# A Context-Aware Approach for Enhancing Data Imputation with Pre-trained Language Models

Anonymous ACL submission

## Abstract

This paper presents a novel approach named **C**ontextually **R**elevant **I**mputation leveraging pre-trained **L**anguage **M**odels (**CRILM**) for handling missing data in tabular datasets, complementing existing numeric-estimation methods. Instead of relying on traditional numerical estimations, CRILM uses pre-trained language models (LMs) to create contextually relevant descriptors for missing values. This method aligns datasets with LMs' strengths, allowing large LMs to generate these descriptors and small LMs to be fine-tuned on the enriched datasets for enhanced downstream task performance. Our evaluations demonstrate CRILM's superior performance and robustness across MCAR, MAR, and challenging MNAR scenarios, with up to a 10% improvement over the best-performing baselines. By mitigating biases, particularly in MNAR settings, CRILM improves downstream task performance and offers a cost-effective solution for resource-constrained environments.

## 1 Introduction

> *'Well! I've often seen a cat without a grin,' thought Alice; 'but a grin without a cat! It's the most curious thing I ever saw in all my life!'*
>
> *Lewis Carroll, Alice's Adventures in Wonderland (1865)*

Missing data in tabular datasets is a ubiquitous problem often arising from real-life data collection processes (Kumar et al., 2017). Handling missing data is crucial for downstream machine learning (ML) tasks, necessitating data imputation to fill in missing entries with plausible values. However, imputation that overlooks the data context can introduce unintended biases, leading to aberrant model behavior (Schelter et al., 2018, 2021; García-Laencina et al., 2010; Stoyanovich et al., 2020; Yang et al., 2020; Abedjan et al., 2018).

Data may be missing because it was never collected or because collected data was lost. These causes are driven by **domain-specific contexts**. For example, in the medical domain, data might not be collected due to various reasons, such as a patient's characteristics not being recorded during a visit, some tests not being performed, intentional omissions by patients, or the difficulty and danger of acquiring certain information (Yoon et al., 2017; Alaa et al., 2018; Yoon et al., 2018b). Data loss can occur through application or transmission errors or due to data integration errors.

Typically, imputation methods estimate missing values based on observed data, such as a patient's blood pressure and heart rate (Yoon et al., 2018c). However, missing data do not always depend on the observed data. Rubin's widely used categorization of missingness mechanisms identifies three cases (Rubin, 1976): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR, the missingness is independent of the data, whereas in MAR, the probability of being missing depends only on observed values. In MNAR, the probability of missingness depends on unobserved values, and imputation in this case can introduce significant biases to the data. *Therefore, to achieve accurate imputation, it is crucial for methods to account for the specific context of the missingness.*

Existing imputation methods use various numeric estimation techniques to capture the data context, preserving joint and marginal distributions of the imputed data. Many methods, including traditional statistical approaches and machine/deep learning methods, aim to learn the joint distribution of the data either implicitly or explicitly (Van Buuren et al., 2006; Yoon et al., 2018a; Gondara and Wang, 2018; Mattei and Frellsen, 2019; Nazabal et al., 2020; Zhao et al., 2023). However, these methods have several limitations: often requiring fully observed training data, being challenging to implement, needing separate models for each feature, and lacking support for column-specific modeling. See Section 3 for more details. Moreover, most approaches, with notable exceptions (Kim and Ying, 2018; Mohan and Pearl, 2019), primarily address MCAR and MAR data, struggling with the more challenging yet prevalent MNAR case (Muzellec et al., 2020).

Parallel to numeric estimation-based imputation, we explore alternative methods for capturing data context to handle missing values. Specifically, we examine whether it is possible to bypass modeling the data distribution entirely. In scenarios where numeric-estimation methods may introduce bias, such as in MNAR settings, or prove inadequate, we develop an approach that avoids direct estimation of missing values. Instead of estimating missing values directly, we investigate whether an artificial intelligence (AI) model can implicitly handle missingness through its prior general knowledge.

To address these challenges, we explore the potential of utilizing general-purpose pre-trained language models (LMs) (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a; OpenAI, 2023) for handling diverse missingness in tabular datasets. These models possess expansive knowledge (Raffel et al., 2020; Roberts et al., 2020), reasoning capabilities (Chowdhery et al., 2022; Wei et al., 2023; Bhatia et al., 2023), and extensive linguistic expertise (Petroni et al., 2019; Mahowald et al., 2024), and have demonstrated exceptional performance across various downstream natural language processing (NLP) tasks (Bubeck et al., 2023; Raffel et al., 2020; Yang et al., 2024a). Our aim is to leverage the advanced capabilities of LMs to enhance the performance of downstream tasks on tabular data with missing values.

To achieve this goal, we approach the downstream task by treating it as an NLP problem and harnessing the capabilities of LMs to handle missing values. We propose a novel method named **C**ontextually **R**elevant **I**mputation leveraging pre-trained **L**anguage **M**odels (**CRILM**), which operates through a *dual-phase process*. Initially, large LMs (LLMs), such as those with more than 10 billion parameters, generate contextually relevant natural language descriptors for missing values. For instance, in the UCI Wine dataset (Aeberhard and Forina, 1991), a contextually relevant descriptor for missing values in the feature malic acid could be: *Malic acid quantity missing for this wine sample*. These descriptors replace missing values, transforming numeric datasets into natural language contextualized formats, thereby aligning the data with the strengths of LMs and augmenting their processing capabilities.

Subsequently, these missingness-aware textual datasets are used for solving downstream tasks such as classification, modeled as NLP tasks. The textual datasets serve as the foundation for fine-tuning smaller pre-trained LMs such as those with less than 10 billion parameters, showcasing a unique and effective use of language models beyond their conventional applications. By incorporating **contextually relevant descriptors for missing data**, CRILM addresses variability and specificity across different domains and navigates the complexities of various missingness mechanisms.

Recently, Transformer-based (Vaswani et al., 2017) methods have been proposed to handle missing values in tabular data, such as masked Transformer for generating synthetic tabular data (Gulati and Roysdon, 2023) and pre-training LMs using enriched tabular data (Yang et al., 2024b). However, these approaches overlook diverse missingness patterns, raising questions about their ability to address the biases introduced by the imputation methods and whether downstream task performance improves as a result. Through the innovative integration of LMs into the data imputation process, CRILM aims to deliver a more nuanced, accurate, and reliable method for handling missing data in a context-aware fashion, essential for improving the quality of downstream NLP tasks.

Our approach offers a **cost-effective solution** by leveraging publicly available LLMs for zero-shot inference and employing smaller LMs for downstream tasks, which can be efficiently fine-tuned in low-resource environments. This feasibility is demonstrated through experiments using accessible resources like ChatGPT-3.5 for inference and smaller LM-based fine-tuning, ensuring efficient implementation.

To evaluate CRILM's effectiveness, we analyze its performance across three missing data mechanisms—MCAR, MAR, and MNAR (Rubin, 1976). CRILM is compared against various existing imputation methods, investigating different phrasing choices for missingness descriptors in LM-based tasks. We also explore the influence of decoder-only and encoder-decoder pre-trained LMs on downstream transfer learning, assessing their impact on task performance. Our empirical studies address two key research questions (RQs):

- **[RQ1]**: To what extent does CRILM effectively perform in imputing missing values across distinct missingness mechanisms (MCAR, MAR, and MNAR), compared to existing methods, in terms of accuracy and robustness on varied datasets?
- **[RQ2]**: How do feature-specific versus generic missingness descriptors impact the performance of LM-based downstream tasks?

The contributions of this work are multifaceted. Firstly, CRILM introduces an innovative imputation approach for missing values in tabular datasets, running parallel to existing numeric-estimation-based methods. By utilizing LMs to generate context-specific descriptors for missing data, CRILM sets a new benchmark in data imputation, departing from traditional numerical methods. Secondly, our empirical evaluation highlights CRILM's superior performance over existing methods across varied datasets and missingness patterns, particularly excelling in MNAR settings where biases introduced by numeric-estimation-based techniques can be significant. Specifically, CRILM demonstrates a substantial performance lead of up to 10% over the best-performing baseline imputation method in the challenging MNAR scenarios. Thirdly, we advance the understanding of the NLP capabilities of pre-trained LMs by demonstrating their potential in handling complex data imputation tasks. Additionally, the cost-effectiveness of our approach, achieved by leveraging smaller LMs for transfer learning, enhances its practicality and accessibility. Lastly, our analysis comparing feature-specific and generic descriptors offers insights into optimizing LM performance for imputation tasks, emphasizing contextual accuracy. These contributions advance data preprocessing techniques and open novel pathways for leveraging LMs in addressing complex data science challenges.

## 2 Method

### 2.1 Problem Formulation

Consider a tabular dataset represented by a matrix $\mathbf{X}$ consisting of a collection of $n$ instances (rows) where each instance $\mathbf{X^i}$ is a $d$-dimensional random variable: $\mathbf{X^i} = (X_1^i, ..., X_d^i)$ (thus $d$ columns). These variables are continuous and/or categorical. The dataset $\mathbf{X}$ has an observed portion denoted by $\mathbf{X_O}$ and a missing portion denoted by $\mathbf{X_M}$. The missingness pattern in $\mathbf{X}$ is denoted by $\mathbf{M}$, which is a matrix of the same dimensions as $\mathbf{X}$ in which cells have a value of 1 if missing and 0 otherwise.

CRILM takes $\mathbf{X}$ and transforms it into a missingness-aware contextualized natural language dataset $\mathbf{X_{missingness\_aware}}$ by replacing the missing values by contextually relevant descriptors. Our goal is to demonstrate the efficacy of CRILM via the performance of a downstream classification task by fine-tuning an LM using $\mathbf{X_{missingness\_aware}}$.

### 2.2 Generating Missing Values

We construct synthetic datasets with up to 30% missing values by applying the following three missingness mechanisms on complete datasets: MCAR, MAR and MNAR. The implementations of these mechanisms are modified from (Jäger et al., 2021).

**MCAR.** It is introduced by randomly removing 30% of the observations from each feature.

**MAR.** First, we select all observations within the 30th percentile range of an independent feature, typically the first column in the dataset. Then, we randomly remove 60% of the values from each corresponding (dependent) feature within this subset, ensuring that missingness is related to the independent feature but random within the dependent features.

**MNAR.** We remove the observations of a feature if the observations fall within the 30th percentile range of the feature value.

### 2.3 Description of CRILM

Figure 1 illustrates the CRILM process, which encompasses four stages: (1) constructing a contextualized natural language dataset, (2) generating suitable descriptors for missing values, (3) creating a missingness-aware contextualized dataset, and (4) adapting an LM for downstream tasks. We detail these stages below.

**Constructing a Contextualized Natural Language Dataset.** We construct a contextualized natural language dataset from a numeric dataset $\mathbf{X}$ containing missing values. The objective is to generate contextually suitable description of each attribute and its measures in natural language. For instance, a record from the UCI Wine dataset (Aeberhard and Forina, 1991) with numeric input and output attributes is contextualized as follows: *"The*

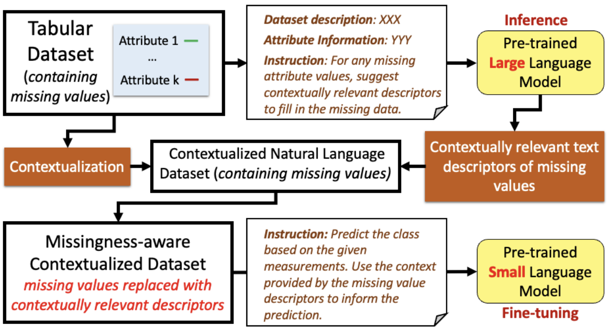

Tabular Dataset (containing missing values)
Attribute 1
...
Attribute k
Dataset description: XXX
Attribute Information: YYY
Instruction: For any missing attribute values, suggest contextually relevant descriptors to fill in the missing data.
Inference
Pre-trained Large Language Model
Contextualization
Contextualized Natural Language Dataset (containing missing values)
Contextually relevant text descriptors of missing values
Missingness-aware Contextualized Dataset
missing values replaced with contextually relevant descriptors
Instruction: Predict the class based on the given measurements. Use the context provided by the missing value descriptors to inform the prediction.
Pre-trained Small Language Model
Fine-tuning

Figure 1: An overview of CRILM.

*alcohol content in the wine is 12.47. The level of malic acid in the wine is 1.52 ... The class of the wine is classified as class 1 wine."*[1] This step converts numeric values into detailed descriptions, preparing the dataset for embedding missing value descriptors.

**Generating Suitable Descriptors for Missing Values.** Unlike conventional imputation methods that estimate missing values from observed data using numerical methods, we utilize contextually relevant descriptors of missing values for imputation. We generate these descriptors by a conversational LLM (e.g., OpenAI's ChatGPT-3.5 (Achiam et al., 2023)). We prompt the LLM with a dataset description and instruct it to generate missing value descriptors, such as: *"For any missing attribute values, suggest contextually relevant descriptors to fill in the missing data."* This method relies on LLM's extensive knowledge base and linguistic capabilities to produce appropriate missing value descriptors. A list of feature-specific contextually relevant missing-value descriptors for selected datasets are provided in Appendix A.4.

**Creating a Missingness-Aware Contextualized Dataset.** We construct the missingness-aware contextualized natural language dataset, denoted as $\mathbf{X}_{\mathbf{missingness\_aware}}$, by replacing the missing values with generated descriptors. This process ensures that each data instance is "aware" of its missing attributes, thereby enhancing the downstream LM's ability to learn from incomplete data by providing explicit context. Additionally, we use distinct descriptors for different features in the dataset that contain missing values. This approach implicitly informs the downstream LM to handle the missingness of each feature in a contextually appropriate manner, ultimately improving the performance of the downstream task.

**Adapting an LM for Solving Downstream Tasks.** The final step involves fine-tuning a pre-trained small LM with the missingness-aware, contextually-rich dataset. During the fine-tuning process, we incorporate specific task instructions and strategies for handling missing data. For instance, in classification tasks, we include instructions such as: *"Predict the class based on the given measurements. Use the context provided by the missing value descriptors to inform the prediction."* This approach ensures that an LM effectively utilizes the contextual information embedded in the descriptors, thereby enhancing its predictive performance despite the presence of missing data. Using smaller LMs for fine-tuning not only makes the process cost-effective but also allows for efficient adaptation to the specific characteristics of the dataset and task at hand.

## 3 Related Work

The challenge of missing data in tabular datasets has led to the development of numerous imputation methods, broadly categorized into those modeling feature distribution and those that do not. The latter category includes methods such as distribution matching and traditional non-parametric methods. In the former category, two distinct types of imputation methods exist: those treating features separately and those treating them jointly. Separate feature treatment methods, like Multivariate Imputation by Chained Equations (MICE) (Van Buuren et al., 2006; van Buuren and Groothuis-Oudshoorn, 2011), which is an iterative method as well as a discriminative method, specify a univariate model for each feature based on others, with other notable iterative methods also existing (Heckerman et al., 2000; Raghunathan et al., 2001; Gelman, 2004; Liu et al., 2014; Zhu and Raghunathan, 2015). Joint treatment methods aim to learn a joint distribution of all features, with recent developments including deep learning-based generative methods like GAIN (Yoon et al., 2018a), utilizing Generative Adversarial Nets (Goodfellow et al., 2014), although their effectiveness varies compared to traditional methods (Jäger et al., 2021). Other types of generative models that are based on Denoising Autoencoders (Vincent et al., 2008), have been proposed (Gondara and Wang, 2018; Rezende et al., 2014; Mattei and Frellsen, 2018; Nazabal et al., 2020; Ivanov et al., 2019; Richardson et al., 2020a; Mattei

[1] The Python script used for contextualization is provided in the Supplementary Material.

and Frellsen, 2019), though most of these models either rely on fully-observed training data or are suitable only for the MCAR data. Another recent approach, Distribution Matching (DM) (Muzellec et al., 2020), bypasses direct modeling of data distributions. A notable DM method is Transformed Distribution Matching (Zhao et al., 2023), which is suitable for real-world data with complex geometry. Non-parametric methods like k-nearest neighbors (k-NN) imputation (Troyanskaya et al., 2001), which is a discriminative method, and MissForest (Stekhoven and Bühlmann, 2012), which is an iterative and discriminative method, have shown effectiveness compared to sophisticated methods (Emmanuel et al., 2021; Jäger et al., 2021), particularly in the MAR setting (Jarrett et al., 2022). Additionally, simple imputation approaches like mean substitution (Hawthorne and Elliott, 2005) provide basic alternatives. More details are provided in Appendix A.1.

## 4 Experiments

We systematically assess CRILM's efficacy in addressing the research questions outlined in Section 1 through a series of experiments. Utilizing two types of LMs—decoder-only and encoder-decoder—we evaluate the performance of LMs fine-tuned with missingness-aware contextual datasets in downstream classification tasks post-imputation. Specifically, we investigate three types of missingness mechanisms: MCAR, MAR, and MNAR. For comparison with the baseline methods, we first impute the numeric datasets using existing methods (described further below). Then, the datasets are transformed into contextualized natural language datasets using the method described in Section 2.3, which are used for fine-tuning the LMs.

**Datasets.** We evaluate CRILM's performance using six real-life multivariate classification datasets from the UCI repository (Dua and Graff, 2017), which are selected based on their prior usage in existing numeric imputation-based studies (Muzellec et al., 2020; Yoon et al., 2018a; Camino et al., 2019; Gondara and Wang, 2018; Lu et al., 2020; Hallaji et al., 2021; Nazabal et al., 2018; Zhao et al., 2023). This selection ensures a fair comparison with previous research efforts. Dataset statistics are provided in Appendix A.3.

**Baseline Imputation Methods.** We compare CRILM against a diverse set of imputation approaches by focusing on the following six baseline methods: (1) Mean substitution (Hawthorne and Elliott, 2005) (simple imputation method), (2) k-NN (Troyanskaya et al., 2001; Batista and Monard, 2002) (non-parametric and discriminative method), (3) MissForest (Stekhoven and Bühlmann, 2012) (non-parametric, discriminative, and iterative method), (4) MICE (Van Buuren et al., 2006; van Buuren and Groothuis-Oudshoorn, 2011) (discriminative and distribution modeling iterative approach that treats each feature separately), (5) GAIN (Yoon et al., 2018a) (generative and distribution modeling iterative approach that treats features jointly), and (6) Transformed Distribution Matching (TDM) (Zhao et al., 2023) (distribution matching method).

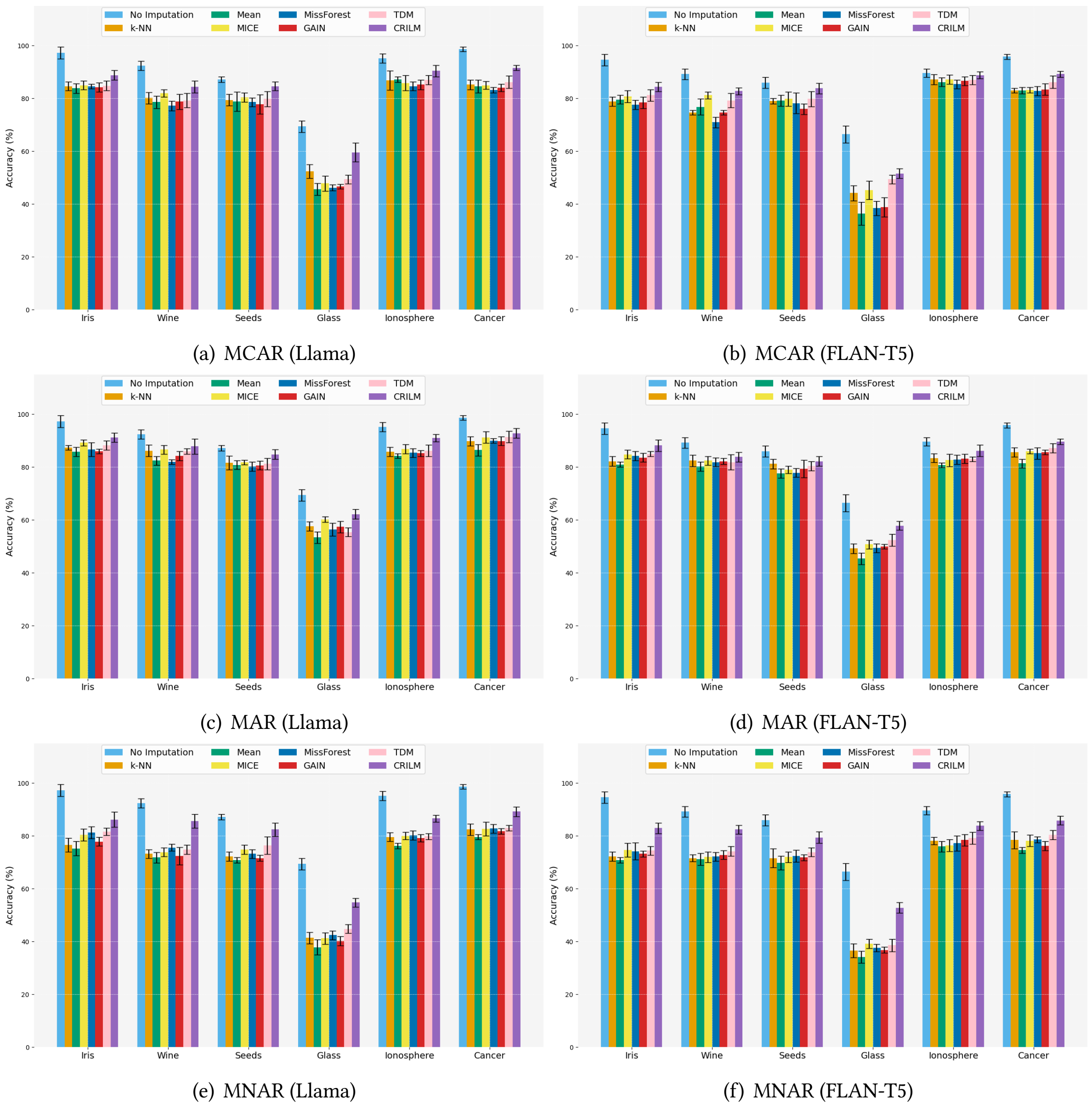
**LMs for Downstream Tasks.** We utilize two types of smaller pre-trained LMs for transfer learning: decoder-only Llama 2 (Touvron et al., 2023b) and encoder-decoder FLAN-T5 (Chung et al., 2022) with 7 billion (7B) and 770 million (770M) parameters, respectively.

**Experimental Settings.** The hyperparameter settings for the various imputation methods and the LMs used in our experiments are detailed below.

*Hyperparameters for Baseline Imputation Methods.* For GAIN, we adhere to the hyperparameters specified in the original publication, setting $\alpha$ to 100, the batch size to 128, the hint rate at 0.9, and the number of iterations to 1000 for optimal performance. MissForest and MICE are configured with their respective default parameters as provided in their PyPI implementations[2], i.e., MissForest: maxiter = 10, ntree = 100, and MICE: m = 5 for the number of multiple imputations. The PyPI MICE implementation utilizes random forests for efficiency. For k-NN, we determine the optimal values for $k$ for each dataset through hyperparameter tuning based on the downstream classification task. For a list of optimal $k$ values, refer to the Appendix A.5. Regarding TDM, we use the original implementation with the reported settings (Zhao et al., 2023).

*Pre-trained LMs for Transfer Lerning.* The Llama model is fine-tuned with the parameter-efficient QLoRA method (Dettmers et al., 2023). The settings are $r = 16$, $\alpha = 64$, $dropout = 0.1$ with the task type set to "CAUSAL_LM". The learning rate is 2e-4, using the "paged_adamw_32bit" optimizer.

[2] https://pypi.org/



(a) MCAR (Llama)

(b) MCAR (FLAN-T5)

(c) MAR (Llama)

(d) MAR (FLAN-T5)

(e) MNAR (Llama)

(f) MNAR (FLAN-T5)

Figure 2: **[RQ1]**: Comparison of CRILM and baseline imputation methods across MCAR, MAR, and MNAR missingness patterns using Llama and FLAN-T5 models. Evaluation involves post-imputation LM-based downstream task performance, with CRILM fine-tuned on missingness-aware contextual datasets and baseline methods on contextual datasets. “No Imputation” cases show LM performance on complete datasets without missing values.

The FLAN-T5 model (Chung et al., 2022) is fine-tuned using an AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate set to 3e-4.

Experiments are conducted with a batch size of 4 across 50 epochs, considering memory constraints during fine-tuning. Two Tesla A40 GPUs are used for distributed training, ensuring efficient processing, with each experiment completing in less than twenty minutes, except for the Breast Cancer dataset with more than 500 instances, which takes about an hour. An estimated training time on a single GPU would require between 45 minutes to 2 hours to complete all experiments. For evaluation, 20% of instances are randomly sampled from each dataset. Models are evaluated five times, and both the average performance and standard deviation are reported for comprehensive analysis.

## 4.1 Results

Figure 2 displays experimental outcomes using two types of downstream LMs across six datasets, benchmarking CRILM against existing imputation methods. Performance metrics for LMs fine-tuned on complete datasets (without missing values, thus no imputation needed) are included for comparison. This approach highlights CRILM’s effectiveness by providing a reference baseline, offering a clear view of its advantages over traditional imputation methods.

**[RQ1]**: *To what extent does CRILM effectively perform in imputing missing values across distinct missingness mechanisms (MCAR, MAR, and MNAR), compared to existing methods, in terms of accuracy and robustness on varied datasets?*

**MCAR**: CRILM demonstrates superior accuracy in imputing missing values across all datasets compared to baseline imputation methods. Both the Llama and FLAN-T5 performed well, with Llama showing a slight advantage (1 to 8% higher accuracy). CRILM's performance under the MCAR assumption, where missingness is independent of any data, suggests that it efficiently leverages contextual information for imputation. This efficacy is particularly evident in its ability to significantly close the gap toward the performance of fully complete datasets, showcasing its effectiveness.

**MAR**: CRILM's adaptability is further highlighted under MAR, where missingness depends on observed data. It outperforms other methods by a considerable margin, indicating its proficiency in utilizing available data points to predict missing values accurately. The Llama consistently exhibits superior performance, similar to the MCAR case (2 to 5% higher accuracy).

**MNAR**: The MNAR scenario, characterized by missingness that depends on unobserved data, poses the most significant challenge. Here, CRILM's performance remains notably superior to traditional imputation methods. This robustness in the face of the most difficult missingness mechanism illustrates CRILM's potential to effectively mitigate biases introduced by MNAR missingness, utilizing the LMs' capacity to infer missing information from complex patterns. Similar to the previous cases, Llama exhibits better performance (2 to 4% higher accuracy)

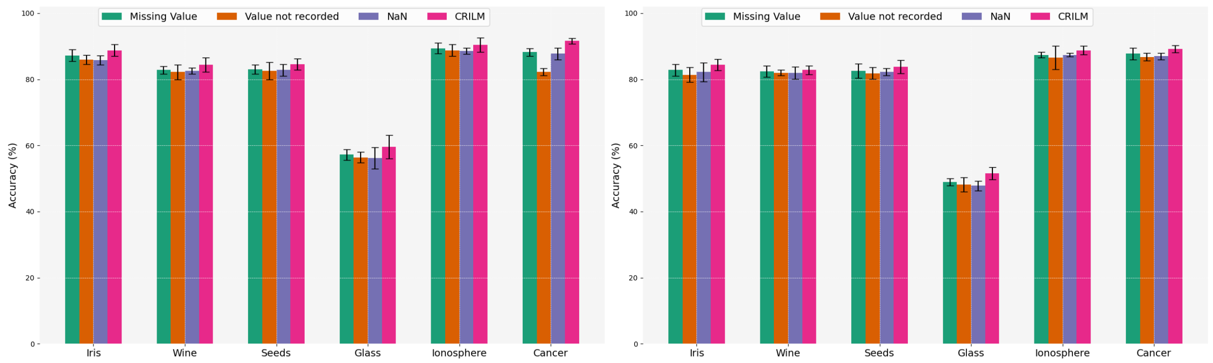
To further demonstrate CRILM's superior performance over traditional baseline imputation methods, particularly in the **MNAR setting**, we assess its efficacy on three challenging datasets: Glass Identification, Seeds, and Wine. These datasets are selected due to the observed lower performance of LMs when utilizing fully complete versions (refer to Figure 2), highlighting their complexity and serving as a rigorous evaluation benchmark for CRILM. According to the results (see Table 1), CRILM consistently outperforms the best baseline methods. The performance gains are 10.0%, 6.0%, and 10.0% for Glass Identification, Seeds, and Wine, respectively, using Llama, and 13.6%, 5.6%, and 8.2% using FLAN-T5. This significant improvement underscores CRILM's effectiveness in addressing the intricacies of MNAR missingness, confirming its position as a robust tool for managing various missing data scenarios. Additional analysis details on MCAR and MAR are provided in Appendix A.2.

Table 1: Comparison of CRILM with leading imputation methods on MNAR missingness across three datasets.

| LM | Data | Best Baseline | CRILM | Gain |
|---|---|---|---|---|
| Llama | Glass | 44.80% (TDM) | 54.80% | +10.0% |
| | Seeds | 76.40% (TDM) | 82.40% | +6.0% |
| | Wine | 75.60% (MissForest) | 85.60% | +10.0% |
| FLAN-T5 | Glass | 39.20% (MICE) | 52.80% | +13.6% |
| | Seeds | 73.80% (TDM) | 79.40% | +5.6% |
| | Wine | 74.20% (TDM) | 82.40% | +8.2% |

**Discussion on RQ1.** CRILM's consistent superiority across diverse missingness patterns and datasets confirms its effectiveness, **addressing RQ1**. This underscores the advantages of integrating contextualized natural language models into imputation, particularly in challenging MNAR scenarios where traditional numeric-estimation methods may introduce biases. The robust performance of CRILM across MCAR, MAR, and MNAR missingness mechanisms highlights its broad applicability, distinguishing it from conventional methods. This generalizability can be attributed to CRILM's missingness-aware data contextualization approach, which effectively taps into the prior knowledge of the pre-trained LMs to implicitly handle missing cases in the data. Notably, Llama (7B) performs slightly better than FLAN-T5 (770M), likely due to its larger model size, which enhances its ability to capture and utilize complex patterns in the data. Furthermore, minimal performance variation across iterations underscores CRILM's stability and reliability, crucial for real-world applications. Its ability to maintain a consistently low error margin highlights its potential as a reliable solution for data imputation.

**[RQ2]**: *How do feature-specific versus generic missingness descriptors impact the performance of LM-based downstream tasks?* Initially, we utilize contextually relevant, feature-specific descriptors for missing values, leading to unique phrases for different features within a dataset. To address RQ2, we aim to determine whether using a uniform,



(a) MCAR (Llama)

(b) MCAR (FLAN-T5)

Figure 3: **[RQ2]**: Impact of feature-specific vs. generic ("NaN", "Missing value", and "Value not recorded") missingness descriptors on LM Performance in MCAR scenario.

yet contextually relevant, descriptor for all features would offer comparable benefits. To this end, we experiment with three consistent descriptors: "NaN", "Missing value", and "Value not recorded". These experiments, focusing on the MCAR scenario, sought to ascertain whether it is more beneficial to use contextually nuanced descriptors or whether a generic descriptor is adequate to harness LMs' general knowledge for managing missing values in datasets.

The experimental findings (Figure 3) illuminate the influence of missing data phrasing on the effectiveness of LMs in addressing such situations. The results reveal a distinct pattern across both types of LMs: generic descriptors, such as "NaN", consistently perform worse than context-specific descriptors designed for each feature and dataset. Among the three fixed descriptors tested, there are some variations in performance. Both "NaN" and "Missing value" outperformed "Value not recorded", with "Missing value" achieving the best results in most cases among the static descriptors.

**Discussion on RQ2.** The findings on RQ2 highlight the importance of context in LMs' handling of missing data. The superior performance of feature-specific descriptors shows that LMs better manage missing data when it is described in a way that accurately reflects the context of the missing information. For example, a descriptor like *"Malic acid quantity missing for this wine sample"* allows an LM to interpret and address the missing data point more effectively than a generic descriptor like *"The level of malic acid in the wine is NaN"*. This preference for context-specific descriptors stems from LMs' extensive linguistic capability. When missing data aligns with the specific context of a feature, an LM can better utilize its knowledge to handle the missing values. However, effectiveness drops when generic labels are used, as they provide minimal contextual information for the LM to draw upon.

**Cost-Effective Implementation of CRILM.** Our method provides an economically viable solution by utilizing publicly available LLMs for zero-shot inference and smaller LMs for downstream tasks, allowing for efficient fine-tuning even in resource-constrained settings. This feasibility is demonstrated through experiments employing accessible resources like ChatGPT-3.5 for inference and single GPU fine-tuning, ensuring experiments are completed within an hour on average, thereby highlighting its cost-effectiveness.

## 5 Conclusion

CRILM demonstrates robust handling of missing data across MCAR, MAR, and notably MNAR mechanisms, consistently outperforming traditional methods. Our experiments highlight CRILM's remarkable effectiveness in MNAR scenarios, achieving up to a 10% performance margin over baseline methods, underscoring its efficacy in the most challenging missingness setting. By leveraging contextualized LMs, CRILM offers a novel imputation method alongside numeric-estimation approaches, particularly beneficial in mitigating biases and enhancing reliability in MNAR case. Its cost-effective implementation, using publicly available LLMs for inference and smaller LMs for downstream tasks, enhances practicality in resource-constrained settings.

Future work will explore extending CRILM to diverse data types such as time-series, images, and unstructured text.

## 6 Limitations

Despite the notable advancements presented by CRILM in addressing missing data within tabular datasets, this work has several limitations. Firstly, CRILM's efficacy depends heavily on the quality and diversity of the training data used to develop the underlying LLMs. In scenarios where LLMs lack exposure to data similar to the specific domain or context of missing information, their ability to generate accurate imputations may be compromised. Additionally, the approach assumes that the descriptive context provided for missing values sufficiently informs the LLM, which may not always be the case. Furthermore, processing large datasets with CRILM, even though we utilize smaller LMs for fine-tuning with contextualized missingness-aware data, may pose scalability challenges, as the fine-tuning process could increase in duration. Moreover, while CRILM performs well across various missingness mechanisms, its application in highly specialized domains where expert knowledge heavily influences data interpretation requires further exploration. Lastly, it is important to note that our evaluation focused on classifying downstream tasks, leaving its efficacy in other task types for future investigation.

## References

Z. Abedjan, L. Golab, F. Naumann, and T. Papenbrock. 2018. *Data Profiling*, volume 10 of *Synthesis Lectures on Data Management*. Morgan & Claypool.

Josh Achiam, Marcin Andrychowicz, Alex Beattie, Jacob Clark, Nitish Drozdov, Adrien Ecoffet, Dario Edwards, Jim Giddings, Ilya Goldberg, Manuel Gomez, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stefan Aeberhard and M. Forina. 1991. Wine. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC7J.

Ahmed M. Alaa, Jinsung Yoon, Scott Hu, and Mihaela van der Schaar. 2018. Personalized risk scoring for critical care prognosis using mixtures of gaussian processes. *IEEE Transactions on Biomedical Engineering*, 65(1):207–218.

G. E. Batista and M. C. Monard. 2002. A study of k-nearest neighbour as an imputation method. In *Frontiers in Artificial Intelligence and Applications*, volume 87, pages 251–260. HIS.

Kush Bhatia, Avanika Narayan, Christopher De Sa, and Christopher Ré. 2023. TART: A plug-and-play Transformer module for task-agnostic reasoning. *arXiv preprint*. ArXiv:2306.07536 [cs].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. 2015. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

R. D. Camino, C. A. Hammerschmidt, and R. State. 2019. Improving missing data imputation with deep generative models. *arXiv preprint arXiv:1902.10666*, pages 1–8.

Kewei Chen, Xiao Liang, Ziqi Zhang, and Zihao Ma. 2022. Gedi: A graph-based end-to-end data imputation framework. *arXiv preprint arXiv:2208.06573*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint*. ArXiv:2204.02311 [cs].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex

Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint*. ArXiv:2210.11416 [cs].

Z. Dai, Z. Bu, and Q. Long. 2021. Multiple imputation via generative adversarial network for high-dimensional blockwise missing value problems. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 791–798.

Z. Dai, Z. Bu, and Q. Long. 2022. Multiple imputation with neural network gaussian process for high-dimensional incomplete data. In *ACML*.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Dheeru Dua and Casey Graff. 2017. UCI machine learning repository.

T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. 2021. A survey on missing data in machine learning. *J Big Data*, 8(1):140. Epub 2021 Oct 27. PMID: 34722113; PMCID: PMC8549433.

Fang Fang and Shu Bao. 2022. Fragmgan: Generative adversarial nets for fragmentary data imputation and prediction. *arXiv preprint arXiv:2203.04692*.

Zehui Gao, Yuan Niu, Jian Cheng, Jie Tang, Tianxing Xu, Pan Zhao, Lei Li, Frank Tsung, and Jianqiang Li. 2023. Handling missing data via max-entropy regularized graph autoencoder. In *AAAI*.

Pedro J. García-Laencina, José-Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. 2010. Pattern classification with missing data: a review. *Neural Comput. Appl.*, 19(2):263–282.

Andrew Gelman. 2004. Parameterization and bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545.

L. Gondara and K. Wang. 2018. Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 260–272. Springer.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, et al. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, Montréal, Canada. Curran Associates, Inc.

Manbir S Gulati and Paul F Roysdon. 2023. Tabmt: Generating tabular data with masked transformers. *Preprint*, arXiv:2312.06089.

E. Hallaji, R. Razavi-Far, and M. Saif. 2021. Dlin: Deep ladder imputation network. *IEEE Transactions on Cybernetics*, 52(9):8629–8641.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. 2015. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402.

Graeme Hawthorne and Peter Elliott. 2005. Imputing cross-sectional missing data: comparison of common techniques. *Australian and New Zealand Journal of Psychiatry*, 39(7):583–590.

David Heckerman, David M. Chickering, Chris Meek, Robert Rounthwaite, and Carl Kadie. 2000. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75.

Bo Huang, Yijun Zhu, Muhammad Usman, Xiang Zhou, and Heng Chen. 2022. Graph neural networks for missing value classification in a task-driven metric space. *TKDE*.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. 2019. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*.

Daniel Jarrett, Bogdan Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. 2022. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9916–9937. PMLR.

S. Jäger, A. Allhorn, and F. Biessmann. 2021. A benchmark for data imputation methods. *Front Big Data*, 4:693674. PMID: 34308343; PMCID: PMC8297389.

Jae-Kwang Kim and Zhiliang Ying. 2018. Data missing not at random: Jae-kwang kim, zhiliang ying editors for this special issue. *Statistica Sinica*.

A. Kumar, M. Boehm, and J. Yang. 2017. Data management in machine learning. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1717–1722, Chicago, Illinois, USA. Association for Computing Machinery. F1277.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. 2019. Misgan: Learning from incomplete data with generative adversarial networks. *arXiv preprint arXiv:1902.09599*.

Jingchen Liu, Andrew Gelman, Jennifer Hill, Yu-Sung Su, and Jonathan Kropko. 2014. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

H.-m. Lu, G. Perrone, and J. Unpingco. 2020. Multiple imputation with denoising autoencoder using metamorphic truth and imputation feedback. *arXiv preprint arXiv:2002.08338*.

Qi Ma and Swarnendu Kumar Ghosh. 2021. Emflow: Data imputation in latent space via em and deep flow models. *arXiv preprint arXiv:2106.04804*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

David A. Marker, David R. Judkins, and Marianne Wingless. 2002. Large-scale imputation for complex surveys. In Robert M. Groves et al., editors, *Survey Nonresponse*. John Wiley & Sons, New York.

Pierre-Alexandre Mattei and Jes Frellsen. 2018. Leveraging the exact likelihood of deep latent variable models. *arXiv preprint arXiv:1802.04826*.

Pierre-Alexandre Mattei and Jes Frellsen. 2019. Miwae: Deep generative modelling and imputation of incomplete data sets. In *Proceedings of the International Conference on Machine Learning*, pages 4413–4423. PMLR.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322.

K. Mohan and J. Pearl. 2019. Graphical models for processing missing data. *Journal of American Statistical Association (JASA)*.

Pedro Morales-Alvarez, Weiwei Gong, Alex Lamb, Steven Woodhead, Simon P. Jones, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. 2022. Simultaneous missing value imputation and structure learning with groups. In *NeurIPS*.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. 2020. Missing data imputation using optimal transport. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera. 2018. Handling incomplete heterogeneous data using vaes. *arXiv preprint arXiv:1807.03653*.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–96.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, pages 1278–1286. PMLR.

Trevor W Richardson, Wencheng Wu, Lei Lin, Beilei Xu, and Edgar A Bernal. 2020a. Mcflow: Monte carlo flow models for data imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14205–14214.

Trevor W. Richardson, Wencheng Wu, Lei Lin, Beilei Xu, and Edgar A. Bernal. 2020b. Mcflow: Monte carlo flow models for data imputation. In *CVPR*, pages 14205–14214.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

D. B. Rubin. 1976. Inference and missing data. *Biometrika*, 63:581–592.

S. Schelter, F. Biessmann, T. Januschowski, D. Salinas, S. Seufert, and G. Szarvas. 2018. On challenges in machine learning model management. *IEEE Data Eng. Bull.*, 41(4):5–15.

Sebastian Schelter, Tammo Rukat, and Felix Biessmann. 2021. JENGA - A framework to study the impact of data errors on the predictions of machine learning models. In *Proceedings of the 24th International Conference on Extending Database Technology, EDBT 2021, Nicosia, Cyprus, March 23 - 26, 2021*, pages 529–534. OpenProceedings.org.

D. J. Stekhoven and P. Bühlmann. 2012. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

J. Stoyanovich, B. Howe, and H. V. Jagadish. 2020. Responsible data management. *Proceedings of the VLDB Endowment*, 13:3474–3488.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. 2001. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525.

S. van Buuren and K. Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Stef Van Buuren, Jean-Paul Brand, Karin Groothuis-Oudshoorn, and Donald B. Rubin. 2006. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint*. ArXiv:1706.03762 [cs].

Roger Vinas, Xiaoxia Zheng, and James Hayes. 2021. A graph-based imputation method for sparse medical records. *arXiv preprint arXiv:2111.09084*.

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on machine learning*, pages 1096–1103.

Shuhui Wang, Jincheng Li, Haoran Miao, Jie Zhang, Jie Zhu, and Jin Wang. 2022. Generative-free urban flow imputation. In *CIKM*, pages 2028–2037.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024a. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Ke Yang, Biao Huang, Julia Stoyanovich, and Sebastian Schelter. 2020. Fairness-aware instrumentation of preprocessing pipelines for machine learning. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA'20)*. ACM.

Yazheng Yang, Yuqi Wang, Sankalok Sen, Lei Li, and Qi Liu. 2024b. Unleashing the potential of large language models for predictive tabular tasks in data science. *Preprint*, arXiv:2403.20208.

J. Yoon, J. Jordon, and M. van der Schaar. 2018a. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR.

Jinsung Yoon, Christina Davtyan, and Mihaela van der Schaar. 2017. Discovery and clinical decision support for personalized healthcare. *IEEE Journal of Biomedical and Health Informatics*, 21(4):1133–1145.

Jinsung Yoon, William R. Zame, Arjun Banerjee, Martin Cadeiras, Ahmed M. Alaa, and Mihaela van der Schaar. 2018b. Personalized survival predictions via trees of predictors: An application to cardiac transplantation. *PloS One*, 13(3):e0194985.

Jinsung Yoon, William R. Zame, and Mihaela van der Schaar. 2018c. Deep sensing: Active sensing using multi-directional recurrent neural networks. In *International Conference on Learning Representations*.

Seongwook Yoon and Sanghoon Sull. 2020. Gamin: Generative adversarial multiple imputation network for highly missing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8464.

Jiaxuan You, Xinyi Ma, Yujie Ding, Mykel J. Kochenderfer, and Jure Leskovec. 2020. Handling missing data with graph representation learning. In *NeurIPS*, volume 33, pages 19075–19087.

He Zhao, Ke Sun, Amir Dezfouli, and Edwin V Bonilla. 2023. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*, pages 42159–42186. PMLR.

Ji Zhu and Trivellore E. Raghunathan. 2015. Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511):1112–1124.

# A Appendix

In this section, we begin with a comprehensive discussion of the related work. Following this, we conduct a comparative analysis of CRILM's effectiveness on a selected set of challenging downstream tasks. Next, we provide a summary of the datasets, along with a list of feature-specific contextually relevant missing-value descriptors for three selected datasets. Lastly, we present the optimal values of $k$ obtained through hyperparameter tuning for k-NN imputation across three missingness patterns—MCAR, MAR, and MNAR—using the Llama and FLAN-T5 models on the six datasets.

## A.1 Related Work

The challenge posed by missing data in tabular datasets has led to the development of numerous imputation methods, broadly classified into two categories: those modeling feature distribution and those that do not model distributions. The latter category includes methods such as distribution matching and traditional non-parametric methods. In the former category, where the focus is on modeling feature distribution, methods aim to model the distribution of missing values while maximizing the observed likelihood (Muzellec et al., 2020). Within this line of approach, two distinct types of imputation methods exist (Zhao et al., 2023): methods that treat features separately and those that treat them jointly.

For methods treating features separately, an **iterative approach** is employed, specifying a univariate model for each feature based on all others. A prominent example is the Multivariate Imputation by Chained Equations (MICE) method (Van Buuren et al., 2006; van Buuren and Groothuis-Oudshoorn, 2011), which adopts a *discriminative approach* for imputation. MICE sequentially imputes missing values for each variable based on the others, cycling through the variables iteratively until predictions stabilize. MICE is particularly effective for handling MCAR and MAR data (Jarrett et al., 2022). Other notable iterative methods include (Heckerman et al., 2000; Raghunathan et al., 2001; Gelman, 2004; Liu et al., 2014; Zhu and Raghunathan, 2015). Given the potential variation in the conditional distribution of each feature, these methods necessitate the specification of separate models for each feature. This approach may prove ineffective, especially in cases where the nature of the missing values remains uncertain.

On the other hand, methods that treat features collectively aim to learn a joint distribution of all features, either explicitly or implicitly. A classical approach for explicit joint modeling assumes a Gaussian distribution for the data, with parameters estimated using EM algorithms (Dempster et al., 1977). Recent developments have seen the utilization of deep learning-based *generative methods* such as Denoising Autoencoders (DAE) (Vincent et al., 2008) and Generative Adversarial Nets (GAN) (Goodfellow et al., 2014). Generative methods can be categorized into implicit and explicit modeling. Implicit models include imputers trained as generators in GAN-based frameworks (Yoon et al., 2018a; Li et al., 2019; Yoon and Sull, 2020; Dai et al., 2021; Fang and Bao, 2022). However, these models produce imputations that are only valid for the MCAR data (Yoon et al., 2018a; Li et al., 2019; Yoon and Sull, 2020). A notable GAN-based method is GAIN (Yoon et al., 2018a), specifically designed for imputing missing data without the need for complete datasets. In GAIN, the generator outputs the imputations, while the discriminator classifies the imputations on an element-wise basis. However, GAIN can be quite difficult to implement in practice (Muzellec et al., 2020). Moreover, it often falls short compared to more traditional machine learning methods such as the non-parametric k-nearest neighbors (k-NN) in terms of performance (Jäger et al., 2021). Explicit generative models refer to deep latent-variable models trained to approximate joint densities using variational bounds. Most of these models either rely on fully-observed training data (Gondara and Wang, 2018; Rezende et al., 2014; Mattei and Frellsen, 2018) or are suitable only for the MCAR data (Nazabal et al., 2020; Ivanov et al., 2019; Richardson et al., 2020a). MIWAE (Mattei and Frellsen, 2019) is an exception in this category that adapts the importance-weighted autoencoders (Burda et al., 2015) objective to approximate maximum likelihood in MAR settings. However, its accuracy depends on the assumption of infinite computational resources. Additionally, with the exception of methods that use separate decoders for each feature (Nazabal et al., 2020), generative methods generally do not support column-specific modeling. Other approaches in the category of methods that learn a joint distribution include those based on matrix completion (Mazumder et al., 2010; Hastie et al., 2015), graph neural networks (You et al., 2020; Vinas et al., 2021; Chen et al., 2022; Huang et al., 2022; Morales-Alvarez et al., 2022; Gao et al., 2023), normalizing flows (Richardson et al., 2020b; Ma and Ghosh, 2021; Wang et al., 2022), and Gaussian processes (Dai et al., 2022).

Distribution Matching (DM) methods represent a recent alternative approach that bypasses the need for modeling data distributions directly (Muzellec et al., 2020; Zhao et al., 2023). The core idea behind DM is that any two batches of data (including those with missing values) originate from the same underlying data distribution. Therefore, an effective method should impute the missing values to ensure that the empirical distributions of the two batches are closely matched. In (Muzellec et al., 2020), the authors achieve DM by minimizing the optimal transport (OT) distance, with the cost function being the quadratic distance in the data space between samples. Another notable method, suitable for real-world data with complex geometry, is Transformed Distribution Matching (TDM) (Zhao et al., 2023). TDM performs OT-based imputation in a transformed space, where the distances between transformed samples better reflect their underlying similarities and dissimilarities, respecting the data's inherent geometry.

Non-parametric methods like k-NN imputation (Troyanskaya et al., 2001; Batista and Monard, 2002) and random forest imputation, such as MissForest (Stekhoven and Bühlmann, 2012), have demonstrated effectiveness in comparison to other sophisticated imputation methods (Emmanuel et al., 2021; Jäger et al., 2021). The k-NN method employs a discriminative algorithm that utilizes the similarity between instances, typically measured by Euclidean distance, to impute missing values, offering flexibility in handling both continuous and categorical data. Conversely, MissForest is an *iterative method* harnessing the power of random forests, excelling in datasets with complex interactions and non-linear relationships, often surpassing other methods in terms of accuracy and robustness. MissForest is particularly adept in the MAR setting (Jarrett et al., 2022).

Finally, simple imputation approaches like mean substitution (Hawthorne and Elliott, 2005) and hot deck imputation (Marker et al., 2002) provide basic alternatives.

Table 2: Performance Comparison of CRILM with leading imputation methods using Llama across three challenging datasets. Best performing baseline methods are in **bold**.

| Dataset | Best Baseline | CRILM | Gain |
|---|---|---|---|
| **Glass Identification** | | | |
| MCAR | 52.40% **(k-NN)** | 59.60% | 7.2% |
| MAR | 60.20% **(MICE)** | 62.20% | 2.0% |
| MNAR | 44.80% **(TDM)** | 54.80% | 10.0% |
| **Seeds** | | | |
| MCAR | 80.40% **(MICE)** | 84.60% | 4.2% |
| MAR | 81.80% **(MICE)** | 84.80% | 3.0% |
| MNAR | 76.40% **(TDM)** | 82.40% | 6.0% |
| **Wine Quality** | | | |
| MCAR | 82.00% **(MICE)** | 84.40% | 2.4% |
| MAR | 86.60% **(MICE)** | 87.80% | 1.2% |
| MNAR | 75.60% **(MissForest)** | 85.60% | 10.0% |

## A.2 Comparative Analysis of CRILM's Effectiveness

To demonstrate the superior performance of CRILM over traditional baseline imputation methods, we investigate its performance on **three particularly challenging datasets**: Glass Identification, Seeds, and Wine. These datasets were chosen due to the comparatively lower performance exhibited by the LMs when using fully complete versions of the datasets (i.e., no missing values), underscoring their complexity and providing a rigorous testing ground for evaluating CRILM's effectiveness.

### A.2.1 Llama

Table 2 presents a detailed comparative analysis based on Llama. In the MCAR setting, CRILM demonstrates substantial superiority over the best baseline method (k-NN, achieving 52.40% accuracy) with a performance gain of 7.2%. This underscores CRILM's robustness in effectively handling missing data within complex datasets. The challenge intensifies with the Seeds dataset, where CRILM surpasses the top-performing baseline method (MICE) by 4.2% under the MCAR setting. Similar trends are observed in the Wine dataset, where CRILM outperforms the best baseline performance under MCAR by 2.4%.

Under MAR conditions, the performance gaps between CRILM and the best-performing baseline methods are relatively modest—2%, 3%, and 1.2% for Glass Identification, Seeds, and Wine, respectively. This suggests that while the predictability of missingness from observed data in MAR scenarios provides some advantage to traditional imputation methods, CRILM still maintains a performance edge.

The MNAR scenario, characterized by the most complex pattern of missingness, highlights CRILM's distinct advantage. Across all three datasets, CRILM not only outperforms the best baseline methods but does so with remarkable performance gains of 10.0%, 6.0%, and 10% for Glass Identification, Seeds, and Wine, respectively. This substantial improvement underscores CRILM's effectiveness in navigating the intricacies of MNAR missingness, further establishing its status as a robust tool for handling various missing data scenarios.

### A.2.2 FLAN-T5

Table 3 provides a FLAN-T5-based comparative analysis of CRILM against leading imputation methods across the three challenging datasets, echoing similar trends observed with the Llama model. In the Glass Identification dataset, FLAN-T5 exhibits significant improvements with CRILM. Under the MCAR setting, CRILM surpasses the best baseline method (TDM, achieving 45.60% accuracy) by 6.0%, highlighting its robust capability to handle missing data effectively, particularly where traditional methods struggle. The Seeds dataset presents a competitive landscape, where CRILM outperforms the top-performing baseline (MICE) by 4.0% under MCAR conditions. Similarly, in the Wine Quality dataset under MCAR conditions, CRILM achieves a 1.2% performance gain over MICE, reinforcing its reliability.

In the MAR scenario for Glass Identification, CRILM shows a pronounced advantage over the best baseline method (TDM, achieving 52.40%), with a notable gain of 5.4%. This underscores CRILM's efficacy in scenarios where missingness can be predicted from observed data, showcasing its versatility across different missing data patterns. However, in the challenging Seeds dataset, the performance gap narrows, with CRILM outperforming k-NN by 1.0%, indicating its continued edge despite the predictability leveraged by traditional methods. The Wine Quality dataset reflects a similar trend, where CRILM achieves a 1.4% performance gain over k-NN.

In the MNAR condition, known for its complexity, CRILM demonstrates a significant advantage. In the Glass Identification dataset, CRILM outperforms MICE by an impressive 13.6%. This substantial improvement is mirrored in the Seeds and Wine Quality datasets, where CRILM achieves gains of 5.6% and 8.2% over TDM, respectively. These results underscore CRILM's exceptional capability in handling the intricate challenges posed by MNAR missingness, firmly establishing it as a powerful tool for addressing diverse imputation challenges.

Table 3: Performance Comparison of CRILM with leading imputation methods using FLAN-T5 across three challenging datasets. Best performing baseline methods are in **bold**.

| Dataset | Best Baseline | CRILM | Gain |
|---|---|---|---|
| | **Glass Identification** | | |
| MCAR | 45.60% **(TDM)** | 51.60% | 6.0% |
| MAR | 52.40% **(TDM)** | 57.80% | 5.4% |
| MNAR | 39.20% **(MICE)** | 52.80% | 13.6% |
| | **Seeds** | | |
| MCAR | 79.80% **(MICE)** | 83.80% | 4.0% |
| MAR | 81.20% **(k-NN)** | 82.20% | 1.0% |
| MNAR | 73.80% **(TDM)** | 79.40% | 5.6% |
| | **Wine Quality** | | |
| MCAR | 81.20% **(MICE)** | 82.40% | 1.2% |
| MAR | 82.40% **(k-NN)** | 83.80% | 1.4% |
| MNAR | 74.20% **(TDM)** | 82.40% | 8.2% |

## A.3 Dataset Summary

Table 4 provides a summary of the six UCI datasets.

## A.4 Missing-value Descriptors

Table 5 reports the list of feature-specific contextually relevant missing-value descriptors for three selected datasets.

## A.5 Optimal $k$ Values for k-NN Imputation in Various Missingness Patterns

Table 6 shows the optimal values of $k$ for k-NN imputation across three missingness patterns (MCAR, MAR, and MNAR) using the Llama and FLAN-T5 models on six datasets. These optimal values were determined through hyperparameter tuning, where k was varied between 3 and 9, based on the downstream classification task to achieve the best imputation performance for each dataset and missingness pattern combination. This tuning process ensures that the k-NN imputation method is tailored to the specific characteristics and requirements of each dataset, enhancing overall performance.

Table 4: Description of the datasets. N=size of the dataset and d=number of features.

| Dataset | N | d | Description |
|---|---|---|---|
| Iris | 150 | 4 | The dataset contains 3 classes of 50 instances each, referring to types of iris plants. |
| Wine | 178 | 13 | Results of a chemical analysis of wines grown in Italy, with three types represented. |
| Seeds | 210 | 7 | Properties of three varieties of wheat: Kama, Rosa, and Canadian. |
| Glass Identification | 214 | 9 | Classification of types of glass for criminological investigation. |
| Ionosphere | 351 | 34 | Phased array of 16 high-frequency antennas, targeting free electrons in the ionosphere. |
| Breast Cancer Wisconsin | 569 | 30 | Binary classification from digitized images of a fine needle aspirate of breast masses. |

Table 5: Feature-specific contextually relevant descriptors for three selected datasets.

| Dataset | Features containing Missing values | Descriptors of missing values |
|---|---|---|
| Iris | 1. Sepal Length<br>2. Sepal Width<br>3. Petal Length<br>4. Petal Width | 1. Sepal Length: Unavailable<br>2. Sepal Width: Unavailable<br>3. Petal Length: Unavailable<br>4. Petal Width: Unavailable |
| Wine | 1. Alcohol<br>2. Malic acid<br>3.Ash<br>4. Alcalinity of ash<br>5. Magnesium<br>6. Total phenols<br>7. Flavanoids<br>8. Nonflavanoi phenols<br>9. Proanthocyanins<br>10. Color Intensity<br>11. Hue<br>12.OD280/OD315 of diluted wines<br>13. Proline | 1. Alcohol content not provided for this wine sample.<br>2. Malic acid quantity missing for this wine sample.<br>3. Ash content data not recorded for this wine sample.<br>4. Alcalinity of ash information unavailable for this wine sample.<br>5. Magnesium level not specified for this wine sample.<br>6. Total phenols data missing for this wine sample.<br>7. Flavanoids content not available for this wine sample.<br>8. Nonflavanoid phenols quantity not provided for this wine sample.<br>9. Proanthocyanins data missing for this wine sample.<br>10. Color intensity information not recorded for this wine sample.<br>11. Hue value not specified for this wine sample.<br>12. OD280/OD315 data missing for this wine sample.<br>13. Proline content not available for this wine sample |
| Seeds | 1. Area<br>2. Perimeter<br>3. Compactness<br>4. Length of kernel<br>5. Width of kernel<br>6. Asymmetry coefficient<br>7. Length of kernel groove | 1. Kernel area not provided.<br>2. Kernel perimeter information missing.<br>3. Kernel compactness data not recorded.<br>4. Length of kernel data missing.<br>5. Width of kernel data missing.<br>6. Asymmetry coefficient information missing.<br>7. Length of kernel groove information missing. |

Table 6: Optimal $k$ values for k-NN imputation across MCAR, MAR, and MNAR missingness patterns using Llama and FLAN-T5 models on six datasets.

| Dataset | Missing pattern | Model | $k$ | Accuracy (%) |
|---|---|---|---|---|
| **Iris** | MCAR | Llama | 5 | 84.60 |
| | | FLAN-T5 | 5 | 78.80 |
| | MAR | Llama | 5 | 87.40 |
| | | FLAN-T5 | 3 | 82.20 |
| | MNAR | Llama | 7 | 76.60 |
| | | FLAN-T5 | 5 | 72.20 |
| **Wine** | MCAR | Llama | 3 | 80.20 |
| | | FLAN-T5 | 3 | 74.60 |
| | MAR | Llama | 5 | 86.20 |
| | | FLAN-T5 | 5 | 82.40 |
| | MNAR | Llama | 5 | 73.20 |
| | | FLAN-T5 | 3 | 71.60 |
| **Seeds** | MCAR | Llama | 3 | 79.40 |
| | | FLAN-T5 | 3 | 79.00 |
| | MAR | Llama | 3 | 81.60 |
| | | FLAN-T5 | 5 | 81.20 |
| | MNAR | Llama | 3 | 72.20 |
| | | FLAN-T5 | 5 | 71.60 |
| **Glass** | MCAR | Llama | 5 | 52.40 |
| | | FLAN-T5 | 5 | 44.20 |
| | MAR | Llama | 5 | 57.60 |
| | | FLAN-T5 | 5 | 49.20 |
| | MNAR | Llama | 3 | 41.40 |
| | | FLAN-T5 | 5 | 36.60 |
| **Ionosphere** | MCAR | Llama | 5 | 86.80 |
| | | FLAN-T5 | 5 | 87.20 |
| | MAR | Llama | 5 | 85.80 |
| | | FLAN-T5 | 5 | 83.40 |
| | MNAR | Llama | 3 | 79.60 |
| | | FLAN-T5 | 5 | 78.20 |
| **Cancer** | MCAR | Llama | 5 | 85.20 |
| | | FLAN-T5 | 3 | 83.00 |
| | MAR | Llama | 5 | 89.80 |
| | | FLAN-T5 | 5 | 85.60 |
| | MNAR | Llama | 5 | 82.40 |
| | | FLAN-T5 | 5 | 78.40 |