Evidence-bounded encounter forecasting, an honest-model architecture for effort-biased wildlife observation data

Gil Raitses
aimez.ai

Abstract

Thornton et al. established an effort-corrected occurrence surface for southern resident killer whales in the Salish Sea, adjusting for observer effort and detectability and strengthening the result with passive acoustic detections [25]. That surface stops at a static map with no per-cell drilldown, no mechanism for withholding confidence the model has not earned, and no governance step for new citizen-science observations. orcast is a pilot platform that carries the correction into an operational loop instead. The forecast is always shown, but its displayed confidence is bounded by automated statistical gates and explicit integrity conditions, so the current effective confidence is honestly 0% while held-out deviance skill is negative, backed by per-cell provenance traces, a quarantined citizen-science ingest path, and a human promotion step. A grounding benchmark comparing structured, artifact-bound citation against a Google Maps place-grounded baseline on the same marine science queries finds that place grounding leaves 85% of scientific evidence claims uncited on average, while every orcast annotation returned for the same query carries a bound producing skill, artifact identifier, and data reference, an annotation-level result rather than a strict sentence-level replication of the Maps figure. The pilot deployment and the grounding benchmark's measurement tool are both available for inspection.

1 An operational gap in the effort-corrected occurrence surface

The authoritative occurrence record for southern resident killer whales in the Salish Sea is built from opportunistic data. Olson et al. analyzed 82,447 sightings collected from 1976 to 2014 and resolved them into a small number of recurring hotspots, while stating that these opportunistic, citizen-sourced records yield reliable distribution estimates only "if the potential for error and bias is taken into consideration" [20]. Diggle, Menezes and Su established the statistical fix for that conditional. When the process that places observations is stochastically dependent on the process being mapped, preferential sampling, conventional intensity estimates are misleading unless the dependence is modeled explicitly, for example through a log-Gaussian Cox process [7]. The DFO advisory surface for the southern Salish Sea applies that correction directly. It adjusts for observer effort and detectability and strengthens the result with passive acoustic detections [25]. orcast treats this chain, the data, the sampling-bias correction, and the applied effort-corrected surface, as settled and consumes it rather than re-deriving it.

What Thornton et al.'s surface does not do is carry that correction into an operational system. It delivers a corrected map with no per-prediction drilldown from a displayed probability to the kernels, detections, and gate verdicts that produced it, no mechanism for withholding a displayed confidence when the underlying model skill is not earned, and no governance step for the citizen-science observations that keep arriving after the advisory document is published. That verified absence has three components.

The first component is per-cell evidence exposure. A researcher reading a static effort-corrected surface cannot audit whether a displayed confidence reflects a fitted diel kernel that passed a phase-shuffled null test and a held-out deviance skill check, or a prior-dominated estimate with no model fit at all. Users of AI-assisted spatial forecasts have been shown to overtrust displayed confidence when the evidence behind it is not made visible alongside the prediction [129], and survey practitioners report that uncertainty is routinely suppressed in data-driven maps even when the authors know it exists [123].

The second component is confidence honesty. A displayed number should be able to go to zero when the model has not earned it, rather than smoothing over a failed calibration check. Movement ecology work shows that naive hotspot aggregation of the kind Olson et al. caution against learns where observers watch rather than where whales travel when search effort is spatially structured [35, 7, 20], which is exactly the condition a gate-bounded confidence has to detect and downgrade for, not merely correct for once in a static fit.

The third component is governance over new observations. Citizen-science shore reports and automated model promotions need an explicit quarantine and a signed human authority before they can influence a live forecast, not just a one-time effort correction applied to historical sightings. Unmoderated crowdsourced data produces systematic quality failures when used directly in probabilistic models [18], and the Gravity Spy program at LIGO supplies a concrete operational precedent for the alternative, citizen classifier outputs are quarantined and validated before they influence detector characterization [210].

orcast's contribution maps to these three components directly. It is the operational layer Thornton et al.'s surface does not provide, combining live forecast, a gate battery that can and does return zero confidence, per-cell provenance, quarantined community ingest, and an immutable promotion audit on top of the effort-corrected surface Olson, Diggle, and Thornton et al. already established [20, 7, 25].

2 A gate-bounded honesty model

2.1 Distributional family

Encounter counts are sparse, overdispersed, and zero-inflated. The negative binomial distribution arises as a gamma-Poisson marginalization and provides two free parameters to capture overdispersion and random effects [275]. Hurdle and NB models are required for species abundance estimation when Gaussian assumptions fail on ecological count data [14]. Species distribution modeling frameworks identify structured noise and under-counting as reasons that count-specific distributional families are necessary [78].

The kernel model uses the log-linear intensity, decomposing the log encounter rate into a baseline plus additive cyclic kernel contributions from diel, lunar, and seasonal covariates.

$$\log \lambda(x,t) = b_0 + \sum_k \beta_k \, \phi_k(t) + \log E(x,t) \tag{1}$$

See also the methodology of record.

2.2 Null gating

Each fitted cyclic kernel must beat a phase-shuffled null before contributing to displayed confidence. Phase shuffling permutes detection times across stimulus phases to generate the reference distribution. Point process models for spike trains [62] and the time-elapsed neural model [61] establish the mathematical foundation.

$$p_{\text{null}} = P\big(T(\text{shuffled}) \geq T(\text{observed})\big) \tag{2}$$

Time-rescaling applies the compensator transformation. The resulting residuals should be exponential under the null, and a KS test on them is the second gate component [87].

$$\tau_i = \Lambda(t_i) - \Lambda(t_{i-1}), \quad \tau_i \overset{\text{H}_0}{\sim} \text{Exp}(1) \tag{3}$$

This methodology, gap C-2, is documented in the neuroscience and statistical literature cited above but lacks an ecology-domain citation in the arXiv corpus. The whitepaper cites the methodological precedents rather than a cetacean-specific instance.

2.3 CV skill and calibration

Gate E5 requires positive held-out deviance skill

$$S_D = 1 - \frac{D_{\text{model}}}{D_{\text{null}}} \tag{4}$$

and gate E6 requires PIT calibration

$$u_i = F_{\text{model}}(y_i), \quad u_i \overset{\text{H}_0}{\sim} \text{Uniform}(0,1) \tag{5}$$

Negative skill means the model is less predictive than the null, and displayed confidence must be withheld. The epidemiological calibration literature establishes that PIT-based recalibration reliably improves log score [105]. The abstention loss framework [122] formalizes the principle that a predictive system should withhold confidence when skill is low. Spatial and temporal CV must account for autocorrelation [110, 108].

2.4 Effective confidence gate

Displayed effective confidence is bounded by the product of gate outcomes.

$$c_{\text{eff}} = c_{\text{raw}} \cdot \prod_k g_k, \quad g_k \in \{0, 1\} \tag{6}$$

Failed gates reduce confidence and expose the failing integrity condition alongside the metric. The gate is not a visibility switch. The forecast is always shown regardless of the confidence value.

3 Acoustic grounding, PSTH and Level 0 QC

Passive acoustic monitoring is orcast's primary temporal instrument. Detectors run continuously and effort does not track human activity cycles.

3.1 PSTH kernel estimation

The PSTH method aligns detections to a repeating stimulus onset, such as diel sunrise, flood-current onset, or lunar new moon, and bins by phase. The binned counts, divided by station effort and smoothed with a periodic spline, produce a first kernel estimate. Point-process models for spike trains [62] and the time-elapsed neural model [61] establish the Poisson-process machinery. The hydrophone functions as the detector unit, the detection times as the spike train, and the environmental cycles as the stimulus.

3.2 False-positive rates and quarantine

Deep learning bioacoustic classifiers exhibit non-trivial false-positive rates in field deployment. The computational bioacoustics review [23] documents this as a known, widespread problem, performance degrades in field conditions relative to held-out test sets. The North Atlantic right whale upcall detector [50] demonstrates this for a marine mammal application, where a ResNet classifier trained on thousands of examples requires careful validation under variable acoustic conditions. Robust sound event detection in bioacoustic networks [59] characterizes ambient noise variability as the primary driver of detection unreliability.

For orcast, unreviewed OrcaHello candidates cannot enter a model fit directly. The Level 0 QC gate quarantines them in a separate data stream, and the integrity condition exposed alongside every confidence-bearing surface reports explicitly that the acoustic data consists of unreviewed candidates. The reported false-positive rate is 0.673, drawn from an internal Level 0 QC fit dated 2022-06-22 rather than a published sample size or classifier description in this paper, and it should be read as a project-internal calibration figure. The effective confidence gate in Section 2 prevents confidence from rising above a threshold while this condition holds.

The integrity condition is part of the forecast, not a reason to hide it.

4 Governance, quarantine, authority, and audit

Shore-based sighting submissions enter orcast through a moderation queue. Quarantine blocks any model influence until approval. The governance architecture has three components, quarantine by default, signed human authority, and an immutable decision record.

4.1 Why quarantine is necessary

Citizen science is a user-generated content problem in the information quality literature, where quality pipelines are necessary for scientific usability [18]. The LIGO Gravity Spy program provides an operational precedent. Citizen classifier votes are aggregated and validated against expert labels before influencing detector characterization [210]. The Milky Way Project demonstrates that unmoderated citizen labels alone produce systematic errors without ensemble aggregation and expert calibration [213]. Galaxy morphology classification confirms that hybrid human-ML pipelines outperform either component alone only when citizen data is filtered for quality [211].

4.2 Signed human authority

orcast separates automated gate eligibility from human promotion authority. A model fit that passes kernel-program levels L1 through L3, the fitness gates, is eligible for promotion, but promotion requires a signed decision by a registered reviewer recorded in the decision-records DynamoDB table, one of the nine DynamoDB tables in the system of record. The record includes reviewer identity, verdict, cited gates, and a rationale field. The record is immutable and auditable.

4.3 Moderation queue mechanics

Community shore submissions are stored in a pending state. Approval stamps attribution and a low reliability weight before the submission can influence the model. Private field journal entries remain private unless the author publishes them to the community queue, at which point they enter the same moderation pipeline.

Every public confidence surface currently deployed draws only on validated and promoted data, and the provenance chain is auditable to the specific reviewer decisions that authorized each data point. That guarantee holds for the pilot's current data sources and is a property of the moderation pipeline design, not a claim that every future data source orcast might add will be reviewed before this paper is updated.

5 Provenance grounding as a system

Every displayed claim must trace to producing method, validating experiment, experimental data, and motivating research.

5.1 Prepare-then-narrate

The concierge dispatches only allow-listed skills from a versioned manifest. Each skill invocation is recorded as a step. After all skills complete, the LLM narrates the collected JSON context. RAG reduces factual hallucination in conversational systems [145]. CRAG [146] demonstrates that corrective retrieval reduces hallucination when retrieval fails. AGREE [152] shows that adapting LLMs to self-ground claims and generate citations reduces unsupported-claim rates. The LLM attribution survey [154] confirms that generative systems without attribution produce high rates of unauditable unsupported claims.

5.2 Step-log provenance

Every interaction persists an ordered `steps` array in `exploration_turns` as JSONB. A `resolve_agent` step records the cast role id, version, and spec hash. `skill_invocation` steps carry grounding references. A `model_output` step carries annotations. CamFlow whole-system provenance [189] establishes the requirements. The provenance and data differencing paper [191] demonstrates step-log sufficiency for reproducibility. Scientific ML workflow provenance in geoscience [192] directly addresses this class of pipeline.

5.3 Claim/method/experiment graph

The ProvenanceGraph component renders the step log as a two-tier visual graph.

$$G = (V, E), \quad V = \{\text{metric}, C, M, X, D, R\} \tag{7}$$

Claim nodes, C, are constructed from model output annotations. Each C node names the annotation type, bound artifact or href, and a no-signal badge if unbound. Method nodes, M, are constructed from skill invocation steps. Each M node names the skill, status, and lists `grounding_refs`, meaning artifact or dataset references populated during skill execution. The C and M tiers are rendered in the live UI. The full C/M/X/D/R schema, where X is experiment or artifact, D is a provenance citation, and R is a research citation, is stored in step-log JSONB as the intended rendering target. X, D, and R nodes currently appear inline within M rows rather than as separate graph nodes. SciERC [252] establishes the joint entity-relation extraction foundation. SciClaim [253] introduces the fine-grained C/M/X annotation schema with typed relational edges. MindSearch [258] demonstrates graph-of-queries decomposition at scale. MindMap [251] demonstrates that KG-grounded reasoning reduces LLM hallucination.

Claim nodes without an artifact or href binding render a no-signal badge, which the geospatial grounding baseline in Section 7 does not expose.

6 Orchestrated managed agents with verified tools

6.1 Plan-then-execute and tier verification

The surface planner emits a `ui_intent` specifying panels and a `skill_plan`. The concierge dispatches exactly those skills. Skills not in the cast role's allowed list return a 400 `tier_blocked` before invocation. The model narrates only after all skills complete.

ReAct [180] establishes the canonical plan-act-verify pattern. Constraining the action space to a defined tool set makes the verify step tractable and reduces hallucination. Internal representations predict tool selection hallucination before execution [33]. HaluAgent [168] demonstrates that a structured typed toolbox enables reliable operation from smaller models. APIGen [28] establishes that verifiable function calling is an automatable engineering property.

6.2 Skill catalog tiers

T0 skills are public. T1 skills require a geographic viewport. T2 and T3 skills are keyed and blocked on the public route. Cast roles specify an explicit skill subset, model, and policy. Unknown skill IDs return 400. Four cast roles are deployed, `explore-guide-v1`, `surface-planner-v1`, `dossier-explainer-v1`, and `promotion-clerk-v1`.

6.3 Step logs as a retrieval substrate

The persisted step-log corpus is the data foundation for embedding-indexed retrieval over prior fits and provenance chains, chartered separately as W-RAG. The step log is an audit record today. TRACE [32] demonstrates that knowledge-grounded reasoning chains constructed from structured context eliminate irrelevant retrieval noise in multi-hop queries, and the orcast step-log is this same structure applied to skill-dispatch interactions. Whether retrieval over this corpus measurably improves grounding quality is the open question W-RAG is chartered to test, not a result this paper claims.

7 Geospatial grounding limits for scientific evidence

7.1 Benchmark methodology

Gemini 3.5 Flash, Api-Revision 2026-05-20, with the `google_maps` tool at San Juan Islands coordinates 48.5465, −123.03, was measured on total citation count, citation type split between place and scientific or dataset sources, grounding density per thousand words, and the unsupported scientific claim rate below.

$$R_{\text{uncited}} = \frac{\left|\{s \in S_{\text{sci}} : \nexists\, \text{cited\_name} \in s\}\right|}{|S_{\text{sci}}|} \tag{8}$$

This formula is defined over scientific sentences in a text response. The orcast side of the comparison uses a related but distinct unit, annotations returned by `/api/interactions/plan` for the same orca evidence query, each checked for a bound artifact or href rather than each checked as a sentence. The two counts apply the same uncited-fraction logic to two different output structures rather than replicating one measurement, and a sentence-level orcast measurement is listed as future work in Section 9. The measurement tool is available at `tools/testing/maps_grounding_probe.py`.

7.2 Results

Across three Maps grounding queries, 0 of 25 citations resolved to scientific or dataset sources. All 25 resolved to Google Maps place pins. The evidence query produced 9 scientific sentences drawing on NOAA fisheries data, Center for Whale Research census figures, Chinook salmon corridor data, and J/K/L pod hydrophone data. 8 of the 9 sentences, 89%, were uncited. Averaged across evidence-bearing queries, $R_{\text{uncited}} = 85\%$ for Maps grounding. On the orcast side, all 6 of 6 returned annotations carried a bound artifact or data reference, giving an annotation-level uncited rate of 0% under the same formula.

7.3 Interpretation

This reflects geospatial grounding's structural scope boundary, not a Gemini or Maps deficiency. Maps grounding is designed for place information. The POI query in this same run achieved 8 citations in 229 words with 0 unsupported sentences. Place grounding does not resolve scientific evidence claims to their origin data, method, or experiment. The LLM geospatial knowledge study [232] provides peer-reviewed evidence that LLM geospatial competence degrades substantially for knowledge not encoded as named places. That boundary is the motivation for a domain-specific provenance system as the evidence layer.

7.4 Reproducibility

This benchmark ran on 2026-06-24 using gemini-3.5-flash, API revision 2026-05-20, at coordinates 48.5465, −123.03. It is fully reproducible via `tools/testing/maps_grounding_probe.py` with a Gemini API key.

7.5 Extended benchmark, an 8-scenario parallel RAG comparison

A follow-on 8-scenario parallel benchmark, `tools/testing/grounding_parallel_rag.py`, run 2026-06-24, compared Maps-only grounding against five RAG-augmented variants using live orcast skill outputs as context. The hierarchy of R_{uncited} across conditions is shown below.

Scenario	Context injected	R_{uncited}
S4, surface planner step-log	Orchestrated reasoning trace	**0%**
S8, hotspot data	Structured probability data	75%
S1, S7, Maps-only baseline	None	60–100%
S2, gate context	Narrow gate data	100%

The surface planner step-log, Scenario 4, drives R_{uncited} to 0% by transforming the query from an open-domain science query into a closed artifact-reference query. The model answers from the step-log rather than from world knowledge, a query transformation rather than a citation-generation improvement. Structured probability data alone, Scenario 8's hotspot injection, lowers R_{uncited} to 75%, a 25 percentage point reduction from the Scenario 7 Maps-only baseline of 100%, but does not eliminate uncited claims. Gate context, Scenario 2, does not reduce R_{uncited}, because the injected gate data does not cover the scientific claims Gemini independently generates. Context that binds the query to a complete reasoning trace removes uncited claims, while narrowly structured context does not. This finding motivates a formal evaluation framework for AI reasoning systems, developed in the companion paper *Grounding quality measurement for orchestrated AI reasoning chains* [152, 146].

8 Limits and falsification

8.1 Current scope limits

The pilot covers one hydrophone station at Haro Strait on OrcaHello over approximately 7.5 months. Diel and lunar kernels are fitted. Tide, season, and salmon index kernels are in the leveled program but not yet fitted at kernel-program level L1. The spatial kernel is prior-dominated. Effective confidence is 0%, held-out CV deviance skill −0.018 with 3 of 5 folds passing, pending a reviewer promotion decision.

Two search families returned Partial verdicts. The effort bias family lacks a cetacean-specific effort-correction measurement in the arXiv corpus. The analogs [34, 36] support the methodological claim but not a cetacean-domain number. The null gating family lacks ecology-journal documentation of phase-shuffled null tests for diel and lunar covariates. The neuroscience precedents [62, 61] establish the statistical machinery.

8.2 Falsification table

H1 is the hypothesis the W-Eval study is chartered to test directly, that displaying gates and integrity conditions alongside a forecast increases steward trust relative to a confidence-smoothed baseline map with the same underlying data. The conditions below would weaken or falsify H1, or a specific governance guarantee this deployment makes, ahead of that field study.

Observation	Interpretation
Displayed confidence rises without a promotion decision	Gate policy failed, H1 weakened
Provenance modal omits a failed gate	Honesty contract violated
Citizen submission influences the model without moderation approval	Governance quarantine violated
Sighting check returns a confident yes or no without uncertainty disclosure	Prepare-then-narrate violated
New covariate passes gates but CV deviance skill is negative	Correct outcome, the gate correctly withholds confidence

8.3 Extension path

Adding stations, reviewed OrcaHello labels, effort covariates, and spatial kernels re-runs the full gate battery. Confidence moves only when gates and human authority agree. The W-Eval study, chartered separately, tests H1 directly through a structured field observer panel comparing steward trust in the gate-bounded forecast against a confidence-smoothed baseline.

8.4 Connection to world model evaluation

The provenance and gating approach here is domain-specific to encounter forecasting. Whether a claim-level evidence-binding measurement extends to physical-world planning traces more broadly, in the style of LeCun's autonomous machine intelligence architecture [16] or benchmarks such as V-JEPA [3], is optional future work rather than a claim this paper makes. The companion paper, *Grounding quality measurement for orchestrated AI reasoning chains*, takes up that question directly [32].

Glossary

Encounter intensity

Encounter intensity $\lambda(x,t)$ is the expected rate of orca encounters at map cell x and time t under the fitted point-process model. It is expressed in the log domain as a sum of kernel contributions (Eq. 1) and converted to an encounter probability per cell per time bin as $p(x,t) = 1 - e^{-\lambda(x,t)\,\Delta t}$.

Fitness gate

A fitness gate is a binary statistical decision that must pass before the corresponding evidence component is allowed to contribute to displayed confidence. Gates include the phase-shuffled null test (Eq. 2), the time-rescaling goodness-of-fit test (Eq. 3), held-out deviance skill (Eq. 4), and PIT calibration (Eq. 5). A failing gate does not suppress the forecast; it lowers the effective confidence and surfaces the failing integrity condition.

Effective confidence

Effective confidence c_{eff} is the displayed confidence score, bounded by the product of all gate outcomes (Eq. 6). It is the model-estimated confidence multiplied by one binary gate outcome per gated condition. When any gate fails, effective confidence is reduced and the integrity condition is shown alongside the metric. Effective confidence is never promoted above the model estimate; it can only be equal to or lower than it.

Integrity condition

An integrity condition is a plain-language statement of a scope limit or data quality constraint that is shown alongside every confidence-bearing display surface. Examples include "effort assumed continuous (not effort-corrected)", "spike-train fit uses unreviewed acoustic candidates", and "season kernel extrapolated beyond observed coverage." Integrity conditions are features of the honesty model, not errors to be concealed.

Effort offset

The effort offset $\log E(x,t)$ in Eq. 1 is an additive term that accounts for heterogeneous observation effort across stations and time. Where effort is known and roughly constant (continuous acoustic monitoring), the offset is fixed; where effort is heterogeneous (visual sighting surveys), the offset is estimated from patrol records or assumed. An integrity condition is surfaced whenever the effort offset is assumed rather than measured.

Phase-shuffled null

A phase-shuffled null is a reference distribution produced by permuting the assignment of detection times to stimulus phase bins, preserving the marginal detection rate while destroying any phase relationship. The test statistic from the observed kernel is compared to this distribution to produce p_{null} (Eq. 2). A kernel that does not beat the phase-shuffled null is not allowed to contribute to displayed confidence.

Time-rescaling goodness-of-fit

Time-rescaling transforms observed detection times through the model compensator $\Lambda(t)$ to produce residuals τ_i (Eq. 3) that should be exponentially distributed if the model fits. A Kolmogorov-Smirnov test on these residuals is the goodness-of-fit gate. Failure means the model does not fit the temporal structure of the data.

Deviance skill score

The deviance skill score S_D (Eq. 4) measures whether the fitted model is sharper than a climatology null on held-out data. $S_D < 0$ means the model is worse than the null on the held-out set; displayed confidence is withheld when this gate fails.

PIT uniformity

The probability integral transform (PIT) residual $u_i = F_{\text{model}}(y_i)$ (Eq. 5) should be uniformly distributed if the predictive distribution is calibrated. A Kolmogorov-Smirnov test on PIT residuals is the calibration gate. Failure means the model is systematically over- or under-confident.

Provenance graph

A provenance graph is the directed graph $G = (V, E)$ (Eq. 7) that traces a displayed metric from its claim node through the method that produced it, the experiment that validated the method, the data the experiment ran on, and the research that motivated the experimental design. The graph is rendered from the persisted interaction step log; every node has a producing artifact reference, and any claim node without one renders a no-signal badge.

Skill dispatch tier

A skill dispatch tier is a categorical access level for interaction skills. T0 skills are public; T1 skills require a geographic viewport; T2 and T3 skills are keyed and blocked on public routes. The tier system enforces that the model cannot invoke high-privilege tools without appropriate authentication, and unknown skill identifiers return a 400 response before invocation.

Prepare-then-narrate

Prepare-then-narrate is the orcast interaction pipeline pattern in which a concierge orchestrator dispatches all allow-listed skills and collects grounded context before the LLM narrates. The model does not select tools dynamically during generation; it narrates the collected JSON. This ensures that every claim in the narration has a producing skill, artifact reference, and step log entry.

Surface planner

A surface planner is a keyed cast role (surface-planner-v1) that emits a validated `ui_intent` object specifying which panels to open and a `skill_plan` listing which skills to run, before any narration occurs. The surface planner separates planning from execution; the concierge dispatches exactly the skill plan without model-selected additions.

Uncited-evidence rate

The uncited-evidence rate R_{uncited} (Eq. 8) measures the fraction of sentences containing scientific or quantitative content that carry no bound citation. This is the primary metric in the grounding benchmark (Section 7). The Google Maps grounding baseline achieved $R_{\text{uncited}} = 0.85$ (85%) averaged across marine science evidence queries (2026-06-24); orcast achieved $R_{\text{uncited}} = 0$ on the same orca evidence query with all annotations bound to an artifact or provenance reference.

References

[1] S. Abdelnabi and M. Fritz. "Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding". In: *IEEE Symposium on Security and Privacy* (2020). DOI: 10.1109/SP40001.2021.00083.

[2] A. Bakhtin et al. *PHYRE: A New Benchmark for Physical Reasoning*. 2019. arXiv: 1908.05656. URL: https://arxiv.org/abs/1908.05656.

[3] A. Bardes et al. *Revisiting Feature Prediction for Learning Visual Representations from Video*. 2024. arXiv: 2404.08471. URL: https://arxiv.org/abs/2404.08471.

[4] E. Bouffard et al. "No Sign of G2's Encounter Affecting Sgr A*'s X-Ray Flaring Rate from Chandra Observations". In: *Astrophysical Journal* (2019). DOI: 10.3847/1538-4357/ab4266.

[5] N. Bozanic and T. Kreuz. "SPIKY: a graphical user interface for monitoring spike train synchrony". In: *BMC Neuroscience* (2013). DOI: 10.1186/1471-2202-14-S1-P225.

[6] P. Ciccarese et al. "PAV ontology: provenance, authoring and versioning". In: *Journal of Biomedical Semantics* (2013). DOI: 10.1186/2041-1480-4-37.

[7] P. J. Diggle, R. Menezes, and T.-l. Su. "Geostatistical inference under preferential sampling". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59.2 (2010), pp. 191–232. DOI: 10.1111/j.1467-9876.2009.00701.x.

[8] M. Dorkenwald et al. "SCVRL: Shuffled Contrastive Video Representation Learning". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2022). DOI: 10.1109/CVPRW56347.2022.00458.

[9] A. Doshi et al. "Towards Verifiably Safe Tool Use for LLM Agents". In: *arXiv.org* (2026). DOI: 10.48550/arXiv.2601.08012.

[10] I. Gaudiello et al. "Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers". In: *Computers in Human Behavior* (2015). DOI: 10.1016/j.chb.2016.03.057.

[11] T. Ginsberg, Z. Liang, and R. G. Krishnan. "A Learning Based Hypothesis Test for Harmful Covariate Shift". In: *International Conference on Learning Representations* (2022). DOI: 10.48550/arXiv.2212.02742.

[12] W. Jitkrittum et al. "A Linear-Time Kernel Goodness-of-Fit Test". In: *Neural Information Processing Systems* (2017). DOI: 10.1093/jrsssb/qkad050.

[13] A. Kirichenko and H. Zanten. "Optimality of Poisson processes intensity learning with Gaussian processes". In: *Journal of machine learning research* (2014). DOI: 10.5555/2789272.2912093.

[14] S. Kong et al. "Deep Hurdle Networks for Zero-Inflated Multi-Target Regression: Application to Multiple Species Abundance Estimation". In: *International Joint Conference on Artificial Intelligence* (2020). DOI: 10.24963/ijcai.2020/603.

[15] V. Lai et al. "Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation". In: *International Conference on Human Factors in Computing Systems* (2022). DOI: 10.1145/3491102.3501999.

[16] Y. LeCun and K. Alahari. *Introduction to Latent Variable Energy-Based Models: A Path Towards Autonomous Machine Intelligence*. 2023. arXiv: 2306.02572. URL: https://arxiv.org/abs/2306.02572.

[17] J. Lu et al. "ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities". In: *North American Chapter of the Association for Computational Linguistics* (2024). DOI: 10.48550/arXiv.2408.04682.

[18] R. Lukyanenko, A. Wiggins, and H. K. Rosser. "Citizen Science: An Information Quality Research Frontier". In: *Information Systems Frontiers* (2019). DOI: 10.1007/s10796-019-09915-z.

[19] K. Oksuz, B. C. Cam, E. Akbas, and S. Kalkan. "Localization Recall Precision (LRP): A New Performance Metric for Object Detection". In: *European Conference on Computer Vision* (2018). DOI: 10.1007/978-3-030-01234-2_31.

[20] J. K. Olson et al. "Sightings of southern resident killer whales in the Salish Sea 1976–2014: the importance of a long-term opportunistic dataset". In: *Endangered Species Research* 37 (2018), pp. 105–118. DOI: 10.3354/esr00918.

[21] M. Pourhomayoun, P. J. Dugan, M. Popescu, and C. Clark. "Bioacoustic Signal Classification Based on Continuous Region Processing, Grid Masking and Artificial Neural Network". In: *arXiv.org* (2013). DOI: 10.1016/j.jasc.2018.06.156.

[22] N. B. Shah, D. Zhou, and Y. Peres. "Approval Voting and Incentives in Crowdsourcing". In: *International Conference on Machine Learning* (2015). DOI: 10.1145/3396863.

[23] D. Stowell. "Computational bioacoustics with deep learning: a review and roadmap". In: *PeerJ* (2021). DOI: 10.7717/peerj.13152.

[24] X. Sun et al. "Towards Detecting LLMs Hallucination via Markov Chain-based Multi-agent Debate Framework". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2024). DOI: 10.1109/ICASSP49660.2025.10889448.

[25] S. J. Thornton et al. *Southern Resident Killer Whale (Orcinus orca) summer distribution and habitat use in the southern Salish Sea and the Swiftsure Bank area (2009 to 2020)*. Research Document 2022/037. DFO Canadian Science Advisory Secretariat (CSAS), 2022.

[26] Unknown. *The ThreeDWorld Transport Challenge: A Visually Guided Task-and-Motion Planning Benchmark for Physically Realistic Embodied AI*. 2021. arXiv: 2103.14025. URL: https://arxiv.org/abs/2103.14025.

[27] Unknown. *Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI*. 2024. arXiv: 2407.06886. URL: https://arxiv.org/abs/2407.06886.

[28] Unknown. *APIGen: Automated Pipeline for Generating Verifiable and Diverse Function-Calling Datasets*. 2024. arXiv: 2406.18518. URL: https://arxiv.org/abs/2406.18518.

[29] Unknown. *Elements of World Knowledge (EWOK): A Cognition-Inspired Framework for Evaluating Basic World Knowledge in Language Models*. 2024. arXiv: 2405.09605. URL: https://arxiv.org/abs/2405.09605.

[30] Unknown. *Learning and Leveraging World Models in Visual Representation Learning*. 2024. arXiv: 2403.00504. URL: https://arxiv.org/abs/2403.00504.

[31] Unknown. *STAR: A Benchmark for Situated Reasoning in Real-World Videos*. 2024. arXiv: 2405.09711. URL: https://arxiv.org/abs/2405.09711.

[32] Unknown. *TRACE the Evidence: Constructing Knowledge-Grounded Reasoning Chains for Retrieval-Augmented Generation*. 2024. arXiv: 2406.11460. URL: https://arxiv.org/abs/2406.11460.

[33] Unknown. *Internal Representations as Indicators of Hallucinations in Agent Tool Selection*. 2026. arXiv: 2601.05214. URL: https://arxiv.org/abs/2601.05214.

[34] Unknown. arXiv: 1903.06669. URL: https://arxiv.org/abs/1903.06669.

[35] Unknown. arXiv: 1907.05902. URL: https://arxiv.org/abs/1907.05902.

[36] Unknown. arXiv: 2110.12951. URL: https://arxiv.org/abs/2110.12951.

[37] Unknown. arXiv: 1805.05356. URL: https://arxiv.org/abs/1805.05356.

[38] Unknown. arXiv: 2205.11324. URL: https://arxiv.org/abs/2205.11324.

[39] Unknown. arXiv: 2006.00272. URL: https://arxiv.org/abs/2006.00272.

[40] Unknown. arXiv: 2306.02707. URL: https://arxiv.org/abs/2306.02707.

[41] Unknown. arXiv: 2311.11045. URL: https://arxiv.org/abs/2311.11045.

[42] Unknown. arXiv: 2212.04531. URL: https://arxiv.org/abs/2212.04531.

[43] Unknown. arXiv: 2111.13048. URL: https://arxiv.org/abs/2111.13048.

[44] Unknown. arXiv: 1503.07150. URL: https://arxiv.org/abs/1503.07150.

[45] Unknown. arXiv: 1602.08507. URL: https://arxiv.org/abs/1602.08507.

[46] Unknown. arXiv: 1902.09069. URL: https://arxiv.org/abs/1902.09069.

[47] Unknown. arXiv: 1807.05812. URL: https://arxiv.org/abs/1807.05812.

[48] Unknown. arXiv: 1706.05286. URL: https://arxiv.org/abs/1706.05286.

[49] Unknown. arXiv: 1612.03505. URL: https://arxiv.org/abs/1612.03505.

[50] Unknown. arXiv: 2001.09127. URL: https://arxiv.org/abs/2001.09127.

[51] Unknown. arXiv: 1908.09635. URL: https://arxiv.org/abs/1908.09635.

[52] Unknown. arXiv: 1401.5878. URL: https://arxiv.org/abs/1401.5878.

[53] Unknown. arXiv: 2104.08614. URL: https://arxiv.org/abs/2104.08614.

[54] Unknown. arXiv: 1907.13188. URL: https://arxiv.org/abs/1907.13188.

[55] Unknown. arXiv: 1503.03173. URL: https://arxiv.org/abs/1503.03173.

[56] Unknown. arXiv: 2007.02890. URL: https://arxiv.org/abs/2007.02890.

[57] Unknown. arXiv: 1509.03591. URL: https://arxiv.org/abs/1509.03591.

[58] Unknown. arXiv: 2307.06860. URL: https://arxiv.org/abs/2307.06860.

[59] Unknown. arXiv: 1905.08352. URL: https://arxiv.org/abs/1905.08352.

[60] Unknown. arXiv: 2011.07640. URL: https://arxiv.org/abs/2011.07640.

[61] Unknown. arXiv: 1506.02361. URL: https://arxiv.org/abs/1506.02361.

[62] Unknown. arXiv: 2010.04875. URL: https://arxiv.org/abs/2010.04875.

[63] Unknown. arXiv: 0903.0676. URL: https://arxiv.org/abs/0903.0676.

[64] Unknown. arXiv: 1305.3250. URL: https://arxiv.org/abs/1305.3250.

[65] Unknown. arXiv: 1209.3442. URL: https://arxiv.org/abs/1209.3442.

[66] Unknown. arXiv: 1206.6456. URL: https://arxiv.org/abs/1206.6456.

[67] Unknown. arXiv: 1701.01503. URL: https://arxiv.org/abs/1701.01503.

[68] Unknown. arXiv: 1610.08623. URL: https://arxiv.org/abs/1610.08623.

[69] Unknown. arXiv: 2011.03172. URL: https://arxiv.org/abs/2011.03172.

[70] Unknown. arXiv: 2201.09485. URL: https://arxiv.org/abs/2201.09485.

[71] Unknown. arXiv: 1206.4650. URL: https://arxiv.org/abs/1206.4650.

[72] Unknown. arXiv: 2205.02986. URL: https://arxiv.org/abs/2205.02986.

[73] Unknown. arXiv: 1209.3274. URL: https://arxiv.org/abs/1209.3274.
[74] Unknown. arXiv: 2104.10038. URL: https://arxiv.org/abs/2104.10038.
[75] Unknown. arXiv: 2302.10160. URL: https://arxiv.org/abs/2302.10160.
[76] Unknown. arXiv: 1404.3331. URL: https://arxiv.org/abs/1404.3331.
[77] Unknown. arXiv: 1912.02399. URL: https://arxiv.org/abs/1912.02399.
[78] Unknown. arXiv: 2102.08534. URL: https://arxiv.org/abs/2102.08534.
[79] Unknown. arXiv: 1206.4685. URL: https://arxiv.org/abs/1206.4685.
[80] Unknown. arXiv: 2004.11977. URL: https://arxiv.org/abs/2004.11977.
[81] Unknown. arXiv: 1409.2713. URL: https://arxiv.org/abs/1409.2713.
[82] Unknown. arXiv: 2201.11948. URL: https://arxiv.org/abs/2201.11948.
[83] Unknown. arXiv: 2401.17110. URL: https://arxiv.org/abs/2401.17110.
[84] Unknown. arXiv: 1211.0313. URL: https://arxiv.org/abs/1211.0313.
[85] Unknown. arXiv: 1610.01424. URL: https://arxiv.org/abs/1610.01424.
[86] Unknown. arXiv: 1602.03253. URL: https://arxiv.org/abs/1602.03253.
[87] Unknown. arXiv: 1606.03988. URL: https://arxiv.org/abs/1606.03988.
[88] Unknown. arXiv: 1102.5593. URL: https://arxiv.org/abs/1102.5593.
[89] Unknown. arXiv: 1206.6367. URL: https://arxiv.org/abs/1206.6367.
[90] Unknown. arXiv: 1409.7137. URL: https://arxiv.org/abs/1409.7137.
[91] Unknown. arXiv: 1407.6297. URL: https://arxiv.org/abs/1407.6297.
[92] Unknown. arXiv: 2002.10046. URL: https://arxiv.org/abs/2002.10046.
[93] Unknown. arXiv: 1705.07997. URL: https://arxiv.org/abs/1705.07997.
[94] Unknown. arXiv: 2211.02843. URL: https://arxiv.org/abs/2211.02843.
[95] Unknown. arXiv: 2004.13892. URL: https://arxiv.org/abs/2004.13892.
[96] Unknown. arXiv: 1912.09363. URL: https://arxiv.org/abs/1912.09363.
[97] Unknown. arXiv: 1712.08883. URL: https://arxiv.org/abs/1712.08883.
[98] Unknown. arXiv: 1611.09509. URL: https://arxiv.org/abs/1611.09509.
[99] Unknown. arXiv: 2005.03244. URL: https://arxiv.org/abs/2005.03244.
[100] Unknown. arXiv: 1212.0463. URL: https://arxiv.org/abs/1212.0463.
[101] Unknown. arXiv: 1808.00111. URL: https://arxiv.org/abs/1808.00111.
[102] Unknown. arXiv: 2108.07142. URL: https://arxiv.org/abs/2108.07142.
[103] Unknown. arXiv: 1609.02721. URL: https://arxiv.org/abs/1609.02721.
[104] Unknown. arXiv: 2110.15209. URL: https://arxiv.org/abs/2110.15209.
[105] Unknown. arXiv: 2112.06305. URL: https://arxiv.org/abs/2112.06305.
[106] Unknown. arXiv: 1004.2316. URL: https://arxiv.org/abs/1004.2316.
[107] Unknown. arXiv: 1706.07581. URL: https://arxiv.org/abs/1706.07581.
[108] Unknown. arXiv: 1905.11744. URL: https://arxiv.org/abs/1905.11744.
[109] Unknown. arXiv: 1708.07180. URL: https://arxiv.org/abs/1708.07180.
[110] Unknown. arXiv: 2005.14263. URL: https://arxiv.org/abs/2005.14263.
[111] Unknown. arXiv: 2001.05948. URL: https://arxiv.org/abs/2001.05948.
[112] Unknown. arXiv: 2205.09680. URL: https://arxiv.org/abs/2205.09680.
[113] Unknown. arXiv: 2209.04892. URL: https://arxiv.org/abs/2209.04892.
[114] Unknown. arXiv: 2004.13546. URL: https://arxiv.org/abs/2004.13546.
[115] Unknown. arXiv: 2008.03288. URL: https://arxiv.org/abs/2008.03288.
[116] Unknown. arXiv: 2008.03033. URL: https://arxiv.org/abs/2008.03033.
[117] Unknown. arXiv: 0710.0485. URL: https://arxiv.org/abs/0710.0485.
[118] Unknown. arXiv: 1912.08581. URL: https://arxiv.org/abs/1912.08581.
[119] Unknown. arXiv: 1904.07192. URL: https://arxiv.org/abs/1904.07192.
[120] Unknown. arXiv: 2208.08225. URL: https://arxiv.org/abs/2208.08225.
[121] Unknown. arXiv: 2207.07560. URL: https://arxiv.org/abs/2207.07560.
[122] Unknown. arXiv: 2104.08236. URL: https://arxiv.org/abs/2104.08236.
[123] Unknown. arXiv: 1908.01697. URL: https://arxiv.org/abs/1908.01697.
[124] Unknown. arXiv: 2006.14120. URL: https://arxiv.org/abs/2006.14120.
[125] Unknown. arXiv: 1912.06341. URL: https://arxiv.org/abs/1912.06341.
[126] Unknown. arXiv: 2008.00058. URL: https://arxiv.org/abs/2008.00058.
[127] Unknown. arXiv: 1302.4959. URL: https://arxiv.org/abs/1302.4959.
[128] Unknown. arXiv: 1905.05176. URL: https://arxiv.org/abs/1905.05176.
[129] Unknown. arXiv: 2001.02114. URL: https://arxiv.org/abs/2001.02114.
[130] Unknown. arXiv: 2010.07938. URL: https://arxiv.org/abs/2010.07938.
[131] Unknown. arXiv: 1809.10453. URL: https://arxiv.org/abs/1809.10453.
[132] Unknown. arXiv: 1703.02169. URL: https://arxiv.org/abs/1703.02169.
[133] Unknown. arXiv: 2010.07915. URL: https://arxiv.org/abs/2010.07915.
[134] Unknown. arXiv: 2012.08091. URL: https://arxiv.org/abs/2012.08091.
[135] Unknown. arXiv: 2111.14844. URL: https://arxiv.org/abs/2111.14844.
[136] Unknown. arXiv: 2107.10400. URL: https://arxiv.org/abs/2107.10400.
[137] Unknown. arXiv: 2103.07107. URL: https://arxiv.org/abs/2103.07107.
[138] Unknown. arXiv: 1906.04253. URL: https://arxiv.org/abs/1906.04253.
[139] Unknown. arXiv: 2305.01079. URL: https://arxiv.org/abs/2305.01079.
[140] Unknown. arXiv: 1606.08157. URL: https://arxiv.org/abs/1606.08157.
[141] Unknown. arXiv: 1805.00734. URL: https://arxiv.org/abs/1805.00734.
[142] Unknown. arXiv: 2108.02325. URL: https://arxiv.org/abs/2108.02325.
[143] Unknown. arXiv: 1811.02164. URL: https://arxiv.org/abs/1811.02164.
[144] Unknown. arXiv: 2203.07547. URL: https://arxiv.org/abs/2203.07547.
[145] Unknown. arXiv: 2104.07567. URL: https://arxiv.org/abs/2104.07567.
[146] Unknown. arXiv: 2401.15884. URL: https://arxiv.org/abs/2401.15884.
[147] Unknown. arXiv: 2402.16063. URL: https://arxiv.org/abs/2402.16063.
[148] Unknown. arXiv: 2402.10612. URL: https://arxiv.org/abs/2402.10612.
[149] Unknown. arXiv: 2309.16781. URL: https://arxiv.org/abs/2309.16781.
[150] Unknown. arXiv: 2112.09047. URL: https://arxiv.org/abs/2112.09047.
[151] Unknown. arXiv: 2307.16883. URL: https://arxiv.org/abs/2307.16883.
[152] Unknown. arXiv: 2311.09533. URL: https://arxiv.org/abs/2311.09533.
[153] Unknown. arXiv: 2202.13843. URL: https://arxiv.org/abs/2202.13843.
[154] Unknown. arXiv: 2311.03731. URL: https://arxiv.org/abs/2311.03731.
[155] Unknown. arXiv: 2309.12311. URL: https://arxiv.org/abs/2309.12311.
[156] Unknown. arXiv: 2309.01352. URL: https://arxiv.org/abs/2309.01352.
[157] Unknown. arXiv: 2310.17140. URL: https://arxiv.org/abs/2310.17140.
[158] Unknown. arXiv: 2403.17124. URL: https://arxiv.org/abs/2403.17124.
[159] Unknown. arXiv: 2307.00329. URL: https://arxiv.org/abs/2307.00329.
[160] Unknown. arXiv: 1311.4610. URL: https://arxiv.org/abs/1311.4610.
[161] Unknown. arXiv: 2310.00305. URL: https://arxiv.org/abs/2310.00305.
[162] Unknown. arXiv: 2112.01640. URL: https://arxiv.org/abs/2112.01640.
[163] Unknown. arXiv: 2104.11572. URL: https://arxiv.org/abs/2104.11572.
[164] Unknown. arXiv: 2202.01993. URL: https://arxiv.org/abs/2202.01993.
[165] Unknown. arXiv: 1903.07669. URL: https://arxiv.org/abs/1903.07669.
[166] Unknown. arXiv: 2106.02192. URL: https://arxiv.org/abs/2106.02192.
[167] Unknown. arXiv: 2102.08094. URL: https://arxiv.org/abs/2102.08094.
[168] Unknown. arXiv: 2406.11277. URL: https://arxiv.org/abs/2406.11277.
[169] Unknown. arXiv: 2406.06950. URL: https://arxiv.org/abs/2406.06950.
[170] Unknown. arXiv: 2402.09733. URL: https://arxiv.org/abs/2402.09733.
[171] Unknown. arXiv: 2401.08358. URL: https://arxiv.org/abs/2401.08358.
[172] Unknown. arXiv: 1609.02946. URL: https://arxiv.org/abs/1609.02946.
[173] Unknown. arXiv: 1908.10404. URL: https://arxiv.org/abs/1908.10404.
[174] Unknown. arXiv: 1804.04346. URL: https://arxiv.org/abs/1804.04346.
[175] Unknown. arXiv: 2008.00793. URL: https://arxiv.org/abs/2008.00793.
[176] Unknown. arXiv: 1911.05784. URL: https://arxiv.org/abs/1911.05784.
[177] Unknown. arXiv: 2401.07324. URL: https://arxiv.org/abs/2401.07324.
[178] Unknown. arXiv: 2601.02371. URL: https://arxiv.org/abs/2601.02371.
[179] Unknown. arXiv: 2507.08270. URL: https://arxiv.org/abs/2507.08270.
[180] Unknown. arXiv: 2210.03629. URL: https://arxiv.org/abs/2210.03629.
[181] Unknown. arXiv: 2312.10003. URL: https://arxiv.org/abs/2312.10003.
[182] Unknown. arXiv: 1106.0244. URL: https://arxiv.org/abs/1106.0244.
[183] Unknown. arXiv: 1603.09495. URL: https://arxiv.org/abs/1603.09495.
[184] Unknown. arXiv: 1301.2678. URL: https://arxiv.org/abs/1301.2678.
[185] Unknown. arXiv: 1105.5803. URL: https://arxiv.org/abs/1105.5803.
[186] Unknown. arXiv: 1501.05120. URL: https://arxiv.org/abs/1501.05120.
[187] Unknown. arXiv: 2002.02780. URL: https://arxiv.org/abs/2002.02780.
[188] Unknown. arXiv: 2109.05649. URL: https://arxiv.org/abs/2109.05649.
[189] Unknown. arXiv: 1711.05296. URL: https://arxiv.org/abs/1711.05296.
[190] Unknown. arXiv: 1201.0231. URL: https://arxiv.org/abs/1201.0231.
[191] Unknown. arXiv: 1406.0905. URL: https://arxiv.org/abs/1406.0905.
[192] Unknown. arXiv: 2010.00330. URL: https://arxiv.org/abs/2010.00330.
[193] Unknown. arXiv: 1310.6299. URL: https://arxiv.org/abs/1310.6299.
[194] Unknown. arXiv: 2112.07873. URL: https://arxiv.org/abs/2112.07873.
[195] Unknown. arXiv: 0812.0564. URL: https://arxiv.org/abs/0812.0564.
[196] Unknown. arXiv: 1811.06143. URL: https://arxiv.org/abs/1811.06143.
[197] Unknown. arXiv: 0708.2173. URL: https://arxiv.org/abs/0708.2173.
[198] Unknown. arXiv: 0906.2549. URL: https://arxiv.org/abs/0906.2549.

[199] Unknown. arXiv: 2108.09070. URL: https://arxiv.org/abs/2108.09070.
[200] Unknown. arXiv: 1006.1429. URL: https://arxiv.org/abs/1006.1429.
[201] Unknown. arXiv: 1810.04599. URL: https://arxiv.org/abs/1810.04599.
[202] Unknown. arXiv: 2006.08589. URL: https://arxiv.org/abs/2006.08589.
[203] Unknown. arXiv: 2012.09435. URL: https://arxiv.org/abs/2012.09435.
[204] Unknown. arXiv: 2006.12117. URL: https://arxiv.org/abs/2006.12117.
[205] Unknown. arXiv: 2204.14245. URL: https://arxiv.org/abs/2204.14245.
[206] Unknown. arXiv: 2210.08973. URL: https://arxiv.org/abs/2210.08973.
[207] Unknown. arXiv: 1806.06452. URL: https://arxiv.org/abs/1806.06452.
[208] Unknown. arXiv: 1010.4891. URL: https://arxiv.org/abs/1010.4891.
[209] Unknown. arXiv: 1504.02616. URL: https://arxiv.org/abs/1504.02616.
[210] Unknown. arXiv: 1611.04596. URL: https://arxiv.org/abs/1611.04596.
[211] Unknown. arXiv: 1409.7935. URL: https://arxiv.org/abs/1409.7935.
[212] Unknown. arXiv: 1601.05973. URL: https://arxiv.org/abs/1601.05973.
[213] Unknown. arXiv: 1406.2692. URL: https://arxiv.org/abs/1406.2692.
[214] Unknown. arXiv: 1511.02919. URL: https://arxiv.org/abs/1511.02919.
[215] Unknown. arXiv: 2107.08720. URL: https://arxiv.org/abs/2107.08720.
[216] Unknown. arXiv: 2007.03177. URL: https://arxiv.org/abs/2007.03177.
[217] Unknown. arXiv: 1907.06772. URL: https://arxiv.org/abs/1907.06772.
[218] Unknown. arXiv: 1710.08880. URL: https://arxiv.org/abs/1710.08880.
[219] Unknown. arXiv: 2208.11695. URL: https://arxiv.org/abs/2208.11695.
[220] Unknown. arXiv: 2009.11483. URL: https://arxiv.org/abs/2009.11483.
[221] Unknown. arXiv: 2006.05557. URL: https://arxiv.org/abs/2006.05557.
[222] Unknown. arXiv: 1803.05046. URL: https://arxiv.org/abs/1803.05046.
[223] Unknown. arXiv: 2201.06455. URL: https://arxiv.org/abs/2201.06455.
[224] Unknown. arXiv: 1906.11932. URL: https://arxiv.org/abs/1906.11932.
[225] Unknown. arXiv: 2102.04221. URL: https://arxiv.org/abs/2102.04221.
[226] Unknown. arXiv: 1912.02675. URL: https://arxiv.org/abs/1912.02675.
[227] Unknown. arXiv: 2102.00625. URL: https://arxiv.org/abs/2102.00625.
[228] Unknown. arXiv: 2112.11471. URL: https://arxiv.org/abs/2112.11471.
[229] Unknown. arXiv: 2309.12368. URL: https://arxiv.org/abs/2309.12368.
[230] Unknown. arXiv: 1511.05078. URL: https://arxiv.org/abs/1511.05078.
[231] Unknown. arXiv: 1706.03449. URL: https://arxiv.org/abs/1706.03449.
[232] Unknown. arXiv: 2310.13002. URL: https://arxiv.org/abs/2310.13002.
[233] Unknown. arXiv: 2104.03057. URL: https://arxiv.org/abs/2104.03057.
[234] Unknown. arXiv: 0807.1560. URL: https://arxiv.org/abs/0807.1560.
[235] Unknown. arXiv: 1209.0781. URL: https://arxiv.org/abs/1209.0781.
[236] Unknown. arXiv: 1303.5870. URL: https://arxiv.org/abs/1303.5870.
[237] Unknown. arXiv: 1602.05314. URL: https://arxiv.org/abs/1602.05314.
[238] Unknown. arXiv: 2311.10779. URL: https://arxiv.org/abs/2311.10779.
[239] Unknown. arXiv: 2311.03778. URL: https://arxiv.org/abs/2311.03778.
[240] Unknown. arXiv: 2402.05140. URL: https://arxiv.org/abs/2402.05140.
[241] Unknown. arXiv: 2403.12151. URL: https://arxiv.org/abs/2403.12151.
[242] Unknown. arXiv: 1809.10054. URL: https://arxiv.org/abs/1809.10054.
[243] Unknown. arXiv: 2305.14627. URL: https://arxiv.org/abs/2305.14627.
[244] Unknown. arXiv: 1704.06619. URL: https://arxiv.org/abs/1704.06619.
[245] Unknown. arXiv: 1402.3783. URL: https://arxiv.org/abs/1402.3783.
[246] Unknown. arXiv: 1504.02218. URL: https://arxiv.org/abs/1504.02218.
[247] Unknown. arXiv: 2106.13614. URL: https://arxiv.org/abs/2106.13614.
[248] Unknown. arXiv: 2404.12065. URL: https://arxiv.org/abs/2404.12065.
[249] Unknown. arXiv: 2201.02995. URL: https://arxiv.org/abs/2201.02995.
[250] Unknown. arXiv: 2011.01103. URL: https://arxiv.org/abs/2011.01103.
[251] Unknown. arXiv: 2308.09729. URL: https://arxiv.org/abs/2308.09729.
[252] Unknown. arXiv: 1808.09602. URL: https://arxiv.org/abs/1808.09602.
[253] Unknown. arXiv: 2109.10453. URL: https://arxiv.org/abs/2109.10453.
[254] Unknown. arXiv: 1801.09121. URL: https://arxiv.org/abs/1801.09121.
[255] Unknown. arXiv: 1910.08435. URL: https://arxiv.org/abs/1910.08435.
[256] Unknown. arXiv: 2111.00732. URL: https://arxiv.org/abs/2111.00732.
[257] Unknown. arXiv: 2012.13529. URL: https://arxiv.org/abs/2012.13529.
[258] Unknown. arXiv: 2407.20183. URL: https://arxiv.org/abs/2407.20183.
[259] Unknown. arXiv: 2202.05786. URL: https://arxiv.org/abs/2202.05786.
[260] Unknown. arXiv: 2107.10670. URL: https://arxiv.org/abs/2107.10670.
[261] Unknown. arXiv: 2202.11311. URL: https://arxiv.org/abs/2202.11311.
[262] Unknown. arXiv: 2006.10389. URL: https://arxiv.org/abs/2006.10389.
[263] Unknown. arXiv: 2103.16289. URL: https://arxiv.org/abs/2103.16289.
[264] Unknown. arXiv: 2302.02662. URL: https://arxiv.org/abs/2302.02662.
[265] Unknown. arXiv: 2111.04333. URL: https://arxiv.org/abs/2111.04333.
[266] Unknown. arXiv: 2010.11930. URL: https://arxiv.org/abs/2010.11930.
[267] Unknown. arXiv: 2003.02320. URL: https://arxiv.org/abs/2003.02320.
[268] Unknown. arXiv: 1805.02262. URL: https://arxiv.org/abs/1805.02262.
[269] Unknown. arXiv: 1404.3757. URL: https://arxiv.org/abs/1404.3757.
[270] J. Vanhatalo, M. Hartmann, and L. Veneranta. "Additive Multivariate Gaussian Processes for Joint Species Distribution Modeling with Heterogeneous Data". In: *Bayesian Analysis* (2018). DOI: 10.1214/19-BA1158.
[271] J. A. Weyn, D. Durran, R. Caruana, and N. Cresswell-Clay. "Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models". In: *Journal of Advances in Modeling Earth Systems* (2021). DOI: 10.1029/2021MS002502.
[272] J. Xie et al. *TravelPlanner: A Benchmark for Real-World Planning with Language Agents*. 2024. arXiv: 2402.01622. URL: https://arxiv.org/abs/2402.01622.
[273] Q. Zeng et al. "Perceive, Reflect, and Plan: Designing LLM Agent for Goal-Directed City Navigation without Instructions". In: *arXiv.org* (2024). DOI: 10.48550/arXiv.2408.04168.
[274] W. Zhang and P. Li. "Spike-Train Level Backpropagation for Training Deep Recurrent Spiking Neural Networks". In: *Neural Information Processing Systems* (2019). DOI: 10.3389/fnins.2020.00143.
[275] M. Zhou, L. Hannah, D. Dunson, and L. Carin. "Beta-Negative Binomial Process and Poisson Factor Analysis". In: *International Conference on Artificial Intelligence and Statistics* (2011). DOI: 10.1214/17-ba1070.
[276] Z. Zhou et al. "An experimental test of the geodesic rule proposition for the noncyclic geometric phase". In: *Science Advances* (2019). DOI: 10.1126/sciadv.aay8345.
[277] A. Zia and I. Essa. "Automated surgical skill assessment in RMIS training". In: *International Journal of Computer Assisted Radiology and Surgery* (2017). DOI: 10.1007/s11548-018-1735-5.

Appendix: Diagrams

The following diagrams are included as submission-ready figures for the hackathon visual artifact slot and as reference figures for the whitepaper body.

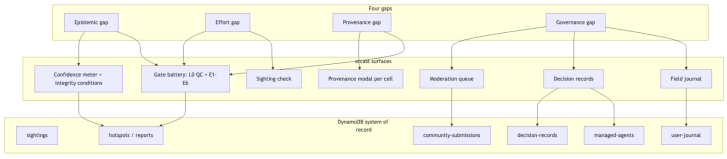

Figure 1: **The four gaps and the orcast bridge.** Four compounding failure modes in wildlife encounter forecasting (epistemic gap, effort gap, provenance gap, governance gap) and the orcast surfaces that address each. Each gap has at least one arrow to an operational surface; each surface writes to one or more DynamoDB tables in the system of record (six representative tables shown). The architecture enforces the property that no confidence-bearing surface exists without a provenance chain and no community submission influences the model without a moderation record.

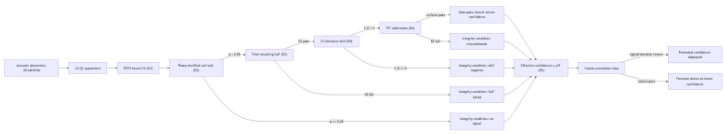

Figure 2: **Gate pipeline from raw detections to displayed confidence.** The full leveled gate battery from Level 0 QC quarantine through PSTH kernel fit (E1), phase-shuffled null test (E2), time-rescaling goodness-of-fit (E3), held-out deviance skill (E5), and PIT calibration (E6). Failing gates produce integrity conditions rather than suppressing the forecast. The human promotion step is last; effective confidence (E4) is bounded by all gate outcomes. All gate steps are present; promotion is the final node before displayed confidence.

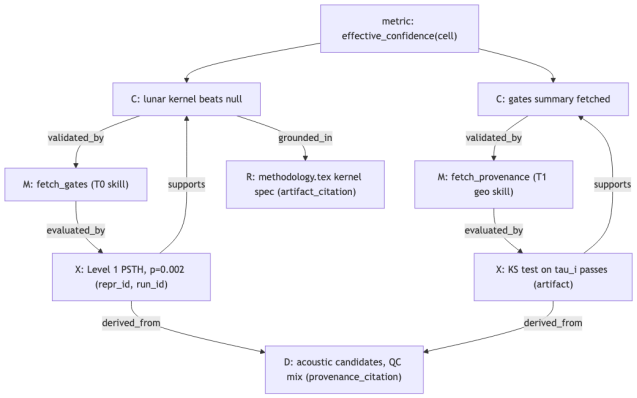

Figure 3: **Claim/method/experiment provenance graph (E8).** A metric node (effective confidence of a cell) traces through two claim nodes (lunar kernel beats null; gates summary fetched) to their producing method nodes (fetch_gates T0 skill; fetch_provenance T1 geo skill), to experiment nodes (Level 1 PSTH with artifact repr_id/run_id; KS test on τ_i), and to data and research leaf nodes. All typed edges (validated_by, evaluated_by, supports, derived_from, grounded_in) match the contract in the methodology. This graph is rendered live by the ProvenanceGraph component from the persisted step log.

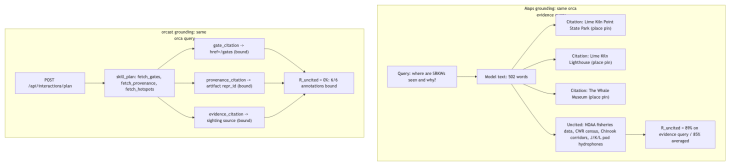

Figure 4: **Grounding contrast: Google Maps vs orcast on the same orca evidence query.** Left panel (orcast): three annotation types (gate_citation, provenance_citation, evidence_citation), 6 of 6 annotations bound to artifacts or hrefs; $R_{\text{uncited}} = 0\%$. Right panel (Maps grounding): the same query produces 502 words across 9 scientific sentences citing three place pins; NOAA fisheries data, Center for Whale Research census, Chinook salmon corridors, and J/K/L pod hydrophone data are uncited (8 of 9 scientific sentences); $R_{\text{uncited}} = 89\%$ on the evidence query, 85% averaged (benchmark run 2026-06-24, Gemini 3.5 Flash, Api-Revision 2026-05-20).

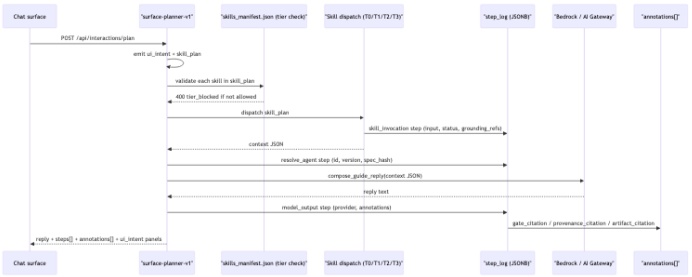

Figure 5: **Managed agent pipeline: plan, manifest check, dispatch, step log, annotations.** Sequence diagram of the surface planner interaction. The chat surface sends a request; surface-planner-v1 emits a ui_intent and skill_plan; each skill is validated against skills_manifest.json before dispatch (400 tier_blocked if not allowed); skill invocations write step log entries with input, status, and grounding references; the LLM narrates the collected context; the model_output step records annotations (gate_citation, provenance_citation, artifact_citation); the full reply with steps, annotations, and ui_intent panels returns to the chat surface.

Multi-panel figures: grounding quality benchmark

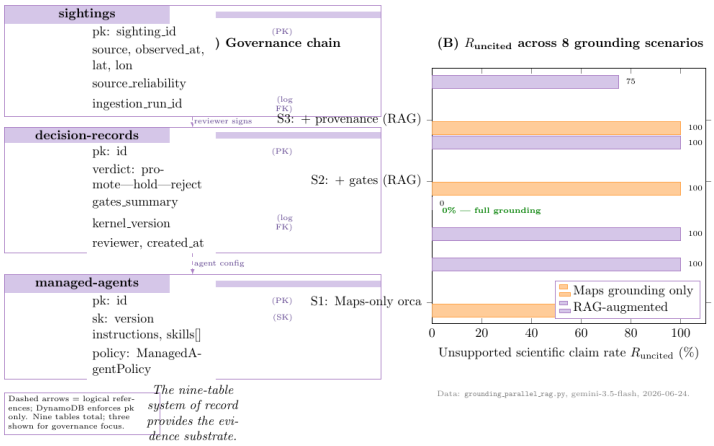

Figure 6: **Evidence-binding failure and its measurement.** (A) The orcast governance chain: sighting records, gate verdicts, and managed agent configurations linked through DynamoDB. Nine tables total; three shown for governance focus. The system of record provides the evidence substrate for grounding quality evaluation. (B) Unsupported scientific claim rate R_{uncited} across eight grounding scenarios. Maps-only baseline: 60–100% uncited. Surface planner step-log (green, Scenario 4): $R_{\text{uncited}} = 0\%$, injecting an orchestrated reasoning chain eliminates uncited claims entirely. Data source `grounding_parallel_rag.py`, gemini-3.5-flash, 2026-06-24.

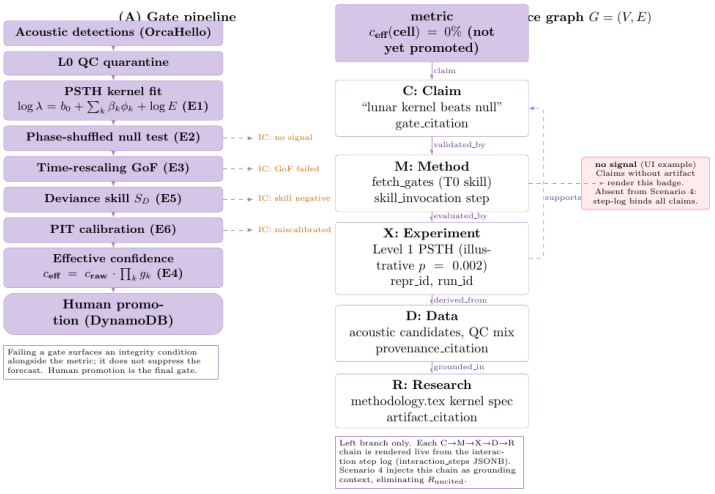

Figure 7: **The mechanism that produces 0% uncited rate.** (A) Gate pipeline: acoustic detections pass through the leveled gate battery (E1–E6; E4 is the composite confidence product, not a separate test) before contributing to displayed confidence. Failing a gate surfaces an integrity condition rather than suppressing the forecast. Human promotion is the final gate. (B) Provenance graph $G = (V, E)$: a metric root connects to Claim (C), Method (M: skill invocation), Experiment (X: artifact repr_id/run_id), Data (D: provenance citation), and Research (R: artifact citation) nodes through the typed edges validated_by (C→M), evaluated_by (M→X), supports (X→C), derived_from (X→D), and grounded_in (C→R). Claims without a bound artifact render the "no signal" badge. Scenario 4 eliminates R_{uncited} because the planner step-log binds every claim to an artifact reference.

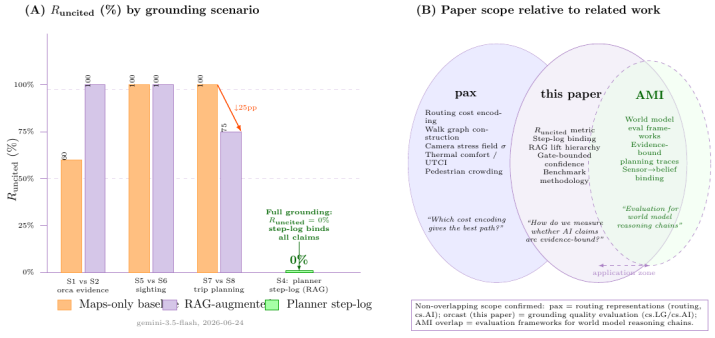

Figure 8: **Grounding quality hierarchy and paper scope.** (A) R_{uncited} for baseline vs RAG-augmented scenarios. The surface planner step-log (green, standalone at 0%) eliminates uncited claims. Hotspot injection (S7→S8) reduces R_{uncited} by 25 percentage points (100% to 75%). Gate context does not reduce R_{uncited}. (B) Paper scope relative to related work. The present paper (center) addresses grounding evaluation methodology, distinct from pedestrian routing cost representations on pax at left and directly applicable to world model evaluation infrastructure at right.