future internet

MDPI

Review

Transforming Data Annotation with AI Agents: A Review of Architectures, Reasoning, Applications, and Impact

Md Monjurul Karim [1], Sangeen Khan [1], Dong Hoang Van [1], Xinyue Liu [2], Chunhui Wang [3] and Qiang Qu [1,*]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; karim@siat.ac.cn (M.M.K.); sangeenkhan2662@gmail.com (S.K.); hoang@siat.ac.cn (D.H.V.)

[2] School of Management, South-Central Minzu University, Wuhan 430074, China; 202221075032@mail.scuec.edu.cn

[3] School of Cyber Science and Technology, Zhejiang University, Hangzhou 310027, China; cctvlaw@163.com

* Correspondence: qiang@siat.ac.cn

Abstract

Data annotation serves as a critical foundation for artificial intelligence (AI) and machine learning (ML). Recently, AI agents powered by large language models (LLMs) have emerged as effective solutions to longstanding challenges in data annotation, such as scalability, consistency, cost, and limitations in domain expertise. These agents facilitate intelligent automation and adaptive decision-making, thereby enhancing the efficiency and reliability of annotation workflows across various fields. Despite the growing interest in this area, a systematic understanding of the role and capabilities of AI agents in annotation is still underexplored. This paper seeks to fill that gap by providing a comprehensive review of how LLM-driven agents support advanced reasoning strategies, adaptive learning, and collaborative annotation efforts. We analyze agent architectures, integration patterns within workflows, and evaluation methods, along with real-world applications in sectors such as healthcare, finance, technology, and media. Furthermore, we evaluate current tools and platforms that support agent-based annotation, addressing key challenges such as quality assurance, bias mitigation, transparency, and scalability. Lastly, we outline future research directions, highlighting the importance of federated learning, cross-modal reasoning, and responsible system design to advance the development of next-generation annotation ecosystems.

Keywords: data annotation; AI agents; large language models; adaptive annotation; responsible AI

check for updates

Academic Editor: Gianluigi Ferrari

Received: 27 June 2025
Revised: 25 July 2025
Accepted: 29 July 2025
Published: 2 August 2025

Citation: Karim, M.M.; Khan, S.; Van, D.H.; Liu, X.; Wang, C.; Qu, Q. Transforming Data Annotation with AI Agents: A Review of Architectures, Reasoning, Applications, and Impact. *Future Internet* **2025**, *17*, 353. https://doi.org/10.3390/fi17080353

1. Introduction

The global AI market is projected to reach USD 1.8 trillion by 2030 with data annotation comprising a significant USD 8.2 billion (https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market, accessed on 20 June 2025). Data annotation is the systematic process of labeling raw, unstructured data to provide meaningful context for AI and ML algorithms (e.g., image labeling, sentiment analysis, or speech recognition). As a result, it has become a critical bottleneck in determining the success of AI systems across diverse domains [1]. High-quality annotated datasets serve as the cornerstone for supervised learning paradigms, enabling algorithms to discern complex patterns and generate reliable predictions in applications ranging from autonomous vehicle perception to medical diagnosis and natural language understanding [2]. LLMs require trillions of training tokens and reflect the exponential growth in AI model complexity while driving an increase in the demand for annotated data by several orders of magnitude [3]. In specialized domains such as communication systems, financial markets, and healthcare, annotated data becomes indispensable for critical tasks including sentiment analysis of user interactions, anomaly detection in network traffic, automated medical coding, and behavioral pattern recognition. This fundamental dependence on high-quality labeled data across an expanding range of AI applications highlights the urgent need for innovative approaches to annotation methodology that can scale with the rapidly evolving demands of modern AI.

Despite its pivotal role in AI development, traditional data annotation remains plagued by fundamental limitations that create significant bottlenecks in the machine learning pipeline. Manual annotation processes are inherently expensive and time-intensive, with costs often exceeding 25 percent of total machine learning project budgets and timelines stretching from months to years for complex datasets [4]. Quality consistency presents another critical challenge, as inter-annotator agreement rates frequently fall below 70 percent for subjective tasks, particularly in domains requiring nuanced judgment such as sentiment analysis, content moderation, and clinical assessment [5]. The scalability crisis becomes acute when considering that state-of-the-art models like GPT-4 require datasets containing hundreds of billions of tokens, while human annotators can typically process only thousands of instances per day [6]. Domain expertise requirements further exacerbate these challenges, with specialized fields like medical imaging or legal document analysis demanding annotators with years of training, significantly limiting the available workforce and inflating costs [7]. Additionally, cognitive biases, fatigue effects, and varying interpretation frameworks among human annotators introduce systematic inconsistencies that compromise dataset quality. These multifaceted challenges, encompassing cost escalation, temporal constraints, scalability limitations, quality variability, and expertise scarcity, collectively constitute a fundamental impediment to AI advancement, necessitating transformative solutions that can address the annotation crisis at scale.

The emergence of AI agents, particularly those enhanced by recent breakthroughs in LLMs [8,9], represents a paradigmatic shift toward addressing these annotation challenges through intelligent automation [10]. AI agents are autonomous computational entities characterized by four fundamental capabilities: autonomy (independent operation without continuous human oversight), reactivity (responsive adaptation to environmental changes), proactivity (goal-oriented initiative-taking behavior), and social ability (collaborative interaction with humans and other agents) [11].

Current AI agents exhibit emergent properties such as self-reflection, wherein agents can assess their past decisions and refine future actions based on previous feedback. For instance, Agent-R [12] is a framework that enables language agents to reflect on erroneous actions and recover correct trajectories using Monte Carlo Tree Search (MCTS). Similarly, LLM agents significantly improve problem-solving performance by reflecting on their mistakes and adjusting their strategies accordingly [13]. On the other hand, collaborative reasoning enables AI agents to jointly solve complex tasks and enhance decision-making through coordinated interactions. For example, AgentsNet [14] evaluates the ability of multi-agent systems to collaboratively form strategies for problem-solving and self-organization. Additionally, the AutoGen [15] enables AI agents to collaborate with distinct roles to solve problems more effectively. These capabilities significantly enhance the efficiency and reliability of annotation workflows. However, while such capabilities have already been implemented in some systems, there remains considerable room for further development. Theoretical advancements suggest that AI agents' self-reflection and collaborative reasoning could be further refined to autonomously adjust strategies based on more complex feedback loops and collaborative dynamics [16]. These developments are expected to transform the future landscape of data annotation systems, wherein AI agents will continuously learn and improve their annotation tasks through dynamic collaboration with both humans and other agents [17].

1.1. Motivation and Contributions

The integration of AI agents into data annotation workflows addresses critical gaps in current methodologies while introducing novel capabilities that extend beyond traditional automation approaches. While the existing literature has explored individual aspects of AI-assisted annotation, such as active learning, weak supervision, and LLM-based labeling, there remains a significant research gap in understanding how autonomous agent architectures can orchestrate comprehensive annotation pipelines [18,19]. Current annotation systems lack the sophisticated reasoning, adaptive decision-making, and collaborative capabilities that AI agents can provide, particularly in complex scenarios requiring multi-step workflows, quality assurance mechanisms, and dynamic guideline interpretation [20]. Moreover, the rapid evolution of LLMs has outpaced the development of systematic frameworks for leveraging their capabilities in annotation contexts, creating an urgent need for comprehensive analysis of agent-driven annotation paradigms.

Our survey addresses key limitations in the existing literature by offering the first in-depth analysis of the transformative role of AI agents in data annotation, bridging theoretical foundations with practical insights. We introduce a novel taxonomic framework for categorizing agents based on their capabilities and analyze how LLM-enhanced agents employ advanced reasoning strategies, such as Chain-of-Thought, Tree-of-Thought, and ReAct, to improve annotation outcomes across modalities. We further examine current trends in agent-driven system design and evaluation, offering insights into their operational efficiency, quality control, and real-world adaptability. In addition, we explore the practical landscape of agent-based tools and highlight ongoing research frontiers that will shape future development in this field. The contributions of our survey are summarized as follows:

- Comprehensive Taxonomy and Technical Analysis: We establish the first comprehensive framework that systematically classifies AI agents in data annotation. This contribution integrates a novel taxonomy based on agent capabilities with a deep analysis of architectural patterns ranging from single-agent systems to multi-agent collaborations to provide unified design principles and implementation guidelines for researchers and practitioners.
- Architectural Design and Evaluation Framework: We systematically examine various architectural approaches for agent-driven annotation systems, including single-agent pipelines, multi-agent collaborations, and human-in-the-loop (HITL) integration, providing practical design guidelines and trade-off considerations. Complementing this, we introduce a holistic evaluation framework with novel metrics and standardized benchmarks to rigorously assess agent performance, economic impact, and quality assurance.
- Real-world Applications and Tools: We present a thorough analysis of the practical ecosystem by examining transformative, real-world applications across diverse industries (e.g., healthcare, finance, and technology). This is paired with a critical assessment of the current landscape of tools, platforms, and frameworks, offering actionable guidance for technology selection and strategic implementation.
- Research Challenges and Future Directions: We systematically identify critical open challenges, including quality assurance, bias mitigation, transparency, privacy, and scalability in AI-agent-driven annotation. Based on this analysis, we outline a forward-

looking research roadmap to guide the future development of robust, reliable, and responsible AI annotation agents.

1.2. Organization

The remainder of this survey is organized as follows. Section 2 reviews related work on AI agents in data annotation, situating our survey within the broader research landscape. Section 3 covers foundational concepts in data annotation, tracing its evolution from manual methods to LLM-driven techniques. Section 4 introduces the core characteristics, classifications, and roles of AI agents in annotation, with an emphasis on LLM-enhanced capabilities. Section 5 outlines our research methodology, including the literature retrieval, selection criteria, and analytical framework. Section 6 discusses how agents enhance annotation workflows through adaptive data selection, quality control, and automation. Section 7 analyzes system architectures, including single-agent, dual-agent, multi-agent, and HITL models. Section 8 presents the evaluation methodologies, covering metrics, economic impact, user experience, and benchmarking. Section 9 highlights real-world applications, tools, and case studies across domains. Section 10 identifies the open research challenges and future directions. Finally, Section 11 concludes with a synthesis of insights and a perspective on the future of agent-driven annotation.

2. Related Work

AI agents powered by LLMs can now draft, refine, and even audit labels with minimal human effort. Chaining reasoning steps and tool calls transforms data annotation from a one-pass manual task into an interactive, quality-controlled workflow. These advances promise faster turnaround and more consistent labels across text, image, and multimodal corpora. We divide the existing and related studies into two major categories: (i) Human-led and Classic-ML Annotation and (ii) Generative-AI/LLM-centric Annotation. The first category focuses on traditional approaches (e.g., expert labeling tools and MTurk-style crowd-sourcing) and generic ML-based approaches (e.g., uncertainty sampling, programmatic labeling, and transfer-learning pipelines). These methods ease cost and scale pain but keep humans in the loop and suffer from noise, cold-start, and domain-shift issues. The second category focuses on generative AI (GenAI) and LLM-based approaches. This section covers GPT-3/4, Llama-2, CLIP, SELF-INSTRUCT, etc. It describes how prompt-engineering, in-context reasoning, and fine-tuning let models produce labels, critique their own outputs, and even synthesize new "informative" examples, while also noting risks of hallucination, cost and explainability gaps.

Existing reviews, however, still cover only fragments of this fast-moving landscape. As Table 1 shows, most surveys concentrate on either data augmentation or modality breadth and overlook agent-based coordination and LLM-centric pipelines.

Table 1. Comparison with the existing literature.

Ref.	Annotation	Multi-agent	LLM	Augmentation	Multimodal
Demrozi et al. [21]	✓	✗	✗	✗	✓
Zhou et al. [22]	✗	✗	✗	✓	✓
Wang et al. [23]	✗	✗	✗	✓	✓
Tan et al. [24]	✓	✗	✓	✓	✗
Xi et al. [25]	✓	✗	✓	✗	✗
Hiniduma et al. [26]	✗	✗	✗	✗	✗
Zha et al. [27]	✗	✗	✗	✗	✓
Liang et al. [28]	✗	✗	✗	✗	✓
Cao et al. [29]	✗	✗	✗	✗	✗
Xu et al. [30]	✗	✗	✗	✓	✗
Ours	✓	✓	✓	✓	✓

2.1. Human-Led and Classic-ML Annotation

The first wave of annotation pipelines still centers on people. Experts using tools such as LabelImg (https://github.com/HumanSignal/labelImg, accessed on 20 June 2025) or Prodigy (https://prodi.gy/, accessed on 20 June 2025) supply the most accurate labels, and crowdsourcing sites like MTurk boost throughput, yet both routes suffer from cost, time, and low agreement; about 30% of crowd labels need rework, and subjective tasks often show agreement below 70%. Classic machine learning helpers lighten the load: active learning selects informative samples, weak supervision converts heuristic rules into labels, and transfer learning lets pre-trained models draft labels in new domains. These methods can cut human effort by half and bootstrap large corpora, but they still require seed labels, produce noisy outputs that must be cleaned, and struggle when the source and target domains differ greatly. In short, the process remains people-centric and capped by cost, noise, and limited scale.

2.2. Generative-AI and LLM-Centric Annotation

Recent generative models and large language models push annotation closer to full automation. Systems built on GPT-3 [31], GPT-4 [32], and Llama [33] now create labels, critique them, and even generate new training examples across text, vision, and multimodal data. Prompt-engineering schemes like SELF INSTRUCT [34] let one model invent both tasks and answers, while agent frameworks add dialogue, planning, and memory so an LLM turns to humans only when rules are unclear, refines labels on the fly, or drives active-learning loops that decide what to annotate next. LLM agents also grow datasets through targeted knowledge distillation and synthetic sample generation, filling domain gaps with minimal human input. These gains bring new risks: hallucinated or biased labels, high computational cost, and opaque reasoning. Robust guardrails and continued human oversight, therefore, remain essential for trustworthy datasets.

2.3. Limitations of the Existing Surveys

AI agents backed by large language models now draft, verify, and correct labels with minimal human effort. When linked in a guided workflow, these agents accelerate annotation, boost consistency, and let human experts focus on hard edge cases. This study is unique because it frames labeling as a closed-loop multi-agent engineering problem. It describes single, dual, and multi-agent workflows that plan, label, and self-audit in one continuous cycle. One study focuses on enlarging corpora through image, text, and paired-data transformations, yet it never addresses who or what produces the labels [22]. Wang et al. [23] cover the full model lifecycle from pre-training through alignment, but annotation quality is listed only as an open challenge. Another study automates labels with a single LLM "oracle", omitting the negotiation and reviewer roles that an agent-centric pipeline would require [24]. Some studies showed that classical machine learning pipelines face similar cost versus quality tensions, yet their solutions remain limited to wearable-sensor data [21]. A recent primer advocates "training-set engineering", yet it offers no guidance on how labels move through the pipeline [27]. Evaluation-focused studies like [29,30] point out benchmark staleness and emergent reasoning, yet they treat failures as downstream diagnostics rather than signals for an agent ensemble to repair its labels.

Existing surveys still overlook several operational and socio-technical gaps that must be bridged before agent pipelines can be deployed safely at scale. First, robustness against adversarial or prompt-injection attacks is rarely covered, even though recent benchmarks such as PromptBench [35] reveal how small perturbations can derail LLM annotators and leak sensitive content [25]. Second, compliance with data-protection regimes is often treated as an afterthought; yet, practical deployments must integrate safeguards like homomorphic encryption and explicit HIPAA/GDPR audit trails to prevent the disclosure of personal data [28]. Third, lifecycle maintenance is missing: contemporary reviews seldom prescribe mechanisms for continuous learning or drift detection, despite calls for HITL administrators that can halt error cascades when label quality degrades over time [26]. Fourth, the human dimension is under-examined; automation bias and job displacement threaten annotator roles, demanding new skills and interface designs that encourage critical oversight rather than rubber-stamping AI suggestions [21]. Fifth, linguistic inclusivity remains unresolved. Some evidence shows that current agent frameworks excel in high-resource languages but struggle to maintain accuracy when ported to low-resource settings [22]. Finally, few surveys quantify the computational and energy costs of privacy-enhanced or security-hardened agents, leaving practitioners without models for balancing throughput, latency, and infrastructure spend [29]. Addressing these blind spots such as robustness, governance, maintenance, labor impact, linguistic equity and cost modeling will be crucial for translating the conceptual promise of agent-based annotation into sustainable industrial practice.

Building on the gaps identified above, this survey makes four specific advances. First, it presents the first complete taxonomy that stitches together manual, classic-ML and generative approaches in one coherent frame, eliminating the fragmentation seen in earlier reviews. Second, it proposes an agent-centric workflow model with planner, annotator and quality checker roles, and validates the model with concrete systems such as Self-Refine and ActiveLLM. Third, it standardizes evaluation by clustering metrics into technical, economic and user viewpoints, enabling a like-for-like comparison of throughput, cost per label and annotator trust across studies. Finally, it maps a future research agenda that targets open questions in ethics, domain transfer and compute efficiency, showing how agent-augmented annotation can tackle cost, scale, and quality constraints that persist in current practice.

3. Revisiting Data Annotation

Data annotation has evolved significantly, transitioning from labor-intensive manual methods to sophisticated AI-driven approaches. Table 2 shows some recent intelligent approaches applied in different domains while utilizing various techniques. In this section, we outline the progression of annotation techniques, categorized into three groups: (i) traditional approaches, (ii) generic ML-based approaches, and (iii) GenAI and LLM-based approaches. We address challenges related to cost, scalability, and quality in data annotation techniques to establish a foundation for understanding the transformative role of AI agents.

Table 2. Examples from recent literature that illustrate modern annotation modalities and key techniques.

Ref.	Annotation Task	Techniques	Evidence/Results	Key Insight
Estévez-Almenzar et al. [36]	Face match labels	User study score check	Humans fix model errors	Hybrid oversight needed
Wei et al. [37]	Multi-annotator classification	EM vs vote approaches	Separation wins with noise	Keep annotator traces
Nasution and Onan [38]	Low-resource NLP labels	Zero/few-shot LLM	LLM near human accuracy	Experts still needed
Mamat et al. [39]	Crop vision tasks	CNN, YOLO, Mask-RCNN	Accuracy up 10–30 pp	Curated datasets essential

3.1. Traditional Approaches

Traditional data annotation methods rely heavily on human effort. These include manual annotation by individual experts and crowdsourcing through distributed workforces. Manual annotation involves trained annotators using specialized tools to label data with high precision. In computer vision, tools like LabelImg, VGG Image Annotator (VIA), and LabelMe (https://github.com/wkentaro/labelme, accessed on 20 June 2025) allow annotators to draw bounding boxes or polygons around objects in images and assign class labels. In natural language processing (NLP), platforms like Prodigy (https://prodi.gy/, accessed on 20 June 2025) support tasks such as named entity recognition and sentiment analysis. These methods excel in accuracy and context understanding, especially for nuanced tasks that require domain expertise. Crowdsourcing platforms (e.g., MTurk from Amazon, Appen, and Prolific) distribute micro-tasks to a global pool of workers to enable the rapid annotation of large datasets at lower costs. For instance, MTurk (https://www.mturk.com/, accessed on 20 June 2025) has been used to annotate product reviews for sentiment analysis by leveraging the collective effort of many annotators. However, crowdsourcing requires careful task design and quality control mechanisms such as majority voting to ensure consistency. Workers often lack specialized knowledge or sustained attention.

Despite their strengths, traditional approaches face significant limitations that hinder scalability and efficiency. Manual annotation is time-consuming and costly. Annotating the ImageNet dataset with over 14 million images required thousands of hours and substantial financial investment [39]. Human annotators are also prone to errors and inconsistencies, particularly in subjective tasks like emotion detection, where inter-annotator agreement falls as low as 60 to 70% [36]. Crowdsourcing, while more scalable, often yields inconsistent results due to diverse annotator backgrounds. One study found that up to 30% of MTurk annotations required correction because of quality issues [38]. Furthermore, both methods struggle with large-scale datasets and tasks that demand deep domain knowledge. Crowdsourcing also raises privacy concerns when handling sensitive data. These challenges have driven the development of ML-based approaches that automate parts of the annotation process, reduce costs, and improve consistency as shown in Figure 1.

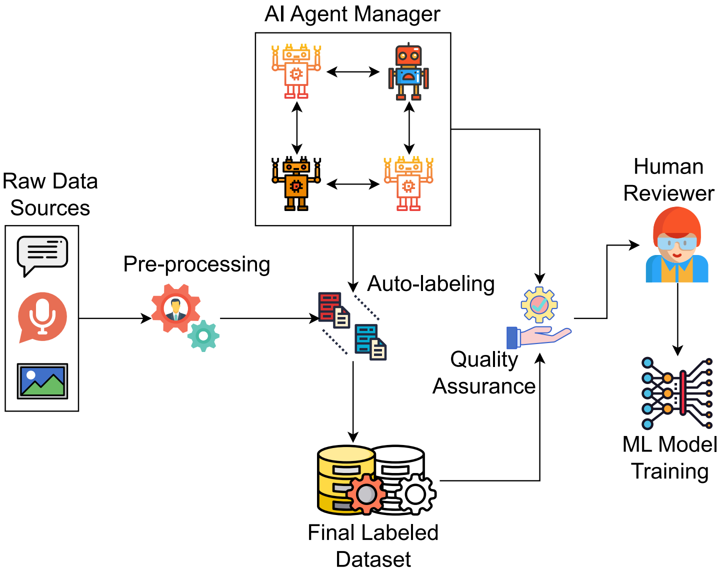

Figure 1. End-to-end AI-assisted data-annotation pipeline. Raw data of diverse modalities (text, images, audio, etc.) are ingested from external sources and pass through a pre-processing stage that standardizes formats and screens out corrupt samples. An auto-labeling module, coordinated by an AI agent manager, generates initial annotations at scale. All provisional labels are routed to a dedicated quality-assurance block; records that fail automated checks are escalated to a human reviewer for expert correction.

3.2. Generic ML-Based Approaches

ML-based approaches leverage algorithms to automate or assist in data annotation and address the inefficiencies of traditional methods. In particular, active learning is a prominent technique in which a model selects the most informative unlabeled data points for human annotation. This minimizes labeling efforts while maximizing model performance. For instance, uncertainty sampling identifies instances where the model is least confident as seen in text classification tasks where active learning has reduced annotation needs by up to 50% [40]. Weak supervision enables the rapid labeling of large datasets using noisy or indirect signals such as heuristics or pre-trained models. Programmatic labeling, a key weak supervision method, involves defining labeling functions that assign labels based on data patterns. One example is identifying entities in text using syntactic rules [41]. Furthermore, transfer learning uses pre-trained models, like BERT for NLP or ResNet for computer vision, to perform initial labeling on new tasks. By fine-tuning these models on a small set of labeled data, they can adapt to specific annotation requirements. This has been demonstrated in applications like named entity recognition [42]. These methods reduce reliance on manual effort and allow more efficient use of human expertise.

However, generic ML-based approaches have notable limitations. To begin with, active learning requires an initial labeled dataset and can be computationally intensive due to iterative model training. Similarly, weak supervision often produces noisy labels and may require additional denoising techniques such as majority voting among multiple labeling functions, which adds complexity [37]. Likewise, transfer learning suffers from negative transfer if the source and target domains differ significantly, leading to suboptimal performance. For example, a model pre-trained for general text classification may struggle with medical texts without extensive fine-tuning [43]. Moreover, these methods still require human input for initial labeling or rule definition. This process can be time-consuming and may demand domain expertise. These challenges highlight the need for more advanced techniques, such as generative AI and large language models (LLMs), which offer greater autonomy and adaptability. They also handle complex annotation tasks with minimal human input, further advancing the efficiency and quality of data annotation.

3.3. GenAI and LLM-Based Approaches

GenAI and LLMs have transformed data annotation by enabling automated labeling across diverse data types and tasks. LLMs such as GPT-3 [31], GPT-4 [32], and Llama-2 [33] interpret natural language instructions to generate annotations for text, images, and multimodal data. For text, LLMs excel in tasks such as sentiment analysis, named entity recognition, and relation extraction. The SELF-INSTRUCT method, for instance, prompts an LLM to generate both instructions and responses, creating annotated datasets for natural language understanding tasks [34]. In computer vision, multimodal LLMs like CLIP (https://github.com/openai/CLIP, accessed on 20 June 2025) enable zero-shot image classification by matching images to textual descriptions. Recent applications include annotating medical images and legal documents, where LLMs use domain-specific prompts to produce accurate labels. LLM-based annotation is generally classified into three categories: annotation generation, assessment, and utilization. Learning strategies such as prompt engineering and fine-tuning enhance LLM performance, making them versatile tools for annotation across domains.

Despite these advancements, LLM-based approaches face challenges that limit their reliability. LLMs can generate incorrect or hallucinated labels, particularly for ambiguous or out-of-distribution data [44]. Fine-tuning on domain-specific data is often necessary but resource-intensive, and the computational cost of running large models can be prohibitive.

Ensuring transparency in LLM decision-making is also critical, as their black-box nature complicates verification.

To address these issues, AI agents improve annotation by integrating interactive capabilities, such as engaging in dialogue with human annotators to resolve ambiguities and refine labels. This interaction improves accuracy while streamlining routine annotation tasks and reducing the computational and resource demands of fine-tuning. AI agents also log decision processes to enhance transparency. Therefore, they represent a promising research direction for improving efficiency and enabling more robust dataset creation.

4. AI Agents: A Primer for Data Annotation

To fully understand how AI agents revolutionize data annotation, it is essential to establish a foundational understanding of their conceptual framework and different types. Table 3 gives an overview of some AI-agent-driven annotation systems.

Table 3. Overview of agent-driven data annotation.

Ref.	Agent Blueprint	Annotation Modality	Core Concept	Remarks
Acharya et al. [45]	Hierarchical, goal-seeking stack (perception, reasoning, planning, action)	Conceptual blueprint for multi-stage, multimodal annotation pipelines	Autonomy, adaptability, long-horizon planning	Conceptual survey; Establish vocabulary for later task-specific agents
Osakwe et al. [46]	Single RL agent (A2C, PPO, DQN) optimizing SRL strategy	On-the-fly behavioral-sequence annotation from text-log traces	LSTM reward model + PPO/AC/DQN; episode-based training	Reward surpass random baseline; Demonstrate adaptive labeling with explicit rewards
Bianchini et al. [47]	HITL: LLM extractor, XR assistant, adaptive MES orchestrator	Procedural-instruction annotation in smart manufacturing (text + XR)	LLM parses docs to structured steps; XR guidance; MES sensor feedback	Qualitative error-rate reduction; Hybrid pipeline feed live industrial instructions
Xu et al. [30]	Multi-role LLM agent (Plan, Tool, Reflect) with self-correction loop	Automated outcome annotation (QA extraction, slot tagging, reasoning traces)	In-context learning, chain-of-thought, feedback refinement	Human-level label quality; Decomposition-based auditable labels
Xi et al. [25]	Generic LLM-agent stack (Brain/Perception/Action)	Taxonomy covering text, vision, and audio annotation scenarios	Memory modules, tool use, multimodal perception; few/zero-shot transfer	Synthesize benchmarks and risks; Provide classification schema for evaluating annotation agents

4.1. Overview of AI Agents

AI agents are autonomous entities that operate through a continuous cycle of perception, reasoning, action, and learning [48]. These agents perceive their environment via sensors or interfaces, analyze data to make informed decisions, execute targeted actions, and adapt their behavior based on feedback from their interactions [25]. The integration of LLMs has significantly advanced these capabilities. Such advancements establish a new paradigm for intelligent automation in data annotation tasks, leveraging sophisticated reasoning frameworks to enhance performance. For example, LLMs augment AI agent functionality across multiple dimensions through advanced reasoning strategies. Their natural language understanding enables the interpretation of complex instructions and the processing of diverse data formats with remarkable sophistication [49]. The reasoning capabilities of LLMs support intricate decision-making and task decomposition. Methods such as Chain-of-Thought (CoT) guide agents through linear, step-by-step logic to improve analytical performance [50]. Additionally, Tree-of-Thought (ToT) extends this approach by exploring multiple reasoning paths in a hierarchical structure, enabling agents to evaluate and backtrack among alternative solutions for complex annotation tasks [30]. LLMs also support few-shot and zero-shot learning, enabling rapid adaptation to new annotation tasks with minimal training examples [51]. Moreover, LLMs generate detailed explanations for their decisions. This enhanced transparency improves trust in automated annotation processes [52], while increased accuracy and a broader scope of addressable annotation challenges result from this integration [53]. Furthermore, LLMs enable natural conversational interactions. These interactions facilitate seamless collaboration between agents and human annotators throughout the annotation workflow [54]. Approach such as Auto-CoT [55] automates the generation of reasoning paths, reducing manual effort in prompt design. Likewise, ReAct [56] combines reasoning and action, allowing agents to interact dynamically with their environment, which further enhances adaptability in annotation workflows. Figure 2 presents the foundational architecture of an LLM-empowered agent system, enabling adaptive data annotation through modular feedback and environmental interaction.

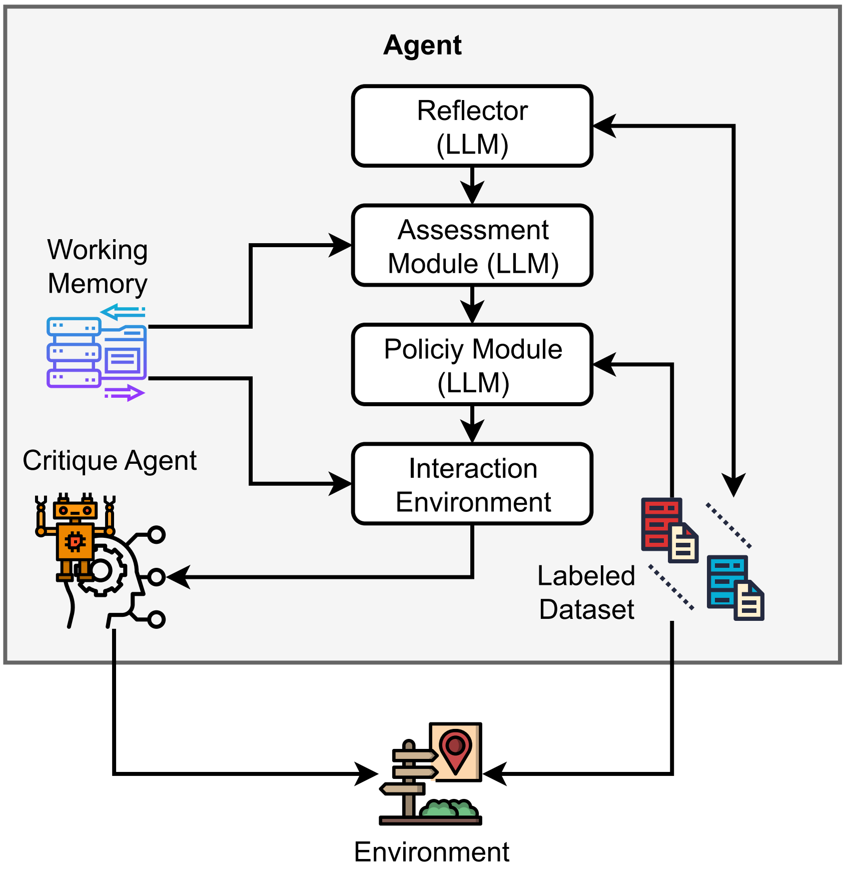

Figure 2. Schematic representation of an agent-driven interaction loop for data annotation. The central agent integrates multiple LLM-based modules: the reflector module evaluates past actions and outcomes; the assessment module analyzes annotation quality; and the policy module determines subsequent strategies. A critique agent provides targeted feedback to enhance decision-making within the interaction environment. The iterative loop culminates in the production of a refined labeled dataset, promoting scalability, consistency, and efficiency in AI-assisted annotation workflows.

4.2. Classification of AI Agents

We categorize AI agents based on their architectural design and functional capabilities, where each type is aligned to specific data annotation tasks:

- Rule-Based Agents: Operate on predefined conditional logic and excel at structured annotation tasks with clear decision boundaries. These agents are commonly used for format validation, basic text classification, and quality control checks in annotation workflows. For example, rule-based agents automatically classify customer support tickets based on keyword patterns or validate the completeness of bounding box annotations [57,58].
- Model-Based Reflex Agents: Maintain internal representations of annotation environments to handle contextually dependent labeling tasks. These agents excel at sequential annotation tasks such as named entity recognition in text or object tracking across video frames, where current decisions depend on understanding previous annotation states [59,60].
- Goal-Based Agents: Plan and execute multi-step annotation strategies to achieve specific data quality and coverage objectives. These agents decompose complex annotation projects into manageable workflows, prioritize annotation tasks based on model training needs, and adapt strategies to meet quality targets. Examples include agents that orchestrate active learning pipelines to maximize model improvement per annotation effort [61].
- Utility-Based Agents: Evaluate annotation decisions by optimizing multiple factors including confidence scores, cost effectiveness, and expected model performance gains. These agents are particularly valuable in active learning scenarios, where they select the most informative samples for annotation while balancing annotation costs against anticipated improvements in model accuracy [61,62].
- Learning Agents: Continuously adapt annotation strategies based on feedback from human annotators, annotation quality metrics, and model performance indicators. These agents improve over time by refining their understanding of annotation guidelines, reducing error rates, and becoming more efficient at identifying cases requiring human intervention [46,47].

4.3. Role of AI Agents in Data Annotation

AI agents introduce autonomous decision-making and goal-driven behavior with minimal supervision. Unlike conventional agents that operate within predefined rules and fixed tasks, AI agents demonstrate genuine agency through independent action and the purposeful pursuit of complex objectives [45]. In the context of data annotation, this evolution addresses major limitations such as inconsistent labeling across annotators, limited scalability with growing datasets, and declining quality over time as requirements change. Whereas LLMs focus on executing specific tasks, AI agents manage end-to-end workflows through multi-agent collaboration, dynamic task decomposition, and persistent memory [63]. Key advancements of AI agents for data annotation include the following:

- Collaborative Workflows: Specialized agents coordinate tasks through role assignment, where retriever agents collect relevant examples, labeling agents apply consistent annotation schemas, and validator agents perform quality checks. This structure enables scalable annotation pipelines that maintain consistency across large datasets.
- Adaptive Quality Control: Systems use confidence scoring to identify uncertain annotations requiring human review, anomaly detection to flag inconsistent labeling patterns, and cross-validation mechanisms in which multiple agents annotate the same samples independently to resolve disagreements and improve reliability.
- Context-Aware Adaptation: Agents adjust annotation strategies based on data characteristics, learn from feedback to refine guidelines, adapt to evolving annotation schemas, and handle domain shifts without requiring complete model retraining.

- Consensus Management: Multi-agent systems detect annotation conflicts, orchestrate consensus-building among human annotators, and maintain annotation histories to track changes in interpretation of ambiguous cases.
- Multimodal Integration: Advanced systems coordinate annotation across text, image, and audio modalities through specialized agents that preserve semantic consistency, resolve cross-modal dependencies, and ensure coherent labeling in unified annotation projects.

5. Research Methodology

This section explains how we identified, selected, and synthesized the literature that underpins our survey. We began by designing a set of Boolean search strings that combined core agent terms with data annotation keywords. Then, all retrieved papers were screened with clear inclusion and exclusion criteria to keep only studies that truly apply AI or LLM agents to annotation tasks.

5.1. Keywords

A set of well-chosen keywords and their combinations were employed to guarantee a thorough literature search. The following keywords were chosen to represent the complex connections between data annotation, AI agents, and their applications:

> (“AI agent” OR “autonomous agent” OR “LLM agent” OR “intelligent agent” AND “data annotation” OR label* OR “ground truth”),
> (“AI agent” OR “LLM agent” AND “generative AI” OR “large language model” AND “active learning” OR “self-refine” OR “prompt engineering”),
> (“AI agent” OR “autonomous agent” OR “LLM agent” AND “quality assurance” OR “inter-annotator agreement” OR “bias detection” AND “data annotation”)

5.2. Inclusion and Exclusion Criteria

Retrieved records underwent assessment against predefined criteria to determine eligibility. Inclusion required that publications be written in English. They have a primary focus on AI/LLM agents applied to data annotation or labeling in any modality. The articles propose, evaluate, or benchmark an agent-based workflow component—such as an active sampling module, quality guardrail, or automated label generator. Exclusion criteria encompassed pure surveys, opinion pieces, or tutorials lacking an original agent method. Works such as books, theses, patents, or non-peer-reviewed blog posts were excluded. Also, papers studying annotation tools without AI-driven agents (e.g., focusing solely on traditional crowdsourcing) and studies limited to single-platform implementations (e.g., image annotation in Labelme) without transferable agent concepts were not considered.

5.3. Research Databases and Selection Process

A multi-stage retrieval strategy was employed to construct a comprehensive review of AI-driven agents for data annotation. The initial phase involved querying eight bibliographic databases: ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, Scopus, Web of Science, Taylor and Francis, and arXiv.

Through systematic database searches, duplicate removal, and two-level eligibility checks, we assembled a robust corpus that supports the analysis of agent-driven annotation workflows, system architectures, and domain deployments as shown in Figure 3. This requirement guided every screening stage and preserved a consistent scope. The resulting evidence base underpins the survey’s conclusions on how AI and LLM agents are transforming data annotation. To maintain a sharp focus for the review, each candidate article had to address at least one of our three research questions (RQ1, RQ2, or RQ3) given below:

- RQ-1: Which peer-reviewed studies introduce AI or LLM agents that automate data-annotation workflows across any modality?
- RQ-2: What multi-agent or single-agent architectures are reported for planning, labeling, and quality checking in annotation pipelines?
- RQ-3: Which real-world application domains, especially those with sensitive data such as healthcare, finance, or content moderation, deploy LLM annotators?

Using the specified keywords, a systematic search produced 183 articles at the start of the selection phase. Then, the duplicate studies were removed, leaving a total of 167 records. After that, each article was reviewed using the inclusion and exclusion criteria. Articles irrelevant to the target areas were eliminated in the first round, bringing the total down to 145. The final 135 articles were chosen after a second round of screening eliminated studies that did not significantly advance our knowledge of AI agents and data annotation. This rigorous technique guarantees that the study is founded on excellent, relevant, and significant research.

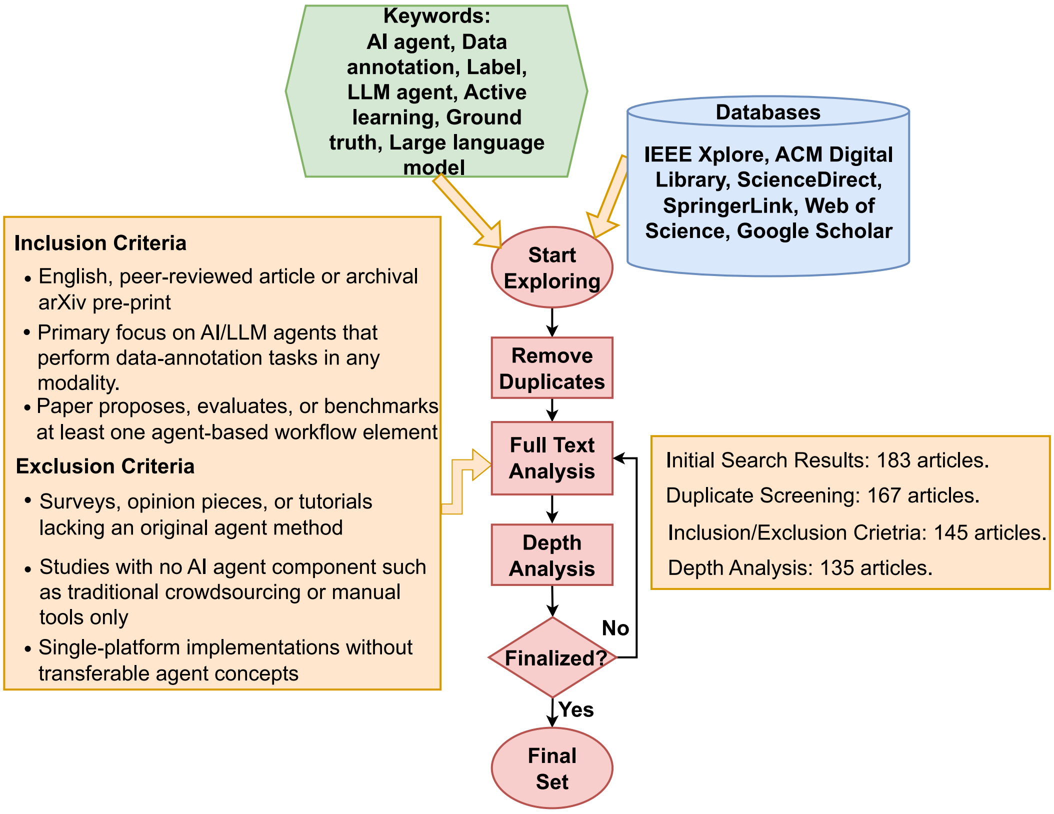
Keywords:
AI agent, Data annotation, Label, LLM agent, Active learning, Ground truth, Large language model
Databases
IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Web of Science, Google Scholar
Inclusion Criteria
• English, peer-reviewed article or archival arXiv pre-print
• Primary focus on AI/LLM agents that perform data-annotation tasks in any modality.
• Paper proposes, evaluates, or benchmarks at least one agent-based workflow element
Exclusion Criteria
• Surveys, opinion pieces, or tutorials lacking an original agent method
• Studies with no AI agent component such as traditional crowdsourcing or manual tools only
• Single-platform implementations without transferable agent concepts
Start Exploring
Remove Duplicates
Full Text Analysis
Depth Analysis
Finalized?
No
Yes
Final Set
Initial Search Results: 183 articles.
Duplicate Screening: 167 articles.
Inclusion/Exclusion Crietria: 145 articles.
Depth Analysis: 135 articles.

Figure 3. Visual representation of the research methodology for efficient literature selection. The process starts with broad keyword searches across six scholarly databases. After duplicate removal, the remaining corpus undergoes a two-stage screening: (i) full-text analysis against the inclusion/exclusion criteria and (ii) depth analysis to confirm methodological relevance. The flowchart visually aligns these steps to provide readers with a transparent and reproducible overview of the study selection protocol.

6. AI Agents for Annotation Workflow and Quality

AI agents transform data annotation workflows by intelligently selecting data, assisting annotators, and automatically ensuring label quality as shown in Table 4. These agents leverage the generative and reasoning capabilities of LLMs to reduce human effort while maintaining or even enhancing annotation accuracy. In this section, we discuss how AI agents improve annotation processes while focusing on adaptive data selection and quality assurance.

Table 4. Agent-centric workflows and their quality controls.

Ref.	**Workflow**	**Mechanisms**	**Impact**	**Domain**	**Remarks**
Rodriguez-Barroso et al. [64]	Post-label aggregation via federated clients	FedAvg on raw, disagreeing labels; preserves annotator privacy	Higher Macro-F_1 than majority vote across 8 NLP tasks	Subjective text (sentiment, hate, toxicity)	Disagreement becomes a *signal*—federated agents upgrade label robustness without extra human passes
Azeemi et al. [65]	Pre-label pool pruning before active learning	KenLM perplexity filter + quantized-LLM scoring	Cuts AL wall-clock time ↓74% while matching quality	Translation, summarization, sentiment corpora	LLM "gatekeeper" trims pools so expensive AL heuristics stay fast yet effective
Bayer et al. [66]	Cold-start and few-shot selection	Zero-shot GPT-4 picks first-batch items, fully auditable	+17–24 pp accuracy, seconds-level runtime	GLUE-style text classification	General-purpose LLM agent eliminates AL cold-start pain at minimal cost
Zhu et al. [67]	In-model label refinement	Uncertainty-weighted fusion; KL regularizer	More robust group-emotion accuracy on 3 benchmarks	Vision (crowd images)	Probabilistic "self-grader" agent down-weights noisy features mid-pipeline
Mishra et al. [68]	End-to-end orchestration across data products	Human checkpoints, KG-edge uncertainty, federated governance	Scales to 1.5 M projects and 27 M tenders while ensuring auditability	Multimodal procurement data	Multi-agent mesh shows how systemic quality gates span every annotation stage

6.1. *Active Data Selection and Adaptive Annotation*

Effective data annotation increasingly leverages adaptive methods that prioritize efficiency, quality, and responsiveness to evolving data challenges. Here, we provide several AI-driven strategies, emphasizing how LLM-based AI agents optimize active sampling, active learning, guideline adaptation, and continuous learning.

- Active Sampling Strategies: Active sampling strategies aim to maximize the information gained per annotation while minimizing labeling effort [67,69]. AI agents enhance this process by intelligently prioritizing data using LLM-driven insights. For example, ActiveLLM [66] addresses the "cold start" problem in few-shot learning by leveraging GPT-4 and similar models to select informative instances even with minimal initial data. The LLM-guided sampling significantly boosts downstream classifier performance, outperforming both standard active learning and other few-shot methods. In parallel, efficiency-oriented approaches like ActivePrune [65] use LLMs to prune large unlabeled pools before selection by employing a two-stage filtering, which not only outperforms prior pruning methods but also cuts total annotation time by up to 74%.
- LLM-driven Active Learning: LLM-driven active learning allows the AI agents to actively select and annotate data in each iteration [68]. One notable framework, LLMaAA [70], treats an LLM as an "active annotator" within the loop to decide what to label next in order to provide high-quality pseudo-labels. Modern approaches [70,71] generate new unlabeled instances that are predicted to be informative, rather than drawing only from a fixed pool. In these approaches, AI agents pose hypothetical or synthesized examples on the fly, expanding the training set with "high-value" queries that a human or the LLM itself can then label [71].
- On-the-fly Guideline Adaptation: In terms of annotation guidelines, AI agents assist in adapting on the fly to improve consistency [72–74]. Complex information extraction

(IE) tasks come with detailed guidelines that vanilla LLMs struggle to follow. To address this issue, Sainz et al. [75] fine-tune an LLM (e.g., GoLLIE model [75]) by following annotation guidelines, which leads to substantial improvements in zero-shot extraction accuracy on unseen IE tasks. Bibal et al. [76] introduce an iterative workflow in which LLM agents help update labeling instructions in response to annotator disagreements. The justifications are analyzed to identify ambiguities or shortcomings in the guidelines.

- Continuous Learning and Model Updates: AI-assisted annotation systems benefit greatly from continuous learning and frequent model updates. In the continuous learning setting, each batch of newly annotated data is immediately used to retrain or fine-tune the model before the next query selection [77]. These incremental updates allow the AI agent to progressively improve its predictions and selection strategy as more data becomes available. Recent research on lifelong learning for LLM-based agents underscores the importance of this adaptivity. For example, an LLM agent with memory and update mechanisms helps it integrate new knowledge without forgetting old knowledge, enabling continuous adaptation to changing data [78]. In effect, the AI annotator becomes smarter and more specialized with each iteration [79].

6.2. Annotation Quality and Consistency

Beyond automating the workflow, AI agents are enhancing the quality and consistency of annotations [80–82]. LLM agents distill domain knowledge into the labeling process and augment data to cover edge cases. Here we present how AI agents serve as diligent reviewers, where they check for bias, resolve conflicts between annotators, and detect errors that humans miss. These advances improve not only the speed of annotation but its reliability and fairness. Figure 4 illustrates the working of the intelligent quality checker agent in the annotation process.

- Knowledge Distillation and Augmentation: AI agents dramatically amplify annotation efforts by distilling knowledge from powerful models and augmenting datasets with synthetic examples [83,84]. The core idea of knowledge distillation in this context is to use LLM to generate labeled data or guidance that trains a small language model (SLM). Liu et al. [85] propose a knowledge distillation scheme where the system first analyzes the SLM weaknesses, then an LLM synthesizes new training examples specifically targeting those weak spots. Alongside distillation, LLM-based agents contribute to data augmentation for annotation. Instead of relying solely on human-curated examples, these LLM agents generate large quantities of plausible data points with labels [85,86].
- Bias and Fairness Checking: Maintaining high annotation quality and consistency is paramount, and AI agents play diverse roles in quality control. One contribution is through detecting and mitigating bias in labels, where LLM agents act as a second pair of eyes to review annotations for potential mistakes or biases. A recent study [64] used GPT-4 to re-evaluate crowdsourced annotations in an event extraction task, and the LLM flagged roughly 24% of the human-provided labels as debatable or likely errors. Another study [87] on political Twitter data found that zero-shot GPT-4 labeling not only achieved higher accuracy but also exhibited equal or lower bias compared to human annotators.
- Inter-annotator Agreement Conflict Resolution: When multiple annotators provide labels for the same item, disagreements often arise, resulting in ambiguity or differing interpretations. AI agents assist in resolving these conflicts to achieve higher inter-annotator agreement (IAA). For example, an LLM agent organizes the rationale and highlights where guidelines might be unclear. After iterative debate and guideline

refinement, the annotations converge, and the guidelines are updated to codify the resolved distinctions [76]. Choi et al. [88] explore chain-of-thought prompting and majority voting among LLM agents to imitate a panel of annotators, producing more consistent labels at scale. This results in higher consistency than individual annotators by aggregating multiple perspectives.

- Automated Quality Control: AI agents excel at the tedious yet critical task of quality control in annotation projects. They perform automated checks on the annotated data to catch errors, inconsistencies, or low-quality labels much faster than manual review. One common technique is to employ an LLM as a quality reviewer: after initial annotation (by humans or another model), the LLM is prompted to verify the label given the input and the guidelines [89]. Another strategy is the "LLM-as-a-judge" approach, where an ensemble of LLMs reviews existing labels and flags those likely to be incorrect. Gat et al. [90] implement this by prompting multiple diverse LLMs to label the same data and measuring their agreement against the original label.

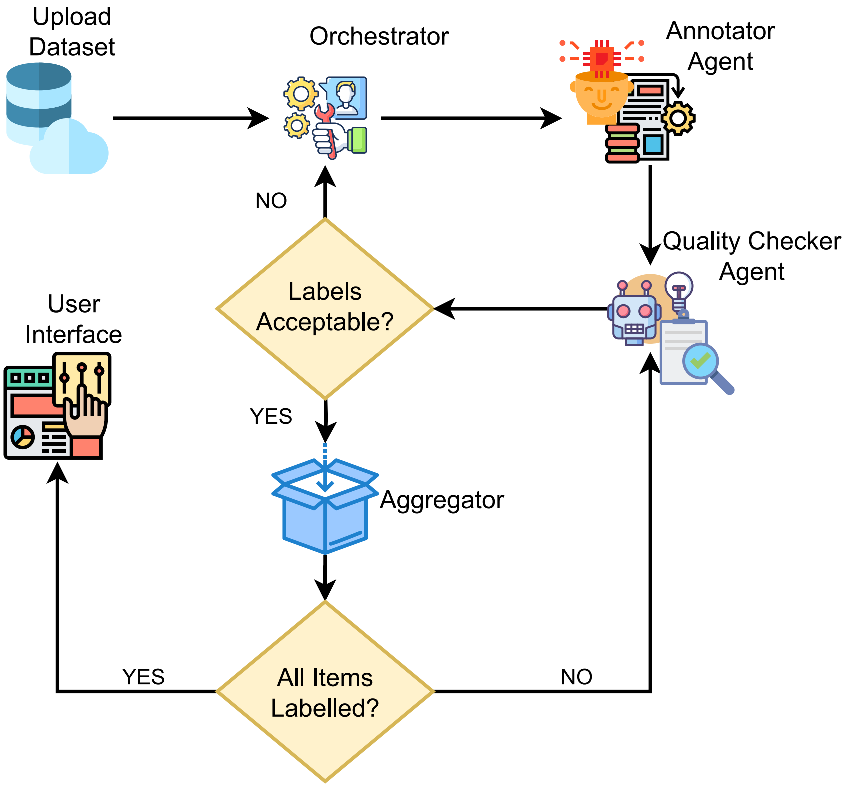

Figure 4. Flowchart of an automated quality-assurance loop for dataset annotation. The process begins with a dataset upload, which is directed by the orchestrator to the annotator agent for label generation. The labels are evaluated by the quality checker agent, leading to a decision point that assesses their acceptability. Unacceptable labels trigger a feedback loop to the orchestrator for re-annotation, with potential human intervention via the user interface. Acceptable labels are consolidated by the aggregator. A subsequent check verifies whether all dataset items have been labeled. If not, the cycle recommences through the orchestrator to ensure comprehensive coverage and iterative improvement in annotation reliability.

7. Architectures and Frameworks for Agent-Driven Annotation

Various architectural paradigms make AI-driven annotation more reliable, efficient, and scalable. These paradigms range from a single agent handling an entire labeling workflow to complex teams of agents collaborating on annotation as shown in Table 5. In this section, we review some emerging frameworks for agent-driven annotation, highlighting their design and benefits.

Table 5. Architectural paradigms and frameworks for AI-driven data annotation.

Ref.	Paradigm	Workflow	Coordination	Strengths	Limitations	Use Cases
Tan et al. [24]	Single-agent sequential pipeline (LLM as Generator, Assessor, Utilizer)	Taxonomy of generation, quality assessment, utilization across many data types	Prompt engineering, self-consistency scoring, filtering, downstream fine-tuning loops	Comprehensive blueprint over diverse modalities and stages	Conceptual survey; no concrete infrastructure specs	Designing new LLM-centric annotation pipelines in NLP or multimodal domains
Kumar et al. [91]	HITL active-learning loop	Data pre-processing, AL query, oracle labeling, model retrain	Uncertainty sampling, diversity sampling, iterative human feedback	Cuts labeling cost, preserves human oversight, domain-agnostic	Manpower-intensive; potential annotator bias	Biomedical imaging, NLP or CV tasks where AL reduces costly expert labels
Schleiger et al. [92]	Multi-agent human–AI collaboration	Shared-objective tasks with sustained two-way interaction; complementarity + shared goal + dialogue	Human intuition + AI computation; performance benefits catalogued	Gains in quality, creativity, safety and enjoyment when criteria met	Only 16 empirical systems; guidelines still emergent	High-stakes decision support where human judgment and AI scale must blend (e.g., clinical triage and manufacturing)
Liu et al. [93]	Single large-model agent with modular subsystems (Task Setting / Planning / Capability / Memory + Reflection)	Signal-analysis annotation: pre-processing, reasoning and planning, tool invocation, memory and reflection	Internal memory, self-questioning reflection, external tool calls (Matlab, RAG), multi-agent interaction bus	Autonomy, tool-use, continuous self-improvement, multimodal reasoning	Domain-specific (ETAR); relies on external tools and prompt engineering	Industrial / defense analytics needing rigorous, auditable annotation of complex sensor data

7.1. Single-Agent Sequential Pipeline

In a single-agent sequential pipeline, one AI agent autonomously performs all stages of the annotation process in a linear fashion. The agent first interprets the task and guidelines, then generates labels for the data, and finally verifies or refines its own outputs before finalizing them. Modularity allows easy integration and scalability [94]. A prominent example is the Self-Refine framework [95], where an LLM functions as both annotator and self-editor. Another example is the Reflexion agent framework [96], in which an LLM agent maintains an internal memory of feedback and "reflections" on its performance. After each annotation attempt, the agent generates a verbal self-critique or summary of errors and stores it in memory. This verbal self-feedback loop enables the single agent to improve its labeling accuracy over time, effectively learning from trial and error without weight updates. In contrast, DAIL [97] features an LLM agent that augments data by paraphrasing inputs and then applies majority voting to its predictions. However, a single agent still exhibits blind spots, reinforcing biases without external feedback, highlighting the importance of feedback loops for continuous improvement [61]. This limitation drives the need for more complex architectures involving multiple agents or validators.

7.2. Dual-Agent Reviewer Models

Dual-agent reviewer models introduce a second agent to critique or validate the outputs of the primary agent, forming a generator–reviewer pair. In this paradigm, one LLM agent produces annotations or intermediate prompts, and a second LLM agent reviews and provides feedback, establishing an iterative refinement loop between two distinct models. For example, Bubeck et al. [98] demonstrate an LLM-as-a-judge approach, in which two separate LLMs generate an annotation and then score or critique the output based on quality metrics. Similarly, Chen and Si [99] present a dual-agent system for automated story annotation. Cohen et al. [100] propose a strategy where the claims of one LLM are interrogated by another LLM acting as an examiner. Coordinating two agents introduces challenges, including confirmation bias when both agents share similar training data and increased computational cost [101].

7.3. Multi-Agent Collaboration

Multi-agent collaboration frameworks extend the concept of dual-agent systems to larger teams of agents that communicate and coordinate to perform annotation tasks. A prototypical multi-agent pipeline is shown in Figure 5. Instead of relying on a single agent, different agents are assigned specialized roles or subtasks and work together to produce final annotations. Collaboration can follow structured formats, such as hierarchical or role-based pipelines, or decentralized formats where agents negotiate outcomes. MetaGPT [102] exemplifies this approach by organizing LLM-based agents in an assembly-line structure, with each agent simulating roles in collaborative tasks. ChatDev [103] arranges agents in hierarchical workflows to decompose complex tasks. TESSA [104] combines general-purpose and domain-specific agents to jointly label data. DyLAN [93] introduces a two-stage architecture with candidate evaluation and structured dialogue for task completion. Multi-agent systems leverage collective intelligence but introduce additional complexity, including the need for communication protocols, risks of conflict or redundancy, and computational overhead [105,106].

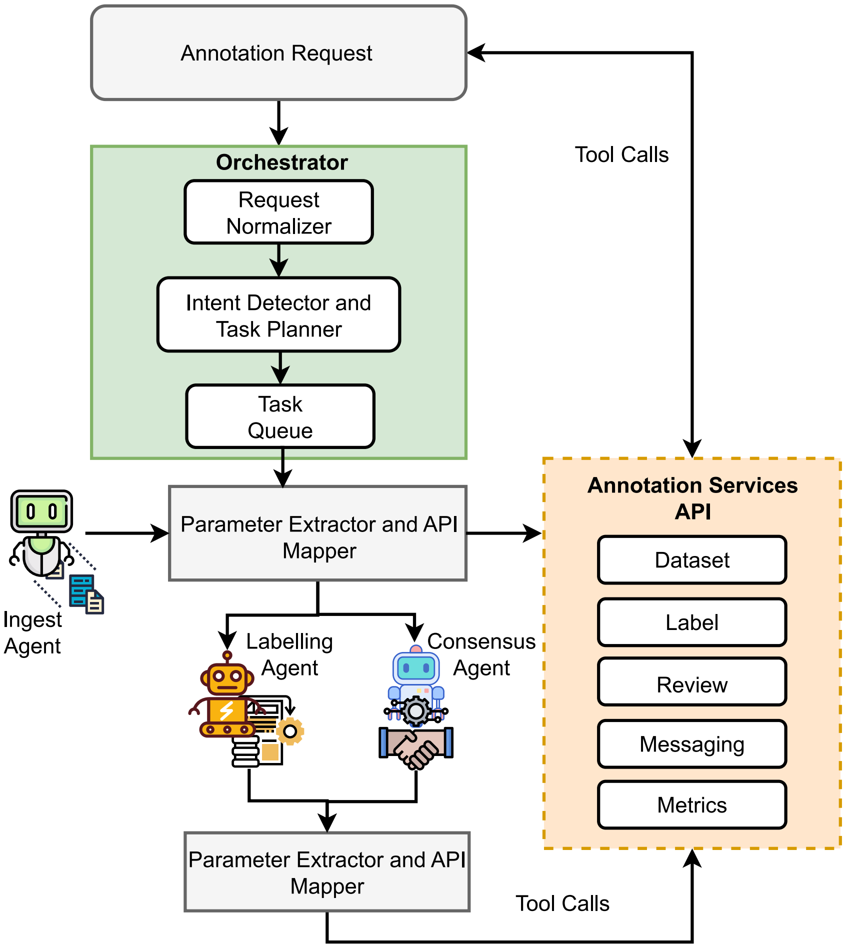

Figure 5. End-to-end workflow of a role-based agent ensemble for data annotation. The process begins with an annotation request entering the orchestrator, which performs normalization, intent detection, task planning, and queuing. Tasks are then routed through a parameter extractor and an API mapper to specialized agents: the ingest agent handles data intake, the labeling agent assigns initial annotations, and the consensus agent resolves discrepancies through validation. These components integrate with the annotation services API, which manages dataset operations including labeling, review, messaging, and metrics evaluation. Tool calls and an additional API mapper support external interactions and ensure consistency within the modular pipeline.

7.4. Human-in-the-Loop as Agent

Although autonomous AI agents have become increasingly capable, human expertise remains essential in many annotation scenarios. The HITL paradigm incorporates human annotators as integral components of the system architecture, working alongside LLM-based agents to ensure quality and manage edge cases [91,92]. One notable framework is CoAnnotating [107], which frames annotation as dynamic work allocation between humans and LLM agents, optimizing efficiency without compromising accuracy. Effective user interface design for seamless human–agent interactions is crucial to these frameworks, ensuring collaboration and oversight [24,108].

8. Evaluating AI Agents in Data Annotation

The systematic evaluation of AI agents in data annotation is a fundamental requirement for enhancing reliability and operational efficiency. In this section, we present a comprehensive discussion of evaluation methodologies that address performance metrics, economic impact assessments, and user experience considerations. Table 6 summarizes the key evaluation metrics and roles of AI agents in data annotation and serves as a foundation for the detailed analysis in this section.

Table 6. Evaluation metrics and roles of AI agents in data annotation.

Ref.	**Type**	**Agent Role**	**Metrics**	**Properties**	**Remarks**
Liu et al. [109]	Benchmark, Performance	Benchmark harness testing LLMs as autonomous agents (8 environments)	Success Rate, F1, reward, avg. turns	CoT prompting; temp=0	GPT-4 leads 27 models; exposes gaps in long-horizon reasoning for OSS LLMs.
Verma et al. [110]	Performance, Economic	Few-shot web agent adaptor (meta-learned planner)	Element Acc., Op. F1, Step SR, Overall SR (+4–7 pp)	Prompt-token cost; seconds-level latency	1–2 multimodal demos boost success 21–45% while keeping compute low.
Schmidt et al. [111]	Performance, Economic	Multi-robot active-learning coordinator	mIoU 2.5 pp, Top-1 Acc.	Data-upload volume 90 %	Synchronized selection matches pool-based AL with a fraction of bandwidth.
Wan et al. [112]	Performance, Economic	LLM generator, assessor, and utilizer pipeline for taxonomy + labeling	Coverage > 99.5 %; pairwise-label Acc./Rel.	Cost \$0.00006 vs \$0.082 per call; human-vs-LLM agreement	Lightweight classifiers match GPT-4 quality at a fraction of runtime cost.
Tutul et al. [113]	Performance, User-centric	Explainable-AI assistant for anxiety annotation	Spearman $\rho = 0.261$ on anxiety bins	8-item Likert trust scale; behavior–trust correlation	Trust rises over time but dips after errors; validates user-centric metrics.
Bolock et al. [114]	Performance, User-centric	Web agent for real-time stimulus display	Reaction-time stats; ANOVA $F = 2.997$ ($p \approx 0.05$)	Scalable in-browser logging; bilingual participant pool	Low-cost agent captures valence shifts from code-switching at scale.

Acronyms: Acc. = Accuracy; Elem. = Element; Op. = Operation; mIoU = mean Intersection-over-Union; pp = percentage points; Rel. = Relevance.

8.1. Performance Metrics for AI Agents

The assessment of AI agents in data annotation requires comprehensive performance metrics that quantify technical effectiveness. Annotation throughput represents a fundamental efficiency indicator, measuring processing speed relative to human benchmarks [114]. Streaming graph-summarization tasks such as OSNet process millions of edges per second, giving a realistic upper bound for agent throughput on dynamic data [115]. Synthetic generators like ROLL create billion-node scale-free graphs in minutes, letting us benchmark label-throughput without privacy constraints [116]. Label quality evaluation typically involves comparison with gold-standard annotations or the analysis of consistency patterns across datasets [117]. The scalability dimension addresses an agent's capacity to handle increasingly complex and voluminous datasets while maintaining performance levels [111]. Recent research emphasizes adaptability across diverse annotation domains as a critical capability metric [110]. Illustrative studies include [53], who evaluated LLMs against human annotators in relevance detection using precision metrics and Cohen's kappa to quantify agreement levels. The innovative "Turking Test" developed by Efrat and Levy [118] examined whether LLM-generated annotations successfully adhered to established guidelines for natural language inference tasks. Zhao et al. [119] implemented a pragmatic evaluation approach by training models on AI-labeled data and comparing their performance against models trained with human-labeled datasets, demonstrating that LLMs achieved 90-95% of human-level accuracy in classification scenarios. These methodologies collectively illuminate both capabilities and limitations of AI agents in generating reliable annotations for machine learning applications.

8.2. Economic and User-Centric Metrics

Beyond technical performance, economic and user-centric metrics provide essential perspectives on annotation agent effectiveness. Cost reduction represents a primary economic advantage, with industry reports indicating AI-driven annotation can reduce expenses and completion time by approximately 50% compared to fully manual approaches in large-scale projects [120]. User satisfaction and trust emerge as critical qualitative indicators that significantly influence adoption trajectories and operational viability [113]. Satisfaction assessment typically employs structured feedback mechanisms, while trust can be measured through the analysis of user override patterns in AI-assisted annotation systems. Research indicates that well-designed agent interfaces minimize annotator fatigue while enhancing productivity, substantially improving workflow experiences. The balance between automation level and human oversight significantly impacts user trust development, with hybrid annotation systems typically demonstrating superior user acceptance compared to fully automated approaches [112]. The integration of these multidimensional metrics enables comprehensive evaluation of real-world implementation viability, ensuring annotation agents deliver both economic efficiency and positive user experiences across diverse operational contexts.

8.3. Standardized Benchmarks and Evaluation Frameworks

The development of standardized benchmarks represents a critical yet challenging requirement for objective comparison of annotation agents. Effective benchmark systems must evaluate accuracy, efficiency, and adaptability across diverse annotation tasks to establish consistent performance assessment frameworks [109]. Domain-specific evaluation instruments have emerged, including knowledge completion assessments for graph annotation and diversity metrics for synthetic dialogue generation as demonstrated in recent literature [121]. Quality filtration constitutes an essential evaluation component with approaches ranging from rule-based heuristics that identify and remove substandard outputs to sophisticated LLM-driven self-consistency verification methodologies [122]. Advanced evaluation techniques increasingly employ LLMs as assessment judges to rank annotation quality, with [123] demonstrating reasonable alignment between LLM judgments and human evaluation standards. While these approaches enable scalable assessment, they necessitate careful implementation to address potential biases inherent in automated evaluation systems. Established standardized frameworks will facilitate rigorous comparative analysis while simultaneously accelerating innovation in annotation agent development.

8.4. Empirical and Quantitative Validation

Empirical studies reveal significant performance variations across different agent architectures. Single-agent systems demonstrate remarkable efficiency gains, with ActiveLLM achieving 17–24% points accuracy improvement over traditional active learning methods while maintaining seconds-level runtime [66]. In contrast, multi-agent collaboration frameworks show superior quality outcomes but at increased computational cost. For instance, federated agent approaches achieve higher Macro-F1 scores than majority voting across 8 NLP tasks, though requiring coordinated processing overhead [64]. The comparative analysis of HITL versus fully autonomous agent systems reveals critical trade-offs. HITL frameworks like CoAnnotating demonstrate optimal efficiency–accuracy balance, reducing annotation time by up to 74% while matching human-level quality through strategic human–agent work allocation [65,107]. Conversely, fully autonomous systems achieve superior scalability, with platforms processing millions of instances at costs as low as CNY (Chinese Yuan) 0.00006 per annotation call compared to CNY 0.082 for traditional methods [112].

Empirical validation studies employ standardized metrics enabling direct performance comparison. SR emerges as a primary effectiveness indicator, with state-of-the-art agent systems achieving 85–95% success rates across benchmark tasks compared to 60–75% for baseline methods [109]. Element accuracy and operation F1 scores provide granular performance insights, revealing that agent adaptability contributes 4–7 percentage point improvements in overall task completion [110]. Economic validation demonstrates substantial cost-effectiveness advantages. Industry deployments report 50% reduction in annotation expenses and completion time compared to fully manual approaches, with some specialized systems achieving up to 80% cost reduction through intelligent human–agent collaboration [124]. Processing speed metrics show orders-of-magnitude improvements, with agent systems handling thousands of annotations per hour compared to dozens for human annotators, while maintaining comparable or superior quality metrics. These quantitative findings are further validated through extensive real-world deployments detailed in Section 9, where case studies demonstrate consistent performance improvements across healthcare, financial services, content moderation, and humanitarian applications, with documented cost reductions of 50-80% and throughput improvements of 3–5x over traditional annotation methods.

9. Real-World Applications, Case Studies, and Tools

The transformative potential of AI agents in annotation is already being realized in a range of real-world domains, supported by both open-source tools and commercial products. Here we highlight several notable applications, case studies, and available tools.

9.1. Content Moderation and Policy Compliance

AI agents have revolutionized content moderation workflows across digital platforms through sophisticated annotation capabilities. Content moderation exemplifies this transformation, where platform-specific policies encoded into agent prompts enable the consistent, high-speed labeling of user-generated content across categories including hate speech, harassment, and inappropriate material [125]. On the other hand, agent-based solution significantly reduce the psychological burden on human moderation teams while maintaining annotation quality [126]. Research studies have demonstrated that LLM-based annotation agents can match crowd worker accuracy for hate speech detection while processing content exponentially faster [107]. Industries have further advanced this paradigm by developing systems where LLMs serve dual roles as both content generators and safety annotators, creating a closed-loop system that ensures policy compliance throughout the content lifecycle [127,128]. These agent-based moderation systems represent a paradigm shift in how platforms handle the ever-increasing volume of user-generated content requiring supervision. Community-aware friend-suggestion engines now fuse text, interaction graphs and topic interest vectors, providing rich multimodal labels that autonomous annotators can exploit for continual learning [129].

9.2. Customer Feedback and Support Ticket Triage

Enterprise adoption of AI annotation agents has accelerated dramatically in customer experience management, where organizations deploy agents to process and categorize massive volumes of unstructured feedback data. E-commerce platforms utilize annotation agents to systematically tag thousands of customer reviews with thematic labels and sentiment indicators, transforming raw textual data into structured insights overnight. Case studies from 2023 demonstrate that support departments implementing LLM annotation agents for ticket triage achieved 90% agreement with manual classification while saving dozens of staff hours weekly. Platforms like UBIAI have commercialized these capabilities, offering integrated systems that automatically label support documents using either zero-shot LLM predictions or domain-specific fine-tuned models [24,130]. These annotation agents excel at imposing meaningful structure on unstructured business data at unprecedented scale, enabling downstream analytics and workflow automation that would be prohibitively expensive through traditional annotation approaches.

9.3. Biomedical and Scientific Data Annotation

The complex, specialized nature of biomedical data annotation has traditionally required scarce expert resources, but AI agents are rapidly transforming this landscape. In genomics, where functional annotation requires labeling DNA sequences with gene functions and regulatory elements, generative AI agents are accelerating the understanding of gene functions as highlighted in a 2023 Lancet commentary [131]. Clinical settings have seen the experimental deployment of specialized healthcare LLMs that automatically annotate radiology reports and clinical notes with medical codes and key observations—tasks typically performed by specialized medical coders. Studies show that LLMs fine-tuned on clinical text can approach professional-level accuracy in diagnosis and procedure code annotation. While these systems undergo rigorous validation to address regulatory and precision requirements, they demonstrate how annotation agents can dramatically reduce the documentation burden in healthcare while maintaining annotation quality comparable to expert annotators. Figure 6 shows a representative suite of medical annotation agents spanning imaging, EHR extraction, and evidence retrieval.

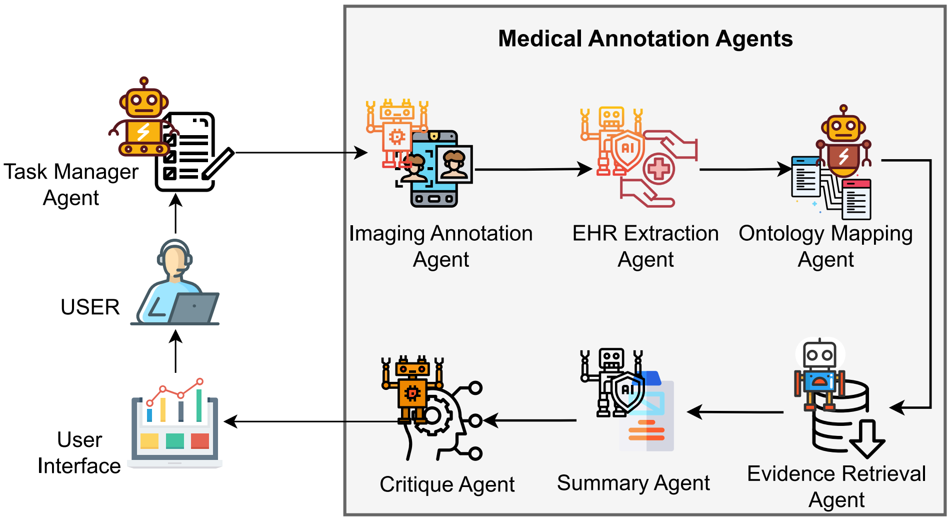

Figure 6. Representative stack of healthcare-focused agents for medical data annotation. The architecture centers on the task manager agent, which coordinates assignments to specialized components: the imaging annotation agent manages medical image annotation, the EHR extraction agent extracts data from electronic health records, and the ontology mapping agent ensures alignment with medical ontologies. Complementary agents include the critique agent for validation, the summary agent for result consolidation, and the evidence retrieval agent for supporting references. These agents enhance reliability through iterative feedback. A user interface enables human interaction, supporting a hybrid system that balances automation with expert oversight in sensitive healthcare domains.

9.4. Multimodal Data Labeling

AI annotation agents have transformed multimodal data labeling through the sophisticated orchestration of specialized models. In computer vision, platforms like Label Studio and CVAT now integrate agent-based assistance for automatic object detection, segmentation propagation, and complex annotation workflows. Overeasy IRIS exemplifies this advancement, combining multiple AI models under LLM guidance to automate image annotation for complex datasets like traffic camera footage. This system achieves remarkable 100x speed improvements while effectively handling challenging visual conditions including rain, night scenes, and motion blur [132,133]. The system processed 5 million surveillance frames, annotating license plates and vehicle types in days rather than months by employing intelligent prompting techniques. In audio annotation, agents now efficiently label speech transcripts with speaker characteristics and sentiment indicators, while also categorizing audio clips for specific events. VisionAgent further demonstrates the evolution of annotation agents by auto-generating code and models from high-level vision task prompts [132,133], illustrating how agents can orchestrate entire annotation pipelines for specialized use cases like manufacturing defect detection. Some locality-aware query operators achieve similar gains for geo-tagged frames and map overlays [134,135].

9.5. Case Studies

Early deployments of AI agent-driven annotation systems have demonstrated significant advantages over traditional manual methods. Case studies across various domains report substantial improvements in annotation speed and cost reductions, enabling large-scale projects that were previously impractical [133]. A key insight is the success of hybrid approaches, where AI agents provide automation and scalability, while human annotators ensure quality and handle complex cases [136]. In specialized fields such as healthcare and legal, fine-tuning LLMs on domain-specific data is critical for achieving the accuracy and reliability needed for high-stakes applications [137]. These findings set the stage for the specific case studies that follow.

- Humanitarian and Low-Resource Settings: AI agents are transforming annotation capabilities in humanitarian response and low-resource linguistic contexts where labeled data scarcity presents significant challenges. In disaster response scenarios, agent-based systems can rapidly process and categorize social media posts by urgency, need type, and location [138] to enable faster humanitarian intervention. These annotation agents demonstrate remarkable cross-lingual transfer abilities, allowing organizations to leverage multilingual capabilities to annotate content in low-resource languages without extensive human translator networks. Kim et al. [139] developed an agent-based annotation framework that strategically directs limited bilingual human resources to only the most uncertain cases identified by the system, achieving annotation coverage that would be impossible through traditional methods. Recent implementations by humanitarian organizations have shown that AI annotation agents can reduce response time by up to 70% while maintaining annotation quality comparable to human experts, effectively democratizing data annotation capabilities beyond well-resourced domains and languages [140].
- Enterprise Adoption and Agent-Based Workflows: Organizations across sectors are implementing sophisticated AI annotation agents tailored to their domain-specific data needs. Financial institutions have deployed fraud detection annotation agents that pre-screen claims by analyzing textual descriptions and transaction patterns, categorizing risk levels before human investigator review. Technology companies have integrated annotation agents throughout their data pipelines, with foundation model developers like OpenAI, Anthropic, and Meta leveraging their own models to generate and annotate massive instruction-following datasets at unprecedented scale [24,141]. The traditional boundaries between data generation and annotation have blurred as agent-based approaches like Self-Instruct [24] enable the automatic creation of labeled examples from minimal seed data. Enterprise adoption has accelerated with the integration of LLM-based annotation agents into existing data workflows, creating hybrid systems where AI agents perform initial annotation at scale while human experts focus on quality assurance, edge case identification, and agent performance improvement through feedback loops. This paradigm shift has positioned AI annotation agents as standard components in enterprise data pipelines, with many organizations reporting 3–5x increases in annotation throughput and significant improvements in dataset quality.

9.6. Tools and Platforms for AI-Driven Annotation

A wide range of tools and platforms is advancing the use of AI agents in data annotation. These solutions use large language models to improve both efficiency and scalability. They include open-source frameworks as well as commercial products, each designed to meet specific annotation needs across different domains.

- LangChain: LangChain, an open-source framework, has gained popularity for building LLM-powered applications. It provides abstractions for chaining prompts and actions, memory, and tool use [24]. It enables the creation of annotation agents through structured workflows where the system first processes annotation guidelines, then applies LLMs with chain-of-thought reasoning to each data item, optionally consults external APIs (such as knowledge bases), and finally produces the appropriate label. LangChain's agents feature robust integration capabilities with the environment (files, databases, and APIs), making them particularly valuable for complex annotation tasks requiring external knowledge sources [142,143].
- Prodigy: Prodigy introduced model-assisted annotation features early in its development and has continuously expanded its capabilities over time. Recently, it integrated

large language models (LLMs) into annotation workflows, supporting both few-shot and zero-shot paradigms by enabling LLMs to suggest labels or generate annotations directly, while retaining human oversight through an accept-or-reject verification mechanism [144]. Prodigy's underlying philosophy centers on "AI as you annotate", an approach designed to simultaneously accelerate the annotation process and deliver measurable quality improvements.

- Labelbox and Scale AI: Labelbox has pioneered AI agent-based annotation workflows, enabling trajectory training and evaluation for agent development where human labelers optimize agent prompts, fine-tune LLMs, and provide critical feedback on agent performance [145,146]. The platform's 2025 enhancements include agent-specific capabilities in the Multimodal Chat Editor for trajectory labeling, allowing teams to create, edit, and annotate complete agent reasoning steps, tool calls, and observations [147]. Similarly, Scale AI has developed an advanced Data Engine featuring AI-based techniques with HITL systems, enabling annotators to identify high-value data for curating agent training sets and implementing automated quality controls that adapt to agent workflows. This human–AI collaborative approach represents a paradigm shift where professional annotator workforces are augmented by AI agents to deliver superior efficiency, quality, and cost effectiveness.
- Cleanlab Studio: Cleanlab has revolutionized data annotation through its Autolabeling Agent [124], which implements an advanced AI agent architecture designed to reduce annotation costs by up to 80%. The system employs active learning algorithms that intelligently identify the most informative data points for human review, while the agent autonomously handles confidently predicted examples. The HITL agent solution continuously learns from expert supervision, allowing teams to iterate through a workflow where the AI agent progressively improves its annotation capabilities based on expert corrections. Its trustworthy language model provides trustworthiness scores for agent-generated annotations, creating a reliability layer that enables organizations to confidently automate up to 99% of annotations while directing human attention only to uncertain cases requiring expert judgment.
- Snorkel: Snorkel features dedicated GenAI agent tools that support trajectory evaluation and annotation, allowing subject matter experts to review AI agent responses and rank them according to quality [148]. The platform's programmatic labeling approach enables users to define labeling functions that can call LLMs under the hood, effectively merging the programmatic labeling paradigm with agent capabilities. Snorkel's Multi-Schema Annotation capability further enhances this functionality by enabling organizations to customize label schemas to collect annotations for different agent evaluation metrics simultaneously. This platform revolutionizes how enterprises capture domain knowledge, build agent evaluation datasets, and fine-tune agent workflows through techniques such as instruction tuning and direct preference optimization.
- Open-source Initiatives: The open-source community continues to drive innovation in agent-based annotation through diverse projects and resources. The LLM4Annotation GitHub repository [24] serves as a comprehensive collection of research papers focused on leveraging agents for annotation workflows. Tools like autolabel [149] enable the creation of agent-based labeling pipelines that can be customized for specific domains, with built-in integration for HuggingFace datasets and models. Additionally, community projects demonstrate the practical implementations of agent-based annotation systems that leverage commercial APIs like OpenAI for common labeling tasks. These open initiatives provide valuable resources for organizations seeking to experiment with agent-based annotation before investing in enterprise solutions,

fostering innovation and accessibility in the rapidly evolving field of AI-assisted data annotation.

- Additional Tools: Several platforms leverage AI agents to streamline and enhance annotation workflows. For instance, CVAT AI (https://www.cvat.ai/, accessed on 20 June 2025) Agents offer customization flexibility for integrating proprietary models into annotation pipelines [150], while Datasaur's LLM Labs playground enables comparative performance testing across multiple language models to optimize annotation workflows [151]. For specialized use cases, GPTBoost (https://www.gptboost.io/, accessed on 20 June 2025) provides purpose-built Annotation Agents targeting sentiment analysis in communication data, and MEGAnno+ implements a balanced human–LLM collaborative system that preserves expert oversight while maximizing automation benefits [139]. Additionally, Agent Label (https://www.agentlabel.ai/, accessed on 20 June 2025) focuses on acceleration-oriented tools designed specifically for machine learning datasets, while established platforms such as V7 (https://www.v7labs.com/, accessed on 20 June 2025) have expanded their offerings to include AI Agents for document processing automation. Furthermore, Labellerr (https://www.labellerr.com/, accessed on 20 June 2025) distinguishes itself through transfer learning approaches that leverage pre-trained models to minimize coding requirements while maintaining annotation quality. These diverse platforms illustrate how agent-based annotation technologies are rapidly evolving to meet specific industry needs through specialized implementations.

10. Research Challenges and Future Directions

While AI agents are revolutionizing data annotation, they also introduce a host of challenges and ethical considerations that must be addressed moving forward. We discuss some of the key concerns and outline future directions for research and practice.

10.1. Quality and Reliability Concerns

AI annotation agents, despite their advanced capabilities, often demonstrate limited reliability when tasked with complex, domain-specific, or out-of-distribution inputs [89]. In specialized areas such as medicine or law, these systems frequently misinterpret technical language, leading to annotation errors that human experts would easily recognize. Hallucination adds further complexity, as agents sometimes produce inaccurate labels or generate justifications not supported by the input data. This behavior includes attributing features to images that do not exist or assigning sentiment to neutral text without justification. Such issues highlight the ongoing risk of deploying fully autonomous annotation systems without human supervision in high-stakes contexts.

Real deployments already show that annotation agents can misfire in varied and costly ways. A study on multilingual classification found that large language models wrongly flagged benign Arabic, French, German, and Hindi posts as hate speech, producing error rates well above human baselines [152]. Follow-up work on HarmonyNet [153] confirmed that even ensemble systems still miss subtle context and let toxic content slip through, while over-blocking neutral messages. Similar accuracy drops were seen when the task moved to low-resource Telugu data [154]. In medicine, ChatGPT added non-existent findings to 60% of its chest-X-ray summaries, risking needless follow-up scans and patient anxiety [155], and another emergency-department trial recorded 2.3% false pneumonia labels on normal images [156]. Researchers showed that a single prompt-injection string can overwrite stored annotations, corrupting entire corpora in one step [157], while a wider survey found that today's security tools catch barely half of such attacks [158]. A study reports that hallucination detection itself is unreliable since current verifiers miss many synthetic facts [159], and that prompt-injection can force an e-commerce annotator into rewriting product tags with marketing slogans, breaking more than fifteen thousand records before rollback [160]. Taken together, these sources show that hallucination, bias, and adversarial control are not theoretical risks but measurable failures that already disrupt annotation pipelines.

Current research prioritizes reliability improvements through refined prompting strategies that minimize ambiguity, evidence-grounding techniques that require annotations to reference specific elements of the input, and the integration of verifier models that assess label validity [54,161]. Additional progress is being made in the development of evaluation benchmarks tailored to edge cases, which allow for targeted stress testing and subsequent fine-tuning based on observed failure patterns. As agents achieve higher reliability on narrowly scoped tasks, organizations can introduce selective automation while retaining human oversight for complex or ambiguous annotation decisions.

10.2. Ethical and Legal Considerations

AI annotation agents often reflect and amplify biases present in their training data. When LLMs internalize stereotypes or problematic associations, these issues appear in their annotation outputs. For instance, an agent trained on biased data sometimes flags content from certain dialects as inappropriate or performs with reduced accuracy when labeling images representing underrepresented demographic groups. This propagation of bias through annotations raises serious concerns, as the resulting labels directly shape downstream models. Researchers have found that, although crowdsourced annotations exhibit recognizable human bias patterns, LLM-based annotations introduce new forms of bias that are often more difficult to detect, particularly when systems are trusted without sufficient scrutiny [162].

AI-driven data-annotation pipelines that rely on LLM agents raise security and privacy concerns such as safeguarding personal information and ensuring decision accountability. A recent review of privacy-preserving techniques for generative AI lists differential privacy, federated learning, homomorphic encryption and secure multi-party computation as effective defenses that keep raw inputs inside the data-controller's trust boundary [163]. Recent studies show that on-premise deployment combined with selective disclosure can mitigate membership-inference and data-reconstruction attacks during labeling tasks [164]. A research for clinical LLMs argues that structured logging is essential for reconstructing how an agent arrived at a label and for measuring bias and accuracy over time [165]. Technical guidance on data provenance in healthcare echoes this point and recommends immutable audit trails that capture prompts, model versions, and chain-of-thought justifications [166]. Recent work on the clinical adoption of LLM systems defines accountability as a triad of auditability, risk management, and redress, reinforcing the need for post-deployment monitoring [167]. Surveys of explainable AI in healthcare find that transparent rationales improve clinician acceptance of automatically generated labels and support downstream bias analysis [168].

Future research must develop rigorous bias auditing protocols tailored to AI-generated annotations. Strategies include the use of diverse calibration datasets, consensus mechanisms involving multiple models with distinct characteristics, and algorithmic bias mitigation applied directly to generated labels. Involving human reviewers from varied backgrounds in the evaluation of sensitive annotation tasks improves the detection of biased outputs. From an ethical governance standpoint, annotation agents require the same fairness assessments and transparency obligations as other AI systems. Standardized documentation formats, such as adapted model cards, support accountability in this context and help formalize responsible deployment practices.

10.3. Transparency and Explainability

Understanding the rationale behind an AI agent's annotation decisions is critical for multiple stakeholders. Developers rely on this insight for debugging, users depend on it to build trust, and regulatory frameworks often require it, especially in sensitive fields such as healthcare, where image classification outcomes must be justified. Although LLM agents are capable of generating explanations for their annotations, there is no assurance that these explanations accurately reflect the model's internal decision-making process. The model may produce reasoning that appears plausible but diverges from the actual computational pathway used to generate the label.

An emerging research direction involves embedding self-explanation directly into the annotation process, treating chain-of-thought reasoning as a core output subject to separate evaluation. Structurally prompting models to express their reasoning before assigning labels, and conditioning label acceptance on the clarity and validity of that reasoning, improves transparency and overall system accountability [169]. Techniques from explainable AI provide further support, including saliency mapping for text inputs and visual segmentation for images, which help identify the features that influenced a model's classification decision. Future annotation systems are expected to present visual explanatory elements, such as heatmaps or highlighted regions, to justify model outputs and support efficient human verification.

10.4. Human Displacement and Evolving Roles

The integration of AI annotation agents raises important questions about the future employment landscape for human annotators, many of whom are part of global workforce networks. In the near term, human roles are shifting from direct labeling to responsibilities such as exception handling, quality control, and more complex judgment tasks. This transition requires annotators to acquire advanced skills for effective collaboration with AI systems, including the interpretation of agent outputs, the ability to provide actionable feedback to improve agent performance, and the discernment needed to decide when to accept or override AI-generated annotations. A related challenge is automation bias, where annotators may uncritically accept AI suggestions, leading to the endorsement of inaccurate outputs.

Specialized training programs for HITL workers and well-designed interfaces that promote critical oversight are essential. One interface strategy involves occasionally concealing AI suggestions to ensure that human assessment capabilities remain active and independent. From a workforce perspective, although some entry-level annotation jobs are becoming less common, new specialized roles are emerging. These include positions such as "annotation manager" or "AI auditor", where human experts supervise AI-based annotation systems. A sustainable approach assigns repetitive annotation tasks to AI agents while human contributors focus on complex decisions and strategic enhancements to the pipeline, including guideline development and data quality analysis. Ensuring responsible deployment will depend on close monitoring of this transition and investment in fair and effective re-skilling initiatives.

10.5. Data Privacy and Security

Many annotation workflows involve sensitive information, including user data, proprietary content, and confidential documents. Transmitting this data to third-party LLM APIs introduces significant privacy vulnerabilities. Even when service providers enforce non-retention policies, which differ considerably across vendors, concerns remain about potential data leakage through model outputs. For example, an LLM is capable of inadvertently revealing elements of confidential input through its reasoning processes or becoming susceptible to prompt-based extraction of sensitive information during later interactions involving similar content.

Deploying models on-premises and using encrypted processing techniques are promising strategies for reducing privacy risks [170]. Techniques such as homomorphic encryption support computation on encrypted data but often result in significant performance trade-offs as demonstrated in visual AI applications [171]. In high-sensitivity environments, organizations find it justifiable to run LLMs locally within secure infrastructure. Other privacy safeguards include data anonymization and selective field masking before agent processing, although these measures occasionally alter the core annotation objective. Regulatory compliance and data governance frameworks, including HIPAA for healthcare data and GDPR for personal information in the EU, increasingly influence how annotation agents are deployed in enterprise systems. Future annotation architectures will prioritize secure environments controlled by the data owner or apply privacy-preserving transformations before processing.

10.6. Scalability and Cost Trade-Offs

While AI annotation agents often prove more economical than human annotators across numerous use cases, the computational requirements of state-of-the-art LLMs introduce non-trivial cost considerations, particularly when processing millions of instances. API usage expenses increase rapidly, and deploying large models on-premises requires substantial infrastructure investment. Organizations face the challenge of optimizing the balance between cost and quality. On the other hand, smaller or open-source models fine-tuned for specific annotation tasks provide cost benefits, but these often come with reduced accuracy compared to more capable models. One effective strategy involves using advanced models to annotate a subset of training data, followed by fine-tuning specialized, more efficient annotation models for broader deployment. Active research investigates how to distill the annotation capabilities of frontier LLMs into compact, task-specific models. Latency also becomes a key factor in real-time or interactive annotation settings, where multi-second processing per item creates operational bottlenecks.

Quantitative analyses highlight stark differences in deployment strategies. For instance, cloud-based deployments (e.g., via APIs like OpenAI's GPT-4) can incur high ongoing costs, with estimates showing daily operational expenses exceeding CNY 700,000 for large-scale LLM services due to reliance on expensive GPUs [172]. In contrast, on-premise or local deployments using open-source models (e.g., Llama variants on dedicated hardware) offer upfront infrastructure costs but can reduce long-term expenses by up to 46% through better control over resources and avoidance of per-query fees [173]. Hybrid approaches, such as local-cloud offloading, further optimize this by running lightweight models on-device for simple tasks and offloading complex ones to the cloud, achieving latency reductions of 50–80% and cost savings of up to 89% in serverless setups [174]. Future research should focus on implementing cascade architectures that apply lightweight heuristics or efficient models for simple decisions, while reserving high-compute LLMs for more complex cases. Model inference (e.g., quantization and model pruning) is a notable direction to reduce memory footprint by accelerating inference speeds and enabling deployment on resource-constrained devices without significant loss in accuracy [175,176].

10.7. Evaluation and Governance

As agent-driven annotation becomes increasingly prevalent, comprehensive governance frameworks are essential to ensure quality, accountability, and responsible deployment. In line with established quality assurance practices in software development, machine learning data pipelines that incorporate AI annotators must implement systematic auditing and validation protocols. Industry standards are needed to certify annotation agents for specific use cases, following models similar to those used for medical devices and laboratory testing procedures. For example, the deployment of AI agents in clinical trial data annotation requires compliance with defined accuracy benchmarks and the integration of continuous monitoring mechanisms. Within organizations, dedicated dashboards should track annotation agent performance using metrics such as human override frequency and error pattern trends. To maintain reliability over time, systems must include robust mechanisms for identifying and correcting performance drift, especially in dynamic environments such as social media, where evolving language can cause annotation errors. These challenges call for safeguards such as continuous learning systems with HITL oversight to prevent the compounding of model errors.

Future research should prioritize the development of standardized evaluation frameworks tailored to annotation agents, along with governance models that balance technological innovation with ethical accountability. In addition, adaptive monitoring systems capable of detecting subtle forms of performance degradation must be designed to ensure long-term reliability. Progress in these areas will depend on close interdisciplinary collaboration among technical developers, domain experts, and policy researchers. Such cooperation is necessary to build governance infrastructures that address the specific demands of agent-based annotation systems while supporting their responsible evolution across domains.

11. Conclusions

This paper presented a comprehensive and systematic survey of AI agents and their transformative role in data annotation methodologies. We systematically classified AI agents based on their architectural designs, functional capabilities, and deployment contexts, establishing a clear taxonomy distinguishing between rule-based agents, learning-based agents, and hybrid LLM-enhanced agents. Our comparative analysis indicates that while traditional manual approaches achieve high precision in controlled settings, AI agents demonstrate superior scalability, consistency, and cost effectiveness, particularly for large-scale datasets. Current evaluation practices primarily emphasize accuracy metrics, with limited attention to broader factors, including annotation speed, cost reduction, and long-term sustainability. Despite significant advances, challenges persist regarding quality assurance, bias mitigation, transparency, and ethical considerations, all of which require ongoing research. The integration of multimodal capabilities, collaborative human–agent frameworks, and LLM-based advanced reasoning represents promising directions for future development. Furthermore, establishing standardized evaluation benchmarks and interoperability protocols will be crucial for the broader adoption of these solutions across diverse domains. With new annotation requirements and complex datasets emerging, AI agents are becoming essential infrastructure for next-generation data annotation ecosystems.

Author Contributions: Conceptualization and Methodology, M.M.K., S.K. and D.H.V.; validation, Q.Q.; investigation, resources, and data curation, M.M.K., S.K. and D.H.V.; writing, original draft preparation, M.M.K.; review and editing, X.L., C.W. and Q.Q.; visualization, S.K. and D.H.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China (No. 2020YFA0909100) and Basic and Applied Basic Research Foundation of Guangdong Province (No. 2023TQ07A264).

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Lists of Abbreviation

AI	Artificial Intelligence
AL	Active Learning
ANOVA	Analysis of Variance
API	Application Programming Interface
Auto-CoT	Automatic Chain-of-Thought
CLIP	Contrastive Language-Image Pre-training
CoT	Chain-of-Thought
CV	Computer Vision
EHR	Electronic Health Record
EM	Expectation-Maximization
EU	European Union
FedAvg	Federated Averaging
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HIPAA	Health Insurance Portability and Accountability Act
HITL	Human-in-the-Loop
IAA	Inter-Annotator Agreement
IE	Information Extraction
KG	Knowledge Graph
KL	Kullback-Leibler
LLM	Large Language Model
MES	Manufacturing Execution System
ML	Machine Learning
MTurk	Amazon Mechanical Turk
NLP	Natural Language Processing
OSS	Open-Source Software
QA	Question Answering
RAG	Retrieval Augmented Generation
ReAct	Reasoning and Acting
RL	Reinforcement Learning
SLM	Small Language Model
SR	Success Rate
ToT	Tree-of-Thought
XR	Extended Reality

References

1. Williams, K.L. The Role of Data in Artificial Intelligence: Informing, Training, and Enhancing AI Systems. In Proceedings of the International Conference on Information Technology-New Generations, Las Vegas, NV, USA, 27–29 April 2005; Springer: Berlin/Heidelberg, Germany, 2025; pp. 107–116.
2. Ghaisas, S.; Singhal, A. Dealing with Data for RE: Mitigating Challenges while using NLP and Generative AI. In *Handbook on Natural Language Processing for Requirements Engineering*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 457–486.
3. Houenou, B. AI Labor Markets: Tradability, Wage Inequality and Talent Development. Available online: https://ssrn.com/abstract=5163063 (accessed on 31 July 2025).
4. Haq, M.U.U.; Rigoni, D.; Sperduti, A. LLMs as Data Annotators: How Close Are We to Human Performance. *arXiv* **2025**, arXiv:2504.15022. [CrossRef]
5. Mirzakhmedova, N.; Gohsen, M.; Chang, C.H.; Stein, B. Are Large Language Models Reliable Argument Quality Annotators? In Proceedings of the Conference on Advances in Robust Argumentation Machines, Bielefeld, Germany, 5–7 June 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 129–146.
6. Zhu, Y.; Yin, Z.; Tyson, G.; Haq, E.U.; Lee, L.H.; Hui, P. Apt-pipe: A prompt-tuning tool for social data annotation using chatgpt. In Proceedings of the ACM Web Conference 2024, Singapore, 13–17 May 2024; pp. 245–255.

7. Yao, B.; Xiao, H.; Zhuang, J.; Peng, C. Weakly supervised learning for point cloud semantic segmentation with dual teacher. *IEEE Robot. Autom. Lett.* **2023**, *8*, 6347–6354. [CrossRef]
8. Sun, G.; Zhan, X.; Such, J. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In Proceedings of the 6th ACM Conference on Conversational User Interfaces, Luxembourg, 8–10 July 2024; pp. 1–6.
9. Zhang, L.; Zhang, Q.; Wang, H.; Xiao, E.; Jiang, Z.; Chen, H.; Xu, R. Trihelper: Zero-shot object navigation with dynamic assistance. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, UAE, 14–18 October 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 10035–10042.
10. Gottipati, S.K.; Nguyen, L.H.; Mars, C.; Taylor, M.E. Hiking up that hill with cogment-verse: Train & operate multi-agent systems learning from humans. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, London, UK, 29 May–2 June 2023; pp. 3065–3067.
11. Tsiakas, K.; Murray-Rust, D. Using human-in-the-loop and explainable AI to envisage new future work practices. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, 29 June–1 July 2022; pp. 588–594.
12. Yuan, S.; Chen, Z.; Xi, Z.; Ye, J.; Du, Z.; Chen, J. Agent-R: Training Language Model Agents to Reflect via Iterative Self-Training. *arXiv* **2025**, arXiv:2501.11425.
13. Renze, M.; Guven, E. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv* **2024**, arXiv:2405.06682.
14. Grötschla, F.; Müller, L.; Tönshoff, J.; Galkin, M.; Perozzi, B. AgentsNet: Coordination and Collaborative Reasoning in Multi-Agent LLMs. *arXiv* **2025**, arXiv:2507.08616.
15. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
16. Park, J.S.; O'Brien, J.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual Acm Symposium on User Interface Software and Technology, Francisco, CA, USA, 29 October–1 November 2023; pp. 1–22.
17. Lewis, P.R.; Sarkadi, Ş. Reflective artificial intelligence. *Minds Mach.* **2024**, *34*, 14. [CrossRef]
18. Chu, S.Y.; Kim, J.W.; Yi, M.Y. Think together and work better: Combining humans' and LLMs' think-aloud outcomes for effective text evaluation. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025; pp. 1–23.
19. Nathani, D.; Madaan, L.; Roberts, N.; Bashlykov, N.; Menon, A.; Moens, V.; Budhiraja, A.; Magka, D.; Vorotilov, V.; Chaurasia, G.; et al. Mlgym: A new framework and benchmark for advancing ai research agents. *arXiv* **2025**, arXiv:2502.14499.
20. Yu, A.; Lebedev, E.; Everett, L.; Chen, X.; Chen, T. Autonomous Deep Agent. *arXiv* **2025**, arXiv:2502.07056.
21. Demrozi, F.; Turetta, C.; Machot, F.A.; Pravadelli, G.; Kindt, P.H. A comprehensive review of automated data annotation techniques in human activity recognition. *arXiv* **2023**, arXiv:2307.05988. [CrossRef]
22. Zhou, Y.; Guo, C.; Wang, X.; Chang, Y.; Wu, Y. A survey on data augmentation in large model era. *arXiv* **2024**, arXiv:2401.15422. [CrossRef]
23. Wang, K.; Zhu, J.; Ren, M.; Liu, Z.; Li, S.; Zhang, Z.; Zhang, C.; Wu, X.; Zhan, Q.; Liu, Q.; et al. A survey on data synthesis and augmentation for large language models. *arXiv* **2024**, arXiv:2410.12896. [CrossRef]
24. Tan, Z.; Beigi, A.; Wang, S.; Guo, R.; Bhattacharjee, A.; Jiang, B.; Karami, M.; Li, J.; Cheng, L.; Liu, H. Large language models for data annotation: A survey. *arXiv* **2024**, arXiv:2402.13446. [CrossRef]
25. Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. The rise and potential of large language model based agents: A survey. *Sci. China Inf. Sci.* **2025**, *68*, 121101. [CrossRef]
26. Hiniduma, K.; Byna, S.; Bez, J.L. Data readiness for AI: A 360-degree survey. *ACM Comput. Surv.* **2025**, *57*, 1–39. [CrossRef]
27. Zha, D.; Bhat, Z.P.; Lai, K.H.; Yang, F.; Jiang, Z.; Zhong, S.; Hu, X. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–42. [CrossRef]
28. Liang, W.; Tadesse, G.A.; Ho, D.; Fei-Fei, L.; Zaharia, M.; Zhang, C.; Zou, J. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat. Mach. Intell.* **2022**, *4*, 669–677. [CrossRef]
29. Cao, Y.; Hong, S.; Li, X.; Ying, J.; Ma, Y.; Liang, H.; Liu, Y.; Yao, Z.; Wang, X.; Huang, D.; et al. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv* **2025**, arXiv:2504.18838. [CrossRef]
30. Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; et al. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv* **2025**, arXiv:2501.09686.
31. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
32. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]

33. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
34. Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N.A.; Khashabi, D.; Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. *arXiv* **2022**, arXiv:2212.10560.
35. Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; Zhang, Y.; Zhenqiang Gong, N.; et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv* **2023**, arXiv:2306.04528. [CrossRef]
36. Estévez-Almenzar, M.; Baeza-Yates, R.; Castillo, C. A Comparison of Human and Machine Learning Errors in Face Recognition. *arXiv* **2025**, arXiv:2502.11337.
37. Wei, J.; Zhu, Z.; Luo, T.; Amid, E.; Kumar, A.; Liu, Y. To aggregate or not? learning with separate noisy labels. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 2523–2535.
38. Nasution, A.H.; Onan, A. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access* **2024**, *12*, 71876–71900. [CrossRef]
39. Mamat, N.; Othman, M.F.; Abdoulghafor, R.; Belhaouari, S.B.; Mamat, N.; Mohd Hussein, S.F. Advanced technology in agriculture industry by implementing image annotation technique and deep learning approach: A review. *Agriculture* **2022**, *12*, 1033. [CrossRef]
40. Nadisic, N.; Arhant, Y.; Vyncke, N.; Verplancke, S.; Lazendić, S.; Pižurica, A. A Deep Active Learning Framework for Crack Detection in Digital Images of Paintings. *Procedia Struct. Integr.* **2024**, *64*, 2173–2180. [CrossRef]
41. Newman, J.; Cox, C. Corpus annotation. In *A Practical Handbook of Corpus Linguistics*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 25–48.
42. Yang, C.; Sheng, L.; Wei, Z.; Wang, W. Chinese named entity recognition of epidemiological investigation of information on COVID-19 based on BERT. *IEEE Access* **2022**, *10*, 104156–104168. [CrossRef]
43. Kumar, M.P.; Tu, Z.X.; Chen, H.C.; Chen, K.C. Enhancing Learning in Fine-Tuned Transfer Learning for Rotating Machinery via Negative Transfer Mitigation. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2533613. [CrossRef]
44. Tu, S.; Sun, J.; Zhang, Q.; Lan, X.; Zhao, D. Online Preference-based Reinforcement Learning with Self-augmented Feedback from Large Language Model. *arXiv* **2024**, arXiv:2412.16878.
45. Acharya, D.B.; Kuppan, K.; Divya, B. Agentic AI: Autonomous Intelligence for Complex Goals–A Comprehensive Survey. *IEEE Access* **2025**, *13*, 18912–18936. [CrossRef]
46. Osakwe, I.; Chen, G.; Fan, Y.; Rakovic, M.; Singh, S.; Lim, L.; Van Der Graaf, J.; Moore, J.; Molenaar, I.; Bannert, M.; et al. Towards prescriptive analytics of self-regulated learning strategies: A reinforcement learning approach. *Br. J. Educ. Technol.* **2024**, *55*, 1747–1771. [CrossRef]
47. Bianchini, F.; Calamo, M.; De Luzi, F.; Macrì, M.; Marinacci, M.; Mathew, J.G.; Monti, F.; Rossi, J.; Leotta, F.; Mecella, M. SAMBA: A reference framework for Human-in-the-Loop in adaptive Smart Manufacturing. *Procedia Comput. Sci.* **2025**, *253*, 2257–2267. [CrossRef]
48. Karim, M.M.; Van, D.H.; Khan, S.; Qu, Q.; Kholodov, Y. AI Agents Meet Blockchain: A Survey on Secure and Scalable Collaboration for Multi-Agents. *Future Internet* **2025**, *17*, 57. [CrossRef]
49. Dong, X.; Zhang, X.; Bu, W.; Zhang, D.; Cao, F. A Survey of LLM-based Agents: Theories, Technologies, Applications and Suggestions. In Proceedings of the 2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC), Wuhan, China, 13–15 September 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 407–413.
50. Huang, Y. Levels of AI agents: From rules to large language models. *arXiv* **2024**, arXiv:2405.06643.
51. Boyina, K.; Reddy, G.M.; Akshita, G.; Nair, P.C. Zero-Shot and Few-Shot Learning for Telugu News Classification: A Large Language Model Approach. In Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7.
52. Faggioli, G.; Dietz, L.; Clarke, C.L.; Demartini, G.; Hagen, M.; Hauff, C.; Kando, N.; Kanoulas, E.; Potthast, M.; Stein, B.; et al. Perspectives on large language models for relevance judgment. In Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, Taipei, Taiwan, 23 July 2023; pp. 39–50.
53. Alizadeh, M.; Kubli, M.; Samei, Z.; Dehghani, S.; Zahedivafa, M.; Bermeo, J.D.; Korobeynikova, M.; Gilardi, F. Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *J. Comput. Soc. Sci.* **2025**, *8*, 17. [CrossRef] [PubMed]
54. Wang, X.; Kim, H.; Rahman, S.; Mitra, K.; Miao, Z. Human-llm collaborative annotation through effective verification of llm labels. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; pp. 1–21.
55. Zhang, Z.; Zhang, A.; Li, M.; Smola, A. Automatic chain of thought prompting in large language models. *arXiv* **2022**, arXiv:2210.03493. [CrossRef]
56. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. React: Synergizing reasoning and acting in language models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.

57. Groeneveld, J.; Herrmann, J.; Mollenhauer, N.; Dreeßen, L.; Bessin, N.; Tast, J.S.; Kastius, A.; Huegle, J.; Schlosser, R. Self-learning agents for recommerce markets. *Bus. Inf. Syst. Eng.* **2024**, *66*, 441–463. [CrossRef]
58. Ransiek, J.; Reis, P.; Sax, E. Adversarial and Reactive Traffic Agents for Realistic Driving Simulation. *arXiv* **2024**, arXiv:2409.14196. [CrossRef]
59. Monadjemi, S.; Guo, M.; Gotz, D.; Garnett, R.; Ottley, A. Human–Computer Collaboration for Visual Analytics: An Agent-based Framework. In Proceedings of the Computer Graphics Forum, Delft, The Netherlands, 28–30 June 2023; Wiley Online Library: Hoboken, NJ, USA, 2023; Volume 42, pp. 199–210.
60. Joshi, R.; Pandey, K.; Kumari, S.; Badola, R. Artificial Intelligence: A Gateway to the Twenty-First Century. In *The Intersection of 6G, AI/Machine Learning, and Embedded Systems*; CRC Press: Boca Raton, FL, USA, 2025; pp. 146–172.
61. Macedo, L. Artificial Intelligence Paradigms and Agent-Based Technologies. In *Human-Centered AI: An Illustrated Scientific Quest*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 363–397.
62. Roby, M. Learning and Reasoning Using Artificial Intelligence. In *Machine Intelligence*; Auerbach Publications: Abingdon-on-Thames, UK, 2023; pp. 237–256.
63. Sapkota, R.; Roumeliotis, K.I.; Karkee, M. AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenge. *arXiv* **2025**, arXiv:2505.10468. [CrossRef]
64. Rodríguez-Barroso, N.; Cámara, E.M.; Collados, J.C.; Luzón, M.V.; Herrera, F. Federated Learning for Exploiting Annotators' Disagreements in Natural Language Processing. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 630–648. [CrossRef]
65. Azeemi, A.H.; Qazi, I.A.; Raza, A.A. Language Model-Driven Data Pruning Enables Efficient Active Learning. *arXiv* **2024**, arXiv:2410.04275. [CrossRef]
66. Bayer, M.; Reuter, C. Activellm: Large language model-based active learning for textual few-shot scenarios. *arXiv* **2024**, arXiv:2405.10808.
67. Zhu, Q.; Mao, Q.; Zhang, J.; Huang, X.; Zheng, W. Towards a robust group-level emotion recognition via uncertainty-aware learning. *IEEE Trans. Affect. Comput.* **2025** . [CrossRef]
68. Mishra, S.; Shinde, M.; Yadav, A.; Ayyub, B.; Rao, A. An AI-Driven Data Mesh Architecture Enhancing Decision-Making in Infrastructure Construction and Public Procurement. *arXiv* **2024**, arXiv:2412.00224.
69. Puerta-Beldarrain, M.; Gómez-Carmona, O.; Sánchez-Corcuera, R.; Casado-Mansilla, D.; López-de Ipiña, D.; Chen, L. A multifaceted vision of the Human-AI collaboration: A comprehensive review. *IEEE Access* **2025**, *13*, 29375–29405. [CrossRef]
70. Zhang, R.; Li, Y.; Ma, Y.; Zhou, M.; Zou, L. Llmaaa: Making large language models as active annotators. *arXiv* **2023**, arXiv:2310.19596. [CrossRef]
71. Xia, Y.; Mukherjee, S.; Xie, Z.; Wu, J.; Li, X.; Aponte, R.; Lyu, H.; Barrow, J.; Chen, H.; Dernoncourt, F.; et al. From Selection to Generation: A Survey of LLM-based Active Learning. *arXiv* **2025**, arXiv:2502.11767.
72. Li, X.; Whan, A.; McNeil, M.; Starns, D.; Irons, J.; Andrew, S.C.; Suchecki, R. A Conceptual Framework for Human-AI Collaborative Genome Annotation. *arXiv* **2025**, arXiv:2503.23691. [CrossRef] [PubMed]
73. Croxford, E.; Gao, Y.; Pellegrino, N.; Wong, K.; Wills, G.; First, E.; Liao, F.; Goswami, C.; Patterson, B.; Afshar, M. Current and future state of evaluation of large language models for medical summarization tasks. *NPJ Health Syst.* **2025**, *2*, 6. [CrossRef] [PubMed]
74. Yang, L.; Sun, X.; Li, H.; Xu, R.; Wei, X. Difficulty aware programming knowledge tracing via large language models. *Sci. Rep.* **2025**, *15*, 11475. [CrossRef]
75. Sainz, O.; García-Ferrero, I.; Agerri, R.; de Lacalle, O.L.; Rigau, G.; Agirre, E. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv* **2023**, arXiv:2310.03668.
76. Bibal, A.; Gerlek, N.; Muric, G.; Boschee, E.; Fincke, S.C.; Ross, M.; Minton, S.N. Automating Annotation Guideline Improvements using LLMs: A Case Study. In Proceedings of the Context and Meaning: Navigating Disagreements in NLP Annotation, Abu Dhabi, UAE, 19 January 2025; pp. 129–144.
77. Rodler, P.; Shchekotykhin, K.; Fleiss, P.; Friedrich, G. RIO: Minimizing user interaction in ontology debugging. *arXiv* **2012**, arXiv:1209.3734. [CrossRef]
78. Zheng, J.; Shi, C.; Cai, X.; Li, Q.; Zhang, D.; Li, C.; Yu, D.; Ma, Q. Lifelong Learning of Large Language Model based Agents: A Roadmap. *arXiv* **2025**, arXiv:2501.07278. [CrossRef]
79. Zhang, G.; Liang, W.; Hsu, O.; Olukotun, K. Adaptive Self-improvement LLM Agentic System for ML Library Development. *arXiv* **2025**, arXiv:2502.02534.
80. Ashktorab, Z.; Pan, Q.; Geyer, W.; Desmond, M.; Danilevsky, M.; Johnson, J.M.; Dugan, C.; Bachman, M. Emerging Reliance Behaviors in Human-AI Text Generation: Hallucinations, Data Quality Assessment, and Cognitive Forcing Functions. *arXiv* **2024**, arXiv:2409.08937. [CrossRef]
81. Wang, X.; Hu, J.; Ali, S. MAATS: A Multi-Agent Automated Translation System Based on MQM Evaluation. *arXiv* **2025**, arXiv:2505.14848. [CrossRef]

82. Ara, Z.; Salemi, H.; Hong, S.R.; Senarath, Y.; Peterson, S.; Hughes, A.L.; Purohit, H. Closing the Knowledge Gap in Designing Data Annotation Interfaces for AI-powered Disaster Management Analytic Systems. In Proceedings of the 29th International Conference on Intelligent User Interfaces, Greenville, SC, USA, 18–21 March 2024; pp. 405–418.
83. Cronin, I. Autonomous AI agents: Decision-making, data, and algorithms. In *Understanding Generative AI Business Applications: A Guide to Technical Principles and Real-World Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 165–180.
84. Upadhyay, R.; Phlypo, R.; Saini, R.; Liwicki, M. Sharing to learn and learning to share; fitting together meta, multi-task, and transfer learning: A meta review. *IEEE Access* **2024**, *12*, 148553–148576. [CrossRef]
85. Liu, C.; Kang, Y.; Zhao, F.; Kuang, K.; Jiang, Z.; Sun, C.; Wu, F. Evolving knowledge distillation with large language models and active learning. *arXiv* **2024**, arXiv:2403.06414. [CrossRef]
86. Ding, B.; Qin, C.; Zhao, R.; Luo, T.; Li, X.; Chen, G.; Xia, W.; Hu, J.; Tuan, L.A.; Joty, S. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 1679–1705.
87. Törnberg, P. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv* **2023**, arXiv:2304.06588.
88. Choi, J.; Yun, J.; Jin, K.; Kim, Y. Multi-news+: Cost-efficient dataset cleansing via llm-based data annotation. *arXiv* **2024**, arXiv:2404.09682.
89. Nahum, O.; Calderon, N.; Keller, O.; Szpektor, I.; Reichart, R. Are LLMs Better than Reported? Detecting Label Errors and Mitigating Their Effect on Model Performance. *arXiv* **2024**, arXiv:2410.18889. [CrossRef]
90. Gat, Y.; Calderon, N.; Feder, A.; Chapanin, A.; Sharma, A.; Reichart, R. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv* **2023**, arXiv:2310.00603.
91. Kumar, S.; Datta, S.; Singh, V.; Datta, D.; Singh, S.K.; Sharma, R. Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access* **2024**, *12*, 75735–75760. [CrossRef]
92. Schleiger, E.; Mason, C.; Naughtin, C.; Reeson, A.; Paris, C. Collaborative Intelligence: A scoping review of current applications. *Appl. Artif. Intell.* **2024**, *38*, 2327890. [CrossRef]
93. Liu, Z.; Zhang, Y.; Li, P.; Liu, Y.; Yang, D. A dynamic LLM-powered agent network for task-oriented agent collaboration. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
94. Yang, J.; Ding, R.; Brown, E.; Qi, X.; Xie, S. V-irl: Grounding virtual intelligence in real life. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 36–55.
95. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-refine: Iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46534–46594.
96. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 8634–8652.
97. Li, D.; Li, Y.; Mekala, D.; Li, S.; Wang, X.; Hogan, W.; Shang, J. DAIL: Data Augmentation for In-Context Learning via Self-Paraphrase. *arXiv* **2025**, arXiv:2311.03319.
98. Bubeck, S.; Chadrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv* **2023**, arXiv:2303.12712. [CrossRef]
99. Chen, Y.; Si, M. Reflections & Resonance: Two-Agent Partnership for Advancing LLM-based Story Annotation. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 20–25 May 2024; pp. 13813–13818.
100. Cohen, R.; Hamri, M.; Geva, M.; Globerson, A. Lm vs lm: Detecting factual errors via cross examination. *arXiv* **2023**, arXiv:2305.13281. [CrossRef]
101. Bandlamudi, J.; Mukherjee, K.; Agarwal, P.; Chaudhuri, R.; Pimplikar, R.; Dechu, S.; Straley, A.; Ponniah, A.; Sindhgatta, R. Framework to enable and test conversational assistant for APIs and RPAs. *AI Mag.* **2024**, *45*, 443–456. [CrossRef]
102. Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S.K.S.; Lin, Z.; Zhou, L.; et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv* **2025**, arXiv:2308.00352.
103. Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; et al. Chatdev: Communicative agents for software development. *arXiv* **2023**, arXiv:2307.07924.
104. Lin, M.; Chen, Z.; Liu, Y.; Zhao, X.; Wu, Z.; Wang, J.; Zhang, X.; Wang, S.; Chen, H. Decoding Time Series with LLMs: A Multi-Agent Framework for Cross-Domain Annotation. *arXiv* **2024**, arXiv:2410.17462.
105. Alam, F.; Biswas, M.R.; Shah, U.; Zaghouani, W.; Mikros, G. Propaganda to Hate: A Multimodal Analysis of Arabic Memes with Multi-agent LLMs. In Proceedings of the International Conference on Web Information Systems Engineering, Doha, Qatar, 2–5 December 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 380–390.

106. Liu, W.; Chang, W.; Shi, C.; Wang, Y.; Hu, R.; Zhang, C.; Ouyang, H. Research on Intelligent Agent Technology and Applications Based on Large Models. In Proceedings of the 2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 6–8 December 2024; IEEE: Piscataway, NJ, USA, 2024; Volume 4, pp. 466–472.
107. Li, M.; Shi, T.; Ziems, C.; Kan, M.Y.; Chen, N.F.; Liu, Z.; Yang, D. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv* **2023**, arXiv:2310.15638.
108. Colucci Cante, L.; D'Angelo, S.; Di Martino, B.; Graziano, M. Text Annotation Tools: A Comprehensive Review and Comparative Analysis. In Proceedings of the International Conference on Complex, Intelligent, and Software Intensive Systems, Taichung, Taiwan, 3–5 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 353–362.
109. Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. Agentbench: Evaluating llms as agents. *arXiv* **2023**, arXiv:2308.03688. [CrossRef]
110. Verma, G.; Kaur, R.; Srishankar, N.; Zeng, Z.; Balch, T.; Veloso, M. Adaptagent: Adapting multimodal web agents with few-shot learning from human demonstrations. *arXiv* **2024**, arXiv:2411.13451.
111. Schmidt, S.; Stappen, L.; Schwinn, L.; Günnemann, S. Generalized Synchronized Active Learning for Multi-Agent-Based Data Selection on Mobile Robotic Systems. *IEEE Robot. Autom. Lett.* **2024**, *9*, 8659–8666. [CrossRef]
112. Wan, M.; Safavi, T.; Jauhar, S.K.; Kim, Y.; Counts, S.; Neville, J.; Suri, S.; Shah, C.; White, R.W.; Yang, L.; et al. Tnt-llm: Text mining at scale with large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 5836–5847.
113. Tutul, A.A.; Nirjhar, E.H.; Chaspari, T. Investigating trust in human-AI collaboration for a speech-based data analytics task. *Int. J. Hum. Comput. Interact.* **2025**, *41*, 2936–2954. [CrossRef]
114. Bolock, A.e.; Abouras, M.; Sabty, C.; Abdennadher, S.; Herbert, C. CARE: A Framework for Collecting and Annotating Emotions of Code-Switched Words. In Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems, Salamanca, Spain, 26–28 June 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 104–116.
115. Qu, Q.; Liu, S.; Zhu, F.; Jensen, C.S. Efficient online summarization of large-scale dynamic networks. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3231–3245. [CrossRef]
116. Hadian, A.; Nobari, S.; Minaei-Bidgoli, B.; Qu, Q. Roll: Fast in-memory generation of gigantic scale-free networks. In Proceedings of the 2016 International Conference on Management of Data, Francisco, CA, USA, 26 June–1 July 2016; pp. 1829–1842.
117. Chang, C.M.; He, Y.; Du, X.; Yang, X.; Xie, H. Dynamic labeling: A control system for labeling styles in image annotation tasks. In Proceedings of the International Conference on Human-Computer Interaction, Washington, DC, USA, 29 June–4 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 99–118.
118. Efrat, A.; Levy, O. The turking test: Can language models understand instructions? *arXiv* **2020**, arXiv:2010.11982. [CrossRef]
119. Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; Singh, S. Calibrate before use: Improving few-shot performance of language models. In Proceedings of the International Conference on Machine Learning. PMLR, Online, 18–24 July 2021; pp. 12697–12706.
120. Smith, A.G.; Han, E.; Petersen, J.; Olsen, N.A.F.; Giese, C.; Athmann, M.; Dresbøll, D.B.; Thorup-Kristensen, K. RootPainter: Deep learning segmentation of biological images with corrective annotation. *New Phytol.* **2022**, *236*, 774–791. [CrossRef] [PubMed]
121. Kim, H.; Hessel, J.; Jiang, L.; West, P.; Lu, X.; Yu, Y.; Zhou, P.; Bras, R.L.; Alikhani, M.; Kim, G.; et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv* **2022**, arXiv:2212.10465.
122. Ho, N.; Schmid, L.; Yun, S.Y. Large language models are reasoning teachers. *arXiv* **2022**, arXiv:2212.10071.
123. Wu, T.; Yuan, W.; Golovneva, O.; Xu, J.; Tian, Y.; Jiao, J.; Weston, J.; Sukhbaatar, S. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv* **2024**, arXiv:2407.19594.
124. Kang, H.J.; Harel-Canada, F.; Gulzar, M.A.; Peng, V.; Kim, M. Human-in-the-Loop Synthetic Text Data Inspection with Provenance Tracking. *arXiv* **2024**, arXiv:2404.18881.
125. Wu, J.; Deng, J.; Pang, S.; Chen, Y.; Xu, J.; Li, X.; Xu, W. Legilimens: Practical and unified content moderation for large language model services. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, Salt Lake City, UT, USA, 14–18 October 2024; pp. 1151–1165.
126. Palla, K.; García, J.L.R.; Hauff, C.; Fabbri, F.; Lindström, H.; Taber, D.R.; Damianou, A.; Lalmas, M. Policy-as-Prompt: Rethinking Content Moderation in the Age of Large Language Models. *arXiv* **2025**, arXiv:2502.18695.
127. Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; Liu, Q. Aligning large language models with human: A survey. *arXiv* **2023**, arXiv:2307.12966. [CrossRef]
128. Wu, S.; Fung, M.; Qian, C.; Kim, J.; Hakkani-Tur, D.; Ji, H. Aligning LLMs with Individual Preferences via Interaction. *arXiv* **2024**, arXiv:2410.03642. [CrossRef]
129. Huang, M.; Jiang, Q.; Qu, Q.; Chen, L.; Chen, H. Information fusion oriented heterogeneous social network for friend recommendation via community detection. *Appl. Soft Comput.* **2022**, *114*, 108103. [CrossRef]
130. Bojić, L.; Zagovora, O.; Zelenkauskaite, A.; Vuković, V.; Čabarkapa, M.; Veseljević Jerković, S.; Jovančević, A. Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Sci. Rep.* **2025**, *15*, 11477. [CrossRef]

131. Harrer, S.; Rane, R.V.; Speight, R.E. Generative AI agents are transforming biology research: High resolution functional genome annotation for multiscale understanding of life. *EBioMedicine* **2024**, *109*, 105446. [CrossRef] [PubMed]
132. Toubal, I.E.; Avinash, A.; Alldrin, N.G.; Dlabal, J.; Zhou, W.; Luo, E.; Stretcu, O.; Xiong, H.; Lu, C.T.; Zhou, H.; et al. Modeling collaborator: Enabling subjective vision classification with minimal human effort via llm tool-use. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 17553–17563.
133. Beck, J.; Kemeter, L.M.; Dürrbeck, K.; Abdalla, M.H.I.; Kreuter, F. Towards Integrating ChatGPT into Satellite Image Annotation Workflows. A Comparison of Label Quality and Costs of Human and Automated Annotators. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 4366–4381. [CrossRef]
134. Qu, Q.; Liu, S.; Yang, B.; Jensen, C.S. Efficient top-k spatial locality search for co-located spatial web objects. In Proceedings of the 2014 IEEE 15th International Conference on Mobile Data Management, Brisbane, Australia, 14–18 July 2014; IEEE: Piscataway, NJ, USA, 2014; Volume 1, pp. 269–278.
135. Cao, X.; Chen, L.; Cong, G.; Jensen, C.S.; Qu, Q.; Skovsgaard, A.; Wu, D.; Yiu, M.L. Spatial keyword querying. In Proceedings of the Conceptual Modeling: 31st International Conference ER 2012, Florence, Italy, 15–18 October 2012; Proceedings 31; Springer: Berlin/Heidelberg, Germany, 2012; pp. 16–29.
136. Tsiakas, K.; Murray-Rust, D. Unpacking Human-AI interactions: From interaction primitives to a design space. *ACM Trans. Interact. Intell. Syst.* **2024**, *14*, 1–51. [CrossRef]
137. Yuan, H. Agentic Large Language Models for Healthcare: Current Progress and Future Opportunities. *Med. Adv.* **2025**, *3*, 37–41. [CrossRef]
138. Qu, Q.; Chen, C.; Jensen, C.S.; Skovsgaard, A. Space-Time Aware Behavioral Topic Modeling for Microblog Posts. *IEEE Data Eng. Bull.* **2015**, *38*, 58–67.
139. Kim, H.; Mitra, K.; Chen, R.L.; Rahman, S.; Zhang, D. Meganno+: A human-llm collaborative annotation system. *arXiv* **2024**, arXiv:2402.18050.
140. El Khoury, K.; Godelaine, T.; Delvaux, S.; Lugan, S.; Macq, B. Streamlined hybrid annotation framework using scalable codestream for bandwidth-restricted uav object detection. In Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, UAE, 27–30 October 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1581–1587.
141. Chen, Z.Z.; Ma, J.; Zhang, X.; Hao, N.; Yan, A.; Nourbakhsh, A.; Yang, X.; McAuley, J.; Petzold, L.; Wang, W.Y. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv* **2024**, arXiv:2405.01769. [CrossRef]
142. Lazo, G.R.; Ayyappan, D.; Sharma, P.K.; Tiwari, V.K. Contextual Science and Genome Analysis for Air-Gapped AI Research. *bioRxiv* **2025**. [CrossRef]
143. Olawore, K.; McTear, M.; Bi, Y. Development and Evaluation of a University Chatbot Using Deep Learning: A RAG-Based Approach. In Proceedings of the International Symposium on Chatbots and Human-Centered AI, Thessaloniki, Greece, 4–5 December 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 96–111.
144. Li, J. A comparative study on annotation quality of crowdsourcing and LLM via label aggregation. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 14–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 6525–6529.
145. Zhou, Y.; Cheng, X.; Zhang, Q.; Wang, L.; Ding, W.; Xue, X.; Luo, C.; Pu, J. ALGPT: Multi-Agent Cooperative Framework for Open-Vocabulary Multi-Modal Auto-Annotating in Autonomous Driving. *IEEE Trans. Intell. Veh.* **2024**, 1–15. [CrossRef]
146. Mots' oehli, M. Assistive Image Annotation Systems with Deep Learning and Natural Language Capabilities: A Review. In Proceedings of the 2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 23–25 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–9.
147. Mazhar, A.; Shaik, Z.H.; Srivastava, A.; Ruhnke, P.; Vaddavalli, L.; Katragadda, S.K.; Yadav, S.; Akhtar, M.S. Figurative-cum-Commonsense Knowledge Infusion for Multimodal Mental Health Meme Classification. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 637–648.
148. Sandhu, R.; Channi, H.K.; Ghai, D.; Cheema, G.S.; Kaur, M. An introduction to generative AI tools for education 2030. *Integr. Gener. Educ. Achieve Sustain. Dev. Goals* **2024**, 1–28. [CrossRef]
149. Ming, X.; Li, S.; Li, M.; He, L.; Wang, Q. AutoLabel: Automated Textual Data Annotation Method Based on Active Learning and Large Language Model. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Birmingham, UK, 16–18 August 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 400–411.
150. Krishnan, N. Advancing Multi-Agent Systems Through Model Context Protocol: Architecture, Implementation, and Applications. *arXiv* **2025**, arXiv:2504.21030. [CrossRef]
151. Aejas, B.; Belhi, A.; Bouras, A. Toward an nlp approach for transforming paper contracts into smart contracts. In *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 751–759.
152. Kastrati, M.; Imran, A.S.; Hashmi, E.; Kastrati, Z.; Daudpota, S.M.; Biba, M. Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with Explainable Artificial Intelligence. *Eng. Appl. Artif. Intell.* **2025**, *149*, 110136. [CrossRef]

153. Raza, S.; Chatrath, V. HarmonyNet: Navigating hate speech detection. *Nat. Lang. Process. J.* **2024**, *8*, 100098. [CrossRef]
154. Khanduja, N.; Kumar, N.; Chauhan, A. Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. *Syst. Soft Comput.* **2024**, *6*, 200112. [CrossRef]
155. Kao, J.P.; Kao, H.T. Large Language Models in radiology: A technical and clinical perspective. *Eur. J. Radiol. Artif. Intell.* **2025**, *2*, 100021. [CrossRef]
156. Ostrovsky, A.M. Evaluating a large language model's accuracy in chest X-ray interpretation for acute thoracic conditions. *Am. J. Emerg. Med.* **2025**, *93*, 99–102. [CrossRef]
157. Almalky, A.M.A.; Zhou, R.; Angizi, S.; Rakin, A.S. How Vulnerable are Large Language Models (LLMs) against Adversarial Bit-Flip Attacks? In Proceedings of the Great Lakes Symposium on VLSI 2025, New Orleans, LA, USA, 30 June–2 July 2025; pp. 534–539.
158. Zhang, L.; Zou, Q.; Singhal, A.; Sun, X.; Liu, P. Evaluating large language models for real-world vulnerability repair in c/c++ code. In Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics, Porto, Portugal, 21 June 2024; pp. 49–58.
159. Heo, S.; Son, S.; Park, H. HaluCheck: Explainable and verifiable automation for detecting hallucinations in LLM responses. *Expert Syst. Appl.* **2025**, *272*, 126712. [CrossRef]
160. Ferrag, M.A.; Alwahedi, F.; Battah, A.; Cherif, B.; Mechri, A.; Tihanyi, N.; Bisztray, T.; Debbah, M. Generative ai in cybersecurity: A comprehensive review of llm applications and vulnerabilities. *Internet Things-Cyber-Phys. Syst.* **2025**, *5*, 1–46. [CrossRef]
161. Kaushik, D.; Lipton, Z.C.; London, A.J. Resolving the Human-Subjects Status of ML's Crowdworkers. *Commun. ACM* **2024**, *67*, 52–59. [CrossRef]
162. Reif, Y.; Schwartz, R. Beyond performance: Quantifying and mitigating label bias in llms. *arXiv* **2024**, arXiv:2405.02743. [CrossRef]
163. Feretzakis, G.; Papaspyridis, K.; Gkoulalas-Divanis, A.; Verykios, V.S. Privacy-preserving techniques in generative ai and large language models: A narrative review. *Information* **2024**, *15*, 697. [CrossRef]
164. Ullah, I.; Hassan, N.; Gill, S.S.; Suleiman, B.; Ahanger, T.A.; Shah, Z.; Qadir, J.; Kanhere, S.S. Privacy preserving large language models: Chatgpt case study based vision and framework. *IET Blockchain* **2024**, *4*, 706–724. [CrossRef]
165. Templin, T.; Fort, S.; Padmanabham, P.; Seshadri, P.; Rimal, R.; Oliva, J.; Hassmiller Lich, K.; Sylvia, S.; Sinnott-Armstrong, N. Framework for bias evaluation in large language models in healthcare settings. *NPJ Digit. Med.* **2025**, *8*, 414. [CrossRef]
166. Sun, L.; Liu, D.; Wang, M.; Han, Y.; Zhang, Y.; Zhou, B.; Ren, Y.; Zhu, P. Taming unleashed large language models with blockchain for massive personalized reliable healthcare. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 4498–4511. [CrossRef] [PubMed]
167. Moreno-Sánchez, P.A.; Del Ser, J.; van Gils, M.; Hernesniemi, J. A Design Framework for operationalizing Trustworthy Artificial Intelligence in Healthcare: Requirements, Tradeoffs and Challenges for its Clinical Adoption. *arXiv* **2025**, arXiv:2504.19179. [CrossRef]
168. Mienye, I.D.; Obaido, G.; Jere, N.; Mienye, E.; Aruleba, K.; Emmanuel, I.D.; Ogbuokiri, B. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Inform. Med. Unlocked* **2024**, *51*, 101587. [CrossRef]
169. Törnberg, P. Best practices for text annotation with large language models. *arXiv* **2024**, arXiv:2402.05129. [CrossRef]
170. Khan, S.; Qiming, H. GPU-accelerated homomorphic encryption computing: Empowering federated learning in IoV. *Neural Comput. Appl.* **2025**, *37*, 10351–10380. [CrossRef]
171. Xie, T.; Harel, D.; Ran, D.; Li, Z.; Li, M.; Yang, Z.; Wang, L.; Chen, X.; Zhang, Y.; Zhang, W.; et al. Data and System Perspectives of Sustainable Artificial Intelligence. *arXiv* **2025**, arXiv:2501.07487. [CrossRef]
172. Dai, X.; Li, J.; Liu, X.; Yu, A.; Lui, J. Cost-effective online multi-llm selection with versatile reward models. *arXiv* **2024**, arXiv:2405.16587.
173. Jiang, Y.; Wang, H.; Xie, L.; Zhao, H.; Qian, H.; Lui, J. D-llm: A token adaptive computing resource allocation strategy for large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 1725–1749.
174. Li, J.; Han, B.; Li, S.; Wang, X.; Li, J. Collm: A collaborative llm inference framework for resource-constrained devices. In Proceedings of the 2024 IEEE/CIC International Conference on Communications in China (ICCC), Hangzhou, China, 7–9 August 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 185–190.
175. Lang, J.; Guo, Z.; Huang, S. A comprehensive study on quantization techniques for large language models. In Proceedings of the 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), Xiamen, China, 27–29 December 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 224–231.
176. An, Y.; Zhao, X.; Yu, T.; Tang, M.; Wang, J. Fluctuation-based adaptive structured pruning for large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 26–27 February 2024; Volume 38, pp. 10865–10873.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.