

14 · Teams and operations

AI Tool Selection Scorecard

An editable weighted scorecard that compares three tools on seven criteria, with privacy knock-out conditions, so you choose by your task rather than by hype.

For: For team leads, business owners and operations staff who must choose between AI tools for a specific task.

Level: Intermediate Version 1.0 · 2026-09-21

Why a weighted scorecard?

Comparing AI tools slides easily into impressions: one tool dazzled in a demo, another is the one everyone talks about, a third is cheaper. A weighted scorecard forces two things before you look at any tool: deciding what actually matters to you, and deciding how much each criterion matters compared with the others. The comparison then becomes a transparent calculation that anyone on the team can check and challenge.

The scorecard does not choose for you. It exposes what you would otherwise have ignored and puts disagreements on the table. When two people on the team disagree about a tool, they usually disagree about the weights, not the tool.

Tools, their capabilities and their prices change quickly. This template deliberately names no prices and recommends no specific tools. The criteria are stable; the scores are yours to give after testing each tool yourself on a date you record in the file (at the time of writing this resource: September 2026).

The seven criteria

Criterion	Core question	How to test it
Task fit	Does it do the task you need?	A sample of 5 to 10 real, anonymised cases from your work, given to all three tools with identical instructions.
Arabic quality	Is the Arabic output usable?	Ask for MSA text, and dialect if your audience needs it; check spelling, phrasing and direction in the exported file.
Privacy and data handling	What happens to what you enter?	Read the terms of use and data policy: training on inputs, retention period, storage location, and whether a data-processing agreement is available.
Export and portability	Can you take your work and leave?	Export a real output and open it in the tool you normally use.

Criterion	Core question	How to test it
Cost fit	Does the cost fit your budget and volume?	Work out the cost for your real number of users and expected volume from the official pricing page on the day you compare.
Reliability	Is quality consistent?	Run the same case three times and log any outage or error during the test week.
Integration	Does it work with your current tools?	Try one complete workflow, from the data source to where the final output lives.

How the formulas work

In the spreadsheet, each criterion has a weight out of 100 (column C) and each tool a score from 1 to 5 (columns D, E and F). Criterion points = weight × score ÷ 5. That way a tool scoring 5 on everything gets exactly 100, and the total reads like a percentage.

- **G2 to I8:** points per criterion per tool, e.g. `C2*D2/5`.
- **C9:** total of weights `SUM(C2:C8)`; C10 tells you if it is not 100.
- **G9 to I9:** total points per tool.
- **Row 11:** you type "yes" or "no": does the tool pass the knock-out conditions?
- **D12 to F12:** the final result; a tool marked "no" shows "Excluded" whatever its points.
- **D13:** the name of the highest-scoring eligible tool, taken from the column headers D1:F1, so rename those headers to your tools.

Knock-out conditions: why the total is not enough

A weighted average lets a tool that is very weak on a critical criterion make up for it elsewhere. That is fine for most criteria but dangerous for some. So before testing, set one or two conditions that allow no compensation, for example:

- If inputs are sensitive (customer, patient or staff data), a tool that does not let you switch off training on inputs or offer a data-processing agreement is excluded.
- If the output will be published in Arabic as is, a tool scoring 1 on Arabic quality is excluded.
- If the budget is a hard ceiling, a tool that exceeds it is excluded however good it is.

How to use it

1. **Start from the task.** Take one task from your AI use map (resource 13). A tool that suits product descriptions may not suit meeting summaries.
2. **Set the weights before seeing the tools.** Agree the weights in a short meeting and save the file. Changing weights after seeing results is the easiest way to fool yourself.
3. **Write the knock-out conditions.** One or two, each in a clear sentence in the file.
4. **Prepare the test sample.** Five to ten real, anonymised cases and fixed instructions given word for word to all three tools.

5. **Test and record evidence.** For every score, note why: an example of an error, a screenshot of the terms, an exported file.
6. **Score with a second person.** Two people score separately, then discuss any gap larger than one point.
7. **Read the result as a suggestion.** A small gap (under 5 points) is not a real advantage; settle it with a two-week trial or on cost.

Blank template (text version)

Tool selection card

Specific task: [from the use map]

Comparison date: [day/month/year]

Candidates: Tool 1 = [name] | Tool 2 = [name] | Tool 3 = [name]

Weights (total 100):

Task fit [] | Arabic quality [] | Privacy [] | Export []

Cost [] | Reliability [] | Integration []

Knock-out conditions (no compensation):

1. [e.g. no data-processing agreement with sensitive inputs]

2. [e.g. a score of 1 on Arabic quality]

Test sample: [number of cases, source, how they were anonymised]

Fixed instructions given to the tools: [paste the text]

Scores 1–5 with short evidence for each:

[Criterion] — Tool 1: [] because ... | Tool 2: [] because ... | Tool 3: [] because ...

Result: [suggested tool] | Next step: [two-week trial / negotiate / retest]

Scorers: [two names]

A complete example: a clinic choosing a tool for patient messages

Illustrative example

"Nabd Clinic" is a fictional clinic in Tripoli that wants a tool to help its receptionist draft appointment reminders and answers to general questions (opening hours, how to prepare for tests). It compared three tools, labelled "Tool 1", "Tool 2" and "Tool 3" in the file.

Weights: task fit 25, Arabic quality 20, privacy 20, export 5, cost 10, reliability 10, integration 10. Privacy got a high weight because some messages might include a patient's name or appointment.

Knock-out condition: any tool that neither lets you switch off training on inputs nor offers a data-processing agreement is excluded.

Criterion	Weight	Tool 1	Tool 2	Tool 3
Task fit	25	4 (20)	5 (25)	3 (15)
Arabic quality	20	3 (12)	4 (16)	4 (16)
Privacy	20	4 (16)	2 (8)	5 (20)
Export	5	4 (4)	3 (3)	4 (4)
Cost	10	3 (6)	4 (8)	2 (4)
Reliability	10	4 (8)	4 (8)	3 (6)
Integration	10	3 (6)	4 (8)	2 (4)
Total	100	72	76	69
Knock-out passed		yes	no	yes
Result		72	Excluded	69

Reading it: Tool 2 had the highest total (76) thanks to task fit and integration, but it failed the privacy condition and was excluded. Between the two eligible tools, Tool 1 leads Tool 3 by only three points, which is a small gap. The clinic decided to trial Tool 1 for two weeks on reminder messages only, without patient names in the first phase, and to reassess afterwards. If Tool 2 changes its terms later, it can re-enter the comparison.

Common mistakes

- **Mistake:** comparing tools "in general" with no specific task. **Fix:** one task and one real sample per scorecard.
- **Mistake:** adjusting weights after seeing the scores so the favourite wins. **Fix:** lock the weights and save a copy of the file before testing.
- **Mistake:** judging privacy from the marketing page. **Fix:** read the actual terms of use and data policy, and note the date you read them.
- **Mistake:** copying prices from an article or a colleague's memory. **Fix:** take cost from the official page on the day you compare, for your real number of users.
- **Mistake:** testing each tool with different instructions. **Fix:** identical instructions for all three, otherwise you are comparing your prompts, not the tools.
- **Mistake:** treating a two- or three-point gap as decisive. **Fix:** small gaps are settled by a short trial, not by the spreadsheet.

Completion checklist

- ☐ I chose one clear task from the use map.
- ☐ We locked weights totalling 100 before testing any tool.
- ☐ We wrote at least one knock-out condition, and a privacy condition if inputs are sensitive.

- ☐ The sample is real and anonymised, and all three tools got the same instructions.
- ☐ Every score has written evidence (an example, a screenshot, an exported file).
- ☐ Two people scored separately and discussed the gaps.
- ☐ We recorded the comparison date, because tools and their terms change.
- ☐ We set the next step: a short trial with a clear measure before any long subscription.

About the author — Mohamad Charafeddine is an entrepreneur from Tripoli, Lebanon: founder of the Passion digital marketing agency, president of Takaful Youth Association, and founder of digital platforms. He trains individuals and teams in applied AI and marketing.

Examples in this file are fictional. Free to use in your work and organisation; credit the source if you republish.
mohamadcharafeddine.com/en/resources/ai-tool-selection/