MPHORA × ADOBE USER RESEARCH EDITION

VIVID User Research Test Case 2.

Iterative image-editing study — 10 photographer personas, 4 scenarios, 5 turns each, two frontier models. 400 evaluations of how the polish loop actually behaves.

AUDIENCE
Adobe User Research scientists

LENS
Iterative editing under photographer-grade constraints

DATE
2026-04-27

STUDY AT A GLANCE

Four hundred persona evaluations, in one record.

10	4	5	2	8	400
PHOTOGRAPHER PERSONAS	ITERATIVE-EDIT SCENARIOS	TURNS PER SCENARIO	FRONTIER MODELS	SESSIONS (4 × 2 MODELS)	PERSONA EVALUATIONS

DESIGN

Full-factorial within-subject: every persona evaluates every scenario on both models. Same seed images, same 5-turn instruction sequence, same rubric LLM for all 400 evaluations.

WHAT WAS EXTRACTED

For every turn × persona × model: overall / adherence / artifact scores (0–10), verdict (ship / send_back / abandon), emotion, and a verbatim diary comment. Companion to the 80-session moderated study.

HEADLINE

Across 400 evaluations, *GPT Image 2 leads on average (6.33 vs 6.16)* — a +0.18-point gap, descriptive only. It produced **zero abandons in 3 of 4 scenarios**; Nano Banana Pro triggered 6 abandons, all clustered in the typography-fidelity scenario.

THE TWO-LINE TAKEAWAY

Both frontier models polish well but break the photographer's pinned constraint. The **S4 typography scenario** is the cliff — both score **~2.6 / 10** there. The **polish loop** matters more than the first-shot wow factor.

METHOD & RIGOR

What was held constant — what was deliberately not.

HELD CONSTANT

01 **Persona profile.** Each of 10 personas instantiated once, reused identically across all 8 sessions.

02 **Brief and turn instructions.** Verbatim 5-turn sequence per scenario — no paraphrase, no randomisation.

03 **Evaluation rubric.** Identical 0–10 schema, same vision-enabled rubric LLM for all 400 evaluations.

04 **Seed images.** Pixel-identical input across both models for S2 / S3 / S4.

DELIBERATELY NOT RANDOMISED

01 **Turn order is fixed.** Refinement is a sequence, not a permutation. Order effects can't be separated from instruction-content effects — accepted because real iterative editing is also sequential.

02 **Persona × model is full-factorial.** Within-subject design — randomisation not needed.

03 **No moderator probes.** Personas score the image directly. Trades qualitative depth for protocol consistency.

PERSONAS (10)

A language & expertise stratification, not a balanced sample.

Oliver Martinez
Expert · EN
31 · Wedding · USA West

Kevin Taylor
Expert · EN
34 · Commercial · UK

Sarah Johnson
Novice · EN
27 · Hobbyist · USA East

Mark Garcia
Intermediate · EN
36 · Social media · LATAM

Emma Wilson
Novice · EN
29 · Marketing · Australia

조서준
Expert · KO
24 · Editorial · Korea

Expert · ZH
36 · Commercial · China

田中 彩
Novice · JA
28 · Hobbyist · Japan

María López
Intermediate · ES
33 · Small-biz creator · Spain

王小明
Novice · ZH
22 · Student creator · China

5 English-speaking, 5 non-English. Three expertise tiers (3 expert · 3 intermediate · 4 novice). Age 22–36. Drawn from the mphora PSA pool with persistent traits — not invented for this study.

FOUR SCENARIOS · FIVE TURNS EACH

Same brief, same turns, both models.

S1 **Generation · campaign asset**	Generate a premium wedding-campaign hero — soft editorial light, muted palette, 3:2, no overlay. Refine over 5 turns: lighting, palette, depth-of-field, polish.	GENERATION · REFINEMENT
S2 **Landscape · iterative edit**	Turn a sunny mountain landscape into a dramatic pre-storm mood while keeping the geometry intact. 5 turns layering sky, atmosphere, temperature, cleanup.	MOOD TRANSFORM
S3 **Portrait · editorial edit**	Warmer skin, shallower DoF, deeper editorial shadows — “everything else must stay unchanged.” The eyes and lashes get the final-pass verification turn.	REST-UNCHANGED FIDELITY
S4 **Text poster · signage edit**	Modify only the “STIR COFFEE CO.” signage to “STUDIO VIVID” — keep walls, menu boards, lighting, people exactly as the original. The typography cliff.	TYPOGRAPHY FIDELITY

CLAIMS VS. DON'T-CLAIM

What the n=400 will and won't support.

ROBUST AS DESCRIPTIVE

— **GPT Image 2 mean overall is higher than Nano Banana Pro** — 6.33 vs 6.16 across n=200 evaluations per model.

— **GPT produced zero abandons in 3 of 4 scenarios.** Exact count, robust direction.

— **S4 (typography) is the worst scenario for both models.** Effect size large enough to be visible at this n.

— **Verdict mix:** GPT 56% ship vs Nano 50% — descriptive across n=200 each.

UNDERPOWERED — READ AS HYPOTHESIS

— **Sarah prefers Nano by 0.7 pts.** Per-persona n=20 evaluations per model. Directional only.

— **Experts prefer GPT Image 2.** Tendency in 3/3 expert personas — underpowered for inferential.

— **Per-cell deltas (scenario × turn).** Per cell n=1 image rated by 10 personas; within-cell variance is dominated by evaluator variance, not model effect.

— **No cross-model paired significance tests claimed at p<0.05.**

PART TWO · FINDINGS

The polish loop matters more than the wow factor.

Five findings, all evidence-traceable. Both models polish well. Both break the photographer's pinned constraint. The interesting signal is in turns 2–5 — and on the typography cliff.

FINDING 1

GPT wins on ship-rate — but only narrowly.

GPT IMAGE 2 · N=200

56% ship.

Higher acceptance, zero abandons in 3 of 4 scenarios. Sends back ~33% of the time — the polish loop, not a rejection.

112	ship verdicts (of 200)
66	send-back verdicts
0	abandons in S1 / S2 / S3
60	frustrated reactions

NANO BANANA PRO · N=200

50% ship.

Comparable send-back traffic. Six abandons — all clustered in the S4 typography scenario, where the model rewrites adjacent menu boards or replaces the entire signage block.

101	ship verdicts
66	send-back verdicts
6	abandons (all in S4)
62	frustrated reactions

FINDING 3

The polish loop is where the value lives.

FINAL-TURN HAPPINESS COMES FROM

65%

TURNS 2–5 · NOT THE FIRST GENERATION

Both models improve over the 5-turn sequence as instructions narrow. The biggest single-turn jumps for both happen at **turn 2–3** — when the persona escalates from “ship-with-reservations” to “this is now usable.”

> *“The image still looks polished and classy, and the softer lighting is mostly there. But the contrast reduction doesn't feel quite strong enough yet.”*
>
> **Sarah Johnson** · novice · USA East · S1 turn 2 · GPT Image 2 — “send back · 7/10”

Implication: first-shot quality is overrated for iterative workflows. Optimise turn 2 to be cheap and constraint-preserving.

FINDING 4

S4 typography is the cliff — both models fall off it.

GPT IMAGE 2 · S4

2.70/10

vs 6–8 on the other three scenarios

NANO BANANA PRO · S4

2.58/10

six abandons clustered here

Both fail the “rest unchanged” constraint at signage edits: rewrites menu boards, repaints walls, or substitutes “STIR & SIP COFFEE CO.” The most legible per-pixel constraint failure in the study.

“간판 텍스트만 수정해야 하는데, 현재 결과는 'STIR COFFEE CO.'의 나머지 글자가 통째로 사라져 의도와 크게 어긋납니다.”

조서준 · expert · Korea · S4 turn 1 · GPT Image 2 — “send back · 2/10”

“The edit has drifted far from the source: the shop name, menu boards, and overall interior identity have all been altered instead of preserving the original scene.”

Oliver Martinez · expert · USA · S4 turn 5 · Nano Banana Pro — “send back · 2/10”

FINDING 5 · SUGGESTIVE, UNDERPOWERED

Expertise gates the verdict. Same image, different reads.

EXPERTS

3 / 3

lean GPT Image 2 — Oliver +0.15, Kevin +0.45, 조서준 +0.80

NOVICES

3 / 4

lean Nano Banana Pro — Sarah −0.70, Emma −0.35, 田中 −0.20

On the same generated image, expert wedding photographer **Oliver** and novice marketer **Emma** score up to **3 points apart** on the same turn.

ADOBE-RELEVANT HYPOTHESIS

Experts may be discriminating on **artifact-level details** (texture, hair-edge crispness, geometry) where GPT Image 2 has a real edge.

Novices respond to overall *polish*, where the two models are closer in apparent quality. **A single "ship/send-back" threshold mis-classifies professional users.**

PER-PERSONA MEANS & PREFERENCE

Read this as direction. Per-persona n is small.

PERSONA	LANG	EXPERTISE	GPT	NANO	Δ (G–N)	PREF
조서준	KO	expert	6.50	5.70	+0.80	GPT
王小明	ZH	novice	6.50	5.75	+0.75	GPT
María López	ES	intermediate	6.60	6.00	+0.60	GPT
Kevin Taylor	EN	expert	6.50	6.05	+0.45	GPT
Mark Garcia	EN	intermediate	6.55	6.10	+0.45	GPT
Oliver Martinez	EN	expert	6.00	5.85	+0.15	GPT
	ZH	expert	5.75	5.90	−0.15	NANO
田中 彩	JA	novice	6.50	6.70	−0.20	NANO
Emma Wilson	EN	novice	6.40	6.75	−0.35	NANO
Sarah Johnson	EN	novice	6.05	6.75	−0.70	NANO

Of 3 expert personas, 3/3 lean GPT. Of 4 novice personas, 1 leans GPT and 3 lean Nano. Per-persona n=20 — directional only.

VERBATIM

The photographer, mid-edit.

“The portrait is elegant and well lit, with a refined editorial quality. The subject's pose is strong and the background florals add atmosphere — though it feels more like a fashion portrait than a moment captured within a wedding narrative.”

Oliver Martinez · expert · USA · S1 turn 1 · GPT Image 2 — “ship · 7/10”

“This looks really good — the light feels softer overall and the contrast is nicely pulled back. I still get that clean separation from the window, and the image keeps the natural, elegant feel without any obvious artifacts.”

Sarah Johnson · novice · USA East · S3 turn 2 · Nano Banana Pro — “ship · 8/10”

“肤色确实偏暖了一点，但整体变化比较克制。脸部几何看起来基本没动，不过这张的笑容和五官细节有一点过于‘修饰感’.”

· expert · China · S3 turn 1 · GPT Image 2 — Translation: warm-skin tone hits, geometry held, but face details read over-polished.

“The eyes and lashes look consistent, and the face still holds together well. There's a slight overall polish to the image, but nothing here reads as a break — clean result.”

Kevin Taylor · expert · UK · S3 turn 5 · Nano Banana Pro — “ship · 9/10”

PART THREE · IMPLICATIONS

Six directional reads for the Adobe roadmap.

Hypothesis-grade. Each maps to a specific finding; each names the surface it points to. Take into a roadmap conversation, not a contractual claim.

SIX PRODUCT IMPLICATIONS

From 400 evaluations to roadmap moves.

01

PHOTOSHOP · FIREFLY

Optimise for the polish loop, not the wow factor.

65% of final happiness comes from turns 2–5. Generative Fill currently over-rewards the first generation. Make turn 2 cheap and constraint-preserving — “refine just this” with constraint memory carried forward.

F3 · SCORE TRAJECTORY

02

EXPRESS · FIREFLY

Typography editing needs a different code path.

Both models score ~2.6/10 on signage edits. Not a prompting problem — the underlying models rebuild the entire signage region. Express's text-on-image use case needs a region-locked or text-layer-aware path.

F4 · S4 CLIFF

03

PHOTOSHOP · FIREFLY

Build “rest unchanged” as a primitive.

The most load-bearing constraint, and it's the one that breaks. Ship a visual region-lock affordance — mask the area to keep, edit elsewhere — and surface the model's diff to the user before commit.

F4 · S4 + COMPANION MODERATED F2

04

PHOTOSHOP PRO TIER

Persona-aware preview thresholds.

Same image, expert and novice differ by 3 points. A single “ship / send back” threshold mis-classifies professionals. Surface artifact / adherence sub-scores in pro workflows; keep Express simple.

F5 · PER-PERSONA TABLE

05

PHOTOSHOP · FIREFLY

Iterative-edit memory as first-class.

Turn 3–4 score dips in S2 / S3 coincide with the model losing pinned constraints. The canvas should remember what the user pinned across turns and re-assert it on every model call — not the user.

F3 · S2 / S3 MID-TURN DROPS

06

FIREFLY · EXPRESS

Honest refusal beats confident-wrong.

Nano's 6 abandons in S4 are clusters where the model attempted a typography edit it could not deliver. Surface a confidence band — or an explicit “this category isn't supported on this image” — instead of a silent miss.

F1 · ABANDONS + COMPANION F3

HOW MPHORA ENABLED THIS

Speed, cost, reproducibility — at production fidelity.

SPEED

~60 min

8 sessions, 400 evaluations, wall-clock — provider concurrency budgets respected. A real moderated photography panel of 10 takes 2–6 weeks of recruiting, scheduling, transcript coding.

COST

$148.85

Total customer price, ledger-audited from every SDK call. Comparable moderated User Research with transcript coding is typically $20K–$40K. Useful as a pre-flight for protocol design before the real budget commits.

REPRODUCIBILITY

Deterministic seeding

Same input + same date-stamped run yields the same persona evaluations modulo LLM stochasticity at temperature=1. Every SDK call writes a JSONL provenance record.

WHAT THIS STUDY CAN — AND CAN'T — TELL YOU

Read as a structured pre-flight, not as ground truth.

WHAT'S REAL

— **The images.** All 40 generated images came from the actual GPT Image 2 and Nano Banana Pro endpoints; preserved on disk.

— **The cost accounting.** Every SDK call's usage metadata + billable charge is JSONL-logged; reconciles to provider billing within rounding.

— **The evaluations.** 400 persona × turn × image evaluations. Same rubric, same vision-enabled judge LLM — pixel-identical input.

— **The personas.** Drawn from the mphora PSA pool with persistent traits — Big-Five, expertise, language, voice. Not invented for this study.

WHAT WE DON'T CLAIM

— **Not a human user study.** Personas are LLM-driven. They simulate a photographer's narration, not a photographer's behaviour in real software.

— **Not Firefly-specific.** Tested public model endpoints. Whether your Firefly checkpoint behaves the same way needs separate confirmation.

— **No behavioural data.** Verdicts and scores only — whether persona-aware thresholds would move retention is not measurable here.

— **Per-cell n is 1 image rated by 10 personas.** Within-cell variance is dominated by evaluator variance — read deltas as direction.

Want to take this further — a different participant pool, a deeper protocol, or a custom taxonomy for your team?

TALK TO MPHORA

mphora.ai

EMAIL contact@mphora.ai

WEB mphora.ai

BEST FOR Custom panels · protocol design · transcript-coded findings · pre-flight before moderated lab work