



AI EXPERIMENT CANVAS

Define a worthwhile AI opportunity, design a credible test, manage material risks, and make an evidence-based decision.

Instructions and a worked example here: cleverbirdtech.org/canvas

Experiment: AI-drafted customer replies Date: Sep 1, 2026
Owner: Customer care lead Version: 1.0
Team: Customer care (3) Review date: Sep 22, 2026

01 OPPORTUNITY

Define the problem before proposing the technology.

Problem or opportunity:

Customers wait 26 hrs for a first reply. Most questions are repeats.

Who experiences it?

Customers waiting on answers; a 3-person care team.

Evidence and current baseline:

4 weeks of inbox data: 410 emails/wk, 26-hr median first reply, ~60% repeat questions.

Desired outcome:

First reply under 8 hrs with satisfaction held or improved.

Out of scope:

Refunds, complaints, damaged orders. No auto-sending.

02 AI FIT

Confirm that AI is appropriate and better than a simpler alternative.

AI capability (check all that apply)

- ☒ Generate ☒ Retrieve ☐ Summarize ☒ Classify
☐ Predict ☐ Recommend ☐ Automate ☐ Act
☐ Other

Why AI?

Same questions, endless phrasings. Templates miss context.

Simpler non-AI alternative:

Saved reply templates and a refreshed FAQ page.

Why AI is preferable:

Templates fit ~30% of emails. AI can adapt to the rest.

Human-led decisions that cannot be delegated:

Refunds, exceptions, and replies to upset customers.

03 PEOPLE AND WORKFLOW

Map who uses, governs, supports, and may be affected.

Primary users:

Care team (3)

Affected parties:

Customers

Owner:

Care lead

Human reviewers:

Sending agent

Tech / data support:

Ops manager

Escalation owner:

Shop owner

Required reviews (privacy, legal, security, domain): Privacy check of vendor data terms

Current workflow (as it happens today)

- Email lands
- Agent checks FAQ
- Writes reply
- Sends reply

Proposed workflow (what changes with AI)

- AI drafts reply
- Agent edits
- Agent approves
- Complaints: human

Human oversight points

- ☐ Before AI use
☐ During AI use
☒ Before output is acted upon
☒ When system fails or confidence is low

04 HYPOTHESIS

State what you believe will improve and what evidence will challenge it.

We believe using AI-drafted replies for routine customer emails will improve median first reply time from 26 hours to under 8 hours because agents edit a draft instead of writing from scratch.

Supported when:

Reply time drops, satisfaction holds, few rewrites.

Unsupported when:

Editing is as slow as writing, or errors reach customers.

Most uncertain assumption:

Drafts will be accurate enough to edit, not rewrite.

05 EXPERIMENT DESIGN

Define the smallest credible test.

What will be tested: Order + shipping replies

Participants or test cases: 2 agents, 150 emails

Baseline or comparison: Prior 4 weeks

Duration and environment: 3 weeks, live inbox

Minimum credible evidence: 150 drafts scored

Tool / model: Business AI assistant

Prompt / configuration / workflow version: v1

Prompt v1 + FAQ v3, training off

Resources needed: ~6 hrs setup

Max acceptable cost: Under \$100/month

Owner: Care lead

06 DATA AND INPUTS

Confirm the information used is suitable, permitted, and protected.

Data and knowledge sources: FAQ, policies, order status

Owner and permission: Ops manager; vendor terms OK

Sensitivity: Names, emails, order IDs; no payment data

Quality / representation concerns: FAQ 8 months old; refresh

Prohibited inputs: Payment details, passwords

Known limitations: No live inventory; may cite old policy

INPUT CHECKLIST

- ☒ Authorized
☒ Relevant
☒ Accurate enough
☒ Representative enough
☒ Noticed appropriately
☒ Traceable to source
☒ Free from prohibited or unnecessary data

PHASE 3: EVALUATE

07 MEASURES AND EVIDENCE

Measure value, performance, behavior, and feasibility separately.

BUSINESS / USER OUTCOME

Metric: Median first reply
Baseline: 26 hrs
Target: Under 8 hrs
Minimum threshold: Under 14 hrs
Evidence source: Help desk report

AI PERFORMANCE AND EVALUATION

Quality measure: Draft score 1-5
Fairness / error measure: Errors per 100
Reliability measure: % usable drafts
Baseline / benchmark: Agent-written replies
Minimum threshold: Avg score 4+

ADOPTION AND APPROPRIATE RELIANCE

Usage measure: % emails drafted
Correctness / overrides: % fully rewritten
User confidence / trust: Weekly 1-5 pulse
Overreliance signal: Sent with no edits
Minimum threshold: 70% use, <10% rewrites

EFFICIENCY AND FEASIBILITY

Time impact: Minutes per reply
Cost impact: Subscription + setup
Effort required: Weekly FAQ upkeep
Operational complexity: Low; no integration
Scalability concern: Holiday volume 3x

Evidence collection method (how results will be recorded, reviewed, and compared): Shared sheet logs each draft, its score and edit time. Reviewed weekly by the care lead.

08 RISKS AND GUARDRAILS

Define prevention, detection, response, and stop conditions.

RISK	PREVENTION	DETECTION	RESPONSE	STOP TRIGGER
Bad policy	Fix FAQ	Spot check	Correct it	Bad refund
Data leak	Training off	Terms	Tell owner	Off-policy
Wrong tone	Route out	Agent reads	Rewrite	Complaint
No checking	Must review	Edit log	Coach	20% unread

Non-negotiable guardrails (must be met at all times):

- Every reply approved
- No payment data in AI
- Refunds stay human

Immediate stop conditions (pause or stop if):

- Wrong policy sent twice
- Any data exposure
- Satisfaction drops 10%+

PHASE 4: DECIDE

09 LEARNING AND DECISION

Make an explicit decision based on the full body of evidence.

EVIDENCE SUMMARY

When: Week 3 review
Performance: Avg draft 4.1/5
Adoption: 78% use, 12% rewrites
Workflow: Reply 26 to 9 hrs
Data: Returns FAQ gaps
Risk: 2 caught, 0 sent
Feasibility: ~6 hrs/week saved
Remaining uncertainty: Holiday volume

DECISION (select one)

- ☐ Scale - evidence supports broader use
☐ Iterate - revise and test again
☒ Narrow - proceed with a smaller use case
☐ Pause - wait for missing evidence or readiness
☐ Stop - value insufficient, risk unacceptable, or AI not the right solution

DECISION RATIONALE

Beat the 14-hr minimum but missed the 8-hr target. Shipping and order drafts were strong. Both errors came from returns questions. Keep what worked; fix the rest.

NEXT ACTION

Action: AI drafts for shipping and order emails only. Fix returns FAQ.

Owner: Care lead

Due date: Oct 6, 2026

DECISION RULE

Decisions require more than data alone. Consider values, evidence, adoption, workflows, data suitability, risk, feasibility, and human oversight.

Be responsible

Act with care and protect people and data.

Be experimental

Test small, learn fast, and iterate.

Be evidence-based

Use the best available evidence, do no harm.



HOW TO USE THE AI EXPERIMENT CANVAS

The canvas helps you decide whether an AI idea deserves more time, money and trust before you commit to it. It works for a solo owner testing one idea and for an enterprise team building the case to scale. Plan on 30 to 60 minutes for a first pass.

BEFORE YOU START

One person owns the canvas. Put their name in the header. Review it with the people affected before the test begins, and again when it's time to decide. If you can't answer a field, don't skip it. Write "unknown" and add it to your remaining uncertainty in section 09. A blank is a finding.

WORK THROUGH THE FOUR PHASES

Frame (01–03). Define the problem before the technology. Check whether a simpler, non-AI fix would do the job. Map who uses the output, who reviews it, and who feels the result.

Design (04–06). Write a hypothesis you could be wrong about. Choose the smallest test that would still convince a skeptic. Confirm the data you need is permitted, protected and fit for purpose.

Evaluate (07–08). Measure four things separately: user outcome, AI performance, adoption and feasibility. A tool can perform well and still go unused. Set your thresholds and stop triggers before the test starts.

Decide (09). Compare results to the thresholds you set. Choose one decision and write down why.

THE FIVE DECISIONS

Scale: the evidence supports broader use.

Iterate: promising, so change one thing and test again.

Narrow: it works for part of the use case, so proceed only there.

Pause: evidence is missing or the organization isn't ready.

Stop: not worth it, too risky, or AI isn't the answer.

Stop is a good outcome. You learned something before it got expensive.

THREE RULES WORTH KEEPING

Set thresholds before you see results. Keep a person accountable for every decision the AI informs. Write down what surprised you, because that's usually the most useful thing you'll learn.

A WORKED EXAMPLE (FICTIONAL)

Harbor & Pine, an 8-person online home goods shop, tested AI-drafted replies to routine customer emails for three weeks. A person approved every reply before it was sent. Reply time fell from 26 hours to 9, which beat the minimum but missed the 8-hour target. Shipping and order drafts were strong. Both errors came from returns questions. **Decision: Narrow.** Keep AI drafts for shipping and order emails, fix the returns FAQ, then retest.

See the full worked example at cleverbirdtech.org/canvas

NEED MORE ROOM?

The Expanded Workbook uses the same framework with space for longer answers. Download it at cleverbirdtech.org/canvas

Want help running your first experiment? Visit cleverbirdtech.com