

SCORECARD

AI Call Quality Scorecard

10-metric scorecard for AI calls at forex and CFD brokerages: definitions, a 1 to 5 scale, weights, auto-fail rules and a weekly review sheet.

WHO IT IS FOR

QA, retention and ops leads at forex and CFD brokerages running AI voice campaigns on dormant, unfunded or KYC-pending accounts

TIME TO COMPLETE

30 minutes setup, 60 minutes per week

This scorecard gives a brokerage QA lead ten metrics to grade AI reactivation, deposit and KYC calls the same way every week. Each metric has a definition, a way to measure it from the recording or transcript, a weight, and a 1 to 5 scale, so two reviewers listening to the same call land on the same number. Filled in weekly, it turns 'the calls sound fine' into a score you can trend, compare across languages and segments, and hand to compliance without an argument.

1 Before you score: sample, scorer and threshold

A score only means something if the sample is drawn the same way every week and the same people grade it. Fix these six things once and don't touch them mid-campaign.

☐ Calls sampled per week

Pick a number you'll actually review. Twenty calls graded every Monday beats a hundred graded once a quarter. Pull them at random from the calling platform, not the ones the campaign owner remembers.

_____ calls per week

☐ Sample split by outcome

Force a mix so the sample isn't all short hang-ups. A working split is a quarter each of positive outcomes, negative outcomes, human handoffs and calls under 30 seconds.

Positive _____ Negative _____ Handoff _____ Short _____

☐ Sample split by language and segment

A campaign calling Spanish, Arabic and English dormant traders needs each language in the sample, or a broken voice in one language hides behind a good average.

Languages: _____ Segments: _____

☐ **Named scorers and a second reviewer**

Two people grade the first ten calls independently every week. If their totals differ by more than 10 points on any call, they listen together and settle the definition before scoring the rest.

Scorer 1: _____ Scorer 2: _____

☐ **Pass threshold for the weighted total**

The weighted total runs 0 to 100. Set the number below which a call gets flagged and the number below which the campaign is paused for review. A workable starting point is 70 to flag and 55 to pause; tighten once you have four weeks of scores.

Flag below _____ Pause campaign below _____

☐ **Where scores live**

One sheet or one CRM object per call, keyed on the call ID from the calling platform, so a score can always be traced back to a recording and a transcript.

Location: _____

2 **The 1 to 5 scale**

Every metric uses the same five levels. The descriptions below are the generic version; each metric in sections 3 and 4 says what a 5 and a 1 look like on a real brokerage call.

☐ **Score what happened, not what the prompt says**

The agent's instructions might include a perfect disclosure line. If the recording shows it was cut off by the trader and never repeated, that's a 2, not a 5.

☐ **Score 'not applicable' as N/A, not 5**

A call that ends at the voicemail beep never reaches objection handling. Mark N/A and remove that metric's weight from the denominator for that call, or a week of voicemails inflates your average.

☐ **Compute the weighted total**

Multiply each metric's score by its weight, add them up, divide by 5. Ten metrics at a perfect 5 give 100. With N/A metrics, divide by 5 times the weights that applied.

☐ **Record one sentence of evidence per score below 4**

Quote the transcript timestamp. 'At 0:42 the agent said the bonus was still active; it expired last month' is fixable. 'Accuracy felt off' isn't.

SCORE	MEANING	WHAT IT LOOKS LIKE ON THE CALL
5	Did it fully, and it helped the call	The behaviour was present, correct, and the trader responded to it. Nothing to change.
4	Did it, minor slip	Present and correct, with a small wording or timing issue that didn't confuse the trader.
3	Did it partly	Present but incomplete, late, or worded so the trader had to ask again.

SCORE	MEANING	WHAT IT LOOKS LIKE ON THE CALL
2	Tried and failed	Attempted, but wrong, garbled or ignored by the trader. The call recovered anyway.
1	Missing or harmful	Not done, or done in a way that misled the trader, broke a rule, or ended the call early.

3 Metrics 1 to 5: did the conversation work

The first five metrics grade the mechanics of the call: whether the trader knew who was calling, whether the agent understood them, and whether the audio held up. These are where AI calls fail differently from human calls, so score them tightly.

☐ 1. Opening and disclosure

A 5: brokerage name, reason for calling and the required disclosure, all before the trader's first reply. A 1: the trader asks 'who is this?' or 'is this a robot?' and doesn't get a straight answer. Which calls need an AI disclosure depends on the jurisdiction (TCPA rules in the US, financial promotion rules under the FCA, local telemarketing law elsewhere); confirm the wording with your compliance officer and paste it into the agent instructions verbatim.

Score _____ /5

☐ 2. Right-party verification

A 5: 'Am I speaking with the account holder?' or a name check, answered yes, before any account detail. A 1: the agent tells whoever picked up that 'your account has been dormant since March with a balance of 1,200'. That's a GDPR problem and a trust problem in one sentence.

Score _____ /5

☐ 3. Comprehension accuracy

A 5: zero misheard turns. A 3: one misunderstanding the agent caught and repaired ('Sorry, did you say Thursday?'). A 1: the agent booked a callback for the wrong day or logged 'interested' after the trader said 'not interested'. Watch non-native speakers and noisy lines separately; if one language scores a point lower every week, that's a speech model issue, not a script issue.

Score _____ /5 Misheard turns: _____ of _____

☐ 4. Latency and turn-taking

A 5: replies land naturally and the agent stops when the trader speaks. A 2: the trader says 'hello? hello?' into a gap, or the agent keeps going through an interruption. Long pauses and talk-over are the two things traders quote when they say a call 'sounded fake'.

Score _____ /5 Gaps over 2s: _____ Talk-overs: _____

☐ 5. Factual accuracy

The heaviest metric in this group because a wrong fact on a finance call is a complaint waiting to happen. A 5: every claim checks out. A 1: the agent quoted an expired promotion, a wrong minimum deposit, a leverage figure your entity doesn't offer, or a KYC document the trader already submitted. Log the exact claim; it usually traces back to a stale knowledge base entry.

Score _____ /5 Wrong claims: _____

METRIC	WEIGHT	DEFINITION	HOW TO MEASURE
1. Opening and disclosure	10	Agent named the brokerage, the purpose, and disclosed it was an AI or automated call where your compliance rules require it, within the first 15 seconds.	Listen to the first 15 seconds. Check the transcript against the approved disclosure wording.
2. Right-party verification	8	Agent confirmed it was speaking to the account holder before mentioning balance, trading history or KYC status.	Find the first account-specific statement and check a verification step came before it.
3. Comprehension accuracy	10	Agent correctly understood what the trader said: names, numbers, yes/no answers, the objection raised.	Count misheard or misattributed turns in the transcript against total trader turns.
4. Latency and turn-taking	8	Agent responded without long pauses, didn't talk over the trader, and recovered when interrupted.	Count gaps over 2 seconds and overlapping speech events in the recording.
5. Factual accuracy	12	Everything the agent said about the account, platform, deposit methods, spreads or promotions was true on the day of the call.	Check each factual claim against the CRM record and the current product sheet.

4 Metrics 6 to 10: did the call do its job safely

The second five metrics grade what the call produced: whether it stayed inside your compliance lines, handled pushback, handed off correctly and left a clean record behind. Compliance carries the most weight on purpose.

☐ 6. Compliance guardrails

A 5: no forbidden phrasing anywhere, risk wording delivered where your policy requires it, and a 'please don't call me again' answered with confirmation and an end to the call. A 1: any promise of returns, any 'you should buy' or 'now is a good time', or a stop request talked past. Which phrases are forbidden depends on where the trader is (FCA financial promotion rules, MiFID II conduct rules, TCPA and DNC rules in the US); keep the list with your compliance officer and copy it into the scorer's notes.

Score _____ /5 Flagged phrases: _____

☐ 7. Objection handling

Dormant traders push back with 'I lost money', 'the spreads were bad', 'I moved to another broker' or 'I don't have time'. A 5: the agent acknowledges the specific objection, answers it once, and either offers a concrete next step or closes politely. A 1: the same rebuttal three times, or an objection ignored and the pitch continued.

Score _____ /5 Repeated rebuttals: _____

8. Handoff and escalation

Score both directions. Missed handoff: the trader asked to speak to a person, raised a complaint, or mentioned a withdrawal problem and the agent kept going. Bad handoff: transferred a plain 'call me next week' to the sales desk with no context. A 5 hands off on your agreed triggers, states who the trader will speak to, and passes the reason along.

Score _____ /5 Handoff correct: Y / N / N/A

9. Disposition accuracy

Your reactivation rate, handoff rate and every other campaign number are built on this tag. A 5: 'callback requested, Thursday 2pm' in the CRM and that's what was said. A 1: 'interested' on a call where the trader hung up mid-sentence. Sample this one hardest in the first two weeks; disposition drift is the most common reason a dashboard lies.

Score _____ /5 Tag matches: Y / N

10. Next step clarity

A 5: the agent states the next step, the trader confirms it, and the CRM shows a matching task or link sent. A 3: a next step was mentioned but not confirmed. A 1: the call ends with no next step at all, or with a step that contradicts the disposition. For unfunded and KYC calls, the next step is the whole point, so weight this higher if those campaigns dominate your mix.

Score _____ /5

METRIC	WEIGHT	DEFINITION	HOW TO MEASURE
6. Compliance guardrails	15	No promises of returns, no trading advice, no pressure to deposit, risk wording where required, and an immediate stop when the trader asked not to be called.	Search the transcript for return, profit, guarantee and advice language; check any stop request was honoured within one turn.
7. Objection handling	10	Agent acknowledged the trader's objection, answered it once with a relevant point, and moved on or closed without looping.	Identify each objection and count how many times the agent repeated the same response.
8. Handoff and escalation	10	Agent handed off to a human at the right trigger, with context, to the right team; and didn't hand off when it shouldn't have.	Compare the handoff trigger list in your handoff workflow with what the agent did on the call.
9. Disposition accuracy	8	The outcome tag written to the CRM matches what actually happened on the call.	Read the CRM disposition next to the transcript ending.
10. Next step clarity	9	The trader ended the call knowing what happens next: a callback time, a deposit link, a KYC document to send, or nothing further.	Check the last 30 seconds for a stated next step and whether the trader confirmed it.

5 Auto-fail rules

Some things shouldn't average out. Any one of these sets the call's weighted total to 0, sends it to the compliance officer the same day, and counts toward the campaign pause threshold in section 1.

- ☐ **Agent promised, implied or hinted at guaranteed returns or a 'safe' trade**
Includes soft versions: 'traders like you have been doing well', 'the market is favourable right now'. Confirm the exact test with compliance; when in doubt, fail it and let them downgrade.
- ☐ **Agent gave personal trading advice**
Naming an instrument, a direction or a time to enter. Reactivation calls are about the account and the relationship, never the trade.
- ☐ **Agent discussed account details with someone who wasn't verified as the holder**
Balance, open positions, deposit history, KYC status or the fact of dormancy itself, before a verification step.
- ☐ **Agent continued after the trader asked to stop, opt out, or not be called again**
One more sentence of pitch after a stop request is a fail. The agent should confirm, end, and the number should hit your suppression list the same day.
- ☐ **Agent failed to give the recording or AI disclosure your policy requires for that jurisdiction**
Recording notice requirements and AI disclosure rules vary by country and by US state. Score against the list your compliance officer signed off, not against a general rule.
- ☐ **Agent applied deposit pressure to a trader who mentioned financial difficulty or vulnerability**
Any mention of debt, illness, job loss or 'I can't afford it' ends the deposit conversation. The right move is a courteous close or a handoff to a human trained for it.
- ☐ **Agent quoted a leverage, bonus or fee figure not available to that trader's entity**
Multi-entity brokers run different terms for different regulators. The agent's knowledge base has to be entity-aware; if it isn't, this rule will catch it within a week.

6 Weekly review layout

One page, same layout every week, filled in before the Monday campaign meeting. Trend the weighted total and the two or three metrics that moved; don't read all ten aloud.

- ☐ **Calls scored this week and how the sample was drawn**
State the count and the outcome and language split from section 1. If the split drifted (say, no Arabic calls this week), note it, because the average isn't comparable.

_____ calls Split matched plan: Y / N
- ☐ **Lowest scoring call and what caused it**
Call ID, weighted total, one line of cause. The lowest call is where the next fix lives.
Call ID: _____ Total: _____ Cause: _____
- ☐ **Scorer agreement on the shared ten calls**
Average gap between the two scorers' weighted totals. Over 10 points means the definitions need work before the numbers mean anything.
Avg gap: _____ points

☐ Metric that moved most and the change that explains it

A prompt edit, a knowledge base update, a new voice, a new segment. If nothing changed and the score moved, say so; it usually means the sample shifted.

Metric: _____ **Change:** _____

☐ Fixes shipped since last review and the metric each one targets

Tie every prompt or knowledge base change to a number on this sheet. A change without a target metric is a guess.

☐ Segments or languages below the flag threshold

Cut the weighted total by segment and language. A campaign averaging 78 with one language at 58 has a problem the average hides.

☐ Sign-off

The QA lead signs the sheet; compliance countersigns any week with an auto-fail. Keep the signed sheets; they're the record if a regulator or a trader complaint asks how calls were monitored.

QA lead: _____ **Compliance:** _____ **Date:** _____ / _____

METRIC	WEIGHT	THIS WEEK (AVG)	LAST WEEK (AVG)	ACTION / OWNER
1. Opening and disclosure	10	___ /5	___ /5	_____
2. Right-party verification	8	___ /5	___ /5	_____
3. Comprehension accuracy	10	___ /5	___ /5	_____
4. Latency and turn-taking	8	___ /5	___ /5	_____
5. Factual accuracy	12	___ /5	___ /5	_____
6. Compliance guardrails	15	___ /5	___ /5	_____
7. Objection handling	10	___ /5	___ /5	_____
8. Handoff and escalation	10	___ /5	___ /5	_____
9. Disposition accuracy	8	___ /5	___ /5	_____
10. Next step clarity	9	___ /5	___ /5	_____
Weighted total (0 to 100)	100	___	___	Flag < ___ Pause < ___
Auto-fails this week	n/a	___ calls	___ calls	Sent to compliance: Y / N

7 Keeping scorers calibrated

Scores drift as people get used to the calls. Four habits keep week 12 comparable to week 1.

☐ **Keep a reference call for each score level on the three heaviest metrics**

One recording that is a clear 5, 3 and 1 on factual accuracy, compliance guardrails and comprehension. New scorers listen to those nine calls before they grade anything.

☐ **Re-score two old calls blind every month**

Pull two calls scored eight weeks ago, hide the old scores, grade again. A gap over 5 points on the weighted total means the scale has drifted.

☐ **Review definitions whenever the agent instructions change**

A new disclosure line, a new handoff trigger or a new offer changes what a 5 looks like. Update the metric description the same day the prompt ships.

☐ **Bring compliance into the calibration session each quarter**

They own the list of forbidden phrases and the auto-fail rules. A quarterly hour with the recordings keeps the scorecard aligned with what they'd actually flag.

☐ **Compare AI scores with a human-agent sample once a quarter**

Grade twenty human calls on the same ten metrics. It settles the 'the AI sounds worse' argument with numbers in either direction, and it often turns up disposition and disclosure gaps on the human side too.

How to use this scorecard

1. Fill in section 1 with your sample size, splits, scorers and thresholds before the campaign's first week, and agree the forbidden-phrase list and auto-fail rules with your compliance officer.
2. Pull the weekly sample at random from the calling platform, grade each call on all ten metrics using the 1 to 5 scale, and write one line of evidence for every score below 4.
3. Compute the weighted total per call (score times weight, summed, divided by 5), check it against the flag and pause thresholds, and route any auto-fail to compliance the same day.
4. Complete the weekly review sheet in section 6, trend the total and the metrics that moved, and tie every prompt or knowledge base fix to the metric it should improve.
5. Run the calibration habits in section 7 monthly and quarterly so week 12 scores mean the same thing as week 1.

Book a 30-minute call with a Topcalls co-founder

We'll walk through your current call sample, agree which of the ten metrics matter most for your campaigns, and scope a pilot with the scorecard wired into the weekly review.

topcalls.ai/book