Autonomously Detecting Defects in Circular Production Using Multimodal Large Language Models and Retrieval Augmentation

Koch, Dominik[1]; Ohland, Marvin[1]; Wen, Di[2]; Peng, Kunyu[2]; Benfer, Martin[1]; Stiefelhagen, Rainer[2]; and Lanza, Gisela[1]

[1] wbk Institute for Production Science, Karlsruhe Institute for Technology (KIT), Kaiserstr. 12, 76131 Karlsruhe, Germany; https://www.wbk.kit.edu/
[2] Computer Vision for Human-Computer Interaction Lab, Karlsruhe Institute for Technology, Adenauerring 10, 76131 Karlsruhe, Germany; https://cvhci.iar.kit.edu/587.php

Abstract. Automated defect detection is a key enabler for scalable inspection in circular production, where varying component conditions and the scarcity of large-scale annotated datasets challenge conventional supervised vision models. This paper investigates whether multimodal large language models (MLLMs) can serve as a flexible alternative for industrial defect detection without task-specific training. To bridge the domain gap of general-purpose models, several retrieval-augmented generation (RAG) architectures are designed to leverage expert-defined examples for visual defect detection on used gears in a circular production context. The study compares baseline prompting with text-based, image based, and self-reflective retrieval strategies that dynamically provide context-relevant examples and descriptions during inference. The results show that MLLMs already achieve good defect detection accuracy out-of-the-box. Integrating RAG increases defect detection accuracy to 95.83% and further improves fine-grained defect classification, with gains of up to 63.94% for specific rare defect types. While the models remain weaker in precise localization and still lack deeper object-level understanding, the results indicate that RAG-enhanced MLLMs are a viable, low-barrier solution for inspection tasks in remanufacturing scenarios with limited expert supervision and diverse object classes. These findings establish a reproducible benchmark for RAG-augmented VLMs in defect classification and demonstrate their viability in low-data, multi-class remanufacturing scenarios.

Keywords: Remanufacturing, Defect Detection, Vision-Language Models, Retrieval-Augmented Generation, Circular Economy, Quality Inspection

1 Introduction

Global material extraction has surged from 30 to over 106 billion metric tons between 1970 and 2024, intensifying resource scarcity and price volatility [1]. Remanufacturing, a core strategy of Circular Production, restores end-of-use products to a “like-new” condition through disassembly, cleaning, inspection, repair, and reassembly [2]. A

critical bottleneck in this process chain is inspection: each returned component must be assessed for defects quickly and accurately to determine its reuse pathway [3].

Traditional Industrial Anomaly Detection (IAD) relies on supervised models that require large, object-specific training datasets and extensive compute time [4]. In remanufacturing, however, product diversity is high, return volumes per variant are low, and defect patterns vary across lifecycles. Prior work has addressed parts of this challenge through synthetic data generation for inspection [5] and adaptive view planning via reinforcement learning [6], but the core problem of limited real defect data persists. Recent advances in Multimodal Large Language Models (MLLMs), also referred to as Vision-Language Models (VLMs), offer a promising alternative. Models such as GPT-4o can jointly process images and text, enabling in-context learning without weight updates [7]. Nevertheless, stand-alone VLMs lack domain-specific inspection expertise. Retrieval-Augmented Generation (RAG) addresses this gap by dynamically supplying task-relevant evidence from an external knowledge base during inference [8].

While several studies have explored VLMs for anomaly detection [9], [10], [11], a systematic comparison of different RAG architectures and retrieval modalities for industrial defect classification remains scarce. This paper addresses this gap by evaluating RAG-augmented VLMs for defect detection and classification in low-data remanufacturing scenarios, comparing text-based, image-based, and self-reflective retrieval modalities, and assessing the impact of retrieval quality on defect detection. As a representative use case, used gear wheels from an angle grinder serve as the test object, as they represent a common component in remanufacturing with well-defined defect morphologies. In the following, section 2 reviews related work on VLMs and RAG for anomaly detection. Section 3 describes the dataset, prompting architectures, and evaluation criteria. Section 4 reports results, followed by a discussion in Section 5 and conclusions in Section 6.

2 Related Work

MLLMs extend classic LLMs by processing text and visual content in a shared context. A typical architecture encodes the input image into patch vectors, compresses them into latent tokens via a resampler, and injects them into the language model through cross-attention, where they are processed alongside text tokens to generate a response grounded in the visual input [9]. RAG addresses key limitations of standalone LLMs, e.g. hallucination, outdated knowledge, and narrow domain coverage, by retrieving task-relevant evidence from an external knowledge base at inference time [8], [10]. A query is embedded and compared against indexed document chunks; the top-k most similar chunks are retrieved and prepended to the prompt, enabling the model to ground its output in current, domain-specific information without retraining.

Fine-tuned VLM architectures have demonstrated strong anomaly-detection capabilities. AnomalyGPT [11] integrates a pretrained image encoder with an LLM via lightweight prompt-learning and feature-matching modules, supporting few-shot inference through memory banks. FabSage [12] combines a learnable prompt expert with a defect map generated from image–text embeddings. While these approaches achieve

competitive performance, they require domain-specific fine-tuning and labelled datasets, resources often unavailable in remanufacturing [3]. RAG mitigates the knowledge gap of generic VLMs by retrieving relevant documents or images at inference time [8], [10]. [13] applied a multimodal RAG pipeline for additive-manufacturing defect detection, improving accuracy by 12%. [12] benchmarked VLMs across 38 product categories and found that few-shot prompting with up to four reference images yields the best trade-off, with diminishing returns beyond. [14] confirmed this using fixed few-shot examples and majority voting. A conceptual advance is Self-RAG [15], where the model uses reflection tokens to decide adaptively whether retrieval is needed and scores the relevance of retrieved passages. [16] proposed a related self-reflection method with iterative factual-knowledge and answer-consistency loops. Both approaches reduce hallucination by filtering irrelevant content before generation.

Existing work either uses fixed examples without investigating retrieval quality or applies RAG in a single modality. No study has systematically compared text-based, image-based, and self-reflective retrieval strategies for VLM-based defect classification in remanufacturing, nor quantified how retrieval accuracy and example diversity translate to downstream performance. The authors prior work on automated visual inspection [6], [17] and semantic 3D product modelling [18] provides complementary geometric and planning capabilities but does not address the classification task targeted here.

3 Methodology

To evaluate RAG-augmented VLMs for defect detection, a study on used gear wheels from angle grinders was conducted as a representative remanufacturing use case. A dedicated image dataset was created by synthetically introducing three defect types into new gear wheels. Five prompting architectures were implemented, ranging from zero-shot to text-based, image-based, and self-reflective RAG and evaluated against this dataset. Section 3.1 describes the image dataset created, Section 3.2 presents the five prompting architectures, from zero-shot to self-reflective RAG. Section 3.3 covers the VLM selection process and the evaluation metrics used to assess performance.

3.1 Dataset Creation

A standardised image-acquisition setup was built around a Basler acA5472-17uc camera mounted at 17 cm above the test object, a gear wheel, with ring illumination ensuring consistent capture conditions. Starting from 15 new gear wheels of an angle grinder, three representative defect types were synthetically introduced through manual machining: Tooth Clipping (TC) – partial or complete loss of tooth material on the outer rim; Surface Scratches (SS) – superficial thin lines or point-like damage in the inner circle; and Dirt (D) – dust, oil, or metal particles distributed across the gear. Defect severity was varied by adjusting treatment duration and pressure. An image was captured after each incremental modification, yielding up to eight scenarios per physical gear wheel and 51 unique defect configurations. For each scenario, four image variations were recorded: standard (frontal), rotated, angled (20° tilt), and blurred. An equal number of

defect-free images was added, resulting in a balanced dataset of 408 images. Single-category and multi-category combinations (TC, SS, D, TCSS, TCD, SSD, TCSSD) are represented. An overview is given in Figure 1. Each image was annotated with a structured ground-truth description comprising a summary stating the number and type of defects, and per-defect detail blocks with category, observation.

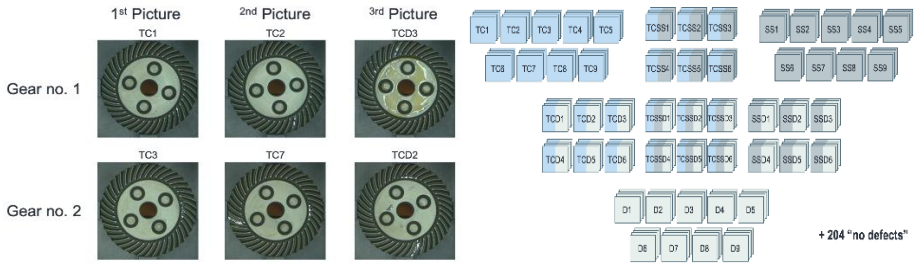

Fig. 1. Overview of the complete dataset of 408 images across two exemplary gear wheels, each captured from three perspectives with incrementally introduced defects (Tooth Clipping, Surface Scratches, Dirt) and their combinations, plus 204 defect-free samples.

3.2 Prompting Architectures

Five architectures were implemented, all sharing a common Zero-Shot prompt backbone. The prompt defines the model's role as a specialised inspection system, prescribes a step-by-step procedure, following evidence that structured prompting improves LLM performance [19], specifies the output structure (summary + per-defect blocks), and lists seven defect-category definitions: three real (TC, SS, D) and three distractors (Rust, Wear Out, Production Errors) to probe hallucination tendencies. The model temperature was set to 0 throughout.

Zero-Shot. The query image and text prompt are provided without any example images. This serves as the lower performance bound.

Fixed-Few-Shot (Baseline). Four manually selected example images, each with its ground-truth description, are prepended to the prompt. Following [12] and [14], $k = 4$ was confirmed as optimal in a preliminary sweep on a 60-image subset. The four examples were chosen to cover all defect categories with maximal diversity. This approach defines the baseline that any RAG architecture must surpass.

Text-based RAG. The query image is first processed with the Zero-Shot prompt to produce a textual description. This description is embedded using BAAI/LLM-Embedder and compared via cosine similarity against a ChromaDB vector store containing all ground-truth descriptions. The four most similar descriptions and their associated images are retrieved and injected into a second, Few-Shot prompt.

Picture-based RAG. The query image is directly embedded with the CLIP model openai/clip-vit-large-patch14-337, selected as the best-performing embedder in a comparison of seven CLIP variants. L2-normalised embeddings are stored in a FAISS IndexFlatIP; cosine-similarity search returns the four most visually similar images plus their ground-truth descriptions.

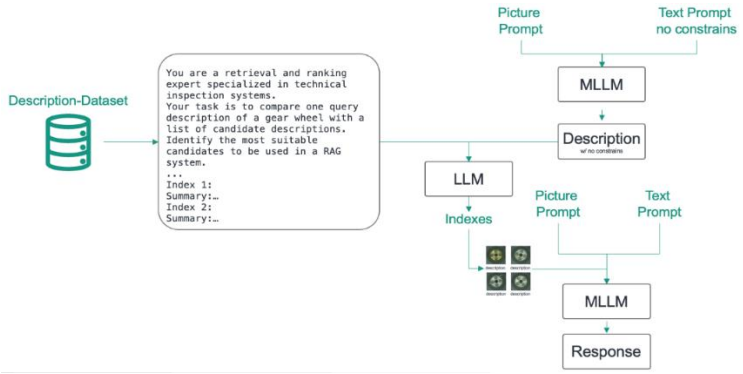

Fig. 2. Schematic representation of the Self-RAG architecture

Self-RAG. Inspired by [15] and [16], a self-reflective loop is introduced. The VLM first generates a free-form description of the query image (no output-format constraints). This description, together with all indexed candidate descriptions, is passed to a reasoning-capable model (OpenAI o1-mini) that ranks and selects the most contextually relevant candidates. The selected images then populate a standard Few-Shot prompt. Unlike cosine retrieval, this mechanism allows the model to reason for defect-category matches and complementarity. The overall pipeline is illustrated in Figure 2.

3.3 LLM Selection and Evaluation Criteria

Five VLMs were benchmarked on a 60-image subset: Llama 3.2-vision, Gemini 2.5 Flash, Gemini 2.5 Pro, Claude 3.7 Sonnet, and ChatGPT-4o. These models were selected as a representative cross-section of frontier multimodal models available at the time of the experiments, balancing state-of-the-art performance with accessible inference costs. ChatGPT-4o achieved the highest Defect Classification Accuracy (53.3% Exact Match) and the only substantial BLEU score (0.37), indicating reliable adherence to the prescribed output format. It was therefore selected for all RAG experiments

Performance is assessed along five dimensions. Defect Detection Accuracy evaluates the binary distinction between defective and non-defective components using precision, recall, and F1 as metrics. Defect Classification Accuracy (DCA) measures how well individual defect categories are identified. It is quantified through Exact Match (all predicted defect labels match the ground truth exactly) and Hamming Loss (fraction of incorrectly predicted labels). Retrieval Accuracy captures how relevant the retrieved examples are, measured by Precision@k (share of retrieved images matching the query's defect categories). Text Similarity assesses how closely the model's output matches the reference description, using Cosine Similarity and BLEU score.

4 Results

4.1 Binary Defect Detection

Table 1 summarises binary detection performance. The Zero-Shot configuration achieves 84.1% accuracy with near-perfect recall (99.7%) but low precision (76.0%), indicating frequent false positives. Adding four fixed examples raises accuracy to 95.1% with perfect precision. Picture-based RAG (95.6%) and Self-RAG (95.8%) yield marginal improvements of 0.5–0.7 percentage points. Text-based RAG drops to 74.5%, performing worse than Zero-Shot.

Table 1. Binary Defect Detection Accuracy across all RAG approaches

RAG Approach	Accuracy	Precision	Recall	F1-Score
Zero-Shot	0.841	0.760	0.997	0.862
Few-Shot	0.951	**1.000**	0.902	0.949
Text-RAG	0.745	0.671	0.961	0.790
Pic-RAG	0.956	0.927	**0.990**	**0.957**
Self-RAG	**0.958**	0.990	0.927	**0.957**

4.2 Defect Classification Accuracy

For exact multi-label classification, the picture changes substantially, as shown in Figure 3. The Zero-Shot Exact Match drops to 44.9%. The Fixed-Few-Shot baseline raises this to 63.1%, a 40.7% increase, highlighting the power of in-context learning [12]. Text-based RAG (47.6%) barely exceeds Zero-Shot, while Picture-based RAG (68.6%) outperforms the baseline by 8.7%. Self-RAG achieves the best result at 73.5%, a 16.5% improvement over the baseline. Under (1 − Hamming Loss), all architectures cluster between 88.8% and 95.0%, indicating that errors are typically single-category mismatches.

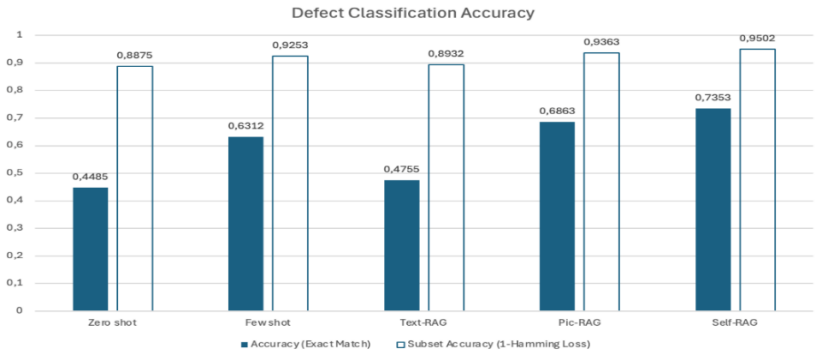

Fig. 3. Defect Classification Accuracy across all RAG approaches

4.3 Retrieval Accuracy

Picture-based RAG achieves the highest retrieval Exact Match, with a clear decline in relevance as rank increases. Text-based RAG surprisingly yields similar per-image retrieval accuracy despite its poor end-to-end classification. Self-RAG retrieves less category-matched images overall but provides greater diversity across the four examples, a property that proves beneficial for classification.

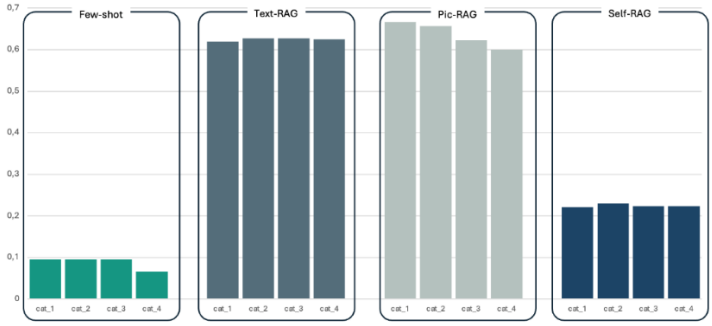

Fig. 4. Retrieval Accuracy of the top k = 4 retrieved images

4.4 Text Similarity and Localisation

Cosine Similarity rises from 0.825 (Zero-Shot) to 0.929 (Few-Shot), with Pic-RAG reaching 0.959. Pic-RAG achieves a BLEU of 0.600, partly inflated by the 50% defect-free images whose retrieved examples share identical descriptions. For localisation, the Fixed-Few-Shot baseline achieves the best result (189.2 px average distance), outperforming all RAG variants. A dedicated red-dot experiment confirmed that poor localisation reflects a fundamental VLM limitation in spatial reasoning rather than failure to detect defects, consistent with findings by [4].

4.5 Example Influence and Self-RAG Analysis

A further experiment investigated whether the VLM genuinely "understands" defects or merely reproduces seen patterns. Three additional Fixed-Few-Shot runs used only examples from a single defect category. When only Tooth Clipping examples are shown, TC is predicted 196 times versus 108 in the ground truth (+81.5%), while Dirt drops to 36 (−66.7%). This demonstrates strong priming by provided examples. To explain why Self-RAG outperforms Text-RAG despite similar architectures, a hybrid experiment fed Self-RAG's initial description into the Text-RAG retriever. Text-based retrieval repeatedly selected the same images, predominantly TCSSD gear wheels whose descriptions textually match the widest range of queries. Self-RAG's LLM-based ranker distributed selections more evenly, providing a healthier mix of matching and contrasting categories. This diversity appears to give the VLM greater "confidence" in distinguishing defect types.

5 Discussion

For binary defect detection, the VLM is already remarkably capable without examples, the visual deviation of a damaged gear wheel from its normal appearance is apparently salient enough for the model's pretrained representations. The jump from Zero-Shot to Few-Shot (+13.1%) is substantial, but further RAG-based gains are marginal (<1%). For a simple "pass/fail" gate, a lightweight Few-Shot setup suffices. For defect classification, RAG delivers meaningful value. The Self-RAG's 16.5% improvement demonstrates that dynamically selected, contextually relevant examples help the VLM disambiguate between similar defect types. Crucially, example diversity, not merely retrieval accuracy, is the key driver. Pic-RAG retrieves more category-matched images than Self-RAG, yet Self-RAG achieves superior classification. A balanced mix of matching and non-matching categories provides the model with richer decision boundaries. The failure of Text-based RAG is instructive. Because the initial Zero-Shot description often mentions multiple defect types (including hallucinated ones), cosine retrieval converges on descriptions containing the same broad vocabulary, typically TCSSD images. The resulting examples are textually repetitive yet visually heterogeneous (standard vs. blurred vs. angled variations), creating a text–image mismatch that confuses the model. From a practical standpoint, cost matters: Zero-Shot costs ~$0.01/image, Few-Shot ~$0.03, and Self-RAG ~$0.30 due to its multi-stage pipeline. For high-throughput remanufacturing, the Fixed-Few-Shot approach offers the best cost–performance trade-off. Self-RAG is better suited for small-batch, high-value components where classification precision justifies the overhead.

5.1 Limitations

The dataset originates from only 15 physical gear wheels with three artificially created defect types. Real remanufacturing defects and multi-object generalisability remain untested. The one-hot encoding of Tooth Clipping, counting any number of TCs as a single positive, simplifies evaluation and may overstate accuracy. The Self-RAG variant relies on the LLM itself for ranking, introducing circular reasoning risks. All experiments use GPT-4o; model-specific biases may not generalise.

6 Conclusion and Future Work

This study has systematically compared five RAG architectures for VLM-based defect detection in remanufacturing. The key findings are: (1) VLMs achieve strong binary defect detection (95.1%) with just four fixed examples. (2) For exact defect classification, Self-RAG achieves 73.5% Exact Match, outperforming the Few-Shot baseline by 16.5%. (3) Example diversity, not retrieval accuracy alone, is the key driver of classification improvement. (4) VLMs are strongly primed by provided examples, suggesting pattern reproduction rather than deep visual understanding. For remanufacturing practice, the Fixed-Few-Shot approach is the most resource-efficient option, requiring only four annotated reference images and no retrieval infrastructure. Where higher

classification accuracy is needed, Self-RAG offers the best results at increased cost. Future work should explore multi-agent architectures with defect-type-specific expert models and develop trained scorer modules to replace LLM-based self-ranking. Integrating the present approach with synthetic data generation [5], semantic 3D models [18], and adaptive view planning [6] presents promising opportunities for end-to-end automated inspection pipelines.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1574 - Project number 471687386.

References

[1] H. Schandl *et al.*, "Global material flows and resource productivity: The 2024 update," *J. Ind. Ecol.*, vol. 28, no. 6, pp. 2012–2031, Dec. 2024, doi: 10.1111/jiec.13593.

[2] T. Tolio *et al.*, "Design, management and control of demanufacturing and remanufacturing systems," *CIRP Ann.*, vol. 66, no. 2, pp. 585–609, 2017, doi: 10.1016/j.cirp.2017.05.001.

[3] C. Nwankpa, S. Eze, W. Ijomah, A. Gachagan, and S. Marshall, "Achieving remanufacturing inspection using deep learning," *J. Remanufacturing*, vol. 11, no. 2, pp. 89–105, Jul. 2021, doi: 10.1007/s13243-020-00093-9.

[4] A. A. Tulbure, D. P. Danciu, E. H. Dulf, and A. A. Tulbure, "A study on multimodal LLM reasoning for defect detection," in *2024 IEEE 30th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, Sibiu, Romania: IEEE, Oct. 2024, pp. 153–157. doi: 10.1109/SIITME63973.2024.10814831.

[5] D. Koch, F. Stamer, and G. Lanza, "Conceptual Framework for Synthetic Data Generation in Remanufacturing," in *Production at the Leading Edge of Technology*, W.-G. Drossel, S. Ihlenfeldt, and M. Dix, Eds., in Lecture Notes in Production Engineering. , Cham: Springer Nature Switzerland, 2025, pp. 317–325. doi: 10.1007/978-3-031-86893-1_35.

[6] D. Koch, J.-P. Kaiser, F. Stamer, R. Stark, and G. Lanza, "Enhancing Visual Inspection in Remanufacturing: A Reinforcement Learning Approach with Integrated Robot Simulation," *Procedia CIRP*, vol. 134, pp. 939–944, Jan. 2025, doi: 10.1016/j.procir.2025.02.228.

[7] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in neural information processing systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[8] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of the 34th international conference on neural information*

processing systems, in Nips '20. Vancouver, BC, Canada: Curran Associates Inc., 2020.

[9] S. N. Wadekar, A. Chaurasia, A. Chadha, and E. Culurciello, "The Evolution of Multimodal Model Architectures," 2024, *arXiv*. doi: 10.48550/ARXIV.2405.17927.

[10] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 27, 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[11] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 3, pp. 1932–1940, Mar. 2024, doi: 10.1609/aaai.v38i3.27963.

[12] X. Jiang *et al.*, "MMAD: The Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection," presented at the The Thirteenth International Conference on Learning Representations, 2025. [Online]. Available: https://openreview.net/forum?id=JDiER86r8v

[13] K. Naghavi Khanghah *et al.*, "Multimodal Rag-Driven Anomaly Detection and Classification in Laser Powder Bed Fusion Using Large Language Models," in *Volume 3A: 51st Design Automation Conference (DAC)*, Anaheim, California, USA: American Society of Mechanical Engineers, Aug. 2025, p. V03AT03A041. doi: 10.1115/DETC2025-168615.

[14] Q. Fang, G. Xiong, F. Wang, Z. Shen, X. Dong, and F.-Y. Wang, "Large Language Models as Few-Shot Defect Detectors for Additive Manufacturing," in *2024 China Automation Congress (CAC)*, Qingdao, China: IEEE, Nov. 2024, pp. 6900–6905. doi: 10.1109/CAC63892.2024.10865554.

[15] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," Oct. 17, 2023, *arXiv*: arXiv:2310.11511. doi: 10.48550/arXiv.2310.11511.

[16] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards Mitigating LLM Hallucination via Self Reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics, 2023, pp. 1827–1843. doi: 10.18653/v1/2023.findings-emnlp.123.

[17] J.-P. Kaiser, J. Gäbele, D. Koch, J. Schmid, F. Stamer, and G. Lanza, "Adaptive acquisition planning for visual inspection in remanufacturing using reinforcement learning," *J. Intell. Manuf.*, vol. 36, no. 7, pp. 4867–4893, Oct. 2025, doi: 10.1007/s10845-024-02478-0.

[18] J.-P. Kaiser, D. Koch, F. Stamer, J. Peeters, and G. Lanza, "Semantic 3D product modelling for automated inspection in remanufacturing processes," *J. Remanufacturing*, vol. 16, no. 1, p. 1, Apr. 2026, doi: 10.1007/s13243-025-00157-8.

[19] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th international conference on neural information processing systems*, in Nips '22. New Orleans, LA, USA: Curran Associates Inc., 2022.