

Are LLM Shopping Agents Behaviorally Distinguishable from Human Shoppers? A Session-Level Study

Arka Dutta^{◇*}, Virinchi Roy Surabhi[♣], Jiayue Tong[♣], Dihan Dai[♣],
Olcay Boz[♣], Amila Weerasinghe^{♣†}

[◇]Rochester Institute of Technology [♣]Amazon

ad2688@rit.edu {viriroy, tongjia, dihandai, olcayboz, weera}@amazon.com

Abstract

Large language models can now operate a web browser and carry out multi-step shopping tasks. Does such an agent leave a behavioral signature in a session that separates it from a person, even when it drives a real browser and presents an ordinary client fingerprint? We study the question at the session level. We first show that weak per-action evidence aggregates into a session-level distinction whose error decays exponentially in the number of actions. We then define six families of behavioral signals, each grounded in a documented difference between goal-directed automation and human browsing, and combine them in a classifier we call HABITS (Human-Agent Behavioral Invariants from Traffic at Session-level). Labeled agent sessions are scarce, so we describe a simulation methodology for driving language models through shopping tasks and instantiate it on public data with realism-controlled planner agents. We evaluate on five public benchmarks spanning e-commerce clickstreams, web-task demonstrations, and mouse dynamics, reimplementing prior methods on the identical corpus. Against agents that imitate human timing, timing-only classifiers fall to 0.35 recall at a 1% false-positive rate and to 0.01–0.02 when timing is fully matched, while HABITS reaches 0.96 on the clickstream corpus. Paired tests separate HABITS from every restricted-view baseline, and equivalence tests show that strong classifiers given the same features are interchangeable, which places the contribution in the behavioral view rather than the estimator. A check against real agent trajectories supports the structural assumption our simulator relies on and finds one realism parameter set too favorably.

1 Introduction

Autonomous agents built on large language models can plan and execute multi-step shopping tasks in

a web browser. Because they drive a real browser their client fingerprint looks ordinary, and a recent measurement study finds that behavioral features separate such agents while static fingerprints do not (Wang et al., 2026). This raises a question independent of any platform: *does an LLM-driven shopping session carry a behavioral signature that distinguishes it from a human session?*

We study the question at the level of a whole session rather than a single action. An agent is efficient by design, and that efficiency leaves regularities a person does not share: steadier timing, more directed navigation, interaction that is generated rather than observed. Any one action can be made to look human, but a session is many actions and small differences accumulate. Section 3 makes this precise: under mild assumptions, aggregating weak per-action evidence yields a distinction whose error decays exponentially in session length.

Data is the binding constraint. Contemporary shopping agents are new and labeled examples scarce, so a study trained on one agent framework risks learning that framework rather than the behavior. We address this with a simulation methodology (Section 4) that drives a range of current language models through realistic shopping tasks while reproducing human-like interaction timing, paired with aggregated, de-identified human sessions from public datasets.

Our contributions are: (i) a theoretical argument for session-level distinguishability; (ii) six families of behavioral signals grounded in prior work on human dynamics, clickstream analysis, and sequence modeling, combined in the HABITS classifier (Section 5); (iii) a realism-controlled simulation methodology; and (iv) a study on five public benchmarks that reimplements prior methods on the identical corpus and reports recall at a fixed low false-positive rate with paired significance tests (Section 6), including a check of our simulator’s assumptions against real agent trajectories (Sec-

* Work done during an internship at Amazon.

† Corresponding author.

tion 6.5).

2 Related Work

Distinguishing automated from human browsing. Separating automated from human web sessions is long studied, from navigation logs (Tan and Kumar, 2002) to clickstream models (Wang et al., 2013) and cross-account synchrony (Cao et al., 2014). Fusing server-side navigation with client-side interaction beats either alone (Iliou et al., 2021), behavior-only features separate automated traffic at scale (Kadel et al., 2024), and automation frameworks leave client-side inconsistencies even in a real browser (Jonker et al., 2019). This literature predates LLM agents, whose behavior is neither as regular as a script nor as varied as a human, which is the gap we study.

LLM agents and how they browse. A think-aloud study reports that web agents update their knowledge and resolve ambiguity differently from humans (Son et al., 2024), and agent benchmarks characterize how trajectories differ across model families (Zhou et al., 2024). These observations motivate several of our signal families: agents plan and execute more directly than people explore, which the navigation-predictability and goal-directedness families measure. Prior work asks whether agents can be told apart on an instrumented site, reporting balanced accuracy or per-agent separation. We ask how separability behaves at the session level as a function of session length, agent realism, and behavioral view, at a fixed low false-positive rate throughout. We also treat the agent as a data-generation instrument, driving many model families through a shared harness so agent realism becomes a controlled variable.

Human dynamics and one-class sequence modeling. Human inter-event times are heavy-tailed and bursty rather than regular (Barabási, 2005; Goh and Barabási, 2008), timing statistics alone can separate automated from human activity (Gianvecchio et al., 2008), and dwell times follow a Weibull law (Liu et al., 2010). Because agent examples are scarce, we adopt one-class modeling on human sessions only (Liu et al., 2008; Ruff et al., 2018; Erfani et al., 2020) and adapt masked and causal language-model objectives (Devlin et al., 2019; Vaswani et al., 2017) to behavioral sequences, as sequence models have been applied to session-level fraud detection (Wang et al., 2017; Hu et al., 2023).

Goal-directedness draws on maximum-entropy inverse reinforcement learning (Ziebart et al., 2008), inverse planning (Baker et al., 2009), and hierarchical options (Kipf et al., 2019).

Evaluation at low false-positive rates. When one class is rare, precision is governed by the false-positive rate (Axelsson, 2000), so evaluation should report recall at a fixed low false-positive rate under group-aware splits and realistic class ratios (Arp et al., 2022), guarding against optimistic estimates from biased splits (Pendlebury et al., 2019). We follow this methodology. Robustness work notes that any behavioral distinction can be imitated, and that imitation must produce a realizable session rather than an arbitrary feature vector (Pierazzi et al., 2020; Tramèr et al., 2020; Acien et al., 2022); our simulation methodology studies distinguishability against increasingly realistic imitation.

3 Problem Setup and Session-Level Distinguishability

A session x is generated by a source of type $c \in \{H, A\}$ (human or LLM agent) and consists of a sequence of interaction actions $a_1, \dots, a_n$. A classifier maps a session to a score and decides by thresholding. We report recall (the fraction of agent sessions above threshold) at a fixed false-positive rate FPR, the fraction of human sessions above threshold.

Definition 1 (Per-action score). A per-action score is a bounded statistic $s_t \in [0, 1]$ computed at step t , larger when the action is more agent-like. Its session mean is $\bar{S} = \frac{1}{n} \sum_t s_t$. Let $\mu_c = \mathbb{E}_c[s_t]$ and $\Delta = \mu_A - \mu_H$.

Proposition 1 (Session-level distinguishability). *Suppose the per-action scores are conditionally independent given the source type, take values in $[0, 1]$, and satisfy $\Delta > 0$. Deciding “agent” when $\bar{S} > (\mu_H + \mu_A)/2$ gives*

$$\max(\text{FPR}, \text{FNR}) \leq \exp\left(-\frac{1}{2}n\Delta^2\right). \quad (1)$$

Proof. $\text{FPR} = \mathbb{P}_H[\bar{S} - \mu_H > \Delta/2]$; since the s_t are independent and bounded, Hoeffding’s inequality (Hoeffding, 1963) gives $\text{FPR} \leq \exp(-n\Delta^2/2)$, and the same bound applies to FNR. $\square$

Even a small per-action difference becomes a reliable session-level distinction once the session is long enough, which is why we work at the session

level and combine weak signals rather than relying on any single action. Section 6.7 tests the predicted trend empirically.

Remark 1 (Correlation and fusion). Real actions are correlated, so the effective sample size is below n and (1) is optimistic as a literal rate; we use it for its shape. Combining families that capture different aspects of behavior contributes closer to independent evidence than several views of one aspect, which is the practical reason fusion helps.

At low prevalence, precision is dominated by the false-positive rate, so we select and report recall at a fixed low FPR rather than accuracy or AUC (Axelsson, 2000).

4 Data and Simulation Methodology

Human sessions are aggregated, de-identified samples from public datasets; we use no individual customer data and report only aggregate quantities. The harder problem is a diverse set of agent sessions, since shopping agents are new and no public corpus exists.

4.1 Simulating realistic agent sessions

We generate agent sessions by having contemporary language models carry out realistic shopping tasks in a browser, and we control the realism of the resulting behavior so the study is not trivially easy. The methodology is a loop with three ideas, sketched in Figure 5 (appendix).

First, we *decouple what from how*: the language model chooses the shopping actions, a separate layer governs interaction realism, and a uniform recorder captures each session. Diversity comes from sweeping public model families as backbones (Claude, GPT, DeepSeek, GLM, Qwen, MiniMax) through the same harness; we do not rank models. The experiments here use the loop’s distilled form, a goal-directed planner whose funnel behavior matches the regularities observed in the LLM trajectories, because a scripted planner is exactly reproducible from a seed while LLM output is not. Our numbers therefore quantify separation from this planner class, an upper bound we test against real agents in Section 6.5.

Second, we *add human-like realism* rather than leaving raw automation in place. A naive agent is easy to separate on uniform timing alone, so we reproduce the distributional shape of human interaction timing drawn from the aggregate human sample. Figure 1 shows the two regimes: machine

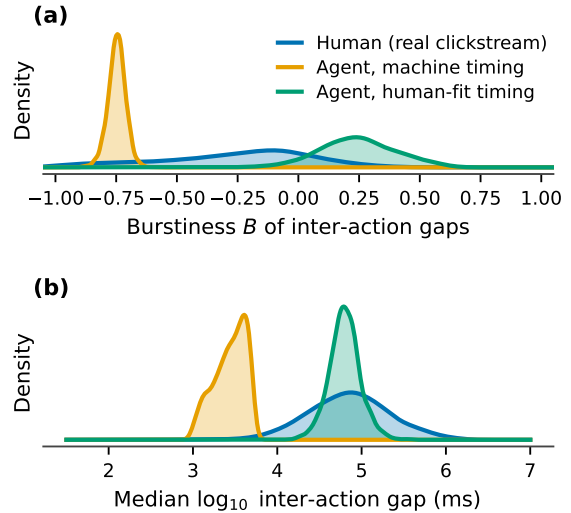


Figure 1: Kernel densities of (a) burstiness B and (b) median $\log_{10}$ inter-action gap on the public e-commerce clickstream. Machine-timed agents (orange) are separable at a glance; agents fit to the human timing prior (green) overlap the human curves (blue), pushing the recoverable signal to the funnel families.

timing sits far from the human curves, while fitting the human prior makes both timing views overlap and forces the classifier onto the non-timing families.

Third, the methodology is a *loop*: the sessions the current classifier finds hardest guide the next round of generation, so the corpus grows toward harder cases. Generation stops before any checkout or payment step.

The study corpus pairs these agent sessions with the human sample, the RetailRocket clickstream (Retail Rocket, 2016) in Setting A. We evaluate with five-fold group-aware cross-validation so no source is split across folds (Arp et al., 2022; Pendlebury et al., 2019), and report recall at a fixed 1% false-positive rate.

5 Behavioral Signal Families

We define six families of behavioral signals over a session (Figure 2), each motivated by a documented difference between goal-directed automation and human browsing. Feature-level definitions are in Appendix B.

Timing regularity. Human inter-action times are bursty and heavy-tailed, whereas automated activity tends toward regular pacing (Barabási, 2005; Goh and Barabási, 2008; Gianvecchio et al., 2008). The anchor feature is the burstiness coefficient

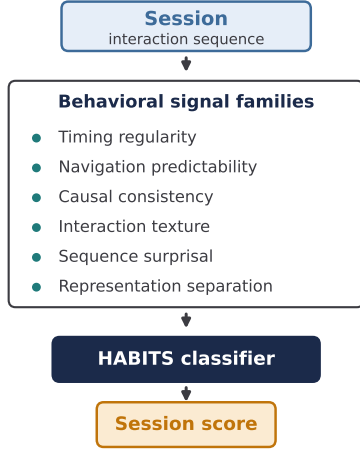


Figure 2: The study pipeline. A session is summarized by six families of behavioral signals, which the HABITS classifier combines into a single session score.

$B = (\sigma - \mu)/(\sigma + \mu) \in [-1, 1]$ over the gaps, near +1 for bursty human activity and near 0 for regular pacing, with dispersion, gap memory, and a think-time ratio around it. People pause before committing, and copying B alone does not reproduce that.

Navigation predictability. Human browsing mixes exploration with the task, whereas a goal-directed agent follows a more predictable path. We compute marginal and conditional action entropy and a Lempel-Ziv compression rate (Shannon, 1948; Ziv and Lempel, 1978; Cover and Thomas, 2006); a directed path has low conditional entropy and compresses well. A goal-directedness score, grounded in maximum-entropy inverse reinforcement learning and inverse planning (Ziebart et al., 2008; Baker et al., 2009; Kipf et al., 2019), adds funnel-progress monotonicity, path efficiency, and sub-goal chunking.

Causal consistency. Human high-value actions are usually preceded by the browsing that motivates them. For a high-value action a_t with prerequisites $\text{pa}(a_t)$, such as a prior view of the item, the core statistic is the break indicator $b_t = 1[\text{pa}(a_t) \not\subseteq \{a_1, \dots, a_{t-1}\}]$. An agent acting on an externally supplied target can commit to an item it never browsed, a break a human session rarely shows.

Interaction texture. Human interaction produces a dense, varied stream of low-level events; automation that does not wait on rendering or move a pointer produces a sparser, more uniform one (Jonker et al., 2019). The family summarizes event

rate and transition structure, plus pointer kinematics where a trace exists.

Sequence surprisal and representation separation. Because human sessions are abundant and agent sessions scarce, two families train on humans only (Liu et al., 2008; Ruff et al., 2018). Sequence surprisal scores a session under causal and masked sequence models of human behavior (Vaswani et al., 2017; Devlin et al., 2019), so an agent whose policy narrows the conditional distribution scores as surprising without any agent training data. Representation separation combines a one-class model of the human region of a learned session embedding with a supervised head over it. Neither is separately validated here, because the public corpora lack the dense-interaction and rich-embedding views they need.

Fusion. The HABITS classifier combines the families into one session score, selecting among standard classifier families (Geurts et al., 2006; Chen and Guestrin, 2016) by cross-validated recall at the target false-positive rate (Section 6.2). Sequence and representation models are fit within each training fold, so no evaluation session informs its own features. Feature-level definitions for every family are in Appendix B.

6 Experiments

We evaluate on five public benchmarks. Rather than quote numbers reported on private data, we reimplement each prior method and run it on the identical corpus, split, and protocol, so every number below is measured by us on public data. All results use five-fold group-aware cross-validation, recall at a fixed 1% FPR, and 95% bootstrap intervals over pooled out-of-fold scores.

6.1 Datasets and settings

Table 1 summarizes the three settings. Setting A is the primary one, testing multi-family fusion against agents that copy human timing; B and C isolate single families.

6.2 Main result: methods on a common corpus

Table 2 reports Setting A: every method on the identical corpus and protocol, with 95% bootstrap intervals. The reimplemented prior methods span a timing-only classifier, an all-feature random forest, and a behavior-only gradient-boosting classifier; ours is HABITS with an extra-trees back end.

Set.	Public source	Sessions	Signal families
A	E-commerce clickstream (Retail Rocket, 2016) + planner agents	5,000 / 2,000	Timing, navigation, causal, goal, structural (17)
B	Web-task demos (Deng et al., 2023) + templated agents	769 / 600	Action-stream (5)
C	Mouse-dynamics, legit vs. impostor (Fülöp et al., 2016)	476 / 405	Interaction texture (11)

Table 1: The three primary settings (sessions given as human/agent; feature counts in parentheses). Setting A uses a public e-commerce clickstream for the human class and a goal-directed planner for the agent class; the planner approximates the human inter-action timing distribution (burstiness $B \approx +0.3$) but keeps agent-like funnel behavior, the hard regime. Section 6.6 widens this to five benchmarks.

Method	R@1% [95% CI]	AUROC
Timing-only NB	0.349 [0.26, 0.42]	0.959
Timing-only LogReg	0.354 [0.30, 0.41]	0.955
All-feature RF	0.961 [0.95, 0.97]	0.998
Behavior GBDT	0.957 [0.94, 0.97]	0.998
HABITS (ours)	0.958 [0.95, 0.97]	0.996

Table 2: Setting A, all methods reimplemented on the identical corpus and five-fold group-aware split. Recall at 1% FPR with 95% bootstrap CI, and AUROC; bold marks the best per column. Numbers are measured by us, not quoted from other papers.

Two findings stand out. First, timing-only classifiers are unreliable against timing-imitating agents: fitting the marginal lognormal of the clickstream’s gaps leaves them at 0.349–0.354, since higher-order structure such as gap autocorrelation remains exploitable. Second, once any strong classifier is given the full behavioral feature set, the classifier choice barely matters: the all-feature random forest, the behavior-only gradient boosting, and HABITS all land at 0.957–0.961, and paired tests show them equivalent (Section 6.9). The contribution is therefore the behavioral feature families and the low-false-positive protocol, not a particular classifier. Thirteen classifiers on the same feature set confirm this (Table 7, appendix): every capable ensemble and margin-based model lands between 0.94 and 0.97, while a single decision tree degenerates at the operating point, 0.000 despite AUROC 0.955, because its coarse probability estimates cannot resolve the extreme tail. A model can look strong by AUROC and be useless at 1% FPR, which is why we select and report at the operating point.

Family removed	R@1%	Δ [95% CI]	Holm p	Alone
Goal-directedness	0.889	−0.069 [−0.095, −0.039]	0.0005*	0.939
Timing regularity	0.947	−0.010 [−0.029, +0.004]	1.00	0.550
Causal consistency	0.955	−0.002 [−0.009, +0.004]	1.00	0.144
Structural redundancy	0.956	−0.001 [−0.006, +0.005]	1.00	0.000
Navigation entropy	0.961	+0.005 [−0.003, +0.010]	1.00	0.652

Table 3: Leave-one-family-out ablation (Setting A, extra-trees), with paired tests. Full-model recall at 1% FPR is 0.957. Δ is the change on removing the family, with a 10,000-resample group-aware paired bootstrap interval; Holm adjusts across the five families. “Alone” is that family used by itself. * marks $p \leq 0.05$: only goal-directedness has a contribution distinguishable from zero. Entropy and timing score well alone but cost nothing on removal, which is redundancy with the remaining families rather than weakness.

6.3 Single-family settings

Settings B and C are single-family probes, not comparable to the cross-benchmark grid because they use different featurizations and splits. Setting B (element-level action strings, whose vocabulary richness formulaic agents lack) is easy at 0.965 [0.895, 0.978]: short, low-diversity agent action sequences are readily separable. Setting C (interaction texture alone, legitimate user versus impostor) is hard at 0.148 [0.096, 0.193]. This mirrors the weakness of timing alone in Setting A: no single modality carries the task.

6.4 Ablation and the effect of agent realism

Table 3 gives the leave-one-family-out ablation for Setting A with a paired test on each delta, which changes what can be claimed. Exactly one family has a contribution distinguishable from zero: removing goal-directedness costs 0.069 recall (Holm $p = 0.0005$). The other four deltas are within noise of zero and not separable from each other, so we do not read an ordering off them; the gap between goal-directedness and the next family is itself significant ($p = 0.0002$), the one comparative statement the data support.

Two of those four are informative in isolation yet cost nothing on removal: navigation entropy reaches 0.652 alone and timing 0.550. That is redundancy rather than weakness. The remaining families carry the same distinction, which is why the fusion degrades gracefully.

Agent realism is what makes the problem hard: without timing humanization every method reaches near-perfect recall. Once timing is humanized, the recoverable signal shifts to goal-directedness, which is structural rather than temporal and so

could remain useful even as agents improve temporal realism; this is why the study controls realism explicitly rather than reporting against a single weak agent.

6.5 Do real agents behave the way our generator assumes?

Our agent class is generated, and the family the ablation finds load-bearing is parameterized with the regularity it measures. Circularity is therefore the sharpest objection to the result, and we test it against real agent trajectories. We use a public multi-shop benchmark (Peeters et al., 2026): all 1,456 task runs across 4 model backends on live storefronts, 41,114 timestamped steps, of which 799 carry enough funnel structure to featurize. This is an agent-side check only. The benchmark ships no human sessions, so it cannot furnish a human-versus-agent recall number and we do not compute one (Appendix A).

Two of the three planted quantities are testable. Funnel monotonicity holds: every usable session respects $\#view \geq \#cart \geq \#purchase$ (799/799 sessions, 95% CI [0.995, 1.000]), so the ordering our generator imposes is not an artifact of the generator. The browse share is not. Real agents browse *more* than we assume, 85.0% of actions against the 0.70 our generator draws ($p < 10^{-50}$), so our simulated agents are more purposeful than real ones and that parameter makes the task easier than reality. The third quantity, purchasing without viewing, is not measurable here: order-confirmation URLs identify an order, not its line items.

The reading is mixed and we report it as such. The structural assumption the load-bearing family depends on survives contact with real agents; one realism parameter does not, and in the direction that flatters us. This bounds the circularity concern rather than dissolving it, and it identifies browse mix as the parameter to fix first. We find no model-family signature in these features (Appendix A), which also means the check is not an artifact of one backend.

Used *alone* (Appendix C), no family matches the fusion: against timing-imitating agents structural redundancy is useless (0.000) and causal consistency nearly so (0.144), timing and entropy are middling (0.551, 0.652), and goal-directedness is the strongest single view (0.939) but still trails the fusion (0.957). The distinction lives in the *combination* of individually insufficient views, as the aggregation argument of Section 3 predicts.

FPR	Recall	SD	Thr. var.	FP budget	FPR bound
0.1%	0.804	0.0375	99%	20	0.0014 (1.45 $\times$)
0.5%	0.932	0.0085	87%	100	0.0059 (1.18 $\times$)
1.0%	0.972	0.0034	69%	200	0.0112 (1.12 $\times$)

Table 4: Our fusion at three operating points on the primary corpus, with the two sources of its uncertainty separated. “Thr. var.” is the share of the variance of recall attributable to estimating the threshold on the human sample rather than to measuring recall on the agent sample; it dominates everywhere and grows as the operating point tightens, so more agent data would barely narrow these intervals. “FPR bound” is the one-sided exact 95% upper limit on the true FPR at the fitted threshold, with its ratio to the nominal target: this is what a deployment can be promised.

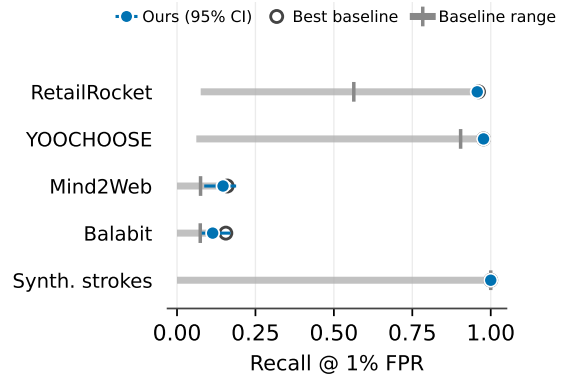


Figure 3: Recall at 1% FPR per benchmark: our fusion (blue, 95% bootstrap CI) against the best baseline (open circle) and the full baseline range (grey band, tick at the median). The fusion is at or within the CI of the best everywhere, while every competing family fails on at least one benchmark. Per-method numbers are in Table 9.

6.6 Cross-benchmark evaluation

To test whether the method and protocol transfer, we repeated the full comparison on five public benchmarks under the identical protocol: two e-commerce action streams with real human sessions (RetailRocket (Retail Rocket, 2016), YOOCHOOSE (Ben-Shimon et al., 2015)), one web-task stream without timestamps (Mind2Web (Deng et al., 2023)), and two mouse-dynamics tasks (the Balabit impostor task (Fülöp et al., 2016) and a synthetic-stroke condition where we reimplement the published generator of Acien et al. (2022) and pair its output against real strokes segmented from Balabit, using no data from that benchmark). In the hard e-commerce regimes the agent draws gaps from a lognormal fit to that corpus’s own human gaps. Corpus details are in Appendix C.

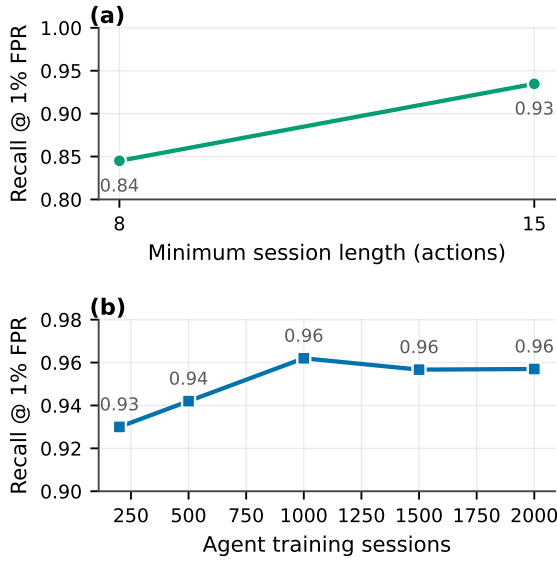


Figure 4: (a) Recall at 1% FPR rises with minimum session length, the trend Proposition 1 predicts (longer bins are underpowered: few human sessions exceed 25 actions). (b) Recall against agent training-set size, flattening by about 1,000 sessions.

Figure 3 summarizes HABITS against the baseline field; the full per-method grid is Table 9 in the appendix. Here HABITS uses gradient-boosted trees: randomized-tree probability estimates tie at 1.0 on pure leaves and zero out recall at a 1% FPR threshold on short-session corpora, while boosted margins do not tie (Table 7 shows classifier choice is otherwise immaterial). Corpus sizes and easy-regime counterparts are in Appendix C. HABITS is the only method that stays near the top on all five benchmarks; every baseline family fails somewhere, with the shallow tree at 0.000 on Balabit, one-class models near zero on four benchmarks, and sequence models unstable. Matching the marginal timing distribution weakens timing-only classifiers without eliminating them, since a lognormal fit leaves gap autocorrelation and think-time exploitable. Absolute recall tracks how much behavioral structure a benchmark exposes: rich funnels (RetailRocket, YOOCHOOSE) support recall above 0.95 against timing-matched agents, while short templated corpora (Mind2Web) leave less to aggregate, the session-length dependence of Section 6.7. On synthetic strokes we reproduce Acien et al. (2022): they are fully separable from real ones, a ceiling shared by all supervised methods and excluded from win counts.

6.7 Session length and the aggregation argument

Proposition 1 predicts that separability improves with session length, since the session mean of weak per-action evidence concentrates as n grows. Figure 4(a) agrees: recall rises from 0.845 on sessions of at least 8 actions to 0.935 at 15 or more. We read this qualitatively, because the filter also shifts the population and few human sessions exceed 25 actions. Figure 4(b) shows recall climbing from 0.930 at 200 agent training sessions to about 0.96 by 1,000, then flattening: the bottleneck is agent diversity, not count.

6.8 What a low-FPR operating point can promise

A point estimate of recall at 1% FPR leaves two questions open: how much of the uncertainty comes from the human sample rather than the agent sample, and what false-positive rate a threshold fitted on finite humans guarantees. Both matter for deployment (Table 4).

Recall at a fixed FPR composes two estimates, a threshold read off the humans and a proportion measured on the agents. Resampling one class at a time separates them. The threshold term dominates everywhere, and more so as the point tightens: 69% of the variance at 1% FPR, 87% at 0.5%, 99% at 0.1%. More agent sessions would barely narrow these intervals. The human sample is the binding constraint, which is the opposite of where effort usually goes. A threshold tuned to 1% FPR on n humans also does not guarantee 1%; it guarantees the one-sided exact upper limit, $1.12\times$ the target at our 20,000-human scale and $1.45\times$ at the 0.1% point. We quote that bound with each operating point, since it is what a deployment inherits.

6.9 Which differences are significant

Every method sees the same sessions and folds, so the comparisons are paired and we test them as such. Comparing two marginal intervals is not a test: overlap does not establish equivalence, and non-overlap overstates a difference when the scores are correlated. We compare the fusion to each baseline with a paired bootstrap on recall at 1% FPR, a DeLong test on AUROC, and exact McNemar at a matched operating point, correcting across baselines with Holm (Appendix C.1).

The fusion separates from every restricted-view baseline. On the clickstream it gains 0.40 to 0.92

Baseline	Δ	Diff. p	Equiv. p	Verdict
Random forest	-0.003	0.158	0.003	EQUIV.
XGBoost	-0.001	0.496	$< 10^{-4}$	EQUIV.
SVM (RBF)	+0.009	0.004	0.163	DIFF. [†]
LogReg, all features	+0.023	0.001	1.00	DIFFERS
Decision tree ($d=4$)	+0.403	0.001	1.00	DIFFERS
Transformer	+0.506	0.001	1.00	DIFFERS
Timing-only NB	+0.552	0.001	1.00	DIFFERS
Timing-only LogReg	+0.580	0.001	1.00	DIFFERS
Isolation forest	+0.756	0.001	1.00	DIFFERS
GRU	+0.804	0.001	1.00	DIFFERS
Deep SVDD	+0.922	0.001	1.00	DIFFERS

Table 5: Difference versus equivalence at 1% FPR on the primary corpus (20,000 humans / 8,000 agents). Δ is ours minus the baseline. The difference test asks whether the two differ; the equivalence test (TOST, margin $\delta = 0.01$, fixed in advance) asks whether they are the same to within one point of recall, so a small value there is a positive finding rather than an absence of one. Random forest and XGBoost come out *equivalent* to the fusion rather than only untested, and every restricted-view baseline differs. [†]SVM is the third case the pair of tests can produce: the difference is real but its interval crosses the margin, so it is nonzero and smaller than one point at once. A dash marks a comparison that resolves neither way.

recall over the sequence models, one-class models, shallow tree, and timing-only classifiers, all at Holm $p \leq 0.001$, and a permutation test agrees on every comparison (Tables 12 and 13). Random-effects pooling across the four corpora puts the gain at +0.72 over timing-only methods, +0.69 over one-class, and +0.53 over sequence models (Table 6).

Against a strong classifier given the same features we can do better than fail to reject. We test for equivalence within a margin of one point of recall, fixed in advance: random forest and XGBoost come out *equivalent* to the fusion (Table 5), and the pooled effect over all nine strong-ensemble comparisons is +0.0007, 95% interval $[-0.001, +0.003]$. The estimator contributes at most a fraction of a point. This supports the paper’s claim: what separates humans from agents is the behavioral view, not the model on top of it, and we can now bound how little the model matters.

6.10 Adaptive evasion and robustness

A detector is only useful if it survives an adversary who knows the features. Holding the trained fusion fixed, we search the generator’s knobs (session length, browse-to-target mix, purchase and cart rates, break rate) to make agents look as human as possible to it. Recall falls from 0.98 to 0.62 on the clickstream and 0.98 to 0.64 on YOOCHOOSE, so

Baseline family	k	Pooled Δ [95% CI]	p	I^2
Timing-only	7	+0.7178 [+0.586, +0.850]	$< 10^{-4}$	0.99
One-class	8	+0.6896 [+0.489, +0.890]	$< 10^{-4}$	0.99
Sequence	8	+0.5329 [+0.217, +0.849]	9×10^{-4}	1.00
Shallow tree	4	+0.0051 [-0.005, +0.016]	0.333	0.92
Linear, all features	3	+0.0152 [+0.002, +0.029]	0.026	0.89
Strong ensemble	9	+0.0007 [-0.001, +0.003]	0.426	0.60

Table 6: Random-effects pooling of the recall gain at 1% FPR by baseline family, over k comparisons across four approved corpora. τ^2 is estimated by DerSimonian–Laird and I^2 reports heterogeneity. The gain over restricted-view families is large and survives pooling; against strong ensembles the pooled effect is +0.0007 with an interval spanning zero and the lowest heterogeneity of any row, which bounds the estimator’s contribution tightly rather than leaving it untested.

a generator that targets the features can evade them. One retraining round on the evasive cohort recovers part of the loss, to 0.65 and 0.72, and moves weight away from goal-directedness, so the families degrade at different rates rather than failing together. Since goal-directedness is the only family the ablation finds load-bearing (Table 3), an adversary who defeats it is the main threat here.

The rest of the protocol is stable (Table 11). A temporal split, training on the earliest sessions and testing on the latest, gives 0.973 and matches the random-split result, so there is no temporal leakage. Hard-negative mining over unused human sessions raises recall from 0.955 to 0.964 at a fixed false-positive rate, beating a random selection of the same size.

7 Conclusion

We asked whether LLM-driven shopping sessions are behaviorally distinguishable from human sessions, using only public data. Weak per-action evidence aggregates into a session-level distinction. Timing-only classifiers fail against timing-imitating agents, while the fused families stay at or near the best method on all five benchmarks. Paired tests place the gain in the behavioral view rather than the estimator: the fusion separates from every restricted-view baseline and is equivalent to strong classifiers given the same features. Testing the ablation rather than ranking it narrows the story to one load-bearing family, and checking that family’s assumption against real agent trajectories supports it while finding one realism parameter set too favorably. The released pipeline lets the question be studied without proprietary logs.

Limitations

This is a study of distinguishability, and several limits follow. The agent corpus, though diverse across model families, is generated by us rather than observed in the wild, so the measured separation is an estimate under our simulation; the realism-controlled loop is designed to keep tightening that estimate. Our experiments validate four of the six signal families directly on public data (timing, navigation predictability, causal consistency, and goal/structural signals in Settings A and the cross-benchmark study) plus interaction texture on the mouse corpora; the sequence-surprisal and representation families are described but not separately validated, because the public datasets lack the paired dense-interaction and rich-embedding views those families need. A further limit is circularity: the agent classes are produced by our own generators, and the family the ablation identifies as load-bearing, goal-directedness, measures a regularity the generators are parameterized to have, so the measured separation is an upper bound conditional on the simulator. Because that single family carries the ablation (Table 3), the circularity concern concentrates on it rather than being spread thinly across six, which makes it a sharper caveat and not a milder one. The cross-benchmark study varies the environment and the human corpus, but not the generator itself, so a single generator gives an optimistic estimate and coverage of generator paradigms remains the binding constraint. The distinguishability bound (Proposition 1) assumes conditionally independent per-action evidence, which real sessions violate, so it is informative about the qualitative shape of the trend rather than a numeric guarantee. The study is two-class (human versus LLM agent) by construction, and the human samples are drawn from a small number of public datasets that may not represent the full diversity of real users.

A final limit concerns data provenance, and it constrains what we can study. The public shopping-agent environments we examined trace their product catalogs to a single source: one benchmark’s catalog was scraped from a commercial retailer without a data licence, and two later environments state that their shopping sites reuse that same content. We therefore exclude those environments and every trajectory collected on them, which is why our e-commerce evidence rests on clickstream logs released by retailers for research rather than on

an interactive storefront. The cost is that we cannot pair real LLM-agent trajectories with human sessions in a shopping environment. No public environment currently satisfies all three of the requirements such a pairing needs: an e-commerce setting, released real agent trajectories, and catalog provenance independent of the scraped source. The nearest candidates each fail one. One populates its storefronts from a web crawl but states no licence for the released data; one synthesises its catalog with a language model but releases no trajectories; and one distributes real agent trajectories permissively but records no timestamps, which leaves the timing family uncomputable. Our agent class is consequently simulated throughout, and validating it against real agent behavior in a provenance-clean shopping environment is the clearest next step for this line of work.

Ethics Statement

This work studies whether LLM-driven shopping sessions are behaviorally distinguishable from human sessions. It uses only public, aggregated, de-identified datasets and simulated agent sessions that we generate; no individual is identified, and no private or proprietary data is used. The intent is scientific understanding of agentic browsing behavior as autonomous agents become common in e-commerce, and the methodology could help platforms design experiences that account for both human and agent traffic. We note a dual-use consideration: any behavioral distinction can in principle be used to exclude agents, and any published distinction can guide an agent toward more human-like behavior. We mitigate this by working only at the aggregate, methodological level, by not releasing any model tuned to a specific platform, and by generating agent sessions only on test tasks that stop before checkout or payment. We believe open study on public data is preferable to leaving the question to proprietary systems, because it lets the research community understand and audit the signals involved.

References

- Alejandro Acien, Aythami Morales, Julian Fierrez, and Ruben Vera-Rodriguez. 2022. BeCAPTCHA-Mouse: Synthetic mouse trajectories and improved bot detection. *Pattern Recognition*, 127:108643.
- Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian

- Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and don'ts of machine learning in computer security. In *USENIX Security Symposium*.
- Stefan Axelsson. 2000. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security (TISSEC)*, 3(3):186–205.
- Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. 2009. Action understanding as inverse planning. *Cognition*, 113(3):329–349.
- Albert-László Barabási. 2005. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.
- David Ben-Shimon, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle. 2015. [RecSys challenge 2015 and the YOOCHOOSE dataset](#). In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 357–358.
- Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. 2014. Uncovering large groups of active malicious accounts in online social networks. In *ACM Conference on Computer and Communications Security (CCS)*.
- Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- C. J. Clopper and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.
- Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley-Interscience.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Rebecca DerSimonian and Nan Laird. 1986. [Meta-analysis in clinical trials](#). *Controlled Clinical Trials*, 7(3):177–188.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
- Bradley Efron. 1987. [Better bootstrap confidence intervals](#). *Journal of the American Statistical Association*, 82(397):171–185.
- Masoud Erfani, Farzaneh Shoeleh, and Ali A. Ghorbani. 2020. Financial fraud detection using deep support vector data description. In *IEEE International Conference on Big Data (Big Data)*.
- Árpád Fülöp, Levente Kovács, Tamás Kurics, and Eszter Windhager-Pokol. 2016. Balabit mouse dynamics challenge data set. <https://github.com/balabit/Mouse-Dynamics-Challenge>.
- Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63(1):3–42.
- Steven Gianvecchio, Mengjun Xie, Zhenyu Wu, and Haining Wang. 2008. Measurement and classification of humans and bots in internet chat. In *USENIX Security Symposium*.
- K.-I. Goh and Albert-László Barabási. 2008. Burstiness and memory in complex systems. *Europhysics Letters (EPL)*, 81(4):48002.
- Wassily Hoeffding. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.
- Sihao Hu, Zhen Zhang, Bingqiao Luo, Shengliang Lu, Bingsheng He, and Ling Liu. 2023. BERT4ETH: A pre-trained transformer for Ethereum fraud detection. In *The ACM Web Conference (WWW)*.
- Christos Iliou, Theodoros Kostoulas, Theodora Tsikrika, Vasilis Katos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2021. [Detection of advanced web bots by combining web logs with mouse behavioural biometrics](#). *Digital Threats: Research and Practice*, 2(3).
- Hugo Jonker, Benjamin Krumnow, and Gabry Vlot. 2019. [Fingerprint surface-based detection of web bot detectors](#). In *European Symposium on Research in Computer Security (ESORICS)*, pages 586–605.
- Jan Kadel, August See, Ritwik Sinha, and Mathias Fischer. 2024. BOTracle: A framework for discriminating bots and humans. In *European Symposium on Research in Computer Security (ESORICS) International Workshops*. Springer. ArXiv:2412.02266.
- Dayeon Kang, Hyejun Jeong, Jade Sheffey, Pubali Datta, and Amir Houmansadr. 2026. Whose agent are you? multi-layer fingerprinting and attribution of autonomous web agents. *arXiv preprint arXiv:2606.20910*.
- Thomas Kipf, Yujia Li, Hanjun Dai, Vinicius Zambaldi, Alvaro Sanchez-Gonzalez, Edward Grefenstette, Pushmeet Kohli, and Peter Battaglia. 2019. CompILE: Compositional imitation learning and execution. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.
- Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding web browsing behaviors through weibull analysis of dwell time. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *IEEE International Conference on Data Mining (ICDM)*.

- Claude Nadeau and Yoshua Bengio. 2003. [Inference for the generalization error](#). *Machine Learning*, 52(3):239–281.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Ralph Peeters, Aaron Steiner, and Christian Bizer. 2026. WebMall: A multi-shop benchmark for evaluating web agents. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ArXiv:2508.13024.
- Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. 2019. TESSERACT: Eliminating experimental bias in malware classification across space and time. In *USENIX Security Symposium*.
- Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. 2020. Intriguing properties of adversarial ML attacks in the problem space. In *IEEE Symposium on Security and Privacy (S&P)*.
- Retail Rocket. 2016. Retailrocket recommender system dataset. Kaggle, <https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset>.
- Lukas Ruff, Robert Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.
- Donald J. Schuirmann. 1987. [A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability](#). *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.
- Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.
- Kihoon Son, Jinhyeon Kwon, DaEun Choi, Tae Soo Kim, Young-Ho Kim, Sangdoo Yun, and Juho Kim. 2024. Unveiling disparities in web task handling between human and web agent. *arXiv preprint arXiv:2405.04497*.
- Pang-Ning Tan and Vipin Kumar. 2002. Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, 6(1):9–35.
- Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ethan Wang, Zubair Shafiq, and Yash Vekaria. 2026. FP-Agent: Fingerprinting AI browsing agents. *arXiv preprint arXiv:2605.01247*.
- Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y. Zhao. 2013. You are how you click: Clickstream analysis for sybil detection. In *USENIX Security Symposium*.
- Shuhao Wang, Cancheng Liu, Xiang Gao, Hongtao Qu, and Wei Xu. 2017. Session-based fraud detection in online e-commerce transactions using recurrent neural networks. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Springer.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*.
- Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. 2008. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*.
- Jacob Ziv and Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536.

A The Real-Agent Check in Detail

Why no recall number. The benchmark releases agent trajectories and no human sessions. Obtaining a recall figure would mean pairing its agents against some other retailer’s clickstream, which varies site, catalogue, year, instrumentation, task motivation and harness simultaneously. There is no single confounder to match on, so no post-hoc control recovers an interpretable number, and we do not report one. Both quantities in Section 6.5 compare agents against a declared simulator parameter, so neither needs a human class.

Mapping trajectories to sessions. The released records are browser-level: one action string per step (click, fill, goto), the resulting URL, and absolute timestamps. We read the funnel stage from the URL and the action together, never the action alone, since a click covers everything from a facet filter to a checkout button. Product pages map to views,

add-to-cart requests and cart pages reached by interaction map to cart events, and only the order-confirmation page maps to a purchase. Search-result and navigation steps are dropped rather than counted as views: calling a search a product view would inflate the browse count and make funnel monotonicity measure our mapping instead of the agent, which is the quantity under test. Consecutive identical events on the same object are collapsed, because an agent filling a checkout form emits one step per field. Product identity is namespaced by shop host, since the benchmark spans four storefronts and an id is unique only within one.

A caveat on timing. The timestamps are genuine wall-clock times, but they are harness times: inter-action gaps are dominated by model-provider latency in a batch run rather than by anything resembling deliberation. We therefore compute no timing feature on this corpus and make no timing comparison against humans; the check is purely structural.

Scale. All 1,456 released task runs were retrieved, giving 41,114 step records, of which 799 runs yield at least three funnel events and are usable. Roughly a tenth of raw steps survive the mapping, the remainder being search and navigation.

No model-family signature at the funnel level.

We also asked whether the backend can be identified from these features across the 4 backends. It cannot: accuracy is 0.285 (95% CI [0.254, 0.318]) against a majority-class baseline of 0.304, which is at chance ($p = 0.88$). We report this as a negative result and build nothing on it. Two readings are consistent with it, and this corpus cannot separate them: either different models produce similar funnel-level behavior, or the 17 features that survive our mapping are too coarse to tell them apart. The second is plausible, since the mapping discards the search and navigation steps where planning style would most plausibly show. An earlier version of this analysis on a partially retrieved corpus appeared to find a signature, at 0.585 against a 0.516 baseline; that was an artifact of having only two of the four backends present, and it disappeared once the corpus was complete. We note it because it is the kind of result a partial download invites.

B Signal Families in Detail

Section 5 states what invariant each family captures. This section gives the feature-level definitions.

Timing regularity. Let $\tau_1, \dots, \tau_{n-1}$ be the interaction gaps with mean μ and standard deviation σ . Alongside the burstiness coefficient $B = (\sigma - \mu)/(\sigma + \mu)$ we compute the coefficient of variation, a lag-1 autocorrelation of consecutive gaps (a memory term), the fraction of very short gaps, and the think-time ratio: the median gap immediately before consequential actions (add-to-cart, checkout) divided by the session median.

Navigation predictability. Over the action stream we compute the marginal Shannon entropy $H_1 = -\sum_a p(a) \log p(a)$, the conditional next-action entropy H_2 , and a compression rate from a Lempel-Ziv parse. The goal-directedness score adds funnel-progress monotonicity (the rank correlation between position in the session and funnel depth), a path-efficiency ratio (shortest path to the deepest reached stage over the realized length), and the entropy of sub-goal chunk lengths.

Causal consistency. From the break indicator b_t the family derives the break count, the position of the first break, and a break density.

Interaction texture. The presence, rate, and low-order transition structure of interaction events, with movement-speed and direction-change statistics and a path-straightness ratio where a pointer trace is available.

Sequence surprisal. Two complementary objectives on human sessions only: a causal model predicting the next event, reported as mean per-step cross-entropy, and a bidirectional masked model reconstructing held-out events from context.

Representation separation. A one-class model of the human region of a learned session representation (reconstruction error) together with a supervised separation over that representation: a cross-fitted human-versus-agent head, a nearest-neighbor agent-density estimate, and a centroid margin. Being partly supervised, this family improves as the agent corpus grows.

C Reproducibility Details

All experiments run on public data with standard libraries (Pedregosa et al., 2011). Classifiers use scikit-learn defaults except where noted (extra-trees and random forest with 200 estimators; gradient boosting with 150; logistic regression with 1,000 max iterations). Cross-validation is five-fold, grouped by session source. Confidence inter-

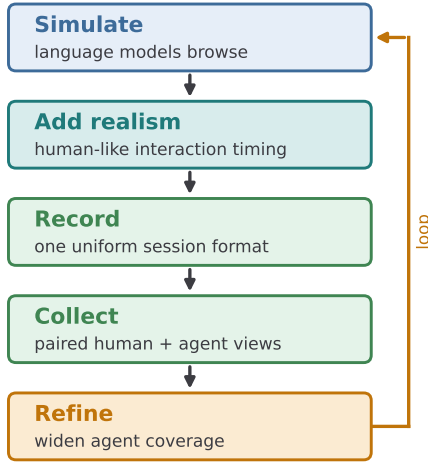


Figure 5: The simulation-and-study loop: language models drive browsing, a realism layer reproduces human-like timing, a uniform recorder captures each session, and the hardest cases guide the next round.

vals are 1,000-sample stratified bootstraps over the pooled out-of-fold scores. The simulation methodology, feature extractors, and evaluation scripts operate on the public datasets described in Section 6 and will be released.

Cross-benchmark study (Section 6.6). Every benchmark runs the same method ladder from one shared implementation, so a method’s number differs across columns only because the data differs. *Data.* RetailRocket: the Kaggle events log (2.76M events); events are split into sessions at 30-minute gaps, sessions of 3 to 60 actions kept, the first 5,000 used, grouped by visitor id. YOOCHOOSE: the first 3M click records of the RecSys 2015 release joined with buys, sessions of 3 to 60 actions, first 5,000 kept; the dataset has no visitor id, so each human session is its own group and the group-aware CV degenerates to stratified CV there. Mind2Web: the first 800 training demonstrations (769 usable with at least 3 actions), operation verbs (CLICK, TYPE, SELECT, HOVER) as tokens; human folds are grouped by website, and the templated agents (3–7 actions, click-dominant verb mix) are grouped in batches. Balabit: genuine training sessions (label 0) plus the challenge’s official legal-versus-impostor test labels, grouped by account. Synthetic strokes: real strokes segmented from Balabit genuine sessions (pauses over 0.5 s split strokes, minimum 20 events, at most 40 strokes per session,

2,600 total) against 4,000 strokes we generate from the function-based method described by Acien et al. (2022) (one third each linear, quadratic, exponential; 20–120 points; 1-pixel coordinate noise; near-constant sampling jitter). We reimplement the generator from its published description and use no data from that benchmark, so no data licence is involved. *Agents.* In the e-commerce hard regimes the agent’s inter-action gaps are drawn from a log-normal whose (μ, σ) are fit to the log-gaps of that corpus’s human sessions (RetailRocket 11.04, 1.59 in ms; YOOCHOOSE 3.99, 1.39 in s); the easy regimes use near-constant machine gaps. Behavior (funnel progression, catalog usage, break rate) is identical across regimes, so the two regimes isolate the timing family. *Methods.* Timing-only baselines see only the timing-family columns. One-class models fit on the human rows of each training fold only. Sequence models read the token stream plus the per-step log gap, truncated to 64 steps (GRU: 16-d embedding, 32-d hidden; transformer: 2 layers, 4 heads, 32-d, mean-pooled; 25 epochs, Adam, learning rate 10^{-3}); Deep SVDD trains 60 epochs at the same rate. The fusion is gradient boosting rather than a randomized-tree ensemble because tree-ensemble probability estimates saturate at exactly 1.0 on pure leaves, producing top-of-ranking ties that zero out recall at a 1% FPR threshold even at high AUROC; boosted margins do not tie. All runs fix seed 42; the per-benchmark JSON results and the table generator are part of the released scripts.

C.1 Significance testing

Because every method is scored on the identical sessions under the identical folds, all comparisons are paired, and we test the paired difference directly rather than inspecting two marginal intervals. The null hypothesis in each test is that the two methods have equal performance on the population the corpus is drawn from.

Recall at a fixed FPR. Recall at 1% FPR is a quantile-thresholded functional, not a sample mean, so it has no closed-form variance. We use a paired bootstrap with $B = 10,000$ resamples. Positives and negatives are resampled separately so class prevalence stays fixed, and when a benchmark supplies group ids we resample groups rather than sessions, so the resampling unit matches the cross-validation unit and correlated sessions from one source cannot inflate significance. Both methods

are scored on the same resample, which removes the between-session variance that dominates each method’s marginal interval. The two-sided p -value is $2 \min\{\mathbb{P}(\hat{\Delta} \leq 0), \mathbb{P}(\hat{\Delta} \geq 0)\}$ over the resamples. A resampling p -value cannot be smaller than $1/B$, so we floor it there and report $< 10^{-4}$ rather than zero.

AUROC. For AUROC we use the DeLong test, which gives the exact asymptotic covariance of two AUCs computed on the same cases from their midrank placement values, so no resampling is needed. We verified our implementation against `sklearn` for the point estimate and confirmed its null calibration empirically: on 400 replications of two equal-strength scorers, 5.8% of p -values fell below 0.05 with a median of 0.47.

Operating point. Once a threshold is fixed, each method either flags a session or not, so we compare the decisions with McNemar’s test on the discordant pairs. Each method’s threshold is set to its own 1% FPR so the comparison is like-for-like, and we count discordance on the positive class, since at a matched FPR the negative-side decisions are constrained by construction. We use the exact binomial form rather than the chi-square approximation because discordant counts are small at a 1% FPR.

Multiplicity. Each benchmark contributes one test per baseline, so we control the family-wise error rate within a benchmark with Holm’s step-down procedure, which is uniformly more powerful than Bonferroni and assumes nothing about dependence between the baselines. We also compute Benjamini–Hochberg adjusted values for reference; the tables report Holm, the conservative choice. One comparison changes conclusion under correction, which is why we apply it.

Testing an ablation. An ablation delta is a difference between two refit models, so it carries sampling noise like any other, and an ordering read off five point deltas is not a finding. We treat each family’s delta as a paired comparison between the full and ablated model on the same sessions and folds, give it a bootstrap interval, and correct across the five families with Holm. Because the interesting claim is comparative, we also test the pairwise differences of the deltas; only the pairs that separate are reported as an ordering. On our primary setting exactly one family separates from zero and

exactly one adjacent pair separates from its neighbour, which is why Section 6 reports a single load-bearing family rather than a ranking of five.

Equivalence, where a difference test is not enough. A large p from a difference test says the data are consistent with no difference; it does not exclude a difference the study was too small to detect. Since several of our comparisons are ties, and we want to make a claim about them rather than report an absence, we also test for equivalence with two one-sided tests (Schuirmann, 1987). Against a margin δ we test $H_0^- : \Delta \leq -\delta$ and $H_0^+ : \Delta \geq +\delta$ and declare equivalence only when both are rejected, so $p_{\text{TOST}} = \max(p^-, p^+)$ and passing at α is exactly the statement that the central $(1 - 2\alpha)$ interval of $\hat{\Delta}$ lies inside $(-\delta, +\delta)$. We fix $\delta = 0.01$, one point of recall, before looking at any p -value; it is below the fold-to-fold spread of every method we report. Table 5 gives the result.

A second route to the same difference. A bootstrap p -value inherits whatever the resampling scheme assumes. As a check that our conclusions are not an artifact of that construction, we repeat each comparison as a paired randomization test under the null that the two methods are exchangeable. Scores cannot be swapped directly, because the two methods have different score distributions and swapping would move each method’s threshold; we therefore rank-transform within method first, which recall at a fixed FPR is invariant to, and swap ranks, which leaves each marginal distribution exactly intact. Flips are applied per group so correlated sessions move together. We use the add-one estimator $(1 + \#\{|\Delta^*| \geq |\hat{\Delta}|\})/(1 + n)$, which cannot report zero. The permutation and bootstrap p -values agree on every comparison in Table 13.

Skew-corrected intervals. Recall at 1% FPR sits near its ceiling, so its sampling distribution is left-skewed and the percentile interval is mis-centred. We report bias-corrected and accelerated intervals (Efron, 1987), which shift which percentiles are read off by a bias term z_0 taken from the resample mass below the estimate and an acceleration a estimated from the skewness of delete-a-block jackknife influence values. On our primary corpus $z_0 = -0.21$ and $a = -0.035$, and the correction moves the lower limit by about half a point of recall.

Variance decomposition and the FPR bound.

Section 6.8 reports these results; here is how they are computed. To split the variance of recall at a fixed FPR we resample one class at a time: negatives only, with the positives held fixed, gives the contribution of estimating the threshold; positives only gives the contribution of measuring recall; the joint resample gives the total, and the residual is the interaction. The promisable false-positive rate is the one-sided exact upper limit $B_{0.95}^{-1}(k+1, n-k)$ on k observed false positives among n negatives (Clopper and Pearson, 1934). The bound narrows only with n , which is the quantitative form of the rule that a target rate of p needs on the order of $1/p$ negatives before the estimate carries meaning.

Pooling across benchmarks. A result on one benchmark can be luck, and averaging across benchmarks would wrongly assume a single common effect when the corpora differ in task and instrument. We pool each per-baseline-family effect with a DerSimonian–Laird random-effects model, which admits a between-benchmark variance τ^2 estimated from Cochran’s Q (DerSimonian and Laird, 1986), and we report I^2 so heterogeneity stays visible. Table 6 gives the pooled effects.

Does the corpus draw matter? Cross-validation accounts for the train/test split but not for the draw of the corpus itself: which agent sessions the generator produced, and how the folds fell. We therefore rebuild the corpus and refit everything at five seeds. Two things come out of it. The spread of recall across seeds is *smaller* than the mean within-run bootstrap interval at every operating point (ratios 0.37 to 0.76, all below one), so the intervals we report already absorb corpus-draw variance rather than understating it. And the sign of the difference against the strong ensembles is not stable across seeds: random forest is ahead of the fusion at 1% FPR on some draws and behind on others, which is the same conclusion the equivalence test reaches by a different route. The one case where the correction changes the reading is XGBoost at 1% FPR, where an uncorrected paired t -test gives $p = 0.074$ and the corrected test gives $p = 0.184$; the uncorrected value would have suggested a difference that does not replicate across draws.

Repeated runs. Where we compare across repeated cross-validation runs, the runs share training data, so the per-run differences are correlated and a naive paired t -test is anti-conservative. We use

the corrected resampled t -test (Nadeau and Bengio, 2003), which estimates the variance of the mean difference as $(1/J + n_{\text{test}}/n_{\text{train}}) S^2$ over J runs, with S^2 the sample variance of the per-run differences, and refers the statistic to t_{J-1} . At five folds the second term is 0.25 and dominates $1/J$, so the denominator grows substantially; we report the corrected and uncorrected statistics together so the size of the correction is visible.

Implementation checks. Statistical code fails quietly: a miscalibrated test still prints a plausible number. Each function is therefore checked against something external to it, and the suite of 46 checks is released with the code. The metric itself is checked against a brute-force search over all cut values; the permutation test is checked for null calibration (3.0% of p -values below 0.05 on 200 replications of two equal-strength scorers, median 0.53) and for power; the exact intervals are checked for coverage; and the equivalence test is checked to pass on identical inputs, fail on a genuine 0.05 gap at $\delta = 0.01$, and pass on that same gap at $\delta = 0.10$. Two defects in our own earlier implementation were found this way and are described in Appendix C.2.

We release the pooled out-of-fold scores alongside the result files, so any of these tests can be recomputed without retraining.

C.2 Two measurement pitfalls at low FPR

Both defects below were in our code, were found by the checks above, and are worth naming because neither announces itself and both are easy to reproduce with standard tools.

First, `sklearn’s roc_curve` defaults to `drop_intermediate=True`, which discards ROC vertices that are redundant for *plotting*. Those vertices are exactly the achievable operating points a low-FPR selection rule chooses among, so with the default the reported recall can be the best *surviving* point rather than the best point. On heavily tied scores, which is what a shallow tree’s `predict_proba` produces, this understates recall by up to 0.04 at no gain in FPR. On our corpora the effect is small because the scores are nearly continuous: of several hundred (method, operating point) cells, one moved, by 0.0008. The fix is one keyword, and the guard is a brute-force search over all cut values.

Second, a threshold and the recall reported at it must describe the same operating point. Our Mc-

Classifier	Recall@ 1%FPR [95% CI]	AUROC
Gradient boosting	0.970 [0.947, 0.977]	0.998
Random forest	0.961 [0.950, 0.971]	0.998
Hist. gradient boosting	0.958 [0.943, 0.968]	0.998
HABITS (extra-trees)	0.957 [0.947, 0.968]	0.996
MLP (64, 32)	0.957 [0.941, 0.967]	0.997
SVM (RBF)	0.956 [0.940, 0.970]	0.996
XGBoost	0.954 [0.943, 0.968]	0.998
Logistic regression	0.944 [0.920, 0.966]	0.997
AdaBoost	0.941 [0.916, 0.960]	0.996
k -NN ($k=15$)	0.928 [0.884, 0.939]	0.991
Ridge (linear)	0.908 [0.876, 0.925]	0.995
Naive Bayes	0.888 [0.850, 0.916]	0.990
Decision tree	0.000 [0.000, 0.000]	0.955

Table 7: Setting A, thirteen classifiers on the identical feature set, corpus, and five-fold group-aware split (95% bootstrap CI). Capable ensembles cluster between 0.94 and 0.97; a single tree degenerates at 1% FPR despite a high AUROC. The features, not the classifier, carry the task.

Benchmark	Hum.	Ag.	R@1% [95% CI]	AUROC
RetailR. (hard)	5,000	2,000	0.957 [0.944, 0.974]	0.998
YOOCH. (hard)	5,000	2,000	0.977 [0.970, 0.983]	0.998
Mind2Web	769	600	0.147 [0.087, 0.188]	0.814
Balabit	476	405	0.114 [0.079, 0.170]	0.766
Synth. strokes	2,600	4,000	1.000 [1.000, 1.000]	1.000
RetailR. (easy)	5,000	2,000	1.000 [1.000, 1.000]	1.000
YOOCH. (easy)	5,000	2,000	1.000 [1.000, 1.000]	1.000

Table 8: Our fusion on each benchmark: corpus sizes, recall at 1% FPR with 95% bootstrap CI, and AUROC. Easy regimes (agents keep regular machine timing) are near-ceiling; the hard regimes bound what action streams alone recover once timing is matched.

Nemar path thresholded with $>$ against a threshold that requires $\geq$, so its discordant counts described a point marginally stricter than the recall reported beside them. The counts were plausible either way, which is why the invariant is now asserted directly: decisions taken at the returned threshold must reproduce the reported recall exactly.

The general lesson is that at a 1% FPR the reported number is determined by a handful of order statistics in the tail of the negative distribution, so conventions that are immaterial elsewhere, tie handling, curve simplification, and the choice between $>$ and $\geq$, become load-bearing.

Family	Method	E-commerce (timing-matched)		Web tasks	Mouse dynamics	
		RetailR.	YOOCH.	Mind2Web	Balabit	Synth.
Timing-only	NB, as in Gianvecchio et al. (2008)	0.349	0.106	–	0.086	1.000
	LogReg, as in Gianvecchio et al. (2008)	0.354	0.061	–	0.074	1.000
	LogReg, all features	0.940	0.971	0.087	0.064	1.000
	Tree, as in Kang et al. (2026)	0.564	0.972	0.075	0.000	1.000
Static features	RF, as in Wang et al. (2026)	0.961	0.977	0.112	0.156	1.000
	Boosted trees (standard)	0.955	0.975	0.160	0.123	1.000
	SVM (RBF)	0.956	0.973	0.038	0.116	1.000
One-class	Isolation forest, as in Liu et al. (2008)	0.153	0.759	0.018	0.007	0.000
	Deep SVDD, as in Erfani et al. (2020)	0.075	0.497	0.038	0.005	0.000
Sequence	GRU, as in Wang et al. (2017)	0.150	0.123	0.000	–	–
	Transformer, as in Hu et al. (2023)	0.636	0.904	0.140	–	–
Fusion	HABITS (ours)	0.957	0.977	0.147	0.114	1.000

Table 9: Recall at 1% FPR for every method on five public benchmarks, identical protocol (hard regimes: agents copy a lognormal fit to each corpus’s own human timing); bold = best per column. “As in” rows reimplement the classifier of the cited human-versus-automation or behavioral-fraud detection work on our features; unmarked rows are standard baselines. Timing-only rows are omitted on Mind2Web (no timestamps), sequence rows on the mouse tasks (no token vocabulary). HABITS is at or within the bootstrap CI of the best on every benchmark; no baseline family is consistent across all five.

Family used alone (Setting A)	Recall@1%FPR
Goal-directedness	0.939
Navigation entropy	0.652
Timing regularity	0.551
Causal consistency	0.144
Structural redundancy	0.000
HABITS (all families)	0.957

Table 10: Each family used alone (extra-trees). No single family matches the fusion, and the weakest are near-useless in isolation.

Robustness check	Recall@1%
Temporal split (early → late)	0.973
Evasion, clickstream (base/evaded/hardened)	0.98 / 0.62 / 0.65
Evasion, YOOCHOOSE (base/evaded/hardened)	0.98 / 0.64 / 0.72
Hard-neg. mining (5 rounds)	0.955 → 0.964

Table 11: Robustness suite on the e-commerce click-stream corpora. A temporal split (train on the earliest sessions, test on the latest) matches the random-split result, so there is no temporal leakage; an adaptive evasion attack lowers recall on both corpora but one hardening round recovers part of it; and iterative hard-negative mining tightens the human tail without raising FPR.

Benchmark	Ours / best base	Δ [95% CI]	Holm p
RetailRocket (hard)	0.957 / 0.961	-0.004 [-0.013, +0.007]	1.00
YOOCHOOSE (hard)	0.977 / 0.977	+0.000 [-0.002, +0.002]	1.00

Table 12: Our fusion against the *strongest* baseline on each benchmark, recall at 1% FPR. Δ is the paired difference with a 10,000-resample stratified group-aware bootstrap CI; p is Holm-adjusted across the baselines within a benchmark. * marks $p \leq 0.05$. The fusion separates from every restricted-view baseline (Table 13) but is indistinguishable from the strongest full-feature classifier on all benchmarks.

Baseline	Δ TPR [95% CI]	Holm p	Perm. p	Δ AUC	OR
Random forest	-0.003 [-0.01, +0.00]	0.158	0.004	+0.001	0.4
XGBoost	-0.001 [-0.01, +0.00]	0.496	0.418	-0.000	0.8
SVM (RBF)	+0.009 [+0.00, +0.01]	0.004*	$< 10^{-3}$	+0.002	4.6
LogReg, all features	+0.023 [+0.02, +0.03]	0.001*	$< 10^{-3}$	+0.001	9.9
Decision tree ($d=4$)	+0.403 [+0.20, +0.44]	0.001*	$< 10^{-3}$	+0.006	404.2
Transformer	+0.506 [+0.34, +0.64]	0.001*	$< 10^{-3}$	+0.026	506.9
Timing-only NB	+0.552 [+0.52, +0.59]	0.001*	$< 10^{-3}$	+0.041	158.9
Timing-only LogReg	+0.580 [+0.55, +0.61]	0.001*	$< 10^{-3}$	+0.044	186.7
Isolation forest	+0.756 [+0.69, +0.83]	0.001*	$< 10^{-3}$	+0.034	6046.0
GRU	+0.804 [+0.78, +0.83]	0.001*	$< 10^{-3}$	+0.421	586.0
Deep SVDD	+0.922 [+0.90, +0.94]	0.001*	$< 10^{-3}$	+0.064	3687.5

Table 13: Full paired comparison at 1% FPR on the primary corpus. Δ TPR carries a 10,000-resample group-aware paired bootstrap interval; Holm adjusts across the eleven baselines; the permutation column is an independent randomization test on the same difference and agrees with the bootstrap on every row; OR is the McNemar conditional odds ratio on discordant positives, an effect size for the operating point. * marks Holm $p \leq 0.05$.