Frontier Fit

Participant workbook

Test where AI helps, where it harms, and where human judgment belongs.

Independent candidate prototype prepared by Amir Daniel | August 22, 2026

Independent candidate prototype created using publicly available HBS research. Not an official Harvard Business School product. Do not enter confidential, personal, regulated, or proprietary information into a public AI tool.

Your outcome

By the end of this working session, you will leave with:

- one real workflow decomposed into testable tasks;
- one evidence-based hypothesis about AI fit;
- one representative human-versus-human-AI comparison;
- one operating mode with a named human checkpoint;
- one seven-day experiment with a success metric and stop rule.

The commitment

In the next seven days, I will test _____

before allowing AI to _____.

The workflow I brought

Workflow: _____

Desired business outcome: _____

What makes it important now? _____

Research in one page

The selected study, *Navigating the Jagged Technological Frontier*, examined how 758 Boston Consulting Group consultants performed on realistic knowledge-work tasks. After a baseline assessment, participants were randomly assigned to no AI, GPT-4, or GPT-4 plus a brief prompt-engineering overview.

On an 18-subtask innovation and go-to-market assignment designed to sit within the tested model's frontier, AI users completed 12.2% more tasks and worked 25.1% faster; the HBS AI Institute's 2026 summary reports work rated about 32% higher in quality.

On a brand-strategy task designed to sit outside the tested frontier, the no-AI group was correct 84.5% of the time. Correctness was 70.6% with GPT-4 and 60% with GPT-4 plus the overview. The wrong recommendation could still look coherent.

The transferable idea: similar-looking tasks can sit on opposite sides of AI capability, even within one workflow. The boundary is difficult to see and moves over time.

What the research does not prove

- It does not map every current model or every job.
- It does not make the frontier permanent.
- It does not prove that lower performers always gain more.
- It does not show that prompting alone solves task-fit or verification problems.
- It does not turn one successful demo into evidence for scale.

Sources: peer-reviewed paper and HBS AI Institute interpretation.

Prototype translation requires HBS/faculty validation.

The Frontier Fit Loop

1. Decompose

Break the workflow into observable tasks. Name inputs, outputs, decisions, owners, handoffs, and consequences.

Question: What action begins and ends?

2. Predict

Record a hypothesis before the AI output changes your mind: promising, uncertain, or restricted.

Question: What evidence makes us believe AI will help?

3. Test

Compare representative human-only/current-process and human-AI cases using the same rubric.

Question: Is the output correct and useful across typical and edge cases?

4. Assign

Choose AI-led, human-AI, or human-led authority. Name the accountable human, evidence source, escalation, and stop rule.

Question: Who decides, using what evidence?

5. Monitor

Track drift, exceptions, overrides, and business results. Re-test after meaningful change.

Question: What moved the frontier?

> The Frontier Fit Loop is an original candidate-created instructional translation, not a framework named or endorsed by the study authors. Requires HBS/faculty validation.

Before: quick capability check

Scenario 1

A team says an AI drafting tool reduced memo time by 40% in three demos. What is the best next step?

- [ ] Deploy to everyone because the time saving is large.
- [ ] Compare representative current-process and human-AI cases using a quality rubric and failure criteria.
- [ ] Ask users whether they liked it.
- [ ] Improve the prompt until the output is longer.

Why? _____

Scenario 2

Which signal most strongly suggests a task needs human-led authority?

- [ ] The task is repetitive.
- [ ] The output is easy to reverse.
- [ ] Errors are hard to detect and create irreversible external commitments.
- [ ] The model answers quickly.

Why? _____

Confidence - circle one

I can decompose a workflow into testable tasks. 1 2 3 4 5

I can design a fair comparison. 1 2 3 4 5

I can decide where final authority belongs. 1 2 3 4 5

Step 1 - Map the current workflow

Desired outcome: _____

Current baseline - cycle time, quality, cost, rework, or risk: _____

Who receives the final output? _____

What is the consequence of a plausible but wrong output? _____

#	Task - one observable action	Inputs	Decision or output	Owner	Failure consequence
1					
2					
3					
4					
5					
6					
7					

Checkpoint: if you wrote a department, job, or broad capability, make it smaller. “Marketing” becomes “classify inbound leads against agreed criteria.”

Step 2 - Predict frontier fit

Selected task: _____

My hypothesis: [ ] Promising [ ] Uncertain [ ] Restricted

Task cues

Cue	Low	Medium	High	Evidence or note
Output can be checked against an independent source	[ ]	[ ]	[ ]	
Task repeats with similar structure	[ ]	[ ]	[ ]	
Success can be scored consistently	[ ]	[ ]	[ ]	
Critical context is explicit and available	[ ]	[ ]	[ ]	
Novelty or ambiguity is required	[ ]	[ ]	[ ]	
Error is hard to detect	[ ]	[ ]	[ ]	
Decision is irreversible or externally binding	[ ]	[ ]	[ ]	

What would prove my hypothesis wrong? _____

Whose context or evidence might be missing? _____

Do not treat this page as certification. It is a testable prediction.

Step 3 - Build a representative test

Test cases

Typical case 1: _____

Typical case 2: _____

Typical case 3: _____

Edge case: _____

Ambiguous or adversarial case: _____

Conditions

Current-process or human-only baseline: _____

Human-AI condition: _____

Same input packet? [ ] Same time window? [ ] Same scoring rubric? [ ]

Blinded scoring where practical? [ ]

Acceptance criteria

Correctness or quality: _____

Time or throughput: _____

Failure severity: _____

Rework or override: _____

Stop rule

Stop or narrow the pilot immediately if _____

_____

Remember: a polished output is not a passing score.

Step 4 - Assign authority

Choose one operating mode

AI-led with sampled human review

Use only when performance is repeatedly strong, errors are easy to detect and reverse, and consequences are low.

Human-AI collaboration

Use when evidence is mixed, context matters, or consequences are moderate. Require independent verification.

Human-led with optional AI support

Use when evidence is weak, errors are hard to detect, or the decision is novel, high-stakes, regulated, or irreversible.

My mode: _____

AI may: _____

AI may not: _____

Accountable human: _____

Independent evidence source: _____

Escalate when: _____

Who bears the cost if the output is wrong? _____

Step 5 - Seven-day experiment

Experiment name: _____

Owner: _____ Review date: _____

First representative case: _____

Model/tool and configuration: _____

Data allowed: _____

Baseline measure: _____

Success metric: _____

Critical failure/stop rule: _____

Evidence reviewer: _____

Decision rule

- GO if _____
- REVISE if _____
- STOP if _____

Re-test triggers

- [ ] Model or tool changes
- [ ] Prompt/system instruction changes materially
- [ ] Data source or retrieval changes
- [ ] Workflow owner or approval rule changes
- [ ] New customer, legal, security, or accessibility constraint appears
- [ ] Error/override pattern changes

Calendar action completed? [ ] Yes [ ] Not yet

Peer challenge

Partner: _____

Ask these four questions:

1. Where could a polished error survive the proposed review?

2. Is the test representative of the real workflow, including edge cases?

3. Does the metric measure business value and correctness, or only activity and speed?

4. Is the action reversible, and is the human checkpoint placed before the consequence?

Feedback I received

Strongest part of my plan: _____

Assumption to challenge: _____

Missing case or evidence: _____

Change I will make: _____

Completed example - customer escalation summary

Outcome and baseline

Goal: reduce manager time spent preparing escalation summaries without losing critical facts. Baseline: median 28 minutes per case; two-manager review; common rework caused by conflicting records.

Workflow

Gather history -> summarize issue -> classify severity -> recommend remedy -> draft response -> approve/send.

Selected task

Summarize issue. Prediction: promising for routine cases; uncertain when notes conflict, logs are incomplete, or security concerns appear.

Representative test

Ten de-identified historical cases: six typical, two conflicting, one security-sensitive, one missing logs. Compare manager-only and AI-assisted summaries. Two blind reviewers score factual completeness, unsupported claims, clarity, and explicit uncertainty. Track time.

Mode and checkpoint

Human-AI collaboration. AI drafts a structured summary. The case owner verifies every critical fact against source records and marks conflicts. AI cannot classify severity, recommend a remedy, or make an external commitment during this pilot.

Decision rule

GO to a wider pilot if time improves at least 20%, factual completeness is no worse than baseline, critical unsupported claims are zero, and every security-sensitive case is surfaced. REVISE if routine performance passes but edge cases fail predictably and can be routed. STOP if a critical invented fact or hidden security signal survives review.

Why this is a stronger test than a demo

It includes a baseline, representative and difficult cases, independent reviewers, a quality rubric, a bounded AI role, and a decision rule.

Reflection and action plan

The belief I changed: _____

The task I should not automate yet: _____

The evidence I need before increasing AI authority: _____

The human judgment that must remain explicit: _____

The person I need to involve: _____

One accessibility, equity, or stakeholder consequence to test: _____

After-session capability check

I can decompose a workflow into testable tasks. 1 2 3 4 5

I can design a fair comparison. 1 2 3 4 5

I can decide where final authority belongs. 1 2 3 4 5

What in my completed plan demonstrates that capability? _____

Executive reference card

Frontier Fit Loop

Decompose - What action begins and ends?

Predict - What evidence suggests AI will help, harm, or remain uncertain?

Test - Same cases, same rubric, credible baseline, typical and edge cases.

Assign - Who decides, using what evidence, with what escalation and stop rule?

Monitor - What changed in the model, data, workflow, exception pattern, or consequence?

Three operating modes

- AI-led: strong repeated evidence; easy-to-detect, reversible, low-consequence errors.
- Human-AI: mixed evidence or meaningful context; independent verification required.
- Human-led: weak evidence, difficult verification, novel/high-stakes/regulated/irreversible consequence.

Do not scale from

- one successful demo;
- speed alone;
- polished language;
- high confidence;
- usage volume;
- a task label with no representative evidence.

Seven-day test

Owner ____ Case ____ Baseline ____ Quality metric ____ Stop rule ____ Review date ____

Source: Navigating the Jagged Technological Frontier. Candidate-created translation; requires HBS/faculty validation.