

Trajectory Abstraction for the Science of Language Agent Behavior

Tianqiang Yan*

Faculty of Information Technology, Monash University

Tianqiang.Yan@monash.edu

Abstract

Scientific studies of language agents need behavioral variables that support hypotheses across tasks and models. We formulate this research problem as learning and testing a hierarchy of trajectory abstractions. A concrete recursive procedure first measures role- and phase-indexed events, proposes temporally constrained relations, and tests their stability across conditions. It then constructs episode-level motif variables from selected relations and repeats the analysis on those variables. Explicit measurement functions connect every abstraction level to the original trajectories. Observations and randomized protocol experiments assess the resulting hypotheses, while comparisons between intervention realizations determine whether an abstraction should be retained, refined, or restricted. We derive a finite-depth bound for accepted reductions, identify protocol effects on fixed abstractions, and characterize realization disagreement and composition of abstraction error. A finite-sample test makes projected intervention consistency operational, and constructed examples illustrate motif construction and abstraction refinement. The formulation distinguishes this experimental approach from semantic taxonomies, qualitative theory induction, and behavior-model recovery. It specifies a proposed research procedure for discovering generalizable behavioral hypotheses, with literature-relative novelty assessed separately from model-relative surprise.

1 Introduction

Language agents communicate, invoke tools, reflect on feedback, and coordinate actions. Scientific studies of these behaviors seek explanations that remain useful across tasks and models (Rahwan et al., 2019; Chen et al., 2026). Their observations are trajectories, while their conclusions concern patterns such as checking evidence before action

or completing a repair after detecting an error. The scientific problem is to construct the variables in which these patterns can be stated and tested.

A useful abstraction must preserve the distinctions relevant to the question. Consider two messages: one identifies a retrieved result and the intended action; another says “Okay.” Both can be counted as acknowledgments, yet the shared label can conceal differences relevant to the next action. Testing those differences determines whether to retain the count, distinguish message types, or restrict the hypothesis to one protocol. This is an abstraction-construction problem within the scientific study of behavior.

Recent work already provides behavioral vocabularies, inductive analysis, and formal trace abstraction. Actonomy develops an action taxonomy; AutoTraceGT revises grounded codes and relations; ATLAS learns automata from abstract traces (Gao et al., 2026; Lu et al., 2026b; Lopez-Miguel et al., 2026). ReBeCA and CAMO connect measured factors to structural analysis and experiments (Yan et al., 2026; Yu et al., 2026). These developments motivate a concrete formulation of how trajectory variables, relational hypotheses, and executable tests interact during discovery.

Our position is that higher-level behavioral explanations should be constructed through explicit trajectory functions and tested against the contrasts they claim to preserve. We operationalize this position with a recursive procedure: measure events, learn relation candidates, test cross-condition stability, construct motif variables, and evaluate their intervention responses. An accepted reduction becomes the input to the next level. The output is a hierarchy of measurements and tested hypotheses, including unresolved and rejected proposals.

The paper contributes a critical review, a computational formulation of this procedure, and analytical results supporting its design. Randomized assignments identify protocol effects on fixed ab-

*Corresponding author.

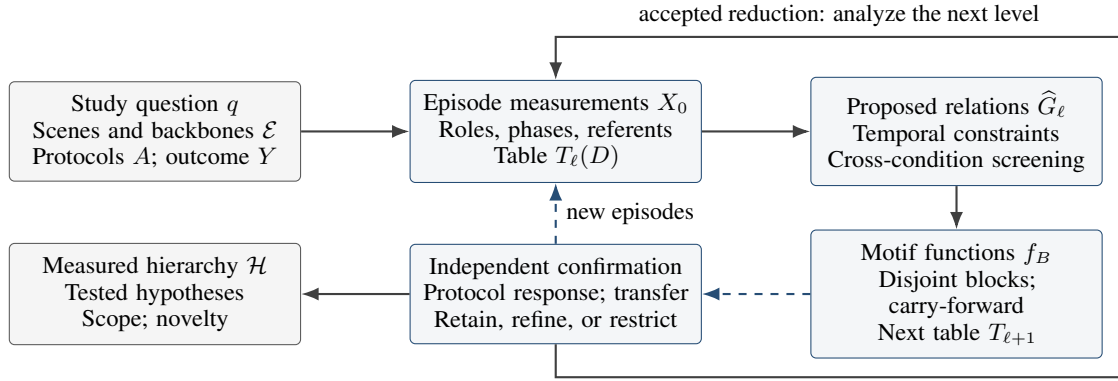


Figure 1: Recursive trajectory abstraction as a scientific research procedure. Episode measurements produce a table at each level. Relations nominate temporal motifs, whose explicit measurement functions define the next table. Independent experiments test these abstractions and their hypotheses across scenes and language-model backbones.

stractions even when the proposed graph is incorrect. Disagreement between realizations gives a criterion for rejecting a common-response claim. A finite-sample comparison implements that criterion, while compression and consistency bounds constrain recursion. The procedure is specified here; empirical implementation and evaluation remain future work.

2 Related Work

We review primary work available through 6 October 2026, using topic searches, reference snowballing, and targeted verification of relevant papers. Appendix A summarizes the selection and comparisons.

2.1 Behavioral Measurement and Induction

ReAct, Reflexion, and LATS expose reasoning, actions, observations, and feedback (Yao et al., 2023; Shinn et al., 2023; Zhou et al., 2024a). ALFWorld, WebArena, SWE-bench, and PARTNR provide repeatable tasks and verifiable outcomes (Shridhar et al., 2021; Zhou et al., 2024b; Jimenez et al., 2024; Chang et al., 2025). AgentBoard analyzes intermediate progress; MAST derives failure categories; AgentDiagnose evaluates trajectory properties using rubrics and visualizations (Ma et al., 2024; Cemri et al., 2025; Ou et al., 2025). These resources supply measurements and experimental settings.

Actonomy organizes agent actions into a reusable three-level taxonomy. AutoTraceGT performs open and axial coding, revises categories and relations, and stops according to a saturation rule (Gao et al., 2026; Lu et al., 2026b). AutoTraceGT therefore already models an inductive study pro-

cess. Our distinction concerns recursive construction of measured motifs from relational hypotheses and confirmation through executable interventions. Category stability and saturation answer different questions from consistency of intervention responses.

Language-based measurements must connect utterances to context and action. Grounding theory, MindCraft, and embodied-dialogue analysis motivate this connection (Clark and Brennan, 1991; Bara et al., 2021; Dongre and Hakkani-Tür, 2026). Interface-changing experiments also distinguish semantic use from interface reliance (Gu et al., 2026). These examples guide the definition of discriminating tests rather than treating semantic labels as explanations by themselves.

2.2 Trajectory and Causal Abstraction

ATLAS categorizes action–observation pairs, normalizes labels, and uses automata learning to recover Markov models; it retains concrete realizations and studies symbolic transfer (Lopez-Miguel et al., 2026). Its recovered transition structure differs from the intervention-preserving response tested here. COAT and IRIS propose variables and evaluate relations from unstructured evidence (Liu et al., 2024; Feng et al., 2025). ReBeCA combines semantic reflection patterns, structure learning, stability checks, and prompt interventions; CAMO constructs factors and performs paired simulator interventions (Yan et al., 2026; Yu et al., 2026). LLawCo tests cooperation laws as policy guidance (Zhou et al., 2026). We build on these concrete operations while making the per-episode map at every recursive level explicit.

Causal abstraction defines consistency of in-

terventions across descriptions (Rubenstein et al., 2017; Beckers and Halpern, 2019; Beckers et al., 2020); causal representation learning concerns the variables themselves (Schölkopf et al., 2021; Geiger et al., 2021). Options, predictive-state representations, and bisimulation supply temporal or predictive criteria (Sutton et al., 1999; Littman et al., 2001; Givan et al., 2003). The present construction selects its criterion according to the behavioral hypothesis and the experiments available to test it.

2.3 Automated Scientific Research

The AI Scientist, Co-Scientist, and AI-Researcher generate and execute research operations with scientific evaluations (Lu et al., 2026a; Gottweis et al., 2026; Tang et al., 2025). AutoDiscovery uses Bayesian surprise to direct hypothesis search (Agarwal et al., 2025). These systems can support the proposed procedure. Its scientific outputs depend on measurements, identification, and confirmation, while surprise ranks belief changes within an explicit model (Itti and Baldi, 2009). Literature-relative novelty requires an independent comparison.

3 Problem Formulation

3.1 Study Setting

A study specifies a question q , environments $\mathcal{E}$, joint protocol assignment A with law P_A , an outcome function g , knowledge K , and budget B . In the categorical instantiation, every scored node has prespecified finite states, including A and the numerical endpoint Y . Counts use fixed bins or truncation. Continuous endpoints require an appropriate likelihood and a fixed finite projection for categorical response tests; their original mean contrasts use integrable numerical endpoints. Environments can be indexed by scene and language-model backbone, $e = (s, m)$. Task, agent composition, interfaces, and persistent memory are fixed or included among the study factors.

An episode has trajectory τ containing timestamped observations, declared plans, messages, tool calls, actions, and execution feedback. Its outcome is $Y = g(\tau)$. Data are

$$D = \{(e_u, A_u, \tau_u, Y_u)\}_{u=1}^N. \quad (1)$$

Assign complete protocol policies randomly before each experimental unit starts. A unit is an episode or a cluster with a prespecified endpoint and weighting; there is no unmodeled interference

or carry-over between units. Background strata are preassignment. Confirmation units are conditionally iid within each tested environment and arm, given completed construction. Successful and failed units are retained.

3.2 Discovery Target

The target is a hierarchy

$$\mathcal{H} = ((X_\ell)_{\ell=0}^L, (a_\ell)_{\ell=0}^{L-1}), \quad (2)$$

where X_ℓ contains measured behavioral variables and a_ℓ constructs the next level. Proposed graphs $\hat{G}_\ell$, selected relations $\mathcal{R}_\ell$, and hypothesis tests are associated with the levels where computed; a terminal graph is optional. All episode, coordinate and response spaces have specified sigma algebras, and all maps are measurable. Every coordinate is an explicit function of an episode. Graph edges themselves are estimates, so they become new variables only through these functions.

A hypothesis specifies a measured relation, outcome contrast, and scope. Descriptive hypotheses concern occurrence; predictive hypotheses concern later outcomes; interventional hypotheses concern feasible manipulations. Results are supported, refuted, or unresolved relative to the tests and assumptions. The search seeks abstractions that reduce unnecessary detail while preserving the contrasts required by q . Generalization is evaluated in held-out scenes, backbones, or their combinations; novelty is assessed against K .

4 Recursive Trajectory Abstraction

4.1 Event Measurements

A measurement map ϕ_0 yields $X_0 = \phi_0(\tau)$. Each coordinate has an event family, agent role, temporal phase, support interval, and availability time. Families include communication, declared planning, tool/action execution, and observed context. Phases follow observable protocol boundaries. For predictive hypotheses, fix $t_{\text{pred}} < t_{\text{target}}$ and use coordinates available by t_{pred} ; feedback revealing an already realized target is excluded. Later coordinates support retrospective analysis. Final outcomes are not measurement inputs.

Executable predicates measure tool arguments and environmental events. Semantic predicates specify referents, coding rules, and independently assessed error. Each coordinate retains its categorical state and the support, referent, and eligibility metadata needed by later maps. Graph learning

uses the categorical states; x_B in a construction map includes their associated metadata. Every macro propagates the metadata required for further construction. Missing or ineligible events are recorded explicitly. For example, D^k denotes detection of a discrepancy in phase k , C^{k+1} a commitment referring to that discrepancy, and R^{k+2} an action implementing its correction. The referent links distinguish one repair sequence from three unrelated events.

4.2 Relation Proposal and Stability

At level ℓ , write $W = s(X_\ell)$ for categorical states and set $V_\ell = (A, W, Y)$. A is one joint assignment node; its factors may be dependent. Graphs have no incoming edges to A , no outgoing edges from Y , and no backward temporal edges. Equal-phase relations require an explicit order or retain their orientation ambiguity. Acyclicity is enforced by the search.

One concrete choice is tier-constrained greedy equivalence search (GES) (Chickering, 2002). For categorical measurements, use a prespecified decomposable score such as

$$s(G) = \sum_{e \in \mathcal{E}_{\text{train}}} [-2\ell_e(G) + d_e(G) \log N_e], \quad (3)$$

where ℓ_e is the maximized multinomial log likelihood, d_e the parameter count, and N_e the episode count. GES approximates score minimization; resampling or alternative scores assess sensitivity. This is a relation-proposal step. Causal interpretation requires the corresponding Markov, faithfulness, and confounding assumptions.

On independent screening episodes, compare each target conditional distribution $P_e(W_v | W_S = w_S)$ across environments. Retain a relation when simultaneous confidence bounds on the largest between-environment TV distance are below a prespecified margin δ_S , over conditioning cells with sufficient support. Fix targets, S , margins, and the tested cell family using build data. Use cell-specific sample counts in simultaneous bounds; required cells with insufficient support remain unresolved. Adaptive screening beyond that fixed family is exploratory. Invariant causal prediction supplies identification results under its own assumptions (Peters et al., 2016); a generic stability pass instead nominates relations. Context-aware discovery is another option (Mooij et al., 2020). Retain uncertainty in edge directions. Candidate blocks come from relations connected to Y or a question-specific target;

temporal precedence alone does not establish an effect (Runge et al., 2019).

4.3 Motif Construction

Let $\mathcal{P}_\ell$ contain proposed connected blocks of at least two behavioral coordinates, excluding A, Y . For each block B , select a function f_B from a finite template family specified for the question. Templates include ordered conjunctions, counts, and categorical progress states. A binary repair motif is

$$M_{\text{repair}} = D^k C^{k+1} R^{k+2}, \quad (4)$$

where the product uses binary state components on eligible episodes with the specified order and shared referent. Its value indicates completion of the defined sequence. An alternative categorical template distinguishes detection, commitment, and completed repair. Ineligible or missing cases retain their own category.

The map’s support covers all required events; it becomes available at the last member’s availability time. Assigned protocols and final outcomes are excluded from f_B . A name such as “evidence-linked repair” is attached after fixing the function. An LLM can propose predicates or names, but the resulting variable must remain computable and validated.

For a concrete selection rule, fit outcome predictors on build data before and after each proposed replacement. On screening data, retain the map when its increase in clipped log loss is at most a prespecified margin δ_Y in each environment and its target relations pass the stability criterion. This is a risk comparison between the fitted predictors. Select a disjoint collection $\mathcal{B}_\ell \subseteq \mathcal{P}_\ell$ in increasing description length, breaking ties by fixed coordinate order. Description length can rank maps that meet the same criteria (Grünwald, 2007); it is a compression preference rather than a causal test. Unmatched variables are carried forward. Writing $U_\ell = \{1, \dots, p_\ell\} \setminus \bigcup_{B \in \mathcal{B}_\ell} B$ for their indices,

$$a_\ell(A, x, Y) = (A, x', Y), \quad (5)$$

$$x' = ((f_B(x_B))_{B \in \mathcal{B}_\ell}, x_{U_\ell}).$$

With $T_0(A, \tau, Y) = (A, \phi_0(\tau), Y)$, the recurrence is

$$T_{\ell+1} = a_\ell \circ T_\ell. \quad (6)$$

Here $T_\ell(D)$ means rowwise application to the archive. Thus each next-level row contains measured motifs. In a constructed two-level example,

$X_0 = (D, C, R, H)$ describes discrepancy detection, referent-linked commitment, corrective action, and a subsequent handoff. The first map forms $M = DCR$ and retains H , giving $X_1 = (M, H)$. A second map forms $J = MH$, measured after the handoff and meaning “completed repair followed by handoff.” It is retained only if its specified predictive and intervention contrasts pass the tests. These products define motif occurrence for subsequent relational hypotheses.

4.4 Intervention Validation

A chosen implementation is a complete measurable policy $\kappa_j : h \mapsto \kappa(j, h)$ on available histories, including a specified action when no eligible event occurs: for example, insert a specified message, require verification before action, or change a role protocol. Its identity is assigned before the unit starts; later realized actions can depend on history. Compliance and collateral changes enter the resulting trajectory. The effect

$$\theta_e(j, j_0) = \mathbb{E}_e[Y^{\kappa_j}] - \mathbb{E}_e[Y^{\kappa_{j_0}}] \quad (7)$$

concerns assignment to those protocols. Identifying a downstream behavior effect requires additional assumptions (Hernán and Robins, 2020; Robins et al., 2000).

Let P_e^i denote the joint assignment–trajectory–outcome law under implementation i , with $\omega(i) = j$, holding non-target factors fixed or using pre-assignment strata. The validation map is $\Phi_\ell^+ = c_\ell \circ T_\ell$, where c_ℓ selects the scientific coordinates and outcome from a learning-table row. It excludes arm identifiers and construction-only timestamps or referent IDs. These remain metadata in the learning table T_ℓ . An abstract response model $Q_{\ell,e}^j$ is fitted independently of confirmation data on the same selected measurable response space. Following causal abstraction (Rubenstein et al., 2017; Beckers et al., 2020), compare

$$\eta_\ell = \max_{e,i} d_{TV}((\Phi_\ell^+)_{\#} P_e^i, Q_{\ell,e}^{\omega(i)}) \quad (8)$$

over a specified finite family of environments and implementations. Push-forward denotes applying the map to draws from the distribution. Implementation-specific hypotheses compare their documented protocol; implementation-independent hypotheses compare multiple realizations of the same j (Figure 2).

For a fixed categorical projection, such as (M_B, Y) , let $r \geq 1$ be its number of states and

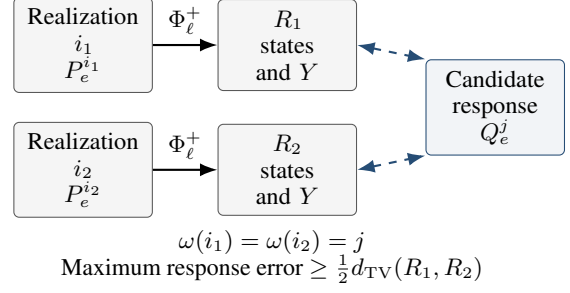


Figure 2: Different realizations of one abstract manipulation must support a common response. Disagreement rejects the common-response claim at the stated tolerance and motivates refinement or scope revision.

$n \geq 1$ iid confirmation units from that law, conditional on construction. The confidence allocation satisfies $0 < \beta < 1$. Coordinate-wise Hoeffding bounds give

$$b(n, r, \beta) = \min \left\{ 1, \frac{r}{2} \sqrt{\frac{\log(2r/\beta)}{2n}} \right\}. \quad (9)$$

With probability at least $1 - \beta$, its empirical distribution $\hat{P}$ satisfies $d_{TV}(P, \hat{P}) \leq b$. Hence $d_{TV}(\hat{P}, Q) + b \leq \epsilon$ certifies the projected tolerance for independently fitted Q . All tests, including rejected proposals and revisions, allocate confidence budgets β_t with $\sum_t \beta_t \leq \beta$. The projection, map, response model, predictors, target laws, tolerances, batch size and allocation are fixed before confirmation; eligibility determined after assignment is retained as a category. Appendix D derives the bounds.

Confirm the jointly selected replacement, since separate block checks do not establish joint preservation. Predictors are fitted before this test; the upper confidence bound on the paired clipped-loss increase must not exceed δ_Y in every specified environment/arm or fixed target mixture. Failed or unresolved criteria trigger additional sampling, refinement, or restriction. Revised functions use fresh fixed-size confirmation batches and new confidence allocations. A layer is accepted only after all required criteria pass. Projected checks do not establish full-trajectory consistency.

4.5 Recursion and Scientific Outputs

Algorithm 1 specifies the construction. Each level uses independent build, screening, and confirmation data. An accepted map becomes the next-level table; lower levels remain available for comparison. The search stops when no reduction meets its

criteria or resources are exhausted. Outputs retain accepted maps and hypotheses, including rejected and unresolved tests.

Render each hypothesis as a measured relation with its contrast and scope, then compare it with K to assess novelty. Condition the prior and likelihood on completed exploration $\mathcal{F}$, including K . For $\pi_h = P(h \mid \mathcal{F})$, independent confirmation defines a posterior π'_h under the fixed hypothesis models. The surprise score $\text{KL}(\text{Ber}(\pi'_h) \parallel \text{Ber}(\pi_h))$ ranks model-relative belief change. Literature novelty is assessed separately; LLM belief elicitation is a heuristic unless the update is justified.

5 Analytical Results

5.1 Randomized Design and Finite Reduction

Proposition 1. If T_ℓ is fixed independently of confirmation data, valid randomized assignment identifies its protocol effects. For a fixed bounded real statistic $V = h(X_\ell(\tau))$, such as the indicator of motif completion,

$$\begin{aligned} \mathbb{E}[V^a - V^{a'} \mid e] \\ = \mathbb{E}[V \mid A = a, e] - \mathbb{E}[V \mid A = a', e]. \end{aligned} \quad (10)$$

This follows from assignment exchangeability, consistency, and positivity. It identifies the assigned protocol's effect on the measured abstraction regardless of graph correctness.

Proposition 2. Let $p_\ell = |X_\ell|$ with $p_0 \geq 1$. Disjoint block replacement with carry-forward gives

$$p_{\ell+1} = p_\ell - \sum_{B \in \mathcal{B}_\ell} (|B| - 1). \quad (11)$$

Every nonempty accepted reduction decreases p_ℓ by at least one. A fixed-inventory reduction chain therefore has at most $p_0 - 1$ transitions. Refinement, splitting, or feature additions outside that chain consume the research budget and are not covered by this bound.

5.2 Realization Disagreement

Proposition 3. If i_1, i_2 realize the same j and R_k is their abstract response distribution, then

$$\begin{aligned} \max_{k \in \{1,2\}} d_{\text{TV}}(R_k, Q^j) \\ \geq \frac{1}{2} d_{\text{TV}}(R_1, R_2). \end{aligned} \quad (12)$$

The triangle inequality proves the bound. It rejects a common-response claim at insufficient tolerance;

possible repairs include refining the block, implementations, response model, or scope.

In a constructed example, X indicates a grounded acknowledgment, W a generic acknowledgment, $Z = X + W$ their count, and $Y = X$ the stipulated outcome. The two protocols

$$\begin{aligned} i_1 &= \text{do}(X = 1, W = 0), \\ i_2 &= \text{do}(X = 0, W = 1) \end{aligned} \quad (13)$$

both realize $j = \text{deliver one acknowledgment}$, with $\omega(i_1) = \omega(i_2) = j$, but produce point masses at $(Z, Y) = (1, 1)$ and $(1, 0)$. Their TV distance is one. A common response at tolerance below $1/2$ is impossible. Distinguishing types requires refining the intervention labels; alternatively the hypothesis can be restricted to one protocol. The example specifies the outcome law; it is not an empirical claim about acknowledgments.

5.3 Composition of Consistency Error

Proposition 4. Fix an environment and preassignment stratum. Let $j_0 = i$ and $j_{\ell+1} = \omega_\ell(j_\ell)$. On specified measurable response spaces, let $R_0^i = (\Phi_0^+)_{\#} P_e^i$ be the actual initial law and $R_\ell^{j_\ell} = Q_{\ell,e}^{j_\ell}$ for $\ell \geq 1$ be specified comparison laws. Suppose measurable outcome-preserving maps satisfy $c_{\ell+1} \circ a_\ell = \bar{a}_\ell \circ c_\ell$ and

$$d_{\text{TV}}((\bar{a}_\ell)_{\#} R_\ell^{j_\ell}, R_{\ell+1}^{j_{\ell+1}}) \leq \epsilon_\ell. \quad (14)$$

Then, with $\Psi_L = \bar{a}_{L-1} \circ \dots \circ \bar{a}_0$,

$$\begin{aligned} d_{\text{TV}}((\Psi_L)_{\#} R_0^i, R_L^{j_L}) \\ \leq \min \left\{ 1, \sum_{\ell=0}^{L-1} \epsilon_\ell \right\}. \end{aligned} \quad (15)$$

TV contraction and induction with the triangle inequality prove the bound. The factorization requires projections retaining sufficient linkage and eligibility summaries. The first tolerance compares actual data to the first model; later tolerances compare specified model laws. The bound concerns the specified adjacent model comparisons. Appendix C gives the proof.

6 Discussion

The formulation specifies how measured motifs become objects of further scientific study. Its practical value should be evaluated against

Algorithm 1 Recursive trajectory abstraction

Require: Study settings; ϕ_0 ; templates and tolerances

```
1:  $T \leftarrow T_0$ ; initialize  $\mathcal{H}$  at  $X_0$ ;  $\ell \leftarrow 0$ 
2: while resources permit do
3:   Acquire independent build and screening data
4:   Fit  $\widehat{G}_\ell$ ; screen relations; store test results
5:   Select disjoint blocks and explicit  $f_B$ 
6:   if no qualifying reduction then break
7:   Form  $a_\ell$  with carry-forward;  $T' \leftarrow a_\ell \circ T$ 
8:   Acquire independent confirmation episodes
9:   Test and record joint predictive loss, responses, and transfer
10:  if required criteria fail or remain unresolved then
11:    Revise  $a_\ell, T'$  or intervention scope
12:    if resources insufficient then break
13:    Acquire fresh fixed-size batch; allocate  $\beta_t$ ; retest
14:    if criteria still unmet then break
15:  end if
16:  Append accepted maps and hypotheses to  $\mathcal{H}$ 
17:   $T \leftarrow T'$ ;  $\ell \leftarrow \ell + 1$ 
18: end while
19: Assess literature novelty and modeled belief change
20: return  $\mathcal{H}$  and tested hypotheses
```

semantic-category induction, automata abstraction, and prediction-driven compression. Matched-budget comparisons can assess independently rejected hypotheses, appropriate unresolved conclusions, calibration, and transfer across scenes and backbones. High-dimensional conditioning and projected-response tests can require many episodes, motivating sparse templates, better experimental design (Myung et al., 2013), and partial identification. Flat semantic coding is sufficient when a question concerns event prevalence or when no validated reduction is available. A hierarchy is useful when a measured motif supports a new relational hypothesis with fewer variables. Descriptive studies need observations, predictive studies need independent future-outcome tests, and causal hypotheses need identified interventions. Compression can remove precisely the distinction that matters to a hypothesis, so retaining a lower-level representation is a scientific outcome. The number of levels is therefore not a quality measure.

7 Conclusion

We formulate scientific discovery of language-agent behavior through a concrete recursive abstraction procedure. Explicit episode functions construct motifs from proposed relations; inde-

pendent tests determine their retention, refinement, and scope. Randomized protocols anchor scientific contrasts, while realization disagreement and consistency bounds constrain the hierarchy. The formulation provides inputs, outputs, construction rules, validation tests, and stopping conditions for subsequent empirical research.

Limitations

The method is proposed and analyzed, with a constructed example; it has not been implemented or empirically evaluated here. The narrative review is bounded by its publication cutoff and selection criteria. Template selection, semantic coding, phase alignment, and intervention feasibility remain study-specific.

GES and invariance procedures require model assumptions that may fail for language-agent trajectories or after coarsening. Hidden state, persistent memory, selection, and incomplete logs can affect identification. Stability and disjoint source coordinates do not establish causality or faithfulness. Protocol effects do not automatically identify the effects of downstream behavioral variables.

The finite-sample bound assumes fixed categorical projections and independent confirmation units; it can be conservative. Projected tests support

only their measured contrasts. The finite-depth result concerns accepted fixed-inventory reductions, and the composition result requires matched layer-wise distributions. Neither establishes scientific completeness or inevitable improvement with depth. Generalization beyond tested conditions and literature-relative novelty require additional evidence.

Ethical Considerations

Trajectory studies can expose private messages and tool outputs. Applications should minimize retained data, document consent and access where relevant, and prevent experimental protocols from causing unauthorized actions. Simulator findings require separate evidence for deployment and human behavior. The example is constructed and contains no human participant data. Human authors remain responsible for scientific claims and automated assistance.

References

- Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, and Peter Clark. 2025. [AutoDiscovery: Open-ended Scientific Discovery via Bayesian Surprise](#). In *Advances in Neural Information Processing Systems*.
- Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. [MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. 2020. [Approximate causal abstractions](#). In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. PMLR.
- Sander Beckers and Joseph Y. Halpern. 2019. [Abstracting causal models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685.
- Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. [Why Do Multi-Agent LLM Systems Fail?](#) In *Advances in Neural Information Processing Systems*.
- Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Motlaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. 2025. [PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks](#). In *International Conference on Learning Representations*.
- Lin Chen, Yunke Zhang, Jie Feng, Haoye Chai, Honglin Zhang, Bingbing Fan, Yibo Ma, Shiyuan Zhang, Nian Li, Tianhui Liu, Nicholas Sukiennik, Keyu Zhao, Yu Li, Ziyi Liu, Fengli Xu, and Yong Li. 2026. [AI agent behavioral science](#). *Humanities and Social Sciences Communications*, 13(1).
- David Maxwell Chickering. 2002. [Optimal structure identification with greedy search](#). *Journal of Machine Learning Research*, 3:507–554.
- Herbert H. Clark and Susan E. Brennan. 1991. [Grounding in communication](#). In *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.
- Vardhan Dongre and Dilek Hakkani-Tür. 2026. [Embodied Multi-Agent Coordination by Aligning World Models Through Dialogue](#). *Preprint*, arXiv:2605.12920.
- Tao Feng, Lizhen Qu, Niket Tandon, and Gholamreza Haffari. 2025. [IRIS: An iterative and integrated framework for verifiable causal discovery in the absence of tabular data](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9400–9428, Vienna, Austria. Association for Computational Linguistics.
- Jie Gao, Kaiser Sun, Jen-tse Huang, Katherine Van Kovering, Sijie Ji, Heyuan Huang, Weiyan Shi, Zhuoran Lu, Ziang Xiao, Daniel Khashabi, and Mark Dredze. 2026. [How to interpret agent behavior](#). Preprint, arXiv:2605.13625v1.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. [Causal abstractions of neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34.
- Robert Givan, Thomas Dean, and Matthew Greig. 2003. [Equivalence notions and model minimization in markov decision processes](#). *Artificial Intelligence*, 147(1–2):163–223.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, and 32 others. 2026. [Accelerating scientific discovery with Co-Scientist](#). *Nature*, 655(8122):487–496.

- Peter D. Grünwald. 2007. *The Minimum Description Length Principle*. MIT Press.
- Weizheng Gu, Chengze Li, Zhuohao Yu, Mengyuan Sun, Zhibang Yang, Wei Wang, Hongrui Jia, Shikun Zhang, and Wei Ye. 2026. [What do agents learn from trajectory-SFT: Semantics or interfaces?](#) In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 37073–37109. PMLR.
- Miguel A. Hernán and James M. Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton.
- Wassily Hoeffding. 1963. [Probability inequalities for sums of bounded random variables](#). *Journal of the American Statistical Association*, 58(301):13–30.
- Laurent Itti and Pierre Baldi. 2009. [Bayesian surprise attracts human attention](#). *Vision Research*, 49(10):1295–1306.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. [SWE-bench: Can Language Models Resolve Real-World GitHub Issues?](#) In *International Conference on Learning Representations*.
- Michael L. Littman, Richard S. Sutton, and Satinder Singh. 2001. [Predictive representations of state](#). In *Advances in Neural Information Processing Systems*, volume 14.
- Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. 2024. [Discovery of the Hidden World with Large Language Models](#). In *Advances in Neural Information Processing Systems*.
- Ignacio D. Lopez-Miguel, Andreas Happe, Jürgen Cito, Ezio Bartocci, Bettina Könighofer, and Martin Tappler. 2026. [ATLAS: Discovering agent strategies through LLM-guided abstraction and automata learning](#). arXiv:2608.14352v1. Accepted at ACM/IEEE MODELS 2026; author version.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026a. [Towards end-to-end automation of AI research](#). *Nature*, 651(8107):914–919.
- Zhuoran Lu, Yangyang Yu, Zhuoyan Li, Yibo Meng, Nan Jiang, Chengxi Zang, Jie Gao, and Ziang Xiao. 2026b. [Using grounded theory for agent behavior analysis at scale](#). arXiv:2608.30391v1. Accepted to Findings of EMNLP 2026.
- Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. [AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents](#). In *Advances in Neural Information Processing Systems*.
- Joris M. Mooij, Sara Magliacane, and Tom Claassen. 2020. [Joint causal inference from multiple contexts](#). *Journal of Machine Learning Research*, 21(99):1–108.
- Jay I. Myung, Daniel R. Cavagnaro, and Mark A. Pitt. 2013. [A tutorial on adaptive design optimization](#). *Journal of Mathematical Psychology*, 57(3–4):53–67.
- Tianyue Ou, Wanyao Guo, Apurva Gandhi, Graham Neubig, and Xiang Yue. 2025. [AgentDiagnose: An open toolkit for diagnosing LLM agent trajectories](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 207–215. Association for Computational Linguistics.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. [Causal inference by using invariant prediction: Identification and confidence intervals](#). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.
- Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex “Sandy” Pentland, and 4 others. 2019. [Machine behaviour](#). *Nature*, 568(7753):477–486.
- James M. Robins, Miguel Ángel Hernán, and Babette Brumback. 2000. [Marginal structural models and causal inference in epidemiology](#). *Epidemiology*, 11(5):550–560.
- Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2017. [Causal consistency of structural equation models](#). In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*.
- Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019. [Detecting and quantifying causal associations in large nonlinear time series datasets](#). *Science Advances*, 5(11):eaau4996.
- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. [Toward causal representation learning](#). *Proceedings of the IEEE*, 109(5):612–634.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language Agents with Verbal Reinforcement Learning](#). In *Advances in Neural Information Processing Systems*.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. [ALFWorld: Aligning Text and](#)

[Embodied Environments for Interactive Learning](#). In *International Conference on Learning Representations*.

Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. [Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning](#). *Artificial Intelligence*, 112(1–2):181–211.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025. [AI-Researcher: Autonomous Scientific Innovation](#). In *Advances in Neural Information Processing Systems*.

Tianqiang Yan, Lizhen Qu, Sihan Shang, Yuheng Li, Song Qiu, Hao Peng, Wenjian Luo, Jue Xie, and Yuan Gao. 2026. [ReBeCA: Unveiling interpretable behavior hierarchy behind the iterative self-reflection of language models with causal analysis](#). *Preprint*, arXiv:2602.06373. Accepted to Findings of the Association for Computational Linguistics: EMNLP 2026; arXiv version 2.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing Reasoning and Acting in Language Models](#). In *International Conference on Learning Representations*.

Xiangning Yu, Yuwei Guo, Yuqi Hou, Xiao Xue, and Qun Ma. 2026. [CAMO: An Agentic Framework for Automated Causal Discovery from Micro Behaviors to Macro Emergence in LLM Agent Simulations](#). *Preprint*, arXiv:2604.14691.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2024a. [Language Agent Tree Search Unifies Reasoning Acting and Planning in Language Models](#). In *Proceedings of the International Conference on Machine Learning*.

Qinhong Zhou, Chuang Gan, and Anoop Cherian. 2026. [LLawCo: Learning Laws of Cooperation for Modeling Embodied Multi-Agent Behavior](#). *Preprint*, arXiv:2606.28182.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024b. [WebArena: A Realistic Web Environment for Building Autonomous Agents](#). In *International Conference on Learning Representations*.

A Review Scope and Comparisons

The review covers trajectory measurement, inductive behavioral analysis, causal abstraction, and automated scientific research. Searches of primary publication and preprint sources were complemented by backward and forward reference searches and verification of specific related works. Selection favored studies specifying measurements, construction procedures, experimental contrasts,

or formal assumptions relevant to language-agent trajectories. The publication cutoff is 6 October 2026; topic-based searches on 7 October checked coverage against that cutoff. We distinguish peer-reviewed publications from author versions and preprints. The synthesis is a narrative review rather than an exhaustive systematic review.

B Measurement Specification

A coordinate is specified by its event predicate, role, phase, referent, support interval, availability time, codomain, and missingness rule. For the repair example, detection requires an observable discrepancy statement; commitment must identify that discrepancy and an intended correction; repair requires execution feedback matching the correction. A label inferred from an utterance is a measurement of expressed content. It does not identify latent intention.

An ordered conjunction requires the declared temporal order and shared referent. Let $\chi_B(x_B)$ indicate eligibility. Eligibility, temporal order, and shared referents are determined from the coordinate metadata retained at the current level. Write $w_v = s_v(x_v)$ for each state component. The macro state is

$$m_B(x_B) = \begin{cases} \prod_{v \in B} w_v, & \chi_B(x_B) = 1, \\ \perp, & \chi_B(x_B) = 0. \end{cases} \quad (16)$$

This equation defines the macro state. The complete record is $f_B(x_B) = (m_B(x_B), k_B(x_B))$, where k_B retains the support, referent and eligibility summaries needed by subsequent maps. The symbol $\perp$ records ineligibility; missing observation can be another category. Successor maps access only the current coordinate records. Products and graph scores use state components; the projection used for a finite-state test excludes identifiers and timestamps. Prospective coordinates access only their available histories. Training may use outcome supervision to select a function, but the frozen function takes neither final outcome nor assigned factor as an input. The next learning table preserves those fields separately. Validation compares behavior and outcome coordinates, with non-target factors fixed or stratified; realization identifiers remain outside this response distribution.

Categorical alternatives can distinguish no detection, detection without commitment, commitment without completed correction, completed repair, and ineligibility. Choice of a binary or categorical

Work	Relevant construction	Relation to this formulation
AutoTraceGT	Grounded coding, category/relation revision, saturation, theory generation.	Already specifies a study process; our target adds recursive measured motifs and protocol-response tests.
Actonomy	Hierarchical action vocabulary and trajectory coding.	Supplies event categories; hypothesis-specific experimental distinctions determine reductions.
ATLAS	Trace abstraction, concretization, automata learning, symbolic transfer.	Recovers transitions; our target is motif construction for tested behavioral hypotheses.
ReBeCA	Semantic patterns, graph learning, stability checks, prompt interventions.	Closely related operations; our formulation makes recursive maps and realization dependence explicit.
CAMO	Factor construction and paired simulator interventions.	Provides experimental factor analysis; our maps define next-level episode measurements.
COAT / IRIS	Variable proposal and relation discovery from unstructured evidence.	Motivates construction; our recursion preserves availability and tests feasible realizations.

Table 1: Selected trajectory-analysis methods and their relation to the proposed scientific target.

map is a testable modeling decision: it changes both compression and the distinctions that later hypotheses can express. If role or phase boundaries differ across scenes, alignment is defined before comparison or the hypothesis scope is restricted.

On observed eligible phase windows, the constructed two-level example gives the state components

$$\begin{aligned}
X_0 &= (D, C, R, H), \\
X_1 &= (DCR, H) = (M, H), \\
X_2 &= MH = J.
\end{aligned} \tag{17}$$

For instance, $(D, C, R, H) = (1, 1, 0, 1)$ yields $(M, H) = (0, 1)$ and $J = 0$; $(1, 1, 1, 1)$ yields $(1, 1)$ and $J = 1$. These values demonstrate the maps.

C Proofs and Identification

C.1 Randomized Assignment

Condition on completed construction, including the fixed measurement function and the unit-level design assumptions in Section 3. Suppose assignment A is randomized within environment e , with positive probability for each compared arm. For a fixed bounded statistic $V = h(X_\ell(\tau))$, consistency gives $V = V^a$ when $A = a$, and assignment

exchangeability gives $V^a \perp A \mid e$. Therefore

$$\begin{aligned}
\mathbb{E}[V \mid A = a, e] &= \mathbb{E}[V^a \mid A = a, e] \\
&= \mathbb{E}[V^a \mid e].
\end{aligned} \tag{18}$$

Subtracting the corresponding identity for a' proves Proposition 1. Randomization identifies assignment effects, including noncompliance. A prompt requiring an action is an assigned protocol; replacing it with a perfect intervention on the resulting action requires additional identification assumptions.

Randomization of design factors does not orient every relation between behavioral variables. Let U be a fair binary variable independent of an assigned factor. Models $X = U, Y = U$ and $X = U, Y = X$ have identical observational distributions. Setting X to zero leaves the first outcome random and fixes the second at zero. Thus behavioral causal effects remain unidentified from these observations, despite randomized assignment of an unrelated factor.

C.2 Reduction Depth

For disjoint blocks, replacing $|B|$ coordinates by one removes $|B| - 1$ coordinates. All unmatched coordinates remain, proving the exact recurrence in Proposition 2. Each block has size at least two, so each accepted transition decreases the positive integer p_ℓ . Hence there are at most $p_0 - 1$ transitions

in a fixed-inventory chain. Retaining lower-level measurements permits later scientific comparison without counting them as new coordinates within that chain.

If blocks partition all coordinates and every block has size at least two, then $p_{\ell+1} \leq \lfloor p_\ell/2 \rfloor$, yielding at most $\lfloor \log_2 p_0 \rfloor$ transitions. This stronger bound does not apply to carry-forward reductions. Introducing features, splitting variables, or rolling back restarts the relevant chain; the research budget bounds those operations separately.

C.3 Realization Dependence

With $d_k = d_{\text{TV}}(R_k, Q)$ for a common response Q , the triangle inequality gives

$$\begin{aligned} d_{\text{TV}}(R_1, R_2) &\leq d_1 + d_2 \\ &\leq 2 \max_k d_k. \end{aligned} \quad (19)$$

This proves Proposition 3. The constructed acknowledgment example has disjoint outcome support under the two protocols, so its TV distance equals one. Its repair options are to retain the distinction between grounded and generic messages or to make an implementation-specific claim.

C.4 Composition

Use the fixed environment, response spaces, comparison laws and factorization in Proposition 4. Let Ψ_0 be the identity and $\Psi_\ell = \bar{a}_{\ell-1} \circ \dots \circ \bar{a}_0$. The assumed layer comparison is

$$d_{\text{TV}}((\bar{a}_\ell)_\# R_\ell^{j_\ell}, R_{\ell+1}^{j_{\ell+1}}) \leq \epsilon_\ell. \quad (20)$$

For a measurable map, TV contracts because each output event has a measurable inverse image. Insert $(\bar{a}_\ell)_\# R_\ell^{j_\ell}$ as an intermediate law:

$$\begin{aligned} d_{\text{TV}}((\Psi_{\ell+1})_\# R_0^i, R_{\ell+1}^{j_{\ell+1}}) \\ \leq d_{\text{TV}}((\Psi_\ell)_\# R_0^i, R_\ell^{j_\ell}) + \epsilon_\ell. \end{aligned} \quad (21)$$

The initial discrepancy is zero. Induction proves the sum bound, capped at one. If two records have identical selected states but differ in linkage, the next motif may be one versus $\perp$. A projection discarding that distinction has no induced map; its comparisons remain level-specific. A common abstract label also does not define a unique actual law when realizations disagree. The higher-level R laws here are explicitly chosen comparison models.

An invariance test’s retention event does not automatically establish causal correctness. If a discovery procedure has a justified conditional error

guarantee α_ℓ at each adaptively chosen level, a union bound controls the probability of any corresponding error by $\sum_\ell \alpha_\ell$. Such a guarantee must be proved for that procedure and its post-abstraction assumptions. Independent data alone do not supply it.

D Finite-Sample Response Tests

Condition on all pretest information. Let P be the fixed law of the declared r -state projection, with $r \geq 1$, $n \geq 1$, and $0 < \beta < 1$. For state v , empirical frequency $\hat{p}_v$ from n iid experimental units drawn from P obeys Hoeffding’s inequality (Hoeffding, 1963):

$$P(|\hat{p}_v - p_v| > t) \leq 2e^{-2nt^2}. \quad (22)$$

Taking a union bound over states with $t = \sqrt{\log(2r/\beta)/(2n)}$ yields

$$d_{\text{TV}}(P, \hat{P}) = \frac{1}{2} \sum_v |p_v - \hat{p}_v| \leq b(n, r, \beta) \quad (23)$$

with probability at least $1 - \beta$. Clipping at one is valid because TV never exceeds one.

For predictive confirmation, use two fixed predictors and clipped losses in $[0, C]$, with $C > 0$ fixed before testing. On the same iid units let $\Delta_u = \ell_{\text{new},u} - \ell_{\text{old},u}$ and $\hat{\Delta} = n^{-1} \sum_u \Delta_u$. Since $\Delta_u \in [-C, C]$, Hoeffding bounds the expected increase by

$$\mathbb{E} \Delta_u \leq \hat{\Delta} + C \sqrt{\frac{2 \log(1/\beta_t)}{n}} \quad (24)$$

with probability at least $1 - \beta_t$. Accept only if this upper bound is at most δ_Y , for the joint replacement in every required target law. The criterion concerns the expected clipped-risk contrast of the fixed predictors.

For a response model Q fitted independently of the confirmation batch, the triangle inequality gives

$$d_{\text{TV}}(P, Q) \leq d_{\text{TV}}(\hat{P}, Q) + b(n, r, \beta). \quad (25)$$

For two independently sampled realizations with confidence allocations $\beta_1 + \beta_2 \leq \beta$, put $b_k = b(n_k, r, \beta_k)$. Let $\hat{d} = d_{\text{TV}}(\hat{P}_1, \hat{P}_2)$. With probability at least $1 - \beta$,

$$\begin{aligned} d_{\text{TV}}(P_1, P_2) &\geq \max\{0, \hat{d} - b_1 - b_2\}, \\ d_{\text{TV}}(P_1, P_2) &\leq \min\{1, \hat{d} + b_1 + b_2\}. \end{aligned} \quad (26)$$

If half the lower bound exceeds a proposed common-response tolerance, Proposition 3 rejects

that response scope. If the upper bounds to independently fitted Q fall below the tolerance for all required arms, the projected criterion is met. Otherwise the conclusion is unresolved.

Before attempt t , the completed history $\mathcal{F}_{t-1}$ fixes maps, projections, models, predictors, target laws, tolerances, sample sizes and allocations. Each fresh batch is conditionally iid from its declared law. Split β_t among every tested arm, environment, cell and quantity, or supply a simultaneous guarantee. If failure event E_t^c has conditional probability at most β_t and $\sum_t \beta_t \leq \beta$ almost surely, then

$$\Pr(\exists t : E_t^c) \leq \mathbb{E} \sum_t \beta_t \leq \beta. \quad (27)$$

Every confirmation attempt uses a fresh fixed-size batch. Completed batches enter the construction history for subsequent independently confirmed proposals. For clusters, this simple proof uses one prespecified finite-valued observation per iid cluster. Episode-weighted cluster analyses specify their corresponding sampling bound separately. A post-assignment missing or ineligible state remains in the distribution; removing it can change the protocol comparison. Tests on (M_B, Y) establish that projected comparison, not equality of unrestricted text-trajectory laws.

E Novelty and Future Evaluation

A hypothesis is represented by its measured variables, claimed relation, contrast, and scope. Comparison with existing work distinguishes replication, scope extension, and previously untested relations. The literature search and its date delimit an underexploration claim. Supported known hypotheses remain legitimate scientific findings.

For a selected hypothesis, condition on completed exploration $\mathcal{F}$. Specify exhaustive alternative models $h, \neg h$, a prior $0 < \pi_h = P(h \mid \mathcal{F}) < 1$, and predictive likelihoods $L_h = P(D_c \mid h, \mathcal{F})$ and $L_0 = P(D_c \mid \neg h, \mathcal{F})$ for the independent batch under the fixed design. Composite models integrate nuisance parameters under their specified priors. If $\pi_h L_h + (1 - \pi_h) L_0$ is positive and finite, then

$$\pi'_h = \frac{\pi_h L_h}{\pi_h L_h + (1 - \pi_h) L_0}. \quad (28)$$

Use $0 \log 0 = 0$ for posterior endpoints. The KL score measures change within that model. It can prioritize investigation but does not replace experimental support or literature comparison.

Future evaluation should compare the proposed construction with fixed action taxonomies, grounded-theory induction, automata learning, and prediction-only compression at matched data and experimental budgets. Controlled tasks with stipulated mechanisms can assess abstraction recovery and appropriate rejection. Held-out scenes and backbones test transfer of measured relations and intervention responses. Required analyses include coding error, eligibility rules, role/phase alignment, template choice, graph-score assumptions, independent data partitions, and calibration of unresolved conclusions.