

COST GUIDE · OCTOBER 15, 2026

AI Agent Cost and Scoping Guide

The AI model is usually the smallest part of the bill: most of the cost is the work around it, connecting your data, testing and keeping the agent safe. Six drivers move the price most, and Netixsol quotes every agent after scoping.

Written by Netixsol, which builds AI agents and DeFi protocols with the same rule: don't ship what you can't verify.

netixsol.com

Who is this guide for?

This guide is for founders, product leads and operations heads deciding whether to build an AI agent, and what it should cost. By the end you will know:

- What you are paying for in an AI agent project.
- The six things that move the price most.
- What a first version, a production agent and a larger platform typically involve.
- What it costs to run an agent each month after launch.
- The questions to answer before you ask anyone for a quote, Netixsol included.

What do you pay for in an AI agent project?

The AI model is usually the smallest part of the bill. Most of the cost is the work around it: connecting your data, testing and keeping the agent safe.

PART	WHAT IT COVERS	WHY IT COSTS
Discovery and scoping	The task, users, data sources, success measures, risks	Decides everything after it; skipping it is the most expensive mistake
Data work	Connecting databases, documents and APIs; cleaning and describing the data	Agents are only as good as what they can read; this is often 30 to 50% of the effort
Agent logic	Prompts, tools the agent can call, multi-step plans, memory	Grows with the number of tools and decisions
Evaluation	A test set of real questions with known answers, accuracy tracking	What turns a demo into something you can trust
Guardrails and access	Permissions, read-only modes, limits, human hand-off, logging	Required wherever the agent touches money, customers or private data
Interface and integration	Chat UI, Slack or WhatsApp, your app or CRM	Depends on where people will use it

PART	WHAT IT COVERS	WHY IT COSTS
Launch and monitoring	Hosting, dashboards, alerts, feedback loop	Agents drift as data and models change

What are the six cost drivers?

Six things move the price more than anything else. If you know where your project sits on each, you can estimate most of the quote.

- **What the agent is allowed to do.** An agent that only answers questions is far cheaper than one that sends emails, changes records or moves money. Every action needs permissions, confirmation steps and tests.
- **How messy the data is.** One clean database is quick. Ten systems, scanned PDFs or undocumented tables mean weeks of data work before any AI is involved.
- **What a wrong answer costs you.** Internal brainstorming tolerates mistakes. Finance, health, legal or customer-facing answers need larger test sets, citations and human review.
- **How many steps and tools.** A single question-and-answer flow is simple. An agent that plans five steps across three tools needs more design, more testing and more monitoring.
- **Where people use it.** A web chat is the cheapest surface. Embedding it in your product, CRM, Slack or WhatsApp adds integration work.
- **Volume and speed.** Thousands of users or near-instant answers change the model choice, caching and hosting, and with them the monthly running cost.

What are the three typical project shapes?

Most projects fit one of three shapes. Start with the smallest one that proves the value, then grow it.

A pilot should end with a clear yes or no: did the agent answer the test set well enough to be worth the next step? If a pilot can't be judged that way, it's scoped too loosely.

SHAPE	WHAT YOU GET	TIMELINE
Pilot	One task, one data source, read-only answers, a test set of about 50 real questions, a simple web chat	4 to 6 weeks
Production agent	Two to five data sources, a few safe actions with permissions, a test set of 200+ questions, monitoring, one integration (your app, Slack or WhatsApp)	8 to 12 weeks
Agent platform	Several agents sharing tools, an admin dashboard, single sign-on, audit logs, several integrations	4 to 6 months

What does an AI agent cost to run after launch?

An agent has a monthly bill as well as a build cost. Ask for both before you sign. The monthly bill has four parts:

- **Model usage.** You pay the model provider per token (roughly, per word read and written). It scales with how many conversations you have and how long they are.
- **Hosting and data.** Servers, the database, and the search index the agent reads from.
- **Monitoring.** Logs, accuracy checks against the test set, and alerts when answers get worse.
- **Improvement.** New questions, new data sources and model upgrades. Plan for a small monthly retainer or a few days a quarter, agreed in the support plan.

How do you estimate the model bill?

Monthly model cost $\approx$ conversations per month $\times$ tokens per conversation $\times$ price per token.

Prices per token differ by model and change often, so check the provider's current pricing page. Caching repeated context and using a smaller model for simple steps often cuts this bill substantially.

How can you lower cost without lowering quality?

- **Start read-only.** Let the agent answer and recommend before it acts. Add actions once the answers have earned trust.
- **One task first.** A narrow agent that does one job well beats a broad one that does five jobs badly, and it costs a fraction.
- **Fix the data once.** Describing your tables and documents clearly helps every future agent, not just this one.
- **Write the test set early.** Twenty to fifty real questions with known answers stop weeks of guessing about whether a change helped.
- **Match the model to the step.** Use a strong model where reasoning matters and a smaller, cheaper one for routine steps.
- **Reuse what exists.** Your existing login, chat tool and dashboards are usually cheaper than building new ones.

What are the red flags in an AI agent quote?

Be careful with a quote that:

- Prices the build but not the monthly running cost.
- Promises an accuracy figure before anyone has seen your data.
- Has no test set or evaluation step.
- Doesn't say who owns the code, prompts and data afterwards.
- Gives the agent write access to your systems with no permission model or audit log.
- Has no plan for monitoring after launch.
- Treats "we use the latest model" as the whole plan.

Scoping worksheet

Answer these before you ask for a quote. With them, any serious team can price your project within days; without them, every quote is a guess.

#	QUESTION	YOUR ANSWER
1	What task or decision should the agent take over? Describe one real example.	
2	Who will use it: your team, your customers, or both? How many people?	
3	Where does the data live (databases, documents, apps), and who may see what?	
4	Should the agent only answer, or also act (send, update, book, pay)?	
5	What does a wrong answer cost you?	
6	Which 20 questions must it always get right?	
7	Where will people use it: web, your app, Slack, WhatsApp, CRM?	
8	What would make the pilot a clear success in numbers (time saved, tickets closed, answers correct)?	
9	Any rules to follow: data location, industry regulation, security reviews?	
10	When do you need a first version, and what's your budget range?	

Worked example: StatGPT

StatGPT is an AI sports analyst Netixsol built. Users ask in plain English and get answers from NBA, AFL, AFLW and cricket player data.

It's built with LangGraph and OpenAI models on a PostgreSQL stats database. The lesson that applies to every agent: most of the work was teaching it the data, and a fixed test set of real questions told us when a change helped. The chat agent took 22 weeks of calendar time, with client reviews setting part of the pace, inside an 81-week platform build by 12 engineers.

DRIVER	WHERE STATGPT SITS
What it may do	Answer only: it reads the stats database and never changes it
Data	Large: season, per-game, advanced and round-by-round stats, which the agent had to learn to query correctly
Cost of a wrong answer	Moderate: a wrong number hurts trust, so answers come from database queries, not from the model's memory
Steps and tools	Multi-step: understand the question, write the database query, check it, explain the result, remember the conversation
Where it's used	A web app with chat and search, in a private beta for the client's investors
Volume	AFL smoke tests in Feb 2026: 1,900 questions, and a median answer of 2.9 s on a run of 1,000; usage figures stay with the client

Next step: a short scoping call

Bring your answers to the worksheet. On a short call we'll tell you which project shape fits, what would make the pilot a clear success, and the next step. We reply within one working day.

Book a call at calendly.com/mirzausman371/30min, or send us your worksheet at netixsol.com/contact.

Netixsol: Protocol-grade engineering for AI, DeFi and digital products. Protocols · Agents · Products.