, M.; Shankarampeta, A.; Gupta, V.; Patil, A.; Cocarascu, O.; and Simperl, E. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. *arXiv preprint arXiv:2311.02216*.

Guo, S.; Shariatmadari, A. H.; Xiong, G.; and Zhang, A. 2024. Embracing foundation models for advancing scientific discovery. In *2024 IEEE International Conference on Big Data (BigData)*, 1746–1755. IEEE.

Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S. B.; Schirrmeister, R. T.; and Hutter, F. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045): 319–326.

Huang, J.; and Chang, K. C.-C. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Huang, S.; Gu, Y.; Hu, X.; Li, Z.; Li, Q.; and Xu, G. 2024. Reasoning factual knowledge in structured data with large language models. *arXiv preprint arXiv:2408.12188*.

Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. 2021. PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1): D1388–D1395.

Matarazzo, A.; and Torlone, R. 2025. A survey on large language models with some insights on their capabilities and limitations. *arXiv preprint arXiv:2501.04040*.

Medupin, C.; Bannister, C.; and Schwartz, J.-M. 2020. Exploring the interactions of physical, chemical and biological variables of an Urban river using network analysis. *Water*, 12(9): 2578.

Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; and Coley, C. W. 2023. Dataset design for building models of chemical reactivity. *ACS Central Science*, 9(12): 2196–2204.

Ramakrishnan, R.; Dral, P. O.; Rupp, M.; and Von Lilienfeld, O. A. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1): 1–7.

Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; and Snurr, R. Q. 2021. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5): 1578–1597.

Ruan, Y.; Lan, X.; Ma, J.; Dong, Y.; He, K.; and Feng, M. 2024. Language modeling on tabular data: A survey of foundations, techniques and evolution. *arXiv preprint arXiv:2408.10548*.

Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; et al. 2023. On the use of real-world datasets for reaction yield prediction. *Chemical science*, 14(19): 4997–5005.

Suadaa, L. H.; Kamigaito, H.; Funakoshi, K.; Okumura, M.; and Takamura, H. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1451–1465.

Van Breugel, B.; and Van Der Schaar, M. 2024. Why tabular foundation models should be a research priority. *arXiv preprint arXiv:2405.01147*.

Wang, F.; Xu, Z.; Szekely, P.; and Chen, M. 2022. Robust (controlled) table-to-text generation with structure-aware equivariance learning. *arXiv preprint arXiv:2205.03972*.

Zhang, Y.; Khan, S. A.; Mahmud, A.; Yang, H.; Lavin, A.; Levin, M.; Frey, J.; Dunnmon, J.; Evans, J.; Bundy, A.; et al. 2025. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1(1): 14.

Supplementary Materials

Generated Dataset Examples

To illustrate the transformation process from structured data to natural-language text, we present representative examples generated by the proposed pipeline. Each sample demonstrates the evolution from the initial template-based rendering directly derived from tabular input values to the linguistically refined form produced by the LLM-based enhancement stage. This comparison highlights how the model increases textual coherence, scientific fluency, and contextual interpretability while preserving the underlying factual content.

(a) QM9 Dataset. The following pair of examples corresponds to the molecule `CC(O)(CO)C(O)C=O` from the QM9 dataset. The first block shows the *template-based* output, where each molecular property is rendered through deterministic sentence templates. The second block presents the corresponding *LLM-refined* text, which integrates the same properties into coherent scientific prose with added interpretive context.

Template-based example for the molecule `CC(O)(CO)C(O)C=O`:

For this molecule, the predicted properties are as follows:

- Rotational constant A: **medium**, measured in GHz.
- Rotational constant B: **low**, measured in GHz.
- Rotational constant C: **medium**, measured in GHz.
- Dipole moment: **medium**, measured in Debye.
- Isotropic polarizability: **low**, measured in Bohr^3.
- HOMO energy: **low**, measured in Hartree.
- LUMO energy: **low**, measured in Hartree.
- HOMO–LUMO gap: **medium**, measured in Hartree.
- Electronic spatial extent: **high**, measured in Bohr^2.
- Zero-point vibrational energy: **medium**, measured in Hartree.
- Internal energy at 0 K (u_0): **very low**, measured in Hartree.
- Internal energy at 298.15 K (u_{298}): **very low**, measured in Hartree.
- Enthalpy at 298.15 K (h_{298}): **very low**, measured in Hartree.
- Free energy at 298.15 K (g_{298}): **very low**, measured in Hartree.
- Heat capacity C_v: **high**, measured in cal/(mol·K).
- Per-atom internal energy at 0 K: **medium**, measured in Hartree.
- Per-atom internal energy at 298.15 K: **medium**, measured in Hartree.
- Per-atom enthalpy at 298.15 K: **medium**, measured in Hartree.
- Per-atom free energy at 298.15 K: **medium**, measured in Hartree.

The LLM-refined version below expresses the same physical quantities in a cohesive paragraph. It automatically rephrases repetitive template structures into connected statements, introduces methodological details, and contextualizes the categorical labels within the dataset distribution. The discrepancy in the heat capacity classification (high $\rightarrow$ very low) is intentionally highlighted, illustrating how linguistic variation may occasionally introduce factual deviations during refinement.

LLM-refined (linguistically enhanced) example for the molecule `CC(O)(CO)C(O)C=O`:

The molecule under consideration, as per its properties derived from the QM9 dataset using DFT at the B3LYP/6-31G(2df,p) level, exhibits a unique set of characteristics. Its rotational constants A and C are classified as medium, both measured in GHz, while B is low. The dipole moment falls into the medium range, also in Debye units. The isotropic polarizability is low, quantified in Bohr^3. In terms of electronic properties, the HOMO energy is categorized as low, the LUMO energy also low, and the HOMO–LUMO gap, a crucial parameter, is medium, all measured in Hartree. The electronic spatial extent is high, represented in Bohr^2, and the zero-point vibrational energy is medium in Hartree. Notably, the properties u_0, u_{298}, h_{298}, and g_{298} at 0 K and 298.15 K, respectively, along with their per-atom counterparts, and the heat capacity C_v at 298.15 K, are all deemed **very low**. These qualitative labels—'very low', 'low', 'medium', 'high', and 'very high'—are dataset-dependent, reflecting the percentile distributions of the QM9 properties. Hence, the description encapsulates both numerical relationships and linguistic interpretability consistent with quantum-chemical conventions.

This comparison demonstrates how the pipeline bridges structured numerical data and natural scientific language: the template output ensures completeness and factual traceability, whereas the LLM-refined version introduces narrative fluency, contextual grounding, and human-readable style suitable for model fine-tuning and downstream evaluation.

(b) QMOF Dataset. The following pair of examples illustrates the textualization process applied to a representative entry from the QMOF dataset. The first excerpt corresponds to the *template-based* rendering, where the material properties are enumerated directly from the structured table. The second excerpt shows the *LLM-refined* version, which integrates these descriptors into cohesive, scientifically framed prose.

Template-based example for the material $Ba_2CuC_6H_{14}O_{16}$:

For this material, the predicted properties are as follows:

- Crystal system: **triclinic**.
- Pore-limiting diameter (PLD): between **0 Å** and **5 Å**.
- Largest cavity diameter (LCD): between **0 Å** and **5 Å**.
- Framework density: **high**, measured in g/cm^3.
- Electronic classification: **semiconductor**.
- Maximum atomic charge: **high**.

LLM-refined (linguistically enhanced) example for the material $Ba_2CuC_6H_{14}O_{16}$:

The MOF, identified as $Ba_2CuC_6H_{14}O_{16}$, adopts a triclinic crystal system, consistent with its crystallographic descriptor. Its geometrical features indicate extremely narrow pore and cavity diameters, both within the 0–5 Å interval, suggesting potential for selective molecular adsorption or separation. The framework density of $Ba_2CuC_6H_{14}O_{16}$ is categorized as **very high**, placing it in the top percentile of densities within the QMOF dataset. Electronically, the material is a **semiconductor** with a band gap estimated between 0–4 eV, according to the QMOF classification scheme. Its maximum atomic charge is also **very high**, based on the PBE-DDEC atomic site charge percentile range. Together, these attributes describe $Ba_2CuC_6H_{14}O_{16}$ as a dense, semiconducting MOF with potentially high catalytic activity arising from its elevated atomic charge—making it a promising candidate for applications in heterogeneous catalysis or gas storage and separation.

The comparison reveals how the LLM-refined description transforms a discrete property list into an integrated scientific narrative. While the structural and electronic characteristics remain largely consistent, the model introduces two notable deviations: the categorical labels for both framework density and maximum atomic charge are shifted from **high** to **very high**. This linguistic intensification reflects the model's tendency to amplify categorical extremes, possibly influenced by contextual correlations within the training distribution. Additionally, the LLM-generated text extrapolates a *band gap range* (0–4 eV) and infers potential functional implications—such as catalytic or adsorption behavior—that are not explicitly encoded in the original tabular data.

These expansions exemplify how the refinement stage enhances interpretive depth and domain fluency but also highlights the need for factual consistency checks to prevent semantic drift. Overall, the QMOF case underscores the dual nature of LLM-based textualization: it improves readability, contextual richness, and scientific expressiveness, while introducing occasional overgeneralizations that warrant post-hoc validation in high-stakes scientific applications.

(c) PubChem Dataset. The following pair of examples demonstrates the transition from a structured, template-based text to a linguistically refined narrative for a representative molecule in the PubChem dataset. Both describe the same underlying properties, yet differ markedly in structure, interpretive scope, and scientific expressiveness.

Template-based example for the molecule `CC1(COC(=O)C1=O)C`:

For this molecule, the predicted properties are as follows:

- Molecular weight: **low**, measured in g/mol.
- Exact mass: **low**, measured in Da.
- Monoisotopic mass: **low**, measured in Da.
- XLogP3: **low**.
- Topological polar surface area (TPSA): **low**, measured in Å^2.
- Tautomer count: **very low**.
- Molecular complexity: **low**.

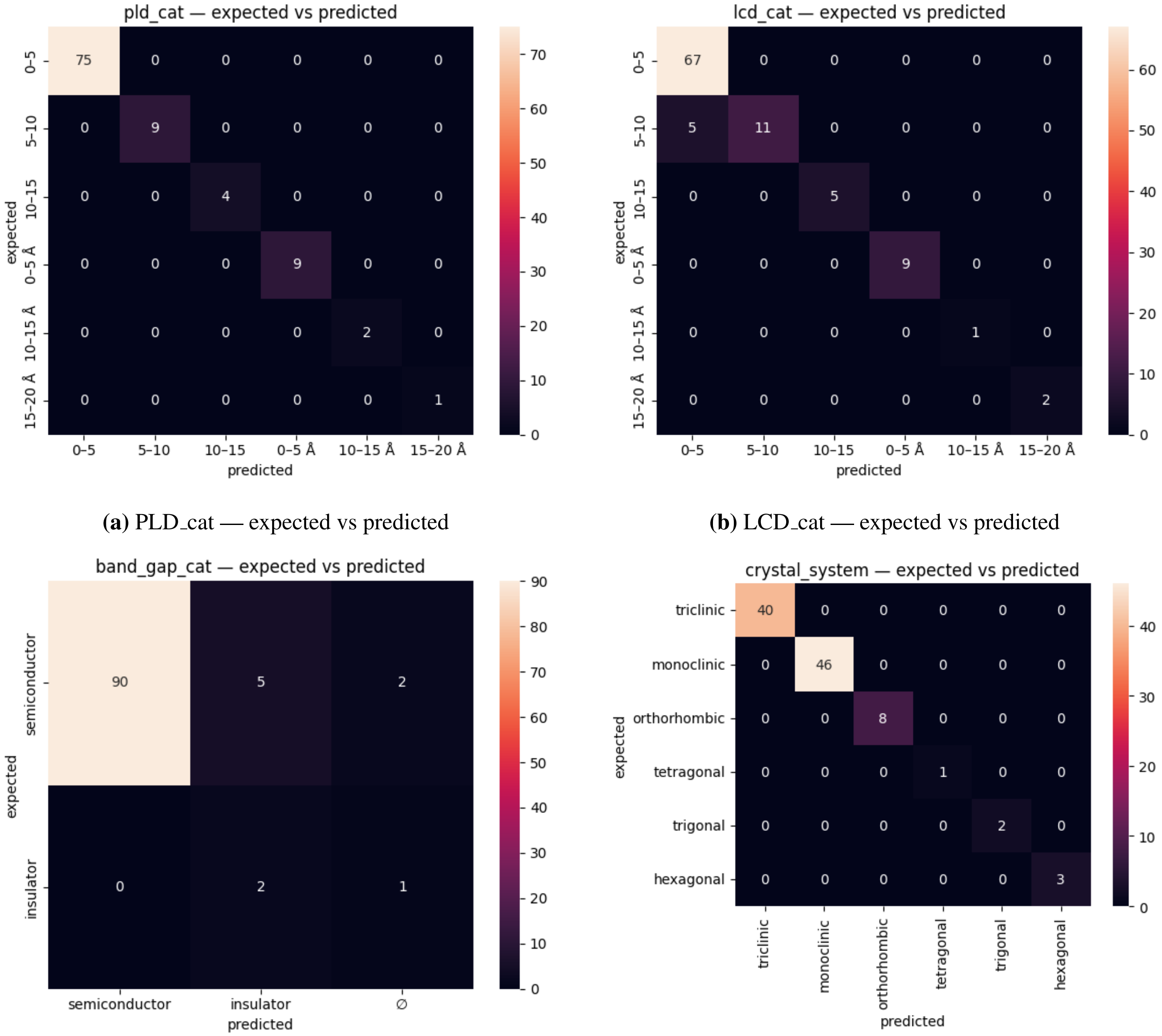
LLM-refined (linguistically enhanced) example for the molecule `CC1(COC(=O)C1=O)C`:

The molecule represented in the grid exhibits a striking simplicity, as evidenced by its **low** complexity score, **monoisotopic mass**, **exact mass**, **molecular weight**, and **XLogP3** value. These attributes fall into the **low** category, based on the dataset-specific quantization strategy employed by PubChem, implying that such values represent the lower end of the distribution for these properties within their extensive chemical dataset. Furthermore, the molecule displays an exceptionally **low** tautomer count, categorizing it under the **'very low'** bracket according to PubChem's standards. Its **topological polar surface area** is similarly low and minimal, as are its hydrogen bond acceptor and donor counts, further underscoring its simplicity. These characteristics suggest a small, hydrogen-dominant molecule with minimal capacity for hydrogen bonding or complex three-dimensional configurations. The low number of heavy atoms, covalent units, stereocenters, and undefined stereocenters corroborates this profile of minimal structural intricacy. However, it is crucial to note that the classification into these categories is dataset-dependent, reflecting the statistical distribution of property values within PubChem's comprehensive chemical information repository.

Both outputs describe identical property categories—low molecular weight, low polarity, and very low tautomer count—yet differ substantially in their communicative structure and scientific intent. The *template-based* text is factual and declarative, designed to ensure deterministic correspondence with the original dataset. The *LLM-refined* output, in contrast, transforms these discrete statements into a coherent chemical narrative: it not only restates the properties but contextualizes them by inferring compositional implications such as hydrogen dominance, limited stereochemical complexity, and weak hydrogen-bonding capability. While all categorical values remain consistent (as indicated in blue), the model introduces inferred descriptors absent from the original data—such as heavy atom count and stereocenter distribution—thereby enriching the interpretation. This shift exemplifies the broader trade-off between factual alignment and linguistic richness: the LLM output preserves scientific correctness while augmenting the interpretive layer that connects structure and function. Ultimately, this transformation advances the model's ability to reason chemically, bridging the gap between tabular molecular descriptors and natural-language scientific discourse.

Additional Confusion Matrices

For completeness, we include the remaining confusion matrices used in the Criterion C2 analysis. Each plot compares the expected and predicted categorical bins for distinct QMOF properties not shown in the main text.



(a) PLD_cat — expected vs predicted

(b) LCD_cat — expected vs predicted

(c) band_gap_cat — expected vs predicted

(d) crystal_system — expected vs predicted

Figure 3: Supplementary confusion matrices for additional QMOF property categories evaluated under Criterion C2 (Fidelity).