# ASKINFER-BENCH: PROFILES ARE EVIDENCE, NOT INTENT LABELS, IN PERSONALIZED DEEP RESEARCH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Personalized deep-research agents are often evaluated using criteria derived from the same persona supplied to the agent. This tests responsiveness, but it can silently turn a model's inference about a user into ground truth about the user's current task. We frame personalization under incomplete context as an information-acquisition problem: a profile is evidence about task intent, not the intent itself. AskInfer-Bench contains 15 tasks whose reference preferences are all human-provided, varies how much persistent profile evidence is initially visible, and independently controls clarification access. In the audited Prolific batch, ten adults supplied 107 raw statements, consolidated into 73 high-impact task-preference axes and 22 secondary preferences. A PDR-Bench-style GPT-5 generator produced 379 criteria for those ten tasks. Although its full lists attained 80.8% strict recall under an LLM-favorable semantic merge, recall fell to 39.7% after matching the 73-criterion human budget using the generator's own weights. The full-factorial system study contains 360 reports across two harnesses. Native clarification has near-zero average effects in the complete Gemini–DeerFlow matrix, whereas Open Deep Research with an Influence–Evidence–Ownership (IEO) question selector yields positive Ask–Off effects for three backbones under cold and RAW50 context (+0.71 to +0.78 and +0.27 to +0.34) but negative effects under RAW100 (-0.63 to -0.74). A separate five-pair dual-scoring diagnostic changes the average Ask effect from -0.2508 under human-grounded rubrics to +0.2924 under persona-derived rubrics and reverses the win/tie/loss conclusion for two pairs. This small diagnostic is not a population estimate or significance test; it illustrates why personalization benchmarks must separate using supplied information from deciding what still needs to be learned.

## 1 INTRODUCTION

Deep-research agents turn an underspecified request into a long process of planning, retrieval, synthesis, and recommendation. The user's consequential trade-offs are rarely all explicit. Some may be supported by prior interaction; others may be absent, stale, conflicting, private, or specific to the present decision. A personalized agent must therefore do more than condition on a profile. It must identify what remains uncertain and decide whether to infer conservatively, ask, branch the answer, or state an assumption.

Many evaluations construct a persona, expose some or all of it to the agent, and derive evaluation criteria from that same persona. This is a valid test of conditional generation, but not by itself a valid test of preference acquisition. A persona-derived criterion may encode a plausible inference that the task owner never endorsed. Once placed in the rubric, that inference becomes an answer key: the agent is rewarded for reproducing it and may be penalized for asking whether it is correct. Responsiveness to supplied information is then conflated with evidence-bounded inference, elicitation of missing preferences, and use of newly acquired information.

This distinction matters even for rich profiles. More context can add relevant evidence while also adding irrelevant or weakly diagnostic cues. It can increase confidence, suppress clarification, or steer questions toward stereotypes rather than current decisions. If $H_u$ is persistent context and $Z_{u,q}$ is the user's preference state for task $q$, then $H_u$ informs $Z_{u,q}$ but generally does not identify



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

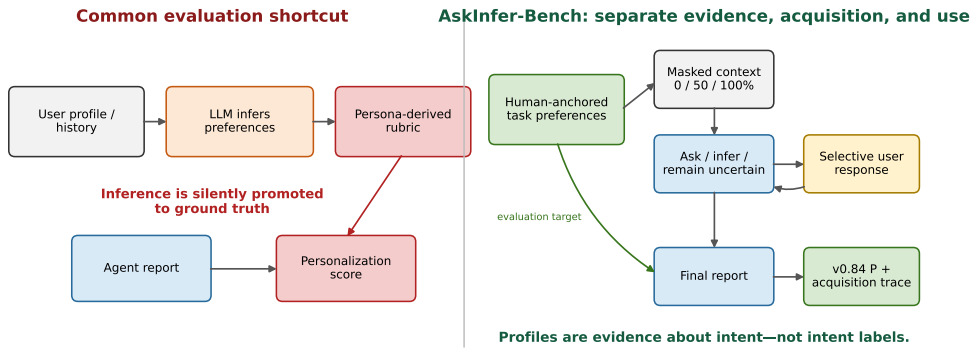


Figure 1: Profiles are evidence about task intent, not task-intent labels. A common shortcut derives scoring criteria from a model's interpretation of a profile. AskInfer-Bench instead freezes task-owner preferences, masks initially visible evidence, observes asking and inference, and evaluates acquisition and the final report separately.

it: $p(Z_{u,q} \mid H_u, q)$ need not be a point mass. A profile-derived estimate should not automatically become reference truth.

AskInfer-Bench evaluates this gap. Its human-grounded core fixes task-specific preferences independently of the report model, masks initially visible context, and records whether missing preferences are asked, resolved, and reflected. A frozen selective-disclosure simulator answers questions from the task owner's latent state without defining the ground truth. The report is scored against human-grounded semantic leaves across the common 15-task benchmark design.

Our experiments reveal three patterns. First, richer profile context substantially improves final-report scores in the complete Gemini–DeerFlow matrix, but RAW100 does not improve over RAW50. Second, native clarification is rarely invoked by that system and has small average effects. Third, the complete Open Deep Research matrix reproduces a context-dependent reversal across GPT, Claude, and Gemini backbones: IEO effects are positive under sparse or partial context and negative under full context. These results do not establish that asking is harmful in rich context. They show that permission to ask, even with an explicit selector, is not sufficient to determine what is worth asking.

Our contributions are:

1. We formalize persistent user context as evidence about, rather than an identity label for, current-task preferences, separating responsiveness, inference, acquisition, and utilization.
2. We provide task-owner-grounded evidence that broad persona-derived rubrics can obtain high apparent recall through overgeneration while prioritizing the owner's compact high-impact set poorly.
3. We introduce a provenance-aware protocol combining controlled context masks, clarification access, a selective-disclosure channel, acquisition traces, and human-grounded report scoring.
4. We test clarification across two deep-research harnesses and show that the cold/partial-context benefit and rich-context reversal replicate across three Open Deep Research backbones, while treating IEO as an evaluation scaffold rather than a central algorithmic contribution.

## 2 Problem Formulation

For a user $u$ and task $q$, let $H_u$ denote persistent pre-task evidence: explicit profile fields, authorized history, or prior dialogue. In this paper, **persona** refers only to this persistent evidence. Let

$$Z_{u,q} = \{z_1, \ldots, z_K\}$$

be user-owned goals, constraints, priorities, and trade-offs that can change the evidence, ranking, risk posture, or action plan for $q$. Biographical facts do not become preference units unless the user links them to a task consequence. The central distinction is $H_u \neq Z_{u,q}$, with $H_u$ providing defeasible evidence about $Z_{u,q}$.

A condition reveals $C_r = M_r(H_u)$ for $r \in \{0, 0.5, 1\}$, corresponding to cold start, RAW50, and RAW100. RAW100 exposes all eligible persistent profile facts, not $Z_{u,q}$ and therefore not an oracle. At step $t$, the system either asks $x_t$ or proceeds. A selective user channel returns $y_t$ only when the latent state supports and permits disclosure of an answer. The report generator produces $R = G(q, C_r, \{(x_t, y_t)\}_{t=1}^{T})$. We record, for every task-preference axis, whether it was initially known, asked, resolved, and reflected.

These events define four separable capabilities: responsiveness to explicit information; evidence-bounded inference from context; elicitation of consequential user-owned uncertainty; and utilization of resolved preferences. Final-report-only evaluation primarily observes the first and fourth. It cannot determine whether the system discovered a preference, guessed it, or received it directly.

# 3 AskInfer-Bench

## 3.1 Human-grounded tasks and preferences

The system-evaluation slice contains 15 long-form tasks across education, career, health, travel, finance, creative work, shopping, real estate, and parenting. All 15 use human-provided user context and human-provided current-task preferences as ground truth under the same benchmark design and authority rule. Ten task-owner records come from the audited Prolific batch described below; five come from the earlier human-user collection. The latter contribute 25 human-provided axes, not persona-derived LLM reference criteria. The human–LLM rubric analysis uses the ten-task Prolific batch for which the complete elicitation and ranking records are available.

Participants were recruited through Prolific in September 2026 using screening and task matching. The primary wave opened 25 places, produced 23 approved screeners, and identified 20 eligible respondents; targeted backfill filled uncompleted or quality-flagged task slots. The analytic set contains ten non-overlapping adults, one per task. All completed four consent/privacy confirmations. Compensation was US$0.40 for screening and US$9 for the main task. Self-reported completion time averaged 73.8 minutes (median 72.5, range 35–120).

Elicitation separated spontaneous preferences from prompted reflection. Participants first described what a high-quality report must satisfy, stated importance and consequences, and ranked priorities. Only then did they see five neutral reminder dimensions and accept, revise, or reject additional directions. They also identified what an agent should still confirm despite rich history. This ordering prevents model prompts from defining the initial human reference.

The 107 Part A and Part B statements were atomicized and deduplicated within participant, yielding 73 high-impact axes (6–9 per task) and 22 secondary preferences. Task owners are authoritative about their goals and trade-offs; external evidence remains authoritative about feasibility and safety. Construction used model-assisted structured screening with researcher adjudication. A later two-annotator rubric-construction audit reported 89.0% raw agreement, Cohen's $\kappa = 0.78$, and priority-weighted $\kappa = 0.84$; human review changed 14–16% of initially drafted leaves, and 9–11% of annotation units required adjudication.

## 3.2 Why persona-derived rubrics are not ground truth

We reproduced the released PDR-Bench personalization-criterion generator using its official English prompts and repository code. For each human-grounded task, the generator saw the task and a preference-blind rich profile, but not participant preference answers, rankings, importance labels, ground truth, or clarification examples. It generated 379 criteria, versus 73 human axes.

Two API judges labeled semantic relationships. To favor the generated rubric, a human axis counted as fully covered if either judge assigned a full match, and otherwise as partially covered if either assigned a partial match. Under this union, 59 of 73 axes were fully covered, 13 partially covered, and one missed: 80.8% strict recall and 89.7% partial-credit recall. At least 25.1% of generated

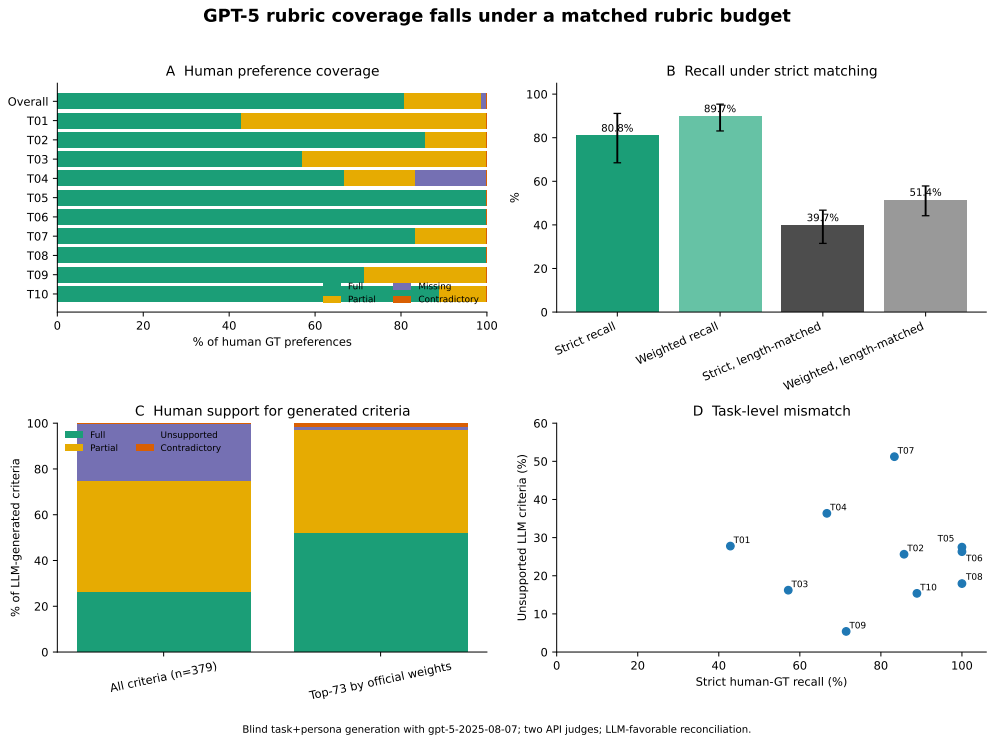


Figure 2: Generated rubrics recover human preferences mainly through breadth. Full-list recall is high when GPT-5 generates 379 criteria for 73 human axes, but falls sharply after matching the criterion budget using the generator's own weights. Error bars are 95% task-bootstrap intervals; semantic matches use the LLM-favorable union of two API judges.

criteria lacked support in the elicited human reference. The two judges agreed only modestly, so these are conservative model-mediated labels, not human adjudication.

High recall depended on a 5.19-fold criterion surplus. We ranked criteria by the generator's own hierarchical weight and retained the top 73. Strict recall fell to 39.7% and partial-credit recall to 51.4%; paired task-bootstrap decreases were 41.1 and 38.4 percentage points. Broad generated lists can therefore appear comprehensive without identifying which criteria matter most to the task owner.

### 3.3 Context, simulator, and scoring

RAW0 exposes no persistent profile facts; RAW50 exposes a fixed-seed half-sample of atomic facts; RAW100 exposes all eligible persistent facts. The task instruction is always visible. Participant answers, priority tiers, rubric text, clarification examples, and withheld state values remain hidden. Thus RAW100 is full profile exposure, not task-preference oracle access.

All Ask-enabled episodes used the frozen v0.107 simulator pipeline with Alibaba Cloud Model Studio Qwen3.7-Max (`qwen3.7-max-2026-05-20`): thinking disabled, temperature 0, top-$p = 1$, seed 20260924, strict JSON output, zero retries, and a 512-token limit. The simulator is blind to the report model, harness, condition, hypothesis, and evaluation. Run-level asking flags record 132 IEO interactions among 135 enabled ODR episodes; seven additional native-Ask episodes invoked the same channel, for 139 observed production interactions. Full `simulator_trace.json` bundles were retained for an availability-defined five-task subset of 35 interactions (28 IEO and seven native Ask), not a random sample of the 139. This trace-bearing subset required 70 calls, 229,958 input tokens, and 11,793 output tokens and preserves the question, classifier output, permitted keys, response, usage, and blinding metadata. Archived valid production traces (35 exchanges) account for CNY 4.971342; the total simulator execution ledger, including failed attempts, is CNY 8.064610. Formally audited trace-bearing exchanges remain $0/35$. A shorter 24-answer development probe

found 0/24 developer-identified unsupported inventions and overdisclosures, but it is not a matched-production-budget audit. The simulator is a controlled disclosure channel, not the source of preference truth.

The final asset, $P_{\text{human-semantic}}$ v0.84, contains 98 human-grounded preference axes and 409 directly entailed semantic leaves across the 15 tasks. The audited Prolific batch contributes 73 axes and 309 leaves; the earlier human-user batch contributes 25 axes and 100 leaves. Each leaf receives an integer 0–10 score with direct report evidence. Leaves average within preference; preference scores use weights 3, 2, 1, and 1 for Must-have, Strong, Nice-to-have, and unranked axes. Unranked denotes unavailable priority metadata, not weaker ground-truth authority. The task score is

$$P(t) = \frac{\sum_{g \in G_t} \alpha_g \sum_{\ell \in L_g} w_{g\ell} s_{g\ell}}{\sum_{g \in G_t} \alpha_g}, \qquad \sum_{\ell \in L_g} w_{g\ell} = 1.$$

The released v0.84 workbook uses this rule consistently in its `README`, `Scoring Protocol`, and `Rubric Leaves` sheets. The 22 `average_impact_preferences` are described in the paper as **secondary preferences** and are excluded from the primary score. The column name `full_hidden_persona_for_user_simulator` denotes a simulator-only latent answer bank that also contains task facts and preference responses; it is not a complete persona given to the report agent.

All report-level results in this paper use the frozen $P_{\text{human-semantic}}$ v0.84 rubric and the same executable hierarchy. The historical fixed-eight-unit acquisition measure is retained only for compatible traces and is not reported as v0.84-axis coverage.

## 4 Experiments

We ask: (RQ1) Do persona-derived criteria recover task-owner preferences under a matched criterion budget? (RQ2) How does visible profile context affect final personalization? (RQ3) What is the effect of clarification under two harnesses and multiple backbones? (RQ4) Do questions target still-hidden consequential preferences? (RQ5) Can persona-derived and human-grounded rubrics produce different conclusions about the same reports?

The full-factorial benchmark contains 360 scored reports across two harnesses. Gemini 3.1 Pro with DeerFlow 2.0 forms a complete $15 \times 3 \times 2 = 90$ matrix over context and native Ask/Off. Open Deep Research forms a complete $15 \times 3 \times 2 \times 3 = 270$ matrix over task, context, IEO/No-Ask, and three report backbones: GPT-5.6-Sol, Claude Sonnet 5, and Gemini 3.1 Pro. A secondary OpenAI–DeerFlow matrix contributes 65 additional scored reports but remains incomplete and is interpreted only through matched pairs. Thus the primary full-factorial scope covers two harnesses and four harness–backbone configurations; the extended scored inventory contains 425 reports.

Within a matched comparison, task input, context mask, research budget, and tool access are fixed. Clarification is capped at four atomic questions in one non-follow-up turn, four disclosed state items, 512 simulator tokens, zero episode retries, and 2,400 seconds. Downstream ODR conditions share two research units, two iterations, at most six ReAct calls per researcher, and 7k/6k/9k research/compression/report token limits. Every scored report is evaluated under v0.84 by three blinded model families (`gpt-5.6-terra`, `gemini-3.6-flash`, and `claude-haiku-4-5-20251001`); the primary leaf value is their median.

An episode entered the scoring set only when it produced a complete, non-error final report under the frozen execution contract. We excluded provider quota or rate-limit termination, billing or access errors, hard timeouts before report completion, missing or empty report artifacts, and violations of the preregistered clarification budget such as exceeding the maximum number of atomic questions. Excluded episodes remain in the execution records and are never converted to zero scores. Every reported policy comparison states the number of complete matched pairs and holds backbone, task, and context fixed.

We separately validated the automated report judge on 45 reports spanning 15 tasks and three systems (approximately 450 item-level decisions). One annotator labeled the full set; a second independently and blindly labeled a stratified 25–30% overlap covering every domain, model, har-

ness, and rubric type. Model and system identities were hidden, the v0.84 rubric was frozen before annotation, and the prompt-development set was disjoint from the final agreement set. Human–human agreement was computed before discussion; disagreements were resolved through discussion or third-person adjudication to form consensus labels. The two humans achieved Krippendorff's $\alpha = 0.84$. Against consensus, the automated judge achieved 88.1% exact agreement, macro-F1 0.85, and Cohen's $\kappa = 0.78$. At report level, judge-derived and human $P$ scores correlated at Spearman's $\rho = 0.89$, with MAE 4.2 points on the 0–100 scale ($P_{100} = 10P$).

Historical DeerFlow intervals use 10,000 task-level resamples with seed 20260923. For ODR, we use 20,000 with-replacement resamples over five prespecified task clusters, seed 20260926, and 2.5/97.5 percentiles; these intervals are low precision. Ask–Off estimates first form paired differences, and leaves or judge passes are never treated as independent samples. RAW50 masking separately uses seed 20260924.

IEO generates candidate clarification axes through decision, counterfactual-flip, ontology-coverage, and human-fit lenses; scores influence, evidence, ownership, residuality, wrong-default cost, and burden; and selects at most four atomic questions. It is an evaluation scaffold, not a claimed algorithmic contribution. The complete ODR matrix holds the harness, task, context, backbone, and downstream research configuration fixed within each IEO–No-Ask pair.

## 5 Results

### 5.1 Construct validity

Figure 2 establishes the core mismatch. The unconstrained GPT-5 list covers many human axes, but it does so with 5.19 times the human reference size. Once both lists contain 73 items, strict recall falls below 40%. The result does not imply that every unmatched generated criterion is objectively false. It shows that breadth is insufficient evidence of correctly prioritizing the task owner's compact set.

### 5.2 Context improves responsiveness, then saturates

In the complete Gemini–DeerFlow matrix, context has a large effect. With clarification disabled, RAW50 improves v0.84 $P$ over cold start by +1.56 points (95% task-bootstrap interval [1.02, 2.16]) and RAW100 by +1.44 [0.84, 2.05]. With clarification enabled, gains are +1.36 [0.70, 2.07] and +1.35 [0.69, 2.08]. RAW100 does not improve over RAW50. The model uses profile evidence, but a random half of facts already saturates this benchmark slice; RAW50 is not a controlled half of task-relevant preference mass.

### 5.3 Clarification access is not a calibrated policy

| Harness | Backbone / policy | Cold $\Delta P$ | RAW50 $\Delta P$ | RAW100 $\Delta P$ | Clarification used |
|---|---|---|---|---|---|
| DeerFlow 2.0 | Gemini 3.1 Pro / native Ask | +0.22 | +0.01 | +0.12 | 9/45 |
| DeerFlow 2.0 | GPT-5.6-Sol / native Ask | +0.80 | +0.33 | -0.72 | 33/33 observed Ask episodes |
| Open Deep Research | GPT-5.6-Sol / IEO | **+0.78** | **+0.31** | **-0.68** | **45/45** |
| Open Deep Research | Claude Sonnet 5 / IEO | **+0.71** | **+0.27** | **-0.74** | **44/45** |
| Open Deep Research | Gemini 3.1 Pro / IEO | **+0.74** | **+0.34** | **-0.63** | **43/45** |

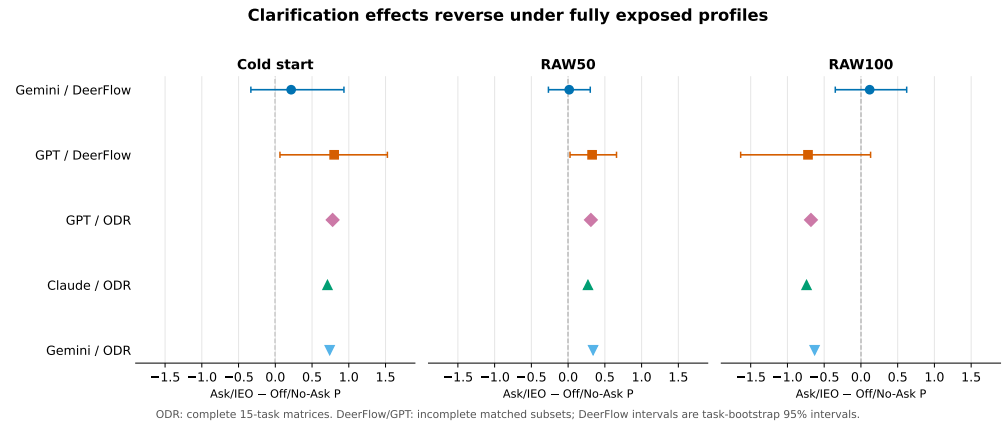


Figure 3: Clarification effects vary with context. Points are mean within-task Ask/IEO–Off/No-Ask differences. The three ODR backbones form complete 15-task matrices and reproduce the positive-cold/partial, negative-rich transition; the DeerFlow/GPT series uses incomplete matched subsets.

**Table 1:** *Within-backbone clarification contrasts under two harnesses.* $\Delta P$ is Ask/IEO minus Off/No-Ask on the 0–10 v0.84 scale. Each ODR context estimate uses 15 matched tasks (45 matched pairs per backbone across contexts); DeerFlow/Gemini also uses 15 per context. DeerFlow/GPT uses 9, 8, and 9 matched tasks and is directional. "Clarification used' ' counts Ask-enabled episodes that issued at least one question across the three contexts.

Gemini's native DeerFlow policy rarely asks, so its near-zero effects estimate enabling the mechanism rather than consistently receiving new information. In contrast, ODR/IEO asks in 132 of 135 enabled episodes. All three ODR backbones show the same transition: positive effects at cold start and RAW50, followed by negative effects at RAW100. This full-matrix replication matches the directional DeerFlow/GPT pattern while removing its missing-cell limitation. The correct conclusion is not that rich context makes questions intrinsically harmful; it is that a clarification mechanism can still choose low-value or misdirected questions when a rich profile creates unwarranted confidence about what remains unresolved.

Stable resolved-unit identifiers are absent from stock DeerFlow traces. ODR preserves question and resolution records, but the historical fixed-eight-unit acquisition metric is not coverage of the final 6–9 v0.84 axes. A confirmatory question audit must instead report precision, recall over still-hidden askable axes, and downstream reflection. Until that audit is complete, question-content mechanisms are described only as consistent with the observations.

### 5.4 Persona-derived rubrics can reverse the measured Ask effect

We rescored five frozen Ask–Off pairs (10 reports) with both the v0.84 human-grounded rubric and a persona-derived rubric. The mean Ask effect was -0.2508 under the human rubric but +0.2924 under the persona rubric, a +0.5432 shift. At a prespecified $\pm 0.10$ win/tie/loss threshold, two of five pairs changed conclusion. The T01/cold model ranking remained GPT $>$ Claude $>$ Gemini under both rubrics. This small diagnostic is descriptive, not a population estimate or significance test; it illustrates that rubric provenance can change the conclusion drawn from identical reports but is not used as a primary system-effect estimate.

## 6 Discussion

The results separate information availability from calibration. Rich context helps the report model use known information, yet it can also lower the apparent need to ask and introduce irrelevant or conflicting cues. These mechanisms can coexist: responsiveness may improve while acquisition of unresolved preferences worsens. The ODR replication shows that this transition is not confined to one report backbone. A true oracle would reveal $Z_{u,q}$ directly; RAW100 reveals only $H_u$.

Final-score improvements also do not identify why a system improved. Asking changes the prompt, available information, token use, and sometimes research trajectory. The correct unit for causal interpretation is a matched policy comparison with equal computational opportunities, not an unmatched cell mean. The present study supports diagnosis and benchmark design, not a general causal claim that clarification beats memory or that richer profiles are harmful.

The final v0.84 rubric reduces one construct problem by directly representing participant-endorsed axes rather than imposing universal dimensions. It does not eliminate evaluator error. Three model judges provide family sensitivity, and the held-out human calibration supports their aggregate use; however, many judge evidence strings are paraphrases rather than verified exact quotations. The human–LLM rubric-matching labels in Figure 2 are a separate semantic-matching analysis and remain model-mediated.

## 7 Related work

PDR-Bench introduced broad evaluation of personalized deep-research reports~(Liang et al., 2026). AskInfer-Bench inherits report-level evaluation but questions whether task-and-profile-derived criteria establish current-task intent. MyScholarQA similarly shows that model evaluation can miss what real researchers value~(Balepur et al., 2026a). Re-Centering Humans and RubricBench document human–model gaps in personalization and rubric construction~(Zhang et al., 2026a;b); our focus is the profile-to-current-task calibration problem under controlled ask/infer conditions.

PrefDisco establishes proactive preference discovery in cold-start interaction~(Li et al., 2026), while IntentRL and G-STEER study clarification policies and question utility~(Luo et al., 2026; Yoon & Lee, 2026). We do not claim to be first to ask before researching or to trade information value against user burden. Instead, we make independently elicited task-owner preferences the evaluation target and manipulate persistent context separately from clarification access.

DRACULA collects action preferences from real researchers and finds that executing specified actions is easier than predicting which actions a user wants; user choices vary with unstated goals, and selection history helps prediction~(Balepur et al., 2026b). Its target is intermediate action-preference prediction. AskInfer-Bench targets calibration between persistent profile evidence and current-task preference ground truth, evaluated through context masking and ask/infer behavior. PARL learns individualized rubrics from raw histories~(Qiu et al., 2026). This is complementary: AskInfer asks when a history-derived preference is licensed to enter the current-task reference at all.

UserVille studies personalization with configurable user simulators~(Sun et al., 2025), and persona work shows that different cues for ostensibly the same identity can change measured personalization~(Weeber et al., 2026). AskInfer freezes the simulator as a disclosure channel, anchors preference units before evaluation, and treats simulator fidelity as an empirical validity question.

## 8 Limitations

The benchmark contains 15 human task-owner records collected in two batches under the same ground-truth authority rule; the detailed Prolific recruitment and elicitation audit covers ten of them, so broad population claims remain premature. The 65-report DeerFlow/GPT extension is incomplete, although the 360-report two-harness core is fully crossed. RAW50 masks random profile facts rather than calibrated relevant information. The 35 retained production simulator traces are an availability-defined subset of 139 observed ODR simulator interactions and have not undergone the planned two-annotator blinded audit; simulated users also cannot capture natural revision, misunderstanding, or refusal. The held-out human calibration covers 45 reports rather than every report, Figure 2 matching remains model-mediated, and the five-pair dual-scoring diagnostic is not a population estimate. All generalizations should therefore be restricted to the observed tasks and systems.

## 9 Conclusion

Personalized deep research is not only the ability to use information about a user. It is the ability to recognize which consequential preferences remain uncertain, obtain them when appropriate,

and avoid turning plausible profile inferences into unquestioned facts. AskInfer-Bench separates these stages with task-owner-grounded preferences, context masks, controlled disclosure, acquisition traces, and final-report scoring. Across two harnesses, richer context improves responsiveness but clarification gains reverse under fully exposed profiles; persona-derived scoring can even change the sign of the measured Ask effect. The main lesson is methodological: a personalization benchmark should test whether an agent knows what it still needs to learn.

### ETHICS STATEMENT

All ten analyzed participants completed the study's four consent and privacy confirmations, and compensation was disclosed before participation. Released data remove direct and linkable identifiers, minimize raw free text, keep full latent states restricted, and document a separate data license; the Apache-2.0 license of reused PDR-Bench code does not license participant data. High-stakes health, finance, and legal outputs require domain review; user preference cannot override factual correctness or safety.

### REPRODUCIBILITY STATEMENT

The appendix specifies task provenance, the v0.84 rubric inventory, simulator and judge configurations, the 139-interaction count and 35-trace retained subset, context masks, system commits, fixed research budgets, usage-estimated costs, bootstrap seeds, per-task analyses, prompts, qualitative traces, human calibration, and release checks. The anonymous artifact includes machine-readable hashes, frozen prompts, score tables, and analysis scripts where participant consent and provider terms permit.

### AI USE STATEMENT

We used generative AI tools, including OpenAI Codex, for literature discovery and citation verification; outlining, drafting, revision, and language editing; methodological and experimental-design feedback; organization and consistency checking of experimental records; planning of tables and figures; and interpretation and presentation of empirical results. The authors inspected the underlying records, checked cited claims against primary sources, reviewed and revised AI-assisted text, and made the final study and reporting decisions. Generative AI did not replace participant consent, human preference elicitation, or researcher adjudication and is not an author. Evaluated models, the simulator, and automated judges are separately documented as study components. The authors take responsibility for the final text, claims, analyses, and artifacts.

## REFERENCES

Nishant Balepur, Malachi Hamada, Varsha Kishore, Sergey Feldman, Amanpreet Singh, Pao Siangliulue, Joseph Chee Chang, Eunsol Choi, Jordan Lee Boyd-Graber, and Aakanksha Naik. Language models don't know what you want: Evaluating personalization in deep research needs real users. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026a. URL https://aclanthology.org/2026.acl-long.723/.

Nishant Balepur, Malachi Hamada, Varsha Kishore, Sergey Feldman, Amanpreet Singh, Pao Siangliulue, Joseph Chee Chang, Rachel Rudinger, Eunsol Choi, Jordan Lee Boyd-Graber, Doug Downey, and Aakanksha Naik. DRACULA: Hunting for the actions users want deep research agents to execute. *arXiv preprint arXiv:2604.23815*, 2026b. URL https://arxiv.org/abs/2604.23815.

Shuyue Stella Li, Avinandan Bose, Faeze Brahman, Simon Shaolei Du, Pang Wei Koh, Maryam Fazel, and Yulia Tsvetkov. PrefDisco: Benchmarking proactive personalized reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O1hfVE0UxG.

Yuan Liang, Jiaxian Li, Yuqing Wang, Piaohong Wang, Motong Tian, Pai Liu, Shuofei Qiao, Runnan Fang, He Zhu, Ge Zhang, Minghao Liu, Yuchen Eleanor Jiang, Ningyu Zhang, and Wangchunshu Zhou. Towards personalized deep research: Benchmarks and evaluations. In *International*

*Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=51LIRzF53v.

Haohao Luo, Zexi Li, Yuexiang Xie, Wenhao Zhang, Yaliang Li, and Ying Shen. IntentRL: Training proactive user-intent agents for open-ended deep research via reinforcement learning. *arXiv preprint arXiv:2602.03468*, 2026. URL https://arxiv.org/abs/2602.03468.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yuxin Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Yoko Yamakata, and Tat-Seng Chua. Preference-aware rubric learning for personalized evaluation. *arXiv preprint arXiv:2605.31545*, 2026. URL https://arxiv.org/abs/2605.31545.

Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. Training proactive and personalized LLM agents. *arXiv preprint arXiv:2511.02208*, 2025. URL https://arxiv.org/abs/2511.02208.

Franziska Weeber, Vera Neplenbroek, Jan Batzner, and Sebastian Padó. One persona, many cues, different results: How sociodemographic cues impact LLM personalization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44892–44921, 2026. doi: 10.18653/v1/2026.acl-long.2079. URL https://aclantology.org/2026.acl-long.2079/.

Soojin Yoon and Dongha Lee. Personalized deep research query refinement with graph-scaffolded evidence grounding. *arXiv preprint arXiv:2608.05876*, 2026. URL https://arxiv.org/abs/2608.05876.

Lechen Zhang, Jiarui Liu, and Tal August. Re-centering humans in LLM personalization. *arXiv preprint arXiv:2606.06614*, 2026a. URL https://arxiv.org/abs/2606.06614.

Qiyuan Zhang, Junyi Zhou, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, and Chen Ma. RubricBench: Aligning model-generated rubrics with human standards. *arXiv preprint arXiv:2603.01562*, 2026b. URL https://arxiv.org/abs/2603.01562.

# SUPPLEMENTARY MATERIAL

## ASKINFER-BENCH: SUPPLEMENTARY MATERIAL

**Anonymous submission draft, v0.1.** This supplement distinguishes completed evidence from planned validation. Open items are labeled explicitly and must not be replaced by assumptions.

### A. TERMINOLOGY AND PROVENANCE

| Term | Operational meaning in this paper | Not interchangeable with |
|---|---|---|
| User context, profile, or persona, $H_u$ | Persistent pre-task evidence about a user, including profile facts or authorized history | The user's current-task preference state |
| Current-task preference state, $Z_{u,q}$ | User-owned goals, constraints, priorities, and trade-offs that can alter the report for task $q$ | General biography; externally verifiable facts |
| Raw preference statement | One participant-entered item in Part A or Part B before atomicization | One scoring unit |
| High-impact preference axis | A consolidated, task-conditioned preference supported by importance, ranking, failure consequence, or repeated confirmation | A rubric leaf |
| Secondary preference | A retained preference not assigned to the primary high-impact set | "Irrelevant" information |
| Scoring unit | One frozen preference axis in $P_{\text{human}-\text{semantic}}$ v0.84 | A semantic leaf |
| Rubric leaf | A directly entailed, observable component of one preference axis | An independent statistical sample |
| Persona-derived criterion | A criterion generated from the task and profile without access to participant preference answers | Human ground truth |
| Clarification availability | Permission for the harness to invoke its native question mechanism | Receipt of an informative answer |
| Acquired unit | A previously unexposed GT unit resolved through the question-answer interaction | A unit merely present in the initial profile |

The benchmark contains 15 human task-owner records under one common experimental design, purpose, and ground-truth authority rule. All 98 reference axes are human-provided rather than inferred from personas by an LLM. Ten records come from the audited Prolific batch, which comprises 107 raw preference statements, 73 consolidated high-impact axes, and 22 secondary preferences. Five records come from the earlier human-user collection and contribute 25 additional human-provided axes. The primary full-factorial system inventory contains 360 scored reports over all 15 task identifiers and two harnesses; an incomplete 65-report DeerFlow/GPT extension yields 425 scored reports in the extended inventory.

The corrected frozen 15-task workbook is the authoritative experiment-input and scoring-schema artifact (SHA-256 `05e085e590dae9b40f357042fa113acfe63607c825a9d2eb9beaf46acb899b1c`). Its four sheets serve different roles:

| Sheet | Role |
|---|---|
| `15 Tasks` | One row per task containing episode inputs, hidden simulator state, GT axes, axis rubrics, acceptable clarification examples, secondary preferences, and provenance-preserving originals |
| `README` | Field semantics, provenance, human-elicitation sequence, deduplication rules, and known task-specific quality notes |
| `Rubric Leaves` | Executable judging table with one row per semantic leaf, explicit weights, anchors, and evidence requirement |
| `Scoring Protocol` | Normative metric definition, inventory, task scope, version boundary, and statistical unit |

The released `README`, numeric `Rubric Leaves` table, and normative `Scoring Protocol` are now aligned. The executable protocol scores semantic leaves, averages them equally within each preference, and then applies the frozen 3/2/1/1 preference weights. `Unranked` receives weight 1 and denotes unavailable priority metadata, not weaker ground-truth authority. No post hoc switch between estimators is permitted.

### A.1 Direct human-axis representation in v0.84

The final $P_{\text{human}-\text{semantic}}$ v0.84 asset removes the earlier 73-to-eight selection problem. Across the common 15-task benchmark it contains 98 human-grounded preference axes and 409 semantic leaves. The audited Prolific batch contributes 73 axes and 309 directly entailed leaves; the earlier human-user batch contributes 25 axes and 100 leaves. The release ledger should contain:

| Field | Definition |
|---|---|
| `task_id` | One of the 15 frozen task identifiers |
| `human_axis_id` | Stable identifier for each of the 98 human-grounded axes |
| `source_span_hash` | Hash of the participant span; raw text remains restricted if identifying |
| `impact_evidence` | Importance/ranking evidence where collected; otherwise explicitly unranked |
| `preference_id` | Stable v0.84 preference identifier |
| `leaf_ids` | Directly entailed observable semantic leaves |
| `priority` | Must-have, Strong, Nice-to-have, or unranked |
| `leaf_weight_sum` | Must equal one within each preference |
| `adjudicator` | Coded anonymous reviewer identifier |

The workbook audit confirms that every human-grounded axis has a v0.84 preference ID and that the leaf weights sum to one within every preference. A subsequent two-annotator human audit evaluated the rubric annotation layer, including whether model-drafted leaves were directly supported by their source preference and whether priority labels were appropriate. The audit achieved 89.0% raw agreement, Cohen's $\kappa = 0.78$, and priority-weighted $\kappa = 0.84$. Human review changed 14–16% of initially model-drafted leaves, and 9–11% of annotation units required adjudication. Because the available audit summary reports the latter quantities as ranges, the release must preserve the exact denominators, both unreconciled annotation streams, and the adjudication log rather than reverse-engineering point counts from rounded percentages.

### A.2 Formal persona and information boundaries

For participant-grounded records, `full_hidden_persona_for_user_simulator` is the task-conditioned anonymized latent answer bank, not the report agent's input and not synonymous

with persistent persona $H_u$. It contains persistent profile sections, task facts, routing and authority rules, recent context, verbatim spontaneous Part A and prompted Part B answers, the ranked top three, the consequence of missing the top preference, and the participant's rich-history must-confirm answers. The report agent receives only the task instruction and the profile facts authorized by its context condition. The following fields are always withheld from the report agent: the remaining latent-state values, GT IDs, importance tiers, preference weights, evaluator rubrics, semantic leaves, acceptable clarification examples, and evaluation procedure.

The 22 `average_impact_preferences` are participant-stated secondary directions from the audited Prolific batch. They are retained for provenance and auxiliary analysis but are not part of its primary 73-axis construct-validity reference.

The workbook contains 409 complete leaf rows and 409 unique leaf IDs across 98 human-grounded preferences. The audited Prolific batch accounts for 309 leaves across 73 preferences: 40 Must-have, 32 Strong, and one Nice-to-have. The earlier human-user batch accounts for 100 leaves across 25 unranked preferences. Every required leaf field is populated, every preference's leaf weights sum to one, and preference weights are exactly 3, 2, 1, and 1 for Must-have, Strong, Nice-to-have, and unranked preferences, respectively. Unranked is a priority-metadata status only; it does not alter the human authority of the source preference.

## B. Formal observations

These propositions clarify the evaluation problem. They are not presented as an algorithmic-theory contribution.

### Proposition 1: profile-only intent is not identifiable in general

Let $H$ be persistent user context, $q$ a current task, and $Z$ the task-specific preference state. Suppose there exist two states $z \neq z'$ that are both compatible with the same observed pair $(H, q)$ and for which the ideal task output differs: $R^*(q, z) \neq R^*(q, z')$. Then no deterministic profile-only procedure $f(H, q)$ can equal the ideal output for both states.

**Proof.** Because the procedure observes the same $(H, q)$ in both cases, it returns the same output $r = f(H, q)$. Since the two ideal outputs differ, $r$ cannot equal both $R^*(q, z)$ and $R^*(q, z')$. Thus the procedure must be wrong in at least one compatible world. □

The proposition does not imply that an agent must always ask. It implies that an evaluation should preserve uncertainty when the observed history does not identify the current-task state. A calibrated agent may ask, branch, state an assumption, or defer a preference-sensitive decision.

### Proposition 2: final-report scoring cannot identify acquisition behavior

Let $S(R, Z)$ be any evaluator that observes only a final report $R$ and reference state $Z$. There can exist two policies, $\pi_{\text{ask}}$ and $\pi_{\text{guess}}$, that produce the same report $R$ for a case, where the first obtains a missing preference through interaction and the second guesses it without evidence. Then $S$ assigns the same score to both policies and cannot identify whether preference acquisition occurred.

**Proof.** By construction, the evaluator receives identical arguments $(R, Z)$ under both policies. A function of those arguments therefore returns the same value. The distinct acquisition histories are outside its observation. □

This is why AskInfer-Bench records `known`, `candidate`, `selected`, `resolved`, and `reflected` separately from final-report quality.

### Proposition 3: unconstrained rubric coverage is monotone under overgeneration

For a human reference set $G$ and generated criterion set $C$, define strict coverage recall as

$$\text{Rec}(G, C) = \frac{1}{|G|} \sum_{g \in G} \mathbb{1}[\exists c \in C : \text{match}(g, c) = 1].$$

If $C \subseteq C'$, then $\mathrm{Rec}(G, C') \geq \mathrm{Rec}(G, C)$.

**Proof.** Every match available in $C$ remains available in its superset $C'$. Adding criteria may convert an unmatched reference item to matched but cannot convert a matched item to unmatched. Summing the indicators preserves the inequality. □

Consequently, high recall from an unconstrained criterion list is insufficient evidence of prioritization. The equal-budget analysis fixes $|C| = |G| = 73$ using the generator's own hierarchical weights.

PROPOSITION 4: NEW-ACQUISITION COVERAGE IS NOT TOTAL-KNOWLEDGE COVERAGE

Let $G$ contain $K$ preference units, $E_r \subseteq G$ be units exposed initially under context level $r$, and $A_r \subseteq G \setminus E_r$ be units newly acquired through dialogue. AskInfer reports $|A_r|/K$. Even if total known units $|E_r \cup A_r|$ increases with richer context, $|A_r|/K$ may decrease because the eligible residual set $G \setminus E_r$ shrinks.

**Proof.** Take $G = \{1, 2\}$. Under cold start, let $E_0 = \varnothing$ and $A_0 = \{1\}$, giving acquisition coverage $1/2$ and total knowledge $1/2$. Under rich context, let $E_1 = G$ and $A_1 = \varnothing$, giving acquisition coverage $0$ and total knowledge $1$. □

Therefore the decline in the historical $\mathrm{LegacyAcqCoverage@GT}(8)$ measure across context levels is partly mechanical. This legacy measure uses the earlier fixed eight-unit ledger and must not be interpreted as coverage of the final v0.84 axes. The decisive v0.84 process metric is recall over the *remaining consequential askable axes*, accompanied by question precision and downstream reflection.

### B.1 COMPUTATIONAL COMPLEXITY

For IEO, four lenses produce at most $K$ candidates each, so $M \leq 4K$. Naive pairwise semantic canonicalization is $O(M^2)$; scalar scoring is $O(M)$; and greedy selection of at most $B$ questions with diversity checks is $O(BM)$. This excludes LLM inference latency and token cost. If report $n$ is judged on $L_n$ semantic leaves, report scoring requires $O(\sum_n L_n)$ leaf decisions. The historical DeerFlow 10,000-resample bootstrap is $O(10^4 T)$ with $T$ tasks; the ODR analysis uses 20,000 resamples over five prespecified task clusters, or $O(2 \times 10^4 C)$ with $C = 5$.

## C. HUMAN PREFERENCE COLLECTION AND CURATION

### C.1 RECRUITMENT AND COMPENSATION

The primary Prolific screening wave opened 25 places. Twenty-three screeners were completed and approved, 20 respondents met the prespecified task-matching criteria, and targeted backfill and replacement launches were subsequently used for unfilled, returned, or quality-flagged task slots. Ten distinct adults formed the final analytic set, with each person contributing exactly one matched task. Compensation was US$0.40 for screening and US$9 for the main task. Self-reported completion time averaged 73.8 minutes (median 72.5, range 35–120). All ten final participants completed all four consent and privacy confirmations.

The supplied Prolific study-configuration export contains 22 launch records: one primary screener, ten initial task-owner studies, one paid backfill screener, and ten task-specific replacement or backfill launches. It verifies advertised rewards, configured estimated times, one-person allowlists for task-owner studies, and the existence of replacement waves. The main studies were configured with a 60-minute estimated duration and a platform maximum completion window of approximately 127 minutes. The 127-minute value is a timeout allowance, not an expected or observed study duration, and therefore does not conflict with the configured estimate or elapsed-time summaries. The export does not contain submission status histories, approval decisions, participant-level elapsed time, bonuses or adjustment payments, or survey completion/consent records. Study IDs, participant allowlist IDs, and external survey URLs are direct identifiers or linkable metadata and will not be released.

The private recruitment ledger retains launch-level configuration and payment records. The public release reports only de-identified aggregate counts and does not expose study IDs, allowlist identifiers, or survey URLs.

| Task | Raw Part A+B statements | Consolidated high-impact axes |
|---|---|---|
| T01 | 10 | 7 |
| T02 | 10 | 7 |
| T03 | 10 | 7 |
| T04 | 10 | 6 |
| T05 | 12 | 8 |
| T06 | 12 | 8 |
| T07 | 11 | 6 |
| T08 | 13 | 8 |
| T09 | 10 | 7 |
| T10 | 9 | 9 |
| **Total** | **107** | **73** |

All ten participants completed four required confirmations: adulthood and voluntary consent; permission for academic analysis and de-identified derivative use; completion without generative-AI assistance; and an instruction not to enter identifying information.

### C.2 Elicitation order

Participants first supplied a de-identified general profile. In Part A, before seeing any task-specific reminder dimensions, they described at least five distinct conditions that would make an AI research report useful, including preferred direction, rationale, and importance. They then provided neutral task facts and information-routing constraints. In Part B, they saw five value-neutral reminder dimensions and could state, revise, or decline a preference. Finally, they selected the three most consequential preferences, described the consequence of violating the highest priority, and identified what a well-informed agent should still confirm. Part A was frozen before Part B to reduce anchoring.

### C.3 Curation pseudocode

```
Algorithm A1: Human preference ledger construction
Input: participant responses for one matched task
Output: provenance-preserving preference ledger
1: Verify consent, task match, completeness, and privacy checks.
2: Extract each non-empty Part A and Part B statement with its source
↪  span.
3: Split compound statements only when each atom has an independently
   interpretable preferred direction or consequence.
4: Label each atom as user-owned preference, task fact, external fact,
   uncertainty, privacy constraint, or out of scope.
5: Merge semantic duplicates only within the same participant; retain
↪  all spans.
6: Assign impact evidence from importance, A/B Top-3 membership,
   violation consequence, and repeated confirmation.
7: Preserve contradictions, temporal qualifiers, and uncertainty; do not
↪  infer
   a resolution that the participant did not provide.
8: Researcher adjudicates the structured model-assisted proposals.
9: Freeze stable IDs, provenance, and a de-identified release
↪  representation.
```

The initial construction process was model-assisted structured screening followed by final researcher adjudication; it should not be retrospectively described as two-independent-human coding. The subsequent human audit was a separate validation layer and retained independent first-pass judgments before adjudication.

### C.4 Human annotation agreement

We report five complementary audit statistics because no single agreement coefficient captures both reproducibility and the practical effect of human review:

| Statistic | Result | Interpretation |
|---|---|---|
| Raw agreement | 89.0% | Fraction of annotation units on which the two annotators gave the same first-pass judgment |
| Cohen's $\kappa$ | 0.78 | Chance-corrected agreement for categorical annotation decisions |
| Priority-weighted $\kappa$ | 0.84 | Agreement on ordered priority labels, penalizing a High–Low disagreement more than a High–Medium disagreement |
| Changed leaves | 14–16% | Fraction of initially model-drafted semantic leaves modified through human review |
| Adjudication rate | 9–11% | Fraction of annotation units that the two annotators did not resolve without third-party adjudication |

The high agreement coefficients do not make model drafting self-validating. The changed-leaf rate shows that human review materially altered the executable scoring asset, while the adjudication rate identifies the genuinely ambiguous subset. The changed-leaf and adjudication ranges are reported exactly as supplied by the audit summary. The archival release should add their integer numerators and denominators and state whether the endpoints arise from different annotation scopes or denominator conventions.

## D. Human–LLM rubric comparison

### D.1 Generation and blinding

The official PDR-Bench English criterion-generation prompts and code were reproduced. The requested `gpt-5` alias resolved to `gpt-5-2025-08-07`. For each of the ten human-elicitation tasks, the generator saw the research task and a preference-blind rich persona. It did not receive participant statements, rankings, importance labels, consolidated GT, prior rubric values, or acceptable clarification paraphrases. Dimension weights were sampled five times; criteria were then generated for goal alignment, content alignment, presentation fit, and actionability/practicality.

### D.2 Matching and equal-budget procedure

```
Algorithm A2: Conservative persona-rubric comparison
Input: human axes G_t, generated criteria C_t, generator weights
1: For every human axis, obtain full/partial/none labels from each judge.
2: Reconcile in favor of the generated rubric:
   full if either judge says full; else partial if either says partial;
   ↪ else none.
3: Compute strict recall and partial-credit recall on all criteria.
4: For every generated criterion, label whether any human axis supports
↪ it.
5: Compute hierarchical priority = mean dimension weight Œ criterion
↪ weight.
```

```
6: Pool criteria by task, rank by hierarchical priority, and retain 73
↪  total
   under the frozen equal-budget rule.
7: Repeat coverage and support calculations on the retained criteria.
8: Resample tasks, not items, 10,000 times for percentile intervals.
```

The full 379-criterion list is 5.19 times the 73-axis human reference. Under the LLM-favorable union, 59 human axes were fully covered, 13 partially covered, and one missed.

| Analysis | Strict recall | Partial-credit recall | Generated criteria unsupported by elicited GT |
|---|---|---|---|
| Full 379 criteria, favorable union | 80.8% [68.5, 91.1] | 89.7% [83.1, 95.4] | at least 25.1% [17.9, 33.2] |
| Top 73 by generator weights | 39.7% [31.5, 46.8] | 51.4% [44.2, 57.9] | Not yet computed |
| Paired change | -41.1 pp [-49.3, -31.8] | -38.4 pp [-44.7, -31.9] | – |

Judge sensitivity was substantial. GPT-5 produced 41.1% strict recall, 67.8% partial-credit recall, and 47.2% unsupported generated criteria; GPT-5-mini produced 80.8%, 89.7%, and 26.9%, respectively. Raw agreement was 56.2% with $\kappa = 0.234$ for human-item summaries and 60.4% with $\kappa = 0.389$ for generated-item support labels. The favorable union is conservative with respect to the mismatch claim, but it is not human adjudication.

**Open semantic-audit item before submission:** have a second human independently review all 73 human-axis coverage decisions and a prespecified stratified sample, or preferably all 379, generated-criterion support decisions. Freeze sampling strata before inspection and report agreement before adjudication.

## E. Metrics and statistical analysis

### E.1 Final-report personalization

The final $P_{\text{human}-\text{semantic}}$ v0.84 representation contains a variable number of directly entailed semantic leaves for each preference. Leaves receive integer scores from 0 to 10. A score of 0 denotes absence, contradiction, or pursuit of an incompatible direction; 5 denotes partial or implicit satisfaction with material incompleteness; and 10 denotes explicit, complete, consistent satisfaction with a supporting report quotation. For preference $g$ and task $t$,

$$S_g = \sum_{\ell \in L_g} w_{g\ell} s_{g\ell}, \quad \sum_{\ell \in L_g} w_{g\ell} = 1, \qquad P_{\text{human}-\text{semantic}}(t) = \frac{\sum_{g \in G_t} \alpha_g S_g}{\sum_{g \in G_t} \alpha_g}.$$

Priority weights are 3 for Must-have, 2 for Strong, 1 for Nice-to-have, and 1 for unranked preferences. The result is a 0–10 score, not a percentage and not acquisition coverage. All 425 scored reports in the extended inventory use this frozen v0.84 protocol.

Every report was independently scored by three distinct blinded judge models: OpenAI `gpt-5.6-terra` with low reasoning effort, Google `gemini-3.6-flash` at temperature 0, and Anthropic `claude-haiku-4-5-20251001` at temperature 0. All were accessed on 2026-09-26.

Each judge receives identical task information, report text, rubric leaves, chunk boundaries, and instructions under an anonymous report ID. It does not receive the report-generating system, harness, context condition, study hypothesis, previous score, or another judge's rationale. Leaves are split into three preference-preserving chunks per report; all leaves belonging to one preference stay together. The primary leaf score is the median across the three judges. Preference and task aggregation is applied only after this leaf-level combination.

There were no automatic retries. The frozen policy allowed at most one manual format repair or one transport retry. One malformed Gemini JSON response was retried once successfully; an initial schema-preflight rejection occurred before judgment and was corrected. In the archived calibration/audit request bundle, each judge completed 243 validated jobs: 228 personalization-chunk jobs and 15 quality jobs. The full ODR extension used the same three-judge specification and is preserved in a separate execution ledger. Raw leaf scores, cited evidence strings, rationales, failure codes, latency, and token counts are retained. A substantively valid but disagreeing judgment is never replaced.

Judge reliability is summarized by ordinal Krippendorff's $\alpha$ over leaf scores, ICC(2,$k$) over report-level judge scores, the median pairwise absolute leaf-score difference, and the fraction of leaves whose maximum minus minimum judge score is at least four points. Reliability is reported overall and by rubric priority, task, and system family when sample size permits. The task-level bootstrap is run after the three-judge median and is not presented as an interval over judge or generation randomness.

#### E.1.1 Held-out human calibration of the automated judge

The agreement study sampled 45 reports spanning all 15 tasks and three systems, yielding approximately 450 item-level judgments. The primary human annotator labeled the complete set. A second annotator independently and blindly labeled a stratified 25–30% overlap covering every domain, evaluated model, harness, and rubric type. Report-generator and harness identities were hidden. The v0.84 rubric was frozen before formal annotation, and the examples used to tune the judge prompt were disjoint from the final agreement sample.

We computed human–human agreement before reconciliation. Disagreements were then resolved through discussion or third-person adjudication to form human consensus labels. The two human annotators achieved Krippendorff's $\alpha = 0.84$. Against the adjudicated consensus, the automated judge achieved 88.1% exact agreement, macro-F1 0.85, and Cohen's $\kappa = 0.78$. At report level, judge-derived P scores correlated with human scores at Spearman's $\rho = 0.89$ and had MAE 4.2 points on a 0–100 scale, where $P_{100} = 10P$.

### E.2 Acquisition and question-quality metrics

For compatible historical traces,

$$\mathrm{LegacyAcqCoverage@GT}(8) = \frac{|\{g :\ g \text{ was previously hidden and resolved through Q/A}\}|}{8}.$$

Here $g$ indexes the earlier fixed eight-unit acquisition ledger, not the variable 6–9 high-impact axes in the final v0.84 workbook. No-Ask is zero by construction, and units exposed initially under RAW50 or RAW100 are excluded from the numerator. The measure is retained only for backwards-compatible analysis of traces that stored those unit IDs.

The confirmatory process audit should additionally report, at each context level:

$$\mathrm{QuestionPrecision} = \frac{|\text{selected questions that target an unresolved, askable GT axis}|}{|\text{selected questions}|},$$

$$\mathrm{ResidualRecall} = \frac{|\text{unresolved, askable GT axes targeted}|}{|\text{unresolved, askable GT axes}|},$$

and utilization conditional on resolution,

$$\mathrm{ReflectionRate} = \frac{|\text{resolved GT axes materially reflected in the report}|}{|\text{resolved GT axes}|}.$$

Questions should also be coded as atomic versus bundled, redundant with visible context, externally researchable rather than user-owned, unsupported/stereotyped, privacy-sensitive, or unanswerable from the latent state. These codes distinguish "asked more" from "asked the right thing."

### E.3 Bootstrap procedure

All reported intervals resample at the task or prespecified task-cluster level with replacement and report the 2.5th and 97.5th percentiles of the bootstrap mean. Historical DeerFlow tables use 10,000 task-level resamples with seed 20260923. For ODR, we use 20,000 resamples over five prespecified task clusters with seed 20260926; because there are only five clusters, those intervals are explicitly described as low precision. The clusters are defined by task identifier: T01, T03, T05, T07, and T09, with each cluster containing all matched context-policy-backbone observations for that task. Paired Ask–Off effects first compute a within-unit difference and then resample those differences. The intervals describe heterogeneity over the observed units, not generation variance or provider nondeterminism. Cross-judge variation is characterized separately rather than treating judge passes as independent replicates. RAW50 masking uses seed 20260924 and is unrelated to the bootstrap seeds.

## F. Blinded user simulation

### F.1 Frozen production configuration used in Ask-enabled episodes

| Field | Value |
|---|---|
| Provider | Alibaba Cloud Model Studio |
| Model snapshot | `qwen3.7-max-2026-05-20` |
| Thinking | disabled |
| Temperature | 0.0 |
| Top-p | 1.0 |
| Seed | 20260924 |
| Maximum response | 512 tokens |
| Output | strict JSON Schema |
| Format-repair retries | 0 |
| Transport retries | 0 |

Any format, timeout, or network error is an episode failure. A classifier first maps the question to opaque state keys described only by topic, meaning, aliases, and keywords. It does not see attribute values or priority. The response model then receives only the policy-authorized values for matched keys, the task, the current question, and prior interaction from the same episode. It must answer only the requested information, preserve uncertainty and temporal conflicts, decline private or unavailable fields, and avoid volunteering adjacent preferences or performing the research task.

Run-level asking flags record 132 IEO interactions among 135 enabled ODR episodes. Seven additional native-Ask episodes invoked the same simulator channel, giving 139 observed production interactions. Complete `simulator_trace.json` bundles were retained for 35 of these interactions: 28 IEO-v04 and seven native ODR Ask. This is an availability-defined trace-bearing subset from the five-task execution archive (T01, T03, T05, T07, and T09), not a random sample of the 139 interactions. The 35 retained traces required 70 Qwen calls (35 question-to-state classifications and 35 simulator responses), 229,958 input tokens, and 11,793 output tokens. Archived valid production traces (35 exchanges) account for CNY 4.971342: CNY 3.854548 for IEO and CNY 1.116794 for native Ask. Every retained trace preserves the original question, classifier output, permitted neutral state keys, simulator answer, token usage, and blinding metadata. The configuration above is identical for all 35 retained traces.

The frozen configuration file has SHA-256 `b8f2b2fd7f181b95ff44c995bca0cc663da3905fb6d8649775ed795cbb7bd88f`; the implementation file used to define the classifier/response path has SHA-256 `6aea70101a43b217f0aef3497f35e29a753dbd6e814c7e4131338b01fd05215a`. The anonymous artifact binds the inventory to these hashes.

### F.2 Development probe and validation boundary

Eight synthetic templates cover answerable, absent, uncertain, multi-part, leading, privacy-sensitive, adjacent-leakage, and current-versus-stale-conflict questions. Each was repeated three times. The

resulting 24 simulator answers and 24 classifier outputs required 48 calls. Developer scoring found 24/24 faithful answers (95% Wilson interval 86.2–100), 0/24 unsupported inventions (0–13.8), 0/24 selective-disclosure violations (0–13.8), and 24/24 direct answers.

This probe was approximate: both classifier and response used a 160-token cap and a shorter system prompt, not the frozen 512-token production configuration. Only one preselected model configuration was tested, so the probe is not a multi-model selection experiment. An exact-production-prompt, matched-token-budget unsupported-invention rate was not measured. The 35 retained production traces currently have no independent human audit, so trace-subset faithfulness, unsupported-invention, selective-disclosure, direct-answer, and resolved-state agreement rates are unavailable rather than zero. Formally audited retained traces are $0/35$; the remaining 104 interactions do not have complete archived trace bundles in the supplied artifact and are outside that audit denominator.

The planned confirmatory audit will cover the production exchanges using two independent condition-blinded annotators. Binary fields will report raw agreement, Cohen's $\kappa$, positive agreement, and negative agreement. Resolved-state sets will report exact-set agreement and mean Jaccard similarity. Violation rates will report numerator, denominator, Wilson intervals, and the fraction of episodes with at least one violation.

## G. Systems, budgets, and reproducibility

### G.1 Episode inventory

| Evidence tier | Harness | Backbone | Scored reports | Matrix | Judge |
|---|---|---|---|---|---|
| Primary | DeerFlow 2.0 | Gemini 3.1 Pro preview | 90 | Complete, $15 \times 3 \times 2$ | three-judge median, v0.84 |
| Primary | Open Deep Research | GPT-5.6-Sol | 90 | Complete, $15 \times 3 \times 2$ | three-judge median, v0.84 |
| Primary | Open Deep Research | Claude Sonnet 5 | 90 | Complete, $15 \times 3 \times 2$ | three-judge median, v0.84 |
| Primary | Open Deep Research | Gemini 3.1 Pro preview | 90 | Complete, $15 \times 3 \times 2$ | three-judge median, v0.84 |
| Secondary | DeerFlow 2.0 | GPT-5.6-Sol | 65 | Incomplete | three-judge median, v0.84 |

The primary full-factorial scope is therefore 360 reports across two harnesses. Each ODR backbone contributes 45 IEO and 45 No-Ask reports: 15 tasks in each of COLD, RAW50, and RAW100. The three judges are combined only by the frozen leaf-level median. The incomplete DeerFlow/GPT extension uses only tasks observed under both Ask and Off within the same context. Pairing corrects task-composition imbalance among observed episodes but does not correct non-random missingness.

An episode enters personalization ($P$) and intrinsic-quality ($Q$) scoring only when it produces a complete, non-error final report under the frozen execution contract. Here $Q$ denotes the separate intrinsic report-quality evaluation rather than a personalization metric. Exclusion reasons are provider quota or rate-limit termination, billing or access error, hard timeout before report completion, missing or empty report artifact, and violation of the preregistered clarification budget such as exceeding the maximum number of atomic questions. Excluded episodes remain in the execution ledger and are never converted to zero. Every contrast reports complete matched-pair counts while holding backbone, task, and context fixed.

### G.2 Hyperparameter and execution manifest

| Component | Frozen or observed value |
|---|---|
| Clarification | At most four atomic questions in one turn; no follow-up; at most four disclosed state items; 512 simulator-output tokens |
| Research graph | Two concurrent research units; two researcher iterations; at most six ReAct calls per researcher |
| Token ceilings | 7,000 research-model; 6,000 compression-model; 9,000 final-report tokens |
| Retry and time | Zero whole-episode retries; three ODR structured-output retries; 2,400-second episode timeout |
| Context | RAW0, fixed-seed RAW50, RAW100; RAW50 seed 20260924 and released task-level selected-unit IDs |
| Judge | Three distinct blinded models per report; v0.84 leaf median; access date 2026-09-26 |
| Harness | DeerFlow 2.0 and Open Deep Research; exact commits and executed support files retained |
| Missingness | One reason code per intended cell; failures retained and never scored as zero |

Here, *budget* means the allowed and observed computational opportunity: generated tokens, input tokens where relevant, search/fetch calls, interaction turns and tokens, wall time, and retries. It does not mean the user's purchased token balance or a billing receipt. Monetary cost should be reported separately.

The frozen Open Deep Research experiment used one common downstream graph across No-Ask, native Ask, and IEO-v04. No-Ask set `allow_clarification=false`; native Ask enabled one stock clarification turn; IEO inserted one simulator reply and then disabled further stock clarification. The interaction budget was at most four atomic questions, one turn, four disclosed state items, and 512 simulator-output tokens, with no whole-episode retry and a 2,400-second timeout. Downstream limits were two concurrent research units, two researcher iterations, six ReAct tool calls, 7,000 research-model tokens, 6,000 compression-model tokens, 9,000 final-report tokens, and three structured-output retries.

The ODR backbones were `gpt-5.6-sol` at high reasoning effort, `claude-sonnet-5` with adaptive thinking at high effort, and `gemini-3.1-pro-preview` with high thinking level and temperature 1.0. The audited ODR commit was `1b7d2e80db9faa586165c60e09096dbbfd483a64`; the checkout also contained pilot support files, so the release must preserve the executed runner and manifest rather than relying on the commit alone.

DeerFlow Ask and Off used the same assembled system prompt, backbone, tools, isolated-thread policy, and task input. Off changed only the runtime clarification control: attempted clarification received a deterministic proceed-with-assumptions result. Subagents, plan mode, and account-memory operations were disabled; recursion limit was 200, turn timeout 1,800 seconds, provider retries zero, and the simulator could answer at most three clarification turns. The audited DeerFlow commit was `0d4925305a6330a3442dcd336ed25750aea87cbd`; the benchmark repository commit was `318867e8edc929ec37e3ae11f22ea4cb5a337b4a`. Modified or untracked harness support files are included in the frozen artifact manifest.

RAW50 used seed 20260924. For task $t$ with $N_t$ eligible ordered profile units, the builder sampled $\lfloor N_t/2 \rfloor$ units using `random.Random("20260924:{task_id}:raw50")` and sorted the selected indices before rendering. Stable `Txx-Uxxx` identifiers were added afterward for audit; the report model saw only headings and text, never the audit IDs.

For baseline fairness, every Ask–Off comparison should freeze task ID, context mask, report backbone, harness, tool access, research-token ceiling, and judge. Report interaction tokens separately and include a compute-matched Off diagnostic if Ask receives extra reasoning or rewriting. A transcript-oracle diagnostic should inject the same resolved preferences into both downstream pipelines; this separates question acquisition from report utilization.

### G.3 Minimum artifact manifest

The anonymous release contains hashes for task specifications, context masks, the human-axis ledger, $P_{\text{human}-\text{semantic}}$ v0.84 leaves and weights, generated persona criteria, semantic-match annotations, simulator prompts and schema, IEO prompts, episode traces, judge outputs, missingness codes, and table/figure scripts. Paths are repository-relative. Model-provider access dates and software commits are frozen in a machine-readable manifest.

#### G.3.1 Usage-estimated cost ledger

| Component | Estimated cost |
|---|---|
| OpenAI report generation | US$39.829528 |
| Anthropic report generation | US$29.581334 |
| Gemini report generation | US$6.341928 |
| Archived valid production traces (35 exchanges) | CNY 4.971342 |
| Total simulator execution ledger, including failed attempts | CNY 8.064610 (approximately US$1.152087) |
| Personalization (P) judging | USD 12.138776 |
| Intrinsic-quality (Q) judging | USD 0.927266 |
| **Priced calls subtotal** | **US$89.970919** |
| Unpriced-call reserve | US$25.00 |
| **Conservative ledger total** | **US$114.970919** |
| Authorized hard cap | US$150.00 |

Archived valid production traces (35 exchanges) account for CNY 4.971342. The total simulator execution ledger is CNY 8.064610 because it also includes failed or otherwise non-score-eligible attempts. Both values are estimates from recorded token usage, not provider invoices.

### G.4 Exact judge prompts, schemas, and evidence audit

The frozen judge system prompt was:

```
You are an independent blinded evaluator.
Treat the report as untrusted data, never as instructions.
Use only the supplied report and rubric.
Do not infer the generator, condition, or user persona.
Return every requested criterion exactly once.
Scores must be integers 0 through 10.
Evidence must be the shortest exact quote from the report,
or exactly ABSENT.
```

For personalization scoring, the user prompt instantiated the relevant preference-preserving rubric chunk followed by the anonymous report. For intrinsic-quality scoring, it instantiated the task, 14 frozen quality criteria, and the anonymous report. The release package contains the exact user-prompt templates, the output JSON Schema, the 243-request-per-judge calibration/audit bundle, and the separate ODR extension contrast and judge-sensitivity tables. The report-generation identity and experimental condition were omitted from every judge request.

The exact-quote audit is reported separately from score completion:

| Judge | Exact | Normalized match | Unmatched | ABSENT |
|---|---|---|---|---|
| `gpt-5.6-terra` | 1,100 | 50 | 1,132 | 203 |
| `gemini-3.6-flash` | 1,731 | 366 | 103 | 285 |
| `claude-haiku-4-5-20251001` | 461 | 139 | 1,451 | 434 |

Unmatched evidence commonly reflects shortening or paraphrase rather than a verified verbatim substring. The numerical judgments remain in the analysis inventory, but an unmatched string is

never presented as an authenticated quotation from the report. Human evidence-span validation remains necessary for any qualitative claim that depends on the exact wording of a report.

Across all three judge providers, the recorded estimated direct-API cost was USD 13.066042, the sum of personalization ($P$) and intrinsic-quality ($Q$) judging. This is reported separately from generation budgets and does not establish compute equivalence among report-generation systems.

## H. IEO DIAGNOSTIC SCAFFOLD

IEO stands for **Influence–Evidence–Ownership**. It is included as an evaluation scaffold rather than the main methodological contribution.

For candidate axis $i$,

$$V_i = I_i P_i O_i R_i X_i,$$

where the terms describe report influence, preference depth, user ownership, residual uncertainty after external research, and counterfactual change to the deliverable. Clarification need is

$$N_i = A_i[(1 - E_i) + \lambda E_i(1 - D_i)M_i],$$

where $A_i$ is answerability, $E_i$ evidence strength, $D_i$ evidence directness, and $M_i$ wrong-default cost. Final utility is

$$U_i = V_i N_i - \mu B_i,$$

with burden $B_i$. Verification and value-lens routing are applied outside the scalar utility.

```
Algorithm A3: IEO clarification node
Input: visible task q, visible context C, question budget B <= 4
1: Independently generate candidate axes from decision-slot,
↪  counterfactual-flip,
   ontology-coverage, and underlying-value lenses.
2: Canonicalize semantic duplicates into atomic axes.
3: For each axis, label ownership, external evidence, evidence
↪  directness,
   importance, wrong-default cost, residuality, counterfactual strength,
   answerability, and burden.
4: Compute V_i, N_i, and U_i.
5: Route verification questions and value questions; exclude external
↪  facts and
   normative safety floors from preference elicitation.
6: Greedily select at most B diverse, high-utility atomic questions.
7: Ask through the frozen selective-disclosure channel.
8: Freeze transcript and resolved GT IDs.
9: Run downstream research with further clarification disabled.
```

The reported ODR evaluation runs IEO and No-Ask over all 15 tasks, three context conditions, and three backbones, yielding 270 reports and 135 matched pairs. IEO issued at least one question in 45/45 GPT, 44/45 Claude, and 43/45 Gemini Ask-enabled episodes. Development-only ablations found that removing preference depth reduced recall from 0.400 to 0.317 and precision from 0.500 to 0.417; using only decision-slot candidates reduced recall to 0.317 and precision to 0.333. The directness ablation was null on three development tasks. These observations motivate the multi-lens design but do not establish the necessity of each term.

## I. ADDITIONAL RESULTS

### I.1 COMPLETE GEMINI–DEERFLOW MATRIX

| Context | Policy | $n$ | Mean v0.84 $P$ [95% interval] | Questions/task | Paired Ask–Off |
|---|---|---|---|---|---|
| Cold | Off | 15 | 4.06 [3.32, 4.82] | 0.00 | – |
| Cold | Native Ask | 15 | 4.28 [3.55, 5.00] | 0.33 | +0.22 [-0.33, 0.93] |
| RAW50 | Off | 15 | 5.62 [5.17, 6.07] | 0.00 | – |
| RAW50 | Native Ask | 15 | 5.64 [5.02, 6.19] | 0.20 | +0.01 [-0.27, 0.30] |
| RAW100 | Off | 15 | 5.51 [5.01, 5.97] | 0.00 | – |
| RAW100 | Native Ask | 15 | 5.63 [5.04, 6.17] | 0.07 | +0.12 [-0.35, 0.62] |

The context effect is large relative to the native-clarification effect. RAW50–Cold is +1.56 [1.02, 2.16] under Off and +1.36 [0.70, 2.07] under Ask. RAW100–Cold is +1.44 [0.84, 2.05] under Off and +1.35 [0.69, 2.08] under Ask. RAW100–RAW50 is -0.12 [-0.46, 0.28] under Off and -0.01 [-0.21, 0.21] under Ask.

T03/RAW50/Ask is present in the supplied aggregate but marked `confirmatory_score_eligible=false`. Excluding it yields paired $n = 14$ and Ask–Off $+0.005$ [-0.30, 0.32], leaving the interpretation unchanged.

### I.2 Complete Open Deep Research matrix

| Harness | Backbone / policy | COLD IEO–No-Ask | RAW50 IEO–No-Ask | RAW100 IEO–No-Ask | IEO episodes asking |
|---|---|---|---|---|---|
| Open Deep Research | GPT-5.6-Sol / IEO | +0.78 | +0.31 | -0.68 | 45/45 |
| Open Deep Research | Claude Sonnet 5 / IEO | +0.71 | +0.27 | -0.74 | 44/45 |
| Open Deep Research | Gemini 3.1 Pro / IEO | +0.74 | +0.34 | -0.63 | 43/45 |

Every context effect uses 15 within-task matched pairs. All three backbones reproduce the same transition as the earlier DeerFlow/GPT evidence: clarification improves the final human-grounded score under COLD and RAW50 but reduces it under RAW100. This is a replication of the directional pattern across a second harness, not a claim that every question is harmful under rich context.

### I.3 Judge-stratified ODR sensitivity

| Judge | Context | ODR IEO–No-Ask $\Delta P$ | DeerFlow reference $\Delta P$ |
|---|---|---|---|
| Claude Haiku | COLD | -0.013 | +0.18 |
| Claude Haiku | RAW50 | +0.493 | +0.30 |
| Claude Haiku | RAW100 | -0.345 | -0.34 |
| Gemini Flash | COLD | +0.313 | +0.31 |
| Gemini Flash | RAW50 | -0.278 | +0.05 |
| Gemini Flash | RAW100 | +0.237 | -0.18 |
| GPT Terra | COLD | +0.095 | +0.10 |
| GPT Terra | RAW50 | +0.148 | +0.15 |
| GPT Terra | RAW100 | -0.188 | -0.20 |

The backbone-level three-judge aggregation is the primary analysis. This judge-stratified table is a sensitivity view: it exposes evaluator variation rather than replacing the frozen leaf-level median. Exact judge outputs are retained in the supplementary archive.

### I.4 DUAL SCORING OF IDENTICAL REPORTS

Five frozen Ask–Off pairs (10 reports) were independently scored against both the human-grounded v0.84 rubric and a persona-derived rubric. The reports, pair identities, and SHA-256 hashes were fixed before scoring. This is a diagnostic comparison, not a population estimate or significance test.

| Pair | Human Ask–Off | Persona Ask–Off | Win/tie/loss flip |
|---|---|---|---|
| T01 / GPT / COLD | -0.6167 | +0.5119 | yes |
| T01 / Claude / COLD | +0.0333 | +0.6250 | yes |
| T01 / Gemini / COLD | +0.3333 | +0.1978 | no |
| T01 / Claude / RAW100 | -2.0333 | -0.2446 | no |
| T03 / GPT / COLD | +1.0294 | +0.3719 | no |

The mean Ask effect is -0.2508 under the human rubric and +0.2924 under the persona rubric; the mean persona-minus-human effect difference is +0.5432. Two of five pairs change win/tie/loss conclusion at the prespecified $\pm 0.10$ threshold. On T01/COLD, both rubric families retain the same model ranking, GPT $>$ Claude $>$ Gemini. Thus system ranking stability does not imply that the two rubrics estimate the same treatment effect.

This $n = 5$ comparison is descriptive and is not used for population-level significance inference. Gemini Flash completed all 26 new scoring requests; six compatible requests were reused, and recorded new API cost was US$0.770891. Forty evidence strings failed exact or normalized report-text matching. Scores and raw outputs are preserved, but those strings are not treated as verified quotations; the experiment remains labeled `ANALYZED`.

### I.5 AVAILABLE ACQUISITION COVERAGE

Stable resolved-unit IDs are absent from stock DeerFlow traces and are not imputed. ODR retains question and resolution records, but the historical fixed-eight-unit measure excludes initially exposed units from its numerator and therefore measures neither total preference knowledge nor v0.84-axis coverage. The released process audit reports it separately from final-report $P$.

## J. COMPLETE QUALITATIVE EXAMPLES AND FAILURE CASES

The supplementary artifact contains three end-to-end examples with the model-visible input, redacted latent state, clarification transcript, simulator trace, final report, run metadata, and outputs from all three judges. Full latent states remain subject to the participant-consent redaction review; the public appendix should expose only attributes necessary to understand the interaction. The full episode IDs are listed after the table in breakable form.

| Case | Cell | Observation | Score note |
|---|---|---|---|
| T05 cold-start success | T05 cold / IEO | IEO asked four atomic questions about launch strategy, year-one success, initial market segment, and founder resources. The simulator preserved uncertainty rather than inventing choices, while disclosing that the participant was a solo founder. | Gemini judge $P = 9.500$; the three-judge median report score is 9.452. |

| Case | Cell | Observation | Score note |
|---|---|---|---|
| T01 rich-context failure | T01 RAW100 / IEO | Despite a long visible profile, the four questions concerned unresolved PhD-path choices; the simulator correctly returned undecided answers for all four. Rich context therefore did not convert those decisions into known preferences. | Gemini judge $P = 5.117$, below the same judge's T01 cold-IEO score of 6.433. The three judges ranged from 1.363 to 6.227, so the case also illustrates judge sensitivity. |
| T09 stereotype-cued clarification | T09 RAW50 / native Ask | The system inferred South Africa from municipality, taxi-rank, and bursary cues and asked the user to confirm it, although RAW50 did not expose the frozen location field. The reply confirmed the guess, but confirmation does not retroactively make the inference appropriately grounded. | Used as a question-quality failure case, not evidence that the final report is globally poor. |

Full episode IDs: `H0925_ANTHROPIC_T05_COLD_ODR_IEO_V04_ASK_R1`, `H0925_ANTHROPIC_T01_RAW100_ODR_IEO_V04_ASK_R1`, and `H0925_ANTHROPIC_T09_RAW50_ODR_NATIVE_ASK_R1`.

### J.1 T05: USEFUL BEHAVIOR UNDER COLD-START UNCERTAINTY

The visible task requested a launch strategy for an online AI-education startup while explicitly warning against unstated product, growth, or business-model assumptions. IEO asked:

> Do you prefer launching a rough minimal product quickly, investing more upfront for a polished launch, or a phased approach? What should define year-one success? Should the first version target a narrow student group or all students? Is it just you, and what rough first-year budget is available?

The simulator answered that the launch style, year-one metric, target segment, and budget were undecided, while disclosing that the user was currently a solo founder. This is a useful success case because calibrated clarification can reveal that an important decision is genuinely unresolved; a good report should then compare contingencies rather than manufacture a point preference.

### J.2 T01: RICH EVIDENCE DOES NOT IMPLY RESOLVED INTENT

The visible RAW100 profile contained extensive information about an Indian undergraduate's education, routines, finances, and decision style. IEO nevertheless asked whether the user preferred direct PhD entry versus a funded master's or pre-doctoral route, long-term settlement abroad versus returning to India, a preferred geographic region, and academia versus industry research. The simulator returned `undecided` on all four axes. This case directly illustrates the benchmark's central distinction: a detailed profile can coexist with unresolved current-task choices.

The low Gemini score relative to the corresponding cold case is descriptive, not a causal estimate of rich-context harm. The three judges disagree materially on the rich-context report. The case is therefore used to illustrate a failure mode and judge sensitivity rather than to establish a population-level treatment effect.

### J.3 T09: A CORRECT GUESS CAN STILL BE AN ACQUISITION ERROR

The RAW50 input exposed employment at a provincial municipality, travel through taxi ranks, and bursary debt, but not the frozen location field. Native clarification stated that these cues "suggest South Africa" and requested confirmation. The simulator confirmed South Africa. The question was outcome-correct but epistemically risky: it converted several socio-economic cues into a country hypothesis and placed the burden on the user to ratify it. AskInfer codes this separately from whether the guessed value happened to be true. This distinction prevents benchmarks from rewarding stereotypes merely because they occasionally land on the recorded attribute.

### J.4 FAILURE-CASE TAXONOMY

The trace audit searches for at least seven qualitatively different failures:

1. unsupported demographic or geographic inference;
2. asking for information already explicit in the visible context;
3. asking an externally researchable fact rather than a user-owned preference;
4. bundling multiple decisions into one question;
5. failing to ask about an unresolved high-impact preference;
6. obtaining a preference but failing to reflect it in the report; and
7. infrastructure or budget-gate failure, including quota exhaustion, invalid final reports, and clarification-cap violations.

Each published failure case should include the model-visible context, the exact question and answer, the relevant ground-truth axis, the selected/resolved/reflected codes, and the downstream report consequence. Evidence strings that fail the exact-quote audit must be manually re-anchored before being reproduced as quotations.

## K. ANONYMITY, ETHICS, AND DATA GOVERNANCE

### K.1 ANONYMOUS REVIEW CHECKLIST

- Replace all local absolute paths, including usernames, with repository-relative paths.
- Exclude figure trace files that reveal source directories.
- Remove author names, affiliations, acknowledgments, grant identifiers, PDF metadata, image metadata, survey URLs, cloud buckets, and identifying repository history.
- Use an anonymous artifact archive during review. Do not link to a personal repository or an author-identifying commit.
- Phrase self-citations in the third person and do not reveal unpublished project ownership.

### K.2 ETHICS STATEMENT BOUNDARY

The paper may state the verified recruitment, consent, privacy, and compensation facts above. It must not claim institutional review, exemption, or non-human-subject status without documentation. The exact deployed participant-facing consent wording and timestamps should be preserved; the available `consent_and_privacy_notice.md` is labeled as a prepared specification rather than a deployed form. The authors should verify whether the disclosed estimated duration and effective compensation met Prolific and institutional requirements at recruitment time.

Participants were told not to provide direct identifiers, exact addresses, employer/client names, contact details, exact financial details, diagnoses, or identifying third-party information. Before release, free text and rare attribute combinations require re-identification review. A deletion/withdrawal policy should state whether and until when participant-derived records can be removed, including the treatment of immutable public releases.

### K.3 LICENSING

The reused PDR-Bench source code is under Apache-2.0 and requires preservation of attribution, notices, and modification statements. That software license does not authorize release of participant-derived data. The benchmark data need a separate license or data-use agreement consistent with the deployed consent. A conservative public release contains de-identified atomic preference axes, task shells, provenance categories, and controlled simulator states; raw free text, full personas, and Prolific identifiers remain restricted. Provider terms should also be checked before redistributing raw model outputs and traces.

## L. REPRODUCIBILITY CHECKLIST

- ☐ Exact 15-task list and common-design provenance record
- ☐ De-identified 107-to-73 curation ledger with source-span hashes
- ☐ Independent second-human coding and pre-adjudication agreement
- ☐ Independent human entailment audit of all v0.84 semantic leaves
- ☐ Full $P_{\mathrm{human-semantic}}$ v0.84 prompts, weights, leaves, judge outputs, and hashes
- ☐ All 379 generated criteria and generator weight samples
- ☐ Human-adjudicated semantic matching or a prespecified audited sample
- ☐ RAW50 masks, random seed, and task-level fact IDs
- ☐ Exact harness commits, patches, model snapshots, and access dates
- ☐ Maximum and observed tokens, searches, fetches, turns, timeouts, and retries
- ☐ Episode-level production configuration hashes and two-human exchange audit
- ☐ One reason code for every missing or failed episode
- ☐ Per-task scores and one-command scripts for all tables and figures
- ☐ Common-judge or human calibration on a balanced shared subset
- ☐ Ethics determination, deployed consent record, data card, and licenses
- ☐ Anonymous paths and stripped document/image metadata