

KNOWLEDGE LIBRARY

What Is an AI Agent?

The Smart Factory from Data to Decision



Author

Ender Öztürk

OETECH Solutions, Bursa, Türkiye

Published

18 October 2026

Version · Language

1.0 · English

ABSTRACT

Let engineers spend their time on engineering, not on collecting data.

You hire an industrial engineer. You expect them to study production resources, expose bottlenecks, make the operator's job safer and easier, and develop better production methods. Yet a large part of their day goes to downloading data from different screens, chasing missing records, merging spreadsheets and preparing reports for meetings.

The person who should be on the shop floor is busy catching up on reports. The person who should be designing improvements is reconciling numbers. And managers often see a problem not when it occurs, but when its results show up in the next report.

Keywords: AI · AI Agent · Smart Factory · MES · Digital Transformation

Suggested citation: Öztürk, E. (2026). What Is an AI Agent? The Smart Factory from Data to Decision (Report No. 2026-04). OETECH Solutions. <https://oetechsolutions.com/nedir/ai-agent-nedir>

Report No.	OETECH-TR-2026-04
Author	Ender Öztürk — Enterprise Solutions Account Manager, OETECH Solutions, Bursa, Türkiye
Published	18 October 2026
Version / Language	1.0 / English
Scope	22 sections, 8 figures, 24 Q&As
Online version	https://oetechsolutions.com/nedir/ai-agent-nedir

This report is the print edition of the online guide in the OETECH Knowledge Library. Example calculations, budgets and timelines are explicitly hypothetical; they are not customer results, price quotes or performance commitments. Images are conceptual illustrations.

CONTENTS

—	Introduction
1	What exactly is an AI agent?
2	Why does data matter so much?
3	Don't turn engineers into data collectors
4	What does a model that knows the factory mean?
5	Which agents can work in a factory?
6	What changes when departments work together?
7	How do you move from an alert to a physical solution?
8	Local, cloud or hybrid: When does each make sense?
9	Does a hybrid setup require moving the entire factory to the cloud?
10	How much hardware does a local model need?
11	Is it possible to run in the cloud on a low budget?
12	How do you calculate the cost of local infrastructure?
13	What should the total project budget include?
14	How much staff time can be saved?
15	How does time saved turn into economic gain?
16	How are energy and resource savings achieved?
17	How should permissions, trust and auditability be established?
18	How do you manage alert fatigue and wrong recommendations?
19	How do you start the project?
20	How is success measured?
21	OETECH's AI Factory approach
—	Sources and calculation notes

A Frequently asked questions

Introduction

The real value of the AI agent approach in manufacturing is that it lets the business get more out of human expertise. A properly built agent examines the data it is authorized to access, investigates deviations from the expected state, prepares an assessment with its supporting evidence and starts the action process within defined limits. Engineers and managers then use the time they gain for better decisions and solutions that work on the shop floor.

This is also the direction OETECH has been working toward for a long time: turning the data generated in areas such as MES, human resources and logistics into a decision support structure that understands how the factory works. The **AI Factory** described on this page means an operating structure built on shared data, with clearly defined permissions and measurable results, instead of disconnected chat screens.

1. What exactly is an AI agent?

An AI agent is a software component that can gather information toward a specific goal, use the appropriate tools, evaluate the results it obtains and choose the next step within its permissions. In Turkish it is also called a "yapay zekâ ajanı" or "yapay zekâ aracısı."

A chat model can only generate an answer to the question you ask. An agent, on the other hand, given a request such as "Investigate the delay on this work order," can query production confirmations, downtime, material status and quality holds; bring the evidence it finds together; say so if data is missing; and draft an investigation task for the responsible person.

Not every automated process is an agent. Monitoring a predefined threshold, sending a regular report or running a fixed calculation is often solved more simply with conventional software. An agent makes sense where the investigation steps change depending on the situation and different sources have to be interpreted together. ¹

Table 1

Structure	What does it do?	Factory example
Report / BI	Presents defined indicators	Daily downtime and production table
Rule engine	Triggers an action when a specific condition occurs	Opening a record when stock falls below the critical level
Forecasting / optimization model	Solves a specific mathematical problem	Demand forecasting, capacity planning, route optimization
AI assistant	Answers user questions, prepares content	Annotated summary of the shift report
AI agent	Investigates with tools, evaluates options and runs a controlled workflow	Examining a delay and preparing evidence and a proposed action

These structures do not have to be alternatives to one another. A good system can use software for exact calculations, a solver for planning and a language model for explanation and investigation, all together.

2. Why does data matter so much?

"We produced 8,000 parts today" tells you the quantity. When you link the same production to the product, work order, machine, mold cavity, operation, time, material lot and quality result, you can answer many more questions.

You can investigate which product struggles on which resource, whether the same defect is concentrated in a particular mold cavity, and whether a delay comes from material or from downtime. Data detail makes visible the relationships needed for the right decision.

However, **more data does not by itself mean more value**. You need data that is relevant to the decision, accurate, timely and shares a common meaning. Wrong timestamps, inconsistent product codes or unexplained downtime records can lead to wrong conclusions, even when they are collected in very large volumes.



Figure 1. Value comes not just from collecting data, but from establishing the relationships between product, time, resource and result.

What questions does detailed data open up?

Table 2

Context collected	Question it can answer
Work order and operation	At which stage is the delay building up?
Machine, mold and mold cavity	Is the loss concentrated on a specific resource?
Actual cycle and ideal cycle definition	Where is the speed loss, and is the comparison fair?
Downtime reason and response time	Which action worked for recurring downtime?
Material lot and quality result	Which product group could a complaint affect?
Competency and training records	Which task needs additional training or support?
Energy meter and production quantity	Under which conditions does energy per unit rise?
Stock, orders and due dates	Which delivery is affected by a material or capacity risk?

Human resources data is not something to be added to production data without limits. Its purpose of use and access boundaries must be defined. Individual health information or private conversations should not be a default data source for an agent that interprets production performance.

3. Don't turn engineers into data collectors

Data collection, coordination and shop-floor follow-up are necessary in any business. The problem is not that these tasks exist, but that expert effort is constantly consumed by them. Supervising a shift as a foreman is also a separate role that requires its own field experience; substituting engineering capacity for it without a plan can reduce the value of both roles.

An engineer's real contribution is to analyze flow, study time and motion losses, evaluate ergonomic arrangements with specialists, compare capacity scenarios and measure the result of a change.



Figure 2. Time saved on report preparation can go toward a safe working layout, process analysis and improvement.

The agent can compile shift records, separate out data inconsistencies, rank the losses that need investigation and prepare an evidence file before a meeting. The engineer then checks whether the proposal fits the physical conditions, talks with the operator and designs the appropriate change.

For example, seeing that the same product moves more slowly at one station is only the starting point. Factors such as a part that is hard to grip, transport distance, fixture access or work sequence need to be examined on the shop floor. The agent can guide this investigation; it does not count as having automatically completed the occupational safety assessment or an actual work study.

4. What does a model that knows the factory mean?

Knowing the factory does not mean loading the entire database into a model once. It means being able to access current sources with the right permissions and interpret the company's concepts correctly.

This knowledge structure has four pillars:

1. **Live business data:** Work orders, confirmations, downtime, stock, quality results and tasks.
2. **Institutional knowledge:** Approved instructions, maintenance documents, control plans, past problem solutions and training content.
3. **Company glossary:** Machine names, product codes, time units, KPI definitions and departmental responsibilities.
4. **Permissions and history:** The records a user can access, approved actions, model version and decision trails.

The approach of searching documents and presenting the relevant passages to the model is called RAG. Structured information such as current production quantity, however, should be retrieved through authorized APIs and validated queries. A shift total should be produced by the calculation layer, not estimated by the language model from text. ²

Fine-tuning can be considered to improve a model's task behavior or output format; it is not a required method for learning every new stock movement. RAG on its own does not guarantee accuracy either: retrieving the wrong document, using an outdated version or pulling in unauthorized content are all still possible.

An answer should show the data time, scope, sources and assumptions alongside it. A statement such as "Last updated record 14:32; two machines are disconnected" is more valuable than a confident sentence that hides the gap.

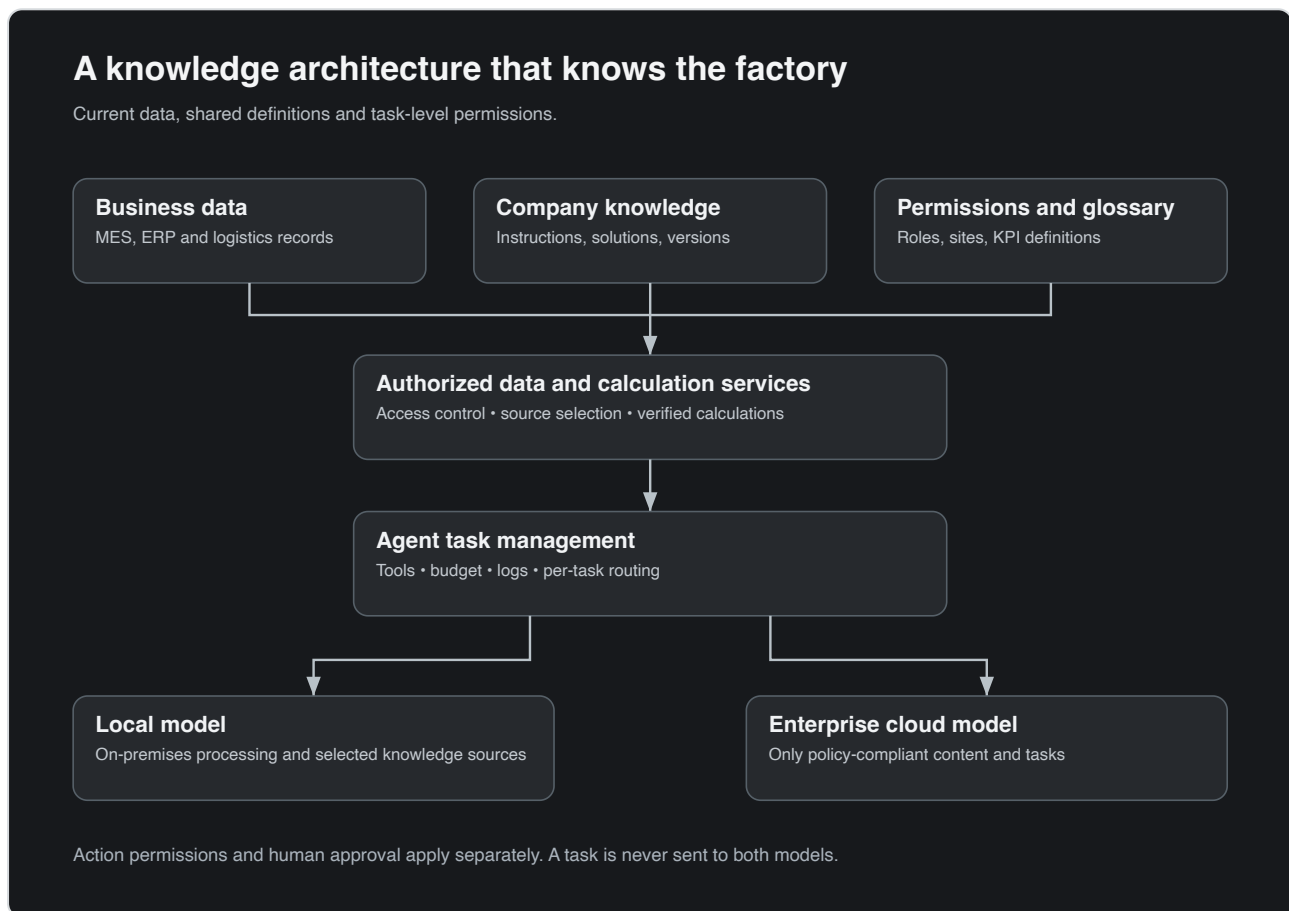


Figure 3. The information the model needs is retrieved according to the task and access permissions. Local and cloud options are chosen per task; not every query is sent to both.

5. Which agents can work in a factory?

The manager and engineer names used here describe the business role the agent supports. They do not give the software professional authority, legal responsibility or the decision rights of a human manager. Data access, tools and action limits are defined separately for each role.



Figure 4. Specialist agents can work in a shared business context; each department's sensitive data is not opened to every role.

Table 3

Agent role	Primary contribution	Example output	Human responsibility
Production engineer	Investigating cycle times, downtime and workflow	Loss list, investigation proposal	Shop-floor verification and process design
Production manager	Shift, capacity and delivery risk	Priority intervention summary	Resource allocation and production decisions
Quality engineer	Reviewing defect, lot and inspection records	Problem file, 8D draft	Root cause verification, technical assessment
Quality manager	Escape risk and open quality actions	Quarantine/release assessment	Authorized quality approval
Planning engineer	Material, capacity and routing scenarios	Alternative sequence and impact comparison	Validation of constraints
Planning manager	Due-date and priority conflicts	Decision options and affected orders	Publishing the plan
Logistics engineer	Investigating stock, internal transport and shipping	Missing material and picking risk	Physical stock verification
Logistics manager	Warehouse, transport and shipping coordination	Shipment readiness summary	Shipping and transport approval
Maintenance / energy agent	Recurring failures and consumption deviations	Draft investigation and maintenance task	Technical intervention and safety
Plant manager	Combining cross-department impacts	Shared risk and action agenda	Operational priorities and decision ownership
Finance manager support agent	Cost, payment term and cash scenarios	Assumption-based cost and cash impact	Accounting verification, payment authority
Human resources agent	Training, competencies and administrative requests	Training needs, summary of the authorized process	Fair evaluation and personnel decisions
Employee support agent	Access to approved resources and referral	Support channels, general well-being information	Psychologist / health professional services

How should the employee wellbeing support agent (workplace psychologist role) be positioned?

Even if a support role is designed under the name workplace psychologist agent, the software should not be put in place of a psychologist's professional services. A better starting point is an assistant that offers approved content, explains how to reach a human specialist and makes it easier for employees to access support processes. Clinical diagnosis, treatment and individualized psychological assessment should not be within the scope of this role.

Private conversations should not be used to generate production performance scores or disciplinary decisions. Methods such as inferring mood from facial expressions are not a default part of this architecture. Employees should be able to understand who has access to which data and how conversations are retained. If a company-level report is needed, data that is properly aggregated so that individuals cannot be identified in small groups should be preferred. The conditions for processing personal data are assessed separately. ⁷⁸

6. What changes when departments work together?

The production agent may identify a speed loss on a machine. The quality agent may have seen an increase in defective products over the same period. The planning agent knows that the order on that machine has a critical due date. On the logistics side, the status of the material needed for alternative production is found.

When these findings are combined in a single event record, the manager is not simply told that "output is low." They are shown which order is affected, the evidence at hand, the missing information and the options that can be implemented.

Agents do not need to constantly query one another. An agent that does a single task well, or a defined workflow, may be enough for most starting points. Many agents bring additional cost, latency and conflict management. The number of agents should grow with real task needs. ¹

Different agents using the same language model reaching the same conclusion does not count as independent evidence. Conflicts should be resolved by going back to the source data, the calculations and an authorized human assessment.

7. How do you move from an alert to a physical solution?

What a business needs is not a constant stream of notifications, but getting the important problem to the right person on time. In a well-designed flow, every event has a scope, a priority, an owner and a closure condition.



Figure 5. Detection, evidence, human decision, shop-floor execution and result verification are tracked through the same event.

The example scenario is purely illustrative: on a production line, cycle-time deviation and quality nonconformities are rising together.

1. **Trigger:** A rule or statistical monitoring system detects the defined deviation. The agent does not need to read the entire database every second.
2. **Data check:** Counters, product changeovers, data latency and missing records are examined.
3. **Investigation:** Related downtime, lot, mold and quality records are retrieved. Similar past events are searched.
4. **Decision file:** The affected work order, evidence, alternative explanations and recommended checks are prepared.
5. **Authorized review:** Production and quality owners review the proposal and, if necessary, decide to hold product or run a technical investigation.
6. **Shop-floor task:** A task is opened with the approved scope; an owner, a deadline and the required safety steps are assigned.
7. **Closure:** After the intervention, cycle time, quality and recurrence are measured. Closing the task is not taken to mean the problem is solved.

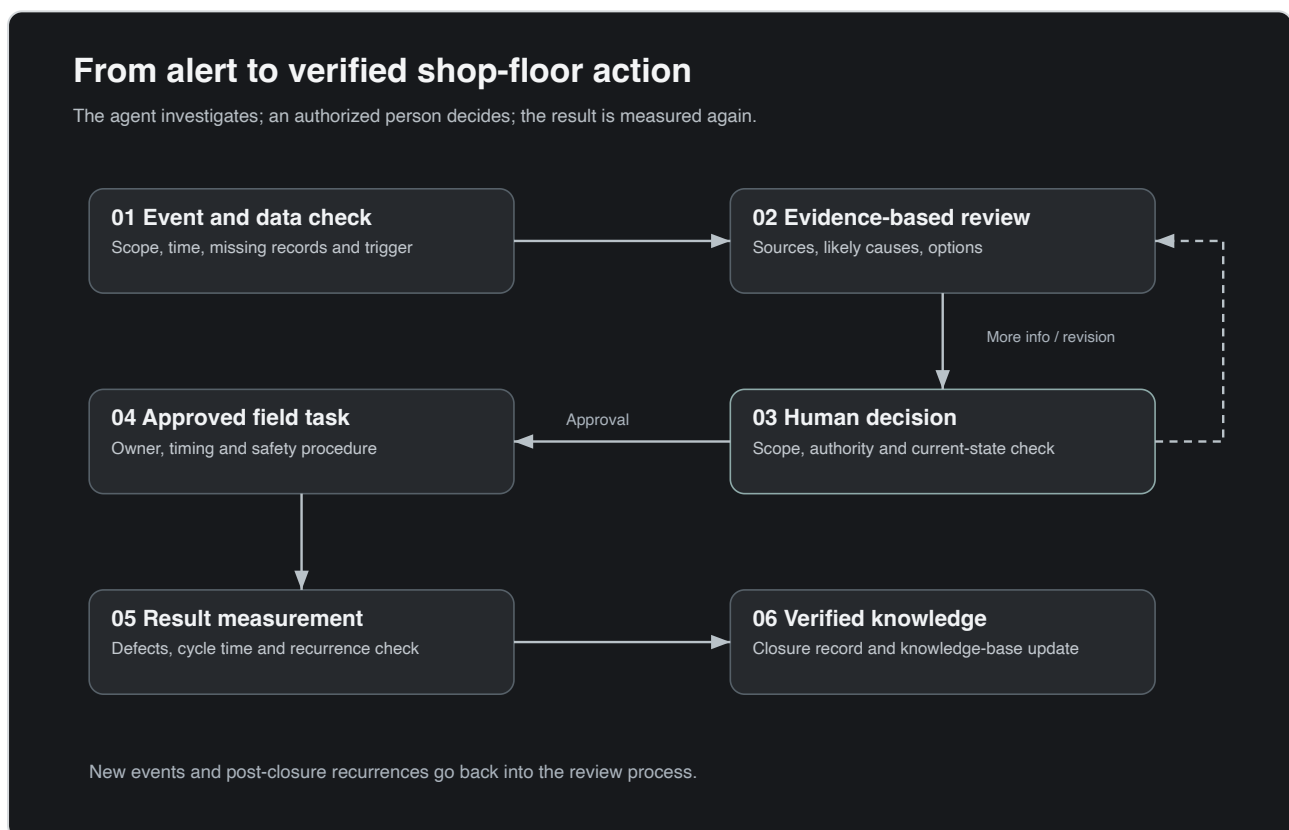


Figure 6. Solid arrows show the main flow; loops back show additional information and re-evaluation.

"Real time" must also be defined. Source data latency, network, query, model queue and approval time together make up the total response time. A language model should not be used in place of millisecond-level machine safety or deterministic PLC control.

8. Local, cloud or hybrid: When does each make sense?

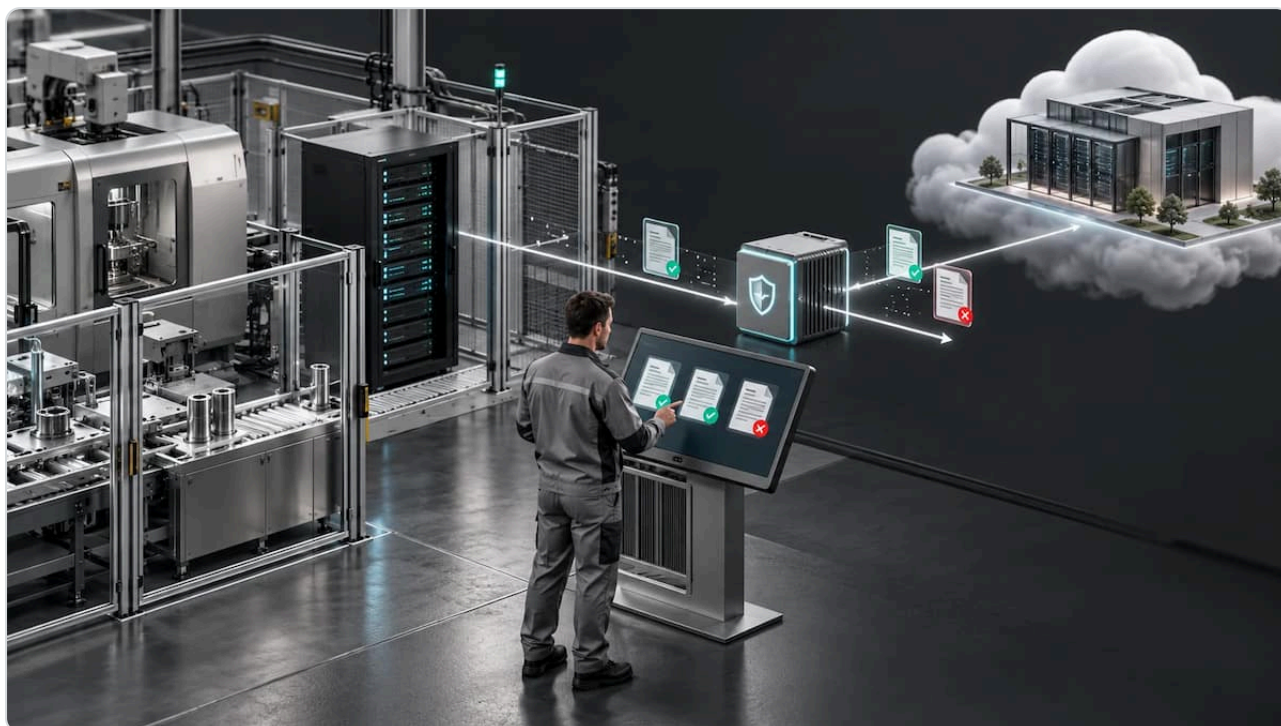


Figure 7. The deployment decision is made by evaluating data sensitivity, usage intensity, model performance and operational capacity together.

Table 4

Topic	Local	Enterprise cloud	Hybrid
Where the model runs	Company hardware	Contracted provider infrastructure	Both environments, depending on the task
Initial hardware need	Usually higher	Purchasing a model server may not be necessary	Can start with limited local capacity
Usage cost	Energy, maintenance, capacity and operations	Token/request, infrastructure and service fees	Combined cost depending on routing
Internet dependency	The model can run if local dependencies are in place	Requires connectivity and service access	Some work can continue locally
Data leaving the site	Can be restricted by design	Submitted content is processed by an external service	Restricted per task
Scaling	Requires hardware capacity and management	Depends on quota, region and service capacity	Both sides must be managed
Suitable starting point	Sensitive/continuous local workloads	Variable, low-volume pilot	Companies with both sensitive data and a need for a powerful model

Local models still process tokens; there may simply be no per-token bill from an external API. Instead, there are hardware, energy and operating costs. The cloud can also start with a low entry cost, but uncontrolled long tasks and retries can inflate the budget.

A "trusted cloud" does not just mean a well-known brand. The chosen service's data use terms, retention period, abuse monitoring conditions, region, subprocessors, access management and contract must be reviewed. A commitment not to use data for training does not mean the data is never stored.

For example, Microsoft's documentation on Foundry models sold by Azure describes specific data use and access limits, while some stateful features do retain records. Amazon Bedrock's documentation likewise restricts model providers' access to the relevant deployment accounts and customer prompts. These statements do not mean that all products and deployment options are the same, or that regulatory compliance is provided automatically. ³⁴

9. Does a hybrid setup require moving the entire factory to the cloud?

No. Instead of the entire historical database, an authorized local service can prepare the result for the question at hand. For example, instead of millions of downtime rows, it can provide a summary with a defined scope and calculation definition, along with the relevant event records.

Sensitive document search and calculations can be kept local; if company policy allows, a specific summary can be sent to a more powerful cloud model. Removing a person's or customer's name does not always achieve anonymization: product codes, dates, quantities and business context can also reveal identities or trade secrets.

The local connector component should use authenticated, defined data services. There is no need to grant broad database permissions from the internet or expose every SQL table to the model. Queues, timeouts, rate limits and records that prevent the same job from being created twice on retries are important.

Which tasks will continue locally when the internet goes down, how cloud-dependent analysis will be deferred, and which decisions will not be made on incomplete analysis must be defined in advance. This behavior should be verified with an actual outage test.

10. How much hardware does a local model need?

The number of agents alone does not determine hardware size. Ten different roles can use the same model server in turn. The real drivers are model size, context length, the number of concurrent requests, response time and the required task performance.

Table 5

Starting level	Purpose	Infrastructure to evaluate
Trial with the existing system	Validate data access, queries and workflow	Existing application server; if suitable, a small local model or a controlled API
Limited local pilot	Specific tasks with a small number of concurrent users	A single model server with adequate memory, SSD, backup and monitoring
Operational use	Multiple roles, regular event flow and accepted latency	Measured capacity, separate data services, job queue and a support plan
Critical service continuity	Maintaining service during failures and maintenance	Redundancy, access separation, fallback and disaster recovery design

Fitting the model weights into memory is not the only requirement. The context cache, concurrent requests and the runtime use additional memory. Low-bit execution can reduce memory use; the quality impact must be measured on the selected tasks. The memory of multiple GPUs also does not automatically combine, regardless of the software used. ⁹

That is why, instead of a "this much GPU is enough for any factory" approach, measurements should first be taken with a small set of tasks. The model's license for commercial use, runtime support and long-term maintainability are also selection criteria. Being open-weight does not mean every use is unconditionally permitted.

11. Is it possible to run in the cloud on a low budget?

It is possible at a certain scope and volume. However, a chat subscription is not the same as enterprise data integration and a complete agent operations service. The model usage fee is only one component of the total budget.

To make the example calculation concrete, let's use the **Claude Haiku 4.5 standard API pricing** verified on 18 October 2026: 1 USD per million input tokens and 5 USD per million output tokens. This model is used only for the sake of the calculation; it is not a recommendation of the best or cheapest option. Regional cloud reseller pricing should not be assumed to be the same. ⁵

The consumption figures below are hypothetical. They assume 22 working days per month, normal usage without caching, and that all billable input/output tokens stay at the amounts given in the table.

Table 6

Usage scenario	Jobs/day	Model calls/job	Input / output per call	Monthly input / output	Model fee only
Limited summaries and queries	20	1	2,000 / 500 tokens	0.88 / 0.22 million	1.98 USD
Regular departmental support	100	3	4,000 / 1,000 tokens	26.4 / 6.6 million	59.40 USD
Intensive, multi-step investigation	500	6	8,000 / 2,000 tokens	528 / 132 million	1,188 USD

Formula: **Monthly input tokens / 1,000,000 × input price + monthly output tokens / 1,000,000 × output price.**

These amounts do not include integration, servers, document search, embeddings, data storage, OCR, additional tool use, support, networking, taxes or exchange-rate effects. Retries, growing conversation history and additional evaluation calls can increase consumption. If the model does not complete the task successfully, the business value of a cheap call drops as well.

To keep costs under control, provide short and relevant data, do not hand deterministic work to the model, cut unnecessary agent turns and set a time/call budget per task. Routing common simple tasks to a small model and hard tasks to a powerful model can also be tested.

12. How do you calculate the cost of local infrastructure?

The calculation below is not a market price or an OETECH quote either. It is a **budget model** that the company should replace with its own quotes:

- Hypothetical investment in a model server: **5,000 USD**.
- Useful life for management accounting: **36 months**; monthly investment share of about **138.89 USD**. This is not a depreciation calculation under tax regulations.
- Assumed average power measured at the outlet: **0.35 kW**.
- Continuous use for 30 days: $0.35 \times 24 \times 30 = \mathbf{252 \text{ kWh/month}}$.
- Hypothetical electricity unit price: **0.15 USD/kWh**; **37.80 USD** per month. This value is not presented as a Turkish electricity tariff.
- Assumed hardware-specific maintenance/spare parts allowance: **50 USD/month**.

Under these conditions, the specified local infrastructure items alone come to about **226.69 USD/month**. Software, integration, staff operations, additional cooling, networking, backup and downtime costs fall outside this example. Using existing hardware does not reduce its capacity and opportunity cost to zero either.

It would not be correct to compare this figure directly with the API fee in the previous table and make a decision. The **total cost of ownership** must be worked out for both sides under the same task performance, volume, latency and continuity conditions. At low volume, starting in the cloud may be eco-

nomical; with continuous and suitably consolidated workloads, the local option may make sense. The outcome is determined by measurement.

13. What should the total project budget include?

The bigger effort, compared with the cost of running the model, usually goes into preparing the data and setting up the workflow. If an agent is not connected to the right user, the right data and the right approval mechanism, a powerful model alone will not solve the problem.

Table 7

Initial setup	Ongoing operations
Process analysis and task definition	Model/API or hardware cost
Data glossary and quality corrections	Data flow and connection maintenance
MES/ERP/HR/logistics integration	Model and document version management
Access, approval and task workflows	Monitoring, support and human review
Test set and acceptance trials	Security, backup and capacity
User training and commissioning	New products, reports and process changes

In a pilot proposal, narrowing the scope is the key step that makes a low budget possible. Instead of an "unlimited agent that knows every department," a starting point such as one that investigates daily production losses from two data sources and creates approved task drafts can be priced clearly.

Cost proposals should show which data sources are ready, how many tasks are covered, usage limits, support scope and acceptance conditions. Validating data access and task performance before purchasing hardware can reduce unnecessary investment.

14. How much staff time can be saved?

When discussing staff savings, two outcomes must be kept apart: **freed-up expert capacity** and **realized cash savings**. If the same team is doing better work, capacity has increased; payroll costs have not automatically gone down.

Let's consider the following workloads in an entirely hypothetical company. The "after" times include human review of the agent's output:

Table 8

Task	People	Before, per person/day	After, per person/day	Gain/day
Merging production data and preparing reports	2	90 min	30 min	120 min
Preparing quality event files	2	60 min	25 min	70 min
Compiling planning and logistics status	2	60 min	30 min	60 min
Management meeting agenda and data preparation	1	45 min	20 min	25 min
Total	7	465 person-min	190 person-min	275 person-min

275 minutes is about **4.58 hours** per day. Over 22 working days a month, **100.83 hours** of capacity are gained. If 12 hours a month are set aside for shared monitoring, model maintenance and additional review, **88.83 net hours** remain.

This equals **11.10 working days** of eight hours, or **0.50 full-time equivalent** of a hypothetical 176-hour month. It is not correct to present this time, spread across different people, as a single head-count reduction. Which improvement the gained time is allocated to should also be tracked.

If the internal loaded hourly cost is assumed, purely as an example, to be TRY 600, the accounting equivalent of the net capacity is about **TRY 53,300/month**. This figure is not a market wage and does not count as an automatic cash gain. The real economic benefit is verified through avoided overtime, avoided external services, additional workload absorbed or measured improvements.

15. How does time saved turn into economic gain?

The source of the gain can be broader than writing a report faster. A delay noticed early, the right material routing, a quality problem that does not recur or a reduction in unnecessary energy consumption can create real operational impact.

However, the same benefit should not be counted twice. If the good-product capacity gained from less scrap is calculated separately, the contribution of those same products should not be added again under another heading. If there is no demand for the additional capacity and no capacity in downstream operations, no sales gain may result.

In the example financial scenario, suppose the following verifiable and additive monthly cash effects are realized:

Table 9

Hypothetical item	Monthly value
Avoided overtime / external reporting costs	TRY 12,000
Net avoided quality and rework costs	TRY 18,000
Energy savings adjusted for production conditions	TRY 8,000
Total achievable benefit	TRY 38,000
Software, support, model, infrastructure and internal operating costs	-TRY 14,000
Net monthly benefit	TRY 24,000

If the initial investment is, entirely hypothetically, **TRY 240,000**, the simple payback period is **10 months**. The calculation assumes the benefit is realized at the same level from the start; it does not include taxes, financing, the time value of money or the ramp-up period. This example is not an OETECH price or a customer result.

The TRY 53,300 capacity value from the previous section is not added to this table. This prevents the same working hours from being counted both as capacity and as a cash gain. A company should not use these values as an investment commitment without its own baseline measurement and pilot results.

16. How are energy and resource savings achieved?

If energy meters, machine states, production and shift information are all available, the agent can investigate idle consumption. It can flag for investigation high consumption during non-production hours, a recurring compressor operating pattern or consumption deviations associated with product changeovers.



Figure 8. Savings are realized through the maintenance, program change or physical improvement carried out after the alert.

Simply using less electricity than the previous month does not prove the agent's success. Factors such as production quantity, product mix, operating time and weather conditions can affect the comparison. A baseline and adjustment method suited to the task should be used.

The AI's own consumption must also be taken into account. For example, assuming the local server above consumes 252 kWh per month, if the gross reduction verified at the facility is 1,000 kWh, subtracting only this additional server load gives a net reduction of **748 kWh**. Any other additional loads are deducted separately. The cloud's energy consumption is not zero either; the fact that it does not appear directly on the local electricity bill does not change this.

In material usage, the agent can make areas such as nonconforming lots, stock aging, rework and unplanned changeovers visible. The owner and result of the action that delivers the physical savings should be recorded separately.

17. How should permissions, trust and auditability be established?

A model can produce a convincingly wrong answer. That is why critical figures should come from a calculation service, sources should be verified and missing information should be stated clearly. Human approval should not be just a button either; the decision-maker must be able to see the scope, the basis and the impact. ⁶

Table 10

Permission level	What can the agent do?	Example
Read	Retrieves and explains authorized information	Comparing downtime reasons
Draft	Prepares a record and submits it for review	Draft maintenance or quality task
Limited automation	Executes a pre-approved low-risk operation	Limited notification entry in an internal task queue
Authorized operation	Makes a call for a specific scope after human approval	Transferring an approved plan revision to the system

Impactful steps such as payments, personnel actions, shipment release or changing machine parameters require separate authorization and verification. If data has changed since approval, conditions must be rechecked before the operation. An approval given in the past is not unlimited permission for a changed scope.

Access control cannot just be written into the prompt text. Permissions at the user, site, department and record level must be enforced in the data service. A command hidden in a document can steer the agent toward leaking data or performing unauthorized actions. Retrieved documents must be handled as data, not instructions; tools and output limits must be enforced in code. RAG alone does not solve this risk. ¹⁰¹¹

In HR and employee support, purpose, access, retention and informing employees must be addressed separately. A local installation alone does not fulfill all personal data obligations either. If there is processing or transfer abroad, the chosen service and the legal conditions must be reviewed separately. ⁷⁸

18. How do you manage alert fatigue and wrong recommendations?

Sending a separate message for every deviation quickly erodes attention. Records related to the same event should be combined, and severity, the affected order and any existing open task should be taken into account. Each alert should have an owner, an expected response time and a person who steps in if it is delayed.

It must remain visible that what the agent calls a "possible cause" is not a proven root cause. Wrong recommendations and rejected alerts should be recorded and used for data or threshold corrections. The confidence score the agent assigns to its own answer should not be the only measure of operational success.

Failed tool calls, outdated documents, empty results and connection outages should also be within the test scope. In these situations, the system should report its limitation and switch to a defined alternative rather than making things up and carrying on.

19. How do you start the project?

The first step is not to define agents for every department at once. It is to choose a task with ready data access, a measurable workload and a clear output owner. Under suitable conditions, a daily production loss summary and quality event file preparation can be two starting candidates.

An example **90-day roadmap** can be used as a planning framework recommended by OETECH; it is not a delivery-time commitment for every company:

Table 11

Period	Work	Exit criterion
Days 1–15	Problem, baseline workload, data and permissions analysis	Task scope and measurable acceptance criterion
Days 16–35	Data services, knowledge sources and the first agent	Sourced, re-evaluable sample outputs
Days 36–55	Historical data testing and shadow operation	Calculation, source, access and cost tests
Days 56–75	Human-approved task workflow and user training	Traceable implementation on a limited part of the shop floor
Days 76–90	Benefit measurement and scope decision	Rationale to continue, adjust or stop

If data infrastructure or process definitions are incomplete, preparation can take longer. Increasing the number of agents at the end of the pilot is not an automatic goal. A failed task can be simplified or converted into conventional automation.

20. How is success measured?

Baseline and pilot conditions should be comparable, and efficiency gains should be measured including human review. A flow that is faster but produces more errors may not actually be reducing the real workload.

Table 12

Metric	What does it show?
Task completion success	Whether the agent actually completes the work
Calculation and source accuracy	Whether the answer is based on the data and the correct calculation definition
Net preparation time	Time saved including human review and correction
Time to first detection and decision	How quickly the problem becomes visible and the decision is made
Verified action rate	How often an alert turns into a useful action
Recurrence after closure	Whether the physical problem decreases
Total cost per task	Combined cost of the model, tools and human effort
Unauthorized access/operation attempts	Whether access limits are enforced
User acceptance and rollbacks	How well the system fits daily work, and erroneous operations

The test set should not contain only easy questions. It should also include examples of missing data, wrong units, shift boundaries, product changeovers, out-of-scope records, outdated instructions and conflicting sources. The same tests should be rerun after a model or integration change.

21. OETECH's AI Factory approach

The process knowledge accumulated through OETECH's MES and automation work, together with its work on production, HR and logistics data, forms the foundation of this approach. The goal here is a structure that does not just report on the company's data, but helps determine which problem will be investigated by whom and which action's result will be measured.

The solution can be deployed locally, in an enterprise cloud or as a hybrid, depending on the company's needs. Roles that start with production and quality can be extended to planning, logistics, maintenance, finance and HR as the data and permission structure becomes ready. Sensitive roles are designed with separate access boundaries.

The set of roles described on this page is a solution framework. It does not mean that every role is ready in every company, that it achieves the same success in all tasks, or that a specific saving is guaranteed. The scope is clarified through a pilot with real data and acceptance tests.

The goal is for the company to make better use of its skilled people: Engineers can do more analysis and improvement; managers don't wait for data to make decisions; the problems operators face become visible on time; and the shop-floor result of every action is tracked.

Related reading: [What Is MES?](#), [What Is OEE?](#), [What Is AI Vision Inspection?](#).

Sources and calculation notes

The sources support the relevant technical and institutional concepts. The example savings, workloads, hardware budget and project timeline in this article are not OETECH customer results or industry averages. Monetary assumptions other than the API pricing are for educational purposes. Prices and service terms should be re-verified at the time of purchase.

1. [Anthropic — Building effective agents](#)
2. [Microsoft Learn — Retrieval augmented generation and indexes](#)
3. [Microsoft Learn — Data, privacy and security for models sold by Azure](#)
4. [AWS — Amazon Bedrock data protection](#)
5. [Claude Platform — Haiku 4.5 pricing and specifications](#)
6. [NIST — Generative Artificial Intelligence Profile, AI 600-1](#)
7. [KVKK \(Turkish Personal Data Protection Authority\) — Guide on Generative AI and the Protection of Personal Data](#)
8. [KVKK — Use of Generative AI Tools in the Workplace](#)
9. [NVIDIA — GPU memory usage for LLM inference](#)
10. [OWASP — Prompt Injection](#)
11. [OWASP — Excessive Agency](#)

Sources and API pricing checked on: 18 October 2026.

Appendix A — Frequently asked questions

A.1 Will an AI agent replace engineers or managers?

It can automate certain tasks, but technical and managerial responsibility does not automatically pass to the software. At the outset, the most meaningful goal is to reduce the data preparation burden and increase experts' capacity for decisions and improvement.

A.2 Is more data always better?

The detail needed for a decision is useful. Purposeless, inaccurate or unrelated data, however, can create cost and confusion. The questions to be answered and the data definitions should be established first.

A.3 Can we start without MES?

You can start with specific document and reporting tasks. Live production decisions, however, require reliable event, identity and process data. MES helps provide this context.

A.4 Do we have to send our entire legacy database to the cloud?

No. You can build a structure that retrieves the relevant, authorized records through local services. The commercial or personal sensitivity of any summary to be sent is assessed separately.

A.5 Do we need to train a model from scratch to teach it about the factory?

For most starting tasks, it is not required. A structure that retrieves data services, the company glossary and approved knowledge sources can be tried first. The need for custom training is determined through task testing.

A.6 Are there no token costs at all with a local system?

If no external API is used, there is no token bill from it. But the model still processes tokens; hardware, energy, support and capacity costs remain.

A.7 Does each agent need its own GPU?

No. Multiple roles can use the same model service. Capacity is measured according to the real concurrent workload and the model's requirements.

A.8 Does using the cloud mean our data goes into model training?

Not for every service. Enterprise services come with different commitments. Training use, retention, monitoring, region and contract terms must each be reviewed.

A.9 Is a local installation automatically secure?

No. Permissions, network, backups, updates, logs and data processing must be managed properly. Data that never leaves the site can still be exposed to the wrong person.

A.10 Will the factory stop if the internet goes down?

Core production and safety processes should not be designed to depend on the agent service. Local work can continue; cloud-dependent analysis can be deferred by a defined rule. Outage behavior should be tested.

A.11 Can an agent find every problem instantly?

No. It can only work within the scope of data it can access that is current and sufficient. Alert speed also depends on data flow and system latency.

A.12 What happens if the agent's recommendation is wrong?

Important operations should have human review and execution controls. Sources, calculations and assumptions should be shown; the error should be recorded and the scope re-evaluated.

A.13 Should an agent stop or adjust a machine on its own?

A language model should not be put in place of direct safety control. Control changes should be carried out through technical risk assessment, validated automation and authorized processes.

A.14 Will the planning agent change the production sequence on its own?

Initially, it is appropriate for it to compare alternatives and submit them for approval. Automatic execution is considered separately, and only with defined constraints, permission limits and a rollback mechanism.

A.15 Can an agent make financial payments?

Preparing a payment draft or a cash scenario is different from transferring money. Financial authority, controls and approvals must be maintained; a language model should not be given an unlimited payment tool.

A.16 Will the HR agent make hiring or firing decisions?

In this architecture, personnel decisions rest with authorized human assessment. The agent can compile information; impactful decisions about an employee should not be made based on a score it produces on its own.

A.17 Can the workplace psychologist agent provide therapy?

The role proposed here is to provide approved general information and support access to human specialists. It does not replace psychologist services, clinical assessment or treatment. Private conversations should not be transferred into performance evaluations.

A.18 Can it work with Turkish and company-specific terminology?

It can work with a suitable model and a company glossary, but real Turkish reports, abbreviations and technical terms must be tested. General language performance does not by itself prove performance on factory tasks.

A.19 What could be a first step on a low budget?

Choose one data source and a narrow task. Start with read and draft-generation permissions. Measure the workload, set usage limits and evaluate hardware investment based on the results.

A.20 How many people's worth of savings should we expect?

A correct number cannot be given before the tasks are measured. Gains should be calculated in person-hours, with human review and maintenance effort deducted. Capacity gains spread across people are not an automatic headcount reduction.

A.21 Should agents constantly talk to each other?

No. Running the relevant roles when an event or task arrives can be simpler. Unnecessary turns create cost and latency; stop conditions should be defined.

A.22 How do we know whether the investment is paying off?

Baseline workload, error rate and cost are recorded. After the pilot, net time, task success and verified physical results are compared. The number of messages generated is not a measure of success.

A.23 What should we do if the pilot does not deliver enough benefit?

Data issues, task selection, model performance and integration are examined separately. The scope can be simplified, conventional automation can be used or the project can be stopped. The purpose of the pilot is to support the investment decision with evidence.

A.24 What should we prepare to get started with OETECH?

Get ready to share two problems you want to solve, your existing reports, data sources, access owners and workload. Success criteria, data boundaries and action approvals should be defined together at the start.



ABOUT THE AUTHOR

Ender Öztürk works on enterprise solutions at OETECH Solutions, focusing on bringing manufacturing execution systems, industrial data and AI applications to the shop floor.

ABOUT OETECH

Founded in 2013 and based in Bursa, OETECH Solutions develops software and hardware for production management, industrial data collection and automation. With IoE+ MES Cloud, HR+ Cloud PDKS and SoSo Inventory, it carries production data from the shop floor to decisions.

© 2026 OETECH Solutions. All rights reserved. Quotation permitted with attribution. ·
oetechsolutions.com