

AI BUYER EVALUATION · 2026

Best VoIP Provider With AI

Every quote says AI, so the word has stopped carrying information. Here are nine tests you can run yourself in a trial week, and the three questions a demo never reaches.

Feature lists converged during 2025. Every Australian cloud phone quote now offers an AI receptionist, transcription, summaries and call scoring, in roughly the same words at roughly the same price. What has not converged is **what happens on the call**, and that is measurable in about two hours with nothing but a phone.

■ The four layers behind any AI phone feature

Layer	What it decides for you	Ask this
Network	Call quality, and the audio the model actually hears. Also whether a fault can be traced end to end.	"Do you operate your own network, or buy wholesale from someone who does?"
Platform	What you can configure, and whether the AI is genuinely in the call path or bolted beside it.	"Is the AI a step inside your call flow, or a forwarded call to a separate product?"
Model	Accuracy, latency, tone, and what happens when the upstream model changes.	"Whose models, in which region, and what happens to my settings when they update?"
Integration	Whether the AI produces work, or only produces text.	"Show me the record it writes, in my CRM, on a live call."

Why the layer question is not academic

Outbound calls start getting labelled as spam and the AI receptionist starts getting hung up on. Is that the network, the platform, or the model? A provider who owns the first two can look at both and answer. A provider who resells both can only raise a ticket with two other companies and relay the replies. **The number of companies between you and the fix is a real product attribute**, and it never appears on a feature list.

■ Nine tests, one afternoon, same script for every provider

- ☐ **Interrupt it.** Talk over it mid sentence. Does it stop and follow your point, or finish its script?
- ☐ **Local proper nouns.** Say three suburbs, two customer surnames, one product name. Read the transcript.
- ☐ **Time the gap.** Under a second reads as conversation. Past two, callers say "hello?" and that gets transcribed.
- ☐ **Check a transcript you know.** Then read the summary built on it. Confidently wrong is worse than absent.
- ☐ **Ask for a human** at the greeting, mid flow, and after giving details. All three should work.
- ☐ **Ring it at 9pm** and on a Saturday. Demos happen at 11am on Tuesdays. Your hard calls do not.
- ☐ **Follow the result into your CRM.** Not their dashboard. Right contact, right fields, inside a minute.
- ☐ **Ask what happens when it is down.** The right answer is a defined fallback, plus alerting.
- ☐ **Ask it something it cannot know.** It should say so and offer a person, not invent an answer.

■ Three questions the demo never reaches

Question	Why it matters	A good answer
Where is the audio processed, and where does it rest?	Processing and storage are different things and can happen in different countries.	A named region for both, in writing, plus a clear no to training on your audio.
How long is it kept, and who can retrieve it?	Retention is a decision, not a default. Too short loses your evidence, too long holds personal information for no reason.	Per record type retention, deletion that deletes, role based access, an inspectable log.
Who can see the call when something is wrong?	Somebody has to look at network, routing and model behaviour together.	Local support who can open your actual call, not a ticket queue and a target response time.

■ Automate freely

- Answering every call so nothing rings out
- Capturing name, number, reason and callback time
- Your twenty most asked published questions
- Transcribing, summarising and filing
- Booking with capacity rules, and confirmations

■ Never automate

- Whether a situation is an emergency
- Approving credit, refunds or expenditure
- Anything affecting rights, housing, care or employment
- A caller in distress, or a vulnerable caller
- Screening someone out

The test for anything in between is reversibility

If a person can fix the mistake within the hour, automate it and review the exceptions. If the customer carries the consequence, a person decides and the AI prepares the decision. Ask each provider to show you where that boundary is set in their console. **If it cannot be set, it has already been set for you** by somebody who has never met your customers.

Turning the tests into a decision

Eliminate anyone who invents answers, will not state where data is processed, or has fixed retention. Rank the rest on the nine behavioural tests, because that is what callers experience every day for the next three years. Use layers owned and support reachability as the tie break, since they decide what happens on the worst day. **Compare price last**, as a five year total divided by sixty months and by seat count. In practice the behavioural spread between providers is far wider than the price spread, and the deciding tests are never the ones in the sales deck.