

From Feed to Agent

Method, states, and lineage

The compact technical companion to the public report: what was observed, how missing states were handled, how the survey layers differ, and how every public result traces back to a frozen protocol.

192

FORMAL OBSERVATIONS

4 surfaces · 2 rounds

149

SURVEY SUBMISSIONS

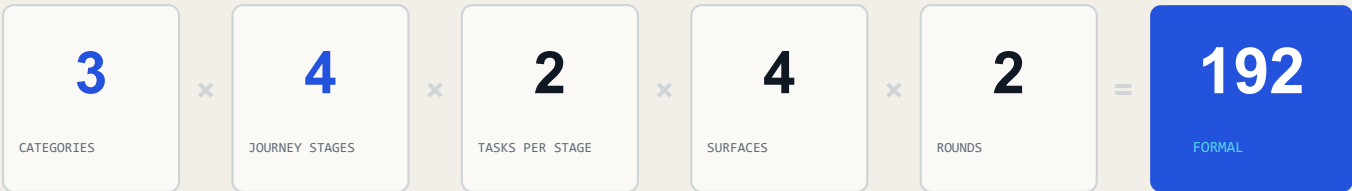
73 primary · 122 expanded

THIS APPENDIX

- Frozen task lattice
- Capture and acceptance
- Response-state model
- Truth refresh
- Survey eligibility
- Audit and uncertainty
- Priority heuristic
- Publication boundary

A frozen task lattice separated protocol from output.

The comparison was defined before formal capture. Every accepted position maps to one category, journey stage, task, surface, and round.



CATEGORIES

Headphones or earbuds

High-consideration consumer-electronics scenarios; robot vacuums and smartwatches or fitness wearables used the same stage and task structure.

JOURNEY

Four stages

Inspire, compare, validate, and shortlist. Each stage contained two frozen tasks with explicit scenario criteria.

SURFACES

Two search + two agent boundaries

TikTok and YouTube were visible search-page observations. Google AI Mode and ChatGPT were complete initial-answer observations.

ROUNDS

August 17 and August 20, 2026

The query set and order were frozen. Round 2 repeated the full lattice to reveal output movement and operational state changes.

PILOT DISCLOSURE

24 pilot observations were used to test the protocol and were excluded before formal analysis. They remain in lineage records, never in formal estimates.

Each surface was judged inside its captured boundary.

The study does not pretend that a visible search page and a complete generated answer are the same artifact.

01

TikTok + YouTube

Titles, captions, metadata, URLs, and visible modules.

Visible search page

Videos were not watched.

02

Google AI Mode + ChatGPT

Answer text, source labels or visible links, and generation state.

Complete initial answer

No follow-up dialogue.

03

Named validation tasks

Six specific product checks with state-preserving coding.

Refreshed truth fields

Not broad recommendation accuracy.

04

Consumer survey

149 raw submissions; primary and expanded eligibility layers.

Anonymous retrospective choices

Not a probability sample.

Formal acceptance required inspectable evidence.

A screenshot could support a row, but it could not replace the mandatory DOM or text evidence and structured record.

01

Freeze

Use the exact query, surface, account context, and run position from the formal order.

02

Capture

Save structured JSON plus mandatory visible DOM or text. Screenshot is auxiliary.

03

Hash

Calculate SHA-256 for the captured evidence and preserve source paths in the record.

04

Code

Apply the frozen criterion rubric and preserve missing or inapplicable states.

05

QA

Run row-level schema, evidence, hash, and round-level completeness checks.

ACCEPTANCE GATE

A row entered the formal result only when its position, evidence, structured fields, hash, and QA state were complete. Access or generation failures were valid response states when documented.

Missing evidence never became a zero.

The analysis stored what was seen, what was not applicable, and why a generated-answer position did or did not contain an answer.

INITIAL_RESPONSE_GENERATED

A complete initial answer appeared and was coded inside the agent-answer boundary.

GENERATION_FAILED_WITH_FALLBACK

The interface showed an explicit generation error and fallback results. Those results were not recoded as an AI answer.

SEARCH_RESULTS_CAPTURED

A visible search-results page was captured for TikTok or YouTube.

NOT_APPLICABLE

A metric did not belong to that evidence family—for example, answer generation on a search page.

40 / 48

GOOGLE AI MODE GENERATED

Round 1 · 24/24

Round 2 · 16/24

Eight Round 2 positions retained generation-failure states with fallback results visible.

Claim checks used four mutually exclusive states.

Official truth fields were refreshed before each round; only six named validation tasks entered this truth set.

60**Verified**

The captured claim matched the refreshed truth field.

36**Unclear / unsupported**

The captured evidence did not support a confident verification.

37**Not covered**

The answer did not address the named check.

3**Incorrect**

The claim contradicted the refreshed truth field.

OBSERVED INCORRECT CLAIMS

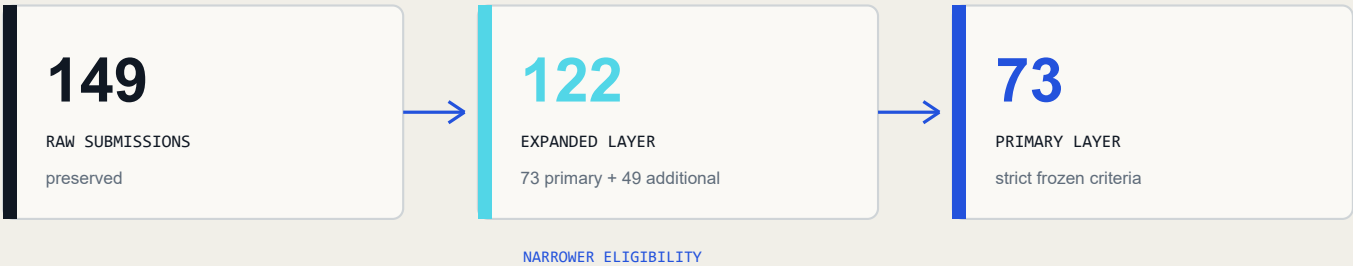
All three were narrow robot-vacuum overclaims in captured Google AI Mode answers: pet-waste avoidance certainty and local-identification versus optional-upload distinctions.

INTERPRETATION BOUNDARY

The truth set supports claim-level findings for the six named tasks. It does not support a general accuracy rank for the four surfaces.

The 122-person sensitivity layer includes the 73-person primary group.

All 149 anonymous submissions are preserved. Eligibility was frozen before analysis, and isolated contradictory items use item-level missingness.



PRIMARY

Strict eligibility

Current U.S. adult, consented, category-aligned, and researched a target consumer-electronics category in the prior 12 months.

SENSITIVITY

Expanded eligibility

Retains 49 additional U.S. adult respondents for a broader sensitivity check while preserving the same core research recency and focus requirements.

HANDOFF PREFERENCE	
Approve each handoff	41.1% / 40.2%
Automatic handoff	35.6% / 37.7%
No cross-product handoff	23.3% / 22.1%
Primary / expanded. Ranking is identical across layers.	

TOP IMPROVEMENT REQUEST

Direct links or citations to original sources: 34.7% primary and 32.2% expanded. Other source and stage comparisons remain exploratory.

Quality checks targeted the frozen unit of analysis.

The coding audit checks protocol adherence. The bootstrap describes uncertainty across frozen query clusters—not across consumers.

48 / 192

STRATIFIED AUDIT
one coder

3,264

METRIC-STATE ROWS
explicit missing states

10,000

BOOTSTRAP RESAMPLES
24 query clusters

01

AUDIT FRAME

The 48-observation audit was stratified across surfaces, categories, stages, and rounds. It checked the stored evidence against the coding protocol and recorded row-level findings.

02

CODER LIMIT

One coder performed the study and audit, so inter-rater reliability is unavailable. The audit supports protocol-consistency claims, not independent coder agreement.

03

INTERVAL UNIT

Coverage intervals resample the 24 frozen query clusters. They describe task-set variability inside the study, not a platform population or consumer population.

04

ROUND STABILITY

Search overlap uses repeated URLs; agent overlap uses repeated tokens. The measures are descriptive and are never treated as one cross-family quality score.

WHAT THE AUDIT DOES NOT PROVE

It does not create a second-coder reliability estimate, validate unobserved video content, or turn the benchmark into a population inference.

The roadmap heuristic is transparent, not predictive.

Numeric scores are retained here and in the workbook so the public story can use plain-language horizons without hiding the assumptions.

PRIORITY INDEX

impact × evidence strength × feasibility ÷ trust risk

IMPACT

1–5

Expected importance of the product job.

EVIDENCE

1–5

Strength and convergence of the research basis.

FEASIBILITY

1–5

Relative buildability for a first product test.

TRUST RISK

1–5

Higher value means more trust exposure; it is the denominator.

01	Consent-based Intent Passport	50.0	Build first
02	Portable Shortlist Snapshot	48.0	Build next
03	Decision Canvas	40.0	Build first
04	Claim-to-Proof Graph	37.5	Build first
05	Verification Queue	30.0	Build first
06	Origin-preserving Evidence Drawer	24.0	Build next

Use the index to expose judgment and debate assumptions. Do not read it as demand, adoption, revenue, or product-market-fit forecast.

Public evidence is useful only when its boundaries travel with it.

The release separates public narrative, technical evidence, and internal QA while preserving numeric consistency across all three layers.

PUBLIC

Safe row-level outputs

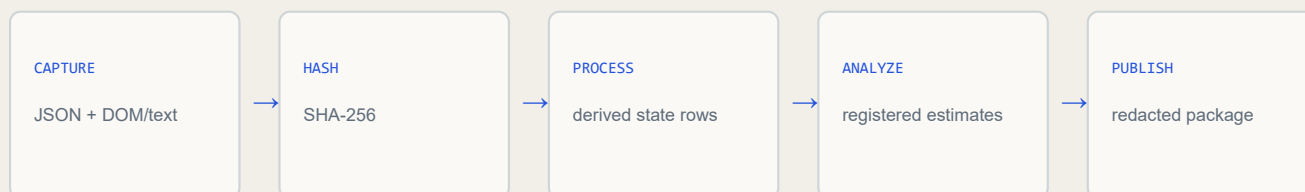
192 observation-feature rows, 3,264 metric-state rows, 136 validation checks, and 48 coding-audit rows.

PRIVATE

Protected source material

Participant-level survey rows, birth fields, timestamps, account identifiers, and raw logged-in evidence remain private.

LINEAGE CHAIN



CORE INTERPRETATION LIMITS

- No causal claim, population estimate, significance test, or universal platform ranking.
- TikTok and YouTube findings apply only to captured visible search-page evidence; videos were not watched.
- Only six named tasks were truth-checked; broad recommendation accuracy remains unmeasured.
- Survey results are anonymous convenience-sample descriptive signals with primary and expanded layers.

PRIMARY CAPABILITY CONTEXT

TikTok Shop newsroom · YouTube Shopping support · Google AI Mode product blog · OpenAI ChatGPT search announcement. Capability sources provide context; observed performance comes from this study.