Harvard Business School

N9-826-218
May 13, 2026

Zoe Cullen
Nicole Tempest Keller

Humans in the Loop (HITL) AI

Data formed the foundation of the AI industry. While advances in algorithms and computing power had played a major role in improving AI capabilities, human input had also been critical. (See **Exhibit 1**.) As AI adoption accelerated across industries, a growing ecosystem emerged to supply this human contribution at scale—operating through a variety of business models and creating new job opportunities across the globe.

The performance of AI systems depended heavily on access to large volumes of accurate, well-structured training data. This need gave rise to the "data annotation" sector which became an essential part of the AI value chain. Data annotation referred to the work of adding labels or explanations to raw data—such as images, text, audio, or video—so that AI systems could learn what the data represented and how to respond to it.

Over time however, simply increasing the quantity of basic data had become less effective. Major AI developers had already scraped and indexed much of the world's publicly available text, making further gains more difficult—a phenomenon referred to as the "data wall."[1] As a result, progress depended less on adding more data and more on new forms of human input that improved data quality, embedded judgment, and captured context. The result was the emergence of a broader market known as human-in-the-loop (HITL) AI—a hybrid approach that combined machine learning with human intelligence to improve model decision making, especially in complex or ambiguous situations where fully automated systems did not perform well.[2] HITL enabled the creation of highly curated datasets, known as "frontier data" that pushed model quality even further.[3]

[1] "AI data wall: Why experts predict AI slowdown and how to break through the plateau [enterprise guide]," SuperAnnotate, November 23, 2024, https://www.superannotate.com/blog/ai-data-wall#:~:text=The%20paths%20to%20SuperData%20vary,happily%20provide%20to%20our%20customers., accessed February 11, 2026.

[2] Jack Lasky, "Human-in-the-loop (HITL)," EBSCO, 2025, https://www.ebsco.com/research-starters/computer-science/human-loop-hitl, accessed February 10, 2025.

[3] David George and Alexandr Wang, "Unlocking AI's Future: Alexandr Wang on the Power of Frontier Data," A16z, September 24, 2024, https://a16z.com/frontier-data-foundries-alex-wang-scale-ai/, accessed February 11, 2026.

Professor Zoe Cullen and Nicole Tempest Keller (California Research Center) prepared this note as the basis for class discussion.

The Human in the Loop Industry

HITL referred to AI systems that depended on human input during training, evaluation, and deployment. Humans provided labels, feedback, and expert judgment that guided model behavior, corrected errors, and stepped in when AI systems encountered ambiguity, edge cases, or decisions that required human judgment. The need for HITL spanned both foundation models and enterprise models. The HITL AI sector was projected to reach $17.1 billion by 2030.[4]

Segments

The HITL industry was divided into several segments: data annotation, reinforcement learning from human feedback (RLHF), model evaluation and safety, and domain experts.

Data annotation: The most basic workflow involved data annotation, which played a key role in data "preprocessing"—the step in which raw and unstructured data was transformed into a clean, structured format suitable for training models. For example, for text, annotation involved capturing intent, sentiment, or topic so that LLMs could learn human language. For images and video, annotation involved labeling and tracking objects so computer vision models could recognize visual elements. Improved image annotations were credited with the enhanced image quality in Open AI's DALL-E 3 model over its predecessor DALL-E 2. OpenAI researcher Gabriel Goh noted, "I think this [annotations] is the main source of the improvements. The text annotations are a lot better than they were [with DALL-E 2]—it's not even comparable."[5] In healthcare, annotated medical images like X-rays, MRIs, and CT scans were essential to train AI models to detect anomalies like tumors, fractures, or infections.[6] Audio annotation involved transcribing speech and tagging sounds (i.e., a car horn, laughter, dog barking).

In 2024, the data annotation industry alone contributed an estimated $5.7 billion to U.S. GDP, comparable in size to established manufacturing sectors such as computer storage devices ($4.2 billion) and computer terminals ($6.5 billion).[7]

Reinforcement learning: More advanced work involved reinforcement learning from human feedback (RLHF), where people reviewed AI-generated answers, compared multiple responses to the same prompt, and judged which were clearer, more accurate, or more appropriate—teaching the system how humans expected it to behave. Some AI labs were also experimenting with "reinforcement gyms" in which models learned how to complete tasks within virtual copies of apps like Salesforce or Amazon.[8] For example, the models might learn how to place an order on Amazon

[4] The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf, accessed January 24, 2026.

[5] Kyle Wiggers, "AI training data has a price tag that only big tech can afford," *TechCrunch*, February 10, 2026, https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford/, accessed February 10, 2026.

[6] The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf, accessed Feburary 10, 2026.

[7] The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf, accessed January 24, 2026.

[8] Stephanie Palazzolo, "Where Reinforcement Learning is Going," *The Information*, July 22, 2025, https://www.theinformation.com/newsletters/ai-agenda/reinforcement-learning-going, accessed January 26, 2026.

or log a new lead in Salesforce without human involvement.[9] By learning on a handful of apps, the models were then able to figure out how to operate in other apps they had not been trained on.

Model evaluation and safety: Another segment of the HITL market focused on model evaluation and safety, which involved testing how AI systems behaved under real-world and adversarial conditions. Human evaluators reviewed model outputs to identify failures, such as hallucinations, biased responses, or unsafe instructions. These were issues that automated testing often failed to detect. This work often involved "red-teaming" in which human specialists actively tried to break AI systems by probing for edge cases and misuse scenarios. For example, OpenAI used extensive red-teaming to uncover safety risks before launching GPT-4.[10] On the model evaluation side, each time models were updated or adjusted, human reviewers needed to check whether those changes unintentionally degraded performance or produced new, hard-to-detect issues. Doing this well required judgment and contextual understanding, making the work more specialized than routine annotation.

Domain experts: At the highest end of the HITL market, domain experts—such as physicians, attorneys, engineers, and mathematicians—applied professional judgment to review and refine AI outputs in specialized fields. Unlike general annotators or evaluators, these contributors were not simply labeling or ranking responses; they were assessing whether model outputs met the standards of a regulated or technically demanding profession. For example, a medical expert might evaluate whether a model's diagnostic reasoning aligned with clinical best practices. A legal expert might assess whether a contract analysis reflected accepted legal interpretations. In mathematics and coding, experts reviewed proofs or complex programs to ensure logical correctness. This work demonstrated what "correct" looked like in domains where errors could carry legal, financial, or safety consequences. Because it relied on deep subject-matter expertise and accountability, domain expert involvement was among the most specialized—and most costly—forms of human-in-the-loop contribution.

People and Pay

The data annotation segment alone directly supported earning opportunities for nearly 200,000 people and an additional 9,000 full-time jobs in the U.S.[11] Approximately 84% of the data annotation workforce held a bachelor's degree or higher—double the share of the U.S. workforce.[12] (See **Exhibit 2**.) Most participated part-time alongside other professional, academic, or caregiving responsibilities, using annotation work to supplement income. (See **Exhibit 3**.) Annotation also served as an upskilling opportunity to enter into AI-related work; over 75% of annotators hoped to apply their new skills to a future position in tech or AI.[13]

---

[9] Stephanie Palazzolo, "Where Reinforcement Learning is Going," *The Information*, July 22, 2025, https://www.theinformation.com/newsletters/ai-agenda/reinforcement-learning-going, accessed January 26, 2026.

[10] Madhumita Murgia, "OpenAI's red team: the experts hired to 'break' ChatGPT," Financial Times, April 14, 2023, https://www.cmu.edu/ambassadors/july-2023/pdf/ft-open-ai_04-14-2023.pdf, accessed February 10, 2026.

[11] The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf, accessed January 24, 2026.

[12] The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf, accessed January 24, 2026.

[13] The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf, accessed January 24, 2026.

Compensation varied widely by segment and task. While simple labeling or validation work paid modest hourly rates, evaluators and scarce domain experts were compensated at significantly higher levels, reflecting the judgment, accountability, and expertise required. (See **Exhibit 4**.)

HITL Companies

There were different types of players in the HITL industry. Large-scale HITL platforms such as Scale AI, valued at $29 billion in 2024,[14] and Appen, which coordinated more than one million annotators across over 170 countries, focused on providing labeled data and evaluation pipelines[15] to support AI model development. In essence, they sold the *output* to labs and abstracted away the process, by recruiting and managing large, contracted workforces themselves.

A different model was a class of companies such as Mercor, which operated as talent marketplaces for specialized knowledge work. Mercor matched companies with domain experts, but the customer was responsible for managing the actual work. Mercor took a percentage of the compensation paid to those experts, reportedly around 30% of a candidate's salary or contract value.[16]

Historically, the majority of spending in the HITL market came from large AI labs, which collectively invested several billion dollars annually in data labeling, evaluation, and feedback. This buyer concentration gave labs significant pricing leverage. However, as AI models improved, high-quality human feedback had become a critical bottleneck, prompting leading labs to bring some of this work in-house or secure it through exclusive partnerships. OpenAI was actively hiring domain experts for its Strategic Deployment team[17] and Meta's $14.3 billion investment in Scale AI[18] highlighted this move toward vertical integration—giving labs more control and squeezing margins for independent data annotation providers.

As labs consolidated capabilities and internalized key functions, attention increasingly turned toward enterprises, where AI adoption was accelerating but operationalization remained incomplete.

Enterprise Challenges with AI

By the mid-2020s, frontier language models such as GPT-4, Claude, and Gemini had demonstrated remarkable progress in language understanding, coding, and reasoning tasks. Trained on vast quantities of data, these systems appeared adaptable to a wide range of enterprise use cases. Organizations rapidly launched AI pilots across functions, expecting productivity gains and measurable cost savings.

Yet early enthusiasm masked operational reality. According to MIT's *GenAI Divide: State of AI in Business 2025* report, despite $30 to $40 billion in enterprise spending on generative AI and broad

---

[14] Dina Bass, "Scale AI Rival Invisible Technologies Valued at Over $2 Billion," *Bloomberg*, September 16, 2025, https://www.bloomberg.com/news/articles/2025-09-16/scale-ai-rival-invisible-technologies-valued-at-over-2-billion, accessed January 25, 2026.

[15] A set of automated and human-in-the-loop processes used to repeatedly test an AI model's performance, accuracy, and safety against standard scenarios and edge cases as models, data, or prompts changed.

[16] "How Does Mercor Make Money," Canvas Business Model, July 12, 2025, https://canvasbusinessmodel.com/blogs/how-it-works/mercor-how-it-works, accessed January 25, 2026.

[17] OpenAI, Careers, https://openai.com/careers/research-scientist-mathematical-sciences-san-francisco/, accessed January 26, 2026.

[18] Kalley Huang Cory Weinberg, "Meta Finalizes $14.3 Billion Deal with Scale AI," *The Information,* June 13, 2025, https://www.theinformation.com/briefings/meta-finalizes-14-3-billion-deal-scale-ai, accessed June 26, 2025.

adoption of AI tools (e.g. ChatGPT) in the workplace among employees, only about 5% of enterprise AI pilots succeeded in producing measurable profit-and-loss impact or reaching scaled production deployment; 95% stalled in experimentation due to integration and workflow challenges.[19] This gap was referred to as the "GenAI Divide." Looking ahead, the outlook was equally sobering. In July 2025, Gartner, a global technology research and advisory firm, forecasted that over 40% of enterprise agentic AI projects would be canceled by the end of 2027, due to escalating costs, unclear business value, or inadequate risk controls.[20]

The problem lay in the fact that when enterprises tried to deploy general purpose LLMs for actual business operations, they quickly discovered critical gaps. A typical model did not know the specific configurations of a manufacturing company's product catalog, an insurance company's proprietary underwriting guidelines refined over decades, a law firm's approach to structuring merger agreements based on thousands of past deals, a hospital system's protocols for handling medication interactions in patients with complex comorbidities, or a logistics company's routing constraints shaped by its fleet, customers, and service-level agreements. This kind of knowledge was embedded in internal systems and day-to-day decision-making, not in the public data used to train general-purpose models.

General-purpose LLMs also lacked awareness of the here and now. LLMs were trained on *historical, public* data, not on the living, breathing operational reality of a specific enterprise. As one enterprise AI architect put it: "Asking a general LLM to help with our business is like asking someone who's read every book about restaurants to run your kitchen. They have great conceptual knowledge, but they don't know your recipes, your equipment, your suppliers, your staff, or what happened during last night's dinner service."

The Full Context Problem

Successful AI deployment in enterprises required access to three types of context that general-purpose LLMs lacked, referred to as the "full context problem."

First was *proprietary domain knowledge*. This was the specialized understanding developed within an organization over years or decades. This included decision frameworks that had been refined through trial and error, edge case handling based on past failures, undocumented "tribal knowledge" passed down among experienced employees, and company-specific terminology, acronyms, and classification systems.

Second was *process integration*. This meant understanding how work actually flowed through the organization. This meant knowing which systems talked to which other systems, where manual handoffs occurred, what checks and approvals were required, and how exceptions were escalated and resolved.

Third was current *operational state*. This was real-time and recent data about what was happening right now. Enterprises needed AI that understood current inventory levels and supply chain status, recent customer interactions and satisfaction trends, latest performance metrics and deviation alerts, and emerging issues flagged by frontline workers.

---

[19] Challapally, A., Pease, C., Raskar, R. and Chari, P., 2025. "The GenAI divide: State of AI in business 2025," MIT Nanda, https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf, accessed January 24, 2026

[20] "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by the End of 2027," Gartner, June 25, 2025, https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027, accessed January 22, 2026.

The Paradox: Human in the Loop—But Training Your Replacement?

To build enterprise specific AI systems that outperformed general LLMs, companies relied heavily on HITL processes. Employees provided domain expertise, institutional knowledge of how things were done, and their decision logic, while forward deployed engineers and data specialists translated that knowledge into agentic workflows. This created a fundamental psychological and economic tension: employees might be documenting their way out of a job.

Unlike external expert annotators who contributed to training foundation models—and whose individual contribution was diluted across millions of training examples—enterprise employees were being asked to systematically codify the specialized knowledge that made them valuable to their employer.

Employee Concerns

Research on technology adoption consistently highlighted job insecurity as a primary barrier.[21] In the context of generative AI, these fears were not hypothetical. A 2025 KPMG American Worker Survey found that 52% of employees were concerned that AI could replace their jobs.[22] In the same year it was reported that AI had been responsible for roughly 55,000 announced layoffs, a figure widely believed to understate the true impact.[23] Separate research sponsored by SAP found that 42% of finance leaders viewed headcount reduction as the clearest way to demonstrate a return on investment from AI spending.[24] Sam Altman also raised alarm bells on this topic. Speaking to the Federal Reserve Board of Governors in 2025, he noted that some job categories would be severely impacted by AI: "Some areas, again, I think just like totally, totally gone."[25] Echoing this sentiment, Jim Farley, CEO of Ford, speaking at the Aspen Ideas Festival in 2025, offered a sobering prediction, "Artificial intelligence is gonna replace literally half of all white-collar workers in the U.S."[26]

Employees sometimes correctly perceived that thoroughly documenting their expertise made them more replaceable. Why should an underwriter share the pattern recognition skills developed over 20 years if the result would be an AI system that made junior employees equally effective—or eliminated the need for the role entirely? A senior claims adjuster at an insurance client put it bluntly: "They want me to train the system on how I evaluate complex claims. But that expertise is why I earn six figures. Once the AI can do it, what's my value?" A translation annotator commented, "The more

---

[21] Taewoo Nam, "Technology usage, expected job sustainability, and perceived job insecurity," Technological Forecasting and Social Change Journal, January 2019, https://www.sciencedirect.com/science/article/abs/pii/S0040162517310387?via%3Dihub, accessed February 11, 2026.

[22] "American Workers are Leading the AI Revolution Despite Questions Around New Career Paths," KPMG, November 24, 2025, https://kpmg.com/us/en/media/news/american-workers-leading-ai-revolution.html, accessed February 10, 2026.

[23] Roberto Torres, "Workers worry about AI job loss amid enterprise adoption," CIODive, February 24, 2026, https://www.ciodive.com/news/workforce-AI-trust-upskilling-CIO/811399/, accessed February 10, 2026.

[24] Alexei Alexis, "Fear of AI-driven job displacement nearly doubles in a year: KPMG," CIODive, November 26, 2025, https://www.ciodive.com/news/AI-job-displacement-kpmg-tech-talent/806468/, accessed February 10, 2026.

[25] Joseph Gedeon, "OpenAI CEO tells Federal Reserve confab that entire job categories will disappear due to AI," *The Guardian,* July 22, 2025, https://www.theguardian.com/technology/2025/jul/22/openai-sam-altman-congress-ai-jobs, accessed February 11, 2026.

[26] Jason Ma, "Ford CEO Jim Farley warns AI will wipe out half of white-collar jobs, but the 'essential economy' has a huge shortage of workers," *Fortune,* July 5, 2025, https://fortune.com/2025/07/05/ford-ceo-jim-farley-ai-white-collar-jobs-essential-economy-skilled-trade-jobs-shortage/, accessed February 11, 2026.

it learns, the more obsolete you become. You're essentially expected to dig your own professional grave."[27]

Beyond economic concerns, expertise often formed core professional identity. Master craftsmen, senior diagnosticians, and expert troubleshooters derived status and satisfaction from being the person others turned to for difficult problems. An AI system that could handle those problems threatened not just their income but their sense of who they were professionally.

Training AI systems also required substantial effort beyond normal job duties. Yet the primary beneficiaries were the company (through efficiency gains) and potentially the AI vendor. Individual employees rarely saw direct compensation or career advancement. Employees had to participate in extensive interviews about their decision processes, review and label thousands of examples, provide detailed explanations of their reasoning, and correct AI mistakes and refine training data. A survey published in 2026 by AI orchestration platform, Zapier, found that the average U.S. employee spent 4.5 hours per week—more than half a work day—revising, correcting, and sometimes completely redoing AI output.[28] Referred to as AI "workslop" this AI output looked polished at first, but lacked the substance or nuance required to complete the task, creating a layer of hidden work.[29]

Where To Go From Here

Ultimately, the success of enterprise AI might hinge less on technical breakthroughs and more on whether companies could secure the cooperation of the employees whose expertise made these systems perform well. At a time when executives openly discussed the linkage between AI investment and headcount reductions, employees were being asked to help codify the very knowledge that gave them value. Human-in-the-loop systems sat at the center of that tension. Whether AI ultimately augmented workers or displaced them would depend not only on model performance, but on how organizations structured incentives, shared benefits, and built trust.

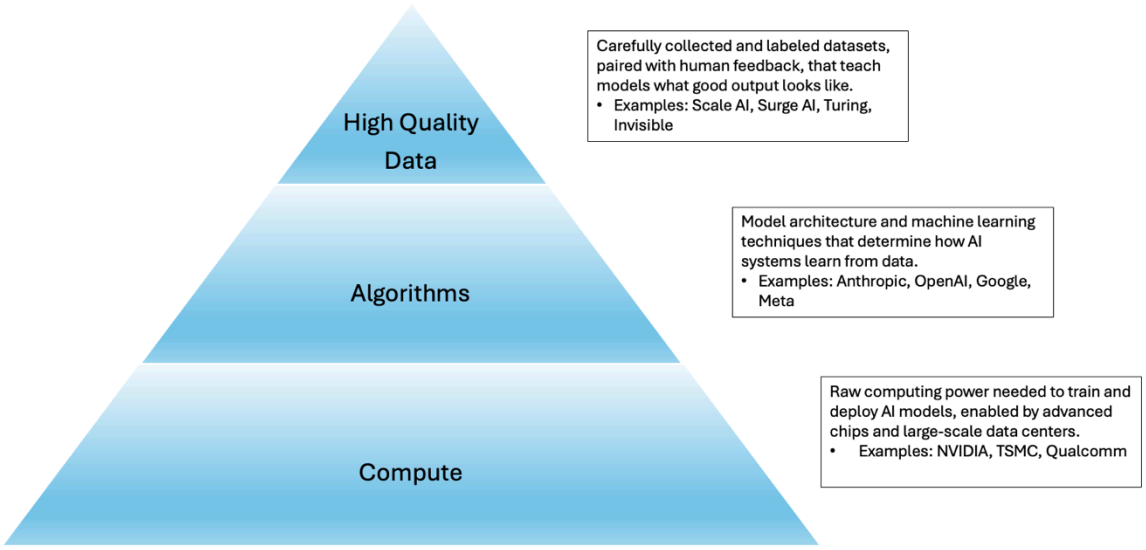
---

[27] Leanne Kolirin, "Like digging 'your own professional grave': The translators grappling with losing work to AI," CNN, January 23, 2026, https://www.cnn.com/2026/01/23/tech/translation-language-jobs-ai-automation-intl, accessed February 11, 2026.

[28] "Zapier Survey Finds Workers Spend 4.5 Hours Per Week Cleaning Up AI Mistakes," Yahoo Finance, January 14, 2026, https://finance.yahoo.com/news/zapier-survey-finds-workers-spend-130000740.html, accessed February 11, 2026.

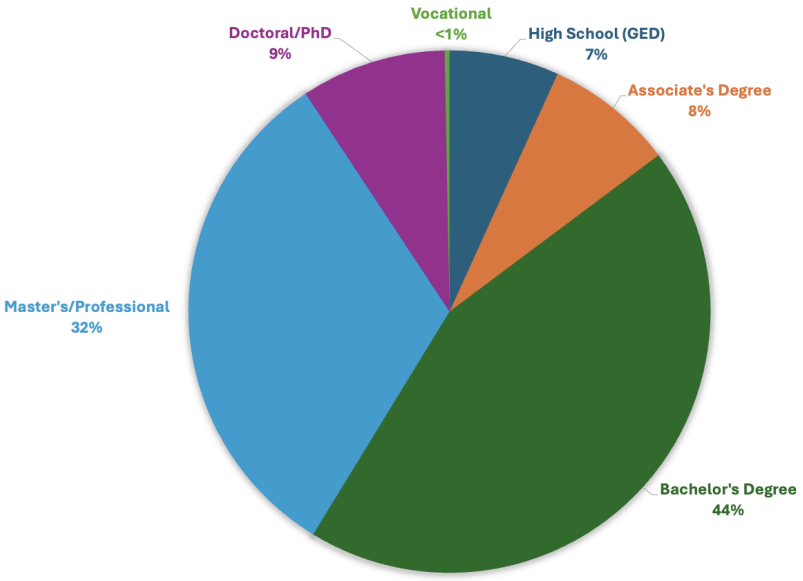
[29] "Zapier Survey Finds Workers Spend 4.5 Hours Per Week Cleaning Up AI Mistakes," Yahoo Finance, January 14, 2026, https://finance.yahoo.com/news/zapier-survey-finds-workers-spend-130000740.html, accessed February 11, 2026.

Exhibit 1 AI Value Tech Stack



Source: Re-created from data from The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf , accessed January 24, 2026.

Exhibit 2 Education Level of HITL Market



Source: Re-created using data from The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf , accessed January 24, 2026.

Exhibit 3 Data annotators engaged in other professional activities

Professional Activities	Percent of Respondents
Full-time job (35+ hours per week)	23%
Freelance only, seeking FT/PT work	16%
Self employed/business owner	15%
Part-time job (<35 hours per week)	12%
Stay-at-home parent/caregiver	10%
Freelance only, not seeking job	9%
Student (college or university)	7%
None of the above	6%
Retired	3%

Source: Re-created using data from The Economic Impact of the Data Annotation Industry, Oxford Economics, December 2025, https://www.oxfordeconomics.com/wp-content/uploads/2025/12/OxfordEconomics_DataAnnotationImpacts.pdf , accessed January 24, 2026.

Exhibit 4 Typical Compensation by HITL Segment

HITL Segment	Typical Pay (Hourly)	Scalability	Strategic Value
Data annotation	$3–15	High	Low, commoditized
RLHF	$15–50	Medium	Moderate
Evaluation & safety	$40–120	Low–Medium	High
Domain experts	$75–300+	Low	Very high

Source: ChatGPT, response to "For the four segments in the HITL market, what is the typical hourly pay?," OpenAI, January 24, 2026.

Endnotes