Dean&Francis
ISSN 2959-6157

Intelligent Data Annotation Based on Generative Artificial Intelligence: Techniques, Analysis, and Future Opportunities

Yanzi Guo[1, *]

[1]Department of Intelligent Science and Technology, Shanghai Lixin University of Accounting and Finance, Shanghai, 201209, China
*Corresponding author: yanzi521456@gmail.com

Abstract:

The explosive development of AI is strongly dependent on the availability of large-scale, high-quality annotated datasets. However, manual labeling is becoming increasingly unsustainable because of high costs and limited scalability, which has led to the incorporation of Generative AI (e. g. LLMs and LMMs) to automate and augment data engineering workflows.This review provides a systematic analysis of this shift, identifying three primary methodologies based on their target application scenarios: Generation-Annotation Integration, Understanding-Annotation, and Interaction-Annotation Augmentation. The paper systematically curates the literature across several domains, including Remote sensing, clinical psychology, and creative arts, to summarize current status on the "Quality-Efficiency-Credibility" triad of generative annotation. Significant cost savings and increases in efficiency are possible with these new methodologies, while associated drawbacks and limitations include Model hallucinations, domain knowledge gaps, and lack of standardized evaluation metrics for generative annotation processes. Finally, the paper proposes a research roadmap to address these fundamental problems with an emphasis on: human-AI collaboration ecosystems, multi-agent architectures, and privacy-preserving local inference. The aim of this work is to create a foundational perspective to develop the next generation of intelligent, trustworthy, and scalable data annotation frameworks.

Keywords: Generative AI; Data annotation; Large language models (LLMs); Human-AI collaboration; Synthetic data.

ISSN 2959-6157

1. Introduction

The growth of artificial intelligence is arguably reliant on large, high-quality datasets. Since the performance of supervised learning models depends directly on the accuracy and fidelity of labels, manual labeling has always been a huge obstacle to developing successful models. This is particularly true in specialized domains, e.g., medical imaging where expert opinion is costly, or in low-resource and privacy-sensitive applications where quantities are limited. Scalability, therefore, becomes a limiting factor in domains ranging from autonomous driving to computational linguistics.

The data annotation process is the foundational component of the machine learning pipeline in many applications, yet it is also one of the most time-consuming and expensive steps in the pipeline. To alleviate this bottleneck, the data annotation workflow including Extractor designing and labeling tasks needs to be automated, simplified, and enhanced. This paper demonstrates how the data annotation process is moving into a new Age with generational artificial intelligence (gen) powered by Large Language Models (LLMs) and Large Multimodal Models (LMMs). These models, possessing deep semantic understanding and reasoning capabilities, can radically transform the data annotation landscape by automating labor-intensive annotation tasks and boosting throughput. In simple terms, this new era of "smart annotation" can not only assign labels but also generate data for edge cases and fuse semantic context. By alleviating human weaknesses such as fatigue and inconsistency, gen can contribute to high-quality results with high scale.

This new discipline has now led to a raft of methodological innovations, including the use of LLMs for text categorization, diffusion models for segmentation, and human-AI collaboration. Empirical results have been equally provocative: GenAI is catching up to the skill level of fellow-human experts in clinical extraction, and its performance on multimodal tasks is strong. Nevertheless, bringing this technology into the real world has progressed faster than quality assessment, which often has a lack of criteria capturing thematic challenges as the factual correctness of LLM-generated labels, or synthetic image novelty. Bridging this gap in assessment is a priority for future progress measurement and scientific maturity.

Apart from that, the state of work today remains rather fragmented and siloed. Specific modalities, limited task definition, or a specific application area are often singled out, neither presenting a consistent framework for systematically classify technical methods, critically evaluate evaluation rigor, and clearly lay out applicability and limitations. This fragmentation is an evident one: as examples, recent works in assorted domains indicate some interesting developments: for instance, satellite image annotation, studies have been done to embed ChatGPT in annotation workflows for comparing of labels quality and costs between human and automated annotators [1]. It has been applied to biomedical text mining, with a cache-augmented generation approach using GPT - 4o combined with PubTator - 3.0 presented to support annotation on datasets of biomedical entities extracted in scientific articles [2]. In the field of IT service management, there is Work on support ticket classification that compares human labels applied on a set of tickets with an automated approach combining active learning and weak supervision [3].

Technology such as deep learning has demonstrated great capacities in data annotation endeavors across various domains and segments. Despite multiple advancements, GenAI related research remains fractured in silos without interoperability, common representations, or methodological consensus. systematic review of the coexistence of GenAI and data annotation is therefore needed to comprehend the current state and identify promising future opportunities. Inherited from this, this paper provides an overview of the existing landscape, proposing a novel typology that distinguishes three major annotation paradigms, i. e., direct generative annotation, human-in-the loop, and synthetic data generation. The paper retroactively summarizes associated technologies across these three categories, emphasizing on application contexts and technological advances while reviewing peer exercise procedures. More importantly, the paper also highlighted the need for automated but human-as-priors metrics regarding implementation cost and scalability properties. In conclusion, the paper delineates missing research steps and point out key challenges, including model hallucination, domain adaptation, and Ethical aspects. By describing technical advancements and emphasizing the need for standardized benchmarks, this paper aims to assist scalable annotation frameworks for efficient GenAI-assisted artificial intelligence research.

2. Introduction to mainstream Intelligent Annotation Technologies

Automated annotation is one of the main fields for the advancement that Generative AI application offers." This chapter will see 3 leading technical paradigms. Based on how generative AI approaches from the analysis engineering workflow, the 3 basic concepts, functional approach and characteristics that combine their specific research examples are introduced.

2.1 Generation-Annotation Integration Paradigm: LLM-based Low-Resource Sequence Labeling Enhancement

This work extends the concept of Generation-Label Integration where large generative models reuse extremely data-hungry Neural models and generate Synthetic Data on-the-fly, precisely labeled, defeating Data paucity in domain-specific domains on-the-fly, which tightly couples Data generation and Annotation.

Taking the Know FREE framework targeting low-resource Named Entity Recognition (NER) tasks as an example, this method uses LLMs as a data augmentation engine. Its basic process includes two key steps: Label Extension Annotation and Enriched Explanation Synthesis. The former uses LLMs to generate new sentences containing target entities, while the latter requires LLMs to output detailed entity boundaries, types, and contextual explanations to guide downstream models in accurately understanding annotation rules and entity definitions [4].

From the perspective of structural process analysis, this framework is used to trigger the LLM knowledge, and via well-crafted reminders, allows the model to imitate the language distribution and annotation behavior of the intended domain. In turn, the generated information is merged with native information to train or fine-tune relevant domain-dependent NER models [4].

However, this mode has profound benefits by producing novel data with endogenous perfect labeling and circumventing inherent subjectiveness and inconsistency in conventional manual annotation [4] while significantly overcoming the physical Constraints of data harvesting. However, it can be limited in several ways: Synthetic data may contain inherent semantic distribution biases [4]; and generated samples may manifest low generalization in low-resource domains when let lacks domain knowledge [4]. Additionally, with multimodal genres lacking sufficient contextual constraints, the chance of hallucination, where how the model learns is misaligned and leads to incoherent production, can arise, as at image generation tasks when models hallucinate a saddle beside a horse for a time machine [4].

2.2 Understanding-Annotation Automation Paradigm: LLM as „Crowdworkers“ for Contextual Information Generation

The core of the Understanding-Annotation Automation paradigm lies in treating LLMs as Automated Crowd workers with advanced reasoning and understanding capabilities to batch analyze, predict, and mark existing unlabeled real data [5].

Taking the research on evaluating the potential of LLMs in Contextual Information Generation (CIG) tasks as an example, the study aims to measure the quality and efficiency of large models in performing complex pragmatic annotation tasks. Its structural process mainly relies on the zero-shot or few-shot reasoning capabilities of LLMs. By inputting premise sentences and instructions, the LLM automatically outputs coherent contextual information, which is finally evaluated for quality by human language experts [5].

The paper refers to it from the point of view of the structural process analysis to describe how carefully constructed prompts enable models to avoid the traditional model training or fine-tunning phases and directly conduct efficient large-scale labels over data [5].

Of very high efficiency and great cost reduction, experimental results indicated LLMs generate context solely milliseconds, which saves gigantic time costs with respect to manual annotator effort and achieves comparable quality to human experts for some pragmatic applications [5]. To gain even more robustness, a lot of work has been done to apply an ensemble of LLMs, whose combination results (e. g. through weighted consistency scoring) combine reasoning results from different models; in this way, LLMs\' specific biases are suppressed faster, and resulting outputs are more similar with respect to humans [5]. However, the same is limited because the performance of models is more sensitive to their input prompts and “hallucinated” content remains a task to be tackled [5], which means efficient post-processing and quality checks still need to be designed to ensure the reliability of automated annotation results.

2.3 Interaction-Annotation Augmentation Paradigm: Refined Annotation under Human-LMM Collaboration Framework

The Interaction-Annotation Augmentation paradigm relies on Human-AI Collaboration [6], where generative AI is the intelligent helper, enhancing the Human annotations efforts, accuracy, and experience. It revolves around the potential to leverage human perception and decision-making skills (such as target selection) and Ai generation and knowledge strengths (such as fine label generation) [6].

As a case study, image annotation research conducted under Human-LMM depicts an example of applications of human cognition with low-level repetitive work.

On the topic of human machine labor, for the first step, targets selection in pictures is regarded by humans. Here, they select areas of interest by different methods, like Bounding Box [6]. In the second step, LMMs receive the information sent from human selection of the image

ISSN 2959-6157

and then automatically analyses its content to create the fine-grained labels [6]. In the final step, the whole system starts to take into consideration the intelligent feedback to check or offer recommendations in time after the LMM has created the fine-grained labels [6].

Process Analysis shows that this paradigm greatly lowers Cognitive burden on human annotators, as it delegates Naming tasks with high knowledge demands to LMMs Therefore, in this Sense, it accomplishes Bidirectional Alignment [6], which means efficient conjunction of human intention and model knowledge. In addition, for Text-to-image generation, a perfect balance between high Quality and Diversity can be sought via a mixed fine-tuning approach of Human-in-the-loop (HITL) and Expert-in-the-Loop (EITL) [7].

It also greatly improves annotation quality and credibility. Because LMMs can offer high fidelity, coarse-grained labels for each image—a tiger, not just a cat—or even a particular breed of sheep or cow, this can make downstream applications perform much better than with more generic labels [6]. From the perspective of augmentation, when models are trained with mixed fine-tuning approaches, they are better able to adapt to particular stylistic preferences and requests for ALIGNments [7], although this approach has drawbacks. It's not easy to design and maintain this kind of system [7]. Building efficient interaction interfaces and effective real-time response mechanisms can be challenging.

3. Literature ReferencesEvaluation System and Application Analysis of Generative Annotation

Analysis of the technology of generative annotation is derived from heterogeneous data sources, ranging from publicly-recent geospatial data to private clinical text. Within the realms of computer vision and earth observation, studies have hinged on high complexity geographically-originating satellite imagery provided via GeoBasisDaten/BKG, necessitating domain knowledge to discern fine-grain land-use classification between, say, commercial and brownfield to serve as a reference for specialized visual classification purposes [1]. In the realms of natural language processing and psychology, on the other hand, there is research centered around using pretext response datasets obtained in clinical studies. An advantage of using such unstructured forms of data—contrary to highly-structured, structured databases—is the richness of the qualitative contents, though they also come with stringent privacy and ethical compliance considerations (HIPAA for example) and hence the need to deploy locally-operable, open-source language models for Annotation, to avoid introducing the inherent threats of uploading such data into closed/source cloud-based models [8].

3.1 Datasets: Sources, Characteristics, and Processing

To showcase generative annotation's adaptability, the study also utilizes several datasets that exhibit different data styles and domain-specific concerns. When exploiting sensor data and targeting the domain of computer vision and earth observation, researchers choose satellite imagery provided by GeoBasisDaten/BKG. Being complex earth observation data, this dataset is difficult to classify manually due to the necessity for specific domain expertise. Thus, it will serve as a challenge for the model to cope with specialized visual tasks [1]. To achieve efficient automated annotations, the experiments must establish benchmarks for annotation performances between annotators who have different biases, backgrounds, and specialties, including machine learning engineers, lay persons without any profession affiliation, and Chat GOOT.

Alternatively, in the context of privacy-sensitive domains, the study leverages publicly available free-text response datasets from psychology. In contrast to structured data, this information contains no fixed representations but is qualitatively richer and subject to stringent personal privacy and ethical standards (such as HIPTAA). Such constraints make it necessary to employ locally deployable open-source languages model ensembles to provide data-security and rule-based compliance enforcement, since data cannot be exchanged with an external model provider [8].

3.2 Evaluation Metrics: A Trinity Evaluation Perspective

Generalizing generative annotation needs a multidimensional approach instead of only focusing on accuracy to establish a trinity "Quality-Efficiency-Credibility". Currently, in the community, a general set of metrics is followed, where Annotation Quality use typical supervisory learns measures like F1-Score, Accuracy and Precision were introduced, in order to measure the correctness of annotations obtained from the expert ground truth. In terms of testing reliability and stability of the annotation, metrics such as Fleiss' Kappa and Gwet's AC1 are following, where to our knowledge, Gwet's AC1 wins because of robustness in the treatment of datasets with unbalanced label distribution or rare classes to quantify the extent of agreement amongst models and between models and human [5,8].

In open-ended text generation tasks where ground truth may be harder to specify, researchers use Metrics that are Reference-Free such as Contextual Approsiability (CATS) to evaluate the semantic quality, relevance, and reads of the Generated text using theories of language, for example, without the requirement of a strict ground truth [5].

Lastly, System Efficiency values the functionality of the technology in real-world settings. System Efficiency values, relative to the cost of real labor, how much the technology saves organizations on both time costs and expenditure, which is probably the most economical way to evaluate the viability of automatic annotation systems [1].

3.3 Experimental Results Analysis and Application

By synthesizing the research from these various studies, a comprehensive picture emerges of generative annotation's performance nuances and economic benefits. Efficiency and Cost-Effectiveness emerge as key considerations, with satellite image classification tasks showing that costs associated with using models like ChatGPT-4 (around $0.01/item) are dramatically less than those associated with expert human annotators ($0.55/item). Moreover, automated systems can provide text generation in seconds, illuminating their potential for large-scale data creation [1,5].

In terms of Quality and Domain Knowledge, a "Middle-Ground" effect may hold. While they cannot yet compete with experienced expert labels on sophisticated, knowledge-intensive tasks like F1, they are far better than non-professional annotators, which confirms their potential as useful high-Quality baselines to bridge the costly expertise and erratic crowdcuration [1].

Methodological strategies are also pivotal for the success. Prompt Engineering like long prompts with priority instructions has been found to significantly improve the accuracy in identifying similar classes. Moreover, Ensemble Strategies comprising predictions from multiple local models has been noted as an effective way to increase robustness and credibility in areas of highly sensitive data annotation, when it seems dangerous to use a single model to make predictions [1,8].

Summoning up, generative annotation technology is moving from theory to reality and has successfully integrated the reasoning and generation strength of LLMs and LMMs into data pipelines. Further enhancement in terms of domain adaptability improvement of the models, the development of more robust reference-free evaluation metrics and more efficient and secure "human-AI collaboration" mechanisms remains to follow.

4.Discussion on Challenges and Future Prospects

4.1 Challenges Regarding Reliability and Domain Specificity

Despite their proven efficiency, the application of Gen-AI in annotation tasks encounters critical reliability challenges. One of the initial issues consists of the hallucination and semantic alignment dilemma, where the models start to produce believable yet incorrect or unrelated responses. An example is found within the realm of gamified text labeling tasks, where models have been reported to incorrectly interpret the context of a term because of a complete absence of narrative limitations, resulting in inaccurate guesses that a saddle refers to a horse whenever mentioning a time machine [9]. The realm of sticker image generation follows a similar trend, with standard models struggling to synchronize their creativity with specific stylistic or culture influences without significant human-in-the-looper ship fine-tuning [7]. Furthermore, there is a competence gap that persists even in domain-stacked sectors. Although LLMs have shown potential in broad classification issues, they encounter difficulties when confronted with the inherent semantic distribution biases present in low-resource domains such as named entity recognition, necessitating complex knowledge fusion structures to compete with the standards set by domain experts [4]. This discrepancy is particularly critical in settings where mistakes are costly, such as in satellite imagery assessment or biomedical studies where domain knowledge requirements are non-negotiable [1,2]. Lastly, the necessity to work with closed-source commercial models poses a barrier for privacy-sensitive tasks, as evidenced in clinical psychology research, where data privacy regulations drive models to rely on local, and often lower-performing, open-source(model) ensembles [8].

4.2 Future Directions Towards Agentic and Collaborative Workflows

To address the limitations, the domain is evolving in the direction of more advanced, agentic and collaborative frameworks. In future research, there will be growing interest in Multi-Agent Systems, where specialized AI agents collaborate to deal with different facets of the problem at hand. 3 of 8 The Meta Designer framework epitomizes such a trend, using separate agents for layout, texture, and user feedback in the synthesis of complex artistic typography to hint that future annotation systems will be more likely to be modular rather than monoidal [10]. Also, there is evolving from mere automation to deep Human-Ai Collaboration. New structured frameworks are emerging in which humans manage higher-level cognitive activities (such as object selection in the picture)

ISSN 2959-6157

whereas AI does most of the grunt work in generating the labels and reducing humans' cognitive burden and fatigue [6]. 5 Style Tailoring 6 also emphasize the significance of directly incorporating human feedback loops into the fine-tuning mechanism to retain variety in the output and ensure that they align with humans' tastes and preferences [7]. In addition, in order to streamline and cut costs in the world of science, approaches such as Cache-Augmented Generation are growing to catch and reuse valid annotations, hence greatly lowering the computation overhead for repetitive tasks [2]. Generally, the incorporation of all these advanced workflows—multi-agent frameworks, human collaboration, and retrieval-automated techniques—will eventually be required to develop resilient, effective, and dependable annotation systems.

5. Conclusion

This review offers an exhaustive examination of the revolutionary influence of Generational Artificial Intelligence (GenAI) within the data annotation industry. By systematically assessing ten prominent state-of-art studies, the paper catalogs the technological evolution from basic automation to advanced systems able to parse intricate modalities, ranging from satellite imagery, biomedical data, and creative visual content.

GenAI has also fundamentally altered the economics and scalability of data generation, as evidenced by multiple application realms; this technology has proven to cut annotation costs by hundreds of times while dramatically enhancing workflow throughput. In niche sectors such as support ticket classification and biomedical entity identification, automated solutions have not only outpaced manual labor but also emerged as powerful assistants that amplify human annotators' efficiency. Moreover, recent developments such as KnowFREE for low-resource contexts and ensemble methods for privacy-preserving text labeling exemplify GenAI's adaptability in resolving challenges that conventional methodologies routinely fail.

However, this revolution is not without complexities. While GenAI brings unprecedented efficiency, its trustworthiness hinges on the model architecture. Robust deployments now rely on hybrid Human-in-the-loop (HITL) techniques, where human intuition acts as a counterbalance to reduce hallucinations and style inconsistencies. Interactive, multi-agent systems holding the possibility to dynamically control software behavior through persistent interaction represent the future of annotation.

As the Generative Annotation field matures, it is transforming into a unified subfield of data engineering - moving from simple label prediction to synthetic data generation, knowledge distillation, and collaborative intelligence. With the maturation of evaluation metrics such as CATS to provide reference-free quality assessment, GenAI stands to become the cornerstone of modern AI development, retooling data annotation from a logistical obstacle into an endless source of fuel for the upcoming wave of scientific innovation.

5.1 Limitations and Future Work

Though the contribution of each study, this review has few limitations. Firstly, given the fast-paced nature of AI research, the ten included studies are but an interim snapshot of a rapidly evolving field, and newer frameworks may be available since this collation.

Secondly, this review focuses on the positive aspects of GenAI, albeit without fully discussing the potential danger of "model collapse," which can occur when future AI models are predominantly trained on synthetic data. Future research should investigate longitudinal studies on long-term high-quality evaluation of AI-generated labels, as well as legal implications surrounding data ownership and privacy in the age of generative engineering.

References

[1] Beck J, Kemeter L M, Dürrbeck K, AbdallaM H I and Kreuter F, Toward Integrating ChatGPT Into Satellite Image Annotation Workflows: A Comparison of Label Quality and Costs of Human and Automated Annotators, in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 4366-4381, 2025.

[2] Giuliani C, Benadi G, Engel F, Werner J, Watter M, Schwarzer G, Groß O, Zeiser R, Binder H, Kaier K. Identifying Biomedical Entities for Datasets in Scientific Articles: 4-Step Cache-Augmented Generation Approach Using GPT-4o and PubTator 3.0. JMIR Form Res. 2025 Nov 20;9:e73822. doi: 10.2196/73822. PMID: 41264807; PMCID: PMC12633840.

[3] Fuchs S, Schnellbach J, Wittges H, Krcmar H. Human vs. Automated data annotation: Labeling the data set for an ML-driven support ticket classifier. Data & Knowledge Engineering, 2026, 162: 102534.

[4] Lai P, Gan J, Ye F, Wang Y, Cui B. Improving Low-Resource Sequence Labeling with Knowledge Fusion and Contextual Label Explanations. ArXiv, 2025, abs/2501.19093.

[5] Martínez-Murillo I, Miró Maestre M, Suárez A, Lloret E, Moreda P. Assessing the potential of LLMs as crowdworkers for contextual information generation. Information Processing & Management, 2026, 63(3): 104486.

[6] Zhang H, Fu X, Carroll J M. Augmenting Image Annotation: A Human-LMM Collaborative Framework for Efficient Object Selection and Label Generation. ArXiv, 2025, abs/2503.11096.

[7] Sinha A, Sun B, Kalia A, Casanova A, Blanchard E, Yan D, Zhang W, Chen J, Shah H, Yu L, Singh M K, Ramchandani A, Sanjabi M, Gupta S, Bearman A, Mahajan D. Text-to-Sticker: Style Tailoring Latent Diffusion Models for Human Expression. European Conference on Computer Vision, 2023.
[8] Qiu J, Guo D, Papini N, Peace N, Fitterman-Harris HF, Levinson CA, Hartvigsen T, Henry TR. Labeling Free-text Data using Language Model Ensembles. 2025.
[9] Althani F, Madge C, Poesio M. Using In-context Learning to Automate AI Image Generation for a Gamified Text Labelling Task. In Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024, Torino, Italia: ELRA and ICCL; 2024:21-31.
[10] He J, Cheng Z, Li C, Sun J, He Q, Xiang W, Chen H, Lan J, Lin X, Zhu K, Luo B, Geng Y, Xie X, Hauptmann A G. MetaDesigner: Advancing Artistic Typography through AI-Driven, User-Centric, and Multilingual WordArt Synthesis. ArXiv, 2024, abs/2406.19859.