

AI Voice Call A/B Test Plan Template

A/B test plan template for AI voice calling at forex and CFD brokers: hypothesis, one variable, control vs variant, sample size, metric, decision rule.

WHO IT IS FOR

Retention, campaign and ops leads at forex and CFD brokerages running AI voice calls to dormant traders, leads or unfunded accounts

TIME TO COMPLETE

30 minutes

Most brokerage calling tests fail before the first dial: two things change at once, nobody wrote down the metric, and the result gets argued over in a meeting. This template fixes the plan on one page before launch. You write the hypothesis, pick exactly one variable, define control and variant, size each arm from your own dormant book, name the primary metric and the guardrails, set the duration and stopping rules, and agree the decision rule with whoever signs off. Filled in, it's the sheet you hand to compliance, the retention lead and the calling platform admin, and the record you keep for the next test.

1 Hypothesis and the one variable

Write the test as a sentence you could be wrong about. If you can't name the single thing that changes, you're running two campaigns, not a test.

☐ Test name and ID

Use the ID in the calling platform campaign name, the CRM tag and the log at the end of this template, so the exports join without a spreadsheet detective.

ID: AB- _____ Name: _____

☐ Hypothesis, written as 'If we change X for segment Y, then metric Z moves by at least N'

Example: 'If we open with the trader's last instrument instead of a generic greeting for 90 to 180 day dormant accounts, positive outcome rate rises from 12% to 15%.' The numbers are yours; the shape is fixed.

If we _____, then _____ by at least _____

☐ The one variable that changes

Pick from: opening line, offer, call window, voice or language, attempts per trader, days between attempts, handoff trigger, voicemail script. One. The disclosure and consent lines stay fixed in every arm.

Variable: _____

☐ **Segment the test runs on**

State the rule, not the nickname: dormancy window, funded or unfunded, jurisdiction, last platform (MT4, MT5, cTrader). Both arms draw from the same segment, or the segment becomes the variable and you didn't mean it to.

Segment: _____

☐ **Why this variable, in one line**

Tie it to a number you already have. 'Connect rate is fine at 38% but conversations end in the first 20 seconds' points at the opening line, not the offer.

Reason: _____

☐ **What you'll do with a win**

If the answer is 'nothing, we're curious', pick a different variable. A test earns its minutes when the winning arm can roll out to the rest of the book.

Rollout plan: _____

2 Control and variant

The control is what you run today. The variant differs in the one variable and nothing else. Fill the table, then have someone who didn't write it check the 'held constant' column.

☐ **Control (A), as run today**

Copy the current campaign settings from the calling platform. If 'today' isn't stable yet, run the control alone for a week first; you can't measure a change against a moving baseline.

A: _____

☐ **Variant (B), one change only**

Describe the change in the words the AI agent will actually say or the setting that actually moves. 'Warmer opening' isn't a variant; the new opening line is.

B: _____

☐ **Random assignment method**

Assign each trader to an arm before the first dial, by account ID, not by list order or by which agent picks up the record. Alternating rows or 'A in the morning, B in the afternoon' both leak the call window into the result.

Method: _____

☐ **Compliance review of both scripts**

Both arms carry the same AI disclosure, recording notice and opt-out path. Financial promotion wording (FCA financial promotion rules, MiFID II where it applies) gets reviewed on the variant as well as the control. Confirm with your compliance officer or counsel before launch.

Reviewed by: _____ **Date:** _____

ELEMENT	CONTROL (A)	VARIANT (B)	HELD CONSTANT?
Opening line	_____	_____	Yes / No

ELEMENT	CONTROL (A)	VARIANT (B)	HELD CONSTANT?
Offer or reason to return	_____	_____	Yes / No
Call window (trader's local time)	_____	_____	Yes / No
Voice and language	_____	_____	Yes / No
Attempts per trader and days between them	_____	_____	Yes / No
Handoff trigger to a human	_____	_____	Yes / No
Voicemail behavior	_____	_____	Yes / No
Disclosure, recording notice and opt-out line	Fixed	Fixed	Always
Suppression and DNC rules	Fixed	Fixed	Always

3 Sample size and split

Size the test from your own baseline, not from a benchmark. A test that ends with 9 positive outcomes in one arm and 11 in the other tells you nothing, whatever the percentages look like.

☐ Baseline rate of the primary metric, from the last 30 days of the control

Pull it from the calling platform's outcome tags for the same segment. If fewer than a few hundred calls sit behind it, note that the baseline is soft and plan a longer test.

Baseline: _____ %

☐ Minimum lift worth acting on

The smallest change that would make you roll out the variant. Smaller lifts need more calls to detect, so pick the number you'd actually act on, not the one that makes the test short.

Minimum lift: _____ points

☐ Confidence level and calculator used

Pick a level, record it, and use any standard two-proportion sample size calculator with the baseline, the minimum lift and the confidence. Write the tool's name so the next test uses the same one.

Confidence: _____ % Tool: _____

☐ Calls needed per arm, from the calculator

Per arm: _____ calls

☐ Traders needed per arm, given your attempts per trader

Calls per arm divided by planned attempts per trader, rounded up. Then check the segment is large enough after suppression to fill both arms without reusing anyone.

Per arm: _____ traders

☐ Segment size after suppression and DNC scrub

If the segment can't fill both arms, widen the dormancy window, raise the minimum lift, or run the test longer. Don't shrink the arms and hope.

Available: _____ traders

☐ Planned calling cost for the test

Total calls times average billed minutes times your per-minute rate. On Topcalls that rate is \$0.35/min all-inclusive, so 2,000 calls averaging 1.5 billed minutes come to about \$1,050 for the whole test.

Est. cost: \$ _____

4 Primary metric and guardrails

One primary metric decides the test. Guardrails can pause it or veto a rollout, but they don't pick the winner. Define each one before launch, with its formula and where the number comes from.

☐ Primary metric for this test

Pick the fastest metric that still reflects money. Opening-line tests read on positive outcome rate within days; offer tests usually need reactivation or deposit rate and a longer window.

Primary: _____

☐ Attribution window for the primary metric

Same window for both arms, counted from each trader's last attempt, not from the campaign start date.

Window: _____ days after last attempt

☐ Outcome tags the AI agent may use, fixed for both arms

'Will fund', 'wants callback', 'KYC help needed', 'not interested', 'do not call again'. If the tag list differs between arms, the primary metric isn't comparable.

Tags: _____

☐ Guardrail thresholds that pause an arm

Set them from the control's history. A variant that wins on outcomes and doubles opt-outs doesn't roll out.

Opt-outs > _____ per 100; complaints > _____

☐ Where the numbers live and who pulls them

One sheet, one owner, both arms side by side. Exports from the calling platform and the trading platform join on account ID and the test ID.

Sheet: _____ Owner: _____

METRIC	ROLE	FORMULA	DATA SOURCE
Positive outcome rate	Primary (fastest read)	Calls tagged with an agreed next step / conversations	Calling platform outcome tags
Reactivation rate	Primary (slower)	Accounts that logged in or traded inside the window / traders reached	MT4/MT5 or cTrader activity export, joined on account ID

METRIC	ROLE	FORMULA	DATA SOURCE
Deposit rate	Primary (slowest)	Accounts that deposited inside the window / traders reached	Payments system or CRM
Connect rate	Guardrail	Answered calls / dial attempts	Calling platform call log
Opt-out requests per 100 conversations	Guardrail (pause)	Opt-outs / conversations x 100	Calling platform plus CRM suppression log
Complaints, by type	Guardrail (pause)	Count per arm	Support desk and compliance log
Human handoff acceptance	Guardrail	Handoffs accepted by the desk / handoffs offered	Calling platform transfer log plus desk calendar
Cost per positive outcome	Secondary	Billed calling cost / positive outcomes	Usage report plus outcome tags

5 Duration and stopping rules

Decide when the test ends before it starts. Checking the dashboard on day two and stopping when the variant is ahead is how teams roll out a coin flip.

☐ Planned start and end dates

Cover at least one full week per arm so weekday and weekend call windows both appear in each arm. Avoid a start that overlaps a major market event, a platform migration or a bonus promotion.

Start: _____ End: _____

☐ Stop rule: sample reached

The number from section 3. The test ends when both arms hit it, not when one does.

Stop when each arm has _____ calls

☐ Stop rule: guardrail breached

Any threshold in section 4 pauses the breaching arm the same day. Log the pause, review it with compliance, then restart or kill.

Pause if: _____

☐ Stop rule: operational failure

Wrong list loaded, disclosure missing from an arm, assignment leaked, handoff queue unstaffed. Any of these invalidates the arm; don't try to rescue the data.

Kill if: _____

☐ No early declaration

Write it down: nobody declares a winner before the end date and the sample, however good day three looks. If the business needs an answer sooner, shorten the test on paper and re-size it, don't cut it live.

☐ **Interim reads and who sees them**

A daily guardrail read is fine. Keep the primary metric split out of the daily view so it can't drive a premature call.

Read on: _____ **Seen by:** _____

☐ **Freeze list: what nobody changes during the test**

Scripts, offer, call windows, attempt rules, suppression files, voice, handoff routing. If something must change, both arms change on the same day and the change is logged.

Frozen: _____

6 Decision rule and sign-off

Agree the outcome table before launch, so the result reads itself and the meeting is about rollout, not interpretation.

☐ **Minimum lift carried over from section 3**

Roll out if lift is at least _____ **points**

☐ **Rollout mechanics if the variant wins**

Who edits the live campaign, which segments it applies to, and when the old control is switched off. On Topcalls the change is a campaign edit, so name the person with access.

Rollout on: _____ **by:** _____

☐ **Where the result is recorded**

One line in the test log below plus the filled plan, stored where the next campaign owner will find it. Tests that live in someone's inbox get run twice.

Record: _____

☐ **Sign-off before launch**

Three signatures on the plan, not on the result. Compliance confirms both scripts, retention confirms the metric and the decision table, ops confirms the list, the split and the freeze.

Compliance: _____ **Retention:** _____ **Ops:** _____

RESULT AT END OF TEST	DECISION	WHO SIGNS
Variant beats control by at least the minimum lift, guardrails clean	Roll the variant out to the full segment; log it as the new control	Retention lead, compliance
Variant beats control by less than the minimum lift	Keep the control; note the direction and try a bigger change next test	Retention lead
No difference, or control wins	Keep the control; retire the variant idea for this segment	Campaign owner
Variant wins on the primary metric but breaches a guardrail	No rollout; fix the cause and re-test, or drop it	Compliance, retention lead

RESULT AT END OF TEST	DECISION	WHO SIGNS
Test invalidated (operational failure, assignment leak)	Discard the data; re-run with the fault fixed	Campaign owner

7 Test log

Keep every test on one sheet. Twelve lines of this table are worth more than any vendor's benchmark, because they're your book, your segments and your traders.

☐ Next test candidate, from the loser's notes

Every test produces a hypothesis for the next one. Write it while the result is fresh, not at the next planning meeting.

Next: _____

☐ Variables already tested on this segment

Stops the new campaign owner re-running last quarter's opening-line test under a different name.

Done: _____

TEST ID	VARIABLE AND SEGMENT	RESULT VS CONTROL	DECISION	ROLLED OUT ON
AB- —	_____	___ points	Roll out / Keep control / Re-test	_____
AB- —	_____	___ points	Roll out / Keep control / Re-test	_____
AB- —	_____	___ points	Roll out / Keep control / Re-test	_____
AB- —	_____	___ points	Roll out / Keep control / Re-test	_____
AB- —	_____	___ points	Roll out / Keep control / Re-test	_____
AB- —	_____	___ points	Roll out / Keep control / Re-test	_____

How to use this template

1. Fill sections 1 and 2 first. If the hypothesis needs two sentences or the variant changes two rows of the table, split it into two tests.
2. Pull the baseline from the last 30 days of the control and size both arms in section 3 before you touch the list.
3. Fix the primary metric, the guardrails and the decision table, then collect the three sign-offs. Nothing goes live unsigned.
4. Load both arms as separate campaigns on the calling platform with the test ID in each name, assign traders by account ID, and run to the end date.
5. Read the result against section 6, fill one line of the log, and start the next plan from the loser's notes.

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

We size your dormant book, pick the first variable worth testing, and scope a pilot where both arms run on Topcalls with the outcome tags and exports this plan needs.

topcalls.ai/book