The Economics of Agentic AI: Token Cost, Human Labour, and the Future of Work

Nathan Gomes

Independent Research Paper

May 24, 2026

Abstract

This research paper examines the economic conditions under which agentic artificial intelligence (AI) becomes cheaper or more expensive than human labour. A simple comparison between token prices and hourly wages can make AI appear obviously cheaper, especially for repetitive digital tasks. However, agentic AI systems are not limited to one prompt and one response. They can plan, use tools, retrieve documents, call external APIs, revise outputs, and complete multi-step workflows, which means their true cost includes more than tokens. The central research question is therefore not whether a single AI response costs less than a human worker's time, but when the total operating cost of agentic AI remains lower than the total cost of human labour. Based on current API pricing, labour wage data, and research on AI adoption and infrastructure demand, this paper argues that agentic AI will usually be more economical for repetitive, high-volume, low-risk digital tasks. However, the cost advantage weakens when workflows require long context, premium reasoning models, repeated tool calls, retries, human supervision, privacy controls, compliance, or error correction. The most likely future is not complete job replacement. Instead, agentic AI will displace tasks inside jobs, reduce some repetitive labour demand, and increase the value of workers who can supervise, verify, and manage AI systems.

Keywords: agentic AI, token economics, human labour, automation, artificial intelligence, future of work

The Economics of Agentic AI: A Cost Comparison of Token Usage and Human Labour

Artificial intelligence is no longer limited to answering simple questions. Many current systems are becoming agentic, meaning they can pursue goals through several steps, use external tools, retrieve information, and produce outputs that support real business workflows. This matters

because businesses usually adopt technology for economic reasons. A company does not only ask whether AI is impressive. It asks whether AI can complete work more cheaply, quickly, safely, and reliably than human workers.

The public conversation about AI often focuses on whether AI will "take jobs." That question is important, but it is too broad. A more useful research question is economic and operational: Will the cost of AI tokens be more expensive than human labour now or in the future when AI becomes prominent in almost every industry? This question is especially important for agentic AI because agents do not simply generate one response. They may plan, search, call tools, read files, generate outputs, check their own work, and retry when they fail. Every step can add tokens and cost.

This paper argues that agentic AI will generally remain cheaper than human labour for repetitive, digital, high-volume tasks, even if AI becomes widely used across industries. However, AI will not be cheaper for every kind of work. The full economic comparison must include tokens, tools, compute, infrastructure, human review, error correction, risk, and compliance. The strongest conclusion is that AI will be cheaper than humans for many tasks, but not always cheaper than humans for complete jobs.

Preliminary Knowledge: What Agentic AI Is

Agentic AI is a type of artificial intelligence system designed to act toward a goal rather than only respond to a single prompt. A standard chatbot is usually reactive: a user asks a question, and the model answers. An agentic AI system is more active. It may interpret the user's goal, break the work into smaller steps, decide which tool to use, retrieve documents, call an API, compare results, revise its output, and ask for human approval before taking a final action.

For example, in an IT support setting, a user might write, “My VPN is not connecting after I approve multi-factor authentication.” A normal chatbot might provide generic troubleshooting steps. An agentic AI support assistant could classify the issue as VPN authentication, search a knowledge base, ask for the operating system, check whether there is an outage, draft a support ticket, recommend escalation, and summarize the case for a technician. This is why agentic AI is relevant to business operations: it can automate parts of a workflow rather than only produce text.

Agentic AI usually includes five major parts. First, it needs a goal or instruction. Second, it uses a language model to interpret the goal and generate reasoning. Third, it uses memory or retrieved context, such as documents, databases, previous tickets, or user preferences. Fourth, it uses tools, such as search, spreadsheets, ticketing systems, email, code execution, or databases. Fifth, it may involve human-in-the-loop review, where a person approves important actions before they are completed.

Table 1

Core Parts of an Agentic AI System

Component	Explanation	Example in business
Goal	The task the agent is trying to complete	Resolve a support ticket
Planning	The agent breaks the task into steps	Classify issue, search knowledge base, draft response

Tools	External functions the agent can use	Web search, file search, ticketing system, database
Memory or context	Information used to guide the task	Company policies, past tickets, user history
Model calls	Requests sent to the AI model	Analyze issue, generate answer, revise output
Human review	Human approval for important actions	Technician approves escalation or customer response

How Agentic AI Differs From Generative AI and Standard LLMs

Agentic AI is closely related to generative AI and large language models (LLMs), but it is not the same thing. A large language model is the underlying model that processes text and predicts likely responses. Generative AI is the broader category of systems that create new content, such as text, images, code, audio, or video. A chatbot is one common interface for using generative AI. Agentic AI uses an LLM or other AI model as one part of a larger workflow, but it adds planning, tool use, memory, and multi-step execution.

This distinction is important for cost analysis. A normal generative AI chatbot may only need one prompt and one response. An agentic AI system may need several model calls to complete the same business task. For example, if a user asks a chatbot to summarize a ticket, the chatbot may produce the summary in one step. If an agentic AI system handles the ticket, it may read the ticket, classify it, search a knowledge base, compare possible causes, draft troubleshooting steps, decide whether escalation is needed, and create a final ticket summary. The second workflow is more useful, but it consumes more tokens.

Table 2

Comparison of Standard Generative AI and Agentic AI

System type	Main purpose	Typical behaviour	Token cost pattern
Large language model	Process and generate language	Predicts or generates text based on input	Depends on prompt and output length
Generative AI chatbot	Answer questions or create content	Responds to a user prompt in one or a few turns	Usually one input plus one output
AI assistant	Helps with tasks interactively	Answers, rewrites, summarizes, or analyzes with user guidance	Increases with conversation length and uploaded context
Agentic AI system	Complete multi-step goals	Plans, uses tools, retrieves data, retries, and may ask for approval	Higher because each step can require another model call
Multi-agent system	Coordinate several AI agents	Multiple agents divide work, critique, or specialize	Highest because several agents may each use tokens

What Tokens Are and How Token Usage Works

Tokens are the units that large language models use to process and generate language. A token can be a word, part of a word, punctuation, or a space. OpenAI explains that tokens are the building blocks of text processed by AI models and that 100 tokens is roughly equal to 75 English words, although the exact count depends on the text, language, punctuation, numbers, and code (OpenAI Help Center, 2026). AI providers usually charge based on the number of tokens processed by the model. Input tokens are the text the model reads. Output tokens are the text the model generates. Cached input tokens may be cheaper because the provider can reuse repeated context more efficiently.

Token usage matters because an AI agent can consume tokens in several places. It uses tokens when it reads the user's prompt, system instructions, retrieved documents, tool results, previous conversation history, and examples. It also uses tokens when it writes a response, drafts a report, generates a ticket summary, or explains its reasoning. If the agent makes a mistake and retries, token usage increases again.

A basic chatbot might use one input and one output. An agentic workflow may use many rounds of input and output. For example, a support agent might first classify the ticket, then search internal documentation, then summarize search results, then generate a troubleshooting plan, then check the answer, and finally draft a response. Each stage can be a separate model call. This makes agentic AI more useful but also more expensive than a simple chatbot.

Table 3

Approximate Token Requirements for Common Tasks

Task type	Example task	Approximate input tokens	Approximate output tokens	Why the amount changes
Very short chatbot answer	User asks one simple question and receives a short answer	50 to 200	50 to 200	Minimal context and short output
Email or message draft	Write a professional email from a short request	200 to 800	200 to 600	More output text increases cost
One-page summary	Summarize a short article or meeting note	1,000 to 2,500	300 to 800	The model must read the source text before summarizing

Support ticket triage	Classify a user issue and draft troubleshooting steps	2,000 to 8,000	500 to 2,000	Includes user issue, instructions, categories, and response
Document-based answer	Answer using an uploaded policy or knowledge base excerpt	5,000 to 30,000	500 to 3,000	Retrieved documents become input tokens
Tool-using agent workflow	Search files, compare results, draft final answer, and self-check	10,000 to 100,000+	1,000 to 10,000+	Multiple model calls and tool results increase total tokens
Multi-agent workflow	Several agents plan, critique, and revise a task	50,000 to 500,000+	5,000 to 50,000+	Each agent may read and write its own context

Note. These are approximate ranges for explaining cost behaviour, not fixed values. Exact token counts depend on model tokenizer, language, formatting, code, length of retrieved documents, tool results, and number of retries.

Table 4

How Token Costs Accumulate in an Agentic Workflow

Workflow step	Token source	Why it adds cost
User request	Input tokens	The model reads the problem
System instructions	Input tokens	The model follows rules and safety instructions
Planning	Input and output tokens	The model creates steps
Tool results	Input tokens	Search or database results are sent back to the model

Draft answer	Output tokens	The model writes the response
Self-check or retry	Input and output tokens	The model evaluates or revises its work
Final response	Output tokens	The model produces the final deliverable

Current Evidence on AI Token Pricing

Current API pricing shows why AI is economically attractive for companies. OpenAI's API pricing page lists prices per one million tokens and separates input, cached input, and output prices. For example, current OpenAI pricing lists GPT-5.4 nano at $0.20 per one million input tokens and $1.25 per one million output tokens, while a more expensive model such as GPT-5.5 is listed at $5.00 per one million input tokens and $30.00 per one million output tokens for standard short-context use (OpenAI, 2026). Google's Gemini API pricing also shows a wide range of costs. Gemini 2.5 Flash is listed at $0.30 per one million input tokens and $2.50 per one million output tokens for text, image, and video input in the paid tier (Google AI for Developers, 2026).

These prices show that model choice matters. A company using a smaller or cheaper model for routine tasks can keep costs very low. A company using premium reasoning models for every task will pay much more. Agentic AI economics therefore depends on routing: simple tasks should go to cheap models, while complex or risky tasks should go to more capable models with human review.

Table 5

Examples of Published AI API Prices

Provider and model	Input price per 1M tokens	Output price per 1M tokens	Best use
OpenAI GPT-5.4 nano	$0.20	$1.25	Low-cost routine tasks
Google Gemini 2.5 Flash	$0.30	$2.50	Fast general workflows
OpenAI GPT-5.5	$5.00	$30.00	Complex reasoning or high-value work

Note. Prices are examples from official provider pricing pages available in May 2026. Actual costs can change and may differ by context length, batch processing, region, caching, tools, and enterprise agreements.

Current Evidence on Human Labour Costs

To compare AI with human labour, it is useful to examine jobs that are likely to be affected by agentic AI, such as customer service and IT support. The U.S. Bureau of Labor Statistics reported that the median hourly wage for customer service representatives was $20.59 in May 2024 (Bureau of Labor Statistics, 2025a). The BLS also reported that the median annual wage for computer user support specialists was $60,340 in May 2024, which is about $29.01 per hour if divided by 2,080 full-time work hours (Bureau of Labor Statistics, 2025b).

Human labour costs are not only wages. Employers may also pay payroll taxes, benefits, training, management, equipment, software, office space, and quality assurance. A simple cost comparison can add 25% overhead to direct wage cost. This is not a perfect estimate, but it makes the comparison more realistic than using wages alone.

Table 6

Approximate Human Labour Cost per Five-Minute Task

Role	Wage basis	Direct five-minute labour cost	Estimated cost with 25% overhead
Customer service representative	$20.59/hour	$1.72	$2.15
Computer user support specialist	$29.01/hour	$2.42	$3.02

Note. Five-minute cost is calculated by dividing hourly wage by 12. Overhead is estimated at 25% for comparison purposes.

Method: Scenario-Based Cost Model

This paper uses a scenario-based cost model rather than an experiment with a live company. The model compares the estimated cost of completing one task with AI against the estimated cost of completing the same task with human labour. The AI cost model includes input tokens, output tokens, tool calls, retries, and human review. The human cost model includes wage time and estimated overhead.

The basic AI formula is:

Total AI cost = input token cost + output token cost + tool cost + retry cost + human review cost + error correction cost

The basic human labour formula is:

Total human labour cost = time spent on task × hourly wage + overhead

The purpose of this model is not to predict the exact cost for every company. Instead, it shows how the economics change depending on task complexity, model choice, and review requirements.

Scenario Analysis: Simple Digital Task

The first scenario is a simple support request. Assume the AI agent uses 8,000 input tokens and 2,000 output tokens. This is enough for a user request, system instructions, a small amount of retrieved context, and a useful response.

Table 7

Estimated AI Cost for a Simple Agentic Support Task

Model scenario	Input tokens	Output tokens	Estimated token cost per task
Low-cost model example	8,000	2,000	About $0.004 to $0.008
Mid-cost model example	8,000	2,000	About $0.03
Premium model example	8,000	2,000	About $0.10

Note. Estimates are based on the published prices in Table 3. They do not include all possible tool, storage, integration, or supervision costs.

Compared with the human cost in Table 4, AI is clearly cheaper for the simple task. A customer service worker costs about $2.15 for a five-minute task after estimated overhead. An IT support worker costs about $3.02 for a five-minute task after estimated overhead. Even a premium AI model costing around $0.10 in tokens is cheaper than the direct human labour cost for a routine five-minute task.

This is the main reason companies are interested in agentic AI. When tasks are repetitive, digital, and high-volume, the marginal cost of AI can be far lower than the marginal cost of human time.

Scenario Analysis: Tool-Using Agent

Agentic AI becomes more expensive when tools are added. OpenAI's pricing page lists web search at $10.00 per 1,000 calls for some models, with search content tokens billed at model rates. It also lists file search tool calls at $2.50 per 1,000 calls (OpenAI, 2026). These are small costs per call, but they matter at scale.

Table 8

Estimated Cost of a Tool-Using Agentic Task

Cost item	Example quantity	Approximate cost
Premium model token use	8,000 input + 2,000 output tokens	$0.10
One web search call	1 call	$0.01
One file search call	1 call	$0.0025
Estimated AI cost before review	One task	$0.1125

Even after adding one web search and one file search call, the AI task remains far cheaper than five minutes of human labour. However, this assumes the agent completes the task successfully on the first attempt and does not require substantial human review. The economics become weaker if the agent retries several times or if every answer must be checked by a human.

Scenario Analysis: Complex Agentic Work

A complex agentic task may require long context, multiple tool calls, and human review. For example, an AI agent asked to investigate a technical incident may read logs, search documentation, compare multiple sources, generate a report, and ask a human to approve the conclusion. Assume the task uses 100,000 input tokens and 30,000 output tokens with a premium model. Using GPT-5.5 standard short-context prices as an example, that token usage would cost about $1.40. If the task requires five web searches, that adds about $0.05. If a human spends two minutes reviewing the result at an IT support wage with overhead, that adds about $1.21. The estimated total becomes about $2.66.

Table 9

Estimated Cost of a Complex Agentic Task

Cost item	Example quantity	Approximate cost
Premium model input tokens	100,000 tokens	$0.50
Premium model output tokens	30,000 tokens	$0.90
Five web search calls	5 calls	$0.05
Two minutes of IT review with overhead	2 minutes	$1.21
Estimated total	One complex task	$2.66

This is still cheaper than a human spending 15 to 30 minutes on the task. However, if the task is high-risk and requires 20 minutes of review, legal approval, or correction after errors, AI may no

longer provide a large cost advantage. This shows the real conclusion: AI cost advantage is strongest when human review is light and error risk is low.

Break-Even Analysis

The break-even point is the level where AI cost equals human labour cost. For a five-minute customer service task costing about $2.15 with overhead, even a $0.10 AI task is far below the break-even point. For a five-minute IT support task costing about $3.02 with overhead, the gap is even larger.

Table 10

Break-Even Comparison for One Five-Minute Task

Task type	Estimated human cost	Estimated AI cost	Cost advantage
Simple customer service task	$2.15	$0.004 to $0.10	AI is much cheaper
Tool-using support task	$2.15 to $3.02	About $0.11	AI is much cheaper
Complex IT task with light review	$6.04 to $18.12 for 10 to 30 minutes of human work	About $2.66	AI is usually cheaper
High-risk task with heavy review	Varies	Can rise quickly	Depends on review, errors, and liability

Figure 1

Cost Comparison for a Routine Five-Minute Task

Cost category	Approximate cost
Low-cost AI agent	Less than $0.01
Premium AI agent	About $0.10
Customer service human labour with overhead	About $2.15
IT support human labour with overhead	About $3.02

Note. This figure is presented as a data chart rather than a decorative graph so that the values remain readable in Google Docs.

Will AI Tokens Become More Expensive in the Future?

The future cost of agentic AI depends on two opposing forces. The first force pushes costs down: model efficiency, hardware improvements, competition, caching, smaller models, batch processing, and local deployment. The second force pushes costs up: more AI usage, more complex agents, more tool calls, greater electricity demand, and demand for premium reasoning models.

There is strong evidence that the first force is powerful. Stanford's 2025 AI Index reported that the inference cost for a system performing at the level of GPT-3.5 dropped more than 280-fold between November 2022 and October 2024 (Stanford HAI, 2025). This means that a fixed level of AI capability became dramatically cheaper in less than two years. If this trend continues, the cost of many routine AI tasks should keep falling.

However, the second force is also real. The International Energy Agency reported that data centres accounted for around 1.5% of global electricity consumption in 2024, or 415

terawatt-hours. It projects that data centre electricity consumption will more than double to about 945 terawatt-hours by 2030, with AI as the most important driver of growth alongside other digital services (International Energy Agency, 2025). This shows that AI infrastructure demand is increasing rapidly.

These facts do not mean that AI token costs will automatically become more expensive than human labour. They mean that total AI spending may rise even while the cost per token or cost per task falls. This is similar to other technologies: when something becomes cheaper, companies use much more of it. The price per unit can decrease while total spending increases because usage grows.

The most likely future is that routine AI inference becomes cheaper, but premium agentic AI remains expensive for complex tasks. Companies will respond by using a mix of models. A cheap model may handle classification, summarization, and simple drafting. A stronger model may handle complex reasoning. Humans may review high-risk outputs. This layered approach is more economical than using the most expensive model for every task.

Why AI Will Usually Be More Economical for Repetitive Digital Work

Agentic AI is likely to be more economical than human labour in repetitive digital work for four reasons. First, the marginal cost of tokens is very low compared with wages. A task that costs a human worker several dollars in time may cost an AI system cents or less. Second, AI can scale quickly. Once the system is built, it can handle many tasks without hiring and training new workers at the same rate. Third, AI can operate continuously, which is useful for 24-hour support environments. Fourth, AI can improve human productivity by drafting, summarizing, searching, and preparing work before a human makes the final decision.

Customer support, ticket classification, document summarization, email drafting, basic data extraction, and knowledge-base search are strong examples. These tasks are digital, text-based, repetitive, and relatively easy to verify. The lower the risk and the easier the verification, the stronger the economic case for AI.

Why AI Will Not Be More Economical for Every Job

AI becomes less economical when mistakes are costly. In legal, medical, financial, cybersecurity, safety, and management decisions, an incorrect answer can cause serious harm. In these situations, the cost of human review, compliance, and liability may be greater than the token savings.

AI is also weaker when work requires physical presence, emotional intelligence, deep accountability, negotiation, or trust. For example, AI can help a customer service worker draft an answer, but a human may still be needed to handle an angry customer, make an exception, or protect a business relationship. AI can help an IT technician diagnose a problem, but a human may still need to check hardware, access secure systems, or make a judgment about risk.

This means the economic question should not be "AI or humans?" The better question is "Which parts of the job should AI handle, and which parts still require humans?" In many industries, the winning model will be human-plus-AI rather than AI-only.

Job Replacement Risk

Table 11

Replacement Risk by Work Type

Work type	AI cost advantage	Risk of AI error	Human judgment required	Replacement risk
FAQ customer support	High	Low	Low	High
Ticket classification	High	Low to medium	Low	High
Routine report drafting	High	Medium	Medium	Medium to high
IT troubleshooting	Medium to high	Medium	Medium	Medium
Data analysis	Medium	Medium to high	High	Medium
Client relationship management	Low to medium	High	High	Low to medium
Legal, medical, or financial decisions	Low	Very high	Very high	Low
Physical labour	Low	High	High	Low

The highest replacement risk is found in tasks that are repetitive, digital, low-risk, and easy to verify. The lowest replacement risk is found in tasks that require responsibility, human trust, physical work, or high-stakes judgment. This supports the idea of task displacement. AI is more likely to remove parts of jobs than to eliminate all human labour.

Discussion: Will Token Costs Outcost Human Labour?

Based on current evidence, token costs are unlikely to outcost human labour for routine digital tasks, even if AI becomes common across industries. The cost per simple AI task is too low compared with wages. Even when tool calls are included, agentic AI can still be far cheaper than human labour for repetitive work.

In the future, AI usage will increase across industries, and data centre demand will grow. This may create energy, infrastructure, and supply constraints. However, these pressures do not necessarily mean token prices will rise above human labour costs. Competition among AI providers, smaller models, caching, improved chips, and software optimization are likely to continue reducing the cost of routine inference. Stanford's reported 280-fold drop in GPT-3.5-level inference cost is strong evidence that AI cost per unit of capability has been falling quickly.

The more realistic risk is not that every AI task becomes more expensive than human labour. The more realistic risk is that companies underestimate the full operating cost of agentic AI. A poorly designed agent may call tools too often, use too much context, retry unnecessarily, or require too much human correction. In that case, AI becomes less economical. Companies need cost controls such as model routing, usage limits, caching, monitoring, and human approval only where it is actually needed.

Conclusion

Agentic AI will likely be more economical than human labour for many repetitive, digital, high-volume tasks. Current token prices show that AI can complete simple support, summarization, classification, and drafting tasks for cents or less, while human labour for the same task often costs dollars. Even when tool calls are included, AI can remain cheaper for low-risk workflows.

However, token cost is not the entire cost of AI. The real cost includes tools, retries, integration, human review, privacy, compliance, infrastructure, and error correction. As AI becomes prominent across industries, total AI spending and data centre energy demand will rise. Still, this does not mean tokens will generally become more expensive than human labour. The cost per unit of AI capability has been falling, and companies will likely use cheaper models for routine work and stronger models only for complex tasks.

The most scientific answer is therefore conditional. AI will be more economical than humans when tasks are repetitive, digital, high-volume, low-risk, and easy to verify. Humans will remain more economical and necessary when work requires judgment, accountability, physical presence, emotional intelligence, legal responsibility, or complex decision-making. Agentic AI is unlikely to end work entirely. It is more likely to change the structure of work by automating tasks, reducing some labour demand, and increasing the importance of human oversight and AI management.

References

Bureau of Labor Statistics. (2025a). Customer service representatives: Occupational Outlook Handbook. U.S. Department of Labor. https://www.bls.gov/ooh/office-and-administrative-support/customer-service-representatives.htm

Bureau of Labor Statistics. (2025b). Computer support specialists: Occupational Outlook Handbook. U.S. Department of Labor. https://www.bls.gov/ooh/computer-and-information-technology/computer-support-specialists.htm

Google AI for Developers. (2026). Gemini Developer API pricing. https://ai.google.dev/gemini-api/docs/pricing

International Energy Agency. (2025). Energy and AI: Executive summary. https://www.iea.org/reports/energy-and-ai/executive-summary

McKinsey & Company. (2025). The state of AI: Global survey 2025. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

OpenAI. (2026). API pricing. https://developers.openai.com/api/docs/pricing

OpenAI Help Center. (2026). What are tokens and how to count them? https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

Stanford Institute for Human-Centered Artificial Intelligence. (2025). Artificial Intelligence Index Report 2025. https://hai.stanford.edu/ai-index/2025-ai-index-report