

Content Learning for Metadata Extraction and Enrichment of Historical Hand-Drawn Illustrations

Rajae El-Idrissi¹[0009–0008–9705–3550], Juba Agoun¹[0000–0003–4911–0617], Jérôme Darmont¹[0000–0003–1491–384X], and Sabine Loudcher¹[0000–0002–0494–0169]

Université Lumière Lyon 2, UR ERIC, Lyon, France
{rajae.el-idrissi, juba.agoun1, jerome.darmont,
sabine.loudcher}@univ-lyon2.fr

Abstract. The digitisation of cultural heritage collections requires automated metadata generation to improve resource discovery and reuse. This paper presents the Content Learning for Metadata Extraction and Enrichment (CLEAD), a multimodal approach for generating metadata from illustrations in digitised archaeological diaries. CLEAD combines illustration detection and segmentation with vision-language and large language models to generate and enrich illustration descriptions using page-level textual context. Experiments on the DataLAC and IlluHisDoc datasets show that integrating visual and textual information improves the semantic quality and relevance of the generated metadata. These results demonstrate that CLEAD effectively supports the indexing and retrieval of visual heritage resources in digital archives and data lakes.

Keywords: Data Lakes · Metadata · Artificial Intelligence · Vision-Language Models · Historical Documents · Document Analysis

1 Introduction

Data lakes are widely used to store large, heterogeneous digitised collections, but storing data alone is insufficient. Without metadata, a data lake becomes a so-called data swamp that cannot be found or reused [7, 20]. This issue appears across diverse domains where collection growth outpaces metadata creation [7].

The description of each data entity within a data lake is often created manually and is therefore time-consuming. Several research studies have proposed machine learning approaches to support metadata generation [1, 2]. In this paper, we address the problem of metadata generation for image entities within a data lake, with the aim of automatically producing image descriptions. In this work, we focus on a specific type of image, namely hand-drawn illustrations. To the best of our knowledge, no existing approach is capable of automatically generating descriptions for hand-drawn illustrations. We selected this particular type of image because it reflects a real-world challenge encountered in the DataLAC project, which aims to develop a data lake for organizing, querying, and analyzing the archaeological archives of a site in Spain [6].

In the DataLAC project we handle excavation diaries, which mix handwritten text, drawings and sketches called illustrations. Though Handwritten Text Recognition (HTR) can transcribe texts [16, 18], illustrations remain undescribed. A vision-language model can caption an illustration, but it doesn’t take into account the archaeological context (e.g., excavation sector, related artefact) recorded in the surrounding text (Figure 1).

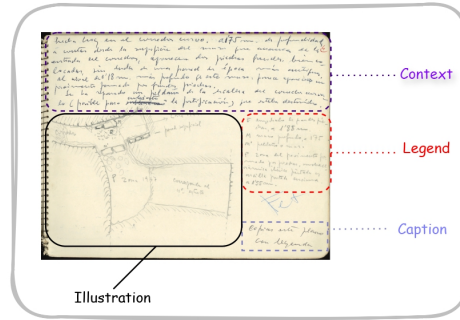


Fig. 1: Example excavation diary page

Combining visual content with page-level textual information transforms generic captions into meaningful metadata that supports retrieval and analysis. This work addresses two questions: how can vision-language models accurately describe illustrations, and how can these descriptions be enriched with textual context? To address these questions, we propose the Content Learning for Metadata Extraction and Enrichment (CLEAD), an approach that detects, segments, describes and enriches illustrations through multimodal analysis.

The remainder of this paper is structured as follows. Section 2 reviews related works. Section 3 presents the CLEAD approach. Section 4 describes two corpus and an evaluation protocol. Section 5 reports experimental results. Finally, Section 6 concludes and discusses future research.

2 Related Work

Automatically generating metadata has become a key research direction for managing large, heterogeneous digital collections. Early approaches relied on supervised learning, clustering, and topic modeling [3], while more recent work has shifted toward deep learning and large language models to produce richer, semantically grounded metadata with less human effort.

In multimodal settings, Boukhers and Bouabdallah [2] combined Bi-LSTM and Maks R-CNN to extract metadata from mixed- content documents, and He

and Wang [8] used Bi-LSTM networks to standardise metadata across heterogeneous sources. LLM-based frameworks such as AutoDDG [25] and LEDD [1] further demonstrated that models like GPT and LLaMa can generate structured metadata directly from raw documents, substantially reducing annotation costs. For semantic enrichment, Mihindukulasooriya et al. [13] paired large language models with knowledge graphs to align metadata with domain ontologies, while Lobo et al. [12] used foundation models to produce human-readable semantic descriptions. Despite this progress, most existing work targets textual or tabular content. Automatic description of visual embedded in historical handwritten documents remains underexplored. CLEAD addresses this gap by combining vision-language models with page-level transcription context to generate enriched metadata for illustrations in digitised archaeological diaries.

3 CLEAD Approach

CLEAD consists of three sequential stages that incrementally enrich illustration metadata: identification and segmentation of visual regions within heterogeneous page layouts, followed by automated generation of context-sensitive descriptions linking visual and semantic data to each illustration (Figure 2).

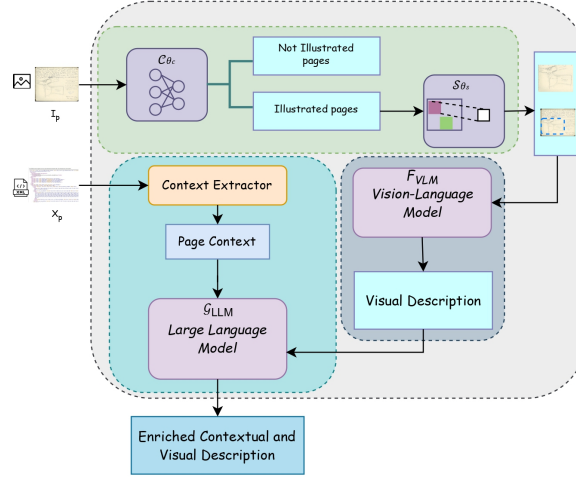


Fig. 2: CLEAD approach: from illustration detection to enriched metadata

3.1 Formalization

The CLEAD approach generates enriched illustration metadata for each page in a digital collection of archaeological diaries through three sequential steps.

For any given page, e.g., page number p , the page image is denoted as I_p and its XML/ALTO transcription as $\mathcal{X}_p$. The progressive development of the final illustration metadata M_p within CLEAD is formalised by Equation 1.

$$M_p = \mathcal{G}_{\text{LLM}} \left(\mathcal{F}_{\text{VLM}} \left(\mathcal{S}_{\theta_s} (\mathcal{C}_{\theta_c} (I_p)) \right), \mathcal{X}_p \right) \quad (1)$$

In Equation 1, each element below corresponds to a distinct stage of the approach.

1. **Illustration Classification** $\mathcal{C}_{\theta_c}$. This initial stage automatically determines whether a page image I_p contains any illustrations, using a specific classifier with parameters θ_c .
2. **Illustration Segmentation** $\mathcal{S}_{\theta_s}$. For pages identified as containing illustrations, this stage further isolates and outlines the corresponding illustration regions by employing a segmentation model with parameters θ_s .
3. **Visual Description and Enrichment** $\mathcal{F}_{\text{VLM}}, \mathcal{G}_{\text{LLM}}$. This stage first utilises a vision-language model $\mathcal{F}_{\text{VLM}}$ to generate initial visual captions. Subsequently, a large language model $\mathcal{G}_{\text{LLM}}$ refines these descriptions by incorporating contextual information from the page transcription $\mathcal{X}_p$ to ensure relevance and accuracy.

This formalism captures the full CLEAD approach, from visual detection to context-aware semantic enrichment of archaeological diary illustrations.

3.2 Illustration Identification and Segmentation

The first stage of CLEAD identifies and localises illustrations on digitised diary pages using a supervised ResNet classifier [22], which effectively distinguishes illustrated pages from text-only pages despite challenges such as handwriting variability and image degradation. Where pages contain illustrations, a U-Net [19] segmentation model then isolates the precise contours of each illustration, generating distinct regions that serve as the foundation for subsequent metadata enrichment, as formalised by $\left(\mathcal{S}_{\theta_s} (\mathcal{C}_{\theta_c} (I_p)) \right)$.

3.3 Visual-Based Illustration Description

CLEAD’s second step creates descriptive metadata for each identified illustration, focusing on visual content. This stage uses the function $\mathcal{F}_{\text{VLM}}$, where advanced vision-language models, such as BLIP-2 [10], CLIP [17], GIT [23], ViT-GPT2 [14] and large-scale multimodal systems like LLaVA-1.5 [11], Gemini 2.5 [5] and GPT-4o [9], automatically generate natural language descriptions that accurately capture the salient visual characteristics of each illustration. Such models train on large-scale image and text corpora, which enables CLEAD to produce context-free captions as the initial layer of metadata. This visually grounded metadata is critical for effective indexing, retrieval and organisation of illustrations within extensive digital diary collections.

3.4 Contextual Enrichment with Page Transcriptions

In the third stage, formalised as G_{LLM} , initial visual descriptions are enriched with contextual information extracted from page-level transcriptions, including captions, labels, and XML/ALTO metadata. Gemini 2.5 [5], GPT-4o, GPT-4-turbo [9], and GPT-3.5-turbo [24] combine visual descriptions with textual context to generate more precise and interpretable metadata for indexing and retrieval. By generating visual captions first and enriching them with transcription data in a second step, CLEAD preserves visual grounding, limits transcription noise, and ensures compatibility with models that do not support contextual inputs, while enabling independent evaluation of each stage.

4 Experimental Protocol

4.1 Datasets

To evaluate the effectiveness of the CLEAD approach, experiments are conducted on two datasets: the DataLAC corpus and the IlluHisDoc benchmark [15].

DataLAC (Archaeological Excavation Diaries) The DataLAC dataset [6] originates from the digitisation of archaeological archives from the Puig de Sant Andreu (Ullastret) excavation site¹², comprising thirteen excavation diaries written between 1947 and 1977 by Miquel Oliva i Prat and collaborators. These documents mix handwritten notes, sketches, plans, photographs, and technical observations related to archaeological fieldwork. The digitised collection includes 2,558 scanned pages, of which 563 contain illustrations such as drawings, excavation plans, stratigraphic profiles, artefact representations, and maps. Illustrations were manually extracted by domain experts, and the dataset provides partial XML/ALTO transcriptions describing page content and layout.

IlluHisDoc (Historical Document Benchmark) The IlluHisDoc, introduced by Monnier and Aubry [15], is a public benchmark designed for illustration segmentation in historical documents. The dataset contains 400 annotated pages from various manuscript and printed sources, including scientific diagrams, illuminations, and mixed-layout documents. Pixel-level annotations enable accurate evaluation of segmentation methods across diverse document layouts and illustration styles.

4.2 Preprocessing and Training

Stratified splits address class imbalance, and data augmentation is applied through rotations, scaling, brightness adjustments, and horizontal flips. Polygonal annotations are converted into binary masks for segmentation training. XML/ALTO

¹ <http://datalac.huma-num.fr/>

² <http://www.macullastret.cat/>

files are parsed to extract captions and surrounding text, with relevant sentences selected based on spatial proximity to illustration regions. Non-English segments are translated using Helsinki-NLP MarianMT models [21]. ResNet is fine-tuned from pretrained weights using cross-entropy loss and Adam $lr = 10^{-4}$ over 3 epochs, while the U-Net architecture is trained with the same loss and Adam $lr = 10^{-3}$ for 100 epochs. All experiments ran on a single NVIDIA RTX A6000 GPU (48 GB VRAM), and the implementation is publicly available³.

4.3 Prompting Procedure

State-of-the-art vision-language models are evaluated using a prompt-based approach without additional fine-tuning. Prompts were designed in collaboration with archaeologists to produce detailed, factual, and domain-relevant descriptions (Table 1), and are available in the project repository³. The evaluated models – GIT, ViT-GPT2, BLIP-2, CLIP, Gemini 2.5, GPT-4o, and LLaVA-1.5 – are tested under zero-shot and few-shot settings, with BLIP-2 and CLIP combined with FlanT5 [4] to improve output quality. For contextual enrichment, visual descriptions and XML/ALTO metadata are provided to Gemini 2.5, GPT-4o, GPT-4-turbo, and GPT-3.5-turbo through dedicated prompts.

Table 1: Prompting Protocol in Visual Description

Model	Zero-shot	Few-shot	Prompt
Git	✓	✓	Prompt 1 and 2 and 4
ViT-GPT2	✓	✓	Prompt 1 and 2 and 4
BLIP-2	✓	✓	Prompt 1 and 2 and 4
CLIP	✓	✓	Prompt 1 and 2 and 4
Gemini 2.5	✓	×	Prompt 1 and 2
GPT-4o	✓	×	Prompt 1 and 2
LLaVA-1.5	✓	✓	Prompt 1 and 3

4.4 Evaluation Metrics

To evaluate the performance of CLEAD, we use a combination of quantitative metrics and qualitative expert assessment, reflecting the different tasks involved in the approach.

For illustration classification, we report *precision*, *recall*, and *F1-score*. For segmentation, we use the *Dice coefficient*, *mean Intersection over Union (mIoU)*, and *pixel accuracy*. As the CLEAD approach includes generative description and contextual enrichment stages, metadata quality is assessed through expert evaluation. Archaeological specialists evaluated the outputs according to visual accuracy, archaeological relevance, contextual consistency, and usefulness for

³ <https://git.msh-lse.fr/rajaeel-idrissi/clead-approach>

indexing and retrieval. The protocol, designed in collaboration with domain experts, was applied to 34 illustrations for visual description and 14 illustrations for contextual enrichment. Experts examined both the illustrations and their associated transcriptions before selecting the most relevant model output based on overall quality, accuracy, and interpretability.

5 Results and Discussion

5.1 Visual Description

Both model architecture and prompting strategy significantly influence the quality of generated visual descriptions. On DataLAC, smaller models such as GIT and ViT-GPT2 fail to produce accurate archaeological descriptions even with few-shot prompting, indicating limited generalisation to this domain. BLIP-2 and CLIP combined with FlanT5 generate more relevant and context-specific descriptions, particularly when expert examples are included in the prompts. Among large multimodal models, Gemini 2.5 and GPT-4o consistently produce the richest descriptions when using structured prompts, effectively capturing the salient visual characteristics of each illustration such as drawing type, layout, visible annotations, and possible archaeological context. LLaVA with few-shot Prompt 3 also produces results comparable to archaeologist-written descriptions. Expert evaluation confirms these findings, with Expert 1 unanimously selecting Gemini 2.5 and Expert 2 unanimously selecting GPT-4o as the most coherent, informative, and contextually appropriate model across all 34 evaluated illustrations. Results on IlluHisDoc further confirm that prompt engineering and model selection are critical factors, with GPT-4o combined with a structured descriptive prompt producing the most accurate and informative descriptions by effectively distinguishing illustrations from surrounding textual content.

Prompt 1: Expert Prompt for Visual Description

You are an archaeological archivist. Describe this illustration in detail: 1. What objects or scenes are shown?

2. Drawing style and technique

3. Any visible text or labels

4. Likely archaeological context

5. Purpose of the drawing. Be factual and detailed.

5.2 Contextual Enrichment

The contextual enrichment experiments demonstrate that combining page transcriptions with visual descriptions significantly improves the semantic quality of the generated metadata. Across all evaluated models – Gemini 2.5, GPT-4o,

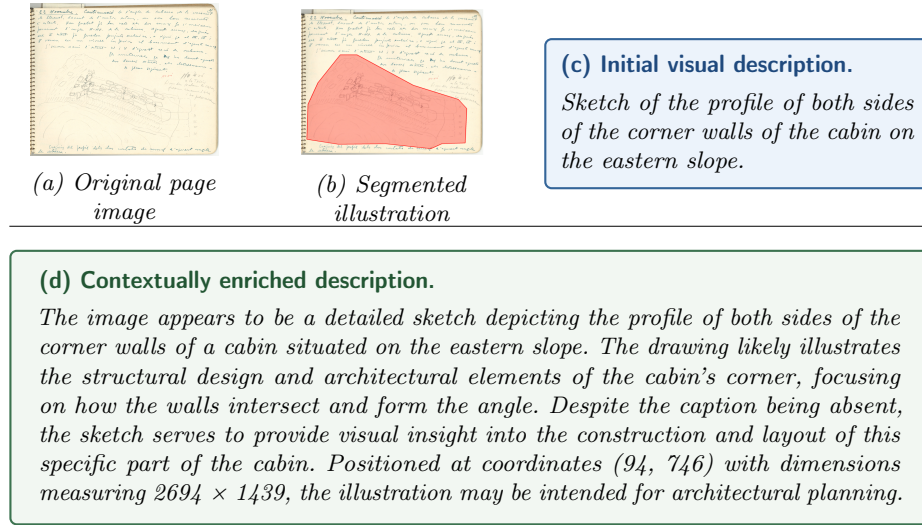


Fig. 3: Example of the CLEAD Approach: from illustration detection to enriched description. Starting from a digitised historical document (a), the U-Net segmentation model isolates the illustration region (b). A vision-language model then produces an initial visual description (c), subsequently enriched with archival context (d).

Prompt 1: Expert Prompt for Enriching with Context

Input contains English + Spanish/Catalan. Output English only.

*Description (English): **Visual Description***

*Metadata (Catalan/Spanish): **Paper context***

Enrich description with metadata. Output English only.

GPT-4-turbo, and GPT-3.5-turbo – both enrichment prompts produce high-quality outputs, though with notable differences. While Prompt 2 generates more elaborate interpretations incorporating broader archaeological context, Prompt 1 provides clearer and more concise descriptions better aligned with expert archaeological documentation practices. GPT-4o achieves the best overall performance, selected by the expert in 13 out of 14 cases (92.9%), confirming its superior ability to integrate visual and textual cues into coherent and contextually relevant metadata. Gemini 2.5 and GPT-4-turbo produce no satisfactory outputs, while GPT-3.5-turbo is selected in only one case, highlighting the importance of model capability for this task. As illustrated in Figure 3, the enrichment stage refines the initial visual description by incorporating contextual information from the page transcription. GPT-4o again produces the most coherent enriched descriptions on IlluHisDoc, though some limitations remain in generating numerical measurements and spatial dimensions.

6 Conclusion and Future Work

This paper presents CLEAD, an approach for the automatic generation of enriched metadata. By combining vision-language models with page transcription context, CLEAD produces more informative and semantically meaningful metadata than visual analysis alone. Experiments on the DataLAC and IlluHisDoc datasets show the effectiveness of integrating visual and textual information for improving metadata quality and supporting the indexing and retrieval of visual heritage resources. Future work will focus on incorporating domain knowledge and expert validation mechanisms to further enhance metadata accuracy and reliability.

References

1. An, Q., Ying, C., Zhu, Y., Xu, Y., Zhang, M., Wang, J.: Ledd: large language model-empowered data discovery in data lakes. arXiv preprint arXiv:2502.15182 (2025)
2. Boukhers, Z., Bouabdallah, A.: Vision and natural language for metadata extraction from scientific pdf documents: a multimodal approach. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries. pp. 1–5 (2022)
3. Cherradi, M., El Haddadi, A.: Enhancing data lake management systems with lda approach. *Journal of Data Science and Intelligent Systems* **3**(1), 58–66 (2025)
4. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022), <https://arxiv.org/abs/2210.11416>
5. Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T., Jacobsson, H., Szpektor, I., Jiang, N.J., Hahn, C.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (07 2025). <https://doi.org/10.48550/arXiv.2507.06261>
6. El-Idrissi, R., Makhlouf, E., Sahrane, M.R., Romeo, D., Loudcher, S.: Datalac: intelligence artificielle dans un lac de données pour valoriser les archives archéologiques. In: RNTI (2026)
7. Hai, R., Koutras, C., Quix, C., Jarke, M.: Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering* **35**(12), 12571–12590 (2023)
8. He, F., Wang, D.: Cloze: a building metadata model generation system based on information extraction. In: Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. pp. 109–118 (2022)
9. Islam, R., Moushi, O.M.: Gpt-4o: The cutting-edge advancement in multimodal llm (07 2024). <https://doi.org/10.36227/techrxiv.171986596.65533294/v1>
10. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), <https://arxiv.org/abs/2304.08485>
12. Lobo, E., Hassanzadeh, O., Pham, N., Mihindukulasooriya, N., Subramanian, D., Samulowitz, H.: Matching table metadata with business glossaries using large language models. arXiv preprint arXiv:2309.11506 (2023)
13. Mihindukulasooriya, N., Dash, S., Bagchi, S., Chowdhury, M.F.M., Gliozzo, A.M., Farkash, A., Glass, M.R., Gokhman, I., Hassanzadeh, O., Pham, N., et al.: Unleashing the potential of data lakes with semantic enrichment using foundation models. In: ISWC (Posters/Demos/Industry) (2023)
14. Mishra, S., Seth, S., Jain, S., Pant, V., Parikh, J., Jain, R., Islam, S.M.: Image caption generation using vision transformer and gpt architecture. In: 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT). pp. 1–6. IEEE (2024)
15. Monnier, T., Aubry, M.: docextractor: An off-the-shelf historical document element extraction. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). p. 91–96. IEEE (Sep 2020). <https://doi.org/10.1109/icfhr2020.2020.00027>
16. Nockels, J., Gooding, P., Ames, S., Terras, M.M.: Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of transkribus in published research. *Archival Science* **22**, 367 – 392 (2022). <https://doi.org/10.1007/s10502-022-09397-0>
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Rakesh, S., Reddy, P.K., Prashanth, V., Reddy, K.S.: Handwritten text recognition using deep learning techniques: A survey. *MATEC Web of Conferences* (2024). <https://doi.org/10.1051/mateconf/202439201126>
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (01 2015)
20. Sawadogo, P., Darmont, J.: On data lake architectures and metadata management. *Journal of Intelligent Information Systems* **56**(1), 97–120 (2021)
21. Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.A., Nieminen, T., Raganato, A., Scherrer, Y., Vazquez, R., Virpioja, S.: Democratizing neural machine translation with opus-mt (2023), <https://arxiv.org/abs/2212.01936>
22. Tomandl, A., Fuchs, K.P., Federrath, H.: Rest-net: A dynamic rule-based ids for vanets. In: 2014 7th IFIP Wireless and Mobile Networking Conference (WMNC). pp. 1–8 (2014). <https://doi.org/10.1109/WMNC.2014.6878854>
23. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language (2022), <https://arxiv.org/abs/2205.14100>
24. Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., Huang, X.: A comprehensive capability analysis of gpt-3 and gpt-3.5 series models (2023), <https://arxiv.org/abs/2303.10420>
25. Zhang, H., Liu, Y., Santos, A., Hung, W.L., Freire, J.: Autoddg: Automated dataset description generation using large language models. *Proceedings of the ACM on Management of Data* **4**(1 (SIGMOD)), 1–27 (2026)