

Test definition

Complete the test definition before collecting results. Use one observation record per question, application, repetition, and attempt. Record unknown when a setting is not exposed and not applicable when a field does not apply.

Test ID and version

Unique identifier and version for this test.

Decision

We need to know whether ____ so we can decide ____.

Test mode

Unbranded discovery / branded accuracy / multi-turn journey.

Target business

Name, domain, and accepted variants. Keep in the assessment record; do not supply during discovery.

Exact prompt

Complete wording, punctuation, and substituted placeholders.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Test definition

Complete the test definition before collecting results. Use one observation record per question, application, repetition, and attempt. Record unknown when a setting is not exposed and not applicable when a field does not apply.

Prior conversation

None (fresh conversation), or reference the complete ordered transcript.

Other supplied context

Instructions, attachments, website content, memory, and custom instructions.

Application or API

Provider, product, interface, and model/version if exposed.

Account state

Signed in/out; personalization and memory settings; unknown if not exposed.

Language and location

Prompt language, requested geography, configured locale/location; record unknown if not exposed.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Test definition

Complete the test definition before collecting results. Use one observation record per question, application, repetition, and attempt. Record unknown when a setting is not exposed and not applicable when a field does not apply.

Search configuration

Enabled / disabled / automatic / unknown. Record observed search activity separately for each run.

Collection schedule

Planned dates, times, and timezone.

Repetitions

Planned repetitions per question and application.

Conversation reset

Fresh conversation for each independent repetition; preserve context within a multi-turn journey.

Assessment rule

Define mention, citation, recommendation, and factual-accuracy criteria separately before collection.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Test definition

Complete the test definition before collecting results. Use one observation record per question, application, repetition, and attempt. Record unknown when a setting is not exposed and not applicable when a field does not apply.

Supporting facts

Authoritative sources, captured facts, and capture dates.

Entity matching

Accepted identifiers and treatment of ambiguous business matches.

Unavailable runs

Record errors, refusals, blocked access, and incomplete answers separately; do not count them as absence.

Retry rule

Conditions for retry, maximum attempts, and which attempt counts. Preserve every original and link each retry to it.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Observation record – one per attempt

Test ID / version: _____ Record reference: _____

Test ID / version

Question ID

Repetition number

Attempt ID / retry of

Started at

ISO date/time and timezone.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Observation record – one per attempt

Test ID / version: _____ Record reference: _____

Application/API and model

Record unknown if not exposed.

Language/location

Search observed and evidence

Ran / did not run / unknown; activity display or tool-call reference.

Run status and reason

Completed / unavailable / partial; preserve the reason.

Full response/transcript reference

Retain the entire answer and preceding conversation.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Observation record – one per attempt

Test ID / version: _____ Record reference: _____

Cited URLs and supporting excerpts

Save the source URL and the passage supporting the assessed claim.

Target match

Confirmed / ambiguous / absent.

Mention

Yes / no / unclear / unassessable.

Citation

Yes / no / unclear / unassessable.

Recommendation

Yes / no / unclear / unassessable.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Observation record – one per attempt

Test ID / version: _____ Record reference: _____

Factual accuracy

Accurate / inaccurate / mixed / unassessable / not tested.

Assessment evidence and reviewer

Quote or reference supporting evidence; record reviewer and date.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Results summary

Test ID / version: _____ Record reference: _____

Planned repetitions

Total attempts including retries

Completed answers counted

Apply the predeclared counting rule.

Unavailable / partial / ambiguous

Counts and reasons; keep visible.

Mentions

Positive / assessable answers.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.

Results summary

Test ID / version: _____ Record reference: _____

Citations

Positive / assessable answers.

Recommendations

Positive / assessable answers.

Factual accuracy

Accurate / assessed claims.

Deviations

Changes to prompts, settings, schedule, or assessment rules.

Conclusion and limits

Under these conditions ____; this does not establish ____.

Repeated runs describe variability for the tested questions and conditions. They are not additional independent businesses and do not estimate the percentage of customers who will see the business.