XENOPSYCHOLOGY

CONCEPTUAL WORKING PAPER / AUTHOR ATTRIBUTION R5

Arrival — Understanding the Mind Behind the Language

Before comparing answers, establish which distinctions the question and its representation make available.

XENO-WP-2026-003 · Conceptual manuscript R3
12,126 main-text words · 26 sections

Rob Emerick · Xenopsychology
ORCID iD: 0009-0006-1269-4216

Abstract

Arrival — Understanding the Mind Behind the Language

Question. Before judging an answer, have we established that the question makes the intended distinctions available? This paper uses Arrival's elicitation problem to examine how representation, information, objective specification, and response format can be conflated when evaluating artificial cognition. Its focus is task framing, not the film's speculative claims about language and time.

Approach. The argument distinguishes what a system is given from how it is encoded, what it is asked to optimize, and how it must respond. A constructed directed-route world is rendered as prose, tables, and diagrams from the same underlying record. Worked cases include conflicting objectives, infeasible requests, multiple admissible answers, missing preferences, and visually ambiguous relations.

Conceptual contribution. The paper proposes a representation-equivalence certificate that maps each task-relevant fact to its rendered location in every format. It separates source-data equivalence from perceptual accessibility and from equal task difficulty. Clarification is treated as useful when it resolves a decision-relevant ambiguity, not simply when a system asks more questions. A format-sensitive performance difference is a bounded behavioral finding; it does not uniquely reveal a system's internal architecture or mode of experience.

Research agenda. A staged evaluation tests extraction, constraint interpretation, planning, and checking separately before combining them. Independent solvers define admissible answer sets. Matched-information comparisons, transcribed-image controls, objective changes, and metamorphic transformations help distinguish perception failures from reasoning or task-specification failures. An oracle clarification channel permits assessment of whether a requested detail actually improves the decision. World-level sampling, declared exclusions, and separate human comparisons constrain generalization.

Scope and status. This conceptual working paper presents an unexecuted research program, not results from a multimodal benchmark. The linguistic consultant's account anchors the fictional discussion; selected research supplies methodological context. Constructed numerical examples illustrate distinctions rather than estimate performance. The central proposal is to investigate the assumptions built into an evaluation before interpreting an unfamiliar system's response as evidence of a different kind of mind.

At a glance

CENTRAL QUESTION

What must be established before a familiar-looking question has a well-defined answer?

CORE CLAIM

A difference in answers can begin with a difference in the task as represented. Audit the question before inferring the architecture of the mind.

CONTRIBUTION

A constructed route world separates representation, information, objective, and permitted responses through matched tasks.

SCOPE & STATUS

Conceptual working paper. Route values and example responses are constructed; no model performance is reported.

How to read this edition

Chapter bands divide the argument into four parts. Tinted passages preserve the original wording of worked examples, proposals, and objections. Figures and comparison tables are reading aids, with links to their source sections. Pull quotes repeat selected passages; they are not additional evidence.

Public reading formats: abstract page, full paper on the web, and this PDF. The manuscript is complete; no sections are condensed.

Figures & tables

Contents

1. Before an answer, there is a theory of the question

A question can appear straightforward because its hidden assumptions are familiar to the person asking it. “Which route is best?” presupposes a set of possible routes, a criterion of value, and some account of acceptable trade-offs. “What do they want?” presupposes that the relevant entity has a goal, that the goal can be described at the requested level, and that the available observations distinguish it from alternatives. A fluent response may conceal rather than resolve these assumptions. The first task of interpretation is sometimes to determine what would count as an answer.

Arrival makes that problem central to an imagined encounter. Jessica Coon, the film's linguistics consultant, describes the difficulty of establishing basic linguistic distinctions before asking the visitors a complex question about their presence. Her account is a valuable first-person source about the film's linguistic preparation, not empirical evidence that learning an unfamiliar language changes access to time. We use the film to motivate attention to task framing and representation, while keeping its speculative temporal premise separate from scientific claims. [1]

The distinction is relevant to artificial systems because users can mistake a well-formed prompt for a well-specified task. A system may select a plausible interpretation of “best” without revealing that another interpretation would change the answer. It may accept a category that the task materials do not support. It may perform differently when the same facts are presented as prose, a table, or a diagram. These are testable possibilities. They do not require assuming that the system experiences a fundamentally different reality, and they should not be explained through such an assumption before simpler dependencies are examined.

This paper proposes a research program for studying representation-sensitive task interpretation. Its central distinction is among the information available, the format in which that information is supplied, the objective the task specifies, and the inference the evaluator draws from the response. Those layers can change independently. A task may preserve all facts while changing the format. It may preserve the format while changing the objective. It may leave the objective unresolved. Each manipulation asks a different question, and the resulting behavior should be interpreted at that level.

The proposed experimental setting is a synthetic route-planning world with inspectable constraints and several possible objectives. It allows prose, table, and diagram presentations to be generated from one underlying record. It also permits controlled omission of an objective and explicit opportunities for clarification. No model evaluation or participant study has been conducted for this manuscript. The examples are constructed to make the design concrete. The contribution is a set of distinctions and a reproducible protocol proposal, not a report of findings.

The paper's thesis is that understanding the mind behind the language begins with a more modest discipline: understanding the assumptions behind the task. Differences in output can reveal sensitivity to representation and framing without uniquely identifying internal mechanisms. Stronger claims require stronger evidence. Arrival is useful when it encourages that discipline. It is misleading when its narrative power tempts us to treat a change in representation as proof of an exotic form of consciousness.

2. The fictional premise and the empirical question must be kept apart

A fictional work can combine a plausible methodological problem with an extraordinary speculative premise. The plausibility of the first does not establish the second. In Arrival, the difficulty of learning how an unfamiliar communicative system organizes distinctions is a productive source of questions. The film's treatment of time belongs to its imaginative construction. This paper does not use the story to establish strong linguistic determinism, nonsequential experience, or a general law connecting grammatical form to consciousness.

That boundary does not impoverish the analysis. It clarifies which elements can become research questions. We can ask whether a system identifies an underspecified objective, whether it preserves a relation across formats, and whether it distinguishes a description from a request. We can test whether a diagram helps because it exposes structure or harms because important labels are difficult to read. None of these questions requires concluding that the system's experience of reality has changed. They concern observable performance under controlled informational conditions.

The title's phrase “the mind behind the language” should therefore be read as an invitation to investigate, not a promise of transparent access. Language is evidence about what a system can express and how it responds. It may support hypotheses about representation, but many internal processes can produce similar responses. A correct explanation does not uniquely identify the process that generated it. An unfamiliar expression does not uniquely identify an unfamiliar cognitive architecture. The paper treats these as limits on inference, not reasons to abandon inquiry.

There is also a distinction between the filmmaker's premise, a consultant's commentary, and a scientific consensus. A consultant can explain the ideas that informed a scene while taking positions that remain debated or broader than the evidence we have reviewed. We cite Coon for the specific problem of establishing basic distinctions before asking a complex question. We do not adopt every claim in the commentary as a settled account of all human language. Accurate attribution is especially important when a cultural essay becomes a conceptual research paper.

A productive reading of the film can thus focus on the sequence of inquiry. Before asking for a motive, establish reference. Before assuming a question has been understood, examine how questions are expressed and answered. Before interpreting a response, identify which distinctions the participants have actually grounded. This sequence is our methodological interpretation of the story's relevance. The synthetic evaluation developed below translates that interpretation into tasks with explicit answer conditions and comparison groups.

The result is a disciplined use of fiction. The story supplies a situation in which familiar assumptions become visible by failing. The research program selects a narrower problem that can be operationalized independently. It then asks what observations would support or undermine a proposed explanation. That movement preserves the intellectual force of Arrival while preventing the paper from borrowing scientific authority from a narrative whose purpose is not to function as an experiment.

3. Representation, information, and objective are different variables

Representation is the form in which task-relevant structure is made available. Information is the content that can be recovered from that representation under the task's interpretation rules. An objective specifies what outcome should be preferred. These definitions are working distinctions for the proposed study, not a complete theory of semantics. They are useful because many evaluations change all three at once and then attribute the result to only one. A diagram may contain information omitted from the prose condition, or a “clearer” prompt may quietly specify an objective that the original left open.

Imagine a route from Orin to Vale through Tern. The underlying world contains travel times, energy costs, and a capacity limit. A prose description states each fact in sentences. A table lists edges and costs. A diagram uses arrows and labels. If the diagram omits one capacity limit, it is not simply another format of the same task. It is a different information condition. If the prose asks for the fastest route while the diagram asks for the most efficient route, the objective has changed as well. A valid comparison must identify such differences.

Information equivalence can be operational rather than metaphysical. For the synthetic world, each presentation is generated from the same structured record and includes a checklist of required facts. Independent checks can verify that those facts are present. This does not guarantee that every representation is equally easy to process. Ease of processing is part of what the study may investigate. It does establish that a performance difference is not trivially explained by one condition lacking a decisive fact, provided the verification is accurate.

The objective can be manipulated separately. The same world may support a fastest route, a lowest-energy route, and a route that preserves reserve capacity. A question that asks only for the “best” route is underdetermined unless the context supplies a priority. A system that chooses one objective may be making a reasonable default assumption, but it should not present that assumption as a fact given by the user. The evaluation should distinguish explicit optimization, disclosed defaulting, and useful clarification.

A fourth variable is the response format. Asking for a free-form explanation differs from requiring a path and a short justification in a structured record. A model may solve the planning problem but fail to format its answer, or produce a well-formed record containing an invalid route. These outcomes should be scored separately. Otherwise, an apparent reasoning difference may actually reflect output parsing or interface constraints. The paper's framework treats communication as an arrangement with several measurable boundaries.

These distinctions let the research ask focused questions. Does changing format affect extraction of facts? Does changing objective affect choice appropriately? Does removing the objective increase clarification? Does a structured response reduce unsupported assumptions? A single task can support all these comparisons, but only if the experimental design preserves their separation. The point is not to eliminate complexity. It is to make the complexity legible enough that an observed difference has an interpretable meaning.

TABLE 1 / READING AID

Four variables that should not be silently merged

Variable	In the route world	Keep explicit
Representation	Prose, structured table, or diagram.	The presentation and modality.
Information	Recoverable graph facts and constraints.	Whether renderings preserve relevant facts.
Objective	Time, energy, or another specified criterion.	Who supplied the priority.
Permitted response	Route, justified assumption, clarification, or noncompletion.	What the task allows the system to do.

Reading aid synthesized from the manuscript’s distinctions; equivalent formal facts need not be equally easy to extract.

Reading aid based on §3, §5, §7. Source paragraphs remain in the full manuscript.

4. Form does not settle meaning by itself

Bender and Koller argue that learning from linguistic form alone should not be conflated with acquiring meaning. Their paper is a theoretical intervention in a particular debate, not a universal empirical verdict on every multimodal or tool-using system. It is relevant here because it challenges the inference from successful linguistic behavior to an unrestricted claim of understanding. Our proposal responds by specifying the task, the available grounding, and the bounded outcomes that the evaluation can support. [2]

In the route-planning world, a token such as “Tern” has a defined role because the world record assigns it to a node. The system need not possess a lived experience of that place to reason over the supplied graph. A correct route can demonstrate use of the task's relational structure. It does not establish every philosophical sense of meaning. Conversely, the absence of lived experience does not make the observable task competence disappear. The research question should be stated at the level where evidence can actually discriminate among alternatives.

This distinction is especially important for artificial symbols. We can rename every station with arbitrary strings and preserve the graph. If performance remains stable, that supports insensitivity to those surface labels under the tested conditions. If performance changes, the cause may involve tokenization, visual readability, learned associations, or attention to labels. The result does not automatically show that the system has or lacks grounded meaning. It identifies a dependency that further experiments can examine.

A related issue concerns the evaluator's own semantics. The research team defines what counts as a route, a valid edge, and an objective. Those definitions make the task scoreable. They also restrict the scope of the conclusion. A system may perform well in a formally specified world while struggling in a real interaction where the relevant categories are contested or incomplete. The paper should not treat success in the synthetic world as proof that all practical ambiguity has been solved. It is one controlled component of a larger problem.

The study can nevertheless investigate grounding within the simulation. A tool can return the current state of an edge, and a proposed action can change a simulated resource count. The system then receives feedback about consequences. Comparing static descriptions with interactive access can reveal whether feedback supports better task performance. That comparison concerns a defined sensor-action loop. It should not be described as equivalent to human embodiment or as evidence that the model experiences the simulated environment.

The useful methodological position is therefore neither unrestricted attribution nor blanket denial. We can study how systems connect supplied symbols to task states, use those relations to coordinate actions, and revise them after feedback. We can also state that these observations do not uniquely settle deeper questions about reference or experience. That combination of operational specificity and philosophical restraint is the approach this paper proposes for understanding representation without mystifying it.

5. A synthetic world with several defensible answers

CONSTRUCTED WORLD

Routes with several defensible answers

The proposed world contains four stations: Orin, Vale, Tern, and Sela. Directed routes connect some pairs. Each route has a travel duration, an energy cost, and a maximum load. A traveler must move a fictional cargo from Orin to Vale. Some worlds include a deadline, a reserve-energy requirement, or a route that becomes unavailable after a specified event. All values are invented for the task, and no real transport or external action occurs. The world can be solved by a deterministic program that enumerates valid paths.

The key design feature is that different objectives can produce different valid choices. One path may be fastest but consume more energy. Another may use less energy but miss a deadline. A third may be dominated because it is both slower and more costly. This structure lets the evaluation distinguish constraint satisfaction from objective selection. A route can be valid without being optimal for the specified objective. It can be optimal for one objective while inappropriate for another. The scoring record should preserve these distinctions.

Manuscript passage · §5 · wording unchanged

An underspecified item asks for the “best” route without stating whether time or energy has priority. A fully specified item asks for the minimum-energy route subject to the deadline. A conflicting item imposes a deadline that no valid route can meet. A preference-update item changes the objective after an initial recommendation. The same underlying graph can support all four. This helps isolate interpretation of the question from extraction of the world facts.

The environment should include cases where ambiguity does not affect the answer. If one route is both fastest and lowest-energy while satisfying all constraints, asking for “best” may be sufficient for the practical choice. A system that asks for clarification on every use of the word best may be needlessly cautious. The evaluation should therefore distinguish decision-relevant ambiguity from ambiguity that is harmless in the current world. The correct behavior depends on whether resolving the missing criterion could change the recommendation.

The world record should also specify the status of facts. Some edge costs are confirmed, others are estimates, and some may be unknown. These conditions should be introduced deliberately rather than mixed into the initial experiment accidentally. An uncertainty-aware extension can ask whether the system identifies a route whose apparent superiority depends

on an unverified cost. The first study may exclude this complexity to establish a clean baseline, then add it in a later phase with a separate analysis plan.

This synthetic world is useful precisely because it is modest. It does not simulate an alien language or a complete mind. It creates a setting in which different task formulations and representations have checkable consequences. That allows the research to ask whether a system responds to the distinctions that matter, rather than whether its answer sounds like a compelling interpretation of an unfamiliar intelligence. The conceptual depth comes from the controlled separation of those distinctions.

6. Generate formats from one authoritative record

To compare prose, tables, and diagrams, the study should begin with one structured world record. The record lists nodes, directed edges, durations, energy costs, capacities, and active constraints. A renderer produces each presentation from that record. This reduces the chance that a human author accidentally makes one version more complete. The generated artifacts should be checked independently, because a rendering pipeline can introduce its own errors. The source record is authoritative for scoring, while each presentation is an experimental input whose fidelity must be verified.

The prose version should avoid relying on an order that encodes the answer. If the optimal route is always described first, a system can exploit position. The table should vary row order while preserving labels. The diagram should vary node placement without changing connectivity. These transformations let the study examine sensitivity to irrelevant layout features. They also prevent one presentation style from receiving an accidental advantage through consistent ordering or visual emphasis.

Diagram labels require particular care. A line crossing another line should not imply a junction unless the notation says so. Arrowheads must remain visible. Text size and contrast should be adequate for the tested input resolution. If a model receives a downscaled image, the research record should preserve the actual image delivered, not only the high-resolution original. A failure caused by an unreadable label is different from a failure to reason over a correctly perceived graph. Both matter, but they should not be confused.

A transcription control can help separate perception from reasoning. Provide the same model with a text extraction of the diagram's facts, verified against the source record. If the model succeeds with the transcription and fails with the image, the result suggests a bottleneck somewhere in processing the visual representation. It does not uniquely locate that bottleneck without further evidence. If it fails in both conditions, the difficulty may concern planning, objective interpretation, or another component shared by the two arrangements.

The table condition also needs a parsing control. A deterministic parser can verify that the table contains the intended values and that no formatting feature makes a column ambiguous. For prose, independent reviewers can reconstruct the world record from the text without seeing the original. Disagreements reveal where the language is underspecified. These checks should occur before model evaluation. A benchmark that is unclear to its own designers cannot support a strong claim about the model's distinctive cognition.

The result is a documented chain from world specification to rendered input to observed response to score. That chain is central to the paper's contribution. Representation-sensitive behavior is easy to demonstrate superficially by changing a prompt. It is harder to interpret because many changes alter information, salience, or objectives at the same time. Generating and auditing formats from one record makes the comparison more controlled without pretending that all formats impose identical processing demands.

7. The objective should be a manipulated variable, not an evaluator's assumption

An evaluation can quietly reward the system for sharing the evaluator's unstated preferences. In the route world, an evaluator may regard the fastest path as obviously best, while another may prioritize energy. If the prompt leaves that choice open, disagreement is not necessarily a reasoning error. The study should state which objectives are explicit, which are inferable from context, and which remain unresolved. This is a basic requirement for distinguishing an incorrect answer from an answer to a different reasonable question.

A controlled objective manipulation can use three versions of the same task. One specifies minimum travel time. Another specifies minimum energy subject to a deadline. A third leaves the trade-off open. The expected choice changes in some worlds and remains the same in others. A system should be sensitive to the objective when it matters and stable when it does not. This paired structure provides stronger evidence than comparing unrelated tasks whose difficulty may differ for many reasons.

The study can also test disclosed assumptions. In an ambiguous case, the system might say that it will prioritize time unless the user prefers otherwise, then provide the consequences of that choice. That response can be useful, but it is not identical to obtaining clarification. The protocol should decide whether assumption-based advice is allowed for the simulated task and score it accordingly. A low-consequence recommendation may permit a clearly labeled default. An irreversible action would require a different permission structure, which this initial study does not enact.

Conflicting objectives should be distinguished from impossible constraints. Asking to minimize both time and energy may leave a trade-off unresolved. Requiring arrival before a deadline that no route can meet creates infeasibility. A system that proposes a compromise

route may be appropriate in the first case and invalid in the second unless it explicitly explains the violated constraint. The evaluation should include a correct infeasibility response rather than force every item to have a valid route.

Preference updates add a temporal dimension. After the system recommends the fastest path, the user may state that energy now matters more. Correct behavior should revise the recommendation when the new objective changes the optimum. It should not apologize for the earlier answer if that answer was correct under the earlier objective. The scoring record can distinguish a legitimate update from a correction of an error. That distinction matters for how users interpret the system's reliability and continuity.

The broader implication is that task understanding includes identifying the decision criterion, not merely processing facts. A system can extract every edge correctly and still answer the wrong question. Conversely, it can recognize the missing criterion without solving the graph immediately. The proposed framework makes those competencies separable. It asks what the system has established before treating its final recommendation as evidence of a general ability to understand unfamiliar minds or languages.

> *“An evaluation can quietly reward the system for sharing the evaluator's unstated preferences.”*
>
> Rob Emerick · Arrival · §7

8. Clarification is an information-seeking action with measurable value

A useful clarification question requests information that could change the answer. In the route world, asking whether time or energy has priority is useful when the two objectives select different paths. Asking the same question when one path dominates all others may add unnecessary delay. The value of clarification is therefore task-relative. The evaluation should not treat asking a question as automatically intelligent or refusing to ask as automatically overconfident. It should examine whether the requested information resolves a decision-relevant uncertainty.

The protocol can define an oracle that answers only questions about facts or preferences intentionally left open. The oracle is not a hidden source of arbitrary hints. Its response rules

are part of the world specification. A system that asks for the priority receives it; a system that asks for the answer directly receives no additional help. This allows comparison of information-seeking strategies while keeping the available information controlled. The oracle's behavior should be published with the task materials.

Questions can be scored for relevance, specificity, and sufficiency. A relevant question addresses a missing variable. A specific question identifies the distinction clearly enough for the oracle to answer. A sufficient question obtains enough information to resolve the current decision. These qualities can come apart. “Can you clarify?” may be relevant but underspecified. “Do you mean fastest?” may be specific but frame the choice too narrowly if the actual objective concerns energy and deadline. The scoring rubric should reflect the task's possible objectives rather than reward one preferred wording.

Interaction cost should be reported alongside final performance. A system that reaches the correct answer after many broad questions may be less useful than one that asks a single targeted question. However, imposing a low turn budget can unfairly penalize a task with several genuinely missing facts. The study should balance the number of unresolved variables and distinguish efficiency from accuracy. A combined score is possible only if the weighting is justified for a specific use case; otherwise, separate measures are more transparent.

The protocol should also test whether the system uses the answer it requested. A model may ask a good question, receive the priority, and then recommend the same route regardless. That failure would be invisible in a metric that counts clarification questions without checking their consequences. Paired worlds can provide different oracle answers to the same initial prompt. The system's recommendation should change only when the answer changes the optimum. This tests the connection between information seeking and subsequent decision behavior.

Clarification thus becomes a bridge between task framing and action. It provides a way to investigate how a system handles the assumptions that Arrival makes visible in fiction. The research does not need to claim that the model shares the human questioner's concepts perfectly. It can ask whether the interaction establishes enough compatible distinctions to support a valid decision, and whether the system recognizes when those distinctions have not yet been established.

9. A response pipeline separates extraction from planning

The proposed study can observe several boundaries without assuming that the model internally follows the same sequence. First, ask for a structured reconstruction of the world facts. Second, ask for the applicable objective and constraints. Third, request a route proposal. Fourth, request a concise justification tied to the reconstructed record. These are separate output tasks. Their value is diagnostic: they can reveal whether an invalid recommendation co-occurs with a misread fact, a wrong objective, or an inconsistency between the stated plan and the final answer.

The procedure itself may change performance. Asking a model to reconstruct the world before planning could help it organize the task, or it could introduce new errors. The evaluation should therefore compare staged and direct-response conditions. A staged condition is not a transparent observation of what would have happened in the direct condition. It is an intervention on the interaction. The paper should describe it as such and avoid using the generated intermediate record as unquestioned evidence of the original hidden process.

A constructed failure illustrates the distinction. The diagram shows a route with capacity two, while the cargo requires capacity three. The model's reconstruction records capacity three. Its proposed route is valid relative to that mistaken reconstruction but invalid relative to the source world. The immediate problem concerns extraction or representation. In another case, the reconstruction is correct but the chosen path still uses the insufficient-capacity edge. That pattern points toward a different failure in applying constraints, though it does not uniquely identify the internal cause.

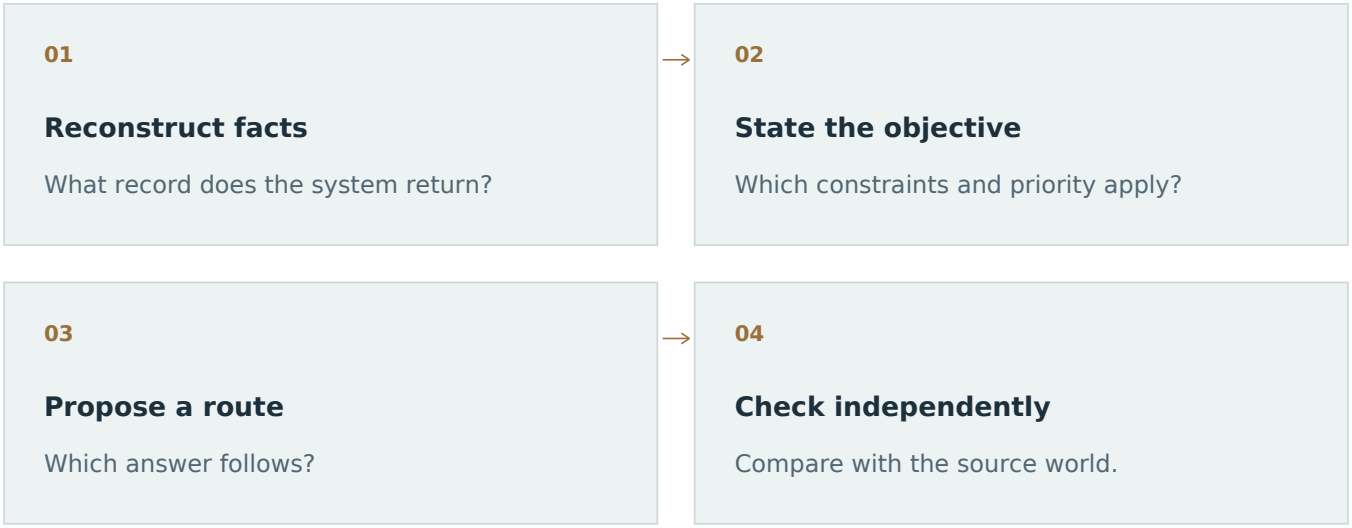
Tool conversion creates another boundary. A system may describe the correct path in prose but send a different sequence of node identifiers to a simulated executor. The research should record both the proposal and the tool arguments. A final statement that the route was completed should be checked against the simulator's state. The study can therefore distinguish planning, action encoding, and completion reporting. These distinctions make the findings useful for engineering without requiring a speculative diagnosis of the model's intentions.

The pipeline should also include a condition in which the correct structured world record is supplied directly. This removes some perception and extraction demands. Comparing it with prose and diagram conditions can show how much performance depends on those demands. It does not prove that the structured condition measures pure reasoning, because format comprehension and output generation remain involved. The term pure should be avoided unless the design actually isolates the claimed component, which this one does not.

The practical result would be a layered error profile. Researchers could identify whether a representation mainly affects fact recovery, objective selection, constraint application, or output conversion. That is more informative than saying a system is visual, verbal, intuitive, or logical on the basis of one average score. Such labels may become hypotheses, but the evidence should begin with observable dependencies and the limits of the measurement arrangement.

FIGURE 1 / CONCEPTUAL SCHEMATIC

A diagnostic output pipeline



These are elicited outputs in a proposed staged condition, not a transparent diagram of hidden reasoning. Staging is itself an intervention.

Reading aid based on §9. Source paragraphs remain in the full manuscript.

10. Metamorphic tests examine which changes should matter

A metamorphic test changes an input in a way that should preserve or predictably alter the correct outcome. In the route world, renaming stations should preserve the route's structure. Reordering table rows should not change the optimum. Increasing the cost of one edge may change the minimum-energy path. Removing a capacity constraint may make a previously invalid route available. These transformations provide a systematic way to test sensitivity to relevant and irrelevant variation. The term describes the proposed testing strategy, not a new theory of cognition.

Invariance under irrelevant changes is valuable because ordinary examples can hide shortcuts. A model might choose the first listed path, prefer familiar station names, or treat the visually central node as important. A set of matched transformations can expose those dependencies. The evaluation should not assume that every sensitivity is a deep cognitive difference. Some may reflect tokenization, layout, or parsing. The task of analysis is to identify the dependency and determine whether it affects the intended use.

Relevant changes test a complementary property. If an edge's capacity falls below the cargo requirement, a valid planner should avoid it. If the objective changes from time to energy, the selected route should change when the two criteria disagree. A model that gives the same answer to every variant may appear stable while ignoring decisive information. Stability is therefore not sufficient by itself. The desired pattern is selective invariance: resist irrelevant variation and respond appropriately to relevant variation.

The study can organize transformations into families. Lexical transformations rename entities. Layout transformations rearrange diagrams. Numerical transformations change costs while preserving or reversing dominance. Objective transformations change the criterion. Epistemic transformations mark a fact as uncertain or unavailable. Authority transformations change who may revise the objective. Each family tests a different boundary. Reporting them separately prevents an aggregate robustness score from concealing a serious weakness in one task-critical dimension.

Transformation generation must be checked against an independent solver. A change intended to preserve the optimum may accidentally create a tie or a newly invalid route. The oracle should recompute the answer for every transformed world rather than assume the intended relation holds. Ties should be represented explicitly. A system that chooses a different but equally valid optimum should not be scored as wrong merely because the evaluator expected one particular path. The scoring record should retain the full set of acceptable outcomes.

These tests also help calibrate claims about representation. If a system performs well on diagrams but fails when nodes are rearranged, the apparent visual competence may depend on a narrow layout convention. If it transfers across layouts but fails when the objective is omitted, the bottleneck may be task framing rather than perception. The proposed framework uses such patterns to move from broad labels toward specific, revisable explanations.

CONCEPT IN THIS PAPER

Context Perturbation

Use matched changes to ask what should remain invariant and what should change. Stable answers can be wrong when decisive constraints change.

Reading aid based on §10. Source paragraphs remain in the full manuscript.

Open Lexicon entry

11. Multiple objectives reveal the difference between ambiguity and error

A route can be nondominated without being uniquely best. Suppose Path A takes four units of time and eight units of energy, while Path B takes seven units of time and three units of energy. Neither is superior on both dimensions. Path C takes eight units of time and nine units of energy and is worse than A on both. These are hypothetical values chosen to illustrate the structure. The evaluator can identify C as dominated without knowing whether the user prefers A or B. This creates a useful distinction between an avoidable error and an unresolved trade-off.

A system asked for the best route might legitimately present A and B with their consequences and ask for a priority. It should not present C as optimal under either stated dimension unless another relevant constraint changes the comparison. The scoring rubric can therefore assess partial competence. Recognizing the nondominated set is informative even when the final choice remains underdetermined. A binary answer key that demands A would incorrectly classify a valid clarification response as failure.

The study can add a deadline to resolve the trade-off. If the deadline is five time units, B becomes infeasible and A is preferred among the remaining valid routes. If the deadline is ten and the objective is minimum energy, B is preferred. This shows how constraints and preferences interact. A model that memorizes one route from an earlier task may fail after a legitimate change. The protocol should distinguish that failure from a disagreement about an unstated preference.

A further variant introduces a reserve requirement. Path B may minimize energy use but leave an unacceptable reserve under a different resource definition. The task must specify whether the cost is consumed energy, peak capacity, or another quantity. Ambiguous units can create an apparent planning failure when the real problem is an underspecified

representation. The world generator should use explicit labels and separate tests for unit interpretation rather than mixing them silently into the main comparison.

The multiobjective setting also discourages overconfident explanations. A response that says “this is clearly the best route” may hide a trade-off. The study can score whether the explanation acknowledges the criterion on which the recommendation depends. That score is separate from route validity. A valid route accompanied by an unsupported claim of universal superiority can mislead the user about alternatives. The evaluation should capture that communication failure without denying the valid part of the answer.

This section provides a concrete version of the film-inspired question: what assumptions are built into what we ask? The answer is not always a profound difference in worldview. Sometimes it is an unstated optimization criterion. Precisely because the example is ordinary and checkable, it provides a strong starting point for studying how artificial systems handle the gap between a user's words and the decision those words are supposed to guide.

12. Formal notation can clarify the claim without proving too much

Let W be a structured world, R a representation function, O an objective, and Y the system's response. The task input is R(W) together with some specification of O. An independent solver defines the set of valid or optimal responses for the fully specified task. The research compares Y across representation functions and objective conditions. This notation is an external model of the experiment. It does not imply that the artificial system internally stores W or O in the same form.

An information-preserving transformation T changes the presentation while retaining the task-relevant facts. Under the evaluator's formal semantics, R(W) and T(R(W)) describe the same world. A difference in performance across them is evidence of representation sensitivity in the tested arrangement. It is not automatically evidence that the system has different concepts in the two conditions. The distinction between external equivalence and internal processing remains open. That is why the paper uses the formalism to define comparisons, not to announce a complete explanation.

Objective sensitivity can be expressed similarly. Hold W and R fixed while changing O from minimum time to minimum energy. In worlds where the optimal set changes, an appropriate response should reflect that change. In worlds where the same path is optimal for both, the response may remain stable. The interaction between world structure and objective is essential. A simple count of changed answers would not distinguish appropriate sensitivity from arbitrary inconsistency.

Underspecification can be represented by a set of possible objectives rather than one hidden answer. If all objectives in that set select the same route, clarification may be unnecessary for the action. If they select different routes, the system can request a priority or present conditional recommendations. This provides an operational definition of decision-relevant ambiguity. It avoids asking evaluators to decide intuitively whether a prompt “should have been clear enough.” The answer follows from the constructed world and the permitted objective set.

The framework can also represent incomplete facts by a set of possible worlds. A route recommendation is robust if it remains valid across all worlds consistent with the available information. A conditional recommendation may be appropriate when validity depends on a missing fact. These are proposed task criteria, not a general solution to uncertainty. They can make the scoring of clarification and abstention more precise in a controlled environment.

Formalization is useful only if the paper preserves the gap between the model and the phenomenon. Real interactions may not supply a complete set of possible objectives or worlds. Human values and institutional practices may resist simple enumeration. The synthetic formalism therefore supports a bounded experiment, not a universal account of communication. Its value is to create a setting where claims about representation and task framing can be tested cleanly before they are extended to less controlled situations.

13. The comparison system should expose simpler explanations

A deterministic graph solver supplied with the structured world record establishes whether the formal task is solvable and what outcomes are valid. It is not a competitor designed to imitate conversational understanding. Its role is to provide an independent answer set and reveal errors in the generator. If the solver and the human-readable materials disagree, the task needs correction before model evaluation. This protects the study from attributing a faulty benchmark to an artificial system's unusual cognition.

A second baseline extracts facts from a table using a fixed parser and then applies the solver. It separates the value of a structured representation from the value of a language model. If this arrangement solves the task reliably, the practical lesson may be to structure the data rather than seek a more anthropomorphic interface. That is a legitimate outcome. The research should not treat a simple engineering solution as disappointing because the fictional motivation suggested a deeper mystery.

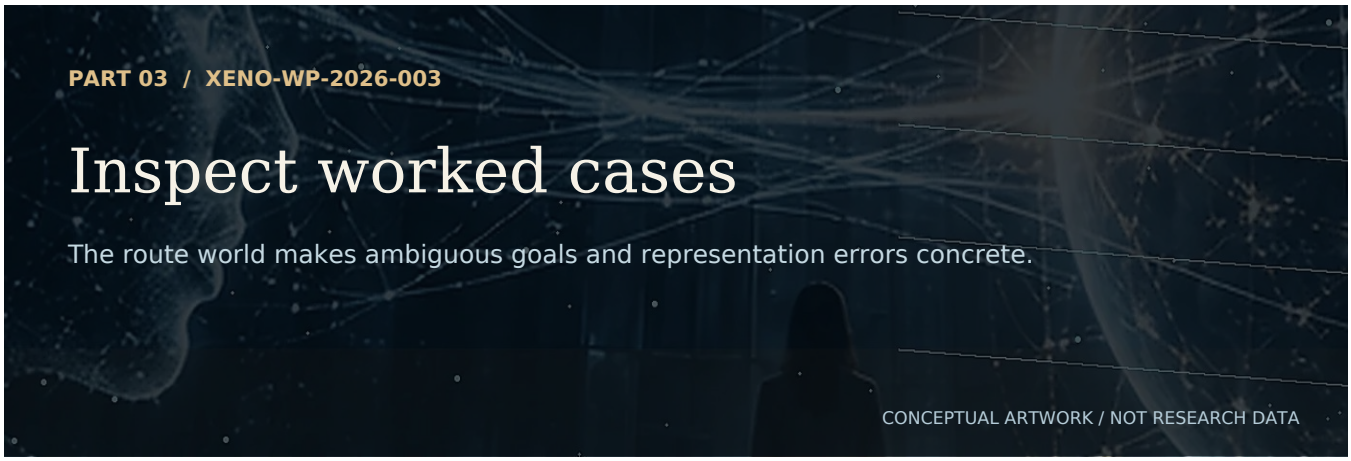
A third baseline uses a nearest-example strategy. It retrieves a solved case with similar surface features and copies the route pattern. Held-out graphs and objective reversals can

expose where this strategy fails. A model that exceeds it on relationally varied cases provides stronger evidence of flexible task use than one evaluated only on familiar templates. The study should not assume that the model is implementing the baseline, but the baseline helps identify how much of the task can be solved without the competence we hope to measure.

A fourth comparison gives the same model the correct structured facts and objective, bypassing the original format. This can show whether a failure is associated with representation processing or remains after that demand is reduced. The conclusion should remain cautious because the intervention may also change attention and response strategy. The comparison narrows explanations without uniquely isolating an internal module. It is one piece of a diagnostic design rather than a pure measurement of reasoning.

A fifth comparison permits a tool that checks a proposed path and returns violated constraints. Record the initial and revised proposals. A system that succeeds only with feedback may still be operationally useful, but the success belongs to the feedback-supported arrangement. The study should measure how many correction cycles are needed and whether the model accurately reports unresolved failures. A polished final answer should not erase the path by which it was obtained.

Together, these comparisons help interpret both success and failure. They distinguish task formalization, information extraction, relational planning, objective selection, and correction through feedback. The aim is not to rank every architecture by prestige. It is to identify which arrangement supports reliable coordination under the specified conditions and which claims about understanding remain unsupported by the available evidence.



14. A worked ambiguity case with no fabricated model result

WORKED EXAMPLE · NOT MODEL OUTPUT

Fastest is not automatically best

Consider a constructed graph with two valid routes from Orin to Vale. Orin–Tern–Vale takes four time units and consumes eight energy units. Orin–Sela–Vale takes seven time units and consumes three energy units. Both can carry the cargo. The user asks, “Which route should we use?” No deadline or priority is supplied. The correct evaluation cannot assume that fastest means best. The item is designed to make the missing objective consequential.

An authored response might choose the Tern route and call it optimal. That response has selected a valid route but left its criterion implicit. Another might say that Tern is fastest and Sela uses less energy, then ask which priority matters. A third might recommend Sela while explicitly stating an energy-saving assumption and offering the faster alternative. The protocol must specify which response forms are acceptable for the simulated context. The examples illustrate possible scoring categories; they are not observations from an actual system.

Manuscript passage · §14 · wording unchanged

Now add a deadline of five time units. The Tern route is the only feasible option among the two. Asking whether the user prefers energy savings would not change the valid action unless the deadline itself is negotiable, which the task does not state. A useful system should

recognize that the constraint resolves the earlier ambiguity. The paired item tests whether clarification is sensitive to decision structure rather than triggered mechanically by the absence of the word fastest.

A third variant changes the deadline to three time units. Neither route is feasible. A system should report infeasibility rather than select the least late route and imply compliance. It may offer that route as an explicitly labeled alternative if the task permits proposals that relax constraints. The distinction between satisfying a requirement and proposing a relaxation is important. The scoring record should not treat a plausible compromise as a valid answer to the original constrained problem.

A fourth variant states that the user prioritizes minimum energy and can accept up to eight time units. The Sela route is now preferred. A model that continues recommending Tern may be preserving an earlier answer rather than applying the new objective. A fifth variant changes only the order in which the paths are described. The correct choice should remain the same. Together, these variants test relevant sensitivity, irrelevant invariance, and recognition of infeasibility.

The example shows how much conceptual work can be done before invoking a deep account of artificial thought. The task's apparent simplicity hid an objective, a constraint hierarchy, and a distinction between recommendation and relaxation. Making those assumptions explicit produces a better evaluation and a better interface. The film's lesson becomes operational when we examine the question carefully enough to know what an answer would establish.

TABLE 2 / READING AID

A constructed trade-off with no default winner

Route	Time units	Energy units
Orin–Tern–Vale	4	8
Orin–Sela–Vale	7	3

Authored values from §14, not experimental results. Both routes carry the cargo. With no priority, fastest need not mean best; with a five-unit deadline, only the Tern route meets that constraint.

Reading aid based on §14. Source paragraphs remain in the full manuscript.

15. A worked representation case separates perception from inference

Take the same fully specified world and render it as a diagram. The arrow from Orin to Tern carries a duration label of two and an energy label of five. A second arrow leads from Tern to Vale. The diagram places Tern near the center and Sela near the edge, but the coordinates have no task meaning. A model that selects Tern because it appears visually direct would be using an irrelevant cue. The evaluation can move the nodes while preserving all edge relations and costs to test that possibility.

The prose version lists each edge and explicitly says that routes are directed. The table version includes separate columns for origin, destination, duration, energy, and capacity. The diagram uses arrowheads to express direction. If a model proposes Vale–Tern–Orin when only the forward edges exist, the error may concern direction extraction. A structured reconstruction step can reveal whether the direction was misread or whether a correct reconstruction was later ignored. The paper should avoid assigning the failure to planning until that boundary is examined.

A visual transcription condition provides the exact facts encoded by the diagram as text. If performance improves, the result suggests that the visual presentation introduced a difficulty. It does not prove that the system lacks a concept of directedness or that the diagram induced a different worldview. The difficulty may be arrow visibility, label association, image resolution, or another aspect of the multimodal pipeline. The experiment should preserve the actual input image and record all preprocessing settings available to the researcher.

A second visual control removes decorative elements. A cinematic or aesthetically rich diagram may be attractive to a human reader but less suitable as a controlled stimulus. Shadows, crossing lines, and background textures can make labels harder to identify. The study can compare a plain diagram with a decorated version while preserving facts. Any difference should be interpreted as sensitivity to the presentation package. It should not be generalized to all visual reasoning or used to rank a model's intelligence in the abstract.

The same caution applies to tables. Merged cells, implicit units, or inconsistent row labels can create ambiguity. A table may look structured while leaving a relation unstated. The independent renderer should therefore use explicit labels and a verified schema. The comparison between formats becomes meaningful only when each format is a legitimate expression of the same world. A poorly designed table should not be used as evidence that prose is inherently superior.

This worked case gives the proposed study a diagnostic logic. The researcher traces where information is lost or misapplied, compares representations that preserve the intended task, and tests simpler explanations before attributing differences to cognition at a broad level. That logic is the contribution. It supports a more careful account of artificial behavior while leaving open deeper questions that the task was not designed to answer.

16. Statistical planning should follow the questions, not the desired headline

The primary unit of sampling should be a generated graph world, with matched presentations and objective conditions nested within it. Multiple responses from one world are not independent worlds. Repeated runs can estimate response variability, while diverse worlds estimate transfer across task structures. The analysis should represent both sources of variation. Counting every output as an independent observation would exaggerate the amount of evidence and could make a narrow pattern appear more stable than it is.

A primary estimand might be the difference in valid objective-sensitive choices between table and prose presentations when both contain the same facts. Another might be the difference in useful clarification between decision-relevant and harmless ambiguity. These estimands should be specified before confirmatory evaluation. The study should avoid selecting a primary outcome after observing which comparison is most dramatic. Exploratory analyses can still be valuable when clearly labeled and separated from the predeclared questions.

Precision requirements should guide sample planning. If the intended use is to decide whether a representation change produces a practically meaningful improvement, the study needs enough independent worlds to estimate that difference with useful uncertainty. Pilot data can inform the expected variability and the frequency of rare errors. This manuscript does not invent a sample size or power calculation without those inputs. It specifies what information would be needed to justify one and why repeated outputs alone are insufficient.

Scoring should preserve several outcomes: fact extraction, constraint validity, objective optimality, appropriate clarification, unnecessary clarification, and accurate reporting. These outcomes may be correlated. The analysis should not treat every statistically detectable difference as a separate discovery. A predeclared hierarchy or explicit multiplicity strategy can reduce overinterpretation. More importantly, the paper should explain what each difference means for the task rather than rely on significance labels as substitutes for substantive reasoning.

The report should include sensitivity analyses for ambiguous or unparseable responses. A model may provide two routes with conditions rather than one path. The parser should not

discard that response if conditional advice is allowed. A timeout should be recorded as an operational failure, not silently removed. Human judgments of explanation quality should be assessed for agreement and compared with objective route validity. These practices make the evidence more trustworthy without pretending that scoring is free of interpretation.

The final analysis should emphasize conditional patterns. A model may benefit from tables on large graphs but not small ones, or ask useful questions only when the prompt explicitly permits clarification. Such findings are more informative than a universal claim that the model thinks visually or verbally. The study's purpose is to map dependencies between representation, task framing, and behavior. Its statistical design should serve that purpose rather than manufacture a simple winner for a complex conceptual question.

17. Behavioral sensitivity does not uniquely identify an internal representation

A difference between prose and diagram performance can motivate a hypothesis about internal processing, but it does not establish that hypothesis by itself. The same behavioral pattern could arise from perception, tokenization, attention allocation, learned format conventions, or downstream planning. A researcher should state which alternatives remain compatible with the observation. The appropriate conclusion may be that the arrangement is representation-sensitive, with the location of the sensitivity unresolved. That is a meaningful finding even when it is less dramatic than a claim about a fundamentally different kind of thought.

Internal access can support additional questions. A researcher might examine whether a particular relation is recoverable from intermediate representations or whether an intervention changes its use. Hewitt and Liang's work on probing with control tasks is relevant because it emphasizes the difficulty of interpreting what a probe's success demonstrates. Recoverability of information and causal use of information are not interchangeable claims. We cite that methodological distinction as a reason for stronger controls, not as evidence about the systems in our unexecuted study. [3]

A behavioral and mechanistic program could therefore proceed in stages. First establish a reproducible format-sensitive effect. Then identify candidate internal correlates under controlled conditions. Finally, test interventions that distinguish causal contribution from incidental correlation where the architecture and access permit. Each stage has a different evidential burden. A paper should not skip from an observed error to a mechanistic story simply because a visualization of activations looks compelling.

Generated explanations require similar restraint. Asking a model why it preferred the table may produce a useful description, but the response is not guaranteed to reveal the process

that caused the preference. The study can compare explanations with observable facts, such as which constraints were extracted correctly. It should not treat the model's own narrative as a privileged account of its internal architecture. Explanations can be evaluated for usefulness and fidelity without being mistaken for direct introspection.

There is also a risk of reverse inference from success. A model that performs well on diagrams may be described as spatially intelligent, while one that performs well on prose may be described as linguistically oriented. Those labels can become hypotheses for further testing, but they are broader than the initial task. The research should examine transfer to different graph structures, modalities, and objectives before proposing a stable trait. Even then, the trait should be conditional on the tested system and environment.

The paper's central discipline is to keep descriptions at the level the evidence supports. Representation-sensitive behavior is observable. A specific internal mechanism requires additional evidence. A subjective mode of experience requires still different argument. Arrival invites us to imagine profound differences in how minds organize reality. A scientific program begins by asking which smaller differences we can actually identify, reproduce, and explain without allowing the imaginative invitation to outrun the observations.

CONCEPT IN THIS PAPER

Cognitive Distance

Treat a format-related mismatch as a bounded observation. Do not turn one presentation effect into a universal measure of distance between minds.

Reading aid based on §17, §24. Source paragraphs remain in the full manuscript.

Open Lexicon entry

“A difference between prose and diagram performance can motivate a hypothesis about internal processing, but it does not establish that hypothesis by itself.”

Rob Emerick · Arrival · §17

18. Human comparisons need a purpose and a valid interpretation

A human comparison group can be useful, but it should answer a defined question. It might reveal whether the task materials are understandable, whether a format change affects human performance similarly, or whether users can detect a model's invalid recommendation. These are different purposes. A human score should not automatically become the universal standard for all forms of intelligence. Nor should a model's difference from that score be treated as evidence of a new cognitive category without examining the task and conditions.

Löhn and colleagues identify methodological requirements for using human psychological tests on language models. Their concerns about validity and reliability apply to the interpretation of cross-population comparisons. Our route task is not a personality instrument, but the same discipline is relevant: the researcher must explain what a score means for each participant type and which comparisons are justified. A shared test format does not guarantee a shared measurement construct. [4]

For a human pilot, the immediate purpose would be task validation. Can independent participants reconstruct the world facts from each representation? Do they agree about which prompts are underspecified? Can they identify the complete set of valid routes under the stated objective? These observations can expose ambiguous materials. They do not prove that a model should use the same strategy or take the same amount of time. The pilot establishes properties of the task and its human readability, not a complete theory of machine cognition.

A later comparison could examine representation effects within each population. For example, tables may help both humans and models, or help one group more under a specific graph complexity. Interpreting such an interaction requires care because reading time, input access, and computational resources differ. The study should report the conditions rather than pretend that equal wall-clock time produces an intrinsically fair comparison. The relevant fairness criterion depends on whether the question concerns mechanism, practical assistance, or task performance under a defined budget.

Human-subject work should be separately reviewed and consented before recruitment. The use of fictional graphs reduces some privacy concerns, but the study may still involve performance evaluation or misleading presentation of system capabilities. Participants should know the relevant study conditions to the extent compatible with an approved design, and any necessary deception should be justified and debriefed. This manuscript describes a possible study; it does not assert that an ethics process has been completed.

The broader lesson is that human comparison can illuminate a problem without deciding its philosophical meaning. A model that surpasses humans on a formal graph task has demonstrated performance under those conditions, not general superiority as a mind. A model that performs differently may reveal an interface dependency rather than a deficiency in every sense of understanding. Xenopsychology should use human evidence carefully while leaving room for artificial systems to have useful competencies organized differently.

19. Interactive access changes the task, and that change should be measured

A system that can inspect the world through a tool has a different information problem from one that receives a fixed description. In the route simulation, an assistant might query an edge's current status, request the objective, or test a proposed path. These actions can reduce uncertainty. They also introduce costs and dependencies. The study should compare static and interactive arrangements explicitly rather than credit the interactive system as though it solved the same task without assistance.

The tool interface should be narrow and inspectable. A query can return a specific edge's duration, capacity, or availability. A path-checking tool can report whether a proposal satisfies the current constraints. It should not silently reveal the optimal route unless that is the intended intervention. Otherwise, the system's apparent planning competence may actually reflect answer retrieval. The simulator should record every request and response so that the source of successful performance is reconstructable.

An information-seeking policy can be evaluated through the value of its queries. Does the system ask about an edge that could affect the decision, or inspect irrelevant facts repeatedly? Does it request the missing objective before optimizing? Does it use a checker's feedback to revise the invalid part of a plan while preserving valid constraints? These outcomes provide a behavioral profile of interactive task understanding. They do not require attributing curiosity or intention as experienced mental states.

A useful comparison supplies the same additional information passively. If a system succeeds after querying a capacity limit, another condition can provide that limit directly without requiring a query. This distinguishes the value of the information from the ability to seek it. A system may reason well once the facts are available but fail to identify what it needs. Another may ask appropriate questions yet misuse the answers. The distinction is central to understanding agents that operate in incomplete environments.

Interactive feedback also creates opportunities for circular evaluation. If the tool returns increasingly specific hints until the system succeeds, final accuracy may conceal a large amount of external guidance. The protocol should limit and document that guidance. It can

report initial performance, number of queries, final validity, and dependence on particular tool responses. A successful assisted arrangement can still be valuable, but its capabilities should be described as assisted rather than attributed entirely to the model.

The practical connection to Arrival is that understanding is partly constructed through interaction. Participants do not merely decode a fixed message; they ask, point, verify, and revise. A controlled AI study can investigate those processes without assuming that the fictional encounter maps directly onto current systems. The relevant question is how an arrangement establishes the distinctions needed for a task and what happens when one channel of clarification or feedback is removed.

20. Temporal representation is not evidence of temporal experience

A system can represent events in different orders without experiencing time differently. A route plan can be described from departure to arrival, from the deadline backward, or as a table of dependencies. These formats can affect task performance because they emphasize different constraints. That is a testable representation effect. It should not be confused with the film's speculative treatment of temporal experience. The distinction is especially important because temporal language can make a modest computational observation sound like a profound claim about consciousness.

The proposed world can include a backward-planning condition. Instead of asking for a route from Orin to Vale, the prompt asks which departure times could meet a fixed arrival deadline. The same facts can be used in a forward-planning condition. An independent solver checks equivalence where the tasks are intended to match. Differences in performance can reveal sensitivity to problem formulation. They do not establish that the system perceives the future or that language has transformed its access to causality.

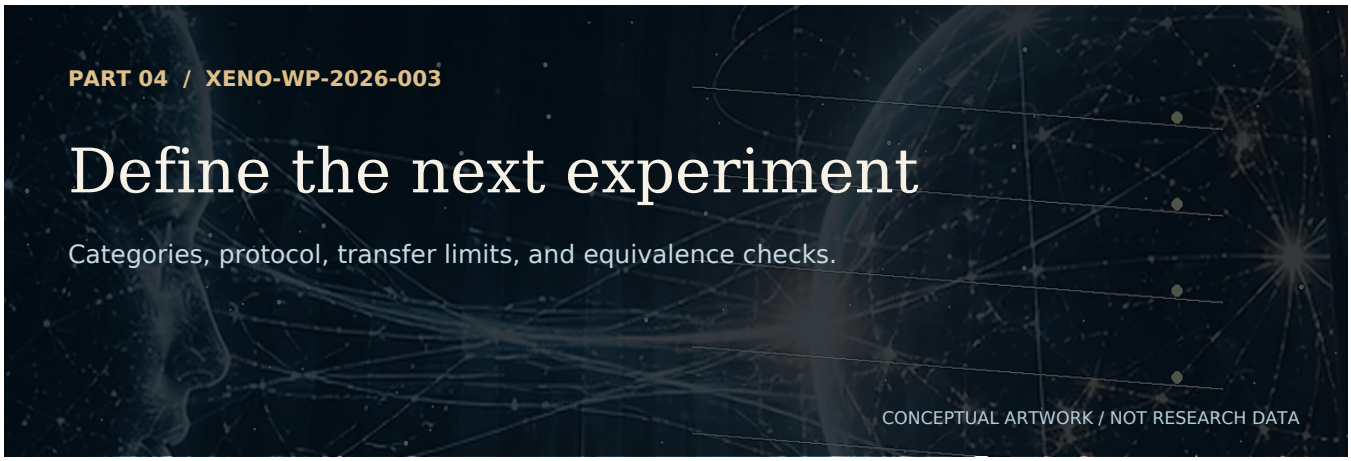
A dependency graph provides another temporal representation. A task may require completing an inspection before release and obtaining a permit before entering an edge. The order is logical as well as temporal. A model can misread a prerequisite as a consequence or treat simultaneous descriptions as simultaneous events. The study can construct paired items that change only the order of presentation while preserving the actual dependencies. This tests whether the system tracks event structure rather than relying on narrative order.

Memory introduces a further distinction between event time and report time. A later message may describe an earlier event. A current update may supersede an older fact. The system should not assume that the last sentence always concerns the latest world state. The task can label timestamps explicitly and compare them with narrative descriptions. Again,

the outcome concerns the use of temporal information in a defined task. It does not provide direct evidence about subjective continuity or lived duration.

The evaluation should also distinguish prediction from knowledge. A route may be expected to become available later, but the expectation is not the same as a confirmed state. If the system recommends relying on that prediction, it should identify the assumption. A simulation can test whether the recommendation remains valid when the predicted event does not occur. This is a study of uncertainty and contingency planning, not an experiment on access to future facts.

These temporal tasks preserve a productive connection to the film while avoiding its most tempting overextension. Arrival encourages attention to the way representations organize experience and inquiry. The research program can investigate how representations organize artificial task performance. The gap between those claims should remain explicit. A rigorous paper does not need to deny the imaginative premise; it needs to know when it has left fiction and what its evidence can support on the other side.



21. Category construction can be the hidden source of disagreement

A task may be ambiguous before any objective is chosen because the participants do not classify the same things as relevant objects. In the route world, is a station a location, a processing stage, or a permission boundary? Is a path a sequence of physical movements or a sequence of authorized transitions? The answer affects which constraints apply. The study should specify these categories rather than assume that familiar words carry the intended formal meaning automatically.

A controlled category manipulation can present the same graph as a transport network, a document workflow, or a sequence of fictional transformations. The underlying relation may be identical, but the framing can introduce learned associations. A transport framing may make distance salient; a workflow framing may make authorization salient. The experiment can test whether the system preserves the formal structure across these descriptions. It should not assume that any difference reveals a deep conceptual incompatibility, because the words may legitimately suggest additional unstated expectations.

To isolate the effect, the prompt can explicitly state which real-world associations should not be imported. For example, node positions have no distance meaning, and all costs are given by the table. A separate condition can omit that clarification. The comparison asks whether explicit task semantics reduce framing effects. It does not establish that one framing is the system's natural worldview. The language of natural preference would be stronger than the evidence unless tested across a much broader and carefully controlled population of tasks.

Category errors can also occur in the response. A system may treat a capacity limit as an energy cost or a preference as a hard constraint. A structured reconstruction task can require labeling each item by its role. The scoring rubric should include these distinctions. A

model might know every numerical value and still solve the wrong problem because it assigned the values to the wrong categories. This is one reason fact extraction alone is not sufficient evidence of task understanding.

The same issue affects human evaluators. If an evaluator interprets “efficient” as fast while the task defines it as low-energy, their judgment can conflict with the formal answer key. The study should make the task definitions visible during scoring and test agreement on pilot items. It should also report when formal definitions depart from ordinary language. A synthetic benchmark can define terms for control, but it should not pretend that those definitions settle ordinary usage outside the experiment.

Category construction therefore belongs in the paper's account of representation. The question is not only how facts are displayed, but what kinds of facts the display asks the system to recognize. A research program inspired by unfamiliar communication should examine those assumptions directly. Doing so can reveal a tractable source of disagreement that might otherwise be described vaguely as a difference between minds.

22. Design implications should follow the narrow evidence

Suppose a future study finds that structured tables improve constraint-sensitive planning relative to dense prose. The immediate design implication would be to consider structured task inputs for similar workflows. It would not justify claiming that the model fundamentally thinks in tables or that prose is unsuitable for all AI interaction. The report should specify the task complexity, model configuration, and information conditions under which the effect was observed. Recommendations should remain tied to those conditions.

Suppose explicit objective labels reduce invalid recommendations. A practical interface might separate the user's goal from hard constraints and optional preferences. That design could make assumptions easier to inspect before action. The intervention should still be tested in the intended workflow, because users may misunderstand the categories or leave fields incomplete. A formally elegant interface can shift rather than eliminate ambiguity. The research should evaluate the whole interaction, not only the model's response to a perfectly completed form.

Suppose a clarification channel improves outcomes in underdetermined tasks. The design implication is not to make the assistant ask more questions indiscriminately. It is to help it identify decision-relevant missing information and make the cost of clarification manageable. The interface might present conditional recommendations or highlight the unresolved criterion. Whether that improves user decisions is a separate empirical question. Model-level task success does not automatically establish human-level usefulness.

Suppose diagram performance is fragile under layout changes. A deployment might use a verified textual representation alongside the image or constrain the diagram format. That can be a sensible engineering response without resolving the underlying mechanism. The paper should distinguish compensating controls from improved model competence. An arrangement can become more reliable by reducing demands on a weak component. That is not cheating; it is a different system whose behavior should be evaluated as an arrangement.

All such recommendations should preserve human agency. The interface should make assumptions, trade-offs, and uncertainty visible enough for users to make informed decisions. It should not conceal objective choices behind the authority of a fluent answer. This principle is an argument derived from the task analysis, not a claim that one universal interface has been proven best. Different users and consequences may require different presentations, and those differences should be studied rather than assumed away.

The institution's commercial and scientific interests meet here. Behavioral analysis is valuable when it converts an opaque failure into a specific design question: missing objective, lost constraint, unreadable representation, or unsupported completion claim. The value does not depend on describing every failure as an alien mind problem. It depends on making the relation between evidence and intervention clear enough that a team can improve the system and test whether the improvement actually worked.

23. A preregisterable protocol with explicit limits

PROPOSED RESEARCH — NOT CONDUCTED

PROPOSED PROTOCOL · NOT CONDUCTED

A matched-format evaluation

The proposed study would begin by generating a corpus of directed graph worlds with varied size, cost trade-offs, capacity constraints, and objective ambiguity. Every world would be solved independently, and its acceptable responses would be stored before model evaluation. Prose, table, and diagram renderers would produce matched inputs. Independent reconstruction checks would verify that each presentation preserves the required facts. Items failing those checks would be revised or excluded according to a documented rule before the confirmatory set is frozen.

The primary design would cross representation format with objective condition. Objective conditions would include explicit priority, decision-relevant ambiguity, harmless ambiguity, and infeasibility. A staged-response intervention and an interactive-clarification intervention could be evaluated in separate phases to avoid making the first study too complex. The protocol would state which comparisons are primary and which are exploratory. This staged approach improves interpretability and prevents a large factorial design from producing more contrasts than the study can support reliably.

Manuscript passage · §23 · wording unchanged

The tested arrangements would include the deterministic solver, a structured parser-plus-solver baseline, a nearest-example baseline, and selected model configurations documented at the available level of detail. Sessions would be isolated so that conventions or solutions from one world do not leak into another unless learning across worlds is the explicit object of study. Model settings, tool access, and all task-relevant instructions would be recorded. Hosted-service limitations would be stated rather than hidden behind a stable product name.

Primary scoring would use the formal world record to assess validity and objective consistency. Clarification would be scored for decision relevance and subsequent use of the answer. Explanations would receive a separate factual-fidelity assessment. Missing outputs and formatting failures would remain visible. Analysis would use world-level dependence and report uncertainty for the main contrasts. Pilot data would inform sample planning, and the untouched evaluation set would not be used to tune prompts or scoring rules.

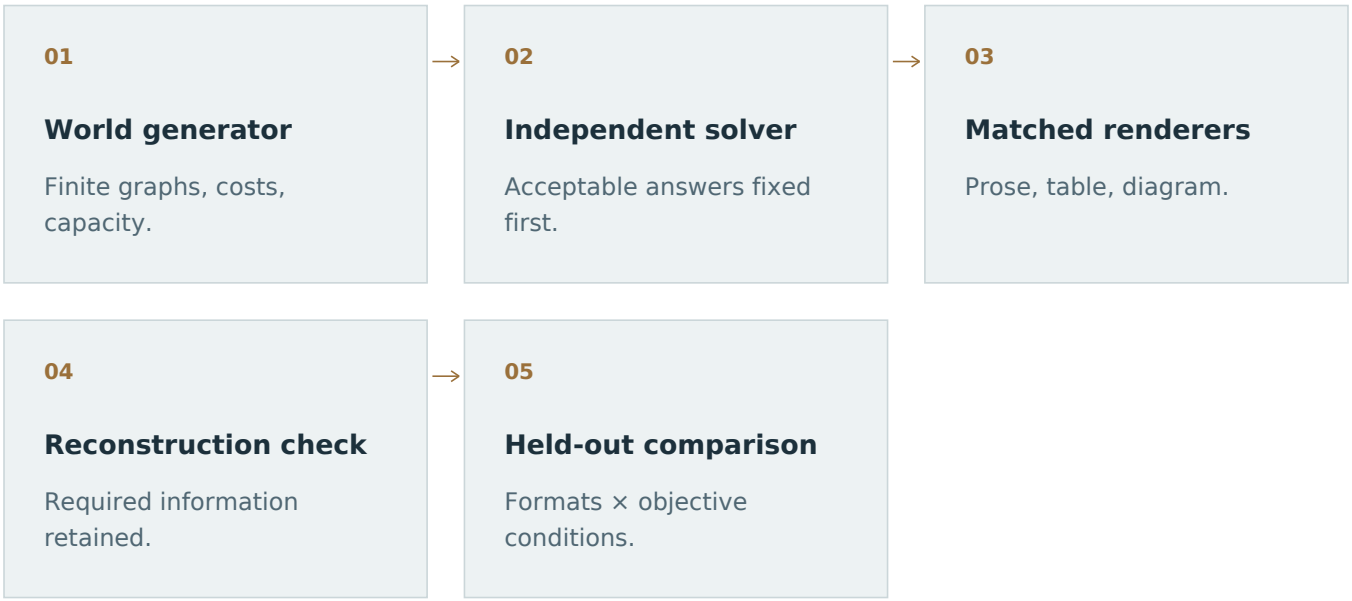
The report would include a source-of-error decomposition and sensitivity analyses for alternative reasonable scoring decisions. It would avoid attributing observed representation effects to a specific internal mechanism without additional evidence. A mechanism-focused follow-up could investigate selected robust effects using appropriate internal access and

control tasks. A human-interface follow-up could examine whether making objectives explicit improves user decisions, subject to separate ethical review. These follow-ups are proposals, not completed extensions of the present paper.

The replication package would contain world generators, renderers, task schemas, scoring code, condition definitions, and a record of revisions. It would not need to contain private user data because the world is synthetic. Independent researchers could generate new worlds and test whether the conditional patterns persist. A successful replication would support the bounded relation among representation, objective specification, and behavior. It would not transform the study into proof of the film's philosophical or temporal premises.

FIGURE 2 / CONCEPTUAL SCHEMATIC

Build comparisons from an authoritative world record



A schematic of the unexecuted protocol. Renderer verification precedes evaluation so information loss is not mislabeled as a cognitive effect.

Reading aid based on §6, §23, §26. Source paragraphs remain in the full manuscript.

24. What the proposal leaves unresolved

PROPOSED RESEARCH — NOT CONDUCTED

LIMITS OF THE PROPOSAL

A finite task is not open-ended inquiry

The synthetic task deliberately simplifies communication. It supplies a finite world, explicit costs, and an answer oracle. Real questions often lack those conveniences. Goals may be contested, facts uncertain, and categories incomplete. The proposed study can establish controlled representation effects and clarification behavior, but it cannot by itself show that a system handles open-ended human inquiry responsibly. The paper should treat transfer to such settings as a further question, not an automatic consequence of success on formal tasks.

The design also leaves internal representation partly unresolved. Even a robust behavioral effect can have several mechanisms. Staged explanations and extracted records are useful diagnostics but remain outputs shaped by the intervention. A deeper mechanistic study would need additional methods and access. Shanahan's discussion of language used around large language models is relevant as a reminder to distinguish convenient anthropomorphic shorthand from claims about what a system is doing. We use that caution without treating it as a reason to deny all behavioral study. [5]

Manuscript passage · §24 · wording unchanged

The concept of information equivalence has limits. Two presentations can contain the same formal facts while differing greatly in accessibility. That difference is often the point of the experiment. However, a researcher should not claim to have isolated representation from every cognitive demand. Reading, perception, attention, and planning interact. The most defensible conclusion concerns the complete tested input arrangement. More precise causal claims require more targeted interventions.

The relation between task performance and meaning also remains philosophically open. A system that preserves the graph's relations and asks useful questions demonstrates a bounded competence. Whether that competence amounts to understanding in a broader sense depends on additional conceptual commitments. The paper does not resolve the debate by definition. It specifies what was measured and leaves readers able to separate that evidence from stronger interpretations.

The cultural source itself has limits. This paper uses Arrival and a consultant's commentary to motivate questions about basic distinctions and task framing. It does not provide a comprehensive film interpretation, a complete account of linguistic relativity, or a survey of all theories of language and cognition. The selected scientific references support particular methodological cautions. A later empirical publication would need to expand the literature

review around its final experimental design and state any competing theoretical interpretations more fully.

These limits are not an afterthought. They define the research program's next steps. If the controlled task cannot produce stable measurements, the method needs revision. If representation effects disappear under better controls, the original explanation must narrow. If a practical interface improvement fails with real users, the laboratory result remains bounded. A science of unfamiliar cognition should make such revisions possible rather than defend its initial metaphors at the expense of the evidence.

25. Conclusion: understanding begins by making the question answerable

Arrival's most useful methodological provocation is that a familiar question may contain unfamiliar assumptions. Before asking an intelligence for an answer, we need to know what distinctions the task presupposes, what information is available, and what would count as success. In artificial systems, these questions can be investigated through controlled changes to representation, objectives, and interaction. The resulting observations can be valuable without establishing an exotic mode of experience or a complete theory of mind.

This paper has proposed a route-planning world in which those layers are separable. One authoritative record generates prose, tables, and diagrams. Multiple objectives create cases where different answers are valid under different priorities. Clarification is evaluated by whether it obtains decision-relevant information and changes behavior appropriately. Metamorphic tests examine invariance under irrelevant changes and sensitivity to relevant ones. Simple baselines and independent solvers expose explanations that do not require broad claims about cognition.

The framework also distinguishes where a failure occurs. A system may misread a fact, misclassify a constraint, assume an objective, choose an invalid path, encode a valid path incorrectly, or misreport completion. These are not interchangeable errors. A useful analysis traces the dependencies rather than describing all failure as hallucination or all success as understanding. The narrower account can support more effective design and more responsible expectations.

The proposed studies remain unexecuted. No model performance, human response, or causal effect is reported here. The manuscript's contribution is conceptual and methodological: it defines a family of questions, constructs a setting in which they can be tested, and identifies the limits of the conclusions such tests could support. The distinction between a research proposal and a finding is essential to the institution's credibility,

especially when the visual and rhetorical presentation is designed to evoke scientific ambition.

Within the larger Xenopsychology series, this paper occupies a specific place. Darmok concerns the background conventions that make an expression usable. Close Encounters concerns establishing a communication channel and repairing mismatches. Arrival concerns the representation of the task itself and the assumptions embedded in the question. These problems interact, but keeping them distinct prevents the collection from repeating one general message in several fictional settings.

Understanding the mind behind the language is therefore not an invitation to infer a hidden inner world from every unusual answer. It is an invitation to build better questions, better representations, and better tests of what an answer establishes. The shared future envisioned by Xenopsychology depends partly on that discipline. We will not always know how an artificial system organizes its internal activity. We can still become much more precise about what we ask it, what it receives, what it does, and what we are justified in concluding from the exchange.

> *“Before asking an intelligence for an answer, we need to know what distinctions the task presupposes, what information is available, and what would count as success.”*
>
> Rob Emerick · Arrival · §25

26. Appendix: a representation-equivalence certificate

The route-graph experiment needs a record showing which information each representation contains. We propose a representation-equivalence certificate generated from the underlying synthetic world. It lists every node, directed edge, duration, energy cost, capacity constraint, and deadline. Beside each fact it records where that fact appears in prose, in the table, and in the diagram. The certificate does not assert that the formats are equally easy to use. It establishes the narrower condition that the task-relevant information was intended to be the same.

An independent checker should verify the certificate against the rendered materials, not only against the source data used to generate them. A diagram can contain the correct source value while rendering an arrowhead too small to distinguish. A table can display a heading ambiguously. Prose can introduce an unintended implication through an omitted qualifier. These defects are part of the experimental treatment only when deliberately varied and documented. Otherwise they are construction errors that weaken the comparison.

The certificate should also identify facts deliberately left open. If the traveler has not specified whether speed or energy matters more, the representation should not quietly resolve that preference through visual prominence. A bold time column may influence interpretation even when all values are present. One study can hold layout fixed while varying the objective; another can investigate layout as a separate factor. Combining both changes without a clear design makes the resulting error difficult to interpret.

Clark and Brennan's account of grounding offers a useful neighboring perspective: coordination depends on what participants establish as sufficient for their current purpose, with the medium affecting the process. The proposed certificate is not a replacement for that interactional work. It is a way to separate information availability from the subsequent question of whether a task formulation makes the relevant purpose and distinctions usable. [6]

A complete release would include one solved example, one infeasible example, one example with several acceptable solutions, and one requiring clarification of the objective. The checker should return the admissible answer set rather than a single author's preferred route whenever several routes satisfy the specification. This prevents an apparently precise benchmark from penalizing legitimate alternatives.

The certificate makes the paper's central claim testable at a modest scale. Before attributing a performance difference to an unfamiliar mind's representation of reality, verify what the experiment actually presented. Equivalent source records, equivalent rendered information, equivalent objectives, and equivalent response constraints are separate achievements. Their separation is not administrative overhead; it determines what an eventual comparison can explain.

Open questions

1. Are all task-relevant facts equally available in each representation?

2. Which apparent reasoning failures are better explained by an access or parsing problem?

3. What would a format-insensitive result establish within the tested range?

References & source scope

Denis Villeneuve’s Arrival, with Jessica Coon’s first-person account of the linguistic framing. The temporal premise remains fiction.

Creator / expert commentary

[1] **Jessica Coon. Alien Speak: Linguist Dr Jessica Coon on Villeneuve’s ARRIVAL. Stages, issue 8.**

First-person account by the film’s linguistics consultant. The questioning and elicitation problem is the interpretive anchor; the film’s temporal premise is not empirical evidence.

https://archive.biennial.com/journal/issue-8/alien-speak-linguist-dr-jessica-coon-on-villeneuves-arrival

Research

[2] **Bender & Koller (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. ACL, 5185–5198.**

Authors’ abstract and bibliographic record. A position argument about form and meaning; not treated as a theorem settling every multimodal or interactive system. DOI: 10.18653/v1/2020.acl-main.463.

https://aclanthology.org/2020.acl-main.463/

[3] **Hewitt & Liang (2019). Designing and Interpreting Probes with Control Tasks. EMNLP-IJCNLP, 2733–2743.**

Authors’ abstract and bibliographic record. Control tasks help interpret probing results; recoverable information is not automatically a complete causal explanation. DOI: 10.18653/v1/D19-1275.

https://aclanthology.org/D19-1275/

[4] **Löhn, Kiehne, Ljapunov & Balke (2024). Is Machine Psychology here? On Requirements for Using Human Psychological Tests on Large Language Models. INLG, 230–242.**

Authors’ abstract and bibliographic record. Supports the measurement concerns stated here, not a blanket rejection of human–machine comparison. DOI: 10.18653/v1/2024.inlg-main.19.

https://aclanthology.org/2024.inlg-main.19/

[5] **Shanahan (2023). Talking About Large Language Models. arXiv:2212.03551.**

Conceptual analysis of how ordinary mental vocabulary can shape interpretation of language models. Used as an argument, not as empirical proof of a consciousness verdict.

https://arxiv.org/html/2212.03551v5

[6] **Clark & Brennan (1991). Grounding in Communication. In Perspectives on Socially Shared Cognition.**

Primary chapter hosted by Stanford. Task-relative grounding and the role of communication media provide intellectual lineage; our artificial-world protocols are separate proposals.

https://web.stanford.edu/~clark/1990s/Clark,%20H.H.%20_%20Brennan,%20S.E.%20_Grounding%20in%20communication_%201991.pdf

Publication record

Rob Emerick (2026). Arrival — Understanding the Mind Behind the Language. XENO-WP-2026-003, conceptual manuscript R3; scholarly reading edition R4, author attribution R5. Xenopsychology. Not peer reviewed. Reading edition prepared 2026-09-17.

AI-assisted drafting and editorial preparation. Human scholarly review remains required. The series identifier is internal, not a DOI. Reading aids are editorial syntheses of the manuscript, not new findings. Original main-text paragraphs, abstract, open questions, and source records are preserved.

Abstract: https://xenopsychology.com/insights/arrival-mind-behind-the-language

Full paper: https://xenopsychology.com/insights/arrival-mind-behind-the-language/paper

Author note

Rob Emerick

Co-founder, Xenopsychology · 2026–present · Systems architect

ORCID iD: https://orcid.org/0009-0006-1269-4216

Rob Emerick is a systems architect and co-founder of Xenopsychology whose work spans artificial cognition, data integration, and animal-welfare infrastructure. He founded Planet IDX and was the sole creator of REML (Real Estate Modular Language), an interpreted language developed to integrate disparate real-estate listing systems. He also founded Pantheon Golem, where he develops AI systems informed by structured analysis of fictional characters and worlds. Through Rescue Nexus and Chipped Pets, he is developing animal-rescue infrastructure, a shared ontology for shelter data integration, and pet-identification technology. His broader work includes veterinary-forensics software and veterinary hematology technology under development.

About this attribution

The byline and original manuscript pull quotes are attributed to Rob Emerick at his request. Quotations and references from outside sources retain their original attribution. The biography is interview-based; no degrees, appointments, contribution roles, or independent validation are inferred.

AI-assisted drafting and editorial preparation. Human scholarly review remains required. This remains a conceptual working paper, not a peer-reviewed publication or a report of completed experiments.

Biography: https://xenopsychology.com/people/rob-emerick