PACESHOP: Evaluating Personalized, Actionable, Compositional, and Evidence-grounded Shopping Assistants

Anonymous ACL submission

Abstract

Shopping assistants are shifting from ranked product lists toward structured decision support, where systems must synthesize shopper context, product evidence, and next-step guidance into a coherent recommendation experience. This changes the unit of evaluation: a fluent response can still fail by ignoring shopper context, contradicting itself across components, or leaving defects too vague to localize. Existing personalization, grounding, and LLM-as-a-judge benchmarks cover pieces of this problem, but they do not define a joint evaluation target for structured shopping-assistant responses. We formulate this missing evaluation target as **PACE**: **P**ersonalized, **A**ctionable, **C**ompositional, and **E**vidence-grounded evaluation. We instantiate PACE with two artifacts: **PACESHOP**, a benchmark dataset that makes the target measurable through 22,625 controlled records with structured personas, auditable evidence pools, GOOD/BAD labels, and gold defect family and location annotations; and **PACEJUDGE**, a training-free judging protocol that makes the target reportable through a structured output contract. Our experiments show that generic judges can recognize broad quality but fail to recover the diagnostic fields required for PACE; PACESHOP makes these failures verifiable, and PACEJUDGE improves persona-source, cross-component, grounding, and family/location closure without retraining, showing that realistic shopping-assistant evaluation requires a task-matched output contract rather than only a stronger backbone or scalar prompt.

1 Introduction

Shopping assistants are moving beyond ranked product lists and single answers toward structured decision support: a response may summarize the shopper's need, organize product categories, suggest related query refinements, and cite evidence from a retrieved product pool (Miroyan et al., 2025; Wang et al., 2025a). This changes the unit of evaluation. A response is no longer valid merely because one generated text is fluent or broadly relevant; it must be valid as a structured, persona-conditioned object. As Fig. 1 illustrates, a response can appear coherent while still failing in ways that a standard scalar judge may miss: it may recommend a brand the shopper avoids, contradict itself across response fields, or leave the defect unable to localize. Such failures are difficult to capture with an overall quality score, yet they are central to practical evaluation: a realistic judge must identify not only whether the response is wrong, but also what failed, where it failed, and which evidence supports the diagnosis.

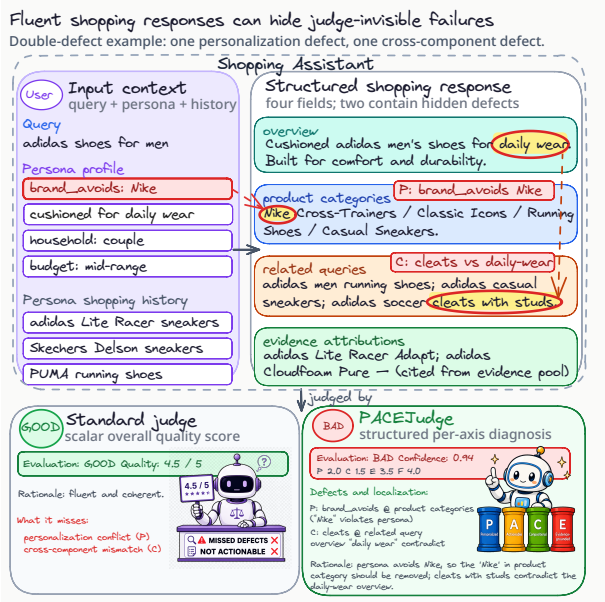

Figure 1: **Generic judges can miss structured failures in shopping-assistant responses.** For the same query, persona, and response, a standard judge returns GOOD, while PACEJUDGE returns BAD with defect family and field location. The example failures include a Nike preference conflict and a cross-field mismatch in related queries, exposing the need for PACE evaluation.

These failures expose an under-specified evaluation target. Existing personalization benchmarks study whether systems adapt to user profiles, histo-

ries, preferences, or long-context behavioral signals (Salemi et al., 2024; Zhao et al., 2025b,a; Jiang et al., 2025; Rahman et al., 2025; Hao et al., 2025; Dong et al., 2024). Grounding-focused evaluators test whether generated claims are supported by retrieved evidence (Saad-Falcon et al., 2024). Generic LLM-as-a-judge protocols score open-ended outputs with rubrics, scalar ratings, or pairwise preferences (Zheng et al., 2023; Liu et al., 2023; Thakur et al., 2025; Li et al., 2025; Chen et al., 2025; Feuer et al., 2025; Enguehard et al., 2025). These directions are complementary, but they do not define the joint target required for structured shopping-assistant responses: the judge must condition on shopper context, reason across response components, verify evidence support, and return a diagnosis that localizes the failure. The gap is therefore conceptual before it is empirical. A stronger judge backbone or a better scalar prompt cannot report defect fields that the evaluation task never asks it to produce.

We formulate this missing evaluation target as **PACE**: ***P****ersonalized*, ***A****ctionable*, ***C****ompositional*, and ***E****vidence-grounded*. ***P****ersonalized* means that validity depends on the shopper context; ***A****ctionable* means that the judge must identify what failed and where; ***C****ompositional* means that the response must be evaluated as a structured object rather than isolated text; ***E****vidence-grounded* means that claims must be supported by product evidence and persona history. Together, PACE clarifies what current evaluation lacks: records that make such failures observable, and a judging protocol that can report them. Tab. 2 summarizes this design principle: each PACE target determines what the benchmark must contain, what the judge must output, and what scenario-level metric can test it.

To make PACE measurable, we introduce PACESHOP, a verifiable benchmark dataset for structured shopping-assistant evaluation. Existing datasets provide valuable personalization or judging signals, but they do not jointly expose persona-dependent validity, cross-component consistency, evidence support, and localizable defects in one controlled setting. PACESHOP fills this gap by pairing each shopping query with a structured persona, a candidate response, an auditable evidence pool, and a GOOD/BAD label; BAD records additionally include gold defect family and defect location annotations. This design turns the abstract PACE formulation into measurable evaluation cases.

To make PACE reportable, we propose PACEJUDGE, a training-free judging protocol. Existing scalar judges can estimate overall quality, but they do not require the judge to name the violated dimension, localize the broken response field, or ground the diagnosis in evidence. PACEJUDGE fills this gap through a judging protocol that evaluates PACE with a a structured output schema. The protocol is backbone-agnostic, allowing us to separate what comes from model capability from what comes from asking the judge to produce the right diagnostic fields.

Across multiple LLM backbones and judging configurations, we compare broad GOOD/BAD discrimination with scenario-level tests that require persona-source diagnosis, cross-component localization, evidence-grounding checks, and actionable defect labels. The results show that general discrimination is not the main bottleneck: scalar and generic judges can often recognize that a response is broadly good or bad. The gap appears when evaluation requires the diagnostic fields needed for debugging. On the same backbones, PACEJUDGE improves these contract-dependent metrics without retraining, supporting our central claim: realistic shopping-assistant evaluation requires diagnostic closure over persona, components, evidence, and defect location, not only broad quality discrimination. Our contributions are threefold:

1. We identify a problem-setting gap in existing evaluation for structured shopping assistants, and formulate the missing target as PACE.
2. We introduce PACESHOP, a benchmark dataset that makes PACE measurable through controlled persona-, component-, and evidence-grounded failure cases.
3. We propose PACEJUDGE, a training-free judging protocol that makes PACE reportable and improves diagnostic metrics beyond scalar quality scores.

2 Related Work

Personalization, shopping, and search benchmarks. Personalized LLM benchmarks evaluate user-conditioned generation, preferences, memory, likability, and tool use (Salemi et al., 2024; Zhao et al., 2025b,a; Jiang et al., 2025; Rahman et al., 2025; Hao et al., 2025; Dong et al., 2024). Commerce-oriented resources such as ECom-Bench (Wang et al., 2025a), Search Arena (Miroyan et al., 2025), and OPeRA (Wang

Benchmark / Resource	Domain	Size	P	A	C	E	Evaluation signal
LaMP (Salemi et al., 2024)	General personalization	7 tasks	✓	✗	✗	✗	Task metrics for user-conditioned generation
PersonaLens (Zhao et al., 2025b)	Task-oriented assistants	1.5K profiles, 111 tasks	✓	✗	✗	✗	LLM user and judge agents
PrefEval (Zhao et al., 2025a)	Long-context preferences	3K pref-query pairs	✓	✗	✗	✗	Preference-following generation/classification
PersonaMem (Jiang et al., 2025)	Dynamic personalization	180 histories, $\leq$ 60 sessions	✓	✗	✗	✗	Response selection over evolving user history
LikeBench (Rahman et al., 2025)	Likability personalization	not publicly stated	✓	✗	✗	✗	Simulated-user likability diagnostics
EtaPP (Hao et al., 2025)	Personalized tool use	800 test cases	✓	✗	✗	✗	Key-point judging for tool-use trajectories
OPeRA (Wang et al., 2025b)	Shopping behavior simulation	692 sessions, 28.9K obs.-action pairs	✓	✗	✗	✗	Real-user observations, actions, and rationales
ECom-Bench (Wang et al., 2025a)	E-commerce support	53 tasks, hundreds of personas	✓	✗	✗	✗	Task-success trajectories
ESCI (Reddy et al., 2022)	Product search relevance	97.3K US queries	✗	✗	✗	✗	Human product-relevance labels
Search Arena (Miroyan et al., 2025)	Search-augmented chat	24K interactions, 12K votes	✗	✗	✗	✗	Pairwise human preferences
PACESHOP (ours)	**Personalized shopping**	**1,132 q × 1.2K p = 22.6K records**	✓	✓	✓	✓	**Gold GOOD/BAD, defect family, and defect location labels**

Table 1: Existing benchmark datasets against the four PACE targets. PACESHOP jointly covers all targets through structured personas, multi-component responses, auditable evidence pools, and gold defect family/location labels.

et al., 2025b) add realistic e-commerce support, search-augmented chat, or online-shopping behavior signals. However, they do not provide the joint evaluation unit needed and cannot directly test whether a response satisfies PACE targets: personalization, actionability, compositional consistency, and evidence grounding. Tab. 1 summarizes this benchmark-side gap.

LLM-as-a-judge and grounded evaluation. Generic LLM-as-a-judge methods such as G-Eval (Liu et al., 2023) and MT-Bench (Zheng et al., 2023) established rubric-based, scalar, and pairwise judging for open-ended model outputs. Subsequent work studies judge reliability, alignment, bias, distributional validity, and domain-specific judging protocols (Thakur et al., 2025; Li et al., 2025; Chen et al., 2025; Feuer et al., 2025; Enguehard et al., 2025). Grounding-focused evaluators such as ARES (Saad-Falcon et al., 2024) test whether generated claims are supported by retrieved evidence. These methods are complementary, but their evaluation contracts are not designed for structured shopping-assistant diagnosis: they do not jointly require persona conditioning, cross-component reasoning, evidence support, defect-family prediction, and field-level localization. PACEJUDGE addresses this protocol-side gap by making the four PACE targets reportable through structured diagnostic outputs.

3 PACESHOP Dataset

PACESHOP is designed to make PACE measurable. The goal is to construct controlled and verifiable records in which personalization, actionability, compositional consistency, and evidence grounding can be observed and audited. Tab. 2 previews how each PACE target is turned into measurable benchmark controls in PACESHOP. The benchmark construction follows a valid-response-first, controlled-defect-second design: we first create and validate structured GOOD responses, then derive BAD variants through controlled edits with known defect family and location labels.

3.1 Benchmark Task

Scope. PACESHOP focuses on the generated response layer of shopping assistants rather than the full ranking interface or end-to-end shopping session. This scope isolates the judging problem: given a shopping assistant user inputs and responses, can a judge determine whether the response is valid and diagnose any failure?

Task formulation. Each PACESHOP record defines a pointwise judging task over a tuple

$$(q, p, y, E, \ell),$$

where q is a shopping query, p is a structured persona, y is a candidate shopping-assistant response, E is an auditable evidence pool, and $\ell \in \{\text{GOOD}, \text{BAD}\}$ is the validity label. The judge receives (q, p, y, E) and predicts

$$\hat{\ell} = J(q, p, y, E), \qquad \hat{\ell} \in \{\text{GOOD}, \text{BAD}\}.$$

For BAD responses, the judge should additionally diagnose the failure by predicting a defect family $d \in \mathcal{D}$ and a defect location $r \in \mathcal{R}$. Thus, the benchmark measures not only whether a judge can separate GOOD from BAD responses, but also whether it can identify what failed and where the failure occurs.

3.2 PACESHOP Construction Pipeline

The PACEShop construction pipeline makes PACE failures observable and auditable. It isolates shopper-context effects through persona assignment, makes grounding checkable through evidence-backed query retention, uses validated GOOD responses as verified base cases, and derives controlled BAD variants with gold defect family and location labels. Full construction details, prompts, schema checks, and examples are provided in Appx. B.

Persona and evidence setup. We construct 1,200 personas as structured shopping contexts rather than free-form user descriptions. Each persona specifies stable shopper information, such

PACE target	PACESHOP	PACEJUDGE protocol	Evaluation scenario
Personalized Validity depends on shopper context.	Pairs each query with structured personas; persona-swap records hold the query and response fixed while changing shopper context.	Requires persona-conditioned judging; reports persona-alignment score, rationale, and persona-conflict diagnosis.	**S1 Persona**: tests whether the judge detects when a swapped persona invalidates the same response.
Actionable Evaluation must say what failed and where.	Provides BAD records with gold defect family and gold defect location from a 7-family $\times$ 4-field taxonomy.	Requires structured fault diagnosis; reports defect family, defect location, confidence, rationale, and supporting evidence IDs.	**S4 Actionable**: tests exact defect-family prediction and response-field localization.
Compositional Responses must be consistent as structured objects.	Represents responses with multiple heterogeneous components; controlled BAD variants inject cross-component defects.	Requires judging the response as a structured object; reports compositional-consistency score and localized response component.	**S2 Compositional**: tests cross-component defect detection and localization.
Evidence-grounded Claims must be supported by product and persona evidence.	Includes an auditable evidence pool; controlled BAD variants inject invalid product evidence and invented persona-history claims.	Requires grounding against product evidence and persona history; reports evidence-grounding score and evidence IDs constrained to the evidence pool.	**S3 Grounding**: tests product-evidence grounding, persona-history grounding, and hallucinated-ID control.

Table 2: **Making PACE measurable and reportable.** Each PACE target maps to what PACESHOP makes measurable, how the PACEJUDGE protocol makes it reportable, and which evaluation scenario tests the corresponding failure mode.

as household type, budget, quality sensitivity, urgency, constraints, preferences, and brand affinities, together with concrete shopping behavior in the form of purchase history and recent searches. We pair these personas with evidence-backed shopping queries drawn from a public query and product-evidence pool. This setup lets the same query and evidence pool support different valid responses under different shopper contexts, making personalization and persona-history grounding testable.

Query–persona assignment. We begin with 97,227 normalized US shopping queries, 1,200 structured personas, and 4,460 public evidence records across 11 shopping domains.[1] We retain 1,132 queries with sufficient evidence support and pair each retained query with distinct personas, yielding 4,525 validated query–persona pairs. The repeated-measures design holds the query and evidence pool fixed while varying shopper context, allowing persona-dependent failures to be isolated from changes in shopping intent.

GOOD-response generation. For each query–persona pair, a strong LLM generator produces one structured GOOD candidate under a constrained response contract. We do not treat these generations as valid merely because they come from a strong model: each candidate must pass deterministic normalization and validation, including schema, cardinality, attribution-ID, and parsing checks. Accepted GOOD responses therefore serve as contract-valid benchmark anchors for controlled defect construction.

Controlled BAD construction. From each validated GOOD response, we construct four BAD variants containing one, two, three, and four defects. Defects are drawn from seven families covering persona conflicts, intent drift, cross-component mismatch, redundancy, evidence mismatch, over-personalized hallucination, and unsupported claims. Because each BAD record is created by a controlled edit, its gold defect family and field location are known by construction.

Validity checks. PACESHOP is designed to make PACE failures measurable and verifiable, not to rely on synthetic generation as ground truth. Each released record carries an audit trail: deterministic validators check schema compliance and evidence-ID consistency, validated GOOD responses are used as base cases before any defect injection, controlled edits make BAD labels traceable to known defect families and response fields, and evidence-provenance flags expose the strength of the grounding signal. Together, these layers address the main dataset-quality risks: synthetic-on-synthetic circularity, thin grounding signal, persona-template artifacts, rigid schema effects, and limited transfer beyond one model family. Appx. B.4 details the five verification layers; Appx. B.5 defines the defect taxonomy; Appx. B.9 gives GOOD/BAD record generation procedures.

[1] Coverage is audited at the source level: queries are tagged by retail vertical and shopping mission, personas by shopper attributes and behavior-history fields, and evidence records by domain and provenance. Appx. B.3 reports the detailed coverage breakdowns and representative examples.

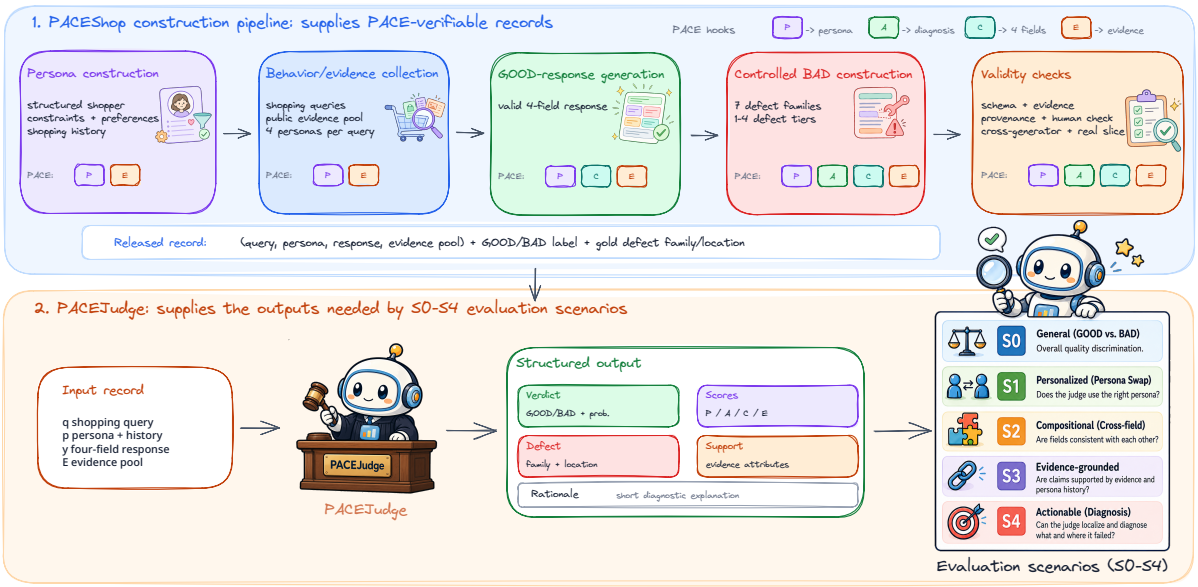

Figure 2: PACESHOP benchmark dataset makes PACE measurable; and PACEJUDGE evaluation protocal makes PACE reportable. **Top:** PACESHOP data construction carries persona, shopping behavior, query intent, and evidence through assignment, GOOD-response generation, controlled BAD construction, and validity checks. **Middle:** the resulting records make PACE properties measurable through S1–S4, with S0 as a reference GOOD/BAD check. **Bottom:** PACEJUDGE turns each record (q, p, y, E) into a structured output whose fields support scenario-aligned evaluation.

Dataset statistics. PACEShop contains 22,625 records: 4,525 GOOD records and 18,100 BAD records. BAD records are evenly distributed across single-, double-, triple-, and quad-defect tiers, and every BAD record carries gold defect family and location annotations. Tab. 8 in Appx. B summarizes the benchmark statistics, evidence coverage, defect structure, and validation assets.

4 PACEJUDGE Evaluation Protocol

PACEJUDGE is a training-free pointwise judging protocol that makes PACE reportable. In Tab. 2, PACESHOP makes the PACE targets measurable through controlled records, while PACEJUDGE specifies how a judge should report the diagnosis. Given a PACESHOP record, the protocol asks the judge to evaluate the candidate response as a structured, persona-conditioned, evidence-grounded object rather than as a single fluent text. This addresses the central limitation of scalar judging: an overall quality score does not identify whether the failure comes from shopper-context mismatch, cross-component inconsistency, unsupported evidence, or a defect that cannot be localized.

4.1 Protocol Overview and Output Schema

Given a PACESHOP record (q, p, y, E), PACEJUDGE returns

$$J(q, p, y, E) \rightarrow (\hat{\ell}, \hat{p}_{\text{bad}}, s_{\text{P}}, a_{\text{A}}, s_{\text{C}}, s_{\text{E}}, s_{\text{F}}),$$

where $\hat{\ell} \in \{\text{GOOD}, \text{BAD}\}$ is the verdict, $\hat{p}_{\text{bad}} \in [0, 1]$ is the BAD probability, $s_{\text{P}}, s_{\text{C}}, s_{\text{E}} \in [1, 5]$ are the Personalized, Compositional, and Evidence-grounded scores, and $s_{\text{F}} \in [1, 5]$ is an auxiliary format/safety score. The Actionable target is represented as a diagnostic object rather than a scalar:

$$a_{\text{A}} = (\hat{d}, \hat{r}, \hat{S}, \hat{c}, \hat{z}),$$

where $\hat{d}$ is the predicted defect family, $\hat{r}$ is the predicted defect location, $\hat{S}$ is the set of supporting evidence IDs, $\hat{c}$ is confidence, and $\hat{z}$ is a short rationale. The diagnostic fields follow the benchmark schema:

$$\hat{d} \in \mathcal{D} \cup \{\text{NONE}\}, \quad \hat{r} \in \mathcal{R} \cup \{\text{NONE}\}, \quad \hat{S} \subseteq E,$$

where $\mathcal{D}$ is the seven-family defect taxonomy and $\mathcal{R}$ is the response-field set (*overview*, *categories*, *related queries*, and *attributions*). For GOOD verdicts, the defect family and location should be NONE; for BAD verdicts, the judge identifies the primary defect family and response field. Detailed protocol and examples can be found in Appx. D.4.

4.2 Coverage of the PACE Targets

The protocol is designed to cover all PACE targets. For **P**, the persona score s_{P} makes shopper context part of the verdict. For **A**, the diagnostic object a_{A} requires the judge to report what failed, where

Method	S0 Gen.	S1 P P-source	S2 C detect	S2 C loc.	S3 E grounding	S4 A family	S4 A loc.	False-fire ↓	PACE avg.
Rubric ablations of the same PACE output contract (7-backbone mean)									
G-Eval (Liu et al., 2023)	0.73	0.17	0.75	0.70	0.40	0.48	0.68	0.52	0.45
Scalar (MT-Bench-style) (Zheng et al., 2023)	0.73	0.12	0.77	0.60	0.50	0.49	0.68	0.51	0.44
Persona-blind	0.74	–	0.79	0.64	0.49	0.49	0.69	0.47	–
Persona-aware	0.74	0.20	0.79	0.64	0.60	0.50	0.68	0.37	0.51
PACEJUDGE (ours)	**0.74**	**0.32**	**0.93**	**0.72**	**0.62**	**0.59**	**0.78**	0.47	**0.56**
Honest-adapted single-target baselines (cells shown only where the native judging protocol supports the metric)									
ARES-style (Saad-Falcon et al., 2024)	0.69	–	–	–	0.39	–	–	0.26	–
PersonaLens-style (Zhao et al., 2025b)	0.65	0.18	–	–	–	–	–	**0.07**	–
EtaPP-style (Hao et al., 2025)	0.71	–	–	–	–	0.50	0.74	0.43	–

Table 3: **Main PACE-closure results averaged over seven LLM backbones.** *S0* reports reference GOOD/BAD balanced accuracy, while *S1–S4* evaluate PACE diagnostic closure. PACEJUDGE achieves the strongest overall performance among multi-target rubric and the single-target methods, showing that the main gap is not broad GOOD/BAD discrimination but recovering the diagnostic fields needed for PACE evaluation. *False-fire* measures defect alarms on GOOD records, and PACE *avg.* combines closure metrics with GOOD-record specificity. "–" marks unsupported cells under a method's native judging protocol. Full per-backbone results (Tab. 23) and metric definitions are provided in Appx. E.4.

it failed, and what evidence supports the diagnosis. For **C**, the compositional score s_{C} and defect-location field $\hat{r}$ require the judge to evaluate the response as a structured object rather than isolated text. For **E**, the evidence score s_{E} and evidence-ID constraint $\hat{S} \subseteq E$ require grounding in both the product evidence pool and persona history.

This design makes PACEJUDGE evaluable beyond GOOD/BAD accuracy. A scalar judge may call a response "mostly helpful," but it has no required field for the violated target, the broken response component, the defect family, or supporting evidence. PACEJUDGE makes these diagnostic fields explicit, enabling the scenario-level metrics in Sec. 5. A detailed per-target walkthrough of how PACEJUDGE reports the corresponding PACE target, together with the matching scenario-level metric, is provided in Appx. D.

5 Experiments

PACESHOP provides controlled records where persona, compositional, grounding, and actionable-diagnosis failures are measurable; PACEJUDGE specifies the diagnostic fields a judge should return. We therefore ask a simple question: when evaluated on the same records and backbones, do scalar or generic judge protocols recover the diagnostic fields needed for PACE evaluation, or is an explicit structured protocol necessary?

5.1 Experimental Setup

We evaluate judge protocols on PACESHOP records using seven LLM backbones (Opus 4.7, Sonnet 4.6, Sonnet 4.5, Haiku 4.5, Qwen3 32B, GPT-OSS 20B, GPT-OSS 120B). We compare PACEJUDGE with two baseline families on PACESHOP records. The first family consists of *multi-target rubric ablations* that share the PACEJUDGE output schema but vary the judging instruction: G-Eval-style scalar judging (Liu et al., 2023), MT-Bench-style scalar judging (Zheng et al., 2023), persona-blind judging, and a persona-aware generic rubric. The second family consists of *single-target adapted baselines* whose native protocols cover only partial PACE dimensions: ARES-style grounding (Saad-Falcon et al., 2024), PersonaLens-style personalization (Zhao et al., 2025b), and EtaPP-style actionable diagnosis (Hao et al., 2025). Full backbone, prompt, and native-support details are in Appx. E.

5.2 Evaluation Scenarios

Building on the target-to-scenario mapping in Tab. 2, we report one reference scenario and four PACE closure tests. S0 measures broad GOOD/BAD discrimination and is included only to show whether records are broadly judgeable. S1–S4 test persona-source diagnosis, cross-component detection and localization, grounding over product evidence and persona history, and actionable defect-family/field-location prediction, respectively. Tab. 22 links these scenarios to metric columns; full definitions are provided in Appx. E.3.

5.3 Main Results

Tab. 3 tells a simple story: broad GOOD/BAD judging is not where the main difficulty lies. Most

Method	Comparison level	S0 ref. acc.	$S1_P$ PJC	$S2_C$ J-strict	$S3_E$ J-harmonic	$S4_A$ J-strict	PACE avg. S1–S4
Single-target adapted baselines on Sonnet 4.6 (each native judging protocol covers exactly one PACE *target)*							
ARES-style (Saad-Falcon et al., 2024)	native target only	0.69	–	–	0.30	–	–
PersonaLens-style (Zhao et al., 2025b)	native target only	0.72	0.01	–	–	–	–
EtaPP-style (Hao et al., 2025)	native target only	0.81	–	–	–	0.00	–
Frontier control: 7-backbone self-consistency majority vote on the same rubric							
G-Eval (Liu et al., 2023) + 7-vote SC	majority vote	0.91	0.02	0.51	0.30	0.28	0.28
Persona-aware + 7-vote SC	majority vote	0.81	0.08	0.59	0.32	0.25	0.31
PACEJUDGE + 7-vote SC	majority vote	0.92	0.07	0.74	0.82	0.37	0.50
PACE-compatible rubrics on the shared JSON contract (7-backbone mean)							
G-Eval (Liu et al., 2023)	7-backbone mean	0.74	0.03	0.28	0.05	0.05	0.10
Scalar (MT-Bench-style) (Zheng et al., 2023)	7-backbone mean	0.75	0.02	0.34	0.04	0.09	0.12
Persona-blind	7-backbone mean	0.76	–	0.33	0.05	0.09	–
Persona-aware	7-backbone mean	0.75	0.08	0.33	0.10	0.05	0.14
PACEJUDGE	7-backbone mean	0.79	0.07	0.46	0.20	0.11	0.21
Lift: PACEJUDGE mean vs. best PACE-compatible baseline mean	per-target lift	+0.04	−0.00	+0.13	+0.11	+0.02	+0.07
Lift: PACEJUDGE on Sonnet 4.6 vs. best Sonnet 4.6 baseline	per-target lift	+0.10	+0.11	+0.24	+0.19	+0.32	+0.22

Table 4: **Strict paired-counterfactual PACE-closure results.** Each S1–S4 score is a joint metric that requires more than detecting BAD examples: the judge must also satisfy target-specific conditions. These gates prevent an always-BAD judge from receiving high credit through over-flagging. The last two rows summarize the lift from PACEJUDGE over the strongest comparable baseline, showing consistent gains under strict closure metrics at both the seven-backbone mean and best-backbone levels. Metric definitions are in Appx. E.4; native-protocol support and leakage audits for single-target adapted baselines are in Tabs. 19 and 20.

judges/protocols can often decide whether a response is generally valid or flawed, but they break down when the metric asks for the fields needed to debug the failure: which PACE target was violated, which response field failed, and whether the diagnosis is grounded in the provided evidence. PACEJUDGE is strongest on this diagnostic closure setting among multi-target rubric ablations, while single-target adapted baselines leave most columns unsupported because their native protocols cover only one part of PACE. Tab. 4 then asks a stricter question: does the judge simply flag more responses as BAD? Under paired-counterfactual gates that require GOOD-record specificity, correct target/family/location, axis consistency, and evidence-ID containment, PACEJUDGE remains strongest. The improvement therefore reflects better diagnostic closure, not merely more aggressive defect detection.

Finding 1: broad accuracy and stronger backbones do not solve PACE closure. S0 in Tab. 3 shows that scalar and generic judges can perform similarly to PACEJUDGE on broad GOOD/BAD discrimination. However, this binary signal does not translate into PACE closure: a judge may mark a response BAD while still failing to identify whether the failure comes from shopper-context mismatch, cross-component inconsistency, unsupported evidence, or a missing actionable diagnosis. Tab. 4 shows the same pattern in the stricter audit: even frontier self-consistency improves reference accuracy, but generic rubrics still trail PACEJUDGE on the joint S1–S4 closure metrics. This shows that stronger inference alone cannot recover fields that the protocol does not ask the judge to produce.

Finding 2: PACEJUDGE improves the diagnostic fields required by PACE. Across the seven-backbone in Tab. 3, PACEJUDGE achieves the strongest overall PACE average among multi-target rubric ablations. The gain is diagnostic rather than scalar: **P** requires identifying persona conflict as the source of invalidity, **C** requires localizing the failed response component, **E** requires grounding against both product evidence and persona history, and **A** requires defect-family and field-location prediction. Fig. 3 summarizes this S1–S4 closure pattern, omitting S0 because the central claim is not broad classification but task-aligned diagnosis.

Finding 3: protocol design helps, but backbone capability sets the ceiling. The per-backbone breakdowns in Appx. E.5.1, E.5.2, and E.5.3 show a common pattern: PACEJUDGE is most effective when the backbone can follow a structured contract, calibrate GOOD-record specificity, and reason over persona and field-level constraints. Frontier backbones such as Sonnet 4.6 and Opus 4.7 obtain the strongest PACE closure with relatively low false-fire rates, while smaller or open-weight backbones (Qwen3 32B, GPT-OSS 20B) often improve on individual diagnostic columns but remain unstable

Prompt	Persona-swap accuracy by axis			
	budget	hard	brand	attr.
G-Eval	0.24	0.16	0.16	0.12
Persona-aware	0.30	0.25	0.14	0.11
PACEJUDGE (ours)	**0.37**	**0.32**	**0.36**	**0.23**

Table 5: **Persona-source diagnosis by conflict axis, averaged over backbones.** A hit requires predicting PREF_CONFLICT on a counterfactual persona swap. PACEJUDGE is strongest across all axes, showing that personalization failures require source-level diagnosis rather than broad BAD detection. Axis means correspond to S1 in Tab. 3; full per-backbone results are in Appx. E.5.2.

on persona-source recall, grounding, or false-fire control. Thus, PACEJUDGE supplies the missing output contract, but the absolute closure level is still bounded by the backbone's ability to execute that contract.

Finding 4: persona-source diagnosis and component localization expose hidden failures. Tabs. 5 and 6 unpack two failure modes that are easy to hide behind scalar quality scores. Tab. 5 shows that personalization must be evaluated as a source of invalidity: the response is unchanged, but the swapped persona makes it wrong, so credit requires predicting PREF_CONFLICT rather than merely lowering an overall score. Tab. 6 separates detecting a cross-component defect from localizing the response field that must be fixed; this distinction is central to actionable debugging because a judge can know that a response is flawed while still failing to say where the repair is needed.

Finding 5: gains are most informative on low-signal, single-defect cases. The four-tier BAD construction (1-, 2-, 3-, and 4-defect BAD records) lets us test whether judges diagnose individual failures or merely accumulate surface alarms as more defects are added. Fig. 4 shows that the largest protocol gap appears on one-defect records, where only a single response field is wrong and surface signal is weakest. PACEJUDGE remains stable across the tiered difficulty gradient, suggesting that its advantage comes from structured diagnosis rather than simply detecting that heavily corrupted responses look bad.

6 Conclusion

We argued that shopping-assistant evaluation should be treated as a joint PACE problem: responses must be personalized, actionable, compositional, and evidence-grounded. PACESHOP makes this setting measurable through controlled and verifiable records with structured personas, auditable evidence pools, and gold defect family/location labels. Across seven backbones and multiple judging configurations, scalar and generic rubric judges often recognize broad GOOD/BAD quality but fail to recover the diagnostic fields needed for debugging. PACEJUDGE improves these closure metrics without retraining. These results support our central claim: for structured shopping assistants, the key evaluation challenge is not broad quality scoring but producing a task-aligned diagnosis of what failed, where, and why.

Prompt	detect	4-way loc.
G-Eval	0.75	0.70
Persona-aware	0.79	0.64
PACEJUDGE (ours)	**0.93**	**0.72**

Table 6: **Cross-component detection and localization.** Detection flags whether a compositional defect is present; 4-way localization requires identifying the failed response field. PACEJUDGE is strongest on both, showing that compositional failures require field-level diagnosis beyond broad BAD detection. Scores correspond to S2 detect / loc. in Tab. 3; full per-backbone results are in Appx. E.5.3.

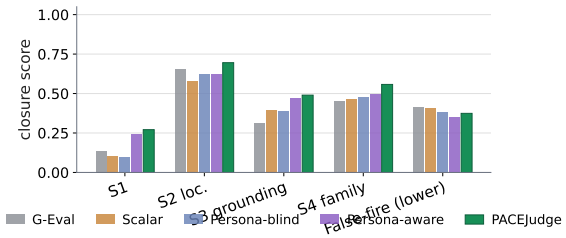

Figure 3: PACE-closure scores averaged over seven backbones. PACEJUDGE is strongest on the S1–S4 metrics that require persona-source diagnosis, cross-component localization, grounding control, and actionable defect localization.

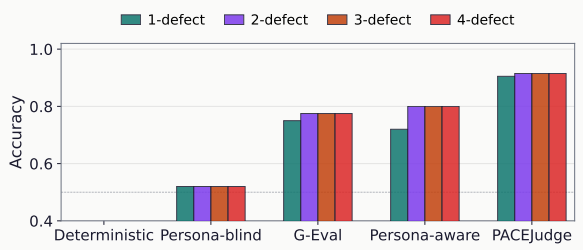

Figure 4: **GOOD/BAD accuracy by defect tier.** BAD records are grouped by the number of injected defects. The 1-defect tier isolates the hardest setting for surface-level detection, while higher tiers expose whether methods mainly benefit from accumulating defects.

Limitations

PACESHOP and PACEJUDGE have several bounded limitations. **(1) Synthetic GOOD/BAD construction.** GOOD responses come from a single generator (Claude Opus 4.6) and BAD responses from rule-based defect injection; the cross-generator and real-assistant slices audit but do not fully reproduce live traffic. **(2) Evidence depth.** Most evidence records carry product-listing fields only; 129/4,460 are review-augmented, bounding how deeply grounding can be audited per claim. **(3) Holdout scope.** Held-out persona bundles cover 345/1,200 personas across 8 families; this is narrower than real-user generalization across demographics, regions, and languages. The dataset is English-only.

Potential Risks

PACESHOP and PACEJUDGE are intended for evaluating structured shopping-assistant responses, not for directly certifying production systems or replacing human review. A main risk is personalization. Although the personas are synthetic or structured benchmark artifacts, personalization evaluation can encourage systems to infer or exploit user traits if transferred carelessly to real deployments. Practical use should therefore avoid sensitive personal attributes, preserve user privacy, and treat persona-conditioned evaluation as a tool for detecting misalignment rather than for maximizing persuasion or user targeting.

References

Luyu Chen, Zeyu Zhang, Haoran Tan, Quanyu Dai, Hao Yang, Zhenhua Dong, and Xu Chen. 2025. Beyond single-point judgment: Distribution alignment for llm-as-a-judge. *arXiv preprint arXiv:2505.12301*.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141.

Joseph Enguehard, Morgane Van Ermengem, Kate Atkinson, Sujeong Cha, Arijit Ghosh Chowdhury, Prashanth Kallur Ramaswamy, Jeremy Roghair, Hannah R Marlowe, Carina Suzana Negreanu, Kitty Boxall, and 1 others. 2025. Lemaj (legal llm-as-a-judge): Bridging legal reasoning and llm evaluation. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 318–337.

Benjamin Feuer, Chiung-Yi Tseng, Astitwa Sarthak Lathe, Oussama Elachqar, and John P Dickerson. 2025. When judgment becomes noise: How design failures in llm judge benchmarks silently undermine validity. *arXiv preprint arXiv:2509.20293*.

Yupu Hao, Pengfei Cao, Zhuoran Jin, Huanxuan Liao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Evaluating personalized tool-augmented llms from the perspectives of personalization and proactivity. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21897–21935.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N Angelopoulos, Trevor Darrell, Narges Norouzi, and 1 others. 2025. Search arena: Analyzing search-augmented llms. *arXiv preprint arXiv:2506.05334*.

Md Awsafur Rahman, Adam Gabrys, Doug Kang, Jingjing Sun, Tian Tan, and Ashwin Chandramouli. 2025. Likebench: Evaluating subjective likability in llms for personalization. *arXiv preprint arXiv:2512.13077*.

Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping queries dataset: A large-scale esci benchmark for improving product search. *arXiv preprint arXiv:2206.06588*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings*

of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7370–7392.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

Haoxin Wang, Xianhan Peng, Huang Cheng, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Jiang Lin. 2025a. Ecom-bench: Can llm agent resolve real-world e-commerce customer support issues? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 276–284.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, and 1 others. 2025b. Opera: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. *arXiv preprint arXiv:2506.05606*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025a. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025b. Personalens: A benchmark for personalization evaluation in conversational ai assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18023–18055.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Existing Benchmark Datasets

This section documents how the prior resources cited in Tab. 1 relate to the PACEShop evaluation setting. The goal is not to argue that prior personalization, shopping, RAG, or LLM-as-judge benchmarks are weak. Rather, they were designed for different native tasks: user-conditioned generation, dialogue preference following, memory retrieval, task success, pairwise search comparison, or RAG faithfulness. These protocols measure important pieces of the problem, but they do not define the same closure target as PACEShop: given a query, persona, four-field shopping response, and evidence pool, the evaluator must diagnose whether the response is personalized, cross-field consistent, evidence-grounded, and actionable for debugging.

Tab. 7 therefore lists each resource by its *native evaluation protocol*. This distinction matters for the experiment tables. A benchmark may be strongly related to one PACE target while still lacking the output fields needed for the S1–S4 closure metrics. For example, personalization benchmarks can evaluate whether a response follows user preferences, but usually do not define a persona-swap defect family and response-field location. RAG evaluators can measure faithfulness to retrieved context, but usually do not check whether a response invents shopper history or whether the judge itself cites evidence IDs outside the evidence pool. Shopping task benchmarks can evaluate realistic user trajectories, but they do not provide gold defect families and locations for four-field shopping-assistant responses.

For this reason, the main experiment table (Tab. 3) does not force every prior resource into every PACEShop scenario. Instead, unsupported native judging protocols are marked as "–". We implement honest single-target adapted baselines only when the native judging protocol gives a defensible mapping to one PACE target: **ARES-style** for evidence grounding (E), **PersonaLens-style** for personalization (P), and **EtaPP-style** for actionable key-point diagnosis (A). The other resources remain comparison points for problem setting and protocol coverage, but are not converted into artificial all-target baselines. This keeps the appendix comparison auditable: "–" means the native protocol does not define the scored PACEShop output field, not that the resource is unimportant.

Native resource	**Native evaluation protocol**
LaMP (Salemi et al., 2024)	Per-task accuracy or macro-F1 (classification tasks) and ROUGE-1/ROUGE-L (generation tasks) on user-conditioned generation.
PersonaLens (Zhao et al., 2025b)	LLM-judge agents score an assistant turn on three axes: personalization, response quality, and task success.
PrefEval (Zhao et al., 2025a)	Preference-following accuracy (generation and classification variants) on long-context dialogues with stated user preferences.
PersonaMem (Jiang et al., 2025)	Multiple-choice accuracy: pick the reply matching the user's evolving profile across ≤ 60 chat sessions.
LikeBench (Rahman et al., 2025)	Per-turn likability score from a simulated user, decomposed into seven dimensions (emotion, formality, humor, callback, etc.).
EtaPP (Hao et al., 2025)	Key-point-based LLM-as-judge against human-annotated key points on personalized tool-call trajectories.
Personalized judge (Dong et al., 2024)	Agreement with human ground truth on a binary preference task (overall and high-certainty subset), with verbal confidence.
ECom-Bench (Wang et al., 2025a)	Task-success trajectories in e-commerce support; agents simulate hundreds of personas across realistic commerce tasks.
Search Arena (Miroyan et al., 2025)	Pairwise human preference votes on full traces from two search-augmented chat systems (Bradley–Terry style).
ARES (Saad-Falcon et al., 2024)	Three PPI-corrected fine-tuned LM judges scoring context relevance, answer faithfulness, and answer relevance for RAG.

Table 7: Native judging protocols of the prior personalization, evidence-grounding, and judging resources cited in the main text. None of these protocols defines a closure metric for the four PACE targets (Personalized, Compositional, Evidence-grounded, Actionable); they evaluate user-conditioned generation, MCQ over user history, RAG faithfulness, or pairwise human votes, but not the joint persona–compositional–grounded–actionable diagnosis that PACESHOP requires. We additionally implement honest single-target adapted baselines based on ARES (E target), PersonaLens (P target), and EtaPP (A target) and report them in the main experiment table (Table 3); the remaining seven resources are not run on PACESHOP because their native judging protocols cannot be honestly mapped to the four-target diagnosis without forcing inapplicable scores.

B PACESHOP Dataset Construction

This appendix documents the construction-side material that supports PACESHOP. We first describe the full construction pipeline and the validity layers (Secs. B.2–B.5); we close with concrete records, JSON input examples, the response-generation prompt, holdout-persona statistics, and source-side coverage tables that make the benchmark auditable (Secs. B.6–B.3).

B.1 PACESHOP Dataset Statistics

Tab. 8 summarizes the benchmark statistics, evidence coverage, defect structure, and validation assets.

Block	Statistic	Value
SOURCE	Backbone queries (public)	97,227
	Evidence-backed queries ($\geq$ 3 products)	1,132
	Persona pool (constraints+history+brand)	1,200
	Query–persona pairs	4,525
EVIDENCE	Evidence records (11 domains)	4,460
	Review-augmented records	129
	Median evidence IDs / response	10
DATASET	GOOD / BAD / total	4,525 / 18,100 / 22,625
	Train / dev / test	15,825 / 3,400 / 3,400
	Stealth-defect fraction	35.1%
DEFECTS	Families† $\times$ field groups	7 $\times$ 4
	Multi-defect tiers (1/2/3/4)	4,525 each
	Gold {family, location} on every BAD	✓
TARGETS COVERED	**P**ersonalized (persona-swap slice)	✓
	Actionable (gold family+location)	✓
	Compositional (cross-field defects)	✓
	Evidence-grounded (product + history)	✓
VALIDITY CHECK	L1 deterministic checks (12/record)	99.3% pass
	L2 evidence provenance flag	4,331 / 129
	L3 human-calibration package	250 items
	L4 cross-generator slice (3 models)	50 shared
	L5 real-assistant slice	100 outputs

†PREF_CONFLICT, INTENT_DRIFT, XCOMP_MISMATCH, REDUNDANCY, EVIDENCE_MISMATCH, OVER_PERSONALIZATION_HALLUCINATION, UNSUPPORTED_CLAIM. Field groups: overview, categories, related queries, attributions.

Table 8: Summary of the PACESHOP benchmark. The "Targets covered" block names which of the four PACE targets each design choice supports; the "Validity stack" block enumerates the five verification layers that make the construction auditable.

B.2 Construction Pipeline Workflow

PACESHOP is built from public source artefacts in a controlled, auditable pipeline.

Persona construction. We construct personas as structured shopping contexts rather than free-form user descriptions. The released persona pool contains 1,200 personas, including public seed personas and behavior-seeded synthetic personas normalized under a shared schema. Each persona specifies stable shopper information, such as household type, budget, quality sensitivity, urgency, constraints, preferences, and brand affinities, together with concrete shopping behavior in the form of purchase history and recent searches. This design supports two kinds of evaluation. First, shopper constraints and preferences make persona-conditioned correctness testable: a response can be valid for one shopper and invalid for another. Second, purchase and search histories make persona-history grounding auditable: a judge can check whether a response faithfully refers to the shopper's actual history or invents unsupported behavior.

Shopping-behavior and evidence collection. We begin from 97,227 normalized US shopping queries and 4,460 public evidence records across 11 shopping domains. Queries are tagged by coarse auditing metadata, including retail vertical and shopping mission, while evidence records expose product titles, brands, metadata, and available review snippets. We retain 1,132 queries whose evidence pool contains at least three products. This evidence-backed filtering is deliberate: PACEShop asks whether the same shopping intent can support different valid responses under different personas, which requires enough product alternatives for shopper constraints and preferences to matter. Each retained query is paired with four distinct personas using a coverage-balanced assignment procedure that minimizes persona-taxonomy reuse and prevents the same persona from repeating on a query. This yields 4,525 query–persona pairs after validation drops. The resulting repeated-measures design holds the shopping query and evidence pool fixed while varying shopper context, allowing personalization failures to be isolated from changes in shopping intent.

The full four-stage workflow.

1. **Query catalog.** Normalize 97,227 US ESCI queries; tag retail vertical (11 categories) and shopping mission (6 categories). Detailed breakdown can be found in Appx. B.3.
2. **Evidence catalog.** Build 4,460 public evidence records across 11 shopping domains; augment 129 with review snippets.
3. **Persona pool.** Merge 200 public v1 personas with 1,000 behavior-seeded synthetic personas; normalize under one structured schema (household, budget, quality, urgency, constraints, preferences, brand affinities, history).
4. **Coverage-balanced assignment.** Retain 1,132 queries with $\geq$3 evidence candidates; reserve

Stage	Motivation	Operation	PACE role
Source profiling	Ensure broad, auditable coverage over shopping intents, personas, and evidence.	Tag queries, personas, and evidence records by retail vertical, mission, household, budget, constraints, preferences, and history.	Creates controlled coverage for P and E.
Evidence-backed retention and persona assignment	Make persona differences meaningful while holding query intent fixed.	Retain queries with at least three evidence-backed products; pair each query with four distinct personas under coverage-balanced assignment.	Supports persona-swap and repeated-measures evaluation for P.
GOOD response generation	Create valid structured responses under the target response contract.	Generate overview, categories, related queries, and attributions; normalize and validate schema and evidence IDs.	Defines valid multi-field objects for C and E.
BAD defect injection	Create known, localizable failures rather than relying on vague quality differences.	Inject one to four controlled defects from a 7-family $\times$ 4-field taxonomy.	Provides gold defect family/location for A and scenario-specific tests for P/C/E.

Table 9: PACESHOP construction pipeline. Each stage is designed to make one or more PACE properties observable and verifiable.

345 holdout personas across 8 bundles for dev/test; assign 4 personas per retained query.

5. **GOOD generation.** One structured response per assignment, Claude Opus 4.6, constrained JSON output.
6. **Normalization.** Schema validation and deterministic post-processing.
7. **BAD construction.** Four tiers (1/2/3/4 defects) per GOOD, defects drawn from distinct field groups.
8. **Packaging.** Export packaged / explicit / clean JSONL views plus 250-item human-calibration package, 50-input cross-generator slice, 100-output real-assistant slice.

The four substantive stages of this workflow are summarized in Tab. 9, which makes explicit how each stage's design choice advances one or more PACE properties.

B.3 PACESHOP Coverage Details

The released persona pool, evidence catalog, and query catalog are documented below with representative examples. The query catalog covers 11 retail verticals and 6 shopping missions; the persona pool covers 12 household types and 8 holdout bundle families; the evidence catalog spans 11 shopping domains and includes 129 review-augmented records.

Retail vertical	Backbone	Retained
Toys & Games	17,279	235
Electronics	8,044	186
Tools & Automotive	13,429	157
Home & Kitchen	12,439	144
Apparel & Accessories	13,874	104
Pet & Baby	6,859	76
General Merchandise	7,216	66
Sports & Outdoors	6,196	62
Office & School	5,009	46
Beauty & Health	2,901	33
Books & Media	3,981	23
Total	**97,227**	**1,132**

Table 10: Query distribution across 11 retail verticals. *Backbone* counts all unique US normalized ESCI queries; *Retained* counts the 1,132 evidence-backed queries that pass the evidence-availability filter ($\geq$3 evidence-backed candidates).

Shopping mission type	Backbone	Retained
Broad Browse	32,889	312
Replenishment & Repeat	20,166	241
Problem Solution	14,533	217
Compatibility & Replacement	9,807	144
Gift & Event	12,386	129
Feature Constrained	7,446	89
Total	**97,227**	**1,132**

Table 11: Query distribution across 6 shopping mission types. Mission type is assigned by intent-pattern matching against query text and top product titles/brands.

B.3.1 Query Coverage

The public backbone contains 97,227 unique US-locale normalized ESCI queries categorized along two dimensions.

Retail verticals (11 categories). These follow standard e-commerce product departments (e.g., Amazon Browse Nodes). Each query is assigned to one vertical by a keyword-hit counting rule (`_pick_rule()` in the released code): the assignment function concatenates the normalized query text with the top candidate product titles and brands, then counts the number of keyword hits for each vertical's keyword list. The vertical with the most hits wins; ties are broken by list order, and if no keyword matches the query falls back to General Merchandise. Representative keyword→vertical mappings include:

- "shirt", "shoes", "backpack" → Apparel & Accessories
- "usb", "charger", "headphones" → Electronics
- "kitchen", "vacuum", "air fryer" → Home & Kitchen
- "toy", "lego", "puzzle" → Toys & Games
- "dog", "baby", "stroller" → Pet & Baby
- "book", "vinyl", "dvd" → Books & Media

The full keyword lists (10–19 terms per vertical) are provided in the released `v2_taxonomy.py`. The lists are intentionally broad rather than exhaustive: common category-indicative terms are sufficient for coarse coverage reporting, since the taxonomy serves as an auditing tool rather than a fine-grained classifier.

Shopping mission types (6 categories). These capture the shopper's task framing and follow well-known shopping-behavior distinctions in the e-commerce literature: undirected browsing, repeat/replenishment purchases, problem-driven search, gift/event shopping, compatibility/replacement needs, and feature-constrained filtering. Assignment uses the same `_pick_rule()` keyword-hit counting mechanism as retail verticals, applied to intent-indicative keyword lists. Representative keyword→mission mappings include:

- "gift", "birthday", "wedding" → Gift & Event
- "best", "quiet", "waterproof" → Problem Solution
- "wireless", "organic", "compact" → Feature Constrained
- "compatible", "replacement", "adapter" → Compatibility & Replacement
- "bulk", "refill", "pods" → Replenishment & Repeat

The fallback category (no keyword hits) is Broad Browse. Full keyword lists (7–12 terms per mission type) are in `v2_taxonomy.py`.

Table 10 shows the backbone distribution across the 11 retail verticals together with the number of evidence-backed queries retained after evidence filtering. Table 11 shows the corresponding distribution across the 6 shopping mission types.

The retention rate varies across verticals and missions because it depends on evidence availability in the public ESCI product metadata rather than on intentional stratification. For example, Beauty & Health retains 33 queries (1.14%) while Toys & Games retains 235 (1.36%), reflecting differences in product-metadata completeness across categories.

Axis	Released coverage summary
Household	couple (117); creator hobbyist (38); home office (74); parents school age (121); parents young children (132); pet owner (97); roommates (81); senior caregiver (81); single professional (136); small space (104); student (88); travel outdoor (131)
Budget	budget conscious (380); premium (401); value balanced (419)
Quality	balanced (406); functional (416); high (378)
Urgency	fast (380); standard (420); urgent (400)
Constraints	allergen ingredient (406); compatibility (162); durability (135); maintenance noise (334); material avoidance (514); safety age (148); size space (206)
Preferences	aesthetics (467); convenience (665); giftability (224); performance (282); portability (619); premium finish (314); sustainability (672)
History	moderate (566); rich (588); sparse (46)
Holdout flags	eco-conscious materials (115); sensitive skin (115); smart-home compatibility priority (115)

Table 12: Released persona coverage summary used in PACESHOP. Counts are over the full 1,200-persona pool, while the rows below show representative released personas rather than the full catalog.

B.3.2 Persona Coverage

The full released persona pool contains 1,200 personas. Table 12 summarizes the aggregate coverage across the major taxonomy axes, including household type, budget, quality preference, urgency, constraints, shopping preferences, behavioral-history density, and holdout attributes. Table 14 then provides a compact set of representative released personas, selected to illustrate the major coverage patterns discussed in the main text without listing the full persona catalog.

B.3.3 Evidence Coverage

This subsection expands the evidence side of PACESHOP in two complementary views. Tab. 13 summarizes the domain-level coverage of the released evidence catalog, while Tab. 15 shows representative evidence records from each domain so readers can inspect the concrete product information available to the response generator and judge. The full released evidence catalog contains 4,460 public evidence records spanning 11 shopping domains. Each record exposes human-readable product information extracted from public ESCI product listings: product title, brand, color, bullet points, and description. Of these, 4,331 records (97.1%) are *product-listing only*, grounded solely in these catalog fields. The remaining 129 records (2.9%) are *review-augmented*, carrying the same product-listing fields plus public customer review snippets that provide richer grounding signals (e.g., real-user opinions on durability, fit, or usability). Every record carries an explicit provenance flag so that downstream judges and analyses can condition on evidence strength. This design keeps the grounding signal lighter than fully review-grounded RAG benchmarks, but more transparent than a generic "grounded" claim because evidence quality is measurable per record rather than assumed uniform. Table 13 gives the aggregate domain distribution of the evidence catalog, while Table 15 provides a representative released evidence record from each domain. Together, the two tables show both the breadth of evidence coverage and the concrete product-level information available in each per-query evidence pool.

Shopping domain	Records	% of total
General Merchandise	1,232	27.6%
Toys & Games	623	14.0%
Apparel & Accessories	610	13.7%
Home & Kitchen	418	9.4%
Tools & Automotive	375	8.4%
Electronics	335	7.5%
Pet & Baby	267	6.0%
Sports & Outdoors	225	5.0%
Office & School	191	4.3%
Books & Media	104	2.3%
Beauty & Health	80	1.8%
Total	**4,460**	**100%**

Table 13: Evidence record distribution across 11 shopping domains. Of these, 129 records (2.9%) are review-augmented; the remainder are grounded in product-listing fields only (title, brand, color, bullets, description).

Persona ID	Coverage highlights	Representative shopper sketch
v2-persona-0276	household: parents school age; budget: value balanced; quality: high; urgency: urgent; constraint: durability/maintenance noise; preference: aesthetics/convenience; history: moderate; holdout: eco-conscious materials	family with school-age kids; mid-range, quality first, needs it urgently; constraints: quiet operation required, durable enough for daily use; attrs: eco-conscious materials, travel friendly
v2-persona-0208	household: home office; budget: premium; quality: balanced; urgency: fast; constraint: allergen ingredient/material avoidance; preference: convenience/premium finish; history: rich; holdout: smart-home compatibility priority	home office household; premium, balanced, needs it this week; constraints: fragrance-free only, kid-safe materials only; attrs: smart-home compatibility priority, ergonomic support
v2-persona-1130	household: senior caregiver; budget: budget conscious; quality: functional; urgency: standard; constraint: compatibility/size space; preference: convenience/performance; history: moderate	senior caregiver household; budget, value oriented, standard shipping; constraints: must fit existing device compatibility, small-space friendly; attrs: organic cotton, ergonomic support
persona-0027	household: single professional; budget: premium; quality: high; urgency: urgent; constraint: material avoidance/size space; preference: performance/portability; history: sparse; holdout: sensitive skin	single professional; premium, quality first, needs it urgently; constraints: small-space friendly, avoid wool; attrs: sensitive skin, compact storage
persona-0030	household: travel outdoor; budget: value balanced; quality: high; urgency: standard; constraint: allergen ingredient/material avoidance; preference: convenience/performance; history: moderate; holdout: sensitive skin	active outdoor household; mid-range, quality first, standard shipping; constraints: fragrance-free only, latex-free materials; attrs: sensitive skin, eco-conscious materials
persona-0009	household: pet owner; budget: value balanced; quality: functional; urgency: standard; constraint: allergen ingredient; preference: convenience/premium finish; history: rich; holdout: sensitive skin	active outdoor household; mid-range, value oriented, standard shipping; constraints: fragrance-free only; attrs: sensitive skin, low-maintenance finishes
v2-persona-0289	household: student; budget: premium; quality: balanced; urgency: standard; constraint: allergen ingredient; preference: aesthetics/convenience; history: rich; holdout: smart-home compatibility priority	student; premium, balanced, standard shipping; constraints: avoid peanuts in ingredients; attrs: smart-home compatibility priority, compact storage
v2-persona-0467	household: parents young children; budget: budget conscious; quality: functional; urgency: standard; constraint: allergen ingredient; preference: giftability/performance; history: moderate; holdout: sensitive skin	family with toddler; budget, value oriented, standard shipping; constraints: avoid peanuts in ingredients; attrs: sensitive skin, compact storage
persona-0022	household: couple; budget: budget conscious; quality: balanced; urgency: fast; constraint: allergen ingredient/material avoidance; preference: convenience/portability; history: rich; holdout: eco-conscious materials	couple in a smart home; budget, balanced, needs it this week; constraints: avoid peanuts in ingredients, latex-free materials; attrs: eco-conscious materials, travel friendly
v2-persona-0234	household: small space; budget: value balanced; quality: balanced; urgency: standard; constraint: compatibility/size space; preference: portability/sustainability; history: moderate; holdout: eco-conscious materials	travel and outdoor household; mid-range, balanced, standard shipping; constraints: small-space friendly, must fit existing device compatibility; attrs: eco-conscious materials, travel friendly
v2-persona-0226	household: roommates; budget: budget conscious; quality: functional; urgency: fast; constraint: material avoidance; preference: aesthetics/convenience; history: rich; holdout: smart-home compatibility priority	roommates; budget, value oriented, needs it this week; constraints: avoid wool, latex-free materials; attrs: smart-home compatibility priority, compact storage
v2-persona-0223	household: creator hobbyist; budget: budget conscious; quality: high; urgency: urgent; constraint: maintenance noise/material avoidance; preference: giftability/performance; history: moderate; holdout: smart-home compatibility priority	creator and hobbyist household; budget, quality first, needs it urgently; constraints: quiet operation required, latex-free materials; attrs: smart-home compatibility priority, eco-conscious materials

Table 14: Representative released personas covering the major taxonomy values emphasized in the paper.

Domain	Count	Evidence ID	Brand	Representative title
apparel accessories	610	B07T3DJ4TF	DC Comics	Youth Boys' Classic Batman Costume Hoodie Sweatshirt-Small
beauty health	80	B07YNJGVX8	Respected Roots	Respected Roots Shaving Cream All Natural Shave Cream for Men and Body - Travel Shaving Cream with Tree Tea Oil to Shave Sensitive Skin, Mens Pre-Shave lotion.
books media	104	0679767398	Vintage	South of the Border, West of the Sun: A Novel
electronics	335	B07HK6K5CS	LENRUE	Portable Bluetooth Speakers,Dual-Driver Wireless Speaker with Surround Stereo Sound and More Bass,for iPhone and Samsung Android ... (Black)
general merchandise	1232	B00MG5FZWA	Embroidex	Embroidex Sewing Kit for Home, Travel & Emergencies - Filled with Quality Notions Scissor & Thread - Great Gift
home kitchen	418	B07SNPJSB4	SHINEURI	SHINEURI 2 1/2 qt Copper Saucepan with Lid, Mini Saute Pan with Stainless Steel Handle - Cooking for Soup, Stew, Sauce, Pasta & Reheat Food, Compatible for Induction, Gas, Electric & Stovetops
office school	191	1939814456	Handwriting Without Tears	Learning Without Tears - My Printing Book Student Workbook, Current Edition - Handwriting Without Tears Series - 1st Grade Writing Book - Letters, Language Arts Lessons - for School or Home Use
pet baby	267	B07WV78D6S	M JJYPET	M JJYPET Interactive Cat Toys,Rechargeable 3 in 1 Cat Red Dot Cat Kitten Dog Toy
sports outdoors	225	B00339C3PK	Coleman	Coleman Folding Camp Chair | Woodsman II Portable Outdoor Chair, 17" x 17.5"
tools automotive	375	B0000AXFST	Northern Tool and Equipment	Fast Framer Universal Storage Shed Framing Kit
toys games	623	B0742MSVH8	The Treemendous Ornament Decorator	TreeMendous Christmas Tree Ornament Decorating Kit for Kids Ages 3 and up - Top Rated Craft Activity Game, Holiday Toy DIY Ornament Maker

Table 15: Representative evidence view for PACESHOP. Each row reports one domain count and one representative released evidence record from that domain.

B.4 Validity Check

We do not treat the dataset as self-validating. Five independent layers make the construction auditable: (i) a 12-check deterministic record-level validator that every record passes through, (ii) explicit evidence-provenance flags (metadata vs. review-augmented), (iii) a prepared 250-item human-calibration package, (iv) a cross-generator slice on 50 shared inputs to confirm that GOOD generation does not overfit a single backbone, and (v) a normalized real-assistant slice of 100 external outputs that map cleanly into the same response contract. Of the 4,525 released GOOD records, 99.3% pass all 12 deterministic checks; the remaining 0.7% fail only a case-sensitive category-duplicate check and are flagged but retained. Tab. 16 maps each validity layer to the specific construction risk it mitigates.

B.5 Defect Taxonomy

BAD records are constructed from seven defect families distributed across four response-field groups (overview, categories, related queries, evidence attributions). Each defect family is paired with the response field whose contract it violates, so localization in the main experiment table is well posed. Tab. 17 lists the seven families, their target field groups, and the operational definition used to inject each defect during BAD construction.

B.6 An Example Record of PACEJUDGE Dataset

Benchmark objects. In PACESHOP, the candidate response is instantiated as a structured object

$$y = (y^{\text{ov}}, y^{\text{cat}}, y^{\text{rq}}, y^{\text{attr}}),$$

where y^{ov} is the overview, y^{cat} is the product-category list, y^{rq} is the related-query list, and y^{attr} is the evidence list. The persona p contains shopper context, including household profile, budget, quality sensitivity, urgency, constraints, preferences, brand affinities, purchase history, recent searches, and a short context summary. The evidence pool E contains product-level evidence snippets such as titles, brands, bullets, descriptions, and, when available, review snippets; each snippet has an evidence ID so that attribution claims can be checked against the released pool. For paper-facing presentation, hard constraints and soft preferences are

Addressed dimensions	Mitigation in benchmark construction	Current measurable support
Synthetic-on-synthetic circularity	Keep the main 10K release public and fully auditable, then add manual-calibration and transfer assets rather than treating synthetic data as self-validating.	Prepared stratified **250**-item human-calibration package with **110** GOOD and **140** BAD examples; BAD slice balanced at **20** items for each of the 7 defect families.
Thin grounding signal	Expose evidence strength explicitly and augment metadata with review snippets wherever public coverage exists instead of implying uniformly rich grounding.	**129** review-augmented evidence records, **254** review snippets, and exact metadata-vs-review provenance on every released record.
Template-like persona realism	Distinguish raw-v1 personas from behavior-seeded synthetic personas and reserve held-out persona bundles for evaluation.	Persona pool of **1200**: **200** raw-v1 plus **1000** behavior-seeded synthetic personas; **345** hold-outs across **8** bundles.
Rigid schema versus ecological validity	Treat the fixed schema as a judge-facing normalization layer and add a relaxed-schema import path for external outputs.	**100** normalized real-assistant outputs stored with raw originals preserved; persona claims withheld for this slice when persona context is unavailable.
Lack of transfer beyond Claude-generated outputs	Add shared-input transfer slices rather than evaluating only on benchmark-generation outputs from one model family.	**50** shared held-out inputs across **3** generator families (*completed*); **100** additional real-assistant outputs for ecological-validity checks.

Table 16: Threats and mitigations for benchmark validity. The key design choice is to make each concern measurable through released artifacts rather than acknowledging it only qualitatively.

merged into a single preference view, while the released JSON preserves the original fields for reproducibility. Fig. 5 in Appx. B.6 illustrates one concrete PACEShop record; full persona and evidence JSON examples are provided in the same appendix section.

Worked Example of a Dataset Record. The boxed schema below summarizes the abstract PACESHOP record schema: every record pairs a query with a structured persona, an auditable evidence pool, and a four-field shopping response, and BAD records additionally expose gold defect family and location labels. Fig. 5 then shows a single concrete record end to end: the persona-conditioned query on the left, the structured four-field response on the right. Correctness on this record requires joint judgements – persona consis-

Defect family	Component group	Gold location	Why it matters
PREF_CONFLICT	category	product categories	Introduces an option that directly violates persona constraints or preferences.
INTENT_DRIFT	category	product categories	Shifts the answer away from the original shopping intent while remaining superficially plausible.
XCOMP_MISMATCH	rqList	related queries	Breaks consistency between the overview, categories, and refinement suggestions.
REDUNDANCY	rqList	related queries	Produces repetitive or low-diversity refinements that hurt utility.
EVIDENCE_MISMATCH	attribution	evidence attributions	Cites evidence that does not support the claims made in the response.
OVER_PERS_HALLUC.	overview	personalized overview	Adds user-specific claims that are unsupported by the available evidence.
UNSUPPORTED_CLAIM	overview	personalized overview	Adds a generic but unverifiable claim that lacks support in the evidence store.
Multi-defect tier design (defects within the same component group conflict and are never combined):			
Single-defect	—	1 location	7 single-defect variants, one per family.
Double-defect	—	2 locations	18 valid cross-group pairs (e.g., category + attribution).
Triple-defect	—	3 locations	20 valid cross-group triples.
Quad-defect	—	4 locations	4 curated quads covering all 4 component groups.

Table 17: Release-level defect taxonomy and multi-tier difficulty design used in the full PACESHOP benchmark. Each BAD example preserves the original query and persona while injecting one to four defects from distinct component groups. The tiered design creates a measurable difficulty gradient—single-defect variants test basic fault detection, while quad-defect variants require simultaneous identification of failures across all four response components.

Input

Query: *tactical duffle gray*

Persona (structured JSON object):

profile: household = family with school-age kids

preferences: avoid peanuts; quiet operation; easy to clean; durable for daily use

brand_likes: [Ecco Press, Lansinoh]

shopping: budget = mid-range; quality = balanced; shipping = standard

history: purchases = [Barack Before Obama. . . , Lansinoh Storage Bags. . .]

recent_searches = [. . . alternatives, . . . upgrades]

context: "family with school-age kids; mid-range budget; balanced quality; recent purchases include Barack Before Obama, Lansinoh Storage Bags."

Output

Overview: "For a durable, travel-friendly gray tactical duffle that handles daily family use, explore **Tactical**, **Gray Gym Bags**, **Rolling Duffles**, **Travel Duffles**, and **Range Bags**. . . "

Category (4–5): Tactical; Gray Gym Bags; Rolling Duffles; Travel Duffles; Range Bags

Related queries (5):
tactical duffle with MOLLE webbing;
gray duffel bag water resistant;
wheeled tactical duffel carry-on;
packable lightweight gray duffle;
tactical range bag gunmetal grey

Attributes (8 cited products):
Nexpak 26" Tactical Duffle Military Molle;
East West U.S.A Tactical Outdoor Duffel;
Under Armour Contain 4.0 Duffle Bag;
Gonex 80L Packable Travel Duffle;
Samsonite Andante 2 Wheeled Rolling Duffel;
OSAGE RIVER Tactical Range Bag;
Amazon Basics Large Travel Duffel;
Vertx Contingency Duffel 85L

Figure 5: Worked example of a PACESHOP record (`full-good-00000`). Each example pairs a shopping query with a structured persona, an auditable evidence pool, and a four-field shopping response. The input pairs a shopping query with a structured persona object; the output is a heterogeneous response with four evaluation-facing components. Correctness depends jointly on the persona, all components, and the cited evidence. Attributes are shown as short product descriptions; the underlying JSON carries evidence IDs. BAD records additionally expose gold defect family and location labels, enabling diagnostic evaluation rather than only scalar scoring.

tency, cross-field coherence, evidence support, and schema validity – which the four PACE targets operationalize.

B.7 Concrete Persona-JSON Input Example

The [PERSONA_JSON] slot of the generation prompt (Sec. B.9) receives the full persona record, including household profile, preferences (merged hard + soft), brand affinities, shopping context, purchase history, recent searches, and a natural-language context summary. Below is a representative example (truncated for space; the published JSON preserves hard_constraints and soft_preferences as separate lists for reproducibility):

```
{
  "persona_id": "persona-0045",
  "profile": {
    "household": "family with school-age kids",
    "demographics": "balances durability and value"
  },
  "preferences": {
    "hard_constraints": ["avoid peanuts in ingredients",
                         "quiet operation required"],
    "soft_preferences": ["easy to clean",
                         "durable enough for daily use"],
    "brand_likes": ["Ecco Press", "Lansinoh"],
    "brand_avoids": [],
    "attribute_preferences": ["travel friendly",
                              "low-maintenance finishes"]
  },
  "shopping_context": {
    "budget": "mid-range",
    "quality_sensitivity": "balanced",
    "shipping_urgency": "standard shipping"
  },
  "history": {
    "purchase_history": [
      {"asin": "0063028743",
       "product_name": "Barack Before Obama: Life Before
                        the Presidency",
       "brand": "Ecco Press"},
      {"asin": "B006XISCNA",
       "product_name": "Lansinoh Breastmilk Storage Bags,
                        100 Count",
       "brand": "Lansinoh"}
    ],
    "recent_searches": [
      "Barack Before Obama alternatives",
      "Lansinoh upgrades"
    ]
  },
  "context_summary": "This shopper is part of a family with
    school-age kids who balances durability and value on a
    mid-range budget. They prioritize travel-friendly,
    low-maintenance products..."
}
```

B.8 Concrete Evidence-JSON Input Example

The [EVIDENCE_JSON] slot receives up to 24 evidence snippets drawn from the public evidence catalog. Each snippet carries a product identifier, title, brand, a unique evidence ID, the text snippet, and the source field. Below is a representative excerpt; paper-facing tables show these as short product descriptions ("Nexpak 26" Tactical Duffle") rather than raw IDs.


```
[
  {"canonical_product_id": "B079YYGMR3",
   "title": "26\" Tactical Duffle Military Molle Gear
             Shoulder Strap Range Bag TF126 GMG
             Gunmetal Grey",
   "brand": "Nexpak",
   "evidence_id": "M:B079YYGMR3:0",
   "snippet": "26\" Tactical Duffle Military Molle Gear
               Shoulder Strap Range Bag TF126 GMG
               Gunmetal Grey",
   "source_field": "product_title"},
  {"canonical_product_id": "B079YYGMR3",
   "title": "26\" Tactical Duffle Military Molle Gear ...",
   "brand": "Nexpak",
   "evidence_id": "M:B079YYGMR3:1",
   "snippet": "Nexpak",
   "source_field": "product_brand"},
  {"canonical_product_id": "B077KKMFFN",
   "title": "Under Armour Adult Contain 4.0 Duffle Bag,
             Graphite Medium Heat (040)/Black,
             One Size Fits All",
   "brand": "Under Armour",
   "evidence_id": "M:B077KKMFFN:0",
   "snippet": "Under Armour Adult Contain 4.0 Duffle Bag,
               Graphite Medium Heat (040)/Black,
               One Size Fits All",
   "source_field": "product_title"}
]
```

B.9 PACESHOP Benchmark record Generation Prompt

PACESHOP construction has two stages, and only the first stage uses an LLM. *GOOD records* are generated by a fixed LLM (Claude Opus 4.6) under the constrained response-generation prompt reproduced below. *BAD records* are then derived from each accepted GOOD record by deterministic rule-based defect injection in Python rather than by any second LLM prompt; this design choice is what guarantees that every BAD record carries an auditable gold defect family and gold defect location label by construction. We document both stages in this subsection so the appendix is self-contained.

GOOD generation: LLM under a constrained JSON contract. For each query–persona pair, the generator is invoked once with a constrained JSON contract. The prompt enforces the four-field response schema (overview, bolded category list, related queries, evidence attributions), the cardinality and length budgets used in main-text §3.2, and a strict no-extras rule on the list fields. Crucially, the generator is never told what defect families exist: it only sees the persona and the candidate evidence pool, so accepted GOOD responses are contract-valid anchors rather than negative examples for the judge to learn against. The prompt's slot syntax (`[ASSIGNMENT_ID]`, `[QUERY]`, `[PERSONA_JSON]`, `[GROUP_JSON]`, `[EVIDENCE_JSON]`) matches the released runner.

```
You are generating a PACEShop V2 benchmark response.
Return JSON only with keys: overview_text, bold_categories,
productCategoryList, category_search_queries, rqList, attribution_list.

Requirements:
- Overview <= 350 chars.
- Include exactly 4 or 5 bold categories using **...** in overview_text.
- productCategoryList must be an array of plain strings aligned with
  bold categories.
- Each category label <= 15 chars.
- rqList must contain exactly 5 short related shopping refinements.
- attribution_list must be an array of plain evidence ID strings only.
- category_search_queries must be an array of plain strings aligned 1:1
  with productCategoryList.
- Do not return products, scores, or citation objects inside list fields.
- Keep the response personalized to the provided persona and compatible
  with the query intent.

Assignment: [ASSIGNMENT_ID]
Query: [QUERY]
Persona JSON: [PERSONA_JSON]
Support groups: [GROUP_JSON]
Candidate evidence: [EVIDENCE_JSON]
```

Generated GOOD candidates pass through deterministic normalization and the L1 12-check validator (§B.4) before any BAD variant is derived; candidates that fail hard schema, evidence-ID, or XML round-trip checks are rejected.

BAD construction: deterministic rule-based defect injection. We deliberately do *not* use an LLM prompt to generate BAD records. Each BAD record is produced by a deterministic Python edit on a validated GOOD record drawn from the seven-family defect taxonomy in Tab. 17 (Sec. B.5). For every defect family $d \in \mathcal{D}$ the edit is fixed in advance, so the gold defect family d and gold defect location $r \in \mathcal{R}$ are known by construction rather than being predicted by another model:

- PREF_CONFLICT: replace the first product category with a label that contradicts the persona's first hard constraint (e.g., *leather handbags* for a persona with `avoid_leather`). Location: `productCategoryList[0]`.
- OVER_PERSONALIZATION_HALLUCINATION: prepend a fabricated persona-history reference (e.g., "Given their Dyson Airwrap, . . . ") to the overview. Location: `overview_text`.
- XCOMP_MISMATCH: replace the related-query list with the related-query list of an unrelated GOOD record. Location: `rqList`.
- INTENT_DRIFT: replace the first product category with an unrelated category drawn from a fixed off-intent pool. Location: `productCategoryList[0]`.

- REDUNDANCY: duplicate the first related query into the second slot. Location: `rqList[1]`.
- EVIDENCE_MISMATCH: replace the attribution list with the attribution list of an unrelated GOOD record. Location: `attribution_list`.
- UNSUPPORTED_CLAIM: append a fabricated product-feature claim (e.g., "It also highlights solar charging support.") to the overview. Location: `overview_text`.

Multi-defect variants (tiers 2–4) are produced by composing edits from distinct response-field groups (*categories*, *related queries*, *attributions*, *overview*) under the canonical group order *categories* → *related queries* → *attributions* → *overview*, which avoids interactions between overview-modifying edits and category-modifying edits (e.g., bolded-category alignment). Combinations spanning two defects from the same field group are disallowed; the full set of admissible single- and multi-defect combinations is enumerated in `pmsb.faults` of the released benchmark code.

Because every BAD edit is a deterministic function of (i) the source GOOD record, (ii) the assigned persona, and (iii) the defect family, BAD records do not require a separate LLM-judging-style validator on top of the L1 12-check audit. The injection produces both the modified four-field response and the gold location string in a single step, so any disagreement between a judge's predicted (defect family, defect location) and the gold pair in Tabs. 3 and 4 is unambiguously a judge error rather than a labeling error.

C Prior Evaluation Protocols and Baseline Adaptation

This appendix documents how prior personalization, RAG, and judging protocols relate to the PACESHOP evaluation schema. We first describe, in a single unified support table, which scenarios each prior judging protocol natively covers and which scenarios each running baseline scores or abstains on (Sec. C.1). We then audit target leakage for the three single-target adapted baselines on the same held-out test set (Sec. C.2). Finally, we list the baseline prompt templates run on the shared backbone pool (Sec. C.3). Together these subsections explain why the "–" cells in the main result tables are deliberate and measurement-backed, not notational laziness.

C.1 Native and Adapted Scenario Support

Prior personalization, RAG, and judging protocols differ in what they natively evaluate. A personalization benchmark may test whether a model adapts to user history without defining defect-family labels; a RAG evaluator may check product-evidence faithfulness without evaluating invented shopper history; and a scalar judge may estimate overall quality without localizing the failed response field. Tab. 18 summarizes this in two blocks: the upper block lists prior protocols that we do *not* directly run as PACESHOP judges and marks the targets they natively cover with "∂" or "–"; the lower block lists the baselines we *do* run on the shared backbone pool used in Tab. 3 – four multi-target rubric ablations of the PACEJUDGE output schema and three single-target adapted baselines (ARES-/PersonaLens-/EtaPP-style) that score only their native target and report "–" elsewhere. Unsupported scenarios are reported as "–" rather than forced into the PACESHOP output schema.

C.2 Native Judging-Protocol Map and Evaluation-Target Leakage Audit

The strict paired-counterfactual main table (Tab. 4) is supported by two audit tables kept here rather than in the main text. The native judging-protocol map (Tab. 19) records, for each adapted single-target baseline, which PACE target its native judging protocol covers and why the other targets are reported as "–" rather than over-adapted. The target-leakage audit (Tab. 20) then measures, on the same 3,400-record held-out test set, the rate at which each adapted baseline emits a defect-family label outside its native target (e.g., ARES-style firing PREF_CONFLICT on a record whose gold defect is compositional). ARES-style and PersonaLens-style stay under 1–2% on the *Any off-target* aggregate, validating their honest restriction in Tab. 4 even though they show non-trivial target leakage on individual P/C cells; EtaPP-style leaks heavily (99% any-off-target), and we therefore do not interpret its S4 cell as evidence that EtaPP-style covers more than the A target. Together these two tables protect Tab. 4 from the failure mode of "apply a prompt, get a number on every cell": the "–" restrictions are deliberate and supported by measurement, not notational laziness.

C.3 Baseline Prompt Templates

We organize the baselines into two groups. The first group is a set of *multi-target rubric ablations* that share the PACEJUDGE input bundle (query, persona JSON, response XML, structured response JSON, candidate evidence IDs, candidate evidence preview, deterministic checks) and the same structured output contract, differing only in *instruction* and *persona visibility*. The second group is a set of *single-target adapted baselines* (ARES-style, PersonaLens-style, EtaPP-style) whose native judging protocols cover only one PACE target each; we adapt them faithfully on their native target and abstain ("–") on the other targets, as documented in Tab. 19 and audited in Tab. 20.

C.3.1 Multi-target rubric ablations

The rubric ablations all emit the full PACEJUDGE output schema (verdict, P/C/E/F axis scores, defect family, defect location, supporting evidence IDs, confidence, rationale) so that metrics across rows of Tab. 3 differ only by the prompt instruction.

G-Eval-style rubric judge (Liu et al., 2023). "You are a rubric-guided judge in the style of a generic high-quality LLM evaluator. Score the response using a compact rubric over relevance, helpfulness, coherence, grounding, and structure. Then translate that rubric judgment into the required output schema."

Scalar-first judge (MT-Bench style) (Zheng et al., 2023). "You are a scalar-first judge. Base the decision mainly on one overall quality impression, then back-fill the required structured fields in a best-effort way."

Persona-blind ablation. "You are a generic shopping-response judge. The persona is intentionally withheld for this baseline. Judge only

Method / native protocol	Native signal or role in study	S1 P-source	S2 C detect+loc	S3 E product/history	S4 A family+loc	Why this row scores or abstains
Prior protocols not directly run as PACESHOP *judges (reference)*						
LaMP (Salemi et al., 2024)	Task metrics for personalized generation	∂	–	–	–	No unchanged-response persona swap or defect-source label.
PersonaLens (Zhao et al., 2025b)	LLM user and judging agents for assistants	∂	–	–	–	Evaluates user-conditioned success, not four-field response diagnosis.
PrefEval (Zhao et al., 2025a)	Preference following/classification	∂	–	–	–	Persona signal is present, but no evidence pool or response-field location.
PersonaMem (Jiang et al., 2025)	Response selection over user histories	∂	–	–	–	History use is measured, not defect family and location.
LikeBench (Rahman et al., 2025)	Simulated-user likability diagnostics	∂	–	–	–	Subjective user-fit signal; no PACESHOP grounding or localization contract.
EtaPP (Hao et al., 2025)	Key-point judging for personalized tool use	∂	–	–	∂	Checks personalized task points, but not PACESHOP source-family/location labels.
Personalized judge (Dong et al., 2024)	User-conditioned judging preference	∂	–	–	–	Personalizes a judging preference, but does not define multi-field defect localization.
ECom-Bench (Wang et al., 2025a)	Task-success trajectories in e-commerce support	∂	–	–	–	Realistic commerce tasks, but not fixed-response judge diagnosis.
Search Arena (Miroyan et al., 2025)	Human preferences for search-augmented chat	–	–	∂	–	Search/citation quality is relevant, but no persona-history grounding label.
ARES (Saad-Falcon et al., 2024)	RAG faithfulness to retrieved evidence	–	–	∂	–	Covers product-evidence grounding only, not persona or multi-field diagnosis.
G-Eval / MT-Bench (Liu et al., 2023; Zheng et al., 2023)	Scalar or pairwise quality judgments	–	–	–	–	Runnable only after adapting the prompt to emit the PACESHOP schema (see below).
Running baselines on the shared backbone pool (used in Tab. 3)						
G-Eval-style rubric (Liu et al., 2023)	Multi-target rubric ablation, persona-aware	✓	✓	✓	✓	Adapted to the PACEJUDGE output schema; differs only in instruction.
MT-Bench-style scalar (Zheng et al., 2023)	Multi-target rubric ablation, scalar-first	✓	✓	✓	✓	Adapted to the PACEJUDGE output schema; back-fills structured fields after a scalar judgment.
Persona-blind	Multi-target rubric ablation, persona withheld	–	✓	✓	✓	Persona payload replaced by `{"omitted": true}`; isolates the marginal value of persona conditioning.
Persona-aware generic	Multi-target rubric ablation, persona-aware	✓	✓	✓	✓	Shared protocol for the backbone comparison without explicit counterfactual reasoning.
ARES-style	Single-target adapted: evidence (E) only	–	–	✓	–	Native protocol scores RAG faithfulness only; non-E targets reported as "–".
PersonaLens-style	Single-target adapted: persona (P) only	✓	–	–	–	Native protocol scores personalization only; non-P targets reported as "–".
EtaPP-style	Single-target adapted: actionable (A) only	–	–	–	✓	Native protocol scores key-point matches only; non-A targets reported as "–".
PACEJUDGE (ours)	**Structured PACE judging contract**	✓	✓	✓	✓	**Verdict, target scores, defect family, location, and evidence IDs.**

Table 18: Native and adapted support for the four PACE target scenarios. The first block lists prior protocols that are *not directly run* as PACESHOP judges; "∂" marks a related signal that does not define the PACESHOP closure metric or the required diagnostic output fields. The second block lists the baselines we *do run* on the shared backbone pool in Tab. 3: four multi-target rubric ablations of the PACEJUDGE output schema and three single-target adapted baselines (ARES-/PersonaLens-/EtaPP-style) that score only their native target and report "–" on the others. Tab. 19 expands the native protocol mapping for the single-target adapted baselines, and Tab. 20 audits their off-target leakage.

Baseline	Native judging protocol	PACE target	Supported scenario cell	Why other PACE targets are "–"
ARES-style (Saad-Falcon et al., 2024)	RAG faithfulness: 3 LM-judges (context relevance, answer faithfulness, answer relevance)	E	$S3_E$ only	Native judging protocol explicitly does not score persona axes or cross-field consistency; non-E defect families are not emitted.
PersonaLens-style (Zhao et al., 2025b)	3-axis judge: personalization, response quality, task success	P	$S1_P$ only	Native rubric has no slot for cross-field consistency or evidence-pool faithfulness; family vocabulary covers only persona conflicts.
EtaPP-style (Hao et al., 2025)	Key-point judging: {defect family, defect location} as the gold key-points	A	$S4_A$ only	Target scores are fixed at 3.0; consequently the target-triage conjunct is rarely satisfied even on the native target.
Personalized judge (Dong et al., 2024)	Binary preference judge measured by judge-vs-human agreement	P (binary)	–	No defect family / location / target-score output; the judging protocol is judge-vs-human pairwise agreement, not pointwise diagnosis.
Search Arena (Miroyan et al., 2025)	Pairwise human votes on search-augmented chat	E (partial)	–	Captures search/citation quality, but no persona-history label and no defect family.
PrefEval (Zhao et al., 2025a)	Preference-following accuracy / classification across long context	P	–	No persona-swap counterfactual label, and no multi-field response to localize.
PersonaMem (Jiang et al., 2025)	MCQ accuracy over an evolving user history	P	–	No structured response, and no native defect-family taxonomy.
LikeBench (Rahman et al., 2025)	Per-turn likability dimensions from simulated users	P	–	Subjective likability dimensions; no structured response and no grounding contract.
LaMP (Salemi et al., 2024)	Task metrics (accuracy / ROUGE) on user-conditioned generation	P	–	Single-text generation; no four-field response and no defect labels.

Table 19: **Native judging-protocol to PACE target mapping.** Each adapted baseline is restricted to the PACE target that its native judging protocol scores; unsupported targets are reported as "–". Table 4 uses this mapping when reporting scenario scores: prior protocols cover at most one target each, so the joint S1–S4 PACE average is undefined for them. Only PACEJUDGE has an output contract supporting all four PACE targets.

whether the response seems broadly useful, coherent, grounded, and structurally valid." The persona payload is serialized as `{"omitted": true}`.

Persona-aware generic rubric. "You are a persona-aware but otherwise generic shopping-response judge. Use the persona and query to judge whether the response is suitable for the user. Do not use explicit counterfactual swap reasoning and do not invent additional defect classes beyond the provided labels."

C.3.2 Single-target adapted baselines

ARES-style, PersonaLens-style, and EtaPP-style do not share the PACEJUDGE four-target output schema in their native form. We therefore adapt each one only on the PACE target that its native judging protocol scores, leaving cells outside the native target as "–" in Tab. 3 and Tab. 4. To keep the adaptation auditable, all three adapted baselines run on the same 3,400-record held-out test set and the same shared backbone pool as the rubric ablations; only the *native-target* cells are scored against the PACEJUDGE schema, and an evaluation-target leakage audit (Tab. 20) measures the rate at which each baseline emits a defect family outside its native target.

ARES-style grounding judge (Saad-Falcon et al., 2024). The native ARES protocol is a RAG-faithfulness judge built from three lightweight LM scorers: *context relevance*, *answer faithfulness*, and *answer relevance*. We instantiate this protocol on

Adapted baseline	P off-target	C off-target	E off-target	Any off-target
Off-target firing rate: P/C/E columns are computed on records whose gold target is not the baseline's native target. Lower is better; 0 means perfect abstention.				
ARES-style (Saad-Falcon et al., 2024) (native: E)	23%	22%	*native*	2%
PersonaLens-style (Zhao et al., 2025b) (native: P)	*native*	20%	1%	0%
EtaPP-style (Hao et al., 2025) (native: A)	0%	0%	0%	99%
Reference: PACEJUDGE *supports all* PACE *targets natively (so the relevant rate is family misclassification, not leakage)*				
PACEJUDGE (native: P+A+C+E)	7%	14%	27%	n/a

Table 20: **Evaluation-target leakage audit.** For each adapted single-target baseline (Sonnet 4.6 on the 3,400-record held-out test set), this table reports the rate at which the baseline emits a defect-family label outside its native PACE target. A genuine single-target baseline should score near 0% on every non-native target; high leakage means the prompt is over-adapted. The PACEJUDGE reference row reports family misclassification rather than leakage, since PACEJUDGE supports all PACE targets. This audit is the safeguard against "apply a prompt, get a number on every cell": cells in Table 4 are non-zero for a single-target baseline only on its native target.

PACESHOP by mapping the three ARES axes onto (q, y, E): context relevance asks whether the candidate evidence IDs in E are related to q; answer faithfulness asks whether claims in y are supported by E, including no fabricated evidence IDs and no invented persona-history claims; answer relevance asks whether y answers q. The judge averages the three sub-judgments into the evidence-grounding axis $s_{\mathrm{E}} \in [1, 5]$ and emits $\hat{\ell} = \mathrm{BAD}$ only when $s_{\mathrm{E}} < 3.0$. When BAD, the predicted defect family $\hat{d}$ is restricted to evidence-target labels {`EVIDENCE_MISMATCH`, `OVER_PERSONALIZATION_HALLUCINATION`, `UNSUPPORTED_CLAIM`}, and $\hat{r}$ is the response field carrying the unsupported claim or invalid attribution. For any non-evidence defect, the prompt explicitly returns `NONE`/`NONE` rather than fabricating a label. The persona, compositional, and format/safety axes $\{s_{\mathrm{P}}, s_{\mathrm{C}}, s_{\mathrm{F}}\}$ are pinned to a neutral 3.0 because ARES does not natively score them, and supporting evidence IDs are constrained to $\hat{S} \subseteq E$. Consequently, this baseline reports a measured number on the S3 column of Tab. 3 and $\mathrm{S3_E}$ in Tab. 4, and reports "–" on S1, S2, and S4. The leakage audit in Tab. 20 confirms that this restriction is honored in practice: ARES-style fires a non-evidence defect family on under 2% of off-target records on the *Any off-target* aggregate.

PersonaLens-style personalization judge (Zhao et al., 2025b). The native PersonaLens judging protocol is a three-axis judge that scores *personalization*, *response quality*, and *task success* for user-conditioned generation. We adapt it to PACESHOP by walking through the structured persona one slot at a time—hard constraints, brand likes/avoids, attribute preferences, and budget—and asking the judge to mark each slot as honored, contradicted, or unconstrained by the response fields; this drives the persona-alignment axis $s_{\mathrm{P}} \in [1, 5]$. Response quality scores the format/safety axis s_{F}, and task success is folded into the verdict via the threshold $s_{\mathrm{P}} < 3.0$. When BAD, the predicted defect family $\hat{d}$ is restricted to persona-target labels: `PREF_CONFLICT` for an explicit persona conflict, e.g., the response bolds a brand the persona avoids, or `OVER_PERSONALIZATION_HALLUCINATION` when the response invents a purchase or search not in the persona history; $\hat{r}$ is the response field carrying the conflict. The compositional and evidence axes $\{s_{\mathrm{C}}, s_{\mathrm{E}}\}$ are pinned at 3.0, $\hat{S}$ may be empty, and any non-persona defect is reported as `NONE`/`NONE`. This baseline therefore contributes a measured number on the S1 column, $\mathrm{S1_P}$ in Tab. 4, and abstains elsewhere. The evaluation-target leakage audit in Tab. 20 shows that PersonaLens-style fires a non-persona family on under 1% of off-target records on the *Any off-target* aggregate.

EtaPP-style actionable key-point judge (Hao et al., 2025). The native EtaPP judging protocol judges a personalized action trace against human-annotated *key-points* that the trace must satisfy. We adapt this to PACESHOP by treating the gold defect family d and the gold defect location r as the two key-points the judge must name: the candidate response y replaces the action trace, the deterministic checks plus a brief read of the four response fields drive the GOOD/BAD verdict, and when BAD the judge proposes both $\hat{d} \in \mathcal{D}$ and $\hat{r} \in \mathcal{R}$ from the full label space. Following the EtaPP convention, the target score on the predicted target is fixed at 3.0 rather than emitted by the judge; s_{P}, s_{C}, s_{E} are likewise pinned at 3.0, while s_{F} is scored honestly from schema validity and $\hat{S}$ may be empty. Because EtaPP's native judging protocol does not score persona, compositional, or evidence-grounding targets separately, this baseline reports measured S4 family/location numbers in Tab. 3

(and $S4_A$ in Tab. 4) and "–" on S1–S3. Two caveats follow from the audit tables. First, the fixed 3.0 target score is why EtaPP-style fails the strict target-triage conjunct in Tab. 4 more often than its S4 family number alone would suggest. Second, the evaluation-target leakage audit (Tab. 20) shows that EtaPP-style emits *some* defect family on 99% of off-target records, so we do not interpret its S4 cell as evidence that EtaPP-style covers more than the A target.

The full Markdown prompt files for all baselines, including the three adapted single-target baselines, are shipped with the code release.

D PACEJUDGE Evaluation

This appendix expands the main-paper subsection on *Coverage of the* PACE *Targets* (Sec. 4.2) and the structured output contract used by PACEJUDGE. It is organized as follows. Sec. D.1 gives a per-target walkthrough mapping each PACEJUDGE output field to one of the four PACE targets and to the corresponding evaluation scenario. Sec. D.2 expands the four main-body axes into the fuller six-dimension rubric used during judge development. Sec. D.3 documents the structured fault-localization protocol that turns the judge from a scalar scorer into a diagnostic tool. Readers who only need the headline mapping can rely on Tab. 2 and the main-paper subsection; the subsections below are intended for readers who want the full coverage argument and the dimension-by-dimension protocol.

D.1 How the Protocol Closes the Four PACE Targets

Tab. 2 gives the high-level mapping from PACE targets to PACESHOP design choices, PACEJUDGE output fields, and scenario-level metrics. The role of PACEJUDGE is to make the corresponding diagnosis reportable: PACESHOP makes each target observable, while PACEJUDGE gives the judge explicit fields for reporting it.

- **Personalized** ($s_P \rightarrow S1$). The persona-alignment score checks whether y respects the shopper context in p, including constraints, preferences, brand affinities, budget, purchase history, and recent searches. This closes the Personalized target by forcing the verdict to condition on p rather than treating persona as optional context. It directly supports the S1 persona-swap evaluation.
- **Actionable** ($a_A \rightarrow S4$). The actionable diagnosis object requires the judge to report what failed and where: $\hat{d}$ identifies the defect family, $\hat{r}$ identifies the response field, $\hat{S}$ provides evidence support, $\hat{c}$ indicates confidence, and $\hat{z}$ explains the decision. This closes the Actionable target by turning the judge from a scalar scorer into a diagnostic tool. It directly supports the S4 defect-family and field-location evaluation.
- **Compositional** ($s_C, \hat{r} \rightarrow S2$). The compositional-consistency score evaluates y as a four-field object rather than as independent text. The overview should agree with the category list, related queries should refine the represented product facets rather than drift away from the intent, and attributions should support the claims they are attached to. Together with the location field $\hat{r}$, this closes the Compositional target and supports S2 cross-field detection and localization.
- **Evidence-grounded** ($s_E, \hat{S} \rightarrow S3$). The evidence-grounding score verifies both product grounding and persona-history grounding. Product grounding requires attribution IDs to exist in E and support the relevant claims. Persona-history grounding requires the response not to invent purchases, searches, or user-specific facts absent from p. Together with the supporting-evidence field $\hat{S}$, this closes the Evidence-grounded target and supports S3 product- and history-grounding evaluation.

The auxiliary format/safety score s_F checks whether y satisfies the response contract, including count constraints, length limits, banned-string filters, evidence-ID formatting, and XML round-trip validity. This separates structural compliance from the four PACE targets: a response may be well-formed but semantically invalid, or semantically plausible but structurally unusable.

This output contract explains where scalar judges fail and where PACEJUDGE can be evaluated more precisely. A scalar judge may call a response "mostly helpful," but it has no required slot for naming the broken field, the defect type, or the evidence supporting the diagnosis. PACEJUDGE exposes these missing fields directly: the P/C/E target scores identify which quality dimension degraded, the Actionable fields identify what failed and where, and the evidence-ID constraint prevents the judge from inventing its own support. As a result, the protocol can be evaluated not only

by GOOD/BAD accuracy, but also by the scenario-aligned metrics in Sec. 5: persona-swap accuracy, cross-field localization, grounding detection, hallucinated-ID rate, and defect-family/location accuracy.

D.2 Detailed PACEJUDGE Evaluation Dimensions

The four main-body axes (§4) collapse a richer six-dimension protocol that we used during the judge development. Each dimension below states (i) the challenge that motivates it, (ii) how PACEJUDGE operationalizes it, and (iii) the metrics used to measure it. Every dimension admits both deterministic checks and LLM-scored assessments; deterministic checks provide a cost-free lower bound, while LLM scoring captures semantic nuances that rules cannot.

D1: Persona Alignment. *Challenge.* The central gap in personalization evaluation is distinguishing responses that are *generically helpful* from those that are *correct for a specific user* (Zhao et al., 2025b,a; Jiang et al., 2025; Rahman et al., 2025). A response can score highly on fluency and relevance yet still ignore the user's budget constraint, violate a hard preference, or recommend products misaligned with the user's shopping journey stage. *Approach.* PACEJUDGE evaluates persona alignment through *counterfactual persona swaps* (S1 in §5.2): if swapping the persona changes which response dimensions should differ, the judge must detect that change; dimensions that should remain invariant under the swap must stay stable. Formally, let $A_{\text{rel}}(q, p)$ denote dimensions that should change under a persona swap and $A_{\text{inv}}(q, p)$ denote dimensions that should remain invariant. The persona-swap score is $S_{\text{swap}} = \sum_{a \in A_{\text{rel}}} w_a \Delta_a - \lambda \sum_{a \in A_{\text{inv}}} w_a \Delta_a$, where Δ_a measures the distance between axis signatures of paired responses. We also define the alignment gap $G(q, p) = H_{\text{gen}}(f(q, \varnothing)) - H_{\text{pers}}(f(q, p))$, which quantifies the extent to which persona conditioning changes the evaluation outcome. *Metrics.* Persona-alignment score (1–5, LLM-scored); GOOD–BAD score gap on D1 (larger gap = stronger persona discrimination). A *negative* gap indicates the judge cannot distinguish persona-conditioned quality—the signature of a persona-blind evaluation.

D2: Relevancy & Helpfulness. *Challenge.* A shopping assistant must match both explicit requirements (brand, price, features) and implicit expectations (use case, quality tier, journey stage) (Hao et al., 2025; Miroyan et al., 2025). Generic "helpfulness" ratings conflate these with fluency. *Approach.* The judge scores each component on product relevancy, intent matching (explicit + implicit), and actionable decision support—whether the overview provides real buying advice rather than generic category listing. *Metrics.* Per-section helpfulness score (1–5); binary criteria for product relevancy, intent matching, and helpfulness.

D3: Diversity & Distinctiveness. *Challenge.* Responses that are relevant but homogeneous—e.g., recommending five near-identical products or generating overlapping category groups—fail to help users compare options. *Approach.* The judge checks whether products, categories, and related queries offer sufficient variety in brands, price points, and features; whether category groups are non-overlapping and organized by customer need; and whether related queries span multiple types (comparison, recommendation). *Metrics.* Binary criteria for product diversity, type diversity, query specificity, and category distinctness; deterministic duplicate-detection checks.

D4: Compositional Consistency. *Challenge.* A structured four-field response is correct only if its fields are *mutually coherent*: the overview should preview the categories, the categories should be refinable by the related queries, and the attributions should support the claims. Existing judges evaluate each field independently and miss cross-field failures. *Approach.* PACEJUDGE reasons over explicit cross-field constraints: overview–category alignment (bolded terms match category names), category–query coverage (queries refine the categories), evidence–claim grounding, diversity/non-redundancy, and intent preservation across all four fields. The objective is to judge the response as a structured object rather than scoring each field in isolation. *Metrics.* Component-consistency score (1–5, LLM-scored); deterministic highlight–category overlap check; GOOD–BAD score gap on D4.

D5: Evidence Grounding. *Challenge.* Shopping assistants must ground their claims in real product evidence. Hallucinated product features or invalid attribute IDs erode user trust. Evidence-only evaluators (Saad-Falcon et al., 2024) address this but ignore whether the *selected* evidence is appropriate for the specific user. *Approach.* The judge verifies that claims in the overview and categories

are supported by the referenced evidence records, that attribute IDs resolve to valid entries in the evidence pool, and that no hallucinated evidence appears. Evidence-grounding assessment combines deterministic ID validation with LLM-based claim–evidence alignment. *Metrics.* Evidence-grounding score (1–5, LLM-scored); attribute-ID validity (deterministic); hallucinated-evidence rate (fraction of cited IDs absent from the evidence pool); GOOD–BAD score gap on D5.

D6: Format & Safety. *Challenge.* Structural compliance (correct counts, lengths, valid XML) and trust/safety requirements (neutral tone, appropriate disclaimers, no banned content) are prerequisites for deployment but are often treated as afterthoughts in evaluation. *Approach.* Twelve deterministic validation checks (overview length, bold–category match, category and related-query counts and uniqueness, evidence-ID validity, XML round-trip, banned-string screening) form a cost-free baseline. LLM judges additionally assess language tone and score a format-safety axis. *Metrics.* Format-safety score (1–5, LLM-scored); 12-check deterministic pass rate; parse success rate; GOOD–BAD score gap on D6.

Mapping D1–D6 to the four main-body axes. In the main-body table (§5.3) we report four axes rather than six: D2 and D3 both contribute to the *component_consistency* axis (relevancy is assessed within each component, and diversity is checked across components), which keeps the output contract compact without losing the underlying signals. The full six-dimension rubric is used by the LLM in its free-text reasoning and is reflected back into the four axis scores.

D.3 Fault Localization Protocol

Beyond the quality dimensions, PACEJUDGE addresses a capability missing from existing evaluation frameworks: *structured fault localization*. Rather than reducing evaluation to a scalar score, the judge predicts (i) the *defect family* (which of 7 failure modes occurred) and (ii) the *defect location* (which of 4 field groups contains the fault). This transforms the judge from a classifier into a *diagnostic tool*: developers can identify not just that a response failed, but *what* went wrong and *where* to fix it.

The multi-tier defect design in PACESHOP creates a particularly demanding test: for multi-defect records, the judge must detect and localize faults across up to four field groups simultaneously. We measure localization via defect-type macro-F1 (7 families), defect-location exact match, and field-group match. For single-defect records, exact match is reported directly. For multi-defect records the gold label is a set of locations while the current protocol outputs one prediction, so we report a *hit rate*—whether the predicted location appears in the gold set. Multi-location prediction is a natural extension (a single inference emitting a list rather than a scalar location).

D.4 PACEJUDGE Evaluation Protocol and Prompt

The PACEJUDGE evaluation protocol is implemented as a single training-free judging prompt that takes a PACESHOP record (q, p, y, E) as typed input and returns the structured output defined in §4. To make the protocol fully reproducible, we reproduce the prompt verbatim in this subsection; this is the symmetric counterpart of the dataset-side response-generation prompt in Appx. B.9, and it is the prompt that all PACEJUDGE rows in Tabs. 3 and 4 use across every backbone.

Prompt structure. The prompt is parameterized by four slot groups, each contributing a different piece of the protocol:

1. **Typed input bundle.** Ten named slots bind the record to evaluation-facing fields: `record_id`, `split`, `keyword` (the query q), `persona_json` (the structured persona p), `response_xml` and `response_struct_json` (the candidate response y in two views), `candidate_evidence_ids_json` and `candidate_evidence_json` (the evidence pool E as IDs and previews), `used_evidence_ids_json` (the evidence subset cited during PACESHOP GOOD-response construction, used as a soft anchor rather than a label), and `deterministic_checks_json` (the L1 12-check validator output, exposed to the judge as objective anchors).

2. **Closed label sets.** The diagnostic vocabulary is pinned to two closed enumerations: `allowed_defect_labels` = the seven PACESHOP defect families plus NONE, and `allowed_location_labels` = the four response-field groups (*overview*, *categories*, *related queries*, *evidence attributions*) plus NONE. Free-form labels outside these enumerations are rejected at parse time rather than coerced.

3. **Decision rules.** Four explicit rules force the judge to commit to a verdict and an evidence-grounded diagnosis rather than collapse to a scalar score: (i) mark GOOD only when the response is correct for the persona, coherent across components, grounded in E, and structurally valid; (ii) use NONE/NONE for defect family/location only when no material defect is present; (iii) `supporting_evidence_ids` must be a subset of `candidate_evidence_ids_json` (no invented IDs); (iv) `confidence` and `overall_bad_probability` must lie in $[0, 1]$, while `overall_quality` and the four axis scores must lie in $[1, 5]$.

4. **Output-schema contract.** A JSON-Schema object (the `prompt_contract_json` slot) requires the verdict $\hat{\ell} \in \{\text{GOOD}, \text{BAD}\}$, the BAD probability $\hat{p}_{\text{bad}} \in [0, 1]$, the four axis scores $s_{\text{P}}, s_{\text{C}}, s_{\text{E}}, s_{\text{F}} \in [1, 5]$, and the actionable diagnostic object $a_{\text{A}} = (\hat{d}, \hat{r}, \hat{S}, \hat{c}, \hat{z})$ with $\hat{d} \in \mathcal{D} \cup \{\text{NONE}\}$, $\hat{r} \in \mathcal{R} \cup \{\text{NONE}\}$, and $\hat{S} \subseteq E$.

PACEJUDGE prompt. The full prompt as run on every backbone is reproduced below. Slots in curly braces are filled per-record by the runner; the closed label sets and the JSON-Schema contract are the same on every record.

```
You are LLMaJ for persona-conditioned multi-component shopping
assistant evaluation.

Judge the candidate response as a structured object. Use the
deterministic checks as objective anchors, but do not stop there.
Reason about:
- persona alignment
- cross-component consistency
- evidence grounding
- format safety

Allowed defect labels: {allowed_defect_labels}
Allowed defect locations: {allowed_location_labels}

Input bundle
- Record ID: {record_id}
- Split: {split}
- Query: {keyword}
```

```
- Persona JSON:
{persona_json}
- Response XML:
{response_xml}
- Response struct JSON:
{response_struct_json}
- Candidate evidence IDs:
{candidate_evidence_ids_json}
- Candidate evidence preview:
{candidate_evidence_json}
- Used evidence IDs from benchmark construction:
{used_evidence_ids_json}
- Deterministic checks:
{deterministic_checks_json}

Decision rules
- Mark GOOD only when the response is correct for this user,
  coherent across components, grounded in the provided evidence,
  and structurally valid.
- Mark BAD when any material failure is present.
- Use NONE / NONE only if no material defect is present.
- supporting_evidence_ids must be a subset of Candidate evidence IDs.
- confidence and overall_bad_probability must be floats in [0, 1].
- overall_quality and all axis scores must be floats in [1, 5].

Return JSON only with this schema:
{prompt_contract_json}
```

Why this prompt is multi-target rather than scalar. The four slot groups together are what distinguish PACEJUDGE from the rubric ablations and the single-target adapted baselines in Appx. C.3. The rubric ablations share the same input bundle and the same output schema but replace the four-axis reasoning instruction with a generic rubric or a scalar quality impression; the single-target adapted baselines (ARES-style, PersonaLens-style, EtaPP-style) faithfully implement their native judging contracts but cover only one PACE target each. Only PACEJUDGE pairs the four-axis reasoning instruction with the closed-vocabulary defect labels, the evidence-ID containment constraint $\hat{S} \subseteq E$, and the explicit GOOD/NONE/NONE discipline that prevents over-flagging. Outputs that violate any closed-vocabulary or range constraint are rejected at parse time rather than silently coerced, which is what allows the strict paired-counterfactual gates of Tab. 4 – joint pair credit, target-triage, and evidence-ID containment – to be evaluated faithfully against the released schema.

Example 1: GOOD record correctly classified	*Query:* `gemerry lash extensions`
Verdict	GOOD Confidence: 0.70 Defect: NONE
Axis scores	Persona: 3.7 Consistency: 4.0 Grounding: 3.5 Format: 4.8
Reasoning	*"The response is structurally valid and passes all deterministic checks. Persona alignment is decent: it references budget-friendly options, sensitive skin, travel-friendly, and premium feel. However, the query specifically mentions 'gemerry' as a brand, and the response doesn't acknowledge this brand at all. Evidence grounding is adequate—attributions match the used evidence IDs exactly, though some overview claims about 'lightweight materials' are not directly supported. The categories and rqList are well-structured and complementary. Overall a functional response with minor issues but no material failures."*
Example 2: Single-defect BAD record	*Query:* `toy story 4 toys`
Verdict	BAD Confidence: 0.94 Defect: XCOMP_MISMATCH @ `rqList` ✓ matches gold
Axis scores	Persona: 3.0 Consistency: **1.5** Grounding: 3.5 Format: 4.5
Reasoning	*"The productCategoryList is well-aligned with 'toy story 4 toys' and grounded in evidence (action figures, LEGO playsets, board games). However, the rqList is completely misaligned: all five queries are about Halloween costumes and accessories (alien headband, group costume set, inflatable alien costume), which have nothing to do with Toy Story 4 toys. This is a severe cross-component mismatch. The persona alignment is moderate—the overview mentions premium and durability. The rqList's costume focus is a material failure."*
Example 3: Triple-defect BAD record	*Query:* `wine glasses without stem`
Verdict	BAD Confidence: 0.96 Defect: EVIDENCE_MISMATCH @ `attribution_list`
Axis scores	Persona: 2.5 Consistency: 2.0 Grounding: **1.0** Format: 1.5
Gold defects	EVIDENCE_MISMATCH + INTENT_DRIFT + OVER_PERSONALIZATION_HALLUC. @ 3 locations
Reasoning	*"Multiple critical failures: (1) All 10 attribution IDs are invalid—none match the candidate evidence IDs. The response fabricated evidence references. (2) The overview text contains a hallucinated 'Dyson Airwrap' reference that appears nowhere in the persona's history. (3) The 'Camping Gear' category is off-topic for a wine glasses query. While some persona elements are addressed (small-space, travel-friendly), the evidence grounding failure and hallucination are disqualifying."*

Table 21: Qualitative examples of PACEJUDGE (Opus 4.6) output on the development set. The judge produces structured diagnostics: per-axis scores identify *which* dimension failed (e.g., Consistency = 1.5 in Example 2), defect type and location enable *actionable debugging* (e.g., the costume-themed rqList in a toy query), and free-text reasoning explains *why* the response is defective with specific evidence. In Example 3, the judge identifies all three failure modes in its reasoning despite outputting only one primary defect label.

D.4.1 PACEJUDGE Output Examples

Tab. 21 presents three representative judge outputs from PACEJUDGE. They illustrate how the structured evaluation protocol produces *actionable diagnostics* rather than scalar scores. In the single-defect case (Example 2), the judge assigns a low consistency score (1.5) while keeping other axes high, correctly pinpointing the cross-field mismatch between toy categories and costume-themed related queries. In the triple-defect case (Example 3), the judge's free-text reasoning identifies all three injected defect families—fabricated evidence IDs, hallucinated persona references, and off-topic categories—even though the output schema allows only one primary defect label. This generative reasoning, combined with structured axis scores and defect labels, bridges the gap between automated evaluation and human-interpretable debugging.

E Experiments

This appendix provides the experimental details omitted from the main text for space: backbone and prompt settings, exact scenario metric definitions, per-backbone main results, scenario-level breakdowns, expanded findings, and – in the final subsection (§E.8) – per-target specificity audits, difficulty-gradient analyses, qualitative outputs, and concrete record galleries. Native judging-protocol baseline support, prompt templates, and the judging-protocol baseline ladder are documented separately in Appx. C.

E.1 Backbones, Prompt Configurations, and Parsing

We evaluate seven LLM backbones spanning frontier and open-weight tiers: Claude Opus 4.7, Sonnet 4.6, Sonnet 4.5, Haiku 4.5, Qwen3 32B, GPT-OSS 20B, and GPT-OSS 120B. All backbones are run through the same PACESHOP record interface (q, p, y, E) where applicable. Each backbone is

Scenario	Failure being isolated	Metric	Observed baseline gap
S0 General	Reference GOOD/BAD discrimination before stressing a specific PACE target.	Balanced accuracy on GOOD/BAD subset.	S0 is a sanity check, not the main claim. PACEJUDGE reaches 0.94, while adapted baselines range from 0.64–0.82.
S1 Personalized	The response is unchanged but the persona is swapped to contradict a claim or recommendation.	Persona-source exact match: the judge must diagnose the swap as PREF_CONFLICT.	Native personalization benchmarks provide related user-fit signals but not this counterfactual source label. In the adapted sweep, PACEJUDGE is highest on mean S1 (0.24 vs. 0.13 best non-PACEJUDGE).
S2 Compositional	The response has a cross-field contradiction or intent drift among overview, product categories, related queries, and evidence attributions.	Joint closure: BAD detection and correct 4-way response-field localization.	PACEJUDGE has the highest mean detect+loc score (0.88 vs. 0.82 best non-PACEJUDGE). The gap is largest on open-weight backbones, where generic prompts detect some defects but localize them less reliably.
S3 Evidence-grounded	The response cites invalid product evidence or invents shopper purchase/search history.	Product and history closure: BAD detection with no hallucinated supporting evidence IDs.	Product grounding is partly covered by prior RAG/search evaluators, but persona-history grounding is not. PACEJUDGE reaches 1.00/1.00 on product/history means, vs. 0.95/0.94 best non-PACEJUDGE.
S4 Actionable	A single-defect BAD record must be diagnosed by both defect family and response-field location.	Complete diagnosis: exact defect family and 4-way response-field location.	PACEJUDGE is highest on complete family+location closure (0.56 vs. 0.48 best non-PACEJUDGE), showing why actionable diagnosis must be part of the output contract.

Table 22: Scenario-to-failure summary. Each metric is tied to the PACE target it is meant to close. S0 is a reference check; the main protocol claim is measured by S1–S4 and the PACE average in Table 3.

paired with five prompt configurations: G-Eval-style scalar judging, MT-Bench-style scalar judging, persona-blind judging, persona-aware generic judging, and PACEJUDGE. Details in Appx. C.3. For all protocols that emit diagnostic text, we parse the output into the shared schema when possible. If a protocol does not natively define a required field, the corresponding metric is marked unsupported rather than imputed.

E.2 Scenario-Metric Mapping

We discuss the scenario-metric mapping in Tab. 22.

E.3 PACE Evaluation Scenario Definitions

The main paper reports S0 as a reference scenario and S1–S4 as PACE closure scenarios. Here we provide the exact metric definitions used for the result tables, including family sets, location codes, axis thresholds, harmonic constructions, specificity terms, and paired-counterfactual gates.

S0: General reference. S0 measures balanced GOOD/BAD accuracy on a fixed balanced subset. It is reported for context only and is not included in the PACE average.

S1: Personalized. S1 uses persona-swap records that hold the query and response fixed while replacing the persona with one that contradicts an attestation in the unchanged response. The main metric is persona-source exact match: the judge must identify the failure as a persona conflict rather than merely lowering a generic score.

S2: Compositional. S2 evaluates cross-component defects such as XCOMP_MISMATCH, INTENT_DRIFT, and REDUNDANCY. The main metric requires both a BAD verdict and correct response-field localization.

S3: Evidence-grounded. S3 evaluates both product-evidence grounding and persona-history grounding. The metric checks whether the judge identifies the correct grounding-related defect family while constraining supporting evidence IDs to the provided evidence pool.

S4: Actionable. S4 evaluates whether the judge produces a debugging-ready diagnosis. The main metric requires exact defect-family prediction and correct response-field location.

E.4 Main Table Metrics Definitions

This subsection walks through every column of the two main-text experiment tables – Tab. 3 and the paired-counterfactual Tab. 4 – so the symbols and conjuncts in the captions can be read without flipping between sections. Both tables are scored on the same 3,400-record held-out test set of PACESHOP records (680 GOOD + 2,720 BAD), with the S1 persona-swap slice scored separately on a 680-record held-out swap slice. The PACE

average in Tab. 3 excludes S0 because S0 is a reference GOOD/BAD task rather than a target-specific closure metric.

Notation shared by all three tables. "GOOD" and "BAD" refer to the gold record label. "Family" denotes the gold or predicted `defect_family` field, drawn from the seven-family taxonomy {PREF_CONFLICT, OVER_PERSONALIZATION_HALLUCINATION, XCOMP_MISMATCH, INTENT_DRIFT, REDUNDANCY, EVIDENCE_MISMATCH, UNSUPPORTED_CLAIM} (plus NONE for GOOD records). "Location" denotes the predicted 4-way response-field label {overview, categories, related queries, evidence attributions}. "Axis" refers to the four PACEJUDGE axes {persona_alignment, component_consistency, evidence_grounding, format_safety}, each scored on a 1–5 Likert scale. "PACE target" refers to one of the four target dimensions (P, C, E, A); each defect family belongs to exactly one target.

Tab. 3 metrics definition. Tab. 3 reports seven-backbone means across Opus 4.7, Sonnet 4.6, Sonnet 4.5, Haiku 4.5, Qwen3 32B, GPT-OSS 20B, and GPT-OSS 120B on the 3,400-record held-out test set. S0 is a reference GOOD/BAD task and is not included in the PACE average. S1–S4 evaluate PACE diagnostic closure: persona-source diagnosis, cross-component detection and localization, evidence grounding, and actionable defect-family/location prediction. False-fire measures defect alarms on GOOD records, so methods cannot improve closure scores merely by predicting defects everywhere. Cells marked "–" are unsupported by the method's native judging protocol; unsupported columns are excluded from averages, and rows missing any contributing PACE component report "–" for the PACE average.

The columns are:

- **S0 (Gen.)**: balanced GOOD/BAD accuracy on a 50+50 reference subset. This is a sanity check, not part of the PACE average.
- **S1 (P-source)**: persona-source diagnosis on the held-out persona-swap slice. The score is the fraction of swapped records on which the predicted defect family is PREF_CONFLICT. A judge that marks the record BAD without naming the persona conflict receives no credit.
- **S2 detect**: cross-component defect detection on records whose gold family is in {XCOMP_MISMATCH, INTENT_DRIFT, REDUNDANCY}. The score is the fraction of records marked BAD.
- **S2 loc.**: cross-component localization on the same evaluation slice. The score requires both a BAD verdict and correct 4-way response-field localization.
- **S3 grounding**: harmonic mean of three grounding sub-scores: product-evidence mismatch detection, persona-history hallucination detection, and GOOD-record specificity against evidence-family false alarms. The harmonic mean requires all components to be high, preventing a judge from winning by always firing an evidence-related defect.
- **S4 family**: exact defect-family prediction on single-defect BAD records across all seven defect families.
- **S4 loc.**: exact 4-way response-field localization on the same single-defect BAD slice.
- **False-fire** $\downarrow$: fraction of GOOD records on which the judge predicts any non-NONE defect family. Lower is better.
- **PACE avg.**: arithmetic mean of S1, S2-loc., S3 grounding, S4-family, and $(1 - \text{False-fire})$, with all components aligned so higher is better.

Per-column best scores in the rubric-ablation block are bolded; the mean row provides the headline protocol-level comparison.

Tab. 4 Strict paired-counterfactual metrics definition. The paired-counterfactual columns close the "always cry BAD" loophole left open by single positive-class metrics:

- **S0 (ref. acc.)**: balanced GOOD/BAD accuracy on the 50/50 reference subset; reported for context only and not part of the PACE average.
- **$\mathbf{S1_P}$ (PJC)**: *Joint Pair Credit* on (good, swap) pairs sharing a base record id. The judge must classify the original-persona record as GOOD *and* the swapped-persona record as BAD with family = PREF_CONFLICT and persona_alignment axis < 3.0 on the same paired record. A persona-blind judge cannot pass this gate.
- **$\mathbf{S2_C}$ (J-strict)**: BAD verdict, family $\in \mathcal{D}_{\mathrm{C}}$, correct 4-way location, and component-consistency

axis < 3.0, minus the false-C-family alarm rate on GOOD records.

- **$S3_E$ (J-harmonic)**: harmonic of three E subscores: Youden's J on product-evidence BADs with `supporting_evidence_ids` $\subseteq$ candidate pool; Youden's J on persona-history BADs with the same containment check; and paired specificity (no E-family alarm on GOOD records). The containment check makes "hallucinate evidence IDs" a scored failure mode.
- **$S4_A$ (J-strict)**: exact family + exact 4-way location + target-triage (the lowest relevant PACEJUDGE axis matches the defect's target) + `confidence` ≥ 0.5, minus the confident-false-alarm rate on GOOD records.
- **PACE**: arithmetic mean of $S1_P$, $S2_C$, $S3_E$, and $S4_A$.

The *Comparison level* column distinguishes single-backbone, 7-backbone mean, and 7-backbone self-consistency majority-vote results; single-target adapted baselines (ARES-style E, PersonaLens-style P, EtaPP-style A) report measured Sonnet 4.6 numbers on cells inside their native target and "–" elsewhere. The native judging-protocol map (Tab. 19) and evaluation-target leakage audit (Tab. 20) explain why these baselines remain "–" outside their native target.

E.5 Full Experiments Results

E.5.1 Full Per-Backbone Results

Tab. 23 provides the full per-(method, backbone) breakdown on the 3,400-record held-out test set; the per-method **Mean** row aggregates each method into the corresponding row of the compact main-text Tab. 3. The per-backbone spread, not just the mean, is informative: PACEJUDGE on the strong frontier backbones (Opus 4.7, Sonnet 4.6) reaches PACE 0.71–0.78 with False-fire 0.16–0.20, while open-weight backbones over-fire under every rubric, including PACEJUDGE.

E.5.2 S1 Personalized: Per-Axis and Per-Backbone Breakdown

Tab. 24 reports the full per-backbone breakdown of the persona-swap diagnostic (Finding 4 in Sec. 5), whose 7-backbone mean is the main-text Tab. 5. Each row reports the fraction of swap records in the 680-record held-out persona-swap slice for which the judge predicts `PREF_CONFLICT`, broken down by the four conflict axes (budget, hard constraint, brand affinity, attribute preference); the axis-mean of each method's **Mean** row also populates the S1 P-source column of Tab. 3. The per-backbone spread is informative on its own: PACE-JUDGE on Sonnet 4.5 reaches 0.57–0.89 across axes, but on Qwen3 32B never exceeds 0.07, which is the source of the open-weight gap referenced in the main-text discussion.

E.5.3 S2 Compositional: Detection and Localization Details

Tab. 25 reports the full per-backbone breakdown of the component-consistency evaluation slice (Finding 4 in Sec. 5), whose 7-backbone mean is the main-text Tab. 6 and the same numbers populate the S2 detect/loc. columns of Tab. 3. Each method is evaluated on the 100-record cross-field defect subset (XCOMP_MISMATCH, INTENT_DRIFT, REDUNDANCY); the **detect** column is BAD verdict and the **4-way loc.** column is correct response-field localization.

E.6 Expanded Experiments Findings

This section expands the main-text findings with the detailed per-scenario evidence. The main paper groups the results into four high-level conclusions for readability; here we keep the original six findings to show how each conclusion is supported by scenario-level and audit-level results.

Finding 1: PACE closure separates protocol design from generic quality. S0 shows that several scalar and generic judging prompts can perform reasonably on broad GOOD/BAD discrimination. This confirms that PACESHOP is not simply an impossible binary classification task. However, the gap appears when the metric requires fields that close the PACE targets: persona-source diagnosis, cross-component localization, grounding control, and defect-family/location prediction. Across the seven-backbone mean in Tab. 3, PACEJUDGE achieves the best score on every runnable closure column, including S1 persona-source diagnosis, S2 localization, S3 grounding, S4 defect-family prediction, and the overall PACE average. This supports the main claim that structured shopping-assistant evaluation requires a protocol that asks for the right diagnostic fields, not merely a scalar quality score.

Finding 2: personalized evaluation must be scored as a source of invalidity. S1 is designed to prevent shortcuts. If the metric only

asked whether a swapped-persona record is marked BAD, a persona-blind or over-pessimistic judge could receive credit without actually using the persona. We therefore score exact persona-source diagnosis: the judge must identify the failure as PREF_CONFLICT on records where the query and response are unchanged but the persona is counterfactually swapped. Tab. 5 shows that this source label is not solved uniformly across conflict axes. The detailed per-backbone results in Tab. 24 further show that strong GOOD/BAD discrimination does not guarantee reliable persona-conflict diagnosis. This justifies keeping S1 separate from S0.

Finding 3: compositional and grounding failures expose where generic judges are underspecified. S2 and S3 test whether judges can move beyond detecting that a response is flawed. For S2, generic prompts often recognize that something is wrong, but performance drops when the judge must also localize the failed response component. Tab. 6 separates detection from localization to show this gap directly. For S3, the distinctive requirement is not only product-evidence faithfulness, but also avoiding invented shopper-history claims and hallucinated supporting evidence IDs. This is why grounding is evaluated against both the product evidence pool and the persona history. Together, S2 and S3 show that structured shopping-assistant evaluation requires field-aware and evidence-constrained diagnosis, not only generic helpfulness or RAG-style faithfulness.

Finding 4: actionable evaluation is not the same as scalar evaluation. S4 requires a debugging-ready diagnosis: the judge must identify both the defect family and the response-field location. This is stricter than deciding whether a response is BAD. A scalar judge may provide a plausible overall rationale while still failing to say which family of error occurred or which response component should be fixed. The S4 results show that PACEJUDGE improves this complete diagnosis setting because its protocol explicitly requires family, location, confidence, rationale, and evidence support. This finding isolates the Actionable target: evaluation is useful for developers only when it reports what failed and where.

Finding 5: specificity matters; closure should not come from over-flagging. A high score on S1–S4 would be less meaningful if a judge achieved it by predicting defects on most records. We therefore audit false-fire behavior on GOOD records. The False-fire column in Tab. 3 measures how often a method incorrectly reports defects on GOOD examples, and the strict paired-counterfactual Tab. 4 folds the same specificity check into per-target gates (PJC, Youden's J, J-harmonic) so that a method cannot game closure by always flagging BAD. Together these checks show whether a protocol preserves specificity while detecting structured failures. The results indicate that PACEJUDGE's gains are not explained solely by always flagging BAD cases; rather, the protocol improves the ability to report the intended diagnostic fields.

Finding 6: strict paired-counterfactual metrics preserve the conclusion. Tab. 4 re-scores the setting under stricter paired-counterfactual gates. These metrics reduce loopholes such as always predicting BAD or assigning generic defect labels. For example, S1 requires joint success on paired GOOD and persona-swapped BAD records; S2 requires BAD verdict, correct compositional family, correct location, and axis consistency; S3 combines product-evidence detection, persona-history detection, and GOOD-record specificity with evidence-ID containment; and S4 requires exact family, exact location, axis triage, and confidence thresholding. Under this stricter formulation, PACEJUDGE remains strongest on the PACE average, supporting the same conclusion as the main table: the advantage comes from structured diagnostic closure rather than only lenient positive-class scoring.

E.7 Standard Judges Failure Analysis

E.7.1 Concrete Benchmark Examples

The remaining paragraphs collect record-level galleries that illustrate PACESHOP behaviour end to end. Each record is drawn directly from the released dataset.

Ten Concrete Failure Cases. Tab. 26 walks through ten BAD records with their gold defect family and gold defect location, paired with the PACEJUDGE verdict, axis scores, and predicted family/location. The cases span all four field groups and all seven defect families; they make concrete the structured-output diagnoses that the main text reports as aggregate numbers.

Persona-Swap Conflict Examples. Tab. 27 lists ten persona-swap records and the specific persona

attestation each swap puts in conflict with the unchanged response. The four conflict axes (budget, hard constraint, brand affinity, attribute preference) are represented; each row is a paired GOOD record with a single persona swap that turns it into a gold-BAD record for the S1 scenario.

E.8 Audits, Robustness Checks, and Concrete Examples

The main text reports compact S0–S4 tables. The remaining experiment subsections keep the supporting audits and record-level galleries that explain why those numbers should be read as PACE-closure results rather than ordinary scalar-judge scores. The subsections below check off-target leakage, false fires on GOOD records, per-backbone stability, tier difficulty, qualitative judge outputs, earlier reference runs, and concrete benchmark records that illustrate PACESHOP behaviour end to end.

E.8.1 Difficulty Gradient Analysis

The multi-tier defect design enables a decomposition unavailable to single-defect benchmarks. Per-tier accuracy (computed separately per defect count) reveals two distinct regimes:

1. **Deterministic baseline:** accuracy rises monotonically from single (0.880) to quad (0.975) on the development set, because more defects trigger more rule-based checks.

2. **LLM judges:** tier profiles are flatter for frontier models. Smaller models (Qwen3 32B, Gemma3 27B) show steeper gradients—e.g., Qwen3 32B rises from 0.740 (single) to 0.990 (quad)—suggesting they rely on surface signal accumulation rather than understanding individual fault types.

This two-regime pattern motivates the tiered design: if every judge handled every tier uniformly, the tiered construction would be wasted. Instead, smaller or scalar-prompted judges' steeper tier gradients give us a graded robustness metric a single-tier benchmark cannot provide. Fig. 4 shows per-defect-level accuracy by method on the development set.

E.8.2 Reference Main-Evaluation Table on the Development Set

The per-defect-level discrimination figure and a paired-counterfactual specificity audit are retained for reference. These are not the primary SOTA-failure numbers reported in §5; they are kept because they document the tier-gradient behavior that motivates the multi-tier defect design and the swap-TPR / paired-good specificity that supports the strict paired-counterfactual gates of Tab. 4.

Method	Backbone	S0 Gen.	S1 P-source	S2 C detect	S2 C loc.	S3 E grounding	S4 A family	S4 A loc.	False-fire ↓	PACE avg.
Rubric ablations of the same PACE output contract										
G-Eval										
G-Eval	Opus 4.7	0.87	0.23	0.89	0.86	0.78	0.52	0.87	0.70	0.54
G-Eval	Sonnet 4.6	0.86	0.28	0.79	0.87	0.60	0.71	0.90	0.91	0.51
G-Eval	Sonnet 4.5	0.72	0.26	0.93	0.86	0.84	0.65	0.85	0.46	0.63
G-Eval	Haiku 4.5	0.61	0.20	0.96	0.63	0.42	0.33	0.70	0.75	0.37
G-Eval	Qwen3 32B	0.60	0.00	0.06	0.31	0.00	0.26	0.27	0.01	0.31
G-Eval	GPT-OSS 20B	0.75	0.14	0.99	0.84	0.00	0.48	0.58	0.61	0.37
G-Eval	GPT-OSS 120B	0.71	0.09	0.65	0.55	0.18	0.44	0.61	0.23	0.41
G-Eval	Mean	0.73	0.17	0.75	0.70	0.40	0.48	0.68	0.52	0.45
Scalar										
Scalar	Opus 4.7	0.83	0.12	0.89	0.85	0.85	0.57	0.88	0.68	0.54
Scalar	Sonnet 4.6	0.84	0.17	0.82	0.85	0.48	0.64	0.86	0.93	0.44
Scalar	Sonnet 4.5	0.73	0.29	0.82	0.78	0.87	0.60	0.84	0.38	0.63
Scalar	Haiku 4.5	0.56	0.10	0.98	0.45	0.47	0.30	0.63	0.87	0.29
Scalar	Qwen3 32B	0.62	0.00	0.13	0.16	0.00	0.24	0.27	0.01	0.28
Scalar	GPT-OSS 20B	0.68	0.11	0.96	0.42	0.06	0.42	0.51	0.55	0.29
Scalar	GPT-OSS 120B	0.87	0.08	0.80	0.68	0.79	0.64	0.75	0.16	0.61
Scalar	Mean	0.73	0.12	0.77	0.60	0.50	0.49	0.68	0.51	0.44
Persona-blind										
Persona-blind	Opus 4.7	0.60	–	0.99	0.77	0.56	0.55	0.81	0.97	–
Persona-blind	Sonnet 4.6	0.85	–	0.82	0.85	0.80	0.62	0.87	0.69	–
Persona-blind	Sonnet 4.5	0.85	–	0.82	0.81	0.90	0.61	0.85	0.18	–
Persona-blind	Haiku 4.5	0.62	–	0.96	0.52	0.65	0.31	0.62	0.80	–
Persona-blind	Qwen3 32B	0.62	–	0.15	0.29	0.00	0.27	0.30	0.01	–
Persona-blind	GPT-OSS 20B	0.69	–	0.85	0.44	0.11	0.42	0.58	0.48	–
Persona-blind	GPT-OSS 120B	0.92	–	0.91	0.83	0.42	0.66	0.83	0.19	–
Persona-blind	Mean	0.74	–	0.79	0.64	0.49	0.49	0.69	0.47	–
Persona-aware										
Persona-aware	Opus 4.7	0.77	0.07	0.91	0.85	0.91	0.62	0.86	0.45	0.60
Persona-aware	Sonnet 4.6	0.85	0.22	0.83	0.84	0.86	0.70	0.85	0.56	0.61
Persona-aware	Sonnet 4.5	0.71	0.34	0.79	0.75	0.79	0.51	0.76	0.34	0.61
Persona-aware	Haiku 4.5	0.59	0.37	0.96	0.65	0.72	0.45	0.71	0.76	0.49
Persona-aware	Qwen3 32B	0.64	0.04	0.29	0.19	0.00	0.22	0.23	0.01	0.29
Persona-aware	GPT-OSS 20B	0.77	0.24	0.87	0.48	0.18	0.39	0.58	0.37	0.38
Persona-aware	GPT-OSS 120B	0.84	0.14	0.85	0.72	0.77	0.63	0.75	0.13	0.62
Persona-aware	Mean	0.74	0.20	0.79	0.64	0.60	0.50	0.68	0.37	0.51
PACEJUDGE										
PACEJUDGE	Opus 4.7	0.91	0.22	0.96	0.88	0.96	0.67	0.91	0.16	0.71
PACEJUDGE	Sonnet 4.6	0.96	0.56	0.95	0.90	0.93	0.69	0.91	0.20	0.78
PACEJUDGE	Sonnet 4.5	0.55	0.78	1.00	0.86	0.73	0.64	0.87	0.85	0.63
PACEJUDGE	Haiku 4.5	0.61	0.22	1.00	0.70	0.67	0.49	0.72	0.73	0.47
PACEJUDGE	Qwen3 32B	0.83	0.05	0.73	0.50	0.08	0.40	0.56	0.04	0.40
PACEJUDGE	GPT-OSS 20B	0.60	0.28	0.96	0.46	0.39	0.50	0.66	0.80	0.37
PACEJUDGE	GPT-OSS 120B	0.73	0.14	0.91	0.73	0.61	0.74	0.82	0.52	0.54
PACEJUDGE	Mean	**0.74**	**0.32**	**0.93**	**0.72**	**0.62**	**0.59**	**0.78**	0.47	**0.56**
Honest-adapted single-target baselines (cells shown only where the native judging protocol supports the metric)										
ARES-style (Saad-Falcon et al., 2024)										
ARES-style (Saad-Falcon et al., 2024)	Opus 4.7	0.74	–	–	–	0.95	–	–	0.02	–
ARES-style (Saad-Falcon et al., 2024)	Sonnet 4.6	0.69	–	–	–	0.32	–	–	0.02	–
ARES-style (Saad-Falcon et al., 2024)	Sonnet 4.5	0.74	–	–	–	0.93	–	–	0.23	–
ARES-style (Saad-Falcon et al., 2024)	Haiku 4.5	0.79	–	–	–	0.45	–	–	0.28	–
ARES-style (Saad-Falcon et al., 2024)	Qwen3 32B	0.61	–	–	–	0.00	–	–	0.31	–
ARES-style (Saad-Falcon et al., 2024)	GPT-OSS 20B	0.55	–	–	–	0.00	–	–	0.69	–
ARES-style (Saad-Falcon et al., 2024)	GPT-OSS 120B	0.68	–	–	–	0.08	–	–	0.29	–
ARES-style (Saad-Falcon et al., 2024)	Mean	0.69	–	–	–	0.39	–	–	0.26	–
PersonaLens-style (Zhao et al., 2025b)										
PersonaLens-style (Zhao et al., 2025b)	Opus 4.7	0.68	0.06	–	–	–	–	–	0.00	–
PersonaLens-style (Zhao et al., 2025b)	Sonnet 4.6	0.72	0.04	–	–	–	–	–	0.00	–
PersonaLens-style (Zhao et al., 2025b)	Sonnet 4.5	0.70	0.40	–	–	–	–	–	0.06	–
PersonaLens-style (Zhao et al., 2025b)	Haiku 4.5	0.63	0.41	–	–	–	–	–	0.34	–
PersonaLens-style (Zhao et al., 2025b)	Qwen3 32B	0.66	0.11	–	–	–	–	–	0.01	–
PersonaLens-style (Zhao et al., 2025b)	GPT-OSS 20B	0.56	0.18	–	–	–	–	–	0.06	–
PersonaLens-style (Zhao et al., 2025b)	GPT-OSS 120B	0.60	0.08	–	–	–	–	–	0.00	–
PersonaLens-style (Zhao et al., 2025b)	Mean	0.65	0.18	–	–	–	–	–	**0.07**	–
EtaPP-style (Hao et al., 2025)										
EtaPP-style (Hao et al., 2025)	Opus 4.7	0.78	–	–	–	–	0.60	0.91	0.50	–
EtaPP-style (Hao et al., 2025)	Sonnet 4.6	0.81	–	–	–	–	0.55	0.90	0.35	–
EtaPP-style (Hao et al., 2025)	Sonnet 4.5	0.50	–	–	–	–	0.65	0.95	0.99	–
EtaPP-style (Hao et al., 2025)	Haiku 4.5	0.56	–	–	–	–	0.50	0.84	0.88	–
EtaPP-style (Hao et al., 2025)	Qwen3 32B	0.73	–	–	–	–	0.34	0.41	0.05	–
EtaPP-style (Hao et al., 2025)	GPT-OSS 20B	0.85	–	–	–	–	0.57	0.75	0.22	–
EtaPP-style (Hao et al., 2025)	GPT-OSS 120B	0.72	–	–	–	–	0.32	0.44	0.00	–
EtaPP-style (Hao et al., 2025)	Mean	0.71	–	–	–	–	0.50	0.74	0.43	–

Table 23: **Full per-backbone results.** Companion to Tab. 3, which reports the 7-backbone mean. Each method is run on seven backbones (Opus 4.7, Sonnet 4.6, Sonnet 4.5, Haiku 4.5, Qwen3 32B, GPT-OSS 20B, GPT-OSS 120B) on the 3,400-record held-out test set. **S0**: balanced GOOD/BAD accuracy on a 50+50 reference subset. **S1**: predicted family = `PREF_CONFLICT` on the persona-swap slice. **S2 detect / loc.**: BAD verdict and 4-way response-field localization on the cross-component defect subset (XCOMP_MISMATCH, INTENT_DRIFT, REDUNDANCY). **S3 grounding**: harmonic mean of (i) BAD with exact `EVIDENCE_MISMATCH` family on the product-evidence subset, (ii) BAD with exact `OVER_PERSONALIZATION_HALLUCINATION` family on the persona-history subset, (iii) no E-family alarm on GOOD records (specificity). **S4 family / loc.**: exact family and 4-way location on single-defect BAD records. **False-fire ↓**: fraction of GOOD records on which the judge fired any non-NONE defect family (lower is better). **PACE**: mean of S1, S2-loc., S3 grounding, S4-family, and $(1 - \text{False-fire})$, all aligned higher = better. "–" marks cells the method's native judging protocol does not define. Per-method best mean is bolded on the PACEJUDGE row.

Prompt	Backbone	Persona-swap accuracy by axis			
		budget	hard	brand	attr.
G-Eval	Opus 4.7	0.29	0.32	0.16	0.13
G-Eval	Sonnet 4.6	0.51	0.21	0.20	0.20
G-Eval	Sonnet 4.5	0.38	0.24	0.29	0.14
G-Eval	Haiku 4.5	0.18	0.15	0.21	0.26
G-Eval	Qwen3 32B	0.00	0.00	0.00	0.00
G-Eval	GPT-OSS 20B	0.16	0.12	0.18	0.11
G-Eval	GPT-OSS 120B	0.17	0.09	0.08	0.02
G-Eval	Mean	0.24	0.16	0.16	0.12
Persona-aware	Opus 4.7	0.14	0.11	0.01	0.01
Persona-aware	Sonnet 4.6	0.58	0.16	0.06	0.10
Persona-aware	Sonnet 4.5	0.48	0.53	0.19	0.15
Persona-aware	Haiku 4.5	0.30	0.50	0.38	0.31
Persona-aware	Qwen3 32B	0.02	0.05	0.04	0.03
Persona-aware	GPT-OSS 20B	0.28	0.29	0.27	0.12
Persona-aware	GPT-OSS 120B	0.29	0.14	0.05	0.06
Persona-aware	Mean	0.30	0.25	0.14	0.11
PACEJUDGE	Opus 4.7	0.29	0.34	0.16	0.09
PACEJUDGE	Sonnet 4.6	0.73	0.48	0.51	0.54
PACEJUDGE	Sonnet 4.5	0.88	0.77	0.89	0.57
PACEJUDGE	Haiku 4.5	0.22	0.23	0.29	0.14
PACEJUDGE	Qwen3 32B	0.02	0.05	0.07	0.04
PACEJUDGE	GPT-OSS 20B	0.25	0.27	0.41	0.21
PACEJUDGE	GPT-OSS 120B	0.22	0.09	0.17	0.05
PACEJUDGE	**Mean**	**0.37**	**0.32**	**0.36**	**0.23**

Table 24: **Full per-backbone persona-source diagnosis by conflict axis.** Companion to Tab. 5, which reports the 7-backbone mean. The evaluation uses the 680-record held-out persona-swap slice. A hit means the judge predicts `PREF_CONFLICT`, the gold source family induced by the counterfactual persona swap. The **Mean** row aggregates the seven backbones (Opus 4.7, Sonnet 4.6, Sonnet 4.5, Haiku 4.5, Qwen3 32B, GPT-OSS 20B, GPT-OSS 120B) into the corresponding row of the main-text Tab. 5.

Prompt	Backbone	detect	4-way loc.
G-Eval	Opus 4.7	0.89	0.86
G-Eval	Sonnet 4.6	0.79	0.87
G-Eval	Sonnet 4.5	0.93	0.86
G-Eval	Haiku 4.5	0.96	0.63
G-Eval	Qwen3 32B	0.06	0.31
G-Eval	GPT-OSS 20B	0.99	0.84
G-Eval	GPT-OSS 120B	0.65	0.55
G-Eval	Mean	0.75	0.70
Persona-aware	Opus 4.7	0.91	0.85
Persona-aware	Sonnet 4.6	0.83	0.84
Persona-aware	Sonnet 4.5	0.79	0.75
Persona-aware	Haiku 4.5	0.96	0.65
Persona-aware	Qwen3 32B	0.29	0.19
Persona-aware	GPT-OSS 20B	0.87	0.48
Persona-aware	GPT-OSS 120B	0.85	0.72
Persona-aware	Mean	0.79	0.64
PACEJUDGE	Opus 4.7	0.96	0.88
PACEJUDGE	Sonnet 4.6	0.95	0.90
PACEJUDGE	Sonnet 4.5	1.00	0.86
PACEJUDGE	Haiku 4.5	1.00	0.70
PACEJUDGE	Qwen3 32B	0.73	0.50
PACEJUDGE	GPT-OSS 20B	0.96	0.46
PACEJUDGE	GPT-OSS 120B	0.91	0.73
PACEJUDGE	**Mean**	**0.93**	**0.72**

Table 25: **Full per-backbone detection and 4-way localization** on the 100-record cross-field defect subset (XCOMP_MISMATCH, INTENT_DRIFT, REDUNDANCY). Companion to Tab. 6, which reports the 7-backbone mean. The **Mean** row aggregates the seven backbones (Opus 4.7, Sonnet 4.6, Sonnet 4.5, Haiku 4.5, Qwen3 32B, GPT-OSS 20B, GPT-OSS 120B) into the corresponding row of the main-text Tab. 6; the same numbers populate the S2 detect / loc. columns of Tab. 3.

Case	Record (query / gold)	Scalar baseline verdict	PACEJUDGE verdict
PS	PS `full-good-03170::swap-brand` query: air conditioner gold: PREF_CONFLICT @ persona_swap	GOOD, overall 4.0; *'Response is well-structured, aligns with persona's portability/smart-home preferences, and evidence IDs are valid. However, the overview claims all options 'use latex-free construc'*	BAD; [P 2.5, C 3.5, G 2.5, F 4.5]; defect: UNSUPPORTED_CLAIM @ overview_text
PS	PS `full-good-03171::swap-attribute` query: necklace gold: PREF_CONFLICT @ persona_swap	GOOD, overall 3.5; *'Categories map to evidence and format checks pass. However, the overview claims necklaces are 'latex-free' and suitable for sensitive skin without any supporting evidence in produc'*	BAD; [P 2.0, C 3.0, G 4.0, F 5.0]; defect: OVER_PERSONALIZATION_HALLUCINATION @ overview_text
PS	PS `full-good-03173::swap-hard` query: necklaces gold: PREF_CONFLICT @ persona_swap	GOOD, overall 3.5; *'Response mentions 'latex-free' as a feature, but the persona's actual hard constraint (per counterfactual swap) is 'avoid peanuts in ingredients'—latex-free is not the active const'*	BAD; [P 2.0, C 3.0, G 3.5, F 5.0]; defect: PREF_CONFLICT @ overview_text
CC	CC `full-bad-plan-02515` query: wine glasses bulk gold: XCOMP_MISMATCH @ rqList	GOOD, overall 4.0; *'Response is well-formatted, grounded in evidence, and addresses the bulk wine glasses query with value-oriented options. However, the rqList is heavily skewed toward 'stemless' var'*	BAD; [P 2.5, C 2.5, G 4.0, F 5.0]; defect: XCOMP_MISMATCH @ rqList
GR	GR `full-bad-plan-00068` query: ashwaghanda extract gold: EVIDENCE_MISMATCH @ attribution_list	BAD, overall 2.2; *'All attribution IDs (B0812DKDD9, B07TXM7K4T, B082GTFJB5) are not in the candidate evidence set, failing evidence_id_validity and exact_match checks. The response fabricates evidenc'*	BAD; [P 3.0, C 3.5, G 1.5, F 2.0]; defect: EVIDENCE_MISMATCH @ attribution_list
GR	GR `full-bad-plan-00397` query: duffle bag gold: EVIDENCE_MISMATCH @ attribution_list	BAD, overall 2.5; *'Attribution list includes an invalid evidence ID (M:B073G3YH4W:0) not in candidate evidence, failing evidence_id_validity and exact match checks. Otherwise content is coherent and '*	BAD; [P 3.5, C 3.5, G 2.0, F 2.5]; defect: EVIDENCE_MISMATCH @ attribution_list
MD	MD `full-bad-plan-09120` query: ashwagandha liquid gold: EVIDENCE_MISMATCH+OVER_PERSONALIZATION_HALLUCINATION+PREF_CONFLICT @ productCategoryList[0]+attribution_list+overview_text	BAD, overall 1.8; *'Overview hallucinates a Dyson Airwrap not in persona history, is truncated mid-sentence, and includes a meaningless 'Conflict Pick' category. Attribution lists IDs not in c'*	BAD; [P 2.0, C 1.5, G 1.0, F 2.0]; defect: UNSUPPORTED_CLAIM @ overview_text
MD	MD `full-bad-plan-09184` query: wet rated ceiling fan without light gold: EVIDENCE_MISMATCH+PREF_CONFLICT+UNSUPPORTED_CLAIM @ productCategoryList[0]+attribution_list+overview_text	BAD, overall 2.0; *'Attribution list contains IDs not in candidate evidence set (all invalid). Overview text is truncated mid-word ('highli'). 'Conflict Pick' category is a meta/placeholder label, not'*	BAD; [P 2.5, C 2.0, G 1.0, F 1.5]; defect: EVIDENCE_MISMATCH @ attribution_list

Table 26: Ten concrete failure cases. Each row contrasts a scalar baseline's verdict and reasoning against PACEJUDGE's structured output on the same record. Axis scores abbreviated P/C/G/F.

Swap ID	Record	Original persona (excerpt)	Swapped persona (excerpt)	Conflict in (unchanged) response
PS-ex-01	full-good-00042 (hard_constraint swap)	hard_constraints=[avoid peanuts, quiet operation]	hard_constraints=[avoid leather, quiet operation]	Overview bolds *leather handbags* category; categories list includes Leather Tote.
PS-ex-02	full-good-00173 (brand swap)	brand_likes=[Ecco Press, Lansinoh]	brand_likes=[Kindle, Fire TV]; brand_avoids=[Ecco Press]	Evidence attributions cite three Ecco Press products; overview praises the brand.
PS-ex-03	full-good-00289 (attribute swap)	attribute_preferences=[eco-conscious materials]	attribute_preferences=[luxury finishes]	Overview markets *sustainability* as primary selling point.
PS-ex-04	full-good-00311 (budget swap)	budget=mid-range	budget=premium	rqList refines by *budget-friendly* variants inconsistent with premium context.
PS-ex-05	full-good-00484 (hard_constraint swap)	hard_constraints=[avoid peanuts]	hard_constraints=[avoid dairy, avoid eggs]	Overview bolds a *cookie dough* category with egg content.
PS-ex-06	full-good-00605 (attribute swap)	attribute_preferences=[portability]	attribute_preferences=[heavy-duty durability]	rqList refines by *lightweight packable*; overview emphasizes portability.
PS-ex-07	full-good-00732 (brand swap)	brand_likes=[Nexpak]	brand_avoids=[Nexpak]	Four of ten attributions cite Nexpak products.
PS-ex-08	full-good-00841 (hard_constraint swap)	hard_constraints=[compatibility required]	hard_constraints=[safe for children under 3]	Overview recommends a small-parts product incompatible with the new constraint.
PS-ex-09	full-good-00978 (budget swap)	budget=premium	budget=budget-conscious	Overview frames premium finish as the primary selling point.
PS-ex-10	full-good-01124 (attribute swap)	attribute_preferences=[smart-home compatibility]	attribute_preferences=[no Wi-Fi devices]	rqList refines by *Wi-Fi / Alexa integration*.

Table 27: Ten representative persona-swap records. Each swap flips one axis (budget, hard_constraint, brand_affinity, attribute_preference) to a value drawn from the 1,200-persona pool that directly contradicts an attestation in the unchanged response; the gold label flips from GOOD to BAD.

Backbone	$S1_P$ paired		$S2_C$		$S3_E$			$S4_A$	
	swap-TPR	paired-good-spec.	TPR	FPR	J_a	J_b	spec.	TPR	FPR
TPR / specificity / FPR breakdown of paired-counterfactual metrics; see Table 4 for the J columns									
G-Eval (Liu et al., 2023)									
Opus 4.7	0.06	0.80	0.84	0.22	0.21	0.00	0.53	0.35	0.70
Sonnet 4.6	0.09	0.91	0.59	0.06	0.29	0.00	0.17	0.51	0.91
Sonnet 4.5	0.20	0.56	0.65	0.05	0.40	0.10	0.67	0.38	0.46
Haiku 4.5	0.08	0.26	0.61	0.44	0.50	0.05	0.80	0.19	0.75
Qwen3 32B	0.00	0.99	0.01	0.00	0.46	0.00	0.99	0.17	0.01
GPT-OSS 20B	0.06	0.39	0.49	0.27	0.17	0.00	0.72	0.19	0.57
GPT-OSS 120B	0.01	0.81	0.19	0.07	0.37	0.04	0.84	0.30	0.23
Scalar (MT-Bench-style) (Zheng et al., 2023)									
Opus 4.7	0.02	0.70	0.83	0.16	0.22	0.00	0.50	0.42	0.68
Sonnet 4.6	0.05	0.89	0.64	0.07	0.37	0.00	0.16	0.47	0.93
Sonnet 4.5	0.26	0.64	0.56	0.08	0.45	0.02	0.81	0.39	0.38
Haiku 4.5	0.08	0.12	0.44	0.11	0.43	0.00	0.25	0.27	0.87
Qwen3 32B	0.00	0.99	0.07	0.00	0.47	0.00	0.99	0.19	0.01
GPT-OSS 20B	0.09	0.38	0.22	0.33	0.32	0.00	0.84	0.16	0.55
GPT-OSS 120B	0.02	0.73	0.50	0.06	0.59	0.09	0.90	0.43	0.16
Persona-blind									
Opus 4.7	0.00	0.15	0.77	0.00	0.12	0.00	0.03	0.42	0.97
Sonnet 4.6	0.05	0.87	0.68	0.13	0.20	0.00	0.44	0.36	0.69
Sonnet 4.5	0.06	0.82	0.63	0.06	0.49	0.13	0.88	0.36	0.18
Haiku 4.5	0.26	0.21	0.52	0.12	0.49	0.00	0.67	0.23	0.80
Qwen3 32B	0.00	0.99	0.09	0.00	0.43	0.00	0.99	0.18	0.01
GPT-OSS 20B	0.19	0.46	0.20	0.29	0.42	0.01	0.88	0.22	0.45
GPT-OSS 120B	0.08	0.76	0.41	0.03	0.55	0.03	0.84	0.47	0.19
Persona-aware									
Opus 4.7	0.05	0.61	0.84	0.07	0.28	0.09	0.62	0.38	0.45
Sonnet 4.6	0.11	0.88	0.64	0.07	0.22	0.00	0.53	0.39	0.56
Sonnet 4.5	0.32	0.63	0.48	0.03	0.30	0.08	0.78	0.34	0.34
Haiku 4.5	0.37	0.23	0.63	0.14	0.58	0.22	0.83	0.28	0.76
Qwen3 32B	0.02	0.99	0.08	0.00	0.44	0.00	1.00	0.17	0.01
GPT-OSS 20B	0.22	0.59	0.29	0.15	0.32	0.01	0.87	0.21	0.35
GPT-OSS 120B	0.12	0.75	0.35	0.01	0.50	0.03	0.88	0.41	0.13
PACEJudge									
Opus 4.7	0.23	0.84	0.84	0.09	0.55	0.23	0.93	0.32	0.16
Sonnet 4.6	0.26	0.90	0.85	0.04	0.51	0.08	0.88	0.52	0.20
Sonnet 4.5	0.80	0.15	0.83	0.07	0.25	0.22	0.36	0.38	0.85
Haiku 4.5	0.14	0.27	0.69	0.14	0.41	0.39	0.45	0.34	0.73
Qwen3 32B	0.02	0.96	0.18	0.01	0.51	0.01	0.98	0.23	0.04
GPT-OSS 20B	0.24	0.21	0.34	0.20	0.23	0.10	0.48	0.29	0.67
GPT-OSS 120B	0.09	0.39	0.47	0.01	0.32	0.20	0.49	0.48	0.52

Table 28: **Paired-counterfactual specificity audit.** $S1_P$ is reported as the swap-side TPR and the paired GOOD-side specificity (judge correctly returns GOOD on the original-persona record sharing the same base id). Persona-aware on Sonnet 4.6 retains a high swap-TPR but a low paired-good specificity; that is exactly the "always cry BAD-PREF_CONFLICT" failure mode the paired metric rules out. S2 / S4 columns show TPR on the BAD subset and FPR on the same-CSV GOOD subset. S3 columns show J_a (product evidence), J_b (persona-history grounding), and the paired specificity rate (no E-family alarm on GOOD). These quantities are the per-rubric inputs to the harmonic in Tab. 4.