FROM GENERATION TO DECISION: PROPER SCORING RULES FOR TYPE-SAFE USER ACTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

User simulation must reproduce behavior on interfaces that mix discrete choices with open-ended text. Existing language-model simulators generate action strings for both, coupling finite choices to token decoding. We introduce **CADET** (Calibrated Decisions over Types), a hybrid simulator that matches its computation to the interaction. A pretrained LM encoder feeds a typed decision head that predicts joint operation-subtype probabilities in one forward pass. An element selector resolves the target, and the predicted operation gates a language generator for text-bearing actions. Proper scoring rules train these probabilities to recover behavioral distributions beyond the most likely action. We evaluate search and category-based browse on two e-commerce benchmarks. On one benchmark, CADET-Brier reduces expected calibration error by 27–58% across seven frozen backbones relative to the best regularized exact-match baseline, with reductions of 71–79% across four LoRA-adapted backbones. On the second benchmark, LoRA CADET achieves 46.7% leaf accuracy, 9.7 percentage points above the majority baseline.

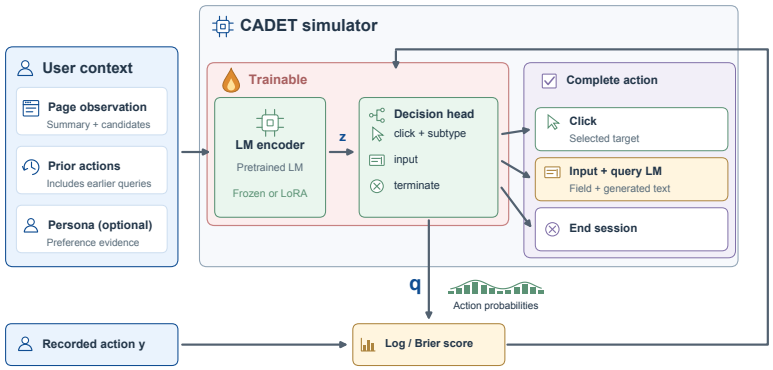

Figure 1: The CADET-based simulator composes a pretrained encoder and a proper-scored typed head with an element selector and a frozen text generator, invoked only for input actions. Appendix W specifies the training and routing boundaries.

1 INTRODUCTION

User simulation on the web must reproduce behavior across interfaces that combine selecting visible targets with entering text (Deng et al., 2023; Lù et al., 2024). We call these *heterogeneous interaction surfaces*. E-commerce provides a concrete testbed: shoppers select products, browse categories, and enter search queries. Simulating these interactions requires both a distribution over choices and the text that some actions need.

Existing shopping simulators prompt or fine-tune language models to generate the next action (Wang et al., 2025; Zhang et al., 2026). Even a finite choice among visible elements then requires token-by-token decoding and parsing back into a valid action. The learning objective raises a separate problem: rewarding exact matches can recover a most likely action without identifying how of-

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

ten plausible alternatives occur (Gneiting & Raftery, 2007). Efficient action selection and faithful behavioral probabilities are therefore distinct requirements.

We introduce **CADET** (**Ca**librated **De**cisions over **T**ypes), a hybrid simulator built on a pretrained LM encoder. A typed head scores discrete choices directly, and a text generator supplies open-ended input when needed. We evaluate this decision interface on search and category-based browse, using two e-commerce benchmarks and comparisons with a Public decision model API and direct/reasoning LMs.

CADET trains the typed probabilities with proper scoring rules. Sampled exact-match reward induces a linear score that concentrates predictions at modes (Gneiting & Raftery, 2007); it can attain Bayes accuracy while leaving behavioral probabilities unidentified. Its logit gradients also weaken for low-probability observations. Section 3.3 separates these identification and optimization effects, motivating our tests of action fidelity, probability quality, and rare-action learning.

Our key contributions are:

- **Direct-decision and hybrid user simulation.** To the best of our knowledge, this is the first web user-simulation study to use a nonautoregressive typed-decision model and to integrate it with on-demand language generation in a hybrid simulator.
- **Learning action distributions.** We train typed decision heads with strictly proper scoring objectives and analyze how exact-match reward affects probability recovery, calibration, and rare-action learning.
- **Search-and-browse evaluation.** We introduce category-choice browse tasks alongside next-action search prediction and compare trained and training-free simulators on fidelity, probability quality, and inference efficiency.

2 Background and Related Work

Behavioral simulation. Human-Computer Interaction (HCI) research on direct manipulation emphasizes acting on visible objects through selection and immediately observable feedback (Shneiderman, 1983). Modern web benchmarks capture the combination of such choices with text entry: Mind2Web represents actions as target–operation pairs, with values for typing and option selection (Deng et al., 2023); WebLINX records clicks, text input, scrolling, and other browser interactions (Lù et al., 2024).

Shopping simulators train language models to generate these actions. Customer-R1 uses GRPO (Wang et al., 2025; Shao et al., 2024b); Shop-R1 combines a difficulty-aware action reward with a rationale reward (Zhang et al., 2026). Their action-level feedback motivates our comparison of scoring objectives for behavioral probabilities.

OPeRA (Wang et al., 2026) provides 692 shopping sessions from 51 users with per-user persona surveys. Its filtered benchmark uses three action types, $\{\mathsf{click}, \mathsf{input}, \mathsf{terminate}\}$, with 13 click subtypes. OPeRA and Wang et al. (2025) evaluate full next-action accuracy by matching the action tuple, including the target element and, for inputs, the typed text. Our objective comparisons evaluate the (type, subtype) leaf, using Customer-R1's *Fine-grained Type Accuracy* taxonomy. OPeRA instead reports click-subtype weighted F1. Section 4.1 reports complete-action prediction separately.

Decision models. We use *decision model* as a context-conditioned predictor that assigns probabilities directly to a declared set of choices, combining a pretrained LM encoder with a learned output head (Devlin et al., 2019). Unlike autoregressive action predictors such as Decision Transformer (Chen et al., 2021) and Gato (Reed et al., 2022), our finite-choice head scores the declared action labels in one forward pass, and a separate generator supplies text-bearing actions.

Calibration via proper scoring rules. A scoring rule is strictly proper when the true distribution uniquely maximizes its expected score (Gneiting & Raftery, 2007; Savage, 1971; Buja et al., 2005); the logarithmic (Good, 1952) and Brier (Brier, 1950) rules are canonical, and strict propriety elicits truthful probabilities while calibration measures forecast reliability (Dawid, 1982). Proper scores have recently rewarded a model's stated scalar confidence (Bani-Harouni et al., 2026; Xu et al., 2024; Damani et al., 2026; Turtel et al., 2025; 2026) and trained generators (Shao et al., 2024a). CADET instead applies categorical heads and proper scores to typed human actions, comparing them with bare, weighted, and regularized sampled exact-match objectives.

Constrained decoding. Grammar-constrained methods (Geng et al., 2023; Willard & Louf, 2023; Beurer-Kellner et al., 2024; Dong et al., 2025) enforce output syntax. Probability quality depends additionally on the scoring and training procedure. Renormalization over a restricted label set can increase overconfidence (Badhe et al., 2026); we evaluate calibration directly on the declared typed support.

Decision-model APIs expose probabilities over declared alternatives. Appendix Q reviews these interfaces and published checks of their probability outputs.

3 METHOD

3.1 A TYPED ACTION SPACE

We first define the typed decision component, then compose its output with a target selector and text module to form complete actions. The OPeRA search taxonomy instantiates the typed support below; Figure 2 contrasts it with the category-choice browse interface.

Let o denote the page observation, h the prior interaction history, and p the persona representation. The objective comparisons predict a *leaf* $\ell = (t, c)$: action type t and subtype c. The support is

$$\mathcal{L} = \coprod_{t \in \mathcal{T}} \{t\} \times \mathcal{C}_t, \qquad \pi_\theta(t, c \mid o, h, p) = \pi_\theta(t \mid o, h, p)\pi_\theta(c \mid t, o, h, p). \tag{1}$$

OPeRA has 13 click subtypes and singleton subtype sets for input and terminate, giving 15 leaves. Normalization within each declared support ensures a valid type–subtype pair in one forward pass (Figure 1). The reported out-of-schema checks concern these leaves.

The simulator composes four modules: an encoder giving pooled features z of (o, h, p) (frozen, or LoRA-adapted); the leaf head $\pi_\theta(t, c \mid z)$; an element selector $\pi_\phi(e \mid z_{o,h}, E(o))$ that scores the page's candidate elements against persona-free context features; and, for input actions, a text module that emits the payload. The element selector is trained by a log score on its own parameters and does not condition on the predicted type; the two are joined only at inference. Section 4.1 evaluates this composition. The training-free comparison selects complete actions from supplied menus.

3.2 DECISION AND TEXT GENERATION

The complete action combines a typed operation, its target, and any required text. Figure 2 shows how the two task interfaces use this decision component.

Search. The output is a click, an input with a field and text payload, or termination. Clickable search suggestions are ordinary target choices. In the implemented hybrid, a predicted input invokes a frozen instruction-tuned generator conditioned on the page summary, prior actions, and persona. The generator uses greedy decoding; its text does not backpropagate into the decision model. Appendix W gives the routing implementation and its evaluation against a no-text baseline.

Browse. At a recorded search step, the model selects a leaf from a fixed product-category taxonomy using page context and prior behavior. Neither dataset labels the category a shopper searched in, so we construct the browse target: for each dataset we build a three-level dependent taxonomy (department $l_1 \rightarrow l_2 \rightarrow$ leaf l_3) grounded in its unique search queries (500 for OPeRA-filtered, 53,232 for ShopCART), with l_1 mirroring departments of popular e-commerce companies. OPeRA has 19 departments and 185 leaves, ShopCART has 29 and 218; every node carries an `other` child and an `unknown` department, so no query is inexpressible. The per-query gold is derived, not human-annotated: OPeRA queries are hand-labeled to a leaf by an LLM annotator (80.2% receive a specific leaf) and the ShopCART queries by embedding nearest-leaf (`all-MiniLM-L6-v2`), so browse is measured as category accuracy on its own cohort. All models within a comparison receive the same taxonomy and use the same scorer.

Frozen-feature training fits the typed head; LoRA also adapts the encoder. Both hold the resulting parameters fixed at evaluation. The next section explains which objective trains the reported probabilities.

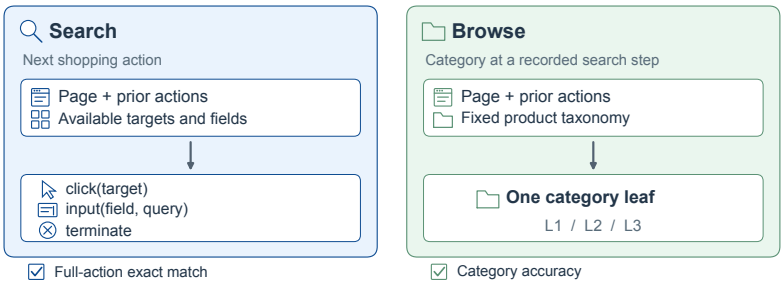

Figure 2: Search predicts a complete shopping action; browse predicts a category at a recorded search step. Their respective metrics are full-action exact match and category accuracy.

3.3 Exact-match reward and distribution identification

We analyze expected sampled-action reward at a fixed state. Averaging over the policy action gives the linear score $S(\pi, y) = \pi(y)$ (Gneiting & Raftery, 2007, Sections 3.1 and 4.1), which the controlled experiments optimize directly. The following proposition gives its finite-action specialization. Evaluation uses top-1 readout with a fixed tie rule.

Proposition 1 (Exact-match reward does not identify the conditional)**.** *Fix a state s, write $P_s^* = P^*(\cdot \mid s)$ and let π_s be the reported policy. Draw $A \sim \pi_s$ and $Y \sim P_s^*$ independently conditional on s. The sampled reward $\mathbf{1}[A = Y]$ has expectation $\langle \pi_s, P_s^* \rangle$, linear in π_s, so its maximizers are exactly the distributions supported on $\mathcal{M}(s) = \arg\max_a P_s^*(a)$. The maximizer is therefore non-unique whenever $|\mathcal{M}(s)| > 1$, and any two conditionals with the same modal set induce the same maximizing set. This set depends on P_s^* only through $\mathcal{M}(s)$; an individual optimum need not reveal every tied mode. Class weights w retain linearity and relocate maximizers to $\arg\max_a w(a)P_s^*(a)$. (Proof: Appendix A.)*

Corollary 1 (Impropriety costs identification, not accuracy)**.** *Let π^{em} be any maximizer of the exact-match objective and $\pi^{\mathrm{proper}} = P^*$ the maximizer of a strictly proper rule. Their top-1 predictions lie in $\mathcal{M}(s)$ for every s, so both attain the Bayes accuracy $\mathbb{E}_s \max_a P^*(a \mid s)$. Their optima are therefore* indistinguishable by accuracy*, while their reported probabilities can differ.*

Entropy and KL regularization retain information about the conditional through invertible transformations, but generally report a different distribution (Appendix A.3). Finite-capacity optimization also matters: low-probability labels can receive weak updates.

Proposition 2 (Exact-match logit gradients vanish on unlikely labels)**.** *Let $\pi = \mathrm{softmax}(z)$ and use Euclidean norms. In the low-probability limit, $\|\nabla_z \pi(a^*)\| = \Theta(\pi(a^*))$: observations assigned low probability have vanishing logit gradients. The log score's gradient norm is bounded below by $1 - \pi(a^*)$ and remains large in this regime. (Proof: Appendix A.)*

These results distinguish population identification from finite-model optimization. Section 4.2 tests both effects; Appendix A gives proofs and numerical checks.

3.4 Personalization

Persona and history condition the action distribution on the user. A KL decomposition predicts that replacing a true persona with an independent persona costs at least as much as averaging over personas (Proposition 3, Appendix A). We test persona removal, mismatch, and measurement noise to identify when the available user information improves prediction.

3.5 Objectives

All objective variants share an architecture and feature representation. **CADET** trains the typed head with a strictly proper scoring rule: the log score $S_{\log}(q, y) = \log q_y$ by default, with the negative Brier loss $S_{\mathrm{Br}}(q, y) = -\sum_i (q_i - \mathbf{1}[i = y])^2$ as a second proper score. Comparison arms use exact-match reward, either bare (C-exact), with class reweighting (the reported $2000\times$ arm), with an entropy bonus, or with a KL penalty to the training marginal. Appendix O records per-arm selection and the available weight specification.

The controlled study directly optimizes each objective's expected score, including q_y for exact-match reward. Every arm uses the same top-1 readout at evaluation, isolating the effect of the objective on its learned distribution.

Entropy regularization appears both in policy-gradient methods (Williams & Peng, 1991; Mnih et al., 2016; Ahmed et al., 2019) and in the supervised confidence penalty of Pereyra et al. (2017). It provides a standard comparison for separating the effects of regularization from those of a proper objective. Label smoothing is a related approach to controlling confidence (Müller et al., 2019).

We evaluate a persona-contrastive InfoNCE term as an auxiliary extension. It differs from the contrasting-prompt preference-pair construction and subsequent RL of Yang et al. (2024). The standalone CADET method uses only the proper score; Appendix Table 5 reports the contrastive extension's lower accuracy and weaker distributional results.

Training-free interface. Frozen models receive page context, prior actions, and a menu of complete actions. The Public decision model API returns a choice and scores; LM baselines return a choice ID. Each prediction is evaluated against one recorded human action, with later states supplied by the human trace. Appendix T specifies the adapters and prompts.

4 Experiments and Results

We organize the evidence into task performance (Table 1), the effect of the training objective (Table 3), and robustness across backbones (Table 4). Search measures complete-action fidelity; browse measures category choice; the objective studies measure typed-action probabilities. Full metric grids and secondary settings are in the appendix.

4.1 Search and browse performance

Table 1 groups results by interface and protocol: menu selection, token decoding, action composition, and category choice. Comparisons within each block retain the stated inputs, output space, and scorer; the appendix contains the full settings.

Search action fidelity. On 902 OPeRA menus with three prior actions, trained CADET_log selectors reach 3.03–9.39% full-action exact accuracy, compared with 7.87% for the Public decision model API. The menus contain 680 of the 902 gold actions, placing a separate constraint on selection. Table 1 keeps the history-5 conditions in their own block; Appendix M gives the trained-selector details.

Composing full actions. Adding a target selector to the typed head yields $16.85 \pm 0.29\%$ full-action accuracy. Routing predicted inputs to the frozen query LM raises it to $17.37 \pm 0.23\%$ over three element-head seeds. Target selection remains the main bottleneck: always predicting click with the same element head already gives 16.93%. These rows use the extracted-element protocol, separate from the menu comparisons. Appendix K gives target coverage; Appendix W gives routing, payload, uncertainty, and timing diagnostics.

Frozen search models. With five prior actions, the Public decision model API scores 9.31% full-action exact on OPeRA, versus 10.75% for Luna low. On ShopCART, it leads the model rows at 11.92%, below always terminating at 28.02%; every model exceeds that baseline in type macro-F1. These menu conditions omit personas. Appendix T gives candidate coverage and adapter-specific inputs.

Token decoding provides a separate output protocol: Gemma-3-12B-it reaches 16.08% full exact zero-shot and 27.61% after LoRA SFT on OPeRA, generating targets and input text. Appendix S reports all four runs.

Browse category prediction. The frozen browse conditions share page context, three prior actions, and the category inventory. The Public decision model API scores 56.6% on OPeRA, tied with Luna medium, while Luna low leads ShopCART at 24.5%. Appendix X gives the history-20 comparison and derived-label diagnostics. Separately fitted ShopCART heads improve under proper scoring, reaching 12.24% with Brier versus 6.98% with entropy regularization. Their cohort alignment and taxonomy inputs differ from the frozen comparison (Appendix N).

4.2 Effect of the training objective

Table 3 isolates the training objective on frozen representations from different pretrained language models for OPeRA and Gemma-3-1B for ShopCART. It compares exact-match reward and entropy/KL regularization with proper log and Brier scores. The synthetic study supplies a known

Table 1: OPeRA search results (%): full-action exact accuracy (Full) and type macro-F1 (F1). Shared menus, history 3, $N = 902$, unless noted. Three-seed means except single-run token decoding.

Method / condition	Full	F1
Training-free baselines		
Always terminate	4.43	2.83
Public decision model API	7.87	32.15
CLM-v0.1-8B	0.33	30.48
Haiku direct[¶]	3.10	9.24
Luna direct[¶]	3.99	32.59
Trained CADET (log) selectors		
Gemma-3-1B	5.03	33.23
Gemma-3-4B	9.39	32.40
Gemma-3-12B	7.54	37.84
Llama-3.2-1B	4.03	34.77
Llama-3.2-3B	3.03	35.50
ModernBERT-base	9.35	31.01
ModernBERT-large	7.76	32.74
OPeRA frozen menus, history 5		
Public decision model API	9.31	32.20
Luna direct	10.75	32.59
LLM token decoding		
Gemma-3-4B-it, zero-shot	8.20	39.95
Gemma-3-12B-it, zero-shot	16.08	36.78
Gemma-3-4B-it, LoRA SFT	24.61	46.89
Gemma-3-12B-it, LoRA SFT	27.61	47.42
Trained action composition[†]		
Always click + element	16.93	31.04
CADET (log) + element	16.85	
+ query LM (hybrid)	17.37	
LoRA leaf arms + element	16.74–16.93	30.99–35.50

[†]Trains leaf/element heads; the query LM is frozen. LoRA range: ModernBERT-large leaf objectives. [¶]OPeRA Haiku direct and Luna direct use frozen menus with history 5. Details: Appendices M, S, W.

Table 2: ShopCART browse category accuracy (%), history 3; $N = 425$, 218 categories.

Method / condition	Category acc.
Training-free models	
Public decision model API	23.5
Haiku direct	21.4
Haiku thinking	20.7
Luna low	24.5
Luna medium	22.8
Fitted browse heads[‡]	
MiniLM similarity	7.06
ModernBERT-base	
exact-match	1.88
+ entropy	6.98
+ KL	5.57
CADET (log)	11.92
CADET (Brier)	12.24

[‡]No history in training, history 3 at test; row-ID alignment with the frozen cohort is unverified. MiniLM uses taxonomy example queries, which the frozen models exclude. Labels are derived from recorded queries. Details: Appendices N, X.

Bold: parameters trained on the study data; blank: not evaluated.

conditional for testing probability recovery; the real-data studies measure leaf accuracy, probability quality, and rare-action recall. Appendices I, O, and P give the constructions and scorer definitions.

Probability recovery and gradient scaling. The synthetic study supplies a known behavioral conditional. Longer exact-match training improves accuracy while worsening its KL to truth; regularization further improves accuracy, whereas CADET better recovers the distribution. A second construction holds Bayes accuracy fixed while changing the tails: CADET's mean total-variation distance to truth is 0.3294, versus 0.6575 for bare exact-match (Appendices I and D). Rare-action trajectories separately track the optimization effect: as exact-match training lowers termination probability, its gradient signal also shrinks (Appendix F).

OPeRA real sessions. CADET (log) matches entropy-regularized exact-match at 47.52% leaf accuracy while reducing ECE from 0.2254 to 0.0845 (Table 3). Both proper scores also recover termination actions that the regularized exact-match arms miss. Appendix Table 5 gives secondary metrics, and Appendix G checks calibration estimation and temperature scaling.

Personalization. Persona effects depend on the available information: the survey slightly improves accuracy on the official OPeRA cohort but lowers it for unseen users. Synthetic removal, mismatch, and noise tests distinguish missing information from misleading information (Appendices E, H, and J).

Selective prediction. At a shared confidence threshold, CADET makes fewer predictions with a lower error rate than entropy-regularized exact-match. The operating points move closer after threshold fitting on held-out calibration data (Appendix V).

ShopCART real sessions. ShopCART tests a second action schema with preceding-page observations, three prior actions, and 1,149 decisions (NEU-HAI, 2025). Proper scores give the best probability metrics among the trained frozen-feature arms, but the training prior remains stronger

Table 3: Effect of the training objective. Accuracy and rare-action recall are percentages; ECE and NLL are lower-is-better. OPeRA (902 decisions, five seeds) ; ShopCART (1,149 decisions, three seeds) use uses frozen features from pretrained open-weight LMs. Rare means termination on OPeRA and `click.refinements` on ShopCART. Both panels use frozen encoders and 15-bin ECE; the ShopCART LoRA result is in Appendix C. Appendices B and C give full metrics.

OPeRA-filtered

Objective	Leaf acc. ↑	ECE ↓	NLL ↓	Rare rec. ↑
Frozen encoder; $N = 902$				
Train prior	20.29	0.0005	2.511	0.0
Exact-match	31.29	0.6567	14.20	0.0
+ entropy	47.52	0.2254	2.062	0.0
+ KL	45.45	0.2876	2.365	0.0
CADET (log)	47.52	0.0845	1.670	9.5
CADET (Brier)	45.96	0.0718	1.725	10.5

ShopCART

Objective	Leaf acc. ↑	ECE ↓	NLL ↓	Rare rec. ↑
Frozen encoder; $N = 1{,}149$				
Train prior	36.99	0.0107	1.430	0.0
Exact-match	37.66	0.6110	15.96	0.0
+ entropy	38.73	0.3692	2.320	0.0
+ KL	37.57	0.4055	2.580	0.0
CADET (log)	38.21	0.1740	1.675	0.0
CADET (Brier)	38.21	0.1626	1.561	0.0

and rare-action recall stays zero. Representation quality is therefore a further bottleneck. A separate LoRA run raises leaf accuracy to $46.71 \pm 0.91\%$, 9.7 points above the majority rate. It uses 107,093 training rows, compared with 23,154 views in the frozen comparison; Appendix C gives the full metrics and seed variation.

4.3 Robustness across backbones

Table 4 tests the objective across seven Gemma, Llama, and ModernBERT backbones with frozen features, and four with LoRA. Each comparison uses 902 OPeRA decisions, three prior actions, no persona, and three seeds. Appendix L records the two scorers, training settings, full metrics, and seed variation.

With frozen features, proper scores give the two lowest ECE and NLL values on every backbone, and a proper rule has the lowest Brier score on all seven. Accuracy leaders vary: proper scores lead on three backbones and entropy regularization on four, with a maximum proper-score deficit of 2.03 points. Bare exact-match collapses to the majority leaf on both ModernBERT sizes. These results extend the accuracy–probability quality distinction beyond the original encoder.

Table 4: Backbone robustness on OPeRA-filtered (902 decisions; three seeds). Panels distinguish frozen features (benchmark scorer) and LoRA (repository scorer). Both use three prior actions and ten-bin ECE. Columns compare bare exact-match, entropy/KL regularization, and CADET log/Brier objectives. Bold identifies the best objective within a backbone and metric; † marks constant predictions. Full metrics and seed variation are in Appendix L.

	Leaf accuracy (%) ↑					ECE ↓				
Backbone	exact	+ent.	+KL	log	Brier	exact	+ent.	+KL	log	Brier
Frozen encoder										
Gemma-3-1B	35.25	39.80	39.69	**41.09**	40.47	0.5388	0.3437	0.3749	0.1859	**0.1830**
Gemma-3-4B	38.29	**42.17**	42.02	42.02	41.35	0.5355	0.3746	0.3890	0.2834	**0.2505**
Gemma-3-12B	38.80	38.95	39.43	**40.35**	39.58	0.5375	0.4302	0.4347	0.3330	**0.2894**
Llama-3.2-1B	39.17	**40.17**	38.47	38.14	37.80	0.5052	0.3635	0.3922	0.3112	**0.2652**
Llama-3.2-3B	38.29	40.47	40.84	**41.13**	40.76	0.5049	0.3800	0.3778	0.3141	**0.2652**
ModernBERT-base	20.29†	**42.39**	41.32	41.28	41.06	0.7971†	0.0747	0.1426	0.0524	**0.0393**
ModernBERT-large	20.29†	**41.54**	41.09	40.17	41.39	0.7971†	0.0918	0.1523	0.0425	**0.0386**
LoRA adaptation										
Gemma-3-1B	37.77	46.86	44.68	**47.01**	45.93	0.5527	0.1673	0.2197	0.0637	**0.0479**
Llama-3.2-1B	39.32	52.37	49.70	**53.36**	52.59	0.5396	0.1897	0.2418	**0.0494**	0.0518
ModernBERT-base	20.29†	40.32	40.87	42.46	**42.90**	0.7970†	0.1644	0.2212	0.0432	**0.0344**
ModernBERT-large	22.88	43.05	41.46	**43.31**	42.39	0.7648	0.1956	0.2303	0.0514	**0.0465**

With LoRA, proper scores achieve the highest mean leaf accuracy and lowest ECE on all four backbones, with an accuracy advantage of 0.15–2.03 points over the best exact-match variant. They retain termination probability mass, although top-1 termination recall remains zero. Together, the

frozen and adapted comparisons show that the probability-quality finding persists across encoder families.

5 LIMITATIONS

Scope. The objective studies isolate single-step scoring effects with frozen features and limited LoRA adaptation; ShopCART LoRA covers only CADET_log. Full-action fidelity also depends on target and text coverage. Sequential RL adds rollout and credit-assignment questions, while the current query generator is frozen.

Personalization. Evidence comes from OPeRA and the simulator; ShopCART has no persona survey. The official OPeRA split shares users between training and test, with only 15 test users; unseen-user results use the regrouped split. Several user-clustered intervals remain wide. Survey-noise findings apply to the simulated measurement process.

6 CONCLUSION

CADET combines direct typed decisions with on-demand text generation for user simulation. Proper-score training improves behavioral probability quality even when top-1 accuracy is unchanged, and this result persists across frozen and LoRA-adapted backbones. Search and category-based browse exercise the decision interface at different output granularities; target selection and query generation connect it to complete actions. The findings support matching computation to the interaction while evaluating both action fidelity and the distribution behind each choice.

REPRODUCIBILITY STATEMENT

Appendix A gives the proofs, Appendix O the hyperparameter selection procedure, Appendix P the metric provenance, and Appendix T the frozen-menu protocols. The retained training-free input packets, settings, raw responses, and predictions support offline metric recomputation. For the Track 1 tables, the current bundle lacks the trained-head code, split manifests, selected configurations, and per-example predictions needed to reproduce them independently. The backbone, LoRA, ShopCART, and ShopCART browse results come from per-seed scorer outputs in the experiment repository; the LoRA reruns, the ShopCART runs, and the browse runs also save per-row probabilities.

AI USE STATEMENT

Large language models supported research ideation and execution, literature retrieval and primary-source verification, drafting and revision, simulator implementation, and theoretical derivation. Algebraic identities were checked numerically; training-free results were checked against retained prediction records. The authors take responsibility for the analysis and content.

REFERENCES

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning (ICML)*, 2019. URL https://proceedings.mlr.press/v97/ahmed19a.html.

Diogo Almeida. Introducing system one models & Jev. https://typesafe.ai/blog/introducing-system-one-models-and-jev, 2026. TypeSafe AI blog, September 15, 2026.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://www.nowpublishers.com/article/Details/MAL-101.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Sanket Badhe, Priyanka Tiwari, and Deep Shah. The silent vote: Improving zero-shot LLM reliability by aggregating semantic neighborhoods. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 511–517. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.gem-main.48. URL https://aclanthology.org/2026.gem-main.48/.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005. URL https://jmlr.org/papers/v6/banerjee05b.html.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/bf8065446507b0d3842838564ac4f1f3-Abstract-Conference.html. arXiv:2503.02623.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Guiding LLMs the right way: Fast, non-invasive constrained generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3658–3673, 2024. URL https://proceedings.mlr.press/v235/beurer-kellner24a.html.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, 2005. URL https://sites.stat.washington.edu/wxs/Learning-papers/paper-proper-scoring.pdf.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 59456–59493, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/615675cc6e94ddb1a783904fb178b5f6-Paper-Conference.pdf.

A. Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982. doi: 10.1080/01621459.1982.10477856.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023. URL https://arxiv.org/abs/2306.06070.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020. doi: 10.18653/v1/2020.emnlp-main.21. URL https://aclanthology.org/2020.emnlp-main.21/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ziyi Xu, Yilong Zhao, Ruihang Lai, and Tianqi Chen. XGrammar: Flexible and efficient structured generation engine for large language models. In *Proceedings of Machine Learning and Systems (MLSys)*, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/hash/5c20ca4b0b20b0bd2f1d839dc605e70f-Abstract-Conference.html. arXiv:2411.15100.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3fc2c60b5782f641f76bcefc39fb2392-Abstract.html.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10932–10952, 2023. doi: 10.18653/v1/2023.emnlp-main.674. URL https://aclanthology.org/2023.emnlp-main.674/. arXiv:2305.13971.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14(1):107–114, 1952. doi: 10.1111/j.2517-6161.1952.tb00104.x.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html. arXiv:1706.04599; also the source for temperature scaling.

JevBench. JevBench: Independent evidence for AI model adoption. https://jevbench.xyz, 2026. Independent benchmark aggregator; primary evaluation cited separately. Accessed September 25, 2026.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Theodore Lee. SemIf (formerly OpenJev): Semantic ifs from open models. https://github.com/TheoLeeCJ/SemIf-OpenJev/tree/23cf1f39fc9534fe81437200959b6dfc7106e45a, 2026. MIT; not affiliated with TypeSafe AI.

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. WebLINX: Real-world website navigation with multi-turn dialogue. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 33007–33056, 2024. URL https://proceedings.mlr.press/v235/lu24e.html.

Chao Ma. Jev vs. the right baselines: A pre-registered independent evaluation. https://github.com/ickma2311/jev-baselines-eval/tree/ca0719a7d89a4b565d074aeac60681538a3c0e9f, 2026. Independent evaluation with analysis and erratum, September 18, 2026.

Matt Mastracci. [Core] structured generation mode for DiffusionGemma model (Jev-like). vLLM pull request #57250, https://github.com/vllm-project/vllm/pull/57250, 2026. Merged September 22, 2026; companion implementation: https://github.com/mmastrac/djev-spark/tree/08b708e51bb8d9f4eba0e85b61dab0be7093d20e.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016. URL https://proceedings.mlr.press/v48/mniha16.html.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

NEU-HAI. ShopCART dataset. Hugging Face dataset, https://huggingface.co/datasets/NEU-HAI/ShopCART, 2025. URL https://huggingface.co/datasets/NEU-HAI/ShopCART/tree/7b1ed6dcc2952234719bf0f3a2536aed20e7b055. Training-free snapshot: revision 7b1ed6d; accessed September 19, 2026.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–41, 2019. URL https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2201.03544.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations (ICLR) Workshop Track*, 2017. URL https://openreview.net/forum?id=HyhbYrGYe.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=1ikK0kHjvj.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022. URL https://proceedings.mlr.press/v151/roelofs22a.html.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7370–7392, 2024. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. doi: 10.1080/01621459.1971.10482346.

Chenze Shao, Fandong Meng, Yijin Liu, and Jie Zhou. Language generation with strictly proper scoring rules. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44474–44488. PMLR, 2024a. URL https://proceedings.mlr.press/v235/shao24c.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024b. URL https://arxiv.org/abs/2402.03300. origin of GRPO.

Ben Shneiderman. Direct manipulation: A step beyond programming languages. *Computer*, 16(8): 57–69, 1983. doi: 10.1109/MC.1983.1654471.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

The vLLM Team and Google DeepMind Team. DiffusionGemma: The first diffusion LLM (dLLM) natively supported in vLLM. https://vllm.ai/blog/2026-06-10-diffusion-gemma, 2026. Project blog, June 10, 2026; 26B discrete diffusion language model.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

TianyuCodings. NanoJev: A nano replica of Jev. https://github.com/TianyuCodings/NanoJev/tree/76fdfc9ecdca45a9bcef17991a07d3041a87685a, 2026. MIT; Qwen3-0.6B backbone; weights at https://huggingface.co/C-Tianyu/NanoJev.

Tim Tomov, Dominik Fuchsgruber, Rajeev Verma, and Stephan Günnemann. Task-aware calibration: Provably optimal decoding in LLMs, 2026. URL https://arxiv.org/abs/2605.10202.

Benjamin Turtel, Danny Franklin, Kris Skotheim, Luke Hewitt, and Philipp Schoenegger. Outcome-based reinforcement learning to predict the future. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=bbhdeL8EUX. Published November 2025; arXiv:2505.17989v4.

Benjamin Turtel, Paul Wilczewski, Kris Skotheim, Ville A. Satopää, and Philip E. Tetlock. How proper scoring rules shape LLM forecasting, 2026. URL https://arxiv.org/abs/2608.28482v2. Version 2.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. Customer-R1: Personalized simulation of human behaviors via RL-based LLM agent in online shopping. In *NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models (MTI-LLM)*, 2025. URL https://openreview.net/pdf?id=1zNmEA6UqC. arXiv:2510.07230v2.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton, and Dakuo Wang. OPeRA: A dataset of observation, persona, rationale, and action for evaluating LLMs on human online shopping behavior simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43942–43960. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2033. URL https://aclanthology.org/2026.acl-long.2033/.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *AAAI Conference on Artificial Intelligence*, 2019. doi: 10.1609/aaai.v33i01.33011658.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models, 2023. URL https://arxiv.org/abs/2307.09702. the Outlines paper.

Ronald J. Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991. doi: 10.1080/09540099108946587.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.13063.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.343. URL https://aclanthology.org/2024.emnlp-main.343/.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. RLCD: Reinforcement learning from contrastive distillation for language model alignment. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2307.12950. arXiv:2307.12950.

Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Simon Zhan, Vianne Gao, Ruochen Jiao, Junze Liu, Kun Qian, Yuxin Tang, Ran Xue, Houyu Zhang, Qingjun Cui, Yufan Guo, and Dakuo Wang. Shop-r1: Rewarding llms to simulate human behavior in online shopping via reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 8406–8423, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0e705ac30e573d1526f81a0fd071a151-Paper-Conference.pdf.

ZhengxuYu. LitJev: A Jev decision layer for off-the-shelf LLMs. https://github.com/zhengxuyu/litjev/tree/e7fb109a7466da9709028eb9c4e9f16eaeb4e2a3, 2026. Apache-2.0; independent reproduction, not affiliated with TypeSafe AI.

A PROOFS

Throughout, s is fixed, $P^*(\cdot \mid s)$ is the data-generating conditional, and $\Delta(\mathcal{A})$ is the simplex over the finite declared action-label support $\mathcal{A}$.

A.1 PROOF OF PROPOSITION 1

The exact-match objective is $R(\pi) = \mathbb{E}_{a\sim P^*(\cdot|s)}[\pi(a \mid s)] = \sum_a P^*(a \mid s)\,\pi(a \mid s)$, which is *linear* in $\pi(\cdot \mid s)$. A linear functional on a compact convex polytope attains its maximum at an extreme point, and the extreme points of $\Delta(\mathcal{A})$ are the point masses δ_a. Hence

$$\arg\max_{\pi\in\Delta(\mathcal{A})} R(\pi) \;=\; \mathrm{conv}\Big\{\delta_a \;:\; a \in \arg\max_{a'} P^*(a' \mid s)\Big\}.$$

With a unique mode, the maximizing distribution is its point mass. More generally, the maximizing set depends on P^* only through its modal set: conditionals with the same modal set induce the same optima. A strictly proper score instead has a unique optimum at the full conditional. For the log score $\mathbb{E}_{P^*}[\log \pi]$, Gibbs' inequality gives $\mathbb{E}_{P^*}[\log \pi] \le \mathbb{E}_{P^*}[\log P^*]$, with equality iff $\pi = P^*$. The negative Brier loss is strictly concave in π with its unique maximizer at P^* (Gneiting & Raftery, 2007).

A.2 PROOF OF COROLLARY 1

Accuracy under deterministic readout is $\mathbb{E}_s\big[P^*(\hat{a}(s) \mid s)\big]$, where $\hat{a}(s)$ is selected from $\arg\max_a \pi(a \mid s)$ using a fixed tie rule. By Proposition 1, an exact-match optimum places all mass on the true modal set, so its top-1 prediction is a Bayes decision. A proper score is maximized at $\pi = P^*(\cdot \mid s)$ and also selects a mode. Both attain the same Bayes accuracy. Finite-sample accuracy differences therefore concern estimation, representation, or optimization; they do not follow from the population identification result alone. Section 4.2 examines these effects experimentally.

A.3 REGULARIZED POPULATION OPTIMA

For $p \in \Delta^{K-1}$ and $\tau > 0$, maximize $\langle p, q\rangle + \tau H(q)$ over the simplex. The stationarity equation is $p_i - \tau(\log q_i + 1) + \lambda = 0$, giving $q_i^H \propto \exp(p_i/\tau)$. Strict concavity gives uniqueness. For $\beta > 0$, $r_i > 0$, and $\sum_i r_i = 1$, maximizing $\langle p, q\rangle - \beta\mathrm{KL}(q\|r)$ similarly gives $q_i^{\mathrm{KL}} \propto r_i \exp(p_i/\beta)$. Both optima are positive. Summing their log equations recovers

$$p_i = \tau \log q_i^H + \frac{1 - \tau\sum_j \log q_j^H}{K}, \qquad p_i = \beta \log(q_i^{\mathrm{KL}}/r_i) + \frac{1 - \beta\sum_j \log(q_j^{\mathrm{KL}}/r_j)}{K}.$$

Thus these Bayes-act maps identify p through a known inverse even though the reports are generally not truthful probabilities. At finite data or incomplete optimization, an inverted estimate need not lie on the simplex. The reported experiments evaluate the native probabilities without inversion.

A.4 PROOF OF PROPOSITION 2

Write $\pi(a \mid s) = \mathrm{softmax}(z)_a$. For the exact-match reward on the observed action a^*,

$$\frac{\partial \pi(a^* \mid s)}{\partial z_j} = \pi(a^* \mid s)\big(\mathbb{1}[j = a^*] - \pi(j \mid s)\big),$$

In the limit $\pi(a^* \mid s) \to 0$, this gives $\|\nabla_z \pi(a^* \mid s)\| = \Theta\big(\pi(a^* \mid s)\big)$. For the negative log score, $\partial(-\log \pi(a^* \mid s))/\partial z_j = \pi(j \mid s) - \mathbb{1}[j = a^*]$. Its observed-action coordinate gives $\|e_{a^*} - \pi\|_2 \ge 1 - \pi(a^* \mid s)$, so the norm is $\Theta(1)$ in this limit. Section 4.2 measures the gradient ratio and its trajectory during training.

Proposition 3 (Ablate–shuffle inequality)**.** *Fix s and draw Θ, Θ' independently from $P(\Theta \mid S = s)$. Write $P_\theta = P^*(\cdot \mid s, \theta)$ and $\bar{P} = \mathbb{E}_{\Theta|s} P_\Theta$. Assume finite expected KL divergences on a common finite support. Then*

$$\mathbb{E}_{\Theta,\Theta'|s}\mathrm{KL}(P_\Theta\|P_{\Theta'}) = \mathbb{E}_{\Theta|s}\mathrm{KL}(P_\Theta\|\bar{P}) + \mathbb{E}_{\Theta'|s}\mathrm{KL}(\bar{P}\|P_{\Theta'}) \;\ge\; \mathbb{E}_{\Theta|s}\mathrm{KL}(P_\Theta\|\bar{P}). \tag{2}$$

Equality holds exactly when $P_\Theta = \bar{P}$ almost surely under $P(\Theta \mid S = s)$.

A.5 Proof of Proposition 3

At a fixed state s, let Θ, Θ' be conditionally independent draws from $P(\Theta \mid S = s)$, with $P_\theta = P^*(\cdot \mid s, \theta)$. For any fixed Q on the common support, expansion gives

$$\mathbb{E}_{\Theta|s}\mathrm{KL}(P_\Theta \| Q) = \mathbb{E}_{\Theta|s}\mathrm{KL}(P_\Theta \| \bar{P}) + \mathrm{KL}(\bar{P} \| Q).$$

This is the KL instance of the Bregman-centroid decomposition (Banerjee et al., 2005). Substitute $Q = P_{\Theta'}$ and average over the independent draw. The final term is nonnegative and vanishes exactly when $P_{\Theta'} = \bar{P}$ almost surely. The ablation cost at each s is the conditional information at that state; its expectation over S is $I(A; \Theta \mid S)$. Learned persona-blind and shuffled models need not equal these oracle conditionals, so the inequality alone imposes no ordering on their finite-sample risks.

A.6 Numerical verification

For $P^* = (0.45, 0.25, 0.15, 0.10, 0.045, 0.005)$, optimization to convergence recovers the truth to TV 0.0002 with the log rule and 0.0000 with Brier. Exact-match converges to the modal point mass (TV 0.55), retaining the same top-1 decision as the proper optima. A $2000\times$ weight on the rarest class shifts the point mass to that class (TV 0.995), illustrating how reweighting changes the selected vertex.

B DETAILED OPERA OBJECTIVE RESULTS

Table 5: OPeRA-filtered objective comparison (902-row cohort, mean of five seeds). Acc is leaf accuracy; TermMass is predicted termination mass over its base rate (1 ideal) and TermRec its recall; SmallP is the fraction of rows with near-zero probability on the observed action. Arrows give the preferred direction and bold the best (ties included); daggered metrics are provisional diagnostics defined in Appendix P.

Model	Acc ↑	TypeMacF1 ↑	TermMass → 1	TermRec ↑	ECE† ↓	AUROC† ↑	NLL† ↓	SmallP ↓
Majority class	0.2029	0.3104	0.0000	0.0000	0.7971	0.5000	22.03	0.797
Train prior	0.2029	0.3104	0.8392	0.0000	**0.0005**	0.5000	2.511	**0.000**
improper reward								
exact-match	0.3129	0.3224	0.0004	0.0000	0.6567	0.5372	14.20	0.514
+ 2000× wts	0.0922	0.0994	0.0001	0.0000	0.9009	0.4978	19.97	0.878
+ temperature	0.3129	0.3224	1.329	0.0000	0.3046	0.5110	3.118	**0.000**
+ entropy	**0.4752**	0.4428	0.2419	0.0000	0.2254	0.7478	2.062	**0.000**
+ KL to ref.	0.4545	0.4415	0.1381	0.0000	0.2876	0.7576	2.365	0.002
proper scoring rule (CADET)								
CADET (log)	**0.4752**	0.4993	1.473	0.0950	0.0845	0.7662	1.670	**0.000**
CADET (Brier)	0.4596	**0.5065**	**1.043**	**0.1050**	0.0718	0.7663	1.725	**0.000**
+ temperature	**0.4752**	0.4993	1.826	0.0950	0.0606	**0.7682**	1.622	**0.000**
persona blind	0.4725	0.4606	**1.043**	0.0400	0.0653	0.7558	**1.615**	**0.000**
contrastive extensions								
InfoNCE alone	0.2328	0.1988	0.0157	0.0000	0.7024	0.3823	8.640	0.243
CADET + InfoNCE	0.4195	0.4809	1.436	0.0950	0.1337	0.7281	1.939	**0.000**

C SHOPCART FULL RESULTS

Table 6 uses the preceding-page observation mapping of the standardized benchmark, three prior actions, and frozen `google/gemma-3-1b-it` features ([mean; last] pooling over the first 4,000 characters at 512 tokens), trained for 40 epochs on 23,154 views from 6,849 sessions. The rare class is `click.refinements` (base rate 0.0296). With twenty prior actions, each arm's leaf accuracy moves by at most 1.0 point and its ECE by at most 0.008. Scoring the CADET_log participant through the benchmark scorer in a separate run gives $38.38 \pm 0.54\%$ with three prior actions and $38.85 \pm 0.13\%$ with twenty.

Table 6: ShopCART objective comparison with preceding-page observations (1,149 test actions, 357 sessions; three prior actions; frozen Gemma-3-1B features; mean ± SD over three seeds). Acc is leaf accuracy (%); CI is the session-clustered 95% interval for seed 0. ECE uses 15 equal-width bins. RareMass is predicted `click.refinements` mass divided by its base rate; rare-class recall is zero for every arm. TermRec is termination recall. The train prior's argmax is the majority leaf. Bold marks the best trained arm in each column.

	Acc ↑	Seed-0 CI	TypeMacF1 ↑	ECE ↓	NLL ↓	Brier ↓	AUROC ↑	RareMass → 1	TermRec ↑
Train prior	36.99	—	18.00	0.0107	1.430	0.7229	0.5000	0.6028	0.0000
exact-match	37.66±0.73	[35.3, 41.2]	28.41	0.6110	15.96	1.2292	0.4916	0.0000	0.2298
+ weights	26.81±0.00	[24.2, 29.3]	17.28	0.7318	12.97	1.4637	0.5359	0.0000	0.0000
+ entropy	**38.73**±0.31	[35.4, 41.6]	**40.03**	0.3692	2.320	0.9214	0.5477	0.2657	**0.2950**
+ KL to ref.	37.57±0.20	[34.5, 40.9]	37.51	0.4055	2.580	0.9515	**0.5550**	0.0104	0.2681
CADET (log)	38.21±0.26	[35.3, 41.6]	39.84	0.1740	1.675	0.7707	0.5382	**0.6114**	0.2888
CADET (Brier)	38.21±0.74	[35.6, 42.1]	39.85	**0.1626**	**1.561**	**0.7605**	0.5444	0.5418	0.2930

LoRA adaptation. The ShopCART LoRA condition adapts `google/gemma-3-1b-it` (rank 16, $\alpha = 32$, learning rate 10^{-4}, batch 8, three epochs, 512 input tokens; 3,641,615 trainable parameters) with CADET_log on 107,093 training rows, excluding the post-action page and the rationale. Seeds 0, 1, and 2 give leaf accuracy 45.78%, 46.74%, and 47.61% (session-clustered intervals $[42.8, 48.8]$, $[43.7, 49.7]$, and $[44.6, 50.6]$; mean $46.71 \pm 0.91\%$), type accuracy 46.91%, 48.22%, and 48.91%, type macro-F1 46.73%, 48.11%, and 48.80%, ECE (ten bins) 0.064, 0.047, and 0.054, NLL 1.161, 1.149, and 1.156, AUROC 0.618, 0.612, and 0.605, and termination recall 0.432, 0.450, and 0.453.

ShopCART LoRA; 107,093 training rows, three-seed means				
Objective	Leaf acc. (%)	ECE	NLL	Rare recall (%)
CADET (log)	46.71	0.0550	1.155	0.0

Superseded post-action protocol. The earlier ShopCART table paired each action with the page recorded after it ($n = 1506$ test views, hashed page features) and reported 0.9159–0.9190 leaf accuracy for the unweighted arms. The post-action page contains the submitted query for 417 of 425 search actions, against 121 of 425 for the preceding page, and it is blank for every termination and add-to-cart action. Rerunning that protocol with frozen Gemma-3-1B features gives 0.917–0.928 for the unweighted arms and reproduces the earlier train-prior row (0.5193 accuracy, 0.2279 type macro-F1). We therefore report only the preceding-page comparison.

D Probability recovery under equal Bayes accuracy

Construction. Let P^*_A be the simulator conditional. At each state, world B retains one selected maximizing action and its probability $p_{\max}$, distributing the remaining mass uniformly over the m other legal actions. For $m > 0$, their probability is $(1 - p_{\max})/m \le p_{\max}$; a singleton support is unchanged. The maximum probability and hence Bayes accuracy are preserved. With tied modes, the modal set can change, so equality of the complete exact-match optimum sets requires a separate check. The recorded top-1 agreement is 1.000; the simulator arrays needed for that stronger check are unavailable. Both worlds share features, legal supports, and personas; only their sampled labels differ. Each arm is trained on each world's labels for 600 epochs over three seeds.

Readout. For a model trained in world w, let TV_{own} be its mean TV to P^*_w and $\mathrm{TV}_{\text{other}}$ its mean TV to the opposite world's truth. The recovery margin is $\mathrm{TV}_{\text{other}} - \mathrm{TV}_{\text{own}}$, averaged over the two training directions. A positive margin means the estimate is closer to its own data-generating conditional; zero means equal average distance.

Table 7: Two conditionals with equal Bayes accuracy whose tails differ by TV 0.2121, three seeds. Margin is the direct probability-recovery margin. The proper rules achieve larger margins than the bare and regularized improper rewards. The reported accuracy-gap column scores both policies against world A's labels; its sign convention and seed aggregation are unavailable. The worlds have identical Bayes accuracy by construction.

	TV(own)	TV(other)	Margin	Reported acc gap
exact-match	0.6575	0.6580	0.0005	0.0334
+ weights	0.8297	0.8351	0.0053	0.0343
+ entropy	0.4887	0.5012	0.0126	0.0300
+ KL to ref.	0.5017	0.5120	0.0103	0.0307
CADET (log)	0.3294	0.3941	0.0648	0.0421
CADET (Brier)	0.3169	0.3779	0.0609	0.0293

Interpretation. The proper rules achieve larger direct-recovery margins. Entropy and KL regularization reduce distance to truth relative to bare exact-match, but their margins remain 0.0126 and 0.0103. In this test, regularization spreads probability mass more successfully than it recovers the differences between the two conditionals. The persona-disjoint split also leaves substantial uncertainty, limiting absolute recovery for every arm. All margins evaluate native reports with data and representation fixed.

E Ablate versus shuffle

Persona ablation and mismatch. We evaluate the oracle inequality and analogous learned-model interventions (Appendix E). In the simulator, the analytic inequality holds at every tested persona strength. Mismatch costs 3.11 times as much as ablation on average, and both costs vanish at zero persona strength. With trained models, the ratio is 3.34.

The OPeRA intervention compares retraining without persona with substituting another user's survey at test time. Appendix E gives the intervention design and NLL diagnostics.

Table 8 compares $\mathbb{E}_\theta \,\mathrm{KL}(P^*(\cdot \mid s, \theta) \,\|\, \bar{P}(\cdot \mid s))$ and $\mathbb{E}_{\theta,\theta'} \,\mathrm{KL}(P^*(\cdot \mid s, \theta) \,\|\, P^*(\cdot \mid s, \theta'))$ analytically. With trained simulator models, the shuffle-to-ablation cost ratio is 3.34 (1.5664 versus 0.4683 nats), compared with the analytic mean ratio of 3.11. The original OPeRA three-seed summary reports NLL of 1.7225 with the correct persona, 1.6798 with persona withheld, and 1.8831 with another user's survey: changes of -0.0427 and $+0.1606$ nats. These values remain provisional diagnostics because the scorer's numerical handling has not been recovered (Appendix P).

Table 8: Analytic evaluation of Proposition 3 in the simulator. The parameter `persona_scale` controls the effect of the latent trait. Ablation costs less than shuffling at every positive tested strength; both costs are zero at `persona_scale`$= 0$.

`persona_scale`	Ablated KL	Shuffled KL	Ratio
0.00	0.00000	0.00000	— (equality)
0.05	0.01159	0.03431	2.96
0.10	0.04278	0.12875	3.01
0.20	0.14487	0.39528	2.73
0.40	0.41517	1.28064	3.09
0.80	0.89102	3.36024	3.77

Interventions. The blind arm is retrained without persona, approximating the marginal predictor in the proposition. The shuffled arm uses a persona-conditioned model with another user's persona substituted at test time, measuring deployment under mismatched information. These interventions distinguish learning without a persona from supplying an incorrect one.

The separate accuracy sweep in Section 4.2 tests the estimation cost of conditioning when the measured persona signal is weak.

F GRADIENT TRAJECTORIES

Rare-action recovery. The exact-match arm predicts `terminate` zero times out of 40, matching the failure reported by Wang et al. (2025). CADET achieves 9.5% recall with the log rule and 10.5% with Brier. The $2000\times$ reweighting further reduces accuracy in this experiment. Entropy regularization increases normalized rare-class mass from 0.0004 to 0.2419, but both regularized improper arms retain zero `terminate` recall. Proper rules improve both rare-class mass and recall on this split.

Small probabilities. The bare reward assigns less than 10^{-6} probability to the observed action on 51.4% of rows. SmallP measures this concentration separately from NLL.

Gradient trajectories on real data. We track probabilities and gradients on OPeRA rows with the rare `terminate` label over three seeds and 160 epochs. Under exact-match reward, $\pi(a^*)$ falls from 0.0804 to 3×10^{-5}, and the gradient norm falls from 0.0853 to 4×10^{-5}, a factor of about $2{,}100$. At the final policy, the log-score gradient on the same rows is 1.3796, roughly four orders of magnitude larger. The separately trained log-score arm retains $\pi(a^*) = 0.2149$. The trajectories show how declining rare-action probability further weakens the exact-match update.

Both gradients are closed-form at the current policy: $\partial\pi(a^*)/\partial z = \pi(a^*)(e_{a^*} - \pi)$ has norm $\pi(a^*)\|e_{a^*} - \pi\|$, and $\partial \log \pi(a^*)/\partial z = e_{a^*} - \pi$ has norm $\|e_{a^*} - \pi\|$. We evaluate both gradients at every checkpoint for both training arms. The log-score column for the exact-match run therefore measures the gradient available under the log objective at the policy the exact-match model actually reached.

Table 9: Gradient trajectories on OPeRA `terminate` rows (base rate 0.0362), three seeds. Exact-match probability and gradient norm shrink together. The log-score gradient remains large when evaluated at the same final policy.

	$\pi(a^*)$ on rare rows		$\|g\|$ exact-match		$\|g\|$ log
	epoch 1	epoch 160	epoch 1	epoch 160	epoch 160
exact-match	0.0804	0.00003	0.0853	0.00004	1.3796
CADET (log)	0.0759	0.2149	0.0702	0.1088	0.9903

G ECE ESTIMATOR ROBUSTNESS

Post-hoc calibration. Temperature scaling fitted on held-out calibration users (Guo et al., 2017) reduces the improper model's ECE from 0.6567 to 0.3046 and its Track 0 KL to the latent conditional from 6.10 to 0.4673. The latter is lower than either regularized improper arm. Temperature scaling preserves the ranking: accuracy remains 0.3129 and `terminate` recall remains zero.

Calibration. The ECE diagnostic uses 15 equal-width bins. Binned estimates (Naeini et al., 2015) have measurement limitations (Nixon et al., 2019). The train-prior predictor achieves ECE 0.0005 while ignoring the state, illustrating why calibration alone is insufficient. LitJev (ZhengxuYu, 2026) also distinguishes typed model-output scores from calibrated probabilities in its documentation.

Estimator choice. Sparse bins can bias plug-in ECE estimates upward (Kumar et al., 2019; Nixon et al., 2019; Roelofs et al., 2022). Each result block specifies its binning rule; NLL and Brier scores provide complementary measures of probability quality.

H TRACK 0 SURVEY-NOISE SWEEP

Persona measurement quality. On the reported OPeRA test cohort, persona conditioning changes accuracy by $+0.0027$ (0.4752 versus 0.4725). OPeRA's prompt-only ablations also show no full-action accuracy gain from personas, while some type metrics improve (Wang et al., 2026, Table 5).

The simulator holds latent persona strength fixed and varies survey noise. The personalization effect ranges from $+7.4$ to -6.6 accuracy points (Appendix H). This sign reversal shows that a small or negative predictive effect can coexist with a useful latent trait. The information estimates quantify this simulated measurement process. Persona sensitivity is 0.0308 even at zero persona signal, motivating a null control for persona-swap tests.

Table 10: Survey quality at fixed latent persona signal ($\texttt{persona_scale}= 0.40$ and $I(A;\Theta \mid S) = 0.4513$ nats). Information columns are estimates; the value 1.008 exceeds the population retention bound. Paired models share the proper scoring rule; one receives the survey and one is trained and evaluated without it. Increasing survey noise changes personalization from beneficial to harmful at the same underlying signal.

Survey noise	Acc (aware)	Acc (blind)	ΔAcc	$\widehat{I}(A;\Theta\|S)$	$\widehat{I}(A;\mathrm{Sv}\|S)$	est. retained
0.05	0.4466	0.3724	+0.0742	0.4513	0.4549	1.008
0.15	0.4342	0.3724	+0.0618	0.4513	0.4181	0.926
0.35	0.4005	0.3724	+0.0281	0.4513	0.3272	0.725
0.75	0.3384	0.3724	−0.0340	0.4513	0.2396	0.531
1.50	0.3069	0.3724	−0.0655	0.4513	0.1240	0.275

The simulator samples latent traits θ and renders a noisy Likert survey. Models observe the survey, so the reachable oracle is $P(a \mid s, \mathrm{survey})$. We estimate it by self-normalized importance sampling over the trait posterior. Retained information is $I(A;\mathrm{Sv} \mid S)/I(A;\Theta \mid S)$, separating the quality of the survey from the strength of the latent persona effect. The displayed information values are estimates. The ratio 1.008 exceeds the population bound $I(A;\mathrm{Sv} \mid S) \le I(A;\Theta \mid S)$ under $A \perp\!\!\!\perp \mathrm{Sv} \mid (S,\Theta)$. Estimator uncertainty has not been established, so these estimates do not locate a retention threshold.

I DIAGNOSTIC DETAILS

Track 0 uses a simulator with known ground truth and leaf marginals fitted to OPeRA's observed shares. Models observe a quantized, noisy survey of a latent trait θ. The reachable target integrates over the survey posterior: $P(a \mid s, \mathrm{survey}) = \mathbb{E}_{\theta\sim p(\theta\mid\mathrm{survey})} P^*(a \mid s, \theta)$. The latent-trait oracle attains 0.4663 accuracy, compared with 0.4298 for the survey-conditioned oracle. KL to the latent-trait conditional has an irreducible floor of 0.0761; we report this floor alongside the reachable accuracy ceiling.

The diagnostic uses three probes in the simulator, where the conditional distribution is known.

Training duration. We train exact-match over $\{60, 200, 600, 1500\}$ epochs. Accuracy reaches 0.3577, below the reachable ceiling of 0.4298, while KL to the latent conditional increases to 14.59. Additional training concentrates the distribution without closing the accuracy gap.

Regularization. At 600 epochs, we compare bare exact-match, exact-match with entropy, exact-match with a KL anchor to the training marginal, and the two proper rules. The concave regularizers improve both accuracy and KL relative to bare exact-match. Proper scores achieve the lowest KL. This comparison separates the benefit of regularization from the distribution recovery obtained with a proper objective.

Table 11: Track 0 diagnostic, 3 seeds, known simulator ground truth. KL is measured against the latent-trait conditional; the survey-conditioned oracle has floor 0.0761. TermMass is predicted rare-class mass divided by its base rate (1.0 ideal). Bayes agree is top-1 agreement with the reachable Bayes decision. Longer exact-match training increases concentration, while regularization improves its accuracy. Proper rules give the lowest KL in the fixed-epoch comparison.

	Acc	KL to latent	TermMass	Bayes agree
Probe 1: exact-match reward, trained longer				
60 epochs	0.3392	6.097	0.0002	0.5525
200 epochs	0.3528	8.221	0.0000	0.5718
600 epochs	0.3549	12.481	0.0000	0.6183
1500 epochs	0.3577	14.594	0.0000	0.6369
Probe 2: making the objective strictly concave (600 epochs)				
exact-match	0.3549	12.481	0.0000	0.6183
+ entropy bonus	0.3806	1.098	1.060	0.6332
+ KL to reference	0.3870	1.638	0.779	0.6480
CADET (log)	0.3632	0.336	1.042	0.6115
CADET (Brier)	0.3667	0.332	1.341	0.6313
reachable ceiling	*0.4298*	*0.0761*	*1.000*	*1.000*

Gradient scaling. We bucket observations by $\pi(a^* \mid s)$ and measure the ratio of exact-match to log-score gradient norm. It increases from 0.0067 in the lowest bucket to 0.5812 in the highest, following the scaling in Appendix A.

J UNSEEN USERS: THE USER-DISJOINT SPLIT

Generalization to unseen users. The regrouped split in Table 12 assigns test users exclusively to test. The regularized improper and proper arms have similar observed accuracy; proper rules have positive rare-action recall and lower reported ECE. `Terminate` recall is 0.1100 and 0.0900 for the proper rules, compared with zero for both regularized improper arms; ECE is lower by a factor of two to three.

The persona-blind model has the highest observed accuracy on this split, 0.4095 versus CADET's 0.3819, and lower ECE. The slight benefit of survey conditioning on the reported test cohort therefore does not extend to this unseen-user evaluation. Track 0 provides a controlled setting for studying one possible source of this pattern: the information retained by the survey (Appendix H).

Table 12: User-disjoint OPeRA evaluation: $n = 1074$ decisions from 13 held-out users. Whole users are assigned exclusively to training, calibration, or test. Hyperparameters are reselected by cross-validation within this split. Intervals are user-clustered. Acc measures leaf accuracy; see Section 4.2.

Model	Acc (95% CI)	TypeMacF1	TermMass	TermRec	ECE†
Train prior	0.1555	0.3097	0.6186	0.0000	0.0200
exact-match	0.2743 [0.138, 0.375]	0.2994	0.0001	0.0000	0.7221
+ entropy	0.3858 [0.312, 0.451]	0.4439	0.3226	0.0000	0.3074
+ KL to ref.	0.3903 [0.312, 0.457]	0.4375	0.1379	0.0000	0.3258
CADET (log)	0.3819 [0.309, 0.432]	0.4687	1.599	0.1100	0.1357
CADET (Brier)	0.3812 [0.315, 0.416]	0.4738	1.335	0.0900	0.1231
persona blind	0.4095	0.4818	1.226	0.0467	0.0898

The regrouped split assigns whole users to training, calibration, or test. It contains 13 test users, compared with 15 in the reported test cohort, and evaluates prediction for people absent from training. Accuracy is lower for every arm; comparisons within this split isolate the objective effect under the same unseen-user setting.

K ELEMENT CANDIDATE COVERAGE

For context, published open-generation accuracy is 0.2151 for GPT-4.1 in OPeRA (Wang et al., 2026, Table 5) and 0.3958 for Customer-R1 SFT+RL (Wang et al., 2025, Table 4). These values describe their respective generation protocols.

The extractor covers 0.7761 of element-bearing gold actions; uncovered targets count as incorrect, and the model lacks an input-text generator. This extension measures the additional challenge of target selection. The controlled objective study measures distribution quality at the leaf level.

The leaf head uses one forward pass. Appendix R retains an original timing summary and specifies its output scopes.

Table 13: Element candidate coverage on the reported 902-row cohort. Coverage is computed over element-bearing click and input actions. The target-free `terminate` action is excluded from that denominator and is fully correct whenever its leaf prediction is correct.

	n	element-bearing	covered	coverage
`click`	786	786	593	0.7545
`input`	76	76	76	1.0000
`terminate`	40	0	0	—
total	902	862	669	0.7761

The element head scores page candidates using hashed character n-grams and the log objective. The extractor omits targets for about 23% of element-bearing actions, which bounds target-selection accuracy for this candidate set. The original summary labels 0.1689 ± 0.0036 as full-action accuracy. The retained element artifact (three seeds, 255 candidates) and the query-integration scorer give 0.1685 ± 0.0029 when every gold input counts as wrong, since the leaf and element heads emit no query text. Element accuracy is 0.2355 on element-bearing rows and 0.3039 on rows whose target is among the candidates. The artifact records 668 covered element-bearing test rows (0.7749), one fewer than Table 13; we have not traced this difference. Predicting a click on every row with the same element head gives 0.1693 ± 0.0042, so target selection, not the leaf objective, sets this full-action accuracy.

L Backbone objective comparison

For published context, Customer-R1 reports 61.20% fine-grained type accuracy after SFT and RL and 56.43% after SFT on its own split (Wang et al., 2025, Table 4). These are its native type–subtype metric and cohort, separate from our full-action comparisons.

Frozen features. Each backbone encodes the standardized leaf-classification packets (contract `standardized-shopping-action/v2`, cohort `continuation902`, three prior actions, no persona). The typed head trains on pooled features while the backbone stays fixed. The benchmark scorer (`benchmark/score.py`) computes leaf and type metrics, ECE with ten equal-width bins, Brier score, and NLL; missing and invalid predictions stay in the denominator. Table 14 lists all six arms. A separate backbone ablation with CADET_log reproduces these CADET_log cells exactly.

LoRA. The LoRA runs use the repository's OPeRA split (4,427 training and 902 test decisions) and the repository scorer, whose ECE uses ten left-closed bins. Per-user label counts on this test split equal those of `continuation902`, but the runs record no row IDs, so the frozen and LoRA results stay in separate tables. The ModernBERT-large row comes from a rerun that saved per-row probabilities; a rerun of Llama-3.2-1B with saved probabilities reproduces its original row exactly. Table 15 gives the full metrics.

M Complete-action selection on shared menus

The trained selectors score each row's complete-action options from the standardized `complete_action_selection` packets with three prior actions (CADET_log; 1,024 input tokens, 60 epochs, batch 8, three seeds); the backbone stays frozen. These packets list the same options, in the same order, as the Public decision model API menus of Section 4.1 on all 902 rows, with the same history length as the API's three-action condition on every row. Both observations contain the page URL, title, and extracted metadata. The API is scored by its evaluation harness and the other rows by the benchmark scorer; both count every row, and an uncovered gold action is a miss. Table 16 reports the comparison.

Table 14: OPeRA objective comparison on seven frozen backbones (902 decisions, three prior actions, no persona; mean of three seeds; benchmark scorer). Arms are exact-match, its entropy/KL-regularized variants, and CADET (log/Brier); ECE uses ten equal-width bins. Bold is the best arm per backbone and metric, † a collapsed constant prediction. Appendix L adds SDs, type metrics, NLL, and Brier.

Arm	Leaf acc. ↑	Type acc. ↑	Type F1 ↑	ECE ↓	Brier ↓	NLL ↓
Gemma-3-1B						
exact-match	35.25±2.83	81.45	38.68	0.5388±0.0385	1.1526	8.1664
+ weights†	8.43±0.00	8.43	5.18	0.9157±0.0000	1.8315	14.6870
+ entropy	39.80±0.29	83.30	41.94	0.3437±0.0041	0.9003	2.9627
+ KL to ref.	39.69±0.11	**84.00**	40.15	0.3749±0.0026	0.9315	3.2941
CADET (log)	**41.09**±0.52	83.19	42.45	0.1859±0.0094	0.7988	2.3316
CADET (Brier)	40.47±0.38	83.19	**42.58**	**0.1830**±0.0046	**0.7938**	**2.2301**
Gemma-3-4B						
exact-match	38.29±2.45	81.71	39.98	0.5355±0.0342	1.1240	8.4178
+ weights†	8.43±0.00	8.43	5.18	0.9157±0.0000	1.8315	14.9986
+ entropy	**42.17**±0.28	85.07	**45.71**	0.3746±0.0048	0.9155	3.5290
+ KL to ref.	42.02±0.59	**85.25**	43.87	0.3890±0.0037	0.9365	3.6311
CADET (log)	42.02±0.78	83.56	44.16	0.2834±0.0078	0.8491	2.7424
CADET (Brier)	41.35±0.48	84.66	45.66	**0.2505**±0.0084	**0.8167**	**2.4258**
Gemma-3-12B						
exact-match	38.80±2.25	83.70	41.37	0.5375±0.0350	1.1268	8.7354
+ weights	8.80±0.64	26.35	16.37	0.8907±0.0242	1.7947	16.5006
+ entropy	38.95±0.51	84.41	43.05	0.4302±0.0084	0.9918	3.9646
+ KL to ref.	39.43±0.13	84.74	43.09	0.4347±0.0027	0.9931	3.8281
CADET (log)	**40.35**±0.55	84.15	45.00	0.3330±0.0063	0.9066	3.1299
CADET (Brier)	39.58±0.22	**84.77**	**45.77**	**0.2894**±0.0041	**0.8707**	**2.6612**
Llama-3.2-1B						
exact-match	39.17±1.26	83.59	41.40	0.5052±0.0011	1.0829	7.4906
+ weights†	8.43±0.00	8.43	5.18	0.9157±0.0000	1.8314	14.3962
+ entropy	**40.17**±1.23	85.29	45.63	0.3635±0.0110	0.9195	3.4088
+ KL to ref.	38.47±1.28	**85.44**	45.04	0.3922±0.0165	0.9519	3.4931
CADET (log)	38.14±1.13	84.74	**45.71**	0.3112±0.0092	0.8794	2.7423
CADET (Brier)	37.80±0.51	85.07	45.63	**0.2652**±0.0054	**0.8457**	**2.5153**
Llama-3.2-3B						
exact-match	38.29±0.75	84.59	41.18	0.5049±0.0193	1.0799	7.4974
+ weights†	8.43±0.00	8.43	5.18	0.9157±0.0000	1.8315	14.4504
+ entropy	40.47±1.17	85.22	**45.97**	0.3800±0.0143	0.9257	3.4602
+ KL to ref.	40.84±1.31	**85.29**	44.48	0.3778±0.0107	0.9215	3.3805
CADET (log)	**41.13**±0.29	83.89	45.06	0.3141±0.0098	0.8783	2.8155
CADET (Brier)	40.76±0.52	84.48	45.52	**0.2652**±0.0045	**0.8371**	**2.5101**
ModernBERT-base						
exact-match†	20.29±0.00	87.14	31.04	0.7971±0.0000	1.5942	11.0608
+ weights†	8.43±0.00	8.43	5.18	0.9157±0.0000	1.8315	13.8753
+ entropy	**42.39**±0.17	87.18	32.18	0.0747±0.0052	0.7395	2.1376
+ KL to ref.	41.32±0.34	87.18	31.62	0.1426±0.0034	0.7605	2.3364
CADET (log)	41.28±0.34	86.95	**34.32**	0.0524±0.0066	**0.7143**	**1.8524**
CADET (Brier)	41.06±0.23	**87.25**	31.92	**0.0393**±0.0016	0.7163	1.9043
ModernBERT-large						
exact-match†	20.29±0.00	87.14	31.04	0.7971±0.0000	1.5942	11.2897
+ weights†	8.43±0.00	8.43	5.18	0.9157±0.0000	1.8315	13.8066
+ entropy	**41.54**±0.34	86.88	30.99	0.0918±0.0043	0.7501	2.1938
+ KL to ref.	41.09±0.32	**87.14**	31.04	0.1523±0.0037	0.7718	2.3799
CADET (log)	40.17±0.17	87.07	**35.50**	0.0425±0.0025	**0.7164**	**1.8581**
CADET (Brier)	41.39±0.17	87.10	31.04	**0.0386**±0.0029	0.7176	1.9048

Table 15: LoRA OPeRA objective comparison, all metrics (repository split, 902 test decisions; mean over three seeds; ± sample SD for leaf accuracy). Type F1 is type macro-F1 (%). ECE uses ten left-closed bins. TermMass is predicted termination mass divided by its base rate (1 is ideal); termination recall is zero for every arm. Bold marks the best non-degenerate arm per backbone and column; † marks a constant prediction across seeds. ModernBERT-large is a rerun with saved per-row probabilities.

Arm	Leaf acc. ↑	Type F1 ↑	ECE ↓	AUROC ↑	Brier ↓	NLL ↓	TermMass → 1
Gemma-3-1B							
exact-match	37.77±0.64	31.04	0.5527	**0.7922**	1.1456	6.5818	0.0005
+ entropy	46.86±0.82	38.64	0.1673	0.7585	0.7383	2.0390	0.4747
+ KL to ref.	44.68±1.17	38.79	0.2197	0.7884	0.7739	2.3648	0.2146
CADET (log)	**47.01**±2.00	**41.57**	0.0637	0.7693	**0.6780**	**1.7552**	**0.9317**
CADET (Brier)	45.93±1.40	41.08	**0.0479**	0.7807	0.6853	1.8062	0.7472
Llama-3.2-1B							
exact-match	39.32±0.63	31.04	0.5396	0.7826	1.1197	6.7343	0.0007
+ entropy	52.37±1.08	39.97	0.1897	0.7658	0.6958	1.9071	0.3824
+ KL to ref.	49.70±0.06	39.82	0.2418	0.7807	0.7446	2.2307	0.1730
CADET (log)	**53.36**±1.12	**46.74**	**0.0494**	0.7731	**0.6128**	**1.5296**	**0.8890**
CADET (Brier)	52.59±0.83	45.61	0.0518	**0.7850**	0.6168	1.5682	0.8085
ModernBERT-base							
exact-match†	20.29±0.00	31.04	0.7970	0.6411	1.5941	10.0690	0.0001
+ entropy	40.32±2.83	31.59	0.1644	0.7937	0.7760	2.1634	0.5819
+ KL to ref.	40.87±1.03	31.32	0.2212	**0.8006**	0.8003	2.4575	0.2380
CADET (log)	42.46±1.25	**33.89**	0.0432	0.7905	**0.7045**	**1.8456**	0.8333
CADET (Brier)	**42.90**±0.48	32.84	**0.0344**	0.7848	0.7074	1.8872	**0.8547**
ModernBERT-large							
exact-match	22.88±4.48	31.00	0.7648	0.7295	1.5339	8.9391	0.0026
+ entropy	43.05±0.55	31.77	0.1956	0.7792	0.7802	2.1632	0.5473
+ KL to ref.	41.46±0.29	**33.26**	0.2303	**0.8023**	0.8029	2.4823	0.2573
CADET (log)	**43.31**±1.51	32.13	0.0514	0.7858	**0.6981**	**1.8246**	**0.8503**
CADET (Brier)	42.39±1.12	31.54	**0.0465**	0.8020	0.7034	1.8612	0.7958

Table 16: OPeRA complete-action selection on shared menus (902 continuation decisions, three prior actions; %). Full is H/N, Covered is H/C; the menus expose the gold action on only 680/902 rows (C), which caps accuracy. Trained rows are CADET_log selectors on frozen features (mean ± SD, three seeds); bold is the best in each column.

System	Parameters	Full ↑	Covered ↑	Type acc. ↑	Type F1 ↑
Always terminate	—	4.43	5.88	4.43	2.83
Public decision model API	frozen	7.87	10.44	55.99	32.15
CLM-v0.1-8B	frozen	0.33	0.44	81.37	30.48
Gemma-3-1B	head trained	5.03±0.23	6.67	60.57	33.23
Gemma-3-4B	head trained	**9.39**±0.17	**12.45**	52.70	32.40
Gemma-3-12B	head trained	7.54±1.36	10.00	73.80	**37.84**
Llama-3.2-1B	head trained	4.03±0.23	5.34	65.04	34.77
Llama-3.2-3B	head trained	3.03±0.61	4.02	66.52	35.50
ModernBERT-base	head trained	9.35±0.26	12.40	**86.95**	31.01
ModernBERT-large	head trained	7.76±0.58	10.29	79.90	32.74

CLM-v0.1-8B is a second frozen decision model: trained projection heads over frozen LM embeddings return option probabilities, with no fine-tuning. It predicts the correct action type on 81.37% of rows but the correct target on only 3 of the 733 clicks it types correctly. On ShopCART packets with three prior actions, whose menus cover the same 767 of 1,149 gold actions as the training-free menus, it reaches 2.35% full exact ($27/1{,}149$), 35.07% type accuracy, and 17.49% type macro-F1, below always terminating (28.02%). On leaf classification it predicts one leaf for every row on both datasets (10.86% on OPeRA; the majority leaf, 36.99%, on ShopCART).

N BROWSE WITH A SEPARATELY FITTED CATEGORY HEAD

This condition trains a separate category head for browse. Each ShopCART search action is labelled with one of 218 L1.L2.L3 taxonomy categories by an embedding annotation of its query, so the gold labels are derived. The head (`TypedPolicy` over frozen ModernBERT-base features; 60 epochs, learning rate 3×10^{-4}) trains on 30,576 examples and is tested on the 425 search actions of the ShopCART test cohort with the benchmark scorer. Training examples omit prior actions, while test packets include three, so training and test inputs differ. Table 17 reports three seeds. The frozen models of Section 4.1 reach 20.7–24.5% on 425 ShopCART search decisions with the same categories and three prior actions; we have not matched their row identifiers to this cohort.

Table 17: ShopCART browse with a separately fitted category head (425 test search actions; 218 derived categories; chance 0.46%). Acc. is category accuracy (%, mean $\pm$ SD over three seeds), with correct counts per seed. ECE uses ten equal-width bins; uniform NLL is $\ln 218 = 5.38$. Entropy is the mean predicted-distribution entropy (nats; $\ln 218 = 5.38$ is uniform), and Leaves is the range of distinct predicted categories across seeds. The benchmark scorer reports no AUROC. Bold marks the best value in each column.

Condition	Acc. ↑	Correct / 425	ECE ↓	NLL ↓	Entropy	Leaves
Zero-shot MiniLM similarity	7.06	30	0.0315	4.94	4.98	21
exact-match	1.88±0.00	8, 8, 8	0.9812	22.52	0.00	1
+ entropy	6.98±0.14	30, 30, 29	0.0579	5.18	5.34	14–18
+ KL to ref.	5.57±0.59	24, 26, 21	**0.0137**	4.94	4.97	10
CADET (log)	11.92±0.83	47, 54, 51	0.0306	**4.54**	4.13	21–23
CADET (Brier)	**12.24**±0.41	51, 51, 54	0.0512	4.86	3.84	23–24

Bare exact-match collapses to a single category on all 425 rows (predicted entropy 0.00 nats, one distinct predicted leaf across every seed), while the proper rules spread mass over 21–24 of the 218 categories at the lowest NLL; this is the browse-task instance of the mode collapse in Proposition 1.

O HYPERPARAMETER SELECTION

Each arm selects epochs, learning rate, and hidden width through 3-fold user-grouped cross-validation within the training pool. The grid is epochs $\in \{40, 80, 160\}$, lr $\in \{10^{-3}, 3 \times 10^{-4}\}$, and hidden $\in \{256, 512\}$. Selection uses accuracy, leaving calibration as an evaluation metric. Grouped folds keep each user on one side of a training–validation partition. The current bundle lacks selected per-arm configurations, including the class-weight vector. The reported $2000\times$ arm therefore denotes a class-reweighting experiment, not a verified implementation of Customer-R1's full reward schedule.

Backbone replacement. Section 4.3 and Appendix L rerun the OPeRA objective comparison with the Hugging Face checkpoints Gemma-3 1B/4B/12B-it (`google`), Llama-3.2 1B/3B-Instruct (`meta-llama`), and ModernBERT base/large (`answerdotai`). Frozen-feature runs train only the typed head and use the benchmark packets and scorer. LoRA runs train the head and rank-16 adapters on `q_proj`, `k_proj`, `v_proj`, and `o_proj` for the decoders and on `Wqkv` and `Wo` for ModernBERT, where `Wo` also matches each MLP output projection. Track 1's table keeps its original encoder.

Trained cohort and model selection. Track 1 uses the reported 902-decision, 15-user OPeRA-filtered cohort (binding details: Appendix P). The reported organization has training (35 users), development (7), and calibration (7) pools. The official session split allows users to appear in both training and test; unseen-user evaluation uses the separate regrouped split in Appendix J. Hyperparameters are selected within the training pool, and temperatures are fitted on held-out calibration

data. InfoNCE negatives and the persona-blind mean vector use training users only. Confidence intervals resample users to account for dependence among actions within their sessions.

P REPORTED METRIC PROVENANCE

Leaf and rare-action metrics. Native leaf accuracy matches the action type and subtype. TermMass is mean predicted termination mass divided by the gold termination base rate (1.0 is ideal); TermRec is termination recall. SmallP is the fraction of rows with observed-action probability below 10^{-6}. The OPeRA table averages five seeds; intervals resample the 15 test users. Per-arm epochs, learning rate, and width use 3-fold user-grouped selection on training data. The typed support gives zero out-of-schema leaf mass for all compared arms.

Cohort binding. The released OPeRA-filtered test artifact has 992 rows in 90 sessions. Our frozen diagnostic removes the earliest $(\mathrm{timestamp}, \mathrm{action_id})$ row in each session to obtain 902 continuation decisions: 786 click, 76 input, and 40 terminate. The trained summary reports the same size and type counts, but its ID manifest is absent. These matching totals do not establish that the trained study selected the same rows. The hash in Appendix T binds only the frozen cohort.

Scorer definitions. Trained-head ECE, NLL, and AUROC values are provisional diagnostics retained from the original summaries. For ECE, 15 equal-width bins are specified, but the calibration event, confidence, bin boundaries, seed aggregation, and interval construction remain unverified. NLL's probability clipping or log-probability handling and AUROC's binary event, ranking score, reduction, and tie rule are also unavailable. The displayed values cannot determine these definitions. Recomputation requires the scorers and per-example predictions.

SmallP has the separately stated definition $n^{-1}\sum_i \mathbf{1}[q_i(y_i) < 10^{-6}]$. If NLL used a floor ε, its definition would be $-n^{-1}\sum_i \log\max(q_i(y_i), \varepsilon)$; the SmallP threshold does not specify ε. For example, the reported NLL 14.20 exceeds $-\log(10^{-6})$, so 10^{-6} cannot be its probability floor. The selective-prediction confidence score, seed aggregation, AURC integration and tie rules, and the complete 21-estimator ECE configuration list also require the original scorers. ShopCART's trained TypeMacF1 class inventory is unverified; its frozen metric is defined separately below.

Q DECISION INTERFACES AND PROBABILITY AUDITS

"System One" decision models. TypeSafe AI's Jev (Almeida, 2026) answers typed questions, including booleans, choices over up to 255 runtime-supplied options, and ordered scores. It returns option probabilities without generating text; its high-cardinality choice path uses separate scoring and choice stages according to the release. Its release calls the training procedure "Reinforcement Learning for Calibrated Decisions".[1] Open implementations include direct option-logit readout (Lee, 2026), trained decision heads (TianyuCodings, 2026), and a structured-read extension (Mastracci, 2026) to DiffusionGemma (The vLLM Team & Google DeepMind Team, 2026). These interfaces expose distributions over declared alternatives through different inference mechanisms, providing the setting for Section 3.3.

Auditing decision probabilities. TypeSafe's release leaves its reward function unspecified (Almeida, 2026). An independent evaluation indexed by JevBench (JevBench, 2026), with released analysis (Ma, 2026), reports error-ranking AUROC 0.734 for Jev and 0.816 for a nano-class baseline on CLINC150, with the reverse ordering on Banking77. The paired CLINC150 difference interval is $[-0.187, 0.008]$. These measurements assess discrimination. The container recipe accompanying Mastracci (2026) reports option-order and cross-question sensitivity: one option draws $p \in [0.03, 0.25]$ regardless of correctness, and a neighbouring question moves an unrelated slot from 0.8 to 0.04. We complement such interface checks with controlled objective comparisons and a separate evaluation of calibration and ranking quality (Section 4.2). The accompanying audit script uses returned probabilities and can therefore also evaluate closed APIs.

[1] This name differs from Reinforcement Learning from Contrastive Distillation in Yang et al. (2024), which constructs preference data.

R INFERENCE COST

Original timing summary. The retained report gives 186.0 ms end-to-end for the typed leaf head, using the same backbone and prompts at batch 1, compared with 608.9 ms for an autoregressive full-action JSON baseline ($3.27\times$). The head uses one forward pass; the baseline generates 39.4 tokens on average. The JSON baseline's out-of-schema rate is 0.2750, compared with zero for the typed leaf output. This measurement covers $n = 80$ prompts on one laptop. Three independent passes at $n \in \{40, 80, 80\}$ give ratios of 3.26–$3.27\times$; no interval was estimated. The output scopes differ: JSON includes the target element, while the timed head returns the (t, c) leaf. A matched full-action timing comparison requires the element stage as well. The timing harness and per-trial records are absent from the current bundle, so these figures remain unrecomputed summaries.

S OPEN-GENERATION COMPARISON

Gemma-3-it generates complete OPeRA actions, including element names and input text. We compare zero-shot inference with LoRA fine-tuning in Table 18. These runs use an open-generation output contract and scorer; Table 19 evaluates finite-menu selection.

Table 18: OPeRA-filtered open-generation results (%, 902 decisions; single runs). LoRA SFT updates model parameters. Bold marks the best value in each column within the zero-shot or SFT group. All metrics favor higher values. The final column retains the reported leaf mapping.

Model	Full exact ↑	Type acc. ↑	Type F1 ↑	Leaf proxy ↑
Zero-shot; no parameter updates				
Gemma-3-4B-it, zero-shot	8.20	**82.04**	**39.95**	15.19
Gemma-3-12B-it, zero-shot	**16.08**	64.86	36.78	**24.83**
Fine-tuned; LoRA SFT				
Gemma-3-4B-it, LoRA SFT	24.61	87.47	46.89	54.77
Gemma-3-12B-it, LoRA SFT	**27.61**	**88.58**	**47.42**	**57.32**

T FROZEN-MENU PROTOCOL DIAGNOSTICS

Table 19: Training-free finite-menu search diagnostics (%). Full exact matches type and complete payload; Type F1 is macro-F1. Leaf proxy maps selected actions to subtypes or families. Bold marks the best value within each dataset, including the unfitted always-terminate baseline.

Model	Full exact ↑	Type acc. ↑	Type F1 ↑	Leaf proxy ↑
OPeRA-filtered diagnostic: 902 decisions, 15 users				
Always terminate	4.43	4.43	2.83	4.43
Public decision model API	9.31	**55.10**	32.20	**28.94**
GPT-5.6 Luna (low effort)	**10.75**	52.33	**32.59**	23.61
ShopCART: 1,149 noninitial decisions				
Always terminate	**28.02**	28.02	14.59	28.02
Public decision model API	11.92	33.42	**21.79**	26.20
Haiku 4.5 (direct)	7.75	35.07	18.14	25.59
GPT-5.6 Luna (low effort)	6.18	**35.51**	21.12	**28.20**

Frozen model adapters. We evaluate frozen models through a shared action-selection interface. For each recorded decision, an adapter constructs a page representation, permitted behavioral context, and an ordered menu of complete actions. The model selects a menu ID, which resolves to an action type, target, and any supplied input text. The hosted *Public decision model API* takes a state object and a typed choice question mapping menu IDs to action labels; it returns an ID and candidate scores. We score its returned ID after verifying that it maximizes those scores, retaining the provider choice in ties. CLI baselines return a menu ID from a JSON prompt. The adapters use provider-specific serialization; their differences are documented in Appendix T. The prompts request a choice without a generated rationale. Each prediction is scored against the recorded human action, and subsequent observations come from the human trace. Model weights remain fixed.

Frozen metric definitions. Type macro-F1 averages over the fixed inventory {`click`, `input`, `terminate`} for OPeRA and {`click`, `type_and_submit`, `terminate`} for ShopCART:

$$\text{MacroF1} = \frac{1}{3} \sum_{t} \frac{2\,\text{TP}_t}{\text{Gold}_t + \text{Pred}_t},$$

with a zero term when its denominator is zero. Failed predictions remain in gold support and contribute no true positives. Full exact, type accuracy, and leaf-proxy accuracy use every cohort row in the denominator.

OPeRA input and selection. The shared observation contains a URL, page title, extracted page metadata, and up to five strictly earlier same-session actions. Named HTML leaves provide click candidates. Input candidates pair a field with a value already present in its source HTML or an earlier input to that field. The menu lists terminate, then inputs, then DOM-order clicks, with a cap of 255 choices and a 60 KB packet limit. The prefix builder stops before adding a choice that would exceed 59,000 serialized bytes, reserving space for packet metadata. On the 902 continuation rows, the choice cap shortens 599 menus and the byte bound shortens 591, with 473 rows affected by both: 717/902 menus are shortened. Before these bounds, all 786 click targets are present; the choice cap removes 119 and the byte bound removes 45 more, leaving 622. Input coverage stays at 18/76 through both stages. Across all 992 rows, the corresponding truncation counts are 654 and 653, with 526 overlapping rows. All reported menu coverage uses the final bounded menu.

The API instruction is `Choose exactly one listed candidate ID.` Luna receives `Use no tools. Return one choice_id.` as a developer instruction and the observation/options JSON as its user message. These directives constrain output without explicitly stating a human next-action objective. Provider system instructions and inference settings differ. In all 992 API records, the returned ID maximizes its scores; 23 records have tied maxima, resolved by the provider's choice.

Excluded Haiku runs. The filtered Haiku runner supplied two positional CLI strings: the instruction and then the observation/options JSON, without a `--system-prompt` flag. A local request-capture test of retained Claude Code 2.1.278 shows that only the first string becomes the user prompt. The JSON observation and labelled option mapping are dropped; the structured-output schema still exposes the allowed IDs. Cache dates and same-day session metadata associate this version with the run, but the benchmark did not save a CLI version per request. The current 2.1.282 binary reproduces the same parsing behavior.

We exclude the affected 992-row Haiku direct run from model comparisons. Its 902-row diagnostic scores were 3.10% full exact, 7.98% type accuracy, 9.24% type macro-F1, and 7.98% leaf proxy. The same helper was used for three filtered Haiku persona conditions with history budgets $k = 0, 3, 20$ (2,976 additional records), which also require rerunning. The separate full-OPeRA and ShopCART runners use different call builders. Luna's exporter supplied the JSON as the actual user message, so its filtered results are independent conditions, not prompt-matched reconstructions of these Haiku calls.

ShopCART contract and baseline. The instruction is `Predict the next shopping action.`, followed by output-format and webpage-as-data instructions. The source is the preceding row's parsed context; its UTF-8 prefix is bounded to keep the API request within 60,000 bytes. This truncates 491 of 1,149 contexts. Options comprise terminate, named elements, and field/text pairs from preceding values or earlier same-session inputs. The largest menu has 245 actions.

The API keeps insertion order; the Haiku and Luna JSON serializers sort choice IDs. These orders differ on 1,147 of 1,149 rows. Canonical hashes sort object keys and therefore do not certify wire-order equality. The menu contains 315/402 clicks, 130/425 inputs, and 322/322 terminations: 767/1,149 complete actions (66.75%). Always choosing terminate gives 28.02% full exact and type accuracy, 14.59% type macro-F1, and 28.02% leaf proxy. The model rows improve macro-F1 but fall below this constant baseline on complete-action accuracy. Haiku produced 1,147 valid results and two terminal errors; both errors remain incorrect in the 1,149-row denominator.

Observation timing. The upstream OPeRA card describes HTML as pre-action. An audit of the filtered test release finds that 9 of 107 input rows already contain the labelled text in the named input's value; all nine remain among the 76 continuation inputs. This is a timing ambiguity, since

a repeated input can also match an earlier value. The OPeRA results are therefore row-associated selection diagnostics. Excluding inputs changes the evaluated actions but does not validate capture timing for the remaining clicks.

Cohorts and coverage. The continuation mask removes the earliest $(\mathrm{timestamp}, \mathrm{action_id})$ row in each session and is applied to saved predictions without new calls. The 902 sorted example IDs have SHA-256:

`63f214f91802a2fae823f6497d243469a122f4de7140d567ac5737b6746036fd`

The corresponding gold counts are 786 clicks, 76 inputs, and 40 terminates. The menu contains 622, 18, and 40 of these complete actions, respectively: 680/902, or 75.39%. On all 992 rows, coverage is 737/992 (74.29%), including 18/107 inputs. Coverage counts the complete action payload.

Coverage and conditional fidelity. Table 20 recomputes exact hits from the saved final choices and exposed menus. Covered-row exact conditions on the complete gold action being expressible; it retains model failures within those rows. Click and input exact use all gold rows of that type and require the whole payload. These diagnostics use the same conditions as Table 19.

Table 20: Saved-menu fidelity decomposition (%). Full is H/N and Covered is H/C. Click/input exact use their gold-type denominators. OPeRA coverage is $680/902$, with click $622/786$ and input $18/76$; ShopCART coverage is $767/1149$, with click $315/402$ and input $130/425$. All gold termination actions are covered.

Model	Full	Covered	Click exact	Input exact
OPeRA-filtered				
Always terminate	4.43	5.88	0.00	0.00
Public decision model API	9.31	12.35	9.03	11.84
GPT-5.6 Luna	10.75	14.26	9.92	2.63
ShopCART				
Always terminate	28.02	41.98	0.00	0.00
Public decision model API	11.92	17.86	25.12	0.00
Haiku 4.5	7.75	11.60	21.64	0.24
GPT-5.6 Luna	6.18	9.26	16.42	0.71

U Additional related context

Improper objectives and reward misspecification. Proxy optimization can diverge from an intended objective (Amodei et al., 2016; Pan et al., 2022; Gao et al., 2023). Skalse et al. (2022) identify reward linearity as a strong restriction on unhackability. Proposition 1 examines linearity in a finite decision space: exact-match reward can attain Bayes accuracy while leaving the conditional distribution unidentified. The finite typed support lets us measure accuracy and distribution recovery separately.

Calibration and personalization in language models. Calibration depends on the model, task, and elicitation method (Desai & Durrett, 2020; Xiong et al., 2024; Minderer et al., 2021). Kadavath et al. (2022) find useful self-evaluation signals, and Tian et al. (2023) find verbalized confidence typically better calibrated than conditional token probabilities in the RLHF models they study. We examine how the training objective shapes the full action distribution. Behavioral simulation and personalization condition on descriptions, memories, or user profiles (Park et al., 2023; Argyle et al., 2023; Salemi et al., 2024). Section 4.2 measures the information supplied by persona conditioning and its effect on accuracy. Decision-focused learning also separates predictive accuracy from downstream decision quality (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022). Here, the central question is which distributional properties remain distinguishable when top-1 accuracy is similar.

V Selective prediction details

Selective prediction uses confidence to decide when a model should answer or defer (El-Yaniv & Wiener, 2010). Softmax thresholding is a standard implementation (Geifman & El-Yaniv, 2017); SelectiveNet learns an integrated reject option (Geifman & El-Yaniv, 2019). This setting connects probability quality to an operational decision.

We first use fixed thresholds to test confidence as reported by the model. This corresponds to an API caller accepting predictions above a chosen probability. Calibration-based methods address a complementary setting: Geifman & El-Yaniv (2017) select thresholds for target risk, while conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) controls marginal prediction-set coverage under exchangeability. Task-aware calibration also links task-level distributions to minimum-Bayes-risk decoding on latent model beliefs (Tomov et al., 2026). We evaluate both fixed thresholds and thresholds fitted on held-out calibration data.

Table 21 accepts predictions with confidence at least $\kappa = 1 - \epsilon$. At $\kappa = 0.80$, bare exact-match answers 94% of cases with 67% error, above the confidence-implied 20% rate. Entropy and KL regularization reduce this excess to 9.5 and 16.6 percentage points. Both proper rules have mean error below ϵ at all three tested thresholds. Fixed-threshold behavior depends on the probability scale, whereas AUROC depends on ranking.

CADET's lower error comes with lower coverage. At $\kappa = 0.80$, it answers about 20% of cases, compared with 40% for the entropy-regularized arm. Fitting thresholds on held-out calibration data brings the operating points closer: at a target error of 0.30, the entropy arm achieves 0.2297 error and

Table 21: Fixed-threshold selective prediction on OPeRA, 5 seeds. Models answer when confidence is at least $1 - \epsilon$, without threshold fitting. Excess is observed mean error minus the confidence-implied rate ϵ. Both proper rules have negative excess at the three thresholds, with lower coverage. These are descriptive means, not risk guarantees.

	$\epsilon = 0.20$ ($\kappa = 0.80$)			$\epsilon = 0.30$ ($\kappa = 0.70$)			$\epsilon = 0.40$ ($\kappa = 0.60$)		
	Cov	Err	Excess	Cov	Err	Excess	Cov	Err	Excess
exact-match	0.937	0.672	+0.472	0.961	0.678	+0.378	0.979	0.682	+0.282
+ entropy	0.404	0.295	+0.095	0.512	0.350	+0.050	0.631	0.409	+0.009
+ KL	0.500	0.366	+0.166	0.595	0.410	+0.110	0.688	0.444	+0.044
CADET (log)	0.196	0.133	−0.067	0.266	0.198	−0.102	0.372	0.263	−0.137
CADET (Brier)	0.194	0.104	−0.096	0.244	0.162	−0.138	0.313	0.234	−0.166

0.3033 coverage, compared with CADET's 0.2354 and 0.3202. Their AURC values are also similar (0.2993 and 0.3107). Proper scoring improves the use of reported probabilities at fixed thresholds; calibration data can recover much of the gap for regularized exact-match models.

W Evaluation contracts and query routing

The complete-action, leaf, and category studies use different output contracts. Appendix P identifies the trained-head scorers; Appendices T and X specify the frozen-model packets, observation timing, history, and task labels. The equations below define the implemented hybrid and its scoring boundary.

Coverage and uncertainty. For a finite-menu block, N counts all registered rows, C the rows containing the complete gold action, and H exact hits. Coverage is C/N, full exact is H/N, and covered-row exact is H/C. Invalid or missing predictions remain in N; a zero denominator is N/A. Paired comparisons use aligned IDs and resample users when identified, or sessions otherwise. Per-condition artifacts record the resampling and inference-cost settings.

Text-module training. The implemented simulator has three separately optimized parts and one frozen part. Let $x = (o, h, p)$, $z_\psi(x)$ the pooled encoder features, and $\ell^* = (t^*, c^*)$, e^*, q^* the recorded leaf, target, and query.

Leaf head (CADET). Parameters θ (two-layer GELU trunk, type and per-type subtype layers) and, in LoRA runs, adapters ψ:

$$\mathcal{L}_{\text{leaf}}(\theta, \psi) = -\log \pi_\theta(\ell^* \mid z_\psi(x)) \quad \text{or} \quad \textstyle\sum_\ell \big(\pi_\theta(\ell \mid z_\psi(x)) - \mathbb{1}[\ell = \ell^*]\big)^2. \tag{3}$$

Supervision is the recorded leaf only. AdamW (weight decay 10^{-4}), one-cycle schedule, gradient-norm clip 1.0. Frozen-feature runs update θ only (settings: Appendix O); LoRA runs update θ and rank-16 adapters (α=32, dropout 0.05) on the q, k, v, o projections (for ModernBERT, the fused `Wqkv` and every `Wo`, which includes the MLP output projection), with learning rate 10^{-4}, batch 8, 3 epochs, and 512 input tokens; embeddings and the remaining weights stay frozen.

Element selector. Parameters ϕ (two affine projections):

$$\pi_\phi(e_j \mid c, E) \propto \exp\big(\langle W_E E_j, W_c c\rangle / \sqrt{256}\big), \qquad \mathcal{L}_{\text{el}}(\phi) = -\log \pi_\phi(e^* \mid c, E), \tag{4}$$

where c is the frozen persona-free context block and E_j hashes character 4-grams of candidate j (256 dimensions, with position). Rows whose target is outside the first 255 candidates are excluded from training and scored wrong. Separate AdamW (learning rate 10^{-3}, batch 256, 40 epochs, clip 1.0, three seeds). $\mathcal{L}_{\text{el}}$ updates ϕ only, and $\mathcal{L}_{\text{leaf}}$ never updates ϕ.

Text module. A frozen instruction-tuned generator g, invoked by the routing gate $r(x) = \mathbb{1}\big[\text{type}(\arg\max_\ell \pi_\theta(\ell \mid x)) = \textsf{input}\big]$, emits $\hat{q} = \text{greedy}(g, \text{prompt}(x))$ with at most 24 new tokens. The prompt contains the persona, prior actions with earlier typed queries, the pre-action page summary and URL, and a fixed instruction; raw HTML and the element list are removed. It is not a supplied-query selector or retriever. There is no text loss ($\mathcal{L}_{\text{text}} = 0$) and no trainable text parameter. Because r is an argmax over stored predictions, $\partial\, \text{score}(\hat{q}, q^*)/\partial(\theta, \psi, \phi) = 0$: text correctness updates neither the encoder, the leaf head, nor the element selector. The decision model is not trained with policy gradients or any other text-dependent feedback.

Composition. The emitted action is $\hat{a} = \textsf{terminate}$ if $\hat{t} = \textsf{terminate}$, $(\textsf{click}, \hat{e})$ for a click, and $(\textsf{input}, \hat{e}, \hat{q})$ for an input, with $\hat{e} = \arg\max_j \pi_\phi(e_j)$. Routing is decided by the predicted type, so a false-negative route yields a non-input action and a false-positive route yields a wrong action; neither is repaired downstream.

Results. On OPeRA-filtered (902 rows, 76 gold inputs), routing on CADET_log predictions sends 75 rows (precision 0.453, recall 0.447). Against a no-text baseline that leaves every input uncovered, strict full-action exact rises from 16.85 ± 0.29 to 17.37 ± 0.23 (mean $\pm$ SD over element seeds; $+0.52$ points, 4–5 newly correct rows, user-clustered 95% interval lower bound 0.0 for every seed). Oracle-routed query exact is $13/76$, versus $12/76$ for copying the previous query. Generation takes 5.0 s for 76 inputs on one A100. These results measure the addition of generated text to an otherwise fixed leaf-and-element predictor. The no-text baseline counts every gold input as wrong.

Implementation. `src/heads.py` (`TypedPolicy`, `ElementScorer`); `src/objectives.py`; `src/train.py`; `src/run_finetune.py`; `src/run_element.py`; `src/run_query_mvp.py`; `src/evaluate_query_integration.py`; `src/run_shopcart_finetune.py`.

Let N_{in} count registered gold-input rows and H_{in} count predictions with the correct input type, field, and literal text after the same scorer normalization. Input exact is $100H_{\text{in}}/N_{\text{in}}$; wrong-type, invalid, and missing predictions are wrong. If T is the number of predicted-input triggers and V the number producing valid queries, trigger rate is T/N and valid-query rate is V/T (N/A when $T = 0$). Full exact and its paired change use the common full cohort N.

X Training-free browse protocol and diagnostics

Task and inputs. We evaluate category prediction at recorded search steps, rather than an interactive rollout through a catalog. OPeRA-filtered contributes all 76 input actions with prior session history in the audited 902-decision continuation cohort (11 users, 36 sessions). ShopCART contributes 425 noninitial `type_and_submit` actions from 233 sessions. Every query has a mapped category, and no eligible test row is dropped. OPeRA uses the same-row released `simplified_html` and URL; ShopCART uses the preceding action row's `parsed_context`. HTML/context text is capped at 48,000 UTF-8 bytes; the current OPeRA URL is retained. History URL/context fields are capped at 1,000 bytes alongside the earlier actions. All models receive identical serialized observations and histories within each condition, plus the same ordered category paths. Taxonomy example queries, gold labels, audit fields, and personas are excluded. All six development/test packet pairs match the registered hashes.

Models and execution. The Public decision model API uses a pinned model version; its requested and resolved identifiers are retained in configurations and prediction records, respectively. Haiku uses `claude-haiku-4-5-20251001` through the Claude Code subscription CLI, with thinking disabled or a 4,096-token thinking budget and output limits of 256 or 8,192 tokens, respectively. Luna uses `gpt-5.6-luna` through the Codex subscription CLI at low or medium effort. Luna's CLI does not expose a resolved model version. CLI tools and session persistence are disabled. These are different reasoning controls, not equal token budgets. All prompts request one category key; the Public decision model API returns a distribution, while the LMs return JSON. No model is fine-tuned. We initially use up to 32 concurrent requests per interface, increasing Luna to 64 after the other interfaces finish. We retry transient failures at most five times after the initial attempt, and retain the first valid answer. An initial Claude CLI stdin startup timeout was repaired by passing the identical prompt as an argument; returned attempts remain in the ledger. Interrupted in-flight calls may have unreturned usage, so returned-response cost is not total billed cost. Reported latency includes CLI and request overhead under concurrent execution. Subscription marginal cash costs are unavailable; they are not estimated from API list prices.

Selection and uncertainty. Each dataset has 150 fixed development decisions at history 3, selected by smallest ID hashes from eligible training rows (both OPeRA training shards and the first ShopCART training shard). The development-selected baseline B is selected on development accuracy among the Public decision model API and the two Haiku conditions, breaking ties by condition ID. This choice is frozen before model test scoring and reused at history 20. Luna is an additional comparison outside that registered selection pool. Accuracy counts all test rows, including failures

as misses. Confidence intervals use 2,000 paired cluster-bootstrap samples (seed 20260923), resampling OPeRA users or ShopCART sessions. ECE uses ten equal-width confidence bins; NLL and multiclass Brier score require a full normalized distribution.

Gold and query visibility. The fixed gold categories are automatically mapped from human query strings, not human category annotations. Some mappings are incorrect; 136 of the 425 ShopCART labels are other/unknown categories. The OPeRA same-row join does not establish that its HTML predates the target action. The gold query is visible in observation/history for 31/76 OPeRA and 173/425 ShopCART rows at history 3; at history 20 the counts are 32/76 and 175/425. The hidden-query diagnostic excludes these case-insensitive literal substring matches. This detects literal visibility, not all semantic leakage. Taxonomy examples are omitted because they contain 56 OPeRA and 24 ShopCART test queries. These task and annotation choices apply to every frozen arm.

Table 22: Browse diagnostics. The lower panel reports returned-probability quality; zero gold probability gives infinite raw NLL. Acc. includes its cluster-bootstrap 95% interval. Hidden is accuracy where the gold query string is absent from the supplied page/history. L1 is department accuracy; distinct counts predicted leaves; latency is median seconds. OB/CB denote OPeRA-filtered/ShopCART; API denotes the Public decision model API.

Block	k	Model	Acc. (%)	95% CI	Hidden (%)	L1 (%)	Leaves	p50 (s)
OB	3	API	56.6	[27.6, 76.8]	48.9	68.4	35	0.6
OB	3	Haiku direct	55.3	[27.9, 70.6]	51.1	67.1	38	9.3
OB	3	Haiku thinking	38.2	[25.4, 47.6]	42.2	68.4	45	17.7
OB	3	Luna low	55.3	[27.9, 72.4]	53.3	67.1	35	25.1
OB	3	Luna medium	56.6	[31.4, 71.6]	53.3	71.1	35	25.0
OB	20	API	59.2	[30.2, 80.2]	47.7	69.7	34	0.6
OB	20	Haiku direct	52.6	[25.0, 67.9]	50.0	64.5	39	9.6
OB	20	Haiku thinking	42.1	[22.8, 56.5]	45.5	68.4	47	17.2
OB	20	Luna low	56.6	[26.0, 74.1]	54.5	68.4	35	24.3
OB	20	Luna medium	56.6	[33.3, 69.8]	52.3	71.1	36	23.9
CB	3	API	23.5	[18.2, 29.2]	21.8	45.6	103	0.5
CB	3	Haiku direct	21.4	[16.4, 26.6]	19.0	44.9	104	10.2
CB	3	Haiku thinking	20.7	[15.9, 25.9]	19.4	43.1	102	17.7
CB	3	Luna low	24.5	[19.6, 29.8]	21.8	45.4	104	28.6
CB	3	Luna medium	22.8	[18.0, 27.9]	19.4	44.0	107	28.7
CB	20	API	24.2	[18.9, 29.9]	23.2	45.9	102	0.5
CB	20	Haiku direct	21.4	[16.4, 26.5]	18.4	43.5	98	10.0
CB	20	Haiku thinking	20.5	[15.7, 25.5]	18.4	44.5	97	17.9
CB	20	Luna low	24.0	[19.3, 29.2]	21.6	45.9	106	27.3
CB	20	Luna medium	24.5	[19.4, 29.9]	21.2	45.2	114	28.1

Returned probability diagnostics for the Public decision model API

Block	k	N	NLL	Brier	ECE	Entropy
OB	3	76	∞	0.629	0.247	0.582
OB	20	76	∞	0.614	0.244	0.535
CB	3	425	∞	1.156	0.477	0.899
CB	20	425	∞	1.151	0.466	0.909

Probability quality and paired comparisons. At history 3, the Public decision model API assigns zero returned probability to the annotated leaf on 19/76 OPeRA and 210/425 ShopCART decisions (19 and 208 at history 20). The smallest positive returned value is 0.01; the zeros characterize the returned scores, whose unrounded values are unavailable. The fixed training-frequency reference has Brier scores of 0.996 on OPeRA and 0.993 on ShopCART. The API improves OPeRA Brier but worsens ShopCART Brier despite higher top-one accuracy. At history 3, the API–Haiku direct accuracy difference is 1.3 percentage points on OPeRA (paired 95% CI: $[-12.5, 10.7]$) and 2.1 on ShopCART ($[-1.6, 5.8]$); both intervals include zero. These frozen-system comparisons establish the category-prediction baseline for subsequent trained-model evaluation.