# The AI tool buying worksheet

Gates, questions for the vendor, a 30-day pilot protocol and an ROI structure that survives the CFO.

**ExperienceFirst** · Companion to the article "How to choose an AI tool for measuring customer experience when you already have VoC" · experiencefirst.lt/en

01 · BEFORE YOU START

# How to use this worksheet

Print it. Take it into meetings. Fill it in while the vendor is talking — not afterwards.

Sequence matters. Most purchases fail because they start with comparison rather than with gates: three vendors are compared, two of which should never have made it that far.

| STEP | WHAT YOU DO | WHEN |
|---|---|---|
| **1. Baseline** | Measure your own signal-to-decision gap (below) | Before any vendor contact |
| **2. Gates** | Four yes/no questions. Fail one, you are out. | Before the demo |
| **3. Questions** | Sent in writing. The quality of the answers filters better than slides. | Before the demo |
| **4. Pilot** | 30 days on your data, against your benchmark | Before signature |
| **5. Comparison** | Only among survivors, and only with pilot numbers | Day 31 |

> **Why not a weighted matrix.** Weights are almost always chosen once you already know which vendor you want, and subjective scores dress themselves up as objectivity. More importantly: depth level and legal compliance are not criteria with percentages — they are gates. A vendor who fails them should not be rescued by points for "support and training".

## Your baseline

Three numbers for the last quarter. Without them you will have nothing to compare the investment against in a year — and the vendor knows it.

| | |
|---|---|
| **Signal volume** | Feedback, complaints, chats, calls — everything you collected |
| **Decision count** | Counts only if it has an owner, a date and a consequence: a changed process, a reallocated budget, a corrected initiative |
| **Cycle time** | Average time from signal to decision |
| **Ratio** | Decisions per 1,000 signals = (decisions ÷ signals) × 1,000 |

If the ratio is below 1, the problem sits between insight and decision. That stretch is organisational, and AI does not yet work there. A tool can still help — but bought as a way to accelerate decisions, not as another layer of listening.

02 · SCREENING

# Four gates

Binary. Fail one and the vendor is out, however good they look everywhere else.

### 1 Depth level ≥ 3

The system must do more than classify (topics, sentiment). It must interpret: identify cause, journey context, size of impact. The proof is not a slide but a real case where the system surfaced a theme nobody had articulated.

☐ Pass ☐ Fail Note: _____

### 2 EU AI Act compliance, in writing

Since 2 February 2025 it is prohibited to use AI to infer employees' emotions in the workplace from biometric data. Recognising a customer's emotion during a call is permitted. Penalties reach €35m or 7% of global turnover. You need written confirmation from the vendor, not a conversation.

☐ Pass ☐ Fail Date confirmed: _____

### 3 Data and taxonomy export

Taxonomy and historical assignments exportable in a machine-readable format on termination. If not, in two years you will be unable to compare your own data against anyone else, and your position at renewal will be zero.

☐ Pass ☐ Fail Contract clause: _____

### 4 Agreement to a 30-day pilot

On your real data, against your benchmark, which the vendor cannot see. A refusal is an answer.

☐ Pass ☐ Fail Pilot start: _____

### Depth levels — mark where the vendor actually stands

| LEVEL | WHAT THE SYSTEM DOES | EVIDENCE (NOT A PROMISE) |
|---|---|---|
| **1. Collection** | Gathers feedback and recordings | |
| **2. Classification** | Topics, sentiment, categories | |
| **3. Interpretation** | Cause, context, size of impact | |
| **4. Recommendation** | Action, owner, priority | |
| **5. Action** | Creates the task in an operational system itself | |

If you already run a working feedback programme, you pay only for level 3 and above. Everything below is either a feature of your existing platform or a commodity whose price falls every year.

03 · QUESTIONS FOR THE VENDOR

## Send in writing before the demo

Not during it. Before. An evasive answer is also data — write it down.

### 1. How did you measure accuracy in our language — on what data, and against what benchmark?

Don't demand a third-party audit report: almost none exist in this market, and asking for a non-existent document gets you "we'll come back to you". The likely answer is "we haven't". That is your information: the number gets created in the pilot.

### 2. Show three cases where the system recommended an action the client had not articulated — and what was done about them.

If the answer is a dashboard screenshot, it is level 2, whatever it is called.

### 3. Who owns the taxonomy and historical assignments after termination? Can we export them as CSV?

Export terms are negotiated in the first contract, not the third.

### 4. Confirm in writing: does the product infer employees' emotional state from voice, video or physiological data? Is it enabled by default in EU deployments? Can it be disabled at tenant level?

In many contact centre platforms customer and employee analysis live in the same module, and the employee-facing part is sometimes on by default.

### 5. Can a recommendation create a task in our operational system without a human in between — and who is accountable if the action is wrong?

The second half of the question matters more than the first. Integration without a line of accountability is a risk, not a feature.

### 6. Who in our organisation will receive the system's conclusion — and does that person hold a budget?

This question is not for the vendor. It is for you. An insight that travels to someone without authority will never become a decision.

04 · PILOT

# The 30-day protocol

A demo on vendor data proves nothing. The only serious test is your data and a benchmark the vendor cannot see.

| | |
|---|---|
| **Hypothesis** | The vendor's model assigns topics and causes to our customer signals accurately enough to base operational decisions on its output. |
| **Change** | 500 real records from our own channels (not demo data). Two people code the benchmark. The vendor cannot see it and processes the same records through their system. |
| **Expected outcome** | Agreement with human coding of ≥ 80% at topic level and ≥ 60% at cause level. In addition, three recommendations go to a business owner with one question: "Would you have done anything about this?" |
| **Success metric** | Accuracy against the benchmark, plus at least one recommendation the owner is prepared to implement without further analysis. |
| **Review date** | 30 days. The decision is made the same day: continue or stop. |

### What this costs your team

The first question you will get from your executive team. Calculate it like this:

| WORK | WHO | EFFORT |
|---|---|---|
| **Benchmark coding** | 2 people | 500 records. Around 60–100 records per hour for short text; 20–30 for call transcripts. **Only a 100-record sample is double-coded**: enough to measure inter-rater agreement, and half the cost of coding everything twice. |
| **Coordination** | 1 person | ~1 day: data extraction, handover, agreeing coding rules |
| **Analysis and decision** | 1 person | ~1 day: comparison against the benchmark, review of three recommendations with the business owner |

Roughly 6–8 working days for the voice channel. Less for text. Compare that with the sum you are about to commit for three years.

> **Why coding rules are agreed before, not after.** If two of your own people disagree on how to assign a topic, you cannot measure any model's accuracy — you will be measuring your own ambiguity. Agreement between humans first; comparison against the system second.

05 · MONEY AND RISK

# An ROI that survives the CFO

The vendor's proposal will almost certainly add saved agent hours, avoided churn and NPS uplift into one impressive figure. Three problems with it:

- **Revenue instead of margin.** "50 customers × €2,000" is revenue. Your CFO writes that off in three seconds.
- **A counterfactual.** Avoided churn cannot be proven without a control group. It will be the first question.
- **The same euro twice.** Saved time and retained customers often come from the same fixed process.

> An ROI calculation before the pilot is a sales artefact, not a financial model. The real number arrives on day 31.

### A defensible structure — two components only

| | |
|---|---|
| **A. Time saved** | Hours **measured** in the pilot (not forecast) × fully loaded cost of employment (salary + taxes + overhead). Not gross salary. |
| **B. One fixed process** | Before-and-after volume + control period + **margin, not revenue**. One proven process convinces more than three forecast ones. |
| **Costs** | Licence + implementation + your team's time + data cleaning |
| **Payback** | Costs ÷ (A + B) in months. Realistic — not three months. |

## What can burn

| RISK | LIKELIHOOD | WHAT TO DO |
|---|---|---|
| **Recommendation reaches someone without authority** | Very high | Before purchase, name who signs off on decisions arising from each type of insight. The risk does not disappear after deployment — it only gets more expensive. |
| **Recommendations not implemented due to budget** | High | Agree in advance on the threshold below which a decision proceeds without a board submission. |
| **Agents feel monitored and stop raising issues** | High | Involve agents in the pilot; communicate that the process is analysed, not the person. A legal question as much as a cultural one. |
| **Messy data** | Medium | Clean before the pilot. Otherwise you measure the state of your CRM, not the accuracy of the model. |
| **Vendor locks in your data** | Medium | Export terms in the first contract. |

## When you want a second pair of eyes

In 45 minutes I review your measurement system, your baseline and the vendors' answers — and tell you whether you need an AI tool at all, and at what level.

**experiencefirst.lt/en · ina@experiencefirst.lt**

Sources: Regulation (EU) 2024/1689 (AI Act), Article 5(1)(f); European Commission guidelines on prohibited AI practices, C(2025) 884 final. This worksheet is not legal advice — for an assessment of your specific case, consult a lawyer. 