Consumer Research Bench (CRB): Evaluating AI Agents for Consumer Research

Independent CRB Working Group

Working Draft v0.6

v0.1 scope: Consumer domain only. Brand, Category, and Culture are retained as roadmap domains.

Preprint · June 2026

ABSTRACT

Consumer research combines evidence collection, behavioral interpretation, qualitative synthesis, quantitative classification, and business recommendation. Recent AI benchmarks evaluate web research agents, long-form research agents, browsing agents, coding agents, and general-purpose assistants. However, these benchmarks do not directly evaluate the capabilities specific to consumer research: discovering organic consumer signals, filtering noisy and heterogeneous data, interpreting motivations and occasions, identifying cohorts, applying research methodologies, integrating quantitative and qualitative evidence, and producing decision-ready insight.

We introduce **Consumer Research Bench (CRB)**, a benchmark for evaluating AI agents on consumer research tasks. We define the benchmark construct as **decision-ready consumer-research quality per unit time, at a fixed evidence-grounding bar**. CRB evaluates whether an AI system can transform fragmented consumer-signal data into evidence-backed consumer understanding with the interpretive depth associated with primary research and the speed of digital signal analysis.

CRB v0.1 contains approximately **12 Consumer-domain research briefs**, each paired with expert-authored reference outputs and atomic sub-instances, yielding roughly **30–40 objective evaluation items**. The benchmark uses two scored tracks. The **Objective Track** evaluates atomic research capabilities against expert-worked ground truth. The **Report Track** evaluates open-ended research outputs using **C-RACE**, a reference-based, criteria-adaptive evaluation framework validated against human experts.

CRB also introduces **RetroSignals**, a frozen consumer-signal environment designed to make evaluation reproducible over time. RetroSignals enables a **data × model ablation** that separates the contribution of consumer-signal data from the contribution of the model itself. In addition to scored outputs, CRB reports diagnostic measures of signal quality, including data breadth, data depth, data condition, relevant signal density, contextual fidelity, quant–qual integration, and trace-level failure modes.

CRB is designed to be administered by an independent set of researchers. Consumer research model providers, data providers, and commercial platforms may participate as evaluated systems or disclosed contributors, but benchmark construction, task selection, grading, and reporting are governed independently. This separation is central to the benchmark's credibility.

1. Introduction

Consumer behavior is increasingly dynamic. Trends can emerge, mutate, and decline over compressed time horizons, while consumer language, category boundaries, and competitive contexts shift continuously. This creates pressure on research teams to produce consumer understanding at a faster cadence without sacrificing interpretive depth or evidentiary rigor.

Traditional primary research methods such as surveys, focus groups, interviews, and ethnography remain valuable because they can explain motivations, beliefs, barriers, rituals, and decision drivers. However, these methods are often slow, expensive, and episodic. Social-listening and market-intelligence tools are faster and useful for monitoring mentions, topics, sentiment, engagement, and share of voice, but they are not primarily designed to evaluate methodology-driven consumer research synthesis. General-purpose AI tools can search and summarize information, but they often lack consumer-specific data structures, research methodology, and domain-specific reasoning standards.

This creates an evaluation problem. If an AI system claims to perform consumer research, how should that claim be tested?

Consumer research differs from general web research in both the nature of its evidence and the standards by which outputs are judged.

First, the relevant evidence is typically distributed across heterogeneous, user-generated sources rather than contained in authoritative documents. Consumer signals may appear in social posts, comments, replies, reviews, marketplace content, creator media, transcripts, community discussions, and search behavior. As a result, the evaluation setting must test an agent's ability to retrieve and reason over fragmented, multi-format evidence.

Second, consumer-signal environments contain substantial noise. Research agents must distinguish organic consumer expression from brand-owned content, promotional material, news coverage, investor commentary, duplicate content, spam, automated activity, and platform-specific artifacts. This makes signal conditioning a core capability rather than a preprocessing detail.

Third, consumer research outputs are interpretive rather than merely factual. A high-quality response must move beyond information retrieval to explain motivations, barriers, occasions, rituals, needs, cohorts, cultural meanings, and implications. The benchmark must therefore evaluate synthesis quality, not only answer correctness.

Fourth, consumer research is methodology-dependent. Different briefs require different research structures: usage and attitude studies, trendspotting, brand diagnosis, campaign measurement, whitespace mapping, digital ethnography, and cohort analysis impose distinct standards for evidence collection, segmentation, interpretation, and reporting.

Fifth, consumer research must integrate quantitative and qualitative evidence. Strong research does not only count signals or quote verbatims; it connects theme prevalence, cohort distribution, temporal movement, and source patterns with consumer language, meaning, motivation, and context.

Finally, the practical value of consumer research lies in decision support. Outputs are useful only insofar as they help stakeholders make informed choices about positioning, innovation, communication, measurement, or further research. CRB therefore evaluates whether an agent produces evidence-grounded conclusions that are not only accurate and insightful, but also usable in applied business contexts.

CRB asks the following central question:

Given a real consumer-research brief and access to messy consumer-signal data, can an AI agent produce a decision-ready, evidence-backed output with consumer-research depth at digital-signal speed?

The benchmark evaluates both final outputs and intermediate capabilities: signal retrieval, signal conditioning, evidence grounding, methodology fidelity, synthesis quality, contextual fidelity, decision usefulness, and speed. It also explicitly evaluates where specialized consumer data does not improve performance, because such cases are necessary for credible interpretation of cases where specialized data does improve performance.

2. Related Work

2.1 Deep Research Agent Benchmarks

Recent deep research benchmarks provide two important methodological precedents.

One line of work evaluates long-form research reports. These benchmarks assess whether agents can produce comprehensive, citation-rich, readable, and analytically useful reports. They introduce reference-based scoring, adaptive evaluation criteria, LLM-as-judge methods, and citation verification. These methods are relevant to CRB because consumer research outputs often take the form of reports, memos, dashboards, or presentation-ready synthesis.

Another line of work evaluates atomic web-research tasks in controlled environments. These tasks include finding numbers, identifying original sources, validating claims, gathering evidence, populating reference classes, and compiling datasets. Such benchmarks are valuable because they support objective or semi-objective scoring using binary accuracy, recall, F1, or probability difference.

CRB builds on both traditions. Its Objective Track adapts atomic research tasks to consumer-signal environments. Its Report Track adapts reference-based report evaluation to consumer research synthesis.

However, CRB differs from general deep research benchmarks in three ways:

1. **Domain**: CRB evaluates consumer research rather than generic web research.
2. **Grounding**: CRB uses organic consumer signal rather than only public web documents.
3. **Output standard**: CRB evaluates decision-ready consumer insight rather than general-purpose reporting.

2.2 Software and Professional-Domain Benchmarks

Software-engineering benchmarks demonstrate the value of verifiable tasks, hidden tests, held-out sets, and execution-based scoring. Professional-domain benchmarks demonstrate the importance of expert-authored tasks and construct validity in specialized domains.

CRB adopts three principles from these benchmark traditions:

1. The benchmark construct should be defined narrowly and explicitly.
2. Task construction should involve domain experts.
3. Public examples should be separated from private held-out evaluation tasks to reduce contamination.

2.3 Social Listening and Market Intelligence

Social-listening and market-intelligence tools are relevant baselines for CRB. These systems are designed to monitor consumer conversation, measure sentiment, detect topics, track share of voice, and surface trends. They are useful for continuous measurement and monitoring.

CRB evaluates a related but distinct capability: whether a system can move from monitoring to research. This requires explaining consumer motivations, occasions, needs, tensions, rituals, cohorts, and implications from organic consumer signal. In CRB, social-listening systems are therefore treated as evaluation subjects rather than prior evaluation methods.

2.4 Evaluation Gap

Existing benchmarks do not directly evaluate the combination of capabilities required for AI-native consumer research:

1. Full-conversation signal discovery.
2. Contextual filtering of non-consumer content.
3. Classification by need, occasion, sentiment, cohort, and stage.
4. Interpretation of consumer motivation and culture.
5. Quantitative and qualitative evidence integration.
6. Methodology-aware synthesis.
7. Decision-ready output.
8. Measurement of specialized consumer-data contribution.

CRB is designed to address this gap.

3. Benchmark Construct and Scope

CRB measures:

Decision-ready consumer-research quality per unit time, at a fixed evidence-grounding bar.

Each part of this construct is intentionally specified.

3.1 Decision-Ready

The output must be useful for a real business or research decision. It should inform what a team might launch, reposition, message, test, measure, monitor, or investigate next.

3.2 Consumer-Research Quality

The output must meet standards associated with consumer research. It should identify motivations, barriers, occasions, rituals, tensions, needs, cohorts, consumer language, cultural meanings, and implications. Surface summaries or generic observations should score poorly.

3.3 Per Unit Time

CRB evaluates both quality and speed. The claim being tested is not simply that an agent can produce good research, but that it can produce grounded, decision-ready consumer understanding within a compressed time frame.

3.4 Fixed Evidence-Grounding Bar

Important claims must be traceable to verifiable consumer signal. A fluent or plausible output that cannot be grounded in evidence should not receive a high score.

3.5 v0.1 Scope

CRB's long-term taxonomy spans four research domains:

1. Brand.
2. Consumer.
3. Category.
4. Culture.

CRB v0.1 is limited to the **Consumer** domain. It focuses on usage, attitudes, motivations, occasions, barriers, cohorts, and emerging behaviors. Beverage-related cases such as matcha, coffee, low-sugar drinks, and ready-to-drink beverages are used as the initial category context.

Brand, Category, and Culture are retained as roadmap domains for future versions.

4. Benchmark Overview

A CRB benchmark instance consists of:

1. A research brief.
2. Business context.
3. Required or suggested methodology.
4. Frozen RetroSignals environment.
5. Task-specific output requirement.
6. Expert reference report or worked answer.
7. Ground-truth evidence checklist where applicable.
8. Scoring rubric.
9. Known failure modes the task is designed to provoke.
10. Validity gates that define run-level failure or score caps.

CRB evaluates both process and output. Two agents may reach similar conclusions but differ substantially in evidence quality, signal coverage, traceability, contextual interpretation, and methodological rigor. For this reason, CRB separates objective evidence-grounded capabilities from open-ended synthesis capabilities.

CRB v0.1 produces two scored outputs:

1. **Objective Track Score**: atomic tasks scored against worked ground truth and evidence checklists.
2. **Report Track Score**: open-ended reports scored through C-RACE and validated against human expert judgment.

Diagnostics such as signal quality, conditioning quality, relevant signal density, contextual fidelity, speed, cost, reproducibility, calibration, and trace failure rates are reported separately. CRB v0.1 deliberately avoids a single composite "overall score" until the two tracks are individually validated.

5. Task Taxonomy

CRB has two task families:

1. Objective atomic tasks.
2. Open-ended report tasks.

These correspond to the two scored tracks.

5.1 Objective Atomic Tasks

Objective atomic tasks adapt general web-research task types to consumer-signal environments. They create a hard-to-dispute evaluation layer and support internal regression testing.

Atomic Task	Description	Scoring
Find Signal	Find a specific, sourced consumer signal	Binary
Derive Metric	Compute a value by combining multiple signals	Binary / accepted range
Validate Claim	Estimate whether a consumer claim is supported	Absolute probability difference
Attribute Driver	Identify the factor driving an observed sentiment or behavior shift	Expert-judged match
Define Cohort	Specify a micro-cohort and its defining signals	Recall against reference set
Populate Set	Compile instances fitting a description	Recall / F1
Compile Dataset	Build a structured table from scattered signals	Precision / recall / F1

Ground truth is created by expert authors at task-construction time through a worked answer or evidence checklist. Tasks without constructible ground truth are assigned to the Report Track rather than the Objective Track.

5.1.1 v0.1 Objective Density

CRB v0.1 is intentionally small, but each research brief can generate multiple atomic sub-instances. This allows the benchmark to maintain objective scoring density without requiring a large number of full briefs.

Proposed v0.1 structure:

Unit	Approximate Count
Consumer-domain research briefs	12
Atomic sub-instances	30–40
Report tasks	6–8
Public sample tasks	2–3
Development tasks	3–4
Private test tasks	5–6

This scope is designed to make v0.1 executable while preserving methodological quality.

5.2 Report Tasks

Report tasks are full research briefs answered as decision-grade research memos or reports.

CRB v0.1 includes three Consumer-domain report families:

5.2.1 Usage and Attitude

Example task:

Why are young urban consumers choosing matcha over coffee? Map motivations, occasions, barriers, and consumer language.

5.2.2 Consumption Occasion Mapping

Example task:

Map emerging occasions for low-sugar sparkling beverages among Gen Z consumers in India.

5.2.3 Trendspotting and Early Signal Detection

Example task:

Identify emerging consumer-created beverage rituals showing early signs of repeatability.

Future versions will extend the taxonomy to Brand Perception Diagnosis, Brand Tracking, Competitive Perception Analysis, Campaign Measurement, Whitespace Mapping, Digital Ethnography, Creator Discovery, Culture Mapping, and Subculture Immersion.

5.3 Design Principle: Include Tie-or-Lose Tasks

At least one quarter of v0.1 tasks should be **data-neutral**: tasks where specialized consumer-signal data is expected to provide little or no advantage over a strong public-web or deep-research baseline.

This is an important design principle. If the benchmark only contains tasks where proprietary or specialized consumer data is expected to help, it risks functioning as a demonstration rather than an evaluation. Including tasks where specialized data provides little advantage makes the observed wins on data-advantaged tasks more credible.

Each task is also tagged with the failure mode it is designed to provoke.

6. RetroSignals: Frozen Consumer-Signal Environment

A key challenge in evaluating research agents is that the underlying information environment changes continuously. Posts are deleted, comments accumulate, search results change, platforms modify ranking systems, and new summaries or analyses appear after a task is created. Without a frozen environment, model comparisons become unstable.

CRB introduces **RetroSignals**, a frozen, task-linked consumer-signal environment.

RetroSignals may include:

1. Social posts.
2. Comments and replies.
3. Reviews.
4. Marketplace content.
5. Community discussions.
6. Search-result pages.
7. Blog and article content.
8. Video metadata.
9. Video and audio transcripts.
10. Creator content.
11. Brand and competitor context documents.

Each signal is stored with metadata such as source type, timestamp, geography where available, language, engagement indicators, parent-child conversation structure, and enrichment labels.

Two arms are served from the same frozen snapshot:

1. **Consumer-signal corpus arm**: enriched and contextually filtered consumer corpus.
2. **Public-web arm**: public snapshot only.

These arms enable the data × model ablation described later.

6.1 Signal Conditioning

CRB evaluates not only whether an agent retrieves data, but whether it retrieves conditioned, research-ready signal.

Signal conditioning includes:

1. Relevance filtering.
2. Spam and bot filtering.
3. Promotional-content detection.
4. Brand-owned versus consumer-owned separation.
5. Duplicate and near-duplicate control.
6. Contextual filtering.
7. Conversation threading.
8. Privacy-safe sentiment, need-state, occasion, stage, and cohort classification.

For example, in a task about nostalgia-driven beverage behavior, a weak system may retrieve earnings reports, memorabilia listings, stock-market discussion, or search-optimized recipe pages. A stronger system should retain consumer-created

recipes, consumption rituals, comments, creator videos, product pairings, and contextual conversation around the behavior.

Conditioning quality is reported as a diagnostic in v0.1. It is not folded into a composite score until the metric is validated.

6.2 Signal Quality Diagnostics

CRB reports three signal-quality diagnostics: **data breadth**, **data depth**, and **data condition**. These diagnostics describe the evidence environment available to a system and the quality of signal used in its output. They are not included in a headline composite score in v0.1.

6.2.1 Data Breadth

Data breadth measures whether the system covers a sufficiently diverse consumer evidence universe.

Evaluation criteria include:

1. Source coverage across social platforms, reviews, marketplaces, video, search, blogs, communities, images, and transcripts.
2. Platform diversity, rather than over-reliance on a single source.
3. Geographic coverage where relevant.
4. Format coverage across text, image, video, audio, comments, reviews, and transcripts.
5. Competitive and substitute coverage.
6. Long-tail coverage of niche but meaningful consumer pockets.

6.2.2 Data Depth

Data depth measures whether the system captures full consumer context rather than only surface-level mentions.

Evaluation criteria include:

1. Conversation depth across posts, comments, replies, and subthreads.
2. Context depth around occasion, emotion, need, trigger, and barrier.
3. Consumer depth around cohort, life stage, psychographic pattern, or behavior pattern.
4. Evidence depth through inspectable raw signals behind each insight.
5. Motivation depth around why consumers behave in a certain way.
6. Tension depth around contradictions, trade-offs, and unmet needs.

6.2.3 Data Condition

Data condition measures whether retrieved evidence is ready for research use.

Evaluation criteria include:

1. Relevance condition.
2. Freshness where required by the task.
3. De-duplication.
4. Spam and bot filtering.
5. Separation of consumer behavior from investor, news, brand, or promotional content.
6. Enrichment completeness.
7. Traceability to source evidence.

Data condition is especially important because poor retrieval can produce plausible but invalid research outputs. A large evidence base is not useful if it is noisy, duplicated, promotional, stale, or untraceable.

6.3 Live versus Frozen Validation

For a subset of tasks, CRB should compare live-corpus performance against RetroSignals performance.

The purpose is to test whether the frozen environment preserves relative model rankings. If relative rankings are stable, RetroSignals becomes the canonical evaluation environment for official scoring.

7. Dataset Construction

CRB tasks are constructed from three sources.

7.1 De-identified Real Research Briefs

The primary source is a library of real consumer research briefs. These briefs provide construct validity because they reflect actual commercial demand.

All briefs are de-identified. Confidential information is removed. Brand names are retained only where public and permissible; otherwise, they are replaced with synthetic equivalents.

7.2 Expert-Crafted Challenge Tasks

Senior consumer-insights researchers, brand strategists, category experts, and cultural analysts create tasks designed to test specific capabilities and failure modes. Experts are credited where possible, following the practice of professional-domain benchmarks.

7.3 Public Consumer Phenomena

Some tasks are built around public trends, campaigns, product launches, or cultural moments. These tasks are useful for public examples, external demos, and easier validation.

7.4 v0.1 Size and Authoring Cost

CRB v0.1 contains approximately 12 Consumer-domain research briefs, each with an expert reference report or worked answer.

This size is intentional. Each task requires:

1. A research brief.
2. A frozen RetroSignals snapshot.
3. A reference answer or report.
4. Evidence checklist.
5. Rubric weights.
6. Tagged failure modes.
7. Human evaluation protocol.
8. Validity gates.

The constraint is not task generation but defensible task authoring. A smaller set of high-quality tasks is preferable to a larger set of weakly specified tasks.

7.5 Contamination Splits

CRB uses three splits:

1. **Public sample**: released for demos and onboarding.
2. **Development set**: used for model and workflow development by approved participants.
3. **Private test set**: never released or trained on; used for official scoring.

Where possible, CRB also includes a date-stamped holdout of consumer conversations collected after relevant model training cutoffs.

Tasks are refreshed quarterly to reduce contamination and saturation.

8. Evaluation Framework

CRB v0.1 produces two scored outputs and a set of diagnostics.

8.1 Scored Outputs

Objective Track Score
Atomic tasks are scored against worked ground truth and evidence checklists.

Report Track Score

Open-ended reports are scored using C-RACE and validated against human experts.

8.2 Diagnostics

Diagnostics are reported but not included in a headline composite score in v0.1.

Diagnostics include:

1. Data breadth.
2. Data depth.
3. Data condition.
4. Relevant signal density.
5. Signal retrieval breadth.
6. Signal conditioning purity.
7. Quant–qual integration.
8. Contextual fidelity.
9. Decision-readiness checklist.
10. Speed.
11. Cost.
12. Trace failure rates.
13. Reproducibility.
14. Calibration.

CRB v0.1 does not define an overall composite score. A premature composite would obscure which capability is actually being measured.

9. Objective Track Score

Each atomic task type is scored against an expert-worked answer or evidence checklist.

Scoring methods include:

1. **Binary scoring** for Find Signal and Derive Metric tasks.
2. **Accepted range scoring** for derived metrics where exact values may vary by interpretation.
3. **Recall or F1** for Define Cohort, Populate Set, and Compile Dataset tasks.
4. **Absolute probability difference** for Validate Claim tasks.
5. **Expert-judged match** for Attribute Driver tasks.

LLM assistance may be used for fuzzy matching, such as recognizing equivalent paraphrases. It should not be used for the core objective judgment unless the judgment protocol has been validated against human expert labels.

9.1 Relevant Signal Density

CRB reports **Relevant Signal Density** as a diagnostic measure of retrieval efficiency:

```
Relevant Signal Density = Relevant evidence items / Total evidence items reviewed
```

This metric distinguishes systems that retrieve large volumes of weak evidence from systems that retrieve a smaller but more useful set of consumer signals. It is especially useful for comparing research agents, public-web agents, and social-listening workflows under similar evidence-review budgets.

10. Evidence Grounding: Statement-Signal Verification

Evidence grounding adapts citation verification to consumer-signal environments.

The system's report is decomposed into claims. Each claim is linked to supporting signal(s). The underlying signal is retrieved. A judge then renders a binary support judgment.

10.1 Metrics

Claim Support Accuracy
Supported statement-signal pairs divided by total statement-signal pairs.

Effective Evidence
Total supported statement-signal pairs divided by number of tasks.

Fabrication Rate
The fraction of cited signals that do not exist or do not say what the report claims.

Triangulation
Whether key insights are supported across multiple signals, sources, or cohorts.

Fabrication Rate is a central trust metric. A system that produces compelling but fabricated consumer evidence should fail regardless of fluency.

A cost-efficient judge model may be used for support judgment if its decisions are validated against human labels.

11. Report Track: C-RACE

The Report Track is scored by **C-RACE — Consumer Research Adaptive Criteria Evaluation**.

C-RACE is a reference-based, adaptive evaluation framework. A judge model generates task-specific dimension weights and criteria, then scores the target report relative to an expert reference report. Isolated scoring is avoided because fluent reports often receive uniformly high scores even when they differ meaningfully in quality.

11.1 C-RACE Dimensions

Dimension	What It Measures
Comprehensiveness	Coverage of relevant groups, occasions, motivations, counter-signals, and category context
Consumer Insight Depth	Quality of interpretation around motivations, tensions, rituals, identity signals, barriers, and behavioral shifts
Methodology Fidelity	Whether the report follows the appropriate research method
Evidence and Traceability	Whether claims are grounded in inspectable consumer signal
Quant–Qual Integration	Whether quantitative patterns are connected to qualitative evidence and interpretation
Contextual Fidelity	Whether findings are interpreted in relation to relevant market, category, cultural, and business context
Decision Readiness	Whether the output provides implications and recommended actions
Readability and Structure	Whether the report is clear, scannable, and usable by business stakeholders

11.2 Dynamic Weighting

Different task types require different weights.

A trendspotting task should weight weak-signal discovery and cultural interpretation more heavily. A usage-and-attitude task should weight comprehensiveness, segmentation, and motivation. A whitespace task should weight strategic synthesis and decision readiness.

For each task, the judge model or expert panel assigns weights across the dimensions before scoring outputs.

11.3 Reference-Based Scoring

The final report score is computed relative to the reference report:

```
S_final(target) = S_int(target) / (S_int(target) + S_int(reference))
```

The reference is the expert-authored gold answer.

Report rankings and proportional differences should be emphasized over absolute score values.

11.4 Reference Report Dependency

C-RACE cannot be run without high-quality expert reference reports. Authoring these references is therefore the key dependency for the Report Track.

11.5 Human-Consistency Validation

C-RACE is credible only if it aligns with human expert judgment.

Using the human evaluation panel, CRB reports:

1. Pairwise Agreement Rate.
2. Overall Pearson Correlation.
3. Filtered Average Pearson Correlation.
4. Filtered Average Spearman Correlation.
5. Intraclass Correlation.

CRB reports whether C-RACE approaches or exceeds human inter-annotator agreement. Until this validation is complete, automated Report Track scores are marked provisional.

12. Quant–Qual Integration

Consumer research outputs are often strongest when quantitative structure and qualitative interpretation are combined. CRB therefore evaluates whether a system can connect measurable patterns with consumer meaning.

Evaluation criteria include:

1. **Sizing**: whether the report quantifies relevant themes, mentions, growth, or confidence where appropriate.
2. **Classification**: whether themes, needs, sentiments, occasions, and cohorts are labeled correctly.
3. **Distribution**: whether patterns are compared across cohort, market, source, or time where relevant.
4. **Qualitative evidence**: whether representative consumer signals support the quantitative pattern.
5. **Synthesis**: whether the system connects numerical patterns to human motivation or behavior.
6. **Confidence**: whether the stated confidence is supported by evidence volume, consistency, and triangulation.

A system should be penalized when it provides confident-sounding numbers without clear grounding, or when it provides qualitative interpretation without showing whether the pattern is isolated, emerging, or widespread.

13. Contextual Fidelity

Consumer signals require interpretation in context. CRB therefore reports **Contextual Fidelity** as part of the Report Track and as a diagnostic.

Contextual Fidelity evaluates whether the system interprets findings in relation to:

1. Category context.
2. Competitive context.
3. Market and geography.
4. Consumer reality.
5. Cultural references and local language.
6. Business objective.
7. Relevant brand, product, or use-case context where provided.

A low-context answer may be factually plausible but strategically weak. A high-context answer should explain not only what consumers are saying, but why it matters in the specific category, market, cohort, or business situation.

14. Human Evaluation Protocol

Each task is reviewed by at least three evaluators:

1. Consumer-insights researcher.
2. Brand or category expert.
3. Research-quality reviewer.

For culture-heavy tasks in future versions, a cultural or local-market expert should also be included.

Evaluators score outputs blindly and do not know which system produced which response.

Human evaluation includes:

1. Absolute rubric scoring.
2. Pairwise preference.
3. Evidence audit.
4. Decision-readiness judgment.
5. Qualitative failure notes.

Inter-rater agreement is computed. Low-agreement tasks are revised or excluded.

These human scores are used to validate C-RACE rather than merely supplement it.

15. Data × Model Ablation

A central goal of CRB is to separate model quality from data advantage.

CRB crosses the model factor with the data factor on the same frozen tasks.

	Public-Web Arm	**Consumer-Signal Corpus Arm**
Generic frontier model	Baseline deep-research capability	Isolates data contribution
Consumer research model under evaluation	Isolates model contribution	Full system

CRB also includes a tool-less control where the model answers from memory without corpus access. This measures how much data helps at all.

The ablation addresses two common interpretations separately:

1. The system performs better because it has better data.
2. The system performs better because it has a better model.

The ablation should also report where specialized consumer data does not improve performance.

16. Speed

Speed is reported jointly with quality, not independently.

16.1 Metrics

Time to Decision-Ready Answer
Wall-clock time until the output meets the grounding bar and answers the brief.

Human Effort
Number of prompts, edits, corrections, or analyst interventions required.

Cost per Task
Compute, retrieval, and tool cost.

CRB reports the depth-at-speed claim as:

Decision-ready Report Track score per unit time at a fixed grounding bar.

The research compression ratio may be used illustratively, but it is not scored because it assumes equivalence between the agent output and human research output, which is precisely what the benchmark is intended to evaluate.

17. Validity Gates

Some failures should not merely reduce a score. They should invalidate a run or cap the maximum score because they violate the minimum evidentiary standards of consumer research.

17.1 Run-Level Failure Conditions

A run should be marked invalid if the system:

1. Fabricates consumer quotes.
2. Fabricates statistics, source counts, or quantitative claims.
3. Cites evidence that does not exist.
4. Produces key claims without inspectable raw evidence.
5. Cannot provide source-level traceability for central conclusions.

17.2 Score-Capping Conditions

A run should receive a capped score if the system:

1. Confuses brand-owned content with organic consumer signal.
2. Treats news, investor commentary, or PR material as consumer behavior.
3. Overgeneralizes from one platform or one viral post.
4. Ignores obvious contradictory evidence.
5. Produces generic insights that could apply to many brands or categories.
6. Provides recommendations without evidence.
7. Fails to explain or follow the intended methodology.

Validity gates make the benchmark harder to game. They also protect against fluent but unsupported outputs receiving high scores.

18. Trace-Level Failure Analysis

CRB evaluates agent traces to diagnose why systems fail.

Each failure mode is mapped to at least one task designed to provoke it.

Failure Mode	Description
Platform Myopia	Over-reliance on one platform or source type
Keyword Literalism	Search behavior limited to explicit brand or category terms, missing consumer-native language
Shallow Listening	Reporting mentions or sentiment without motivation or context
Consumer-Brand Confusion	Treating brand-owned content, news, PR, or investor commentary as consumer signal
Viral Post Overgeneralization	Treating one high-engagement item as representative of broader behavior
Fabricated Verbatims	Inventing or distorting consumer quotes
Weak Triangulation	Making major claims from insufficient or weak evidence
Missed Counter-Signals	Ignoring contradictory or minority evidence
Methodology Drift	Producing an output that does not match the intended research method
Generic Insighting	Producing claims that could apply to many brands or categories
Cultural Misread	Misinterpreting memes, irony, local language, creator context, or subculture
Segmentation Collapse	Treating consumers as a single homogeneous group
Premature Satisficing	Stopping after a plausible but incomplete answer
Data Overclaiming	Making claims beyond what the evidence supports

This taxonomy is both diagnostic and generative: it should inform task design as much as model evaluation.

19. Public, Development, and Private Splits

CRB uses three splits.

19.1 Public Sample

A small set of tasks released for transparency, demos, and onboarding.

19.2 Development Set

Tasks used for model and workflow development by approved participants.

19.3 Private Test Set

Hidden tasks used for official scoring. Prompts, evidence checklists, reference reports, and RetroSignals snapshots are not released.

CRB refreshes tasks quarterly to reduce contamination and saturation.

20. Governance and Participation

CRB is intended to be administered by an independent research group or benchmark consortium.

20.1 Benchmark Maintainer

The CRB Working Group is responsible for:

1. Defining the benchmark construct.
2. Recruiting expert task authors.
3. Maintaining public, development, and private splits.
4. Running official evaluations.
5. Validating automated judges against human experts.
6. Publishing results and limitations.
7. Managing benchmark refreshes.

20.2 Participant Systems

Participant systems may include:

1. Consumer research AI models.
2. General-purpose LLMs with web search.
3. Deep research agents.
4. Social-listening or market-intelligence workflows.
5. Human researcher baselines.

Commercial vendors may submit systems for evaluation, but they do not control scoring, task selection, reference outputs, or leaderboard publication.

20.3 Data Providers

Some benchmark environments may include data contributed by commercial or academic data providers. Any such contribution must be disclosed. Data providers should not be allowed to influence task selection, scoring criteria, or official reporting for tasks where their systems are evaluated.

20.4 Release Strategy

CRB should be released in stages:

1. Internal pilot by the independent working group.
2. Public sample release.
3. Private held-out evaluation.
4. Verified leaderboard.
5. Rolling benchmark refresh.

21. Example Benchmark Instance

Task

Why are young urban consumers choosing matcha over coffee? Produce a consumer-research memo identifying motivations, occasions, barriers, consumer language, emerging rituals, and implications for a premium beverage brand.

Required Output

1. Executive summary.
2. Key motivations.
3. Consumption occasions.
4. Consumer cohorts.
5. Barriers and tensions.
6. Evidence table with raw signals.
7. Opportunity spaces.
8. Brand implications.
9. Confidence assessment.

Strong Answer Characteristics

A strong answer goes beyond the claim that matcha is healthy or trendy. It identifies more specific patterns such as clean energy, mid-afternoon productivity, aesthetic self-expression, coffee replacement, wellness signaling, at-home ritualization, premiumization, creator influence, and taste, price, or authenticity barriers. Each claim should be grounded in inspectable signal.

Scoring

Report Track:

1. C-RACE dimensions.
2. Dynamic weighting toward depth and segmentation for a usage-and-attitude task.
3. Quant–Qual Integration.
4. Contextual Fidelity.
5. Human validation.

Objective Track:

1. Grounding accuracy.
2. Fabrication Rate.
3. Relevant Signal Density.
4. Atomic sub-tasks such as Define Cohort recall or Validate Claim.

Diagnostics:

1. Data breadth.
2. Data depth.
3. Data condition.
4. Occasion specificity.
5. Source breadth.
6. Time to be decision-ready.
7. Cost.

Validity Gates:

1. Fabricated consumer quote: invalid run.
2. Unsupported statistic: invalid run.
3. No inspectable evidence for central claims: score cap.
4. Generic answer without cohort, occasion, or category specificity: score cap.

22. Results Format

No empirical scores are reported until the benchmark is run. Cells should remain blank rather than estimated.

22.1 Report Track and Grounding

System	C-RACE Overall	Depth	Methodology	Quant–Qual	Context	Evidence / Trace	Decision	Grounding Acc.	Fabrication	Time-to-Ready
Consumer Research Model A	—	—	—	—	—	—	—	—	—	—
Consumer Research Model B	—	—	—	—	—	—	—	—	—	—
Deep Research Agent	—	—	—	—	—	—	—	—	—	—
Social Listening Workflow	—	—	—	—	—	—	—	—	—	—
Human Researcher	—	—	—	—	—	—	—	—	—	—

22.2 Signal Quality Diagnostics

System	Data Breadth	Data Depth	Data Condition	Relevant Signal Density	Source Diversity	Conversation Depth
Consumer Research Model A	—	—	—	—	—	—
Consumer Research Model B	—	—	—	—	—	—
Deep Research Agent	—	—	—	—	—	—
Social Listening Workflow	—	—	—	—	—	—
Human Researcher	—	—	—	—	—	—

22.3 Ablation

	Public-Web Arm	Consumer-Signal Corpus Arm
Generic Model	—	—
Consumer Research Model	—	—
Tool-less Control	—	n/a

22.4 Validity Gates

System	Invalid Runs	Score-Capped Runs	Fabricated Quotes	Unsupported Statistics	No Evidence for Key Claims
Consumer Research Model A	—	—	—	—	—
Consumer Research Model B	—	—	—	—	—
Deep Research Agent	—	—	—	—	—
Social Listening Workflow	—	—	—	—	—

23. Limitations

23.1 Vendor and Data-Provider Involvement

CRB is designed to evaluate systems that may be built by commercial vendors. Some vendors may also contribute data, infrastructure, or funding. This creates a potential conflict of interest.

Mitigations include:

1. Independent benchmark governance.
2. Independent task authors who are not employed by evaluated vendors.
3. Blind grading by evaluators who do not know which system produced which output.
4. Pre-registered criteria locked before results are seen.
5. Sealed holdout tasks not exposed to participants.
6. Full method and prompt disclosure for public tasks.
7. Public sample subset.
8. Clear separation between development use and official benchmark claims.
9. Data × model ablation to show whether performance advantages come from data, model, or system design.

No evaluated vendor should control the official test set, scoring criteria, reference reports, or leaderboard publication.

23.2 Subjectivity

Insight quality is partly subjective. Experts may disagree on what counts as strong. CRB mitigates this through multiple judges, pairwise comparison, reference reports, ICC filtering, and reporting rankings and gaps rather than absolute scores.

23.3 Data Bias

Organic conversation may overrepresent vocal, urban, younger, platform-specific, or high-engagement cohorts. CRB should evaluate whether agents acknowledge these limitations.

23.4 Cultural Interpretation

Agents may misread irony, slang, memes, local language, creator context, or subculture. Culture-heavy tasks in future versions should include local-market or cultural experts.

23.5 Frozen versus Live Environment

RetroSignals improves reproducibility but may reduce realism. CRB should pair frozen evaluation with rolling live validation.

23.6 Scale

Expert-authored tasks are expensive. v0.1's small size limits statistical power. Expanding the task set while preserving quality is the main future-work priority.

23.7 Contamination

Public tasks can leak into training data. CRB mitigates this with private held-out tasks, date-stamped data snapshots, task refreshes, and contamination checks.

23.8 Diagnostic Metric Validation

Data breadth, data depth, data condition, relevant signal density, contextual fidelity, and quant–qual integration are useful diagnostics, but they require validation before they can be safely combined into a headline composite score. In v0.1, CRB reports them separately.

24. Privacy and Responsible Data Use

Consumer research agents operate on human expression. Privacy and responsible data use are therefore core benchmark-design requirements.

CRB follows these principles:

1. Personally identifiable information should not be exposed in benchmark outputs.

2. Usernames, handles, and private details should be removed or anonymized.
3. Cohort-level claims should use aggregation thresholds.
4. Sensitive attribute inference should be prohibited unless explicitly supported, privacy-safe, and ethically approved.
5. Quotes should be anonymized or paraphrased where required.
6. Platform terms, data licenses, and source restrictions should be respected.
7. Psychographic and cohort labels should be used cautiously and should not be treated as unsupported individual-level claims.
8. Systems should be penalized for stereotyping, manipulative recommendations, or overconfident psychological profiling.

CRB should reward systems that are evidence-grounded, privacy-safe, transparent, and honest about uncertainty.

25. Broader Impact

A rigorous consumer-research benchmark can support the development of more reliable, transparent, and useful AI research systems. It may lower the cost of understanding consumers and help organizations make better product, marketing, brand, and innovation decisions.

The same capabilities could also enable manipulative targeting, cultural exploitation, privacy-invasive profiling, or over-optimized persuasion. CRB should therefore penalize fabricated consumer claims, unsupported psychological inference, and overstated confidence. It should encourage systems that separate observation from interpretation and clearly disclose uncertainty.

26. Conclusion

Consumer research is becoming an AI-native workflow. Evaluating this workflow requires a benchmark designed for messy organic data, noisy signal, cultural context, methodology-specific reasoning, quant–qual integration, and decision-ready synthesis.

CRB combines:

1. A frozen consumer-signal environment.
2. Expert-authored briefs and reference reports.
3. An objective grounding layer.
4. A human-validated adaptive report rubric.
5. A data × model ablation.
6. Signal-quality diagnostics.
7. Quant–qual and contextual-fidelity evaluation.
8. Validity gates for hard failures.
9. A domain-specific failure taxonomy.
10. Independent governance and safeguards against vendor influence.

CRB v0.1 is scoped to a single domain and approximately 12 tasks so that it can be executed before it is expanded.

The central question is:

Can an AI system conduct consumer research, not merely search the web?

CRB v0.1 provides a methodology for answering that question.

The next step is to run the first benchmark instance and fill the first cell in the results table.

References

[1] FutureSearch / Bosse et al. “Deep Research Bench: Evaluating AI Web Research Agents.” 2025.

[2] Du, M., Xu, B., Zhu, C., Wang, X., Mao, Z. “DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents.” 2025.

[3] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., Narasimhan, K. “SWE-bench: Can Language Models Resolve Real-World GitHub Issues?” 2024.

[4] Guha, N., et al. “LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in LLMs.” 2023.

[5] Zheng, L., Chiang, W.-L., Sheng, Y., et al. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.” 2023.

[6] Mialon, G., et al. “GAIA: A Benchmark for General AI Assistants.” 2023.

[7] Yao, S., et al. “ReAct: Synergizing Reasoning and Acting in Language Models.” 2023.