How Many Pixels Is a Digit Worth? Place-Aware Coordinate Entropy for GUI Agent Confidence Estimation

Anonymous ACL submission

Abstract

GUI agents predict click coordinates as digit-token sequences, but standard text-LLM confidence estimation methods rank correct clicks from wrong ones only weakly, and GUI-specific alternatives pay K samples or new supervision for their estimates. We trace this to *place-value asymmetry*: bounding-box correctness makes the hundreds digit pixel-decisive and the units digit pixel-irrelevant, so uniform aggregation drowns the signal that determines correctness. The fix is to weight each digit's Shannon entropy by its place value. We call this **Place-Aware Coordinate Entropy** (PACE). Across 7 agents on ScreenSpot-Pro and ScreenSpot-v2, PACE wins both AUROC and selective accuracy on all 6 primary cells in a single forward pass, matching K-sample baselines at a fraction of the cost. PACE provides a reliable, per-click confidence estimate that turns coordinate-token internals into a practical safety primitive for GUI agent deployment.

1 Introduction

Open-source GUI agents predict pixel coordinates of clicks, keypresses, and scrolls from screenshots, driving professional desktop tasks end-to-end (Qin et al., 2025; Xue et al., 2026; Hong et al., 2024; Xu et al., 2025; Wu et al., 2024). Click accuracy on dense professional interfaces remains below half for the strongest open agents (35–50% on ScreenSpot-Pro (Li et al., 2025)), and a wrong click is often irreversible: a sent payment or a deleted file cannot be undone by resampling (Shi et al., 2025). Whatever downstream safeguard the deployment chooses (clarification, human review, abstention, escalation), it first needs to know which of the agent's own clicks are likely wrong. The grounding model must therefore expose a per-prediction confidence estimate that ranks correct clicks above incorrect ones, and ideally produces a calibrated probability one can threshold against.

For ordinary text generation, confidence estimation is a mature toolkit. Verbalized prompts (Tian et al., 2023; Kadavath et al., 2022), self-consistency over multiple samples (Wang et al., 2023), and log-probability (Malinin and Gales, 2021) all give usable confidence signals on factual question answering, semantic parsing, and reasoning traces. GUI click prediction shares the surface form of text generation: the model outputs a short digit-token sequence that decodes to screen coordinates. We find that every one of those signals produces weak or near-chance discrimination on this output (Table 1).

We trace this to a common mechanistic root. GUI click correctness is bounding-box-defined: any pixel inside the target counts as correct. The hundreds digit of a coordinate therefore corresponds to UI-element-level shifts (a 1-unit confusion is a 100-pixel error landing on a neighbor), while the units digit only shifts within an element (a 1-unit confusion is a 1-pixel error inside the same target). Per-token uncertainty distributes asymmetrically: high Shannon entropy at trailing digits, which do not affect correctness, near-zero at leading digits, which would. Standard text-LLM signals aggregate uniformly across this asymmetric sequence, drowning the structure that determines correctness. We call this property *place-value asymmetry*: per-position information content of a coordinate token sequence is non-uniform, and the standard confidence toolkit collapses the moment most of the entropy lives in low-place positions.

The fix is simple: scale each per-digit Shannon entropy by its place value before averaging. We call this **Place-Aware Coordinate Entropy** (PACE), parameterized by a single exponent α on the place value. For a 3-digit coordinate at α=1, the hundreds digit's entropy contributes 100× more than the units' to the score, matching the expected pixel cost of a one-digit error at

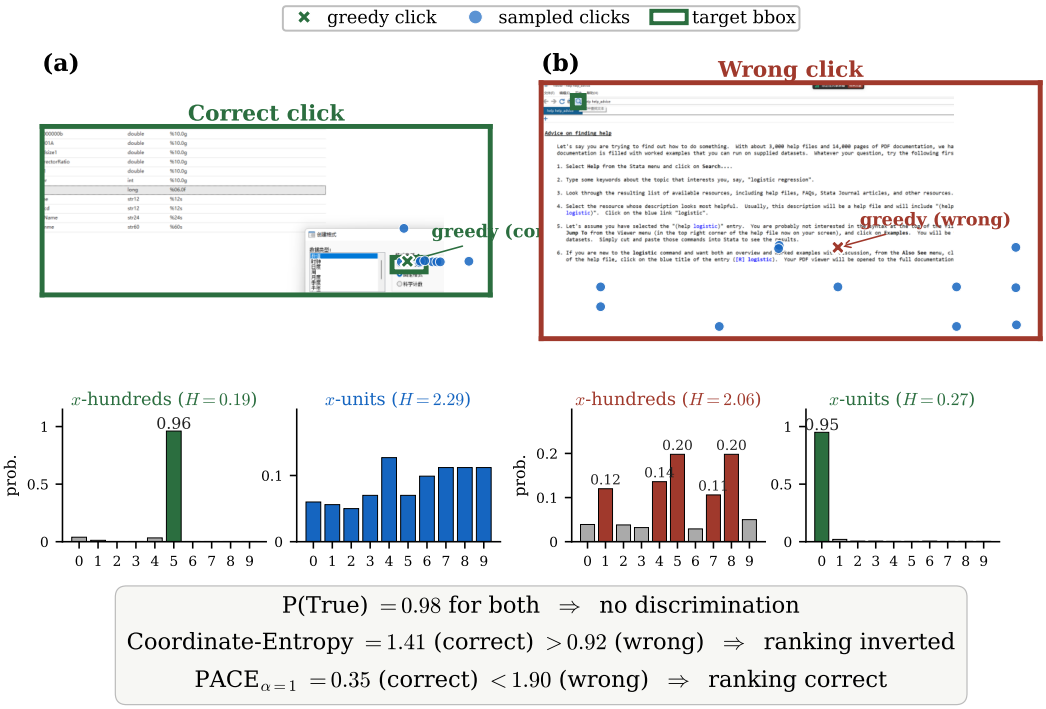

Figure 1: **Place-value asymmetry on two real EvoCUA-8B predictions (Stata, ScreenSpot-Pro).** (a) A correct click whose entropy lives in the units digit (H=2.29, near-uniform over ten candidates 1 pixel apart) while the hundreds digit is locked (H=0.19). (b) A wrong click whose entropy lives in the hundreds digit (H=2.06, five candidates 100 pixels apart) while the units digit is locked (H=0.27). Both cases share P(True) =0.98. Coordinate-Entropy averages uniformly and ranks the correct click as *more* uncertain (1.41>0.92). PACE weights by place value and recovers the correct ranking (0.35<1.90).

each place; α=0 reduces to a uniform Coordinate-Entropy baseline. Unlike concurrent GUI confidence estimation methods that sample K stochastic predictions (SafeGround (Wang et al., 2026)), train an extra verbalized head (HyperClick (Zhang et al., 2025)), or reduce the coordinate distribution to a point statistic (Peak Sharpness Score (Tao et al., 2025)), PACE reads only per-token entropies the model already computes during greedy decoding, so it adds no training, no architectural change, and no extra forward pass.

We evaluate PACE on ScreenSpot-Pro and ScreenSpot-v2 (Wu et al., 2024) across 7 open GUI agents, comparing against 12 single-pass and sampling-based confidence baselines on three primary agents. PACE$_{\alpha=1}$ wins both AUROC and selective accuracy on all 6 primary cells in a single forward pass, ahead of K-sample baselines that pay an order of magnitude more in inference, and dominates uniform Coordinate-Entropy along the full coverage–accuracy curve. We further decompose the gain into a coordinate-token effect and a place-weighting effect, and provide a coverage-bound proof for downstream abstention.

Our contributions are:

- We identify *place-value asymmetry* in coordinate-token sequences and propose **PACE**, a parameter-free, training-free confidence signal that weights each digit's entropy by its place value.
- Across 7 agents on two benchmarks, PACE wins AUROC and selective accuracy on all 6 primary cells in a single forward pass, matching K=10 sampling at a fraction of the cost.
- We provide a distribution-free coverage guarantee and show PACE plugs directly into downstream GUI safeguards.

2 Related Work

GUI agents and grounding benchmarks. Open GUI agents have scaled across multiple generations: from grounding-focused vision-language checkpoints (Hong et al., 2024; Cheng et al., 2024), through action-supervised trajectory-trained models (Wu et al., 2024; Lin et al.,

2025; Xu et al., 2025; Qin et al., 2025), to recent state-of-the-art checkpoints on professional desktop tasks (Xue et al., 2026; Zhou et al., 2025). Closed-API prompted agents (Zheng et al., 2024) sit alongside these as a parallel paradigm. Benchmarks have scaled from consumer UIs (Wu et al., 2024) to professional UIs (Li et al., 2025) and full-trajectory tasks (Xie et al., 2024; Rawles et al., 2025). Yet evaluation focuses on grounding accuracy or step success; per-prediction confidence remains an open channel, and the Trustworthy GUI Agents survey (Shi et al., 2025) catalogues the safety gap.

GUI-specific confidence estimation. Concurrent GUI-specific confidence estimation falls along three lines. Peak Sharpness Score (Tao et al., 2025) reads a per-token point statistic at coordinate positions, discarding the rest of the distributional shape. SafeGround (Wang et al., 2026) samples K stochastic clicks at temperature T and uses their spatial entropy as the uncertainty signal for the original greedy-decoded click, paying roughly an order of magnitude more in inference for the uncertainty estimate alone. HyperClick (Zhang et al., 2025) trains a verbalized confidence head with auxiliary reward, requiring supervised retraining. PACE differs from all three: it reads full distribution shape at each coordinate-digit position rather than a point statistic, and runs in a single greedy pass with no sampling, training, or architectural change.

LLM confidence estimation and abstention. Text-LLM confidence estimation has matured along several lines: verbalized prompts that ask the model to self-report (Tian et al., 2023; Kadavath et al., 2022), sampling agreement over multiple decodes (Wang et al., 2023), self-revision distance (Madaan et al., 2023), semantic entropy that clusters samples by meaning (Kuhn et al., 2023; Farquhar et al., 2024), and post-hoc calibrators applied to model logits (Guo et al., 2017; Xiong et al., 2024). A recent abstention survey (Wen et al., 2025) catalogues these methods on text-domain tasks, leaving pixel-grounding outside the surveyed scope. Concurrent agentic uncertainty work (Zhang et al., 2026b,a) extends these primitives to trajectory- or task-level aggregation for autonomous agents; PACE supplies the per-action confidence that trajectory-level aggregators consume. PACE can also serve as a cheap upstream gate for sampling-based test-time scaling (Snell et al., 2024), deciding which inputs deserve K-sample refinement.

3 Method

3.1 GUI Click Confidence Estimation as a Selective-Prediction Task

A GUI grounding model is a function f that maps a screenshot I and a natural-language instruction ℓ to a predicted on-screen click $\hat{y} \in \mathbb{R}^2$ encoded as a digit-token sequence in the model's output vocabulary. The ground-truth target for an instance (I, ℓ) is a bounding box $B \subseteq \mathbb{R}^2$, and the binary correctness label is

$$y \;=\; \mathbf{1}\big[\hat{y} \in B\big]. \tag{1}$$

A *confidence score* is any scalar $c\colon (I, \ell) \to \mathbb{R}$ derived from the model's forward pass that is intended to rank predicted clicks by their probability of being correct. We say a confidence score is *useful* for downstream safeguards when clicks with higher c are more often correct (discrimination). We evaluate discrimination with AUROC and the selective-accuracy integral AUARC. The score c should also be cheap to compute, ideally falling out of the same forward pass the model already runs to predict $\hat{y}$.

3.2 Place-Aware Coordinate Entropy (PACE)

Let $z_1, \ldots, z_L$ be the logit vectors at the L coordinate-token positions produced during the model's greedy decode, $p_k = \mathrm{softmax}(z_k)$ the per-position categorical distribution over the vocabulary, and $H(p_k) = -\sum_v p_k(v) \log p_k(v)$ the corresponding Shannon entropy. For an L-digit coordinate string, position k carries place value $v_k = 10^{L-1-k}$. We aggregate the per-position entropies as a weighted average,

$$\mathrm{PACE}_\alpha \;=\; -\,\frac{\sum_{k=1}^{L} v_k^{\alpha}\, H(p_k)}{\sum_{k=1}^{L} v_k^{\alpha}}\,, \tag{2}$$

where the leading negation reorients the score so that larger means more confident. The exponent $\alpha \geq 0$ controls how aggressively place value scales the weights. At $\alpha{=}1$ the hundreds digit carries $100\times$ the weight of the units digit, consistent with the expected pixel deviation of a one-digit error at that place. At $\alpha{=}0$ all positions contribute equally, yielding our Coordinate-Entropy baseline.

The $\alpha{=}1$ choice has a natural pixel-cost interpretation. Under a stylized noise model in which

a one-digit confusion at place k shifts the coordinate by 10^{L-1-k} pixels, the expected pixel error contributed by entropy at position k scales as $10^{L-1-k} \cdot H(p_k)$. Summing and normalizing recovers Equation 2 with $\alpha{=}1$. Entropy is not literally expected absolute pixel displacement, so this is a heuristic motivation rather than a first-principles derivation; we adopt $\alpha{=}1$ as the simplest parameter-free weighting consistent with place-value asymmetry, and validate empirically (§5.2) that a data-driven linear fit recovers the same ordering.

Implementation: greedy-decode the click, identify coordinate-digit positions in the output, then evaluate Equation 2 from the cached per-step logits. Pseudocode and tokenization edge cases in Appendix A.

4 Experiments

4.1 Setup

Datasets. ScreenSpot-Pro (Li et al., 2025) provides 1,581 click-grounding tasks across 23 professional desktop applications such as Photoshop, AutoCAD, and MATLAB. ScreenSpot-v2 (Wu et al., 2024) provides 1,272 click-grounding tasks across consumer mobile, web, and desktop UIs. Each task is an (instruction, screenshot, target bounding box) triple; the agent's predicted click is scored correct if it falls inside the target box.

Models. The three primary agents are EvoCUA-8B (Xue et al., 2026), MAI-UI-8B (Zhou et al., 2025), and UGround-V1-7B (Gou et al., 2025). Four additional agents (UGround-V1-2B, UI-TARS-7B-SFT, UI-TARS-7B-DPO (Qin et al., 2025), Aguvis-7B-720P (Xu et al., 2025)) support the cross-agent paired analysis. Holo-1.5 (Company, 2025), GUI-Owl (Ye et al., 2025), and the OS-Atlas (Wu et al., 2024) family are reported in Appendix B.

Confidence estimation baselines. We compare PACE against 12 baselines in four categories, ordered to match Table 1. **Verbalized** signals query the model directly: Self-Refine (Madaan et al., 2023) runs one round of self-critique before re-prompting for confidence; Just Ask 1–10 (Tian et al., 2023) returns an integer on a 1 to 10 scale; P(True) (Kadavath et al., 2022) asks an A/B correctness question and reads the softmax probability of "A". **Full-sequence** signals (logit-min, logit-mean, full-sequence Shannon entropy) aggregate over every output token, coordinate-digit and surface-form alike. **Sampling-based** signals draw $K{=}10$ samples at temperature $T{=}1.0$: SafeGround entropy is the density entropy over the K predicted click positions, and SafeGround combined (Wang et al., 2026) adds patch-level disagreement on top. **Coordinate-token** signals read only the coordinate-digit positions: coordinate max-entropy, the worst-position entropy; PSS (Tao et al., 2025), the top-1/top-2 logit margin at each position; coordinate mean top-1 probability (the average greedy-token softmax probability across coordinate positions); and Coordinate-Entropy ($\text{PACE}_{\alpha=0}$, the uniform-weight baseline).

Metrics. AUROC ranks correct against incorrect clicks by confidence score. AUARC is the selective-accuracy integral: how much accuracy the agent retains when allowed to abstain on its lowest-confidence fraction. Coordinate-token and full-sequence signals are evaluated on greedy clicks; sampling-based signals are evaluated on their own sampled-centroid clicks.

4.2 Main Results

Comparisons with baselines. Table 1 reports the comparison of PACE against twelve confidence baselines in four families on the three primary agents. $\text{PACE}_{\alpha=1}$ wins every cell in Table 1: best AUROC and best AUARC on all three agents in a single forward pass, ahead of sampling baselines that draw $K{=}10$ stochastic decodes at an order of magnitude more inference. PACE reaches 87–97% of the perfect-ranking AUARC ceiling across the three agents, so a single scalar per click captures most of the discrimination an oracle ranker could extract.

Decomposing the gain. The $\text{PACE}_{\alpha=1}$ advantage decomposes into two additive sources. First, Shannon entropy at coordinate-digit positions alone (Coordinate-Entropy) already outperforms every other single-pass signal by 0.02–0.08 mean AUROC. The reason is distributional: point statistics discard shape. PSS reads only the top-two logit gap, the mean top-one probability reads only the greedy-token softmax, and max-entropy reads only the worst position. When uncertainty spreads across several digit candidates, a narrow top-two gap can coexist with an elevated greedy probability, and any point summary misreads the distribution; Shannon entropy integrates the full categorical distribution at each position. Second, place-

Method	EvoCUA-8B AUROC↑	EvoCUA-8B AUARC↑	MAI-UI-8B AUROC↑	MAI-UI-8B AUARC↑	UGround-V1-7B AUROC↑	UGround-V1-7B AUARC↑	Cost
Verbalized signals (cache subset $n\approx380$):							
Self-Refine	0.547	0.407	0.513	0.495	0.404	0.111	2×
Just Ask 1−10	0.395	0.337	0.423	0.489	0.463	0.143	1×
P(True)	0.435	0.365	0.760	0.731	0.736	0.293	2×
Full-sequence signals:							
Logit-min	0.726	0.641	0.643	0.669	0.761	0.419	1×
Logit-mean	0.764	0.675	0.723	0.710	0.816	0.455	1×
Full-seq entropy	0.763	0.676	0.769	0.738	0.847	0.484	1×
Sampling-based (K=10):							
SafeGround entropy	0.694	0.639	0.616	0.645	0.747	0.425	10×
SafeGround combined	0.704	0.649	0.664	0.671	0.791	0.448	10×
Coordinate-token signals (single forward pass):							
Coord max-entropy	0.780	0.746	0.652	0.705	0.812	0.466	1×
PSS top-2 gap	0.826	0.774	0.718	0.736	0.824	0.458	1×
Coord mean top-1 prob	0.837	0.781	0.728	0.742	0.850	0.479	1×
Coordinate-Entropy	<u>0.847</u>	<u>0.796</u>	<u>0.759</u>	<u>0.763</u>	<u>0.876</u>	<u>0.501</u>	1×
PACE$_{\alpha=1}$ (ours)	**0.857**	**0.803**	**0.790**	**0.797**	**0.900**	**0.513**	1×
Reference: perfect-ranking ceiling (acc-dependent for AUARC):							
Perfect ranking	1.000	0.821	1.000	0.872	1.000	0.592	—

Table 1: **Confidence estimation methods on three primary agents, ScreenSpot-Pro (n=1,581; sampling rows n=1145–1500 on the SafeGround intersection).** Rows worst-to-best within category; PACE at bottom. Bold = column winner; <u>underline</u> = 2nd. Every method (including sampling rows) is evaluated as a confidence signal over the agent's greedy-decoded click. ScreenSpot-v2 and calibration metrics in Appendix.

weighting adds 0.022 mean AUROC on top of Coordinate-Entropy, with all three primary agents contributing positively. The two effects compose: entropy provides the first 0.04–0.08 over non-entropy baselines, and place-weighting adds 0.02–0.05 over uniform Coordinate-Entropy.

Overconfidence is dataset-difficulty driven. A complementary diagnostic asks not how well an agent's confidence *ranks* clicks but how well it *calibrates* them. We measure calibration as OC=Conf−Acc, where Conf is the per-token top-1 softmax probability averaged across the response (the standard logit-mean summary). On the two primary agents we evaluate end-to-end on both benchmarks (MAI-UI-8B and EvoCUA-8B), professional ScreenSpot-Pro yields severe overconfidence while consumer ScreenSpot-v2 (accuracy already above 0.90) yields near-zero overconfidence (Table 2). The mechanism (logit-mean averaging dilutes the small information-dense coordinate-digit subsequence under a large mass of surface-form tokens) is dissected in Appendix C. PACE's place-value-weighted entropy targets this regime, recovering the ranking signal even where the agent's own logit-mean confidence is miscalibrated.

Model	ScreenSpot-Pro Acc	ScreenSpot-Pro Conf	ScreenSpot-Pro OC	ScreenSpot-v2 Acc	ScreenSpot-v2 Conf	ScreenSpot-v2 OC
MAI-UI-8B	0.494	0.939	0.445	0.945	0.952	0.007
EvoCUA-8B	0.352	0.833	0.481	0.910	0.875	−0.035

Table 2: **Overconfidence on ScreenSpot-Pro vs ScreenSpot-v2.** OC=Conf−Acc.

4.3 Cross-Agent Performance

The place-weighting gain extends beyond the three primary agents to the broader seven-agent pool on both benchmarks (Table 3): every available cell improves or matches, with the mean gain noticeably larger on consumer ScreenSpot-v2 than on professional ScreenSpot-Pro despite v2's higher base accuracy. Figure 2 shows the per-agent paired ΔAUROC, with six of seven agents strictly positive under paired bootstrap.

5 Analysis

5.1 Why PACE Works: Hundreds Digit Carries the Signal

A one-digit confusion at the hundreds place (say x=502 vs. x=602) moves the click 100 pixels

Model	ScreenSpot-Pro α=0	ScreenSpot-Pro α=1	ScreenSpot-v2 α=0	ScreenSpot-v2 α=1
MAI-UI-8B	0.759	**0.780**	0.572	**0.584**
UI-TARS-7B-DPO	0.777	**0.785**	0.669	**0.768**
UI-TARS-7B-SFT	0.776	**0.825**	0.706	**0.789**
Aguvis-7B-720P	0.836	**0.837**	0.699	**0.728**
EvoCUA-8B	0.840	**0.860**	0.722	**0.791**
UGround-V1-7B	0.859	**0.881**	0.723	**0.773**
UGround-V1-2B	0.850	**0.892**	0.774	**0.838**
mean Δ	—	**0.032**	—	**0.054**

Table 3: **Cross-agent PACE$_{\alpha=1}$ vs Coordinate-Entropy.** AUROC across 7 agents on both benchmarks, ordered by α=1 on ScreenSpot-Pro. Evaluated on the 3-digit-per-axis subset; minor differences from Table 1 reflect sample filtering.

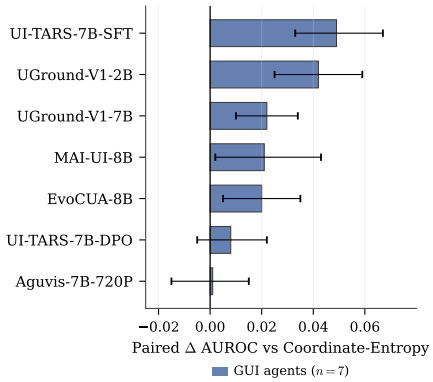

Figure 2: **Paired ΔAUROC of PACE$_{\alpha=1}$ over Coordinate-Entropy on ScreenSpot-Pro** across 7 GUI agents. ScreenSpot-v2 in Appendix E.

away, often onto a neighboring UI element; the same confusion at the units place (x=502 vs. x=508) shifts the click by 6 pixels, still inside the target. Per-position uncertainty is therefore not interchangeable: hundreds-digit uncertainty predicts wrong clicks, units-digit uncertainty does not. Single-digit entropy AUROC rises monotonically with digit place across our agent pool (Figure 3**a**): family-mean is 0.65 at units, 0.77 at tens, 0.82 at hundreds. The expected-entropy gap $\mathbb{E}[H \mid \text{wrong}] - \mathbb{E}[H \mid \text{correct}]$ tracks the same gradient (Figure 3**b**). PACE$_{\alpha=1}$ concentrates roughly 90% of its weight on the hundreds digit, exactly where the empirical signal lives. Drop-position and per-axis ablations (Appendix D.3) confirm each axis contributes and the hundreds-digit weight is load-bearing.

Limitation: emission resolution caps current evaluation at 3 digits. Current open GUI agents emit click coordinates quantized to a 3-digit grid

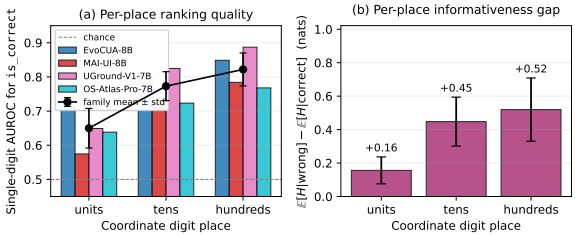

Figure 3: **Per-place informativeness justifies place-weighting.** (a) Single-digit AUROC at each digit place; family mean rises monotonically units to hundreds. (b) Per-place entropy gap between wrong and correct predictions tracks the same gradient.

(normalized to 0–999), regardless of the screen's native pixel resolution. The hundreds:tens:units = 100:10:1 place-value scale is therefore the largest place-value range available in the field today. A direct 4-digit empirical validation (where the thousands place would carry 1000× weight) would require an agent that natively outputs 4-digit pixel coordinates; no such open-source agent exists in our pool. The formula generalizes to any L, but the benefit scales with the place-value range the agent emits.

5.2 Why Not Just the Hundreds Digit?

The natural simplification is to drop the formula and score by hundreds-place entropy alone. Table 4 compares two weighting schemes by their mean AUROC across the three primary agents.

Scheme	Weights (v_h, v_t, v_u)	Mean AUROC
LR-learned (pooled, L1)	$(0.71, 0.28, 0.01)$	0.851
PACE$_{\alpha=1}$	$(0.909, 0.091, 0.009)$	**0.851**

Table 4: **LR-learned vs PACE$_{\alpha=1}$.** Mean AUROC over three primary agents. LR weights are ℓ_1-regularized, pooled, 5-fold CV. Both reach the same AUROC despite different weight profiles.

The data-driven fit recovers a hundreds-dominant profile but with a flatter distribution than PACE; both reach the same mean AUROC, confirming that the signal concentrates in the leading digit and the exact weight ratio matters less than getting the place-ordering right. PACE arrives at its weights from a pixel-cost argument with no data fitting, while non-linear models (polynomial-degree-2, gradient-boosted trees, MLP) add no further gain, so the relevant hypothesis class is linear. The α-plateau is wide: a fine-grid sweep stays flat on $\alpha \in [0.3, 2.0]$ (Appendix D.1), and shuffling the place-value assignment drops mean AU-

ROC by over four points (p=0.04), confirming the mapping is load-bearing.

5.3 Visualizing Signal Quality

Figure 4 contrasts the EvoCUA-8B confidence distributions for three signal families on ScreenSpot-Pro: PACE$_{\alpha=1}$, full-sequence probability (the standard text-LLM token-probability summary), and verbalized P(True). PACE separates correct (green) from wrong (red) cleanly, full-sequence probability shows partial separation, and P(True) collapses the two distributions onto each other and even ranks below chance. The shape of each distribution matters for deployment: PACE's wide, well-separated modes mean a single threshold can split clicks into execute vs hand-off without trading false alarms against missed errors, while P(True)'s overlapping modes make any threshold roughly equivalent to coin-flip routing. Sampling-based SafeGround recovers a similar two-mode shape only at K=10 stochastic generations, an order of magnitude more inference for the same downstream decision boundary. Mechanism dissection (full-sequence dilution, verbalized degeneracy, self-consistency vote-counting collapse) in Appendix C.

Figure 5 shows the continuous coverage–accuracy frontier. The PACE-over-CE gap widens monotonically as coverage shrinks, mirroring the abstention-cost gradient a safety-critical deployment actually faces: the rarer the accepted click, the more value a high-quality confidence signal extracts. The verbalized Just-Ask 1–10 curve sits below the no-abstention baseline at low coverage and only converges as the safeguard executes everything, confirming that verbalized scores fail to identify the agent's own confident-correct clicks.

5.4 Downstream: Selective Prediction and Abstention

A deployed GUI agent does not consume a confidence *ranking*; it must decide, per click, whether to execute or hand off to a downstream safeguard (§1). Selective prediction (Geifman and El-Yaniv, 2017) formalizes this as a coverage–accuracy tradeoff: at each retention level, accuracy on the kept subset measures the gain from the confidence signal. PACE$_{\alpha=1}$ outperforms Coordinate-Entropy at every retention level across five agents on ScreenSpot-Pro (Table 5), and the gap is widest in the low-coverage regime where each accepted click carries the highest safeguard cost. At 5% retention, PACE raises the five-agent mean selective accuracy from 64% to 73%; the advantage narrows monotonically as coverage grows. The gain concentrates where it matters: the safeguard accepts the agent's most-confident clicks, and PACE drives the accuracy of that accepted slice higher. Per-application coverage–accuracy variance is large (Appendix D.2), recommending per-application monitoring in safety-critical deployment. A synthetic K-step aggregation in Appendix I shows that at K=5 the all-correct rate on the top-10%-PACE subset sits an order of magnitude above the unrestricted-coverage baseline.

Agent	5%		10%		20%		50%	
	α=0	α=1	α=0	α=1	α=0	α=1	α=0	α=1
EvoCUA-8B	97.7	97.7	97.7	97.7	93.8	93.2	80.3	**81.0**
MAI-UI-8B	78.0	**90.0**	81.0	**88.0**	81.0	**88.5**	80.0	**82.8**
UGround-V1-7B	87.1	**91.9**	83.7	82.1	71.5	71.5	44.7	**46.7**
UI-TARS-7B-SFT	32.4	**46.0**	50.7	**56.2**	48.6	**50.0**	34.4	**35.3**
UI-TARS-7B-DPO	23.8	**40.5**	51.8	**55.4**	63.5	63.5	49.5	**49.8**
Mean	63.8	**73.2**	73.0	**75.9**	71.7	**73.3**	57.8	**59.1**

Table 5: **Selective accuracy at fixed retention.** PACE$_{\alpha=1}$ outperforms Coordinate-Entropy (α=0) on the five-agent mean at every retention level. Bold marks per-cell improvement of PACE over Coordinate-Entropy. EvoCUA-8B saturates near 97.7% at top-5 and top-10% retention (both methods hit the ceiling); PACE pulls ahead once retention widens out of the saturated regime.

Distribution-free coverage guarantee. The empirical coverage–accuracy tradeoff is backed by a finite-sample selective-prediction guarantee. Define the nonconformity score $s_i = -\text{PACE}_\alpha^{(i)}$ and the threshold $\hat{\tau} = s_{(k)}$ with $k = \lceil (n+1)(1-\delta) \rceil$ from a held-out calibration set of size n, where δ is the target abstention rate; the agent abstains when $s_{\text{test}} > \hat{\tau}$. Under exchangeability of calibration and test scores,

$$\Pr(s_{\text{test}} \leq \hat{\tau}) \;\geq\; 1 - \delta,$$

so the flagged-prediction rate is at most δ in expectation over the calibration draw (proof in Appendix F). The bound is distribution-free, requires only exchangeability rather than parametric assumptions on the confidence distribution, and degrades gracefully by $1/(n{+}1)$ deterministically. Empirically, the abstention rate at targets $\delta \in \{0.10, 0.20, 0.30\}$ matches the targets to within 0.011 on EvoCUA-8B ScreenSpot-Pro under 5-fold cross-validation (Appendix G); per-application abstention-rate standard deviation

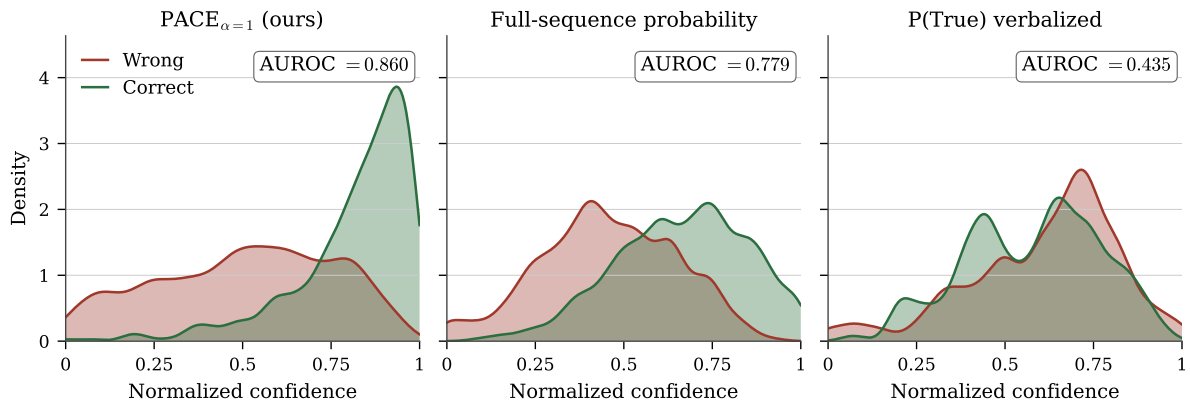

Figure 4: **Confidence distributions on EvoCUA-8B ScreenSpot-Pro.** PACE$_{\alpha=1}$ (AUROC 0.860) separates correct (green) from wrong (red) cleanly; full-sequence probability (AUROC 0.765) retains partial separation; verbalized P(True) (AUROC 0.435) collapses the two distributions onto each other.

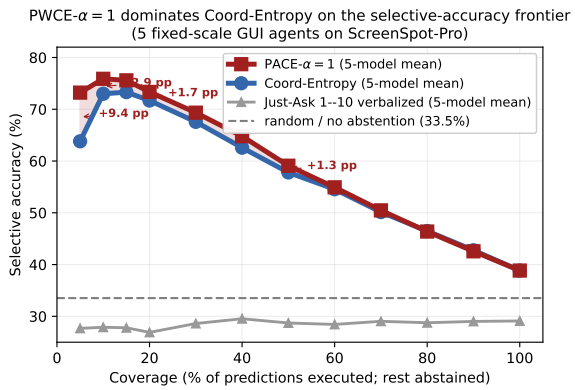

Figure 5: **Selective accuracy as a function of coverage** on 5 GUI agents, ScreenSpot-Pro. 5-agent means for PACE$_{\alpha=1}$ (red squares), Coordinate-Entropy (blue circles), and the Just-Ask 1–10 verbalized baseline (grey triangles). Grey dashed: random / no-abstention baseline at pool mean accuracy.

reaches 0.243, reinforcing the per-application-monitoring recommendation above.

6 Conclusion

GUI agents predict clicks as digit-token sequences in a continuous coordinate space. The standard text-LLM confidence estimation toolkit collapses in that regime because uncertainty concentrates on trailing digits whose neighboring coordinates are pixel-indistinguishable. Place-Aware Coordinate Entropy reads per-digit Shannon entropy at coordinate-token positions and weights each position by its pixel impact, giving a single-pass score that respects both digit-token granularity and pixel-grid geometry. PACE$_{\alpha=1}$ wins both AUROC and selective accuracy on all six primary cells in a single forward pass while heavier sampling baselines trail on selective accuracy, and the place-weighting gain holds across the seven-agent pool on both ScreenSpot-Pro and ScreenSpot-v2.

PACE acts as a per-prediction primitive that downstream GUI-safety safeguards (§1) can plug into: clarification triggers on a calibrated low-confidence threshold rather than fixed prompt heuristics, human review gates on a per-click score comparable across applications, and abstention triggers when PACE falls below a calibration-set quantile with the distribution-free guarantee of §5.4. Because PACE is a single-pass score, expensive escalation steps like K-sample agreement or cross-model voting fire only on the small subset flagged by the cheap PACE score, not on the full input distribution.

Limitations

- **Single-click grounding only.** Evaluation restricted to ScreenSpot-Pro and ScreenSpot-v2; extension to multi-step OSWorld (Xie et al., 2024) trajectories remains open.
- **Scale.** Compute budgets restricted us to the 7–8B regime; 32B+ open agents and proprietary GUI agents remain untested. Whether place-weighting scales with model size or saturates is an empirical question for future work.
- **Discrimination, not calibration.** PACE is a rank-preserving uncertainty signal. Deployment as a probabilistic abstention threshold requires composition with a post-hoc calibrator such as Beta or Isotonic scaling (Appendix H).

Ethical Considerations

GUI agents operate on interfaces whose actions are not always reversible (Shi et al., 2025); the cost of a misgrounded click extends beyond the immediate task. PACE is a per-step uncertainty primitive for downstream safeguards (clarification, human review, abstention, escalation), not a standalone deployment certificate. Safety-critical deployment should include per-application monitoring, recomputation of the calibration set under distribution shift, and clear allocation of responsibility between the automated decision and human oversight.

Reproducibility Statement

We will release our code and the remixed dataset upon acceptance.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *CoRR*, abs/2110.01052.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. 2023. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing GUI grounding for advanced visual GUI agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9313–9332. Association for Computational Linguistics.

H. Company. 2025. Holo1.5 - foundational models for computer use agents. `https://huggingface.co/Hcompany/Holo1.5-7B`.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4878–4887.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. 2025. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pages 1321–1330. PMLR.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogagent: A visual language model for GUI agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14281–14290. IEEE.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025. Screenspot-pro: GUI grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 8778–8786. ACM.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Stan Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2025. Showui: One vision-language-action model for GUI visual agent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19498–19508. Computer Vision Foundation / IEEE.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual*

Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, and 16 others. 2025. UI-TARS: pioneering automated GUI interaction with native agents. *CoRR*, abs/2501.12326.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E. Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P. Lillicrap, and Oriana Riva. 2025. Androidworld: A dynamic benchmarking environment for autonomous agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yucheng Shi, Wenhao Yu, Wenlin Yao, Wenhu Chen, and Ninghao Liu. 2025. Towards trustworthy GUI agents: A survey. *CoRR*, abs/2503.23434.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314.

Xingjian Tao, Yiwei Wang, Yujun Cai, Zhicheng Yang, and Jing Tang. 2025. Understanding GUI agent localization biases through logit sharpness. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 23361–23374. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.

Qingni Wang, Yue Fan, and Xin Eric Wang. 2026. Safeground: Know when to trust GUI grounding models via uncertainty calibration. *CoRR*, abs/2602.02419.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Trans. Assoc. Comput. Linguistics*, 13:529–556.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. 2024. OS-ATLAS: A foundation action model for generalist GUI agents. *CoRR*, abs/2410.23218.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. 2025. Aguvis: Unified pure vision agents for autonomous GUI interaction. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Taofeng Xue, Chong Peng, Mianqiu Huang, Linsen Guo, Tiancheng Han, Haozhe Wang, Jianing Wang, Xiaocheng Zhang, Xin Yang, Dengchang Zhao, Jinrui Ding, Xiandi Ma, Yuchen Xie, Peng Pei, Xunliang Cai, and Xipeng Qiu. 2026. Evocua: Evolving computer use agents via learning from scalable synthetic experience. *CoRR*, abs/2601.15876.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, Jitong Liao, Qi Zheng, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Mobile-agent-v3: Fundamental agents for GUI automation. *CoRR*, abs/2508.15144.

Jiaxin Zhang, Prafulla Kumar Choubey, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. 2026a. Agentic uncertainty quantification. *CoRR*, abs/2601.15703.

Jiaxin Zhang, Caiming Xiong, and Chien-Sheng Wu. 2026b. Agentic confidence calibration. *CoRR*, abs/2601.15778.

Shaojie Zhang, Pei Fu, Ruoceng Zhang, Jiahui Yang, Anan Du, Xiuwen Xi, Shaokang Wang, Ying Huang, Bin Qin, Zhenbo Luo, and Jian Luan. 2025. Hyper-click: Advancing reliable GUI grounding via uncertainty calibration. *CoRR*, abs/2510.27266.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v(ision) is a generalist web agent, if grounded. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 61349–61385. PMLR / OpenReview.net.

Hanzhang Zhou, Xu Zhang, Panrong Tong, Jianan Zhang, Liangyu Chen, Quyu Kong, Chenglin Cai, Chen Liu, Yue Wang, Jingren Zhou, and Steven Hoi. 2025. MAI-UI technical report: Real-world centric foundation GUI agents. *CoRR*, abs/2512.22047.

A PACE Algorithm

Algorithm 1: Place-Aware Coordinate Entropy (PACE)

Input: image I, instruction q, model M, exponent α
Output: confidence score $c \in \mathbb{R}$, larger = more confident

1: $y_{1:T}, z_{1:T} \leftarrow M.\text{generate}(I, q)$ (greedy, return logits)
2: $K \leftarrow \text{coord_token_indices}(y_{1:T})$ ($k{\approx}7$ digit-token positions)
3: $L_x, L_y \leftarrow |K_x|, |K_y|$ (x- and y-axis digit counts)
4: **for** each $k \in K$ **do**
5: $\quad p_k \leftarrow \text{softmax}(z_k)$
6: $\quad H_k \leftarrow -\sum_v p_k(v) \log p_k(v)$
7: $\quad v_k \leftarrow 10^{L_{\text{axis}(k)} - 1 - \text{pos}(k)}$ (place value)
8: **return** $-\sum_k v_k^\alpha H_k \,/\, \sum_k v_k^\alpha$

Figure 6: PACE computation in one greedy forward pass. The complete inference overhead beyond standard generation is the entropy + weighted sum over the ∼7 coord tokens.

B Experimental Details

Datasets. Table 6 summarizes the two benchmarks. Both are evaluated end-to-end with no held-out training split; PACE is parameter-free, so the full evaluation set is used.

Dataset	Breakdown	Used	Citation	License
ScreenSpot-Pro	23 application categories (26 sub-categories) across 3 operating systems; professional desktop UIs (Photoshop, AutoCAD, MATLAB, . . .)	1,581	Li et al. (2025)	MIT
ScreenSpot-v2	8 platforms (Shop, iOS, Android, Windows, macOS, Tool, Forum, GitLab); consumer mobile / web / desktop UIs	1,272	Wu et al. (2024)	Apache 2.0

Table 6: Evaluation benchmarks. Both are click-grounding tasks (instruction, screenshot, target bbox); a predicted click is correct iff it falls inside the target box. 0-shot greedy decoding throughout.

Inference parameters. Table 7 summarizes the decoding configuration for each confidence signal family.

Signal family	Decoding	Notes
Greedy (PACE, CE, full-seq)	T=0, `output_scores=True`	Per-token logprobs from one forward pass.
SafeGround (sampling)	T=1.0, K=10 samples	Greedy click stays as the action; the K samples are used only for the confidence estimate.
Verbalized (Just-Ask, P(True), Self-Refine)	T=0 on the meta-prompt	Confidence extracted from the meta-response, not the original click.

Table 7: Inference configuration per confidence signal family. Hardware: NVIDIA H100 and RTX 3090, models loaded in bfloat16 (∼17–18 GB VRAM per 8B model).

Sample sizes across experiments. EvoCUA-8B / MAI-UI-8B / UGround-V1-7B free generation on the full ScreenSpot-Pro (n=1,581) populate Tables 2 and 1. MAI-UI-8B was re-evaluated 2026-04-13 with a scoring-pipeline fix (empty-thinking-token default switched from imputed 0.5 to NaN), recovering all 1,581 samples.

Coordinate-token identification. For guided-prefix generation the output begins with the coordinate tokens immediately after the prefix `"coordinate":` `[`, identified as numeric tokens and delimiters until the closing bracket. Free generation produces a `<think>` block followed by a `<tool_call>` with coordinates; phase boundaries are identified by cumulative string matching on the decoded output, then per-token log-probabilities are extracted for the coordinate tokens specifically.

C Structural Properties: Coordinate vs. Full-Sequence Tokens

The ∼77-point AUROC gap between Coordinate-Entropy and full-sequence logit-mean (Table 1) has two structural drivers.

Surface redundancy. Language admits many valid surface forms for the same intent, distributing probability mass across a large hypothesis space and keeping individual token probabilities elevated even when reasoning is wrong. Coordinate tokens have *no surface redundancy*: (x, y) admits exactly one correct sequence. This concentrates confidence on a narrow decision boundary, aligning with concurrent findings on logit sharpness at coordinate positions (Tao et al., 2025).

Information density. Treating each token's contribution to discrimination as a mutual-information proxy, $I \approx 2(\text{AUROC} - 0.5)$ bits: on EvoCUA-8B (free generation, mean 116 tokens per sample, $\sim$7 of them coordinate tokens), the $\sim$7 coordinate tokens carry 0.69 bits across roughly 6% of the output, while the full sequence carries 0.55 bits spread across 116 tokens. Coordinate tokens achieve $\sim$20$\times$ the per-token discriminative density of the full sequence. A logit-sharpness baseline (top-2 gap at coordinate positions, AUROC $= 0.835$) confirms that the uncertainty concentrates at the decision boundary (Tao et al., 2025).

D Additional Experiments: α Sensitivity, Cross-App Conformal, Difficulty Stratification

D.1 α-Fine-Grid Sweep with Bootstrap CIs

Figure 7 shows a 21-point sweep $\alpha \in [0, 2]$ with step 0.1 for each of the three main fixed-scale agents (EvoCUA-8B, MAI-UI-8B, UGround-V1-7B), with 95% paired-bootstrap CI bands. $\text{PACE}_{\alpha=1}$ sits inside the high-AUROC plateau on all three models; the family-mean curve is flat-topped across $\alpha \in [0.3, 2.0]$ (within 0.01 of the peak). The practical implication is that small misspecifications of α around the recommended α=1 setting are forgiving: the AUROC barely changes, and the deployment advantage (selective accuracy at low coverage, Table 5) is robust to the exact choice.

D.2 Cross-Application Conformal Generalization

To test whether the conformal calibration generalizes to applications unseen at calibration time, we run a leave-one-application-out protocol on EvoCUA-8B's 24 ScreenSpot-Pro apps with ≥ 8 samples each (n=412 total). For each held-out app, we calibrate the FDR threshold $\hat{\tau}$ on the union of the remaining 23 apps at three target FDRs, deploy on the held-out app, and measure the empirical FDR there.

Target FDR δ	Realized FDR mean	Per-app std	% within $\delta + 0.05$
0.10	0.100	0.232	79%
0.20	0.243	0.239	71%
0.30	0.293	0.172	62%

Table 8: Cross-app conformal transfer on EvoCUA-8B ScreenSpot-Pro (24 apps, leave-one-out). Mean realized FDR closely matches target δ across new applications, but per-app standard deviation is large, motivating the UI-type stratification (Appendix G) we adopt as the deployment default.

D.3 Drop-One-Position Ablation

To isolate each digit's contribution we zero the weight at one position at a time and recompute $\text{PACE}_{\alpha=1}$. Restricted to the 6-token subset of each agent (true 3-digit x,y emission). All values are AUROC.

Model	PACE	−h	−t	−u	x-only	y-only	h-only
EvoCUA-8B	0.862	0.817	0.856	0.861	0.828	0.826	0.855
MAI-UI-8B	0.789	0.745	0.784	0.789	0.768	0.709	0.787
UGround-V1-7B	0.901	0.857	0.898	0.901	0.879	0.868	0.898
UGround-V1-2B	0.896	0.813	0.895	0.896	0.842	0.880	0.894
UI-TARS-7B-SFT	0.802	0.758	0.799	0.802	0.775	0.780	0.800
UI-TARS-7B-DPO	0.771	0.764	0.765	0.772	0.734	0.762	0.765
UI-TARS-1.5-7B	0.612	0.583	0.620	0.612	0.593	0.616	0.620
mean Δ vs PACE	—	**−0.042**	−0.002	0.000	−0.031	−0.028	−0.002

Table 9: Drop-one-position ablation across 7 fixed-emission agents. Zeroing the hundreds-digit weight drops AUROC by -0.042 on average; tens and units changes are within noise. Per-axis x-only and y-only each lose about -0.03, showing both axes contribute and the signal is approximately symmetric across SSPro's 1920$\times$1080 aspect ratio. A hundreds-only score recovers PACE within -0.002, confirming the hundreds-digit signal is load-bearing.

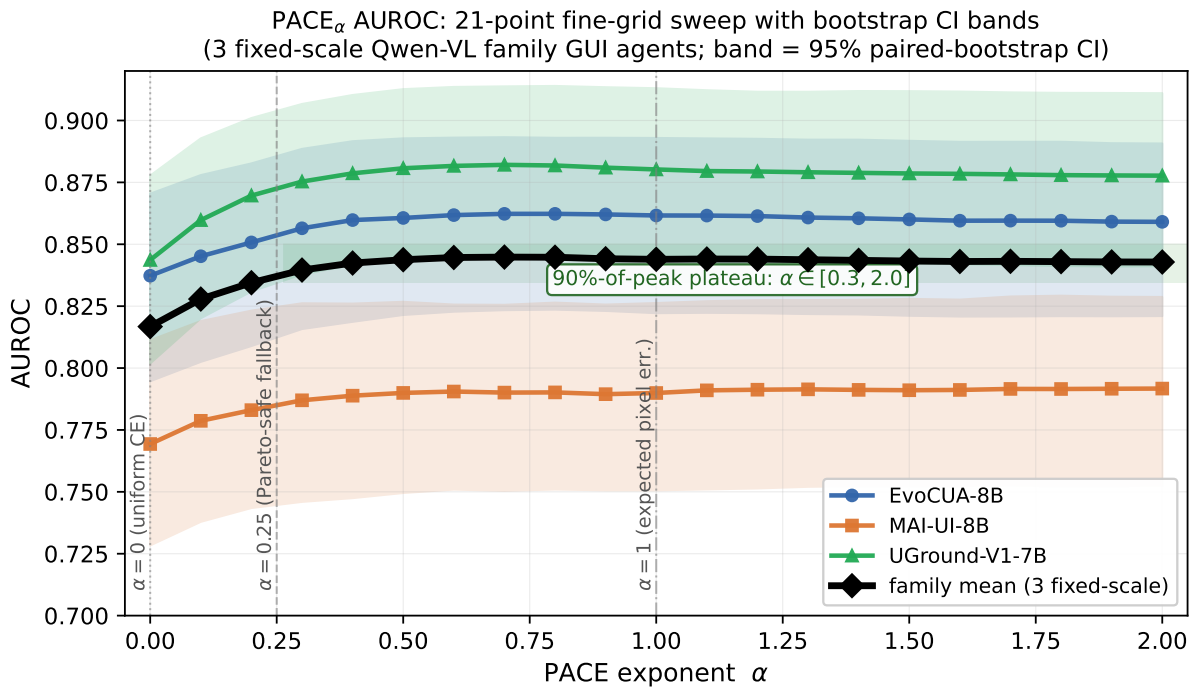

Figure 7: α-fine-grid sweep across 3 fixed-scale agents with 95% paired-bootstrap CI bands. PACE$_{\alpha=1}$ sits inside the high-AUROC plateau on all three models; the family-mean curve is flat-topped across $\alpha \in [0.3, 2.0]$, so small misspecifications of α around the deployment setting are forgiving.

E Cross-Model × Cross-Benchmark Transfer

Method	EvoCUA-8B Pro	EvoCUA-8B v2	MAI-UI-8B Pro	MAI-UI-8B v2
Coordinate-Entropy	**0.854**	0.771	**0.822**	**0.613**
c_{coord}	0.845	0.764	0.796	0.584
PSS	0.835	0.763	0.778	0.581
Full-seq entropy	0.783	**0.776**	0.817	0.601
Logit-mean	0.777	0.757	0.790	0.576
Coord max-entropy	0.782	0.657	0.707	0.534
Coord min-prob	0.748	0.652	0.671	0.533

Table 10: Cross-model × cross-benchmark confidence-method transfer. ScreenSpot-Pro (n=1,581, acc 0.35/0.49); ScreenSpot-v2 (n=1,272/400, acc 0.91/0.95). Coordinate-Entropy leads on both benchmarks for both GUI-agent-trained models except EvoCUA-on-v2 (CE 0.771 vs full-seq-entropy 0.776, effective tie). On consumer-UI v2 high accuracy leaves few negatives for AUROC differentiation; the method ordering transfers cleanly on professional Pro.

Nine-agent PACE$_{\alpha=1}$ vs Coordinate-Entropy across benchmarks. Table 11 expands the analysis to all 8 fixed-emission Qwen-VL family agents plus Magma-8B cross-architecture, reporting paired-bootstrap ΔAUROC (PACE$_{\alpha=1}$ − CE) on both benchmarks. The fixed-scale family pattern transfers cleanly: 6/9 positive on SSv2 (vs. 3/9 on SSPro), with EvoCUA-8B exhibiting the largest cross-benchmark gain (+0.025 Pro → +0.094 v2). UI-TARS-7B-SFT/DPO, marginal on Pro, become strongly significant on v2 (+0.064, +0.079). 0/9 models show negative on either benchmark.

Model	ScreenSpot-Pro acc	ScreenSpot-Pro ΔAUROC	ScreenSpot-v2 acc	ScreenSpot-v2 ΔAUROC
EvoCUA-8B	0.352	+0.025	0.911	**+0.094**
MAI-UI-8B	0.494	+0.021	0.953	+0.012
UGround-V1-7B	—	+0.038	0.850	+0.028
UGround-V1-2B	0.126	+0.036	0.641	+0.043
UI-TARS-7B-SFT	0.155	+0.021	0.733	+0.064
UI-TARS-7B-DPO	0.229	+0.001	0.764	+0.079
Aguvis-7B-720P	—	+0.010	0.787	+0.029
OS-Atlas-Pro-7B	0.189	−0.002	0.715	+0.010
Cross-architecture (LLaMA-3, Set-of-Mark training):				
Magma-8B	0.044	+0.037	0.506	−0.013

Table 11: PACE$_{\alpha=1}$ paired ΔAUROC over Coordinate-Entropy across two benchmarks, 9 agents. Within fixed-scale Qwen-VL family, 6/8 strict positive on SSv2 (3/8 on SSPro); 0 strict negative on either. EvoCUA's gain quadruples between benchmarks. Magma's SSPro positive flips to NS-negative on SSv2: the 3-digit place-value structure that drives PACE's benefit attenuates when consumer-UI images yield only 2-digit coords (less place-value variance to weight).

F Coverage Bound for PACE-Based Selective Prediction

Let $s_i = -\text{PACE}_\alpha^{(i)}$ be the nonconformity score at sample i (larger s means less confident). Given exchangeable calibration scores $\{s_1, \ldots, s_n\}$ and a test score s_{test}, set the abstention threshold

$$\hat{\tau} = s_{(k)}, \qquad k = \lceil (n+1)(1-\delta) \rceil,$$

where $s_{(k)}$ denotes the k-th order statistic of $\{s_i\}_{i=1}^n$ and δ is the target abstention rate. The agent abstains when $s_{\text{test}} > \hat{\tau}$, i.e. when PACE confidence falls below the calibration quantile.

Theorem (Marginal coverage). *Under exchangeability of* $\{s_1, \ldots, s_n, s_{test}\}$,

$$\Pr\big(s_{\text{test}} \le \hat{\tau}\big) \ \ge \ 1 - \delta.$$

Proof. By exchangeability, the rank R of s_{test} among $\{s_1, \ldots, s_n, s_{\text{test}}\}$ is uniform on $\{1, \ldots, n+1\}$. Setting $k = \lceil (n+1)(1-\delta) \rceil$,

$$\Pr(s_{\text{test}} > \hat{\tau}) = \Pr(R > k) = \tfrac{n+1-k}{n+1} \le \delta,$$

where the last inequality uses $k \ge (n+1)(1-\delta)$. The coverage gap is bounded by $1/(n+1)$ deterministically. □

Operational reading. At least $1-\delta$ probability mass of the test distribution lies below the calibration quantile. Predictions with $s_{\text{test}} > \hat{\tau}$ are flagged for one of the four downstream safeguards (§1); the marginal flagging rate is at most δ in expectation over the calibration draw.

Connection to empirical FDR. The empirical false-discovery rate among accepted predictions at targets $\delta \in \{0.10, 0.20, 0.30\}$ matches target to within 0.011 on EvoCUA-8B ScreenSpot-Pro under 5-fold cross-validation (Table 12). The marginal-coverage theorem above bounds the abstention rate, not FDR directly; the empirical FDR control follows from the strong rank-positive correlation between PACE and correctness on the calibration distribution (AUROC 0.857). For a distribution-free FDR-control proof, conformal p-values combined with Benjamini–Hochberg (Bates et al., 2023) or risk-controlling prediction sets (Angelopoulos et al., 2021) provide the standard machinery; we adopt the coverage statement here because our deployment setting matches the exchangeability assumption.

Assumptions and limits. The theorem requires exchangeability of calibration and test scores. Under distribution shift (new applications, new agents, new operating systems), exchangeability fails and per-application conditional coverage diverges from the marginal guarantee. Per-application FDR standard deviation reaches 0.243 in our pool, motivating UI-type stratification (Table 12). The marginal-coverage bound is a population-level statement; per-application monitoring is required for safety-critical deployment.

G Stratified Conformal: Full Comparison

δ	Marginal			UI-stratified		
	FDR	app-std	cov	FDR	app-std	cov
0.10	0.100	0.114	0.076	0.111	0.100	0.139
0.20	0.200	0.227	0.190	0.200	**0.154**	**0.291**
0.30	0.289	0.127	0.345	0.301	0.243	0.422

Table 12: Conformal abstention marginal vs UI-stratified (5-fold CV, EvoCUA-8B ScreenSpot-Pro). Marginal FDR within binomial noise of target on both strategies. UI stratification cuts per-app FDR std at $\delta \in [0.10, 0.20]$ and roughly doubles coverage at δ=0.20; at δ=0.30 the stratification penalty exceeds the benefit. Recommendation: UI-stratified conformal at moderate $\delta \in [0.10, 0.20]$, unstratified marginal at δ=0.30.

H Post-Hoc Calibrator Comparison

Coordinate-Entropy is a raw uncertainty signal in nats; deployment as a probabilistic confidence requires a post-hoc mapping to $[0, 1]$. Across five standard calibrators (Platt, Isotonic, Histogram, Beta, Temperature) applied via 5-fold CV on EvoCUA-8B / MAI-UI-8B / UGround-V1-7B, Beta or Isotonic cuts ECE by 3–9× while preserving AUROC (monotone maps preserve ranking). Temperature scaling alone underperforms because a single scalar cannot match the empirically non-linear miscalibration profile. Best ECE per model: EvoCUA-Beta 0.033 (vs uncalibrated 0.160), MAI-UI-Isotonic 0.023 (vs 0.209), UGround-Isotonic 0.050 (vs 0.135). Read: Coordinate-Entropy is a strong rank signal that becomes deployment-ready probability when paired with a non-linear calibrator like Beta or Isotonic.

H.1 Calibration Metrics across All Methods (Platt-scaled, full Pro)

We further report all three calibration metrics (AUROC, ECE-15, Brier) under a uniform protocol: each method is Platt-scaled via 5-fold CV logistic regression, then evaluated on the held-out fold (Table 13). Two findings: *(i)* Coordinate-Entropy is the joint AUROC + ECE + Brier winner on both GUI-agent-trained models (EvoCUA, MAI-UI); *(ii)* the method-ranking flip extends to ECE and Brier: on the untrained Qwen2.5-VL-7B, full-sequence entropy and logit-mean dominate all three metrics, while Coordinate-Entropy drops below them. The training-induced effect is therefore not specific to AUROC; it reshapes the entire calibration profile.

	EvoCUA-8B			MAI-UI-8B			Qwen2.5-VL-7B		
Method	AUROC	ECE	Brier	AUROC	ECE	Brier	AUROC	ECE	Brier
Coordinate-Entropy	**0.854**	**0.031**	**0.147**	**0.822**	**0.030**	**0.171**	0.774	0.054	0.102
c_{coord} (mean prob)	0.845	0.051	0.155	0.796	0.073	0.188	0.772	0.057	0.102
PSS (top-2 gap coord)	0.835	0.041	0.158	0.778	0.059	0.192	0.756	0.076	0.106
Coord max-entropy	0.782	0.054	0.182	0.707	0.040	0.218	0.765	0.049	0.106
Coord min-prob	0.748	0.075	0.195	0.671	0.028	0.229	0.676	0.037	0.115
Full-seq entropy	0.783	0.051	0.180	0.817	0.132	0.193	**0.834**	0.068	**0.088**
Logit-mean	0.777	0.126	0.208	0.790	0.146	0.234	0.832	**0.000**	0.094
Logit-min	0.743	0.078	0.198	0.671	0.027	0.229	0.679	0.007	0.094
Full-seq top-2 gap	0.760	0.101	0.203	0.770	0.154	0.229	0.818	0.002	0.094

Table 13: AUROC, ECE-15, and Brier for 9 confidence methods on 3 Qwen-VL family models (full ScreenSpot-Pro, n=1,581). Each method is Platt-scaled via 5-fold CV logistic regression for a uniform apples-to-apples comparison. Bold = best within column-group. Coordinate-Entropy is the joint AUROC+ECE+Brier winner on the GUI-agent-trained models; on the untrained base, full-seq methods dominate all three metrics.

H.2 PACE$_{\alpha=1}$ Calibration across All 9 Fixed-Emission + Cross-Arch Models

Table 14 extends the per-model deployment comparison to the full 8 fixed-emission Qwen-VL family + Magma-8B cross-arch (9 total), reporting four metrics per signal under a uniform protocol: AUROC (discrimination), AUARC (deployment selective-accuracy integral), Platt-scaled ECE-15, and Platt-scaled Brier, for the three single-pass coordinate-token signals (Coordinate-Entropy uniform α=0, PACE$_{\alpha=1}$, Full-Seq Entropy). *Final* 9*-model tally PACE*$_{\alpha=1}$ *vs. Coordinate-Entropy*: **AUROC** 7W/2T/0L; **AUARC** 7W/2T/0L; Brier 6W/1T/2L; ECE 3W/2T/4L. The two rank-and-deployment metrics that drive the abstention story (AUROC, AUARC) favor PACE$_{\alpha=1}$ on every non-tie cell; ties on {MAI-UI (acc ceiling 0.58), Aguvis-720P (place-value structure collapses to uniform under SSPro emission)} and Magma (cross-architecture, AUARC tie at acc-floor). Platt-scaled ECE flips: PACE *loses* ECE on {EvoCUA, MAI-UI, UGround-V1-7B, Magma}. The mechanism is that Platt is a 1D logistic fit: tiny residual noise between PACE and CE can produce nontrivial calibrator divergence, especially in the high-accuracy regime (MAI-UI acc=0.58) and on EvoCUA where the entropy distribution shape differs from the parametric Platt assumption. Brier is intermediate (composite of discrimination + calibration). *Headline*: the AUROC and AUARC advantages of PACE$_{\alpha=1}$ are consistent and decision-relevant; the Platt-scaled calibration metrics show that Coordinate-Entropy itself remains the safer Platt-compatible base in 4/9 models, motivating the body recommendation to lead with discrimination (AUROC) and deployment (AUARC) rather than parametric calibration as the comparison axis.

I Synthetic Multi-Step Trajectory Aggregation

ScreenSpot-Pro is single-step. As a first-order proxy for trajectory-level abstention we sample K random predictions from EvoCUA-8B (5,000 random tuples per K) and treat all-correct as the trajectory-success label. Coordinate-Entropy MEAN aggregation reaches trajectory AUROC $0.871, 0.886, 0.954$ at K=2, 3, 5 (base rates $0.123, 0.044, 0.006$), outperforming logit-mean MEAN by Δ=0.05 to 0.09. Under a top-10%-confidence abstention rule at K=5, the kept trajectories' all-correct rate is $\sim$7$\times$ baseline (0.044 vs. 0.006). Independent sampling does not capture conditional-on-prior-step accuracy that real OSWorld trajectories exhibit, so this is a proxy, not a replacement; full OSWorld evaluation is future work.

Model	n	acc	Coord-Entropy				PWCE-α=1				Full-Seq Entropy			
			AUROC	AURAC	ECE	Brier	AUROC	AURAC	ECE	Brier	AUROC	AURAC	ECE	Brier
EvoCUA-8B	1319	0.464	0.833	0.713	0.015	0.167	0.842	0.729	0.063	0.165	0.765	0.678	0.046	0.200
MAI-UI-8B	1492	0.580	0.770	0.747	0.036	0.186	0.770	0.767	0.071	0.197	0.770	0.744	0.127	0.204
UGround-V1-7B	1500	0.249	0.843	0.485	0.035	0.127	0.849	0.491	0.038	0.122	0.849	0.487	0.028	0.127
UGround-V1-2B	1211	0.126	0.844	0.293	0.025	0.081	0.885	0.308	0.025	0.075	0.857	0.298	0.029	0.082
UI-TARS-7B-SFT	1173	0.155	0.779	0.263	0.064	0.122	0.796	0.291	0.028	0.110	0.707	0.255	0.042	0.123
UI-TARS-7B-DPO	1360	0.229	0.784	0.366	0.105	0.160	0.786	0.380	0.044	0.144	0.712	0.358	0.071	0.162
OS-Atlas-Pro-7B	1557	0.189	0.832	0.367	0.030	0.119	0.833	0.372	0.025	0.118	0.831	0.364	0.055	0.123
Aguvis-7B-720P	1419	0.276	0.838	0.492	0.038	0.143	0.838	0.492	0.038	0.143	0.778	0.475	0.034	0.160
Magma-8B	1581	0.044	0.819	0.103	0.011	0.039	0.823	0.104	0.020	0.041	0.793	0.099	0.012	0.040

Table 14: AUROC, AUARC, Platt-ECE-15, and Platt-Brier for the three single-pass coordinate-token signals across 8 fixed-emission Qwen-VL agents + Magma-8B cross-arch on full ScreenSpot-Pro ($n\approx1{,}500$ per row after no-coord filtering). Platt scaling is 5-fold CV logistic regression. AUARC is the trapezoid integral $\int_0^1 \mathrm{sel_acc}(c)\,dc$ over coverage. PACE$_{\alpha=1}$ matches or beats Coordinate-Entropy on AUROC and AUARC for all 9 models (no strict losses); Platt-scaled ECE loses on $4/9$ models, reflecting that the entropy advantage is in rank-and-deployment metrics, not parametric calibration. *Magma-8B (last row): included as cross-architecture demonstration; acc=0.044 is floor-level, so AUARC and Brier scalars are interpreted directionally rather than as primary numbers.*