# Ranking-Guided Autoregressive Modeling for Multimodal Tabular Anomaly Detection

Antonios Georgakopoulos[1,*], Paul Groth[1] and Lise Stork[1]

[1]University of Amsterdam, Amsterdam, The Netherlands

**Abstract**

Tabular data on the Web is frequently heterogeneous, multimodal, and published in a decentralised way, making it prone to structural and semantic inconsistencies. As such datasets increasingly serve as sources for knowledge graph construction, ensuring structural and semantic consistency prior to integration becomes critical. One approach to dealing with these issues is anomaly detection, a technique that identifies irregular patterns in a given dataset. Although several approaches exist for anomaly detection on tables, methods that capture interactions between different modalities, rather than treating each modality separately, are rare. This paper begins to address this gap by introducing a new Multimodal Large Language Model (MLLM) architecture designed for anomaly detection on multimodal tabular datasets. By combining a dual objective that is comprised of an autoregressive and a pairwise ranking loss, the model is able to efficiently learn both patterns of textual and visual data and understand cross-modal correlations. Experimental results show a performance gain of 58.4% for our architecture on a diverse set of anomaly categories. Our code and datasets are available at: https://github.com/Antonis-Georgakopoulos/multimodal-tabular-ad.

**Keywords**

Anomaly Detection, Multimodal Data Quality, Multimodal Large Language Models, Multimodal Tabular Data, Pairwise Ranking Loss, Semantic Consistency, Autoregressive Modeling

## 1. Introduction

Anomalies are data points that deviate from other data points in a way suggesting they may have been generated by a different underlying process [1]. In the context of structured knowledge resources, such deviations often manifest as structural irregularities, implicit schema violations, or semantically inconsistent values [2]. Multimodal tabular data plays a critical role in different downstream applications. For example, large volumes of multimodal tabular data are transformed into knowledge graphs by many organizations to support knowledge integration and reasoning [3]. Moreover, tasks operating on tabular data, such as taxonomy inference, directly contribute to knowledge graph population [4]. Noisy and incomplete tabular data, particularly missing or ambiguous metadata, often hinder the effectiveness of different applications, such as tabular data to knowledge graph matching [5]. Different kinds of noise and errors in tables, such as misspellings and ambiguous mentions, have shown to degrade performance of entity linking algorithms even when they perform well on clean data [6]. Detecting anomalies in data is therefore essential, as their presence can have a detrimental impact on the performance and quality of downstream tasks, such as knowledge graph construction[7].

To tackle this problem, applications can employ *anomaly detection (AD)*, the task of detecting anomalies in data. The current state-of-the-art in anomaly detection for tabular data is led mainly by Transformer-based models [8], diffusion generative models [9], and contrastive methods [10, 11]. However, these approaches have not considered the multimodal nature of web data where tables often contain images or other modalities. To that end, the goal of our paper is to detect anomalies in multimodal tabular data, where each data point combines information from multiple modalities (as illustrated in Figure 1). To address this problem, we capitalise on the capabilities of multimodal large language models (MLLMs) [12] and tackle the problem of modality collapse [13], where models rely

*QKG@ESWC2026: Workshop on Quality of Knowledge Graphs at ESWC 2026, May 11, 2026, Dubrovnik , Croatia*
*Corresponding author.
a.georgakopoulos@uva.nl (A. Georgakopoulos); p.t.groth@uva.nl (P. Groth); l.stork@uva.nl (L. Stork)
0000-0003-0183-6910 (P. Groth); 0000-0002-2146-4803 (L. Stork)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

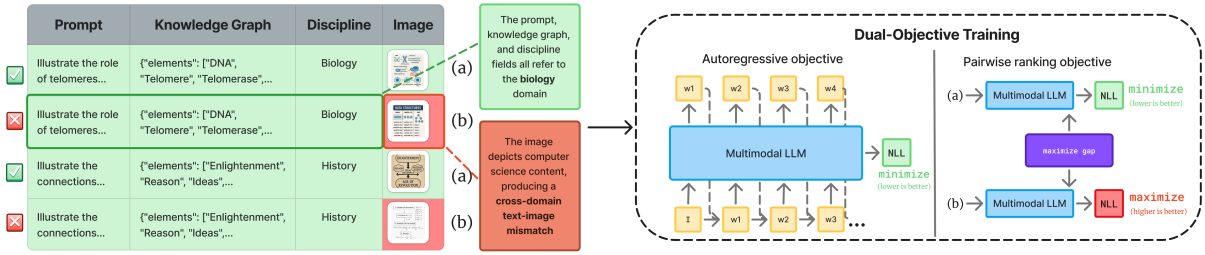


**Figure 1:** An overview of our architecture. Positive (*a*) and negative (*b*) pairs are formed by keeping the original image $I_i$, or randomly swapping it, $I_j$ for $j \neq i$, respectively. An MLLM $M_\theta$ is fine-tuned with a dual objective on these pairs: (1) the **autoregressive objective** minimises the Negative Log-Likelihood ($NLL$) of the positive pairs, and (2) the **pairwise ranking objective** maximises the difference between the $NLL$ of the positive and of the negative pairs.

heavily on one modality and ignore others. To do so, we present a new architecture that is designed to better capture the dependencies between modalities, thereby improving the detection of anomalies in multimodal tabular data.

Our contributions are:

1. We introduce a new architecture that combines autoregressive modeling with a pairwise ranking-based objective in order to capture the cross-modal dependencies in tabular multimodal data.
2. We provide a multimodal anomaly dataset, which extends a text-image benchmark, that contains different categories of anomalies allowing for systematic evaluation of anomaly detection models on multimodal tables.

## 2. Related Work

We first discuss related work with respect to anomaly detection in tabular data and then focus on the task with respect to multimodal tabular data.

### 2.1. Tabular Anomaly Detection

Many different methods have been developed to identify anomalies in structured data. These range from diffusion-based approaches [14], to deep-learning based, such as self-supervised learning approaches [15]. Tree-based methods, such as Isolation Forest [16], detect anomalies by partitioning the feature space — data points that can be isolated more easily are more likely to be outliers. Score-based approaches such as NCSBAD [9], a noise-conditional generative model, and TabADM [14], a diffusion-based generative model, both achieve state-of-the-art results on large-scale benchmarks. Additionally, SORTAD [15], is a self-supervised deep learning method that learns from small edits on tabular datasets and detects unusual rows by using a table-aware score. More recently, LLMs and Transformer-based approaches have come to spearhead the state-of-the-art in tabular anomaly detection. AnoLLM [17], a fine-tuning-based approach, outperforms all classic and deep learning baselines on mixed type tabular datasets. Similarly, a Non-Parametric Transformer for masked-cell reconstruction [8] , achieves superior results on a 31-dataset benchmark. Whereas these methods have proven to work well for unimodal tabular anomaly detection they have not been extensively tested on multimodal data.

### 2.2. Multimodal Tabular Anomaly Detection

The problem of anomaly detection in multimodal tables remains largely underexplored. To the best of our knowledge, [18] is the only approach that targets the same task setting, where rows in e-commerce tabular datasets consist of text and images. The authors inject cross-modal anomalies into four benchmarks and evaluate five baseline methods on these benchmarks for the task of error detection.

They find that current methods remain limited in their capabilities of dealing with cross-modal errors in tabular datasets, motivating our work.

# 3. Method

## 3.1. Problem definition

Let $\mathcal{T} = (x_1, x_2, \ldots, x_n)$ denote a tabular dataset with $m$ columns and $n$ rows. Each row $x_i \in \mathcal{T}$ can contain numerical, categorical and textual types and is associated with one image $I_i$. We then define the task of multimodal tabular anomaly detection as the task of finding anomalous rows in $\mathcal{T}$, given the associated image $I_i$. The goal is to learn a scoring function $S : \mathcal{T} \to \mathbb{R}$, where a higher score indicates a higher likelihood of being an anomaly.

## 3.2. Model architecture and training objective

We extend previous work on tabular anomaly detection [17] by adapting it to multimodal tabular data and introducing a dual-objective loss. An overview of our approach is shown in Figure 1 and described formally below.

Let $M_\theta$ denote a causal multimodal language model with parameters $\theta$. For an input row $x_i$, its serialised token sequence $Tokenise(x_i) = (w_1^{(i)}, \ldots, w_L^{(i)})$ — where $L$ is the length of the input sequence — and a corresponding image $I_i$, the conditional probability that $M_\theta$ defines over the sequence given the image is:

$$p_\theta(Tokenise(x_i)|I_i) = \prod_{t=1}^{L_i} p_\theta(w_t^{(i)}|I_i, w_1^{(i)}, \ldots w_{t-1}^{(i)})$$

The Negative Log-Likelihood (NLL) of the row $x_i$ will then be:

$$NLL_\theta(Tokenise(x_i)|I_i) = -\sum_{t=1}^{L_i} \log p_\theta(w_t^{(i)}|I_i, w_1^{(i)}, \ldots w_{t-1}^{(i)})$$

We fine-tune $M_\theta$ on a dual objective by training on positive tuples $a_i = (x_i, I_i)$ and negative tuples $b_i = (x_i, I_j)$ for $j \neq i$, as visualised in Figure 1. In each batch, for each positive tuple, we create one negative tuple. Our approach utilises two loss functions:

**1. Autoregressive (CLM) loss.** This loss is only computed over the positive samples and is defined as follows:

$$\mathcal{L}_{AR}(\theta) = \frac{1}{B} \sum_{i=1}^{B} NLL_\theta(a_i)$$

Where $NLL_\theta(a_i)$ is defined as $NLL_\theta(Tokenise(x_i)|I_i)$.

**2. Pairwise ranking loss.** To address the problem of modality collapse [13] we introduce a second loss function—the pairwise ranking loss—which encourages a lower NLL score for normal samples, and a higher for anomalous samples. The softplus function, defined as: $softplus(x) = log(1 + e^x)$, provides an auto-calibrated penalty that scales with the magnitude of the error. The ranking loss is then defined as:

$$\mathcal{L}_{rank}(\theta) = \frac{1}{B} \sum_{i=1}^{B} softplus(NLL_\theta(a_i) - NLL_\theta(b_i))$$

Here, B denotes the size of the batch, $NLL_\theta(a_i)$ is defined as $NLL_\theta(Tokenise(x_i)|I_i)$, and $NLL_\theta(b_i)$ as $NLL_\theta(Tokenise(x_i)|I_j)$. The final loss is then defined as the sum of both losses:

$$\mathcal{L}(\theta) = \mathcal{L}_{AR}(\theta) + \mathcal{L}_{rank}(\theta)$$

### 3.3. Input construction and column permutations

We follow the serialisation approach as described in [17] as it achieves state-of-the-art results on unimodal tabular data, and extend it to accommodate visual information. For each row in the dataset we construct a textual sequence: for a row $x_i$ with columns $(c_1, ..., c_M)$, the converted sequence is "$V(c_1)$ *is* $V(x_{i,1})$*, ...,* $V(c_M)$ *is* $V(x_{i,M})$", where $V(\cdot)$ corresponds to the cell or column's literal value. For the image column we prepend the string "$I_i$ is <|image_1|>", used by the MLLM to place the image at that specific point in the sequence. We strategically place the visual information in the front of the sequence as it has been shown to yield improved performance in a single-image reasoning task scenario [19].

Similarly to [17], we randomise column order during training, but keeping the image column fixed, to train a model that is independent of column order. During evaluation we apply a fixed set of $r$ permutations to ensure a stable and robust scoring.

## 4. Experimental Setup

### 4.1. Dataset

For our experiments, we modify the Massive Multi-Discipline Multi-Tier Knowledge-Image Generation Benchmark (MMMG) [20] by artificially introducing anomalies from six anomaly types that we describe below, with a focus on creating anomalous entries that violate the semantic dependencies within a row. Motivated by [21], two anomaly types include out-of-domain data unrelated to the MMMG domain, and for anomalies that swap images and text within the MMMG, we followed the re-pairing strategy of [22], mismatching images and textual inputs between rows in the MMMG. We create six versions of the MMMG test set, each produced by injecting one of the six anomaly types on 30% of its data. The train set contains 11,991 samples, and each test set contains 5,140 samples with 1,542 anomalous instances (30%). Both train and test sets contain three textual, one image, and two categorical features.

- `Random Noise`: A randomly chosen cell is corrupted with random character strings (50-200 random alphanumeric characters).
- `Out-of-Context Domain – Unstructured`: A randomly chosen cell in a column with structured information is replaced with *unstructured* text sampled from a domain outside MMMG (the automotive domain).
- `Out-of-Context Domain – Structured`: A randomly chosen cell in a column with structured information is replaced with *structured* text (respecting the syntax of the content it replaces) sampled from a domain outside MMMG (the automotive domain).
- `Cross-Domain Image Swap`: A randomly chosen image is replaced with one from a row about a different scientific discipline, introducing domain mismatches between the image and the rest of the row.
- `In-Column Text Swap`: A randomly chosen cell from a textual column is replaced with one from the same column, introducing textual localised inconsistencies.
- `In-Column Cell Swap`: A cell is randomly swapped with another from the same column, introducing localised inconsistencies in any modality.

The MMMG is a large-scale multimodal benchmark about scientific knowledge, designed for reasoning through the integration of textual and vision properties. Specifically it aims at evaluating the reasoning capabilities of models on the text-to-image generation task. It contains a wide range of expert-validated pairs of prompt and image that span across 10 academic disciplines (Biology, Chemistry, Mathematics, Engineering, Geography, Economics, Sociology, Philosophy, History, and Literature) and six educational levels (pre-school, primary school, secondary school, high school, undergraduate, and PhD). Each data sample contains a textual prompt, an image, a knowledge graph property that captures the key entities mentioned in the prompt and the relationships between them, and also an annotation property that provides a step-by-step explanation of the entities and their interactions within the knowledge graph. Additionally, two categorical properties that indicate the academic discipline and the education level

**Table 1**
Distribution of corrupted features per anomaly category across the six test sets (5,140 rows each, of which 1,542 (30%) are anomalous). Dashes indicate feature types that are not applicable to a given anomaly category.

| | Corrupted Feature Distribution | | | | | |
|---|---|---|---|---|---|---|
| **Anomaly Category** | **Prompt** | **KG** | **Annotation** | **Image** | **Discipline** | **Education** |
| Random Noise | 522 | 497 | 523 | — | — | — |
| OOC Domain - Unstructured | 543 | 503 | 496 | — | — | — |
| OOC Domain - Structured | 487 | 505 | 550 | — | — | — |
| Cross-Domain Image Swap | — | — | — | 1542 | — | — |
| In-Column Text Swap | 398 | 384 | 390 | — | 314 | — |
| In-Column Cell Swap | 274 | 260 | 240 | 242 | 252 | 200 |

also accompany each sample. The complexity of the knowledge graph and the level of visual detail in the image both increase with higher educational levels. The semantic dependence between the textual, visual, and categorical features in a row makes the MMMG a suitable dataset for our anomaly detection experiments.

Details of the creation of our test sets can be found in our GitHub repository[1]. The test datasets' statistics are described in Table 1.

## 4.2. Multimodal LLM and Fine-Tuning

### 4.2.1. Model description and Training

In this work we adopt the Phi-4-multimodal-instruct model (Phi4MM) [23], a lightweight open-source multimodal foundation model developed for language, vision and audio tasks, containing 5.6 billion parameters. We opt for a lightweight model, as these often perform on par with larger models while being a more cost-effective alternative [24]. We perform parameter-efficient fine-tuning (LoRA [25]) by solely training the vision encoder, projector and MLP projection layers, and using the fixed-budget fine-tuning approach of [17] with a small fixed-step setup with 1500 steps. For a detailed training configuration please refer to our GitHub repository[2]. We train our model in a supervised way using the dual objective, as shown in Figure 1 and Section 3.2. During evaluation, we use the test set $\mathcal{X}' = (x_i', y_i')_{i=1}^{m} \subset \mathcal{T}$, with $m$ being the number of rows and with $y_i' \in \{0, 1\}$ indicating the label, with 0 indicating a normal, and 1 indicating an anomalous row.

### 4.2.2. Baselines

To evaluate the impact of our custom architecture, we define two baselines. For the first baseline (Phi4MM vanilla) we leverage the Phi4MM model without any fine-tuning or task-specific training. The second baseline (Phi4MM CLM) follows a conventional fine-tuning approach where the model is being trained solely with the CLM loss function.

## 4.3. Evaluation and Metrics

Our evaluation is inspired by [17]. We evaluate each test set separately. For each test set, we generate 5 randomly sampled permutations of the textual columns, while keeping the image column fixed in place. We then calculate the token-level NLL of each row and obtain a per-permutation row score. After scoring each row under all five permutations we aggregate by averaging, resulting in the final per-row anomaly score.

[1] https://github.com/Antonis-Georgakopoulos/multimodal-tabular-ad/blob/main/data_pipeline/data_preprocessing.ipynb
[2] https://github.com/Antonis-Georgakopoulos/multimodal-tabular-ad/blob/main/anomaly_detection/config.py

Following the evaluation protocol of [17], we evaluate the performance of our trained model using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) as the primary evaluation metric. We also use the F1@k metric as an additional evaluation metric, with k being the total number of anomalies in the testing dataset.

## 5. Results

Table 2 shows the results of the three methods over all test sets. Our dual-objective method outperforms the baseline models on five out of six categories. The untrained vanilla Phi4MM model shows near-chance performance across all categories, indicating training is required for a better understanding of the row semantics. Our dual-objective approach offers notable performance gains when the tested dataset contains text-only anomalies. This is clear when comparing the performance of Phi4MM CLM with our Phi4MM DO (Dual-Objective) architecture.

**Table 2**
AUC-ROC and F1@k scores for all six testing datasets for three models: Vanilla (no pretraining), trained with Causal Language Modeling (CLM) and our Dual-Objective (DO) approach. K refers to the total number of anomalies in the testing dataset.

| | AUC-ROC | | | F1@k | | |
|---|---|---|---|---|---|---|
| **Anomaly Category** | **Vanilla** | **CLM** | **DO** | **Vanilla** | **CLM** | **DO** |
| Random Noise | 0.45 | **0.72** | 0.64 | 0.25 | **0.51** | 0.45 |
| OOC Domain - Unstructured | 0.39 | 0.53 | **0.99** | 0.21 | 0.30 | **0.93** |
| OOC Domain - Structured | 0.48 | 0.50 | **0.99** | 0.26 | 0.31 | **0.96** |
| Cross-Domain Image Swap | 0.39 | 0.52 | **0.99** | 0.22 | 0.31 | **0.99** |
| In-column Text Swap | 0.48 | 0.51 | **0.82** | 0.28 | 0.29 | **0.68** |
| In-column Cell Swap | 0.47 | 0.51 | **0.78** | 0.25 | 0.29 | **0.63** |

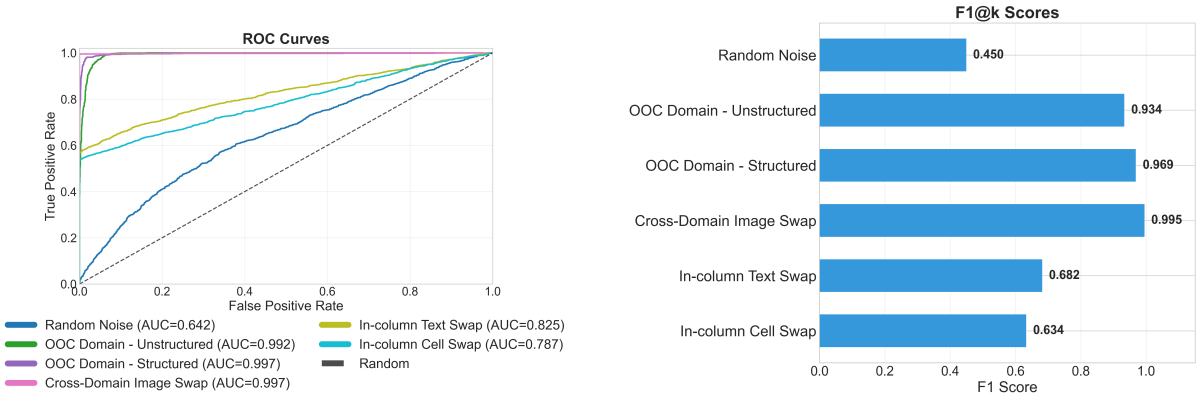


**Figure 2:** ROC curves (left) and F1@k scores (right) for the Dual-Objective model across six anomaly categories.

Additionally, we see a major performance boost on vision-related anomaly scenarios. As evident from the performance of all models on the Cross-Domain Image Swap dataset, the dual-objective approach yields an almost perfect score, suggesting that our approach has improved cross-modal reasoning beyond what the CLM objective can capture. However, a noteworthy exception is the Random Noise category where the CLM baseline outperforms our dual-objective model. One possible explanation for this discrepancy is that unlike the rest of the text-based anomaly categories, which contain plausible-looking values, Random Noise contains random strings that are improbable based on the row context. The dual-objective approach leans more on the semantic and cross-modal understanding, which likely offers a better advantage on contextually subtle anomalies. When an anomaly is trivial, such as in

the case of Random Noise, the CLM baseline benefits from the single objective, which encourages the model to learn representations that are more sensitive to such irregularities.

The F1@k results show a similar pattern. Our dual-objective approach again achieves superior performance over the baselines on five out of six categories. The Random Noise and In-column Cell Swap datasets appear to be the most challenging overall in our evaluation as also shown in Figure 2. While these outcomes are promising, this study represents an initial evaluation of the proposed architecture, validated using a single MLLM and dataset (MMMG). Future work will extend the evaluation across a broader range of models and datasets to further assess its generalisability.

## 6. Limitations and Future Work

While the results of our proposed fine-tuning regime show promise, our work could be further strengthened by addressing several important limitations. As described in [26], the performance of different MLLMs varies greatly depending on the downstream task, not only because of the reasoning capabilities of each model, but also due to the data used during the training phase. By using a diverse range of MLLMs with different number of parameters and inference strengths, we can understand whether our suggested architectural modifications ensure an equivalently high performance across different models on the same task. Therefore, future work will extend the evaluation across a broader range of MLLM models.

To the best of our knowledge, no benchmark or dataset for multimodal tabular anomaly detection was available at the time of conducting this study. As a result, we artificially created such a dataset by by perturbing cell values and altering structural relationships in the MMMG dataset 4.1. Consequently, the final performance is directly related to the types and number of anomalies that we introduced in each test set. To further strengthen our understanding of how well the proposed dual-objective architecture works for this task in other domains or in real-world settings, we aim at evaluating our approach on additional multimodal datasets; with naturally occuring anomalies as well as modified according to the anomaly categories mentioned in this work. Another possible direction would be to introduce a greater variety of anomaly types and more challenging scenarios, such as injecting multiple anomaly types within a single row.

Finally, although our approach improves the detection performance compared to the CLM training paradigm, the method used to determine the anomaly score for each sample does not offer explainability for the resulting predictions. The NLL score is computed over the entire token sequence and then averaged across permutations, so a single scalar value is calculated for each row. Flagging a row as anomalous with a high NLL score would require a practitioner to manually inspect the row to understand the source of the high anomaly score. A promising future direction would be to explore per-feature NLL scoring to understand the source of the anomaly within the row.

## 7. Conclusions

In this work, we presented a dual-objective architecture that enhances cross-modal understanding for the task of anomaly detection in multimodal tabular datasets. By combining autoregressive modeling with a pairwise ranking loss, we show that our approach consistently outperforms plain inference and single objective fine-tuning when tested on six different anomaly types. To address this, we created a multimodal tabular anomaly detection benchmark based on the MMMG dataset, as, to the best of our knowledge, no benchmark for multimodal tabular anomaly detection currently exists. Future work will focus on further validating the generalisability of our approach on a variety of different datasets and anomaly types, with a variety of different MLLMs. As tabular datasets are abundant on the web and increasingly serve as sources for knowledge graph construction [3], ensuring they are consistent and of high quality is essential. We believe that multimodal tabular anomaly detection approach plays a key role by enhancing the quality of web data and thereby the robustness of downstream applications that depend on it.

## References


[1] D. M. Hawkins, Identification of outliers, volume 11, Springer, 1980.

[2] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe, P. Szekely, A study of the quality of wikidata, Journal of Web Semantics 72 (2022) 100679.

[3] J. Arenas-Guerrero, A. Alobaid, M. Navas-Loro, M. S. Pérez, O. Corcho, Boosting knowledge graph generation from tabular data with rml views, in: European Semantic Web Conference, Springer, 2023, pp. 484–501.

[4] Z. Wu, J. Chen, N. W. Paton, Taxonomy inference for tabular data using large language models, in: European Semantic Web Conference, Springer, 2025, pp. 403–422.

[5] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, Semtab 2019: Resources to benchmark tabular data to knowledge graph matching systems, in: European Semantic Web Conference, Springer, 2020, pp. 514–530.

[6] V. Cutrona, F. Bianchi, E. Jiménez-Ruiz, M. Palmonari, Tough tables: Carefully evaluating entity linking for tabular data, in: International Semantic Web Conference, Springer, 2020, pp. 328–343.

[7] V. Ryen, A. Soylu, D. Roman, Building semantic knowledge graphs from (semi-)structured data: A review, Future Internet 14 (2022). URL: https://www.mdpi.com/1999-5903/14/5/129. doi:10.3390/fi14050129.

[8] H. Thimonier, F. Popineau, A. Rimmel, B.-L. Doan, Beyond individual input for deep anomaly detection on tabular data, in: R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, F. Berkenkamp (Eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 48097–48123. URL: https://proceedings.mlr.press/v235/thimonier24a.html.

[9] M. Hirth, E. Kasneci, Anomaly detection by estimating gradients of the tabular data distribution, 2025. URL: https://openreview.net/forum?id=7QDIFrtAsB.

[10] S. Tao, T. Zhu, H. Wang, X. Meng, Semanticmask: a contrastive view design for anomaly detection in tabular data, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, pp. 2370–2378.

[11] J. Gao, C. Tao, Z. Sun, X. Jiang, S. Ma, Semi-supervised anomaly detection through denoising-aware contrastive distance learning, in: Proceedings of the ACM on Web Conference 2025, WWW '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2111–2119. URL: https://doi.org/10.1145/3696410.3714626. doi:10.1145/3696410.3714626.

[12] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in neural information processing systems 36 (2023) 34892–34916.

[13] M. Y. Sim, W. E. Zhang, X. Dai, B. Fang, Can vlms actually see and read? a survey on modality collapse in vision-language models, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 24452–24470.

[14] G. Zamberg, M. Salhov, O. Lindenbaum, A. Averbuch, Tabadm: Unsupervised tabular anomaly detection with diffusion models, arXiv preprint arXiv:2307.12336 (2023).

[15] G. Hay, P. Liberman, Sortad: Self-supervised optimized random transformations for anomaly detection in tabular data, arXiv preprint arXiv:2311.11018 (2023).

[16] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 eighth ieee international conference on data mining, IEEE, 2008, pp. 413–422.

[17] C.-P. Tsai, G. Teng, P. Wallis, W. Ding, Anollm: Large language models for tabular anomaly detection, in: Y. Yue, A. Garg, N. Peng, F. Sha, R. Yu (Eds.), International Conference on Learning Representations, volume 2025, 2025, pp. 7562–7592. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/165bbd0a0a1b9470ec34d5afec582d2e-Paper-Conference.pdf.

[18] O. Ovcharenko, S. Schelter, Towards cross-modal error detection with tables and images, arXiv preprint arXiv:2510.12383 (2025).

[19] G. Wardle, T. Sušnjak, Image first or text first? optimising the sequencing of modalities in large language model prompting and reasoning tasks, Big Data and Cognitive Computing 9 (2025) 149.

[20] Y. Luo, Y. Yuan, J. Chen, H. Cai, Z. Yue, Y. Yang, F. Z. Daha, J. Li, Z. Lian, Mmmg: A massive,

multidisciplinary, multi-tier generation benchmark for text-to-image reasoning, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.

[21] D. Hendrycks, M. Mazeika, T. Dietterich, Deep anomaly detection with outlier exposure, Proceedings of the International Conference on Learning Representations (2019).

[22] G. Luo, T. Darrell, A. Rohrbach, NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6801–6817. URL: https://aclanthology.org/2021.emnlp-main.545/. doi:10.18653/v1/2021.emnlp-main.545.

[23] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen, et al., Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, arXiv preprint arXiv:2503.01743 (2025).

[24] S. Subramanian, V. Elango, M. Gungor, Small language models (slms) can still pack a punch: A survey, arXiv preprint arXiv:2501.05465 (2025).

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.

[26] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, D. Yu, Mm-llms: Recent advances in multimodal large language models, Findings of the Association for Computational Linguistics: ACL 2024 (2024) 12401–12430.