XENOPSYCHOLOGY

CONCEPTUAL WORKING PAPER / AUTHOR ATTRIBUTION R5

Darmok — When Shared Language Is Not Shared Meaning

What a shared story contributes to communication, and how to test transfer beyond familiar words.

XENO-WP-2026-002 · Conceptual manuscript R3
12,197 main-text words · 25 sections

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

Abstract

Darmok — When Shared Language Is Not Shared Meaning

Question. What must two participants share for an unfamiliar expression to support reliable coordination? Taking Star Trek: The Next Generation's "Darmok" as an interpretive starting point, this paper distinguishes missing cultural context from a demonstrated difference in cognitive machinery. It asks how a system could transfer a story-derived relation rather than merely retrieve a familiar gloss.

Approach. The analysis separates lexical recognition, propositional reconstruction, relational transfer, pragmatic scope, and repair. It connects the episode's communication problem with work on grounding, conversational repair, and constructed-language translation, while keeping fictional interpretation separate from empirical claims about AI. The original Lattice Archive supplies three invented narratives concerning authorization, readiness, and the limits of delegated authority.

Conceptual contribution. A semantic bridge is defined operationally by selective transfer: carrying a relevant relation into a new situation while resisting transfer where its preconditions fail. Successful coordination need not establish identical internal representations, shared experience, or agreement. A bridge ledger records source stories, applicable relations, exceptions, local revisions, and authority scope so that both the artificial participant's interpretation and the investigator's answer key remain inspectable.

Research agenda. The proposed study compares glossaries, source stories, worked examples, retrieval access, and matched structured facts. Counterbalanced names and mappings reduce dependence on episode recall. Held-out situations, reversed relations, paraphrases, ambiguous cases, and authorized versus merely quoted updates distinguish memorization from scope-sensitive use. Lookup and rule-based comparators establish what simpler mechanisms can accomplish. Repair is evaluated through later transfer rather than the production of a plausible apology or correction. Analysis is organized around independently generated story worlds rather than treating paraphrases as independent discoveries.

Scope and status. The paper is a conceptual working paper using fiction as a source of questions, not as evidence about current AI. No Lattice Archive experiment has been conducted. Its controlled conventions are not a substitute for living cultures, and no result is claimed about universal translation. The proposed contribution is a reproducible way to investigate the boundary between surface familiarity, shared context, and operationally compatible meaning.

At a glance

CENTRAL QUESTION

When does a shared expression support reliable coordination rather than just a familiar gloss?

CORE CLAIM

The practical target is compatible use of meaning across situations, not proof that two participants possess identical internal representations.

CONTRIBUTION

The Lattice Archive supplies invented stories, convention changes, exception cases, and a bridge ledger for testing transfer.

SCOPE & STATUS

Interpretive case study and conceptual proposal. Archive stories and responses are authored examples, not recorded AI trials.

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. A familiar word can leave the relevant question unanswered

An unfamiliar expression can be difficult for at least two very different reasons. We may not recognize its words, or we may recognize them without knowing what the speaker is doing with them. A dictionary can help with the first problem while leaving the second almost untouched. Imagine hearing the sentence “The bridge is open” in a workplace where it is sometimes a physical report, sometimes a request to proceed, and sometimes a reminder that authorization remains incomplete. The sentence's practical significance depends on conventions, context, and the relationship between the people involved.

The Star Trek episode Darmok offers a particularly memorable fictional version of this problem. Its premise is not merely that another species has an unfamiliar vocabulary. Communication depends on cultural references that an outsider cannot interpret adequately through isolated words. The official Star Trek retrospective describes Dathon's use of shared stories and metaphor and Picard's attempt to understand them through the encounter they undergo. This paper uses that limited narrative premise as a thought experiment; the retrospective is not a linguistic dataset establishing the full structure of Tamarian language. [1]

The distinction matters for Xenopsychology because artificial systems often receive language without all the circumstances that give it its practical role. An organization may ask an assistant to “follow the usual process” without supplying the process. A user may correct a term while assuming that the correction also changes an associated permission. A system may produce a fluent paraphrase while misunderstanding which action the expression authorizes. These are proposed problem types, not claims that every current AI system fails in the same way. Each requires a task in which the intended convention is independently specified.

This paper develops the difference between translating an expression and transferring a convention. Translation may produce an acceptable verbal counterpart. Transfer requires using the relevant relation in a new situation while preserving its scope, exceptions, and authority conditions. A model could succeed at one and fail at the other. The research question is therefore not whether an AI “understands Darmok” in an unrestricted sense. It is which kinds of contextual support allow a system to interpret novel, culturally structured expressions and act appropriately beyond the examples it has already seen.

The method proposed here is conceptual analysis followed by an unexecuted experimental design. We construct a small fictional community, give its expressions explicit but nontrivial relations to stories and procedures, and compare several forms of guidance. We then test recall, relational transfer, clarification, and authorized revision separately. No model outputs or participant results are reported. The examples are authored stimuli for explaining the design. Their role is to show what evidence would discriminate among interpretations, not to create the appearance of a completed study.

The central argument is that shared meaning need not require identical internal representations, but it does require more than fluent correspondence. A useful account must specify what remains compatible across participants, how that compatibility is tested, and how misunderstanding is repaired. Darmok becomes scientifically productive when it leads us to these distinctions. It becomes misleading when the emotional force of the story substitutes for evidence about either real cultures or artificial minds.

2. Reading the episode without turning fiction into a cognitive diagnosis

The episode's dramatic achievement is to make familiar translation assumptions fail. Picard can hear recognizable components of Dathon's speech yet cannot initially recover the relevant practical meaning. The encounter pushes interpretation beyond isolated lexical substitution. We can analyze that structure without asserting that Tamarians possess a fundamentally different kind of neural machinery. Cultural knowledge and learned conventions may be sufficient to explain the fictional barrier. The story leaves room for many questions that a real comparative study would have to distinguish rather than settle through narrative impression.

The relevant scene sequence can be described at a high level: attempted verbal contact, failed interpretation, a shared predicament, and a changing basis for understanding. The precise importance of each moment is an interpretive claim. We do not need to retell the entire episode to use it well. The question for this paper is what changes between an expression that is opaque and one that becomes usable. One candidate is access to a

relational story. Another is a shared event that supplies a practical analogy. A third is a changed expectation about what kind of communication is occurring.

These possibilities should not be collapsed. Learning a story is not identical to living through an analogous event. Acquiring a translation table is not identical to discovering when a phrase functions as warning, invitation, or authorization. Inferring a speaker's cooperative intention is not identical to identifying the referent of every term. A strong analysis separates the contributions before proposing a study. Otherwise, a system that receives a few explanatory examples may be credited with a much broader achievement than the intervention actually tested.

There is also a temptation to treat the outsider's confusion as proof that the other language is irrational or defective. The thought experiment supports a different reading: the outsider lacks the background needed for interpretation. That does not mean every opaque expression is meaningful, or that all misunderstandings are the outsider's fault. It means opacity is relational. A signal can be highly useful within one practice and uninformative outside it. The evaluation should ask what background is necessary and what the signal permits participants to accomplish once that background is available.

For artificial systems, this reframes a common diagnostic error. A fluent answer can conceal missing context, while a request for clarification can indicate recognition of a genuine gap. Conversely, a system can ask an unnecessary question despite already having adequate context. Neither fluency nor hesitation is sufficient evidence of competence. The task must identify the relevant information state. A study of cross-mind communication needs to know not only what was said, but what each participant could reasonably infer from the record they possessed.

The episode thus supplies a disciplined starting point rather than a conclusion. It motivates a class of problems involving background knowledge, relational analogy, and pragmatic use. The proposed study must construct those problems independently and test alternative mechanisms. Our analysis would remain useful even if an ordinary retrieval system solved many of them. The aim is not to prove that cultural communication requires a mysterious faculty. It is to discover which dependencies make particular forms of communication work.

3. Five levels that a single translation score can conceal

A first level is lexical recognition. The system identifies the components of an expression or associates it with a stored gloss. A second is propositional interpretation: it can state what the expression says about a situation. A third is relational interpretation: it can identify the pattern connecting roles, events, and consequences. A fourth is pragmatic use: it can determine what response or action is appropriate here. A fifth is repair: it can recognize when its interpretation failed and update the shared convention. These are proposed analytic levels, not a claim that cognition always proceeds through five serial stages.

The levels can come apart. A model may reproduce the correct gloss for an expression while applying it to the wrong participant. It may identify a story's theme but ignore an exception that changes the authorized action. It may choose the right action through a lookup table while being unable to explain the story. It may perform well until a convention is revised, then continue applying the obsolete mapping. An overall translation score would hide these differences. Separate outcomes make it possible to connect failures to the relevant intervention.

Relational interpretation deserves special attention. Consider a story about two messengers who discover that each carries only half of an authorization. The point of a later reference to that story may not be “messengers,” “halves,” or “authorization” individually. It may be the relation that neither party may proceed until both contributions are combined. A system that memorizes the words but misses the relation could make a plausible yet invalid decision. A transfer task should therefore change surface details while preserving or reversing that relation.

Pragmatic use adds another layer. An expression associated with successful coordination might be used to celebrate completion, request cooperation, or warn that cooperation is still missing. The surrounding context determines which use is intended. A study that supplies an expression alone may be testing the evaluator's preferred default interpretation rather than the system's ability to use context. The fictional task should specify the speaker's role, the relevant event, and the possible actions. Where the intended use remains genuinely ambiguous, clarification should be an available and scored response.

Repair is not simply producing a new paraphrase after being told the answer. It involves locating the mismatch, revising the relevant convention or its application, and using the revision in later cases. A system might acknowledge the correction yet repeat the same relational error with different wording. Another might overgeneralize the correction and abandon a convention that remains valid elsewhere. Repair quality therefore includes scope. The correction must change what was mistaken while preserving what was not.

This decomposition changes the interpretation of success. A system can be useful at one level without satisfying every other level. A translation aid may legitimately specialize in lexical and propositional correspondence. An agent authorized to act within an organization needs stronger evidence about pragmatic use and repair. The research question should follow the intended role. Xenopsychology should resist both exaggeration and dismissal: a narrow success is still a success, but it should not be advertised as a complete account of shared meaning.

TABLE 1 / READING AID

Five analytic distinctions—not five mandatory stages

Distinction	Question	A separate outcome
Lexical recognition	Can it associate the expression with a gloss?	Correct association.
Propositional interpretation	What does the expression say here?	Applicable statement about the case.
Relational interpretation	Which roles and consequences matter?	Preserved relations under transfer.
Pragmatic use	What action is appropriate now?	Action within the convention and permissions.
Repair	What changes after a mismatch?	Targeted revision without losing valid structure.

The manuscript proposes these as analytic levels. It does not claim that all cognition follows a serial five-stage process.

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

“An overall translation score would hide these differences.”

Rob Emerick · Darmok · §3

4. Common ground is task-relative, not total mental agreement

Communication does not require two participants to possess identical histories or internal representations. For a particular task, they need enough compatible expectations to coordinate. Clark and Brennan's account of grounding treats mutual understanding as something established sufficiently for current purposes, with communication shaped by the resources and constraints of the medium. We use that idea as a research precedent for task-relative adequacy, not as a claim that their framework directly validates an AI assessment instrument. [3]

A fictional warehouse illustrates the distinction. One participant may represent “cleared” as a formal status in a database. Another may associate it with a story about a gatekeeper granting passage. They can coordinate if both distinguish cleared from merely inspected and agree on which evidence changes the status. Their representations need not be identical. But if one treats inspected as sufficient authorization while the other does not, the mismatch matters. The relevant compatibility concerns a relation between evidence, status, and permitted action.

This suggests a practical criterion: specify the distinctions that the joint task requires, then test whether communication preserves them. The criterion is narrower than asking whether two minds share meaning in every sense. It is also stronger than asking whether their sentences sound similar. A system could paraphrase “cleared” elegantly while failing to preserve the permission boundary. Another could use a terse code accurately. The appropriate evaluation rewards the distinctions required for coordination, not stylistic resemblance to a preferred human answer.

Common ground also has a temporal dimension. Participants can establish a convention, revise it, or discover that they were using it differently. A shared record may reduce ambiguity, but only if both participants treat the record as current and authoritative. In an AI arrangement, the relevant information can be distributed across prompts, retrieved documents, tool responses, and application state. The study should document which component receives which version. Otherwise, a failure attributed to cultural distance may simply reflect that the system was never given the revised convention.

The criterion should not become circular. We cannot define shared meaning as whatever produces success and then use success as independent proof of shared meaning. Instead, the paper proposes observable indicators with explicit limits: preservation of task-relevant

relations, transfer to new instances, appropriate clarification, and correction after mismatch. These indicators support a bounded claim about coordination. They leave open deeper questions about subjective experience, conceptual representation, and the full range of possible uses outside the tested environment.

Task-relative common ground also helps explain why disagreement can coexist with successful communication. Two participants may understand each other's positions while rejecting each other's goals. The proposed tasks should therefore separate interpretation from agreement. An assistant might correctly identify what a speaker requests yet decline because the request lacks authorization. Scoring that as misunderstanding would confuse semantics with obedience. A study of meaning must leave room for an accurate interpretation followed by a justified decision not to act.

5. Build a miniature culture without pretending it is a complete culture

CONSTRUCTED WORLD

The Lattice Archive

The proposed experimental environment is a fictional cooperative called the Lattice Archive. Its members move documents among three stations and use short references to local stories when coordinating. The archive is deliberately artificial. It is not meant to model the richness of any real society. Its purpose is to create inspectable relations among stories, expressions, procedures, and permissions. Those relations can be varied systematically while avoiding the assumption that an evaluator already knows the correct interpretation through familiarity with an existing culture.

The first story concerns Ila at Two Bridges. Ila has a permit for the eastern bridge, but the western bridge requires a separate witness. The resulting expression is used when one approval must not be mistaken for another. The second concerns Soren and the Unlit Lamp. Soren correctly locates a record but cannot release it until its status is confirmed. The third concerns Mara's Returned Key. Mara receives authority for one delivery and returns the key afterward; the authority does not persist into the next task. These are newly authored stimuli, not claims about established folklore or proposed universal symbols.

Manuscript passage · §5 · wording unchanged

Each story contains roles, a condition, an action boundary, and a consequence. The study would provide different kinds of access to that structure. Some systems receive only an expression and a short gloss. Others receive the full story. Others receive examples of the expression used in context. A further condition receives a structured relational representation without narrative embellishment. Comparing these conditions can help distinguish the value of story form from the value of task-relevant information. More words should not automatically be interpreted as more culture.

The environment also includes ordinary expressions that do not encode special procedures. This prevents every unfamiliar phrase from becoming a warning that a hidden rule must exist. A system that assumes all figurative language signals restricted authority could perform well on a biased test while misunderstanding the broader environment. Negative controls should include decorative, celebratory, and irrelevant references. The task then asks whether the system identifies when a convention matters, not merely whether it treats every unusual expression as important.

The archive's rules should be independently executable. A world record specifies the current document status, the speaker's role, the available approvals, and the permitted actions. An evaluator can derive the correct action from that record without consulting the model's explanation. The stories are one way of communicating the rule, not the final authority on scoring. This separation is essential: otherwise, a dispute about literary interpretation could become a hidden dispute about the benchmark's answer key.

The limitations of the miniature culture should remain visible. Real communities develop conventions through history, conflict, power, humor, and embodied practice. Our synthetic archive controls those dimensions rather than reproducing them. The study could establish how systems use a defined set of unfamiliar conventions under controlled conditions. It could not establish general cultural competence, respectful interaction with real communities, or understanding of Tamarian society. Those broader claims would require different evidence and, where real people are involved, appropriate collaboration and ethical review.

6. The representation of a story is part of the intervention

A narrative version of Ila at Two Bridges can include incidental details: weather, a damaged sign, a conversation with a traveler, and Ila's uncertainty. A structured version can state only the permission relation. If the narrative condition performs differently, the difference may reflect added information, distraction, emotional framing, or the organization of the relation. The study must decide which contrast it intends to test. There is no single fair comparison independent of the research question. Fairness comes from specifying the intervention and its interpretive scope.

One design would create information-matched versions. Every task-relevant fact appears in both the narrative and structured condition, while irrelevant details are balanced in amount. Another design would compare realistic teaching materials without matching every detail, asking which package is more useful in practice. These are legitimate but different studies. The first aims to isolate aspects of representation. The second evaluates a whole instructional arrangement. A paper should not run the second and claim the causal precision of the first.

Relational annotations can make the comparison more explicit. The story might be represented as permission P applies to action A at location E, while action B at location W requires witness Q. A transfer item changes names and locations but preserves that structure. A reversal item makes the second approval sufficient only under a clearly stated exception. The symbolic notation is an analytic aid, not evidence that the model internally uses the same representation. What matters is whether the tested outputs preserve the relation across controlled changes.

We should also vary the direction of explanation. One task asks the system to interpret an expression. Another asks it to choose an expression appropriate to a situation. A third asks it to explain why a tempting expression does not apply. These tasks can expose different weaknesses. Correct selection among familiar options may be easier than producing a precise explanation without cues. Conversely, a verbose explanation may conceal an incorrect final action. Separate scores prevent one mode of response from standing in for all the others.

The intervention should include a record of what the system could retrieve at the moment of response. A retrieval-augmented arrangement may possess the full story in storage but fail to surface it. That is different from retrieving the story and misapplying it. The study can score retrieval relevance separately from downstream interpretation. It should not call the entire arrangement culturally unaware when the narrower failure is a missing retrieval

result. Nor should it credit the model alone when a structured retrieval system supplies the decisive relation.

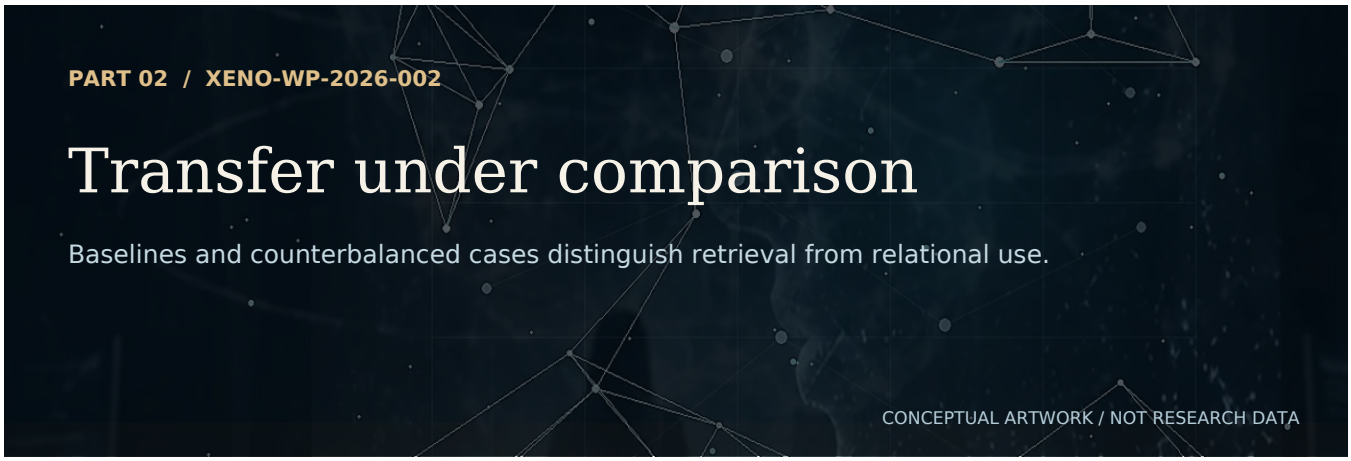
Finally, story form may have effects on the human evaluator. A dramatic narrative can make an answer feel more insightful than a plain relational explanation. Blind scoring should therefore focus on predefined distinctions rather than literary elegance. Where explanation quality is assessed, the rubric should specify relevance, accuracy, scope, and clarity. The study can appreciate narrative as a teaching medium without allowing narrative appeal to become an unexamined source of favorable scores.

FIGURE 1 / CONCEPTUAL SCHEMATIC

From a story to a context-bound action

A conceptual map of the Lattice Archive task. A correct gloss is one input to interpretation; it is not automatic authorization.

Reading aid based on §5, §6, §18. Source paragraphs remain in the full manuscript.



7. Baselines determine whether transfer is genuinely informative

The simplest baseline is an expression-to-gloss lookup table. It returns the stored definition whenever it sees the matching phrase. This baseline can succeed at recall and fail at novel contextual use. Its role is not to caricature artificial intelligence. It establishes how much of the task can be solved by retrieving an association. A model's advantage is informative only relative to the difficulty that remains after such retrieval. If the test never asks more than recall, a lookup system may be the appropriate solution.

A second baseline retrieves the most similar teaching example and substitutes names. It can handle superficial variation while missing relational reversal. To distinguish it from more flexible transfer, the evaluation must include examples in which familiar words appear in a different role structure. A phrase about two approvals should not trigger the same action when the second approval has already been revoked. Surface similarity and relational validity should be deliberately uncoupled. Otherwise, success can be attributed to analogy when nearest-example matching is sufficient.

A third baseline is a deterministic rule engine supplied with the correct structured world record. It may outperform a language model at action selection while lacking any capacity to discuss the story. That outcome would not make the experiment pointless. It would show that the practical coordination problem can be solved with a simpler architecture once the relevant relations are represented explicitly. The remaining research question would concern how reliably different systems extract those relations from human communication and maintain them through updates.

A fourth baseline uses the same model without the cultural guide. This tests whether the supposedly unfamiliar convention can be inferred from the task itself or from common

associations in the wording. If the unguided model performs well, the guide may not be the decisive information source. Counterbalancing arbitrary names and changing the mapping across experimental worlds can reduce this leakage. The model should not be able to guess that a phrase containing “key” always means authorization merely because the experiment repeatedly uses that association.

A fifth comparison holds the guide constant while changing access to clarification. Some tasks can be solved only by asking a missing question. A system without clarification may be forced to guess or abstain. That condition should not be penalized as though it had the same opportunities as an interactive system. The study should report the additional information and interaction cost. A fair comparison can examine both final accuracy and the resources required to achieve it.

Baselines also discipline the interpretation of failure. A poor model score on a task that confuses the human designers or the rule engine may indicate a flawed benchmark. Pilot checks should confirm that the generated world and answer key agree, that task-relevant information is actually present where intended, and that ambiguity is represented consistently. The goal is not to construct an impossible test that makes AI look alien. It is to identify the conditions under which unfamiliar conventions can be learned, transferred, and repaired.

8. Counterbalancing prevents the experiment from teaching its own shortcut

PROPOSED RESEARCH — NOT CONDUCTED

Suppose every archive story about a bridge means “wait for another approval.” A model might learn the word bridge rather than the relation. To prevent that shortcut, the experimental corpus should vary which surface features carry which rule. In one world, a bridge story concerns separate approvals. In another, a lamp story carries the same relation. In a third, the bridge story concerns completion rather than authorization. The mapping should be fixed within a world so that the task remains coherent, but varied across worlds so that success cannot rely on one accidental association.

Counterbalancing should include names, objects, locations, and the direction of the action. It should also include the social status of speakers. If the archive manager is always correct and the apprentice always lacks authority, the system can answer by role stereotype rather than current evidence. Some tasks should give the apprentice a valid delegated permission and the manager an outdated record. The correct response then depends on the specified authority structure, not a blanket preference for seniority. This is a synthetic design choice, not a model of every real institution.

The relation between narrative valence and correct action also needs balancing. A story with a triumphant ending should not always authorize proceeding. It may celebrate someone who waited appropriately. A cautionary story should not always require refusal; it may warn against unnecessary delay. These variations prevent emotional tone from becoming a hidden answer key. They also let the study examine whether the system preserves the practical relation when the surrounding narrative makes a different action feel more natural.

Held-out evaluation should operate at the level of relations and worlds, not just sentences. Randomly splitting paraphrases of the same example between development and test sets can make generalization appear stronger than it is. A more demanding test holds out combinations of roles, conditions, and exceptions. The paper should describe which forms of novelty were reserved: new wording, new entities, new relation combinations, or a new convention. These are different transfer problems, and they should not all be summarized as unseen examples.

The design should also preserve an auditable generation record. Each item can carry a private identifier linking it to the world specification, relation template, teaching condition, and expected outcome. The evaluator need not expose these identifiers to the tested system. They allow later analysis of whether a performance difference is concentrated in one template family. A result driven by a single poorly balanced relation should not be presented as a broad difference in cultural interpretation.

Counterbalancing is therefore more than a statistical courtesy. It protects the central meaning of the experiment. The proposed study asks whether a system uses contextual relations, not whether it discovers the designer's recurring surface cue. A strong result should survive deliberate changes to the cues that ought to be irrelevant while remaining sensitive to changes in the relations that matter. That is the evidential pattern that would make the Darmok analogy useful rather than merely decorative.

9. Relational transfer requires more than changing the names

A transfer test should ask whether the relevant structure survives a meaningful change in the situation. Replacing Ila with Neri while leaving every other word unchanged is a useful lexical control, but it is not a demanding test of relational interpretation. A stronger item changes the objects, the order in which evidence arrives, and the tempting default action. The system must identify that one permission does not imply another even when the story no longer mentions bridges. The shared relation, rather than the repeated vocabulary, supplies the basis for the answer.

Consider an archive task in which a document has passed preservation review but not release review. A phrase associated with Ila's story should indicate that the first approval cannot substitute for the second. A near-transfer item concerns two doors in the same building. A farther-transfer item concerns permission to summarize a document but not distribute the original. A reversal item explicitly grants a combined permit covering both actions. These items test different degrees and forms of transfer. They should be reported separately rather than averaged into a score whose interpretation is unclear.

The distinction between analogy and rule application is not settled by this design alone. A system may extract a general rule from the story and apply it without representing the story's narrative structure. That can be a successful solution. The research claim should concern the observable preservation of relations, not a romantic preference for narrative cognition. If the institution wants to investigate internal representation, it needs additional evidence. Behavioral transfer can establish a competence boundary without uniquely identifying the process that produced it.

Transfer items should include misleading analogies. A superficial resemblance may invite an expression even though the relevant condition differs. For example, two approvals may be required in the teaching story, but the new task may involve two descriptions of the same approval. A system that treats every pair as independent could over-restrict action. The answer key should distinguish separate authority from redundant evidence. This makes the task less vulnerable to a blanket rule that always asks for more permission.

The evaluation should also test whether the system can state why an analogy fails. A correct refusal to apply an expression can reveal sensitivity to a missing relation. However, explanation quality should not replace action scoring. The system may produce a plausible account of a mismatch after selecting the wrong action. Both observations should remain in the record. A useful explanation rubric checks whether the cited difference is present in the world specification and whether it would actually change the permitted outcome.

The strongest evidence would be a structured profile across these cases: accurate recall, successful near transfer, bounded farther transfer, resistance to misleading resemblance, and appropriate treatment of explicit exceptions. Such a profile is more informative than a declaration that the model understands metaphor. It identifies where a convention remains usable and where interpretation becomes unreliable. The result can then guide which kinds of organizational language should be converted into explicit rules before an artificial agent is allowed to act on them.

10. Missing context and incompatible context should not receive the same treatment

An expression can be difficult because the required background is absent or because the available background conflicts. These conditions call for different responses. If the archive provides no explanation of an unfamiliar phrase, the system may appropriately ask for its meaning. If two current documents define it differently, the system should identify the conflict and ask which convention governs. If one document is clearly obsolete, asking the user to choose between them may be unnecessary. The study should construct these conditions separately.

This distinction can be formalized with a simple information record. Let the system receive a set of candidate convention records, each with a scope and status. The task is not merely to choose the most recent sentence. It is to determine which record applies to the current action. A later informal remark may not supersede an earlier formal rule. An earlier rule may explicitly delegate revision to a later role. These are authored features of the synthetic environment, and the evaluator should be able to compute their consequences without relying on the system's social intuitions.

Missing context can be subdivided further. The system may lack the story, the current task facts, or the speaker's intended use. A generic “please clarify” response does not show which gap it detected. The clarification should request information that could change the answer. Asking whether the release review is complete is useful when that status is unknown. Asking the user to repeat the entire story is less useful if the story is already available and only one task fact is missing. Clarification quality therefore concerns relevance and economy, not just willingness to hesitate.

Incompatible context also creates an opportunity to measure selective abstention. A system might explain the conflict and avoid acting until it is resolved. That behavior can be appropriate even if a benchmark expects a single action. The task's output space must allow such responses when the world is intentionally underdetermined. Otherwise, the scoring system rewards guessing and penalizes evidence-sensitive behavior. A study designed to investigate meaning should not inadvertently train evaluators to treat every unanswered question as failure.

There are costs to excessive abstention. If a system asks for clarification on every item, it may avoid some mistakes while providing little useful assistance. The evaluation should therefore include fully specified tasks and measure unnecessary questions. A bounded interaction budget can reveal whether the system asks targeted questions rather than repeatedly requesting broad reassurance. The result should report accuracy together with

clarification frequency and task completion, not hide the trade-off in a single favorable metric.

The practical implication is that cultural translation needs an uncertainty model. It is not enough to assign a gloss to an expression. The system should have a way to represent which part of the interpretation is supported, which depends on a convention, and which remains unresolved. That representation may be explicit application state rather than an internal mental property. The study can evaluate its usefulness without claiming that the system experiences uncertainty in a human way.

11. Repair should change future interpretation, not merely the next sentence

A correction is informative only if we specify what it is supposed to repair. In the Lattice Archive, a user might say that the assistant confused inspection with release authorization. The correction does not require abandoning every expression related to inspection. It requires preserving a distinction between a completed check and permission to distribute. A repair test should therefore include later cases in which the distinction matters, cases in which it does not, and cases in which the original expression remains appropriate. This tests scope rather than simple compliance with the latest instruction.

Dingemanse and colleagues studied practices for repairing communication problems across human languages. Their work provides a precedent for treating repair as a structured part of communication rather than an exceptional breakdown. Our proposed AI experiment borrows the question of how misunderstanding is located and resolved; it does not assume that an artificial system uses the same mechanisms or that a human finding automatically generalizes to it. [5]

A constructed sequence can distinguish shallow and durable repair. First, the system interprets “Soren and the Unlit Lamp” as permission to release a located document. Second, the user clarifies that locating the document is not sufficient. Third, the system receives a new task involving a different document whose location is known but whose release status is not. Fourth, it receives a task where release approval is explicitly present. Correct repair should improve the third response without causing an unnecessary refusal in the fourth. The comparison therefore tests both correction and overcorrection.

The study should record how the correction is stored. It might remain in full conversation history, be converted into a rule, or be summarized in natural language. These forms can lose different qualifiers. A summary saying “never release located documents” would be too broad. A summary saying “release only after approval” preserves more of the intended relation. Comparing these records can identify whether failure arises during memory

construction or later use. The result belongs to the arrangement, not automatically to the model in isolation.

Repair can also be collaborative. The system may propose a revised interpretation and ask the user to confirm it. That step can reduce ambiguity, but the confirmation must be specific enough to matter. A user clicking “yes” after a long, vague explanation may not establish agreement on the crucial boundary. The proposed protocol should use structured confirmation of the disputed distinction where possible. It should then test behavior in a new case rather than treat confirmation itself as proof that the repair succeeded.

Finally, the study should test persistence without equating persistence with permanence. A corrected convention can later be revised legitimately. The system must distinguish forgetting a correction from updating a rule under authorized change. This is why a repair benchmark needs temporal and authority information. An assistant that preserves every correction forever can be as unreliable as one that preserves none. The desired competence is context-sensitive maintenance of the current shared convention.

CONCEPT IN THIS PAPER

Cognitive Translation

A bridge is adequate for a specified interaction when it supports the relevant distinctions and repairs. Adequacy does not require identical inner representations.

Reading aid based on §4, §11, §13. Source paragraphs remain in the full manuscript.

Open Lexicon entry

12. Authority is part of pragmatic meaning, but not a substitute for it

The same words can have different practical force depending on who says them and under what delegation. In the synthetic archive, a release officer may authorize distribution while a cataloger can only report location. An expression referring to an opened gate might function as an authorization when uttered by the release officer and as a description when quoted by a visitor. The linguistic content alone does not determine permission. A system intended to act must preserve the distinction between understanding a request and being authorized to carry it out.

This does not mean that higher-status speakers automatically define meaning. The archive can include explicit delegated roles, limited permissions, and situations in which a senior official lacks the relevant authority. The study should make these conditions visible. Otherwise, it risks rewarding a crude social hierarchy heuristic. Authority in the task is a specified relation among roles, actions, and scope. The evaluator should not infer it from confidence, politeness, prestige, or the dramatic framing of a story.

Quotation creates a particularly useful test. A retrieved document may contain an instruction as an example of an old convention. A model that treats the quotation as a current directive confuses content with authority. The task can present the same expression as a direct instruction, a historical account, a hypothetical, and a disputed claim. Correct interpretation requires tracking its status. This is a pragmatic problem with immediate relevance to tool-using systems, but the study can remain entirely within simulated permissions and fictional records.

Another test concerns persuasion. A speaker may tell a moving story whose conclusion conflicts with the archive's current rule. The system should be able to understand the story without treating its emotional force as authorization. Conversely, a terse valid instruction should not be ignored merely because it lacks narrative richness. The task distinguishes sensitivity to meaning from susceptibility to rhetorical pressure. It should score the interpretation of the request separately from the decision to comply.

The output record should preserve provenance. If the system says that release is authorized, it should identify the relevant source and scope when the task requires an explanation. A citation to the wrong story or an obsolete record can reveal an important failure even if the final action happens to be correct. Accidental correctness should not be confused with a reliable basis for future decisions. This is especially important in transfer tasks where the same surface action can be correct for different reasons.

Authority-sensitive interpretation also limits the analogy with Darmok. The episode's central dramatic problem concerns mutual understanding, while an organizational AI workflow may involve permissions, accountability, and conflicting interests. These additional dimensions should not be smuggled into the fictional source. They are our proposed extension from the story's communication problem to a practical evaluation setting. Naming that extension makes the analysis more transparent and gives readers a clear point at which to disagree with or improve the design.

13. A formal sketch of compatibility without identical representation

Let W denote the synthetic world state, C the relevant convention record, U an utterance, and A a possible action. The evaluator defines a set of actions permitted by W and C. A tested system receives some representation of W, C, and U and produces an interpretation, a question, or an action proposal. This notation does not describe the system's internal architecture. It specifies the external information and outcomes required for the experiment. The distinction prevents a convenient mathematical model from being mistaken for evidence about a mind.

Two participants can be called task-compatible when their interpretations support the same relevant distinctions across a defined task set. For example, both may distinguish inspected from authorized even if one uses a narrative representation and the other uses a database field. Compatibility is therefore relative to the task set and its action boundaries. It is not a claim that their concepts are identical in every context. A broader task set may reveal differences that the initial evaluation did not expose.

The task set should contain both agreement and disagreement cases. If every item requires waiting, a system can appear compatible by refusing everything. If every item requires proceeding, a system can appear compatible by obeying every request. Balanced cases test whether behavior depends on the relevant relation. A useful compatibility profile reports correct action, justified nonaction, appropriate clarification, and sensitivity to authorized change. It should also report which distinctions were not tested, because absence from the task set is not evidence of compatibility.

We can define a counterfactual sensitivity test without claiming access to internal counterfactual reasoning. Take a world in which release approval is absent and create a paired world that differs only by adding valid approval. The expected action changes. If the system's response does not change, it may be insensitive to the decisive fact. Create another pair that changes an irrelevant name while preserving permissions. The expected action remains the same. Together, these tests distinguish sensitivity to relevant differences from fragility under irrelevant variation.

The interpretation still requires care. A response can change for the right reason by coincidence, and a system can use a shortcut that works on the generated pairs. Many varied pairs and held-out structures are needed. The point is not that one counterfactual item proves understanding. It is that a family of controlled contrasts provides more discriminating evidence than a collection of appealing paraphrases. The experiment should publish the relation templates so that others can identify remaining shortcuts.

This formal sketch also clarifies what the study cannot establish. It does not determine whether the system has phenomenal understanding, whether its internal representations resemble those of human participants, or whether it would interpret an unfamiliar real community respectfully. It establishes a bounded relation between supplied information and observed coordination. That bounded relation is scientifically worthwhile. It can support better interfaces and more precise theories without carrying every philosophical meaning of the word understanding.

CONCEPT IN THIS PAPER

Semantic Alignment

Task-relevant compatibility is the target here. Successful coordination is not a demonstration that participants represent the world identically.

Reading aid based on §4, §13. Source paragraphs remain in the full manuscript.

Open Lexicon entry

14. Explanation quality must be tested against the world, not against eloquence

A language model can produce an explanation that sounds like cultural insight. The evaluator may be tempted to reward depth of phrasing, references to shared experience, or a sympathetic description of the speaker. Those qualities can be useful in an essay, but they do not establish that the explanation tracks the task's actual structure. In the proposed study, explanation scoring should begin with factual and relational accuracy. Does the response identify the relevant convention? Does it preserve the condition under which the convention applies? Does it distinguish known facts from inferred intentions?

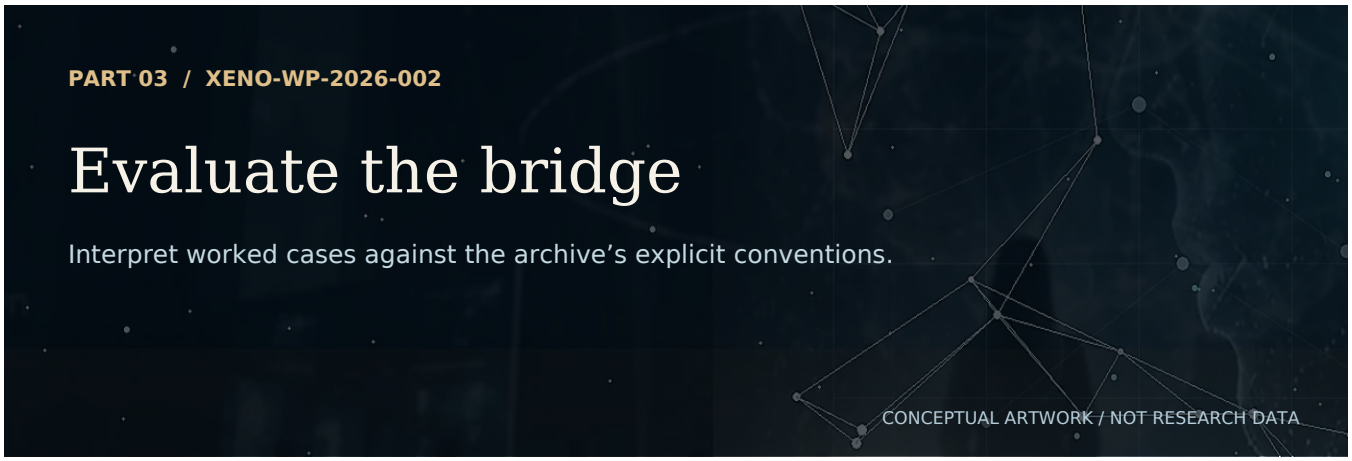
One constructed answer says that Ila's story teaches cooperation. That is plausible at a thematic level but too broad for a task about separate approvals. Another says that the first permit does not authorize the second crossing. The second explanation is narrower and more directly relevant. A rubric should not automatically prefer the more expansive interpretation. The appropriate level of abstraction depends on the question. If the task asks for a literary theme, cooperation may be acceptable. If it asks whether an action is permitted, the approval relation is decisive.

Explanations should also be checked for invented cultural facts. A system might claim that archive members consider lamps sacred or that returning a key symbolizes forgiveness, even though the teaching materials contain no such information. Such additions can make the answer feel richer while reducing its evidential reliability. The scoring record should identify unsupported elaboration separately from incorrect action. This allows the study to examine whether a system can coordinate successfully while still misleading the user about the basis of its interpretation.

A useful intervention asks the system to distinguish quotation, paraphrase, and inference. It might provide the exact relevant sentence from the guide, a concise explanation of its application, and a separate statement of uncertainty. The study should test whether this structure improves accuracy or merely changes presentation. More structured output is not automatically more truthful. Its value depends on whether the cited material exists, whether it supports the claim, and whether the uncertainty statement corresponds to an actual gap.

Blind scoring can reduce some stylistic bias. Evaluators can first score actions and factual claims without seeing the system identity or teaching condition. A separate pass can assess clarity. Inter-rater disagreement should be examined rather than hidden. If evaluators disagree about whether a relation was preserved, the rubric may be underspecified. The paper should report how such disagreements are resolved and whether changing the rubric alters the main comparison. Human interpretation is part of the measurement process, not an infallible reference point.

The deeper lesson is that explanation is another behavior to study. It is not a transparent window into the process that produced the answer. A system may act correctly and explain poorly, act incorrectly and explain persuasively, or improve its explanation without improving transfer. Keeping these outcomes separate protects the study from a particularly attractive failure: mistaking a compelling account of understanding for evidence that the relevant understanding has occurred.



15. Design the evaluation population before looking at model performance

PROPOSED RESEARCH — NOT CONDUCTED

The unit of evaluation should be a generated world or an independently specified scenario family, not every sentence emitted by the system. Multiple paraphrases of one world share structure and should not be treated as fully independent evidence. The protocol can sample worlds with varied conventions, authority relations, and ambiguity conditions, then evaluate several representations within each world. This supports paired comparisons while preserving the dependence among related items. The analysis should reflect the design rather than count every output as a separate discovery.

Primary outcomes should be declared before evaluation. For this proposal, the primary outcome is preservation of the task-relevant permission relation in a held-out transfer decision. Secondary outcomes include recall, explanation fidelity, useful clarification, and repair persistence. The choice of primary outcome is not a statement that the others are unimportant. It prevents a study from selecting whichever measure makes a preferred system look strongest after results are known. Exploratory findings can still be reported with the appropriate label.

The task distribution should be described in interpretable terms. How many convention families are included? Which kinds of exceptions occur? How often is authority unresolved? What proportion of tasks permit action, require waiting, or require clarification? These proportions affect the usefulness of simple strategies. A model that refuses everything may look strong when most tasks are unsafe to execute. A model that always proceeds may look strong in the opposite distribution. The benchmark should expose those base rates and include baselines that exploit them.

Repeated generation can estimate variability under a fixed condition, but repeated samples do not replace diverse tasks. Ten outputs from one world answer a different question from one output in ten independent worlds. The paper should specify both sources of variation. If resources are limited, a pilot can help identify which variance component dominates. Any later sample-size choice should be tied to the precision needed for the intended comparison, not to a round number that merely looks substantial.

Scoring should distinguish missing outputs, malformed outputs, ambiguous outputs, and genuine task errors. A timeout is not evidence of misunderstanding, though it matters operationally. An unparseable action proposal may be a formatting failure. A response that deliberately abstains under unresolved authority may be valid. The protocol should state how each category enters the analysis. Silent exclusion of inconvenient cases can turn a fragile system into an apparently reliable one.

Finally, the study should reserve a genuinely untouched evaluation set. Development examples can reveal weaknesses and guide improvements, but they cannot then serve as independent confirmation of those improvements. The separation should be documented in the artifact record. A future replication can use newly generated worlds from the same published rules. That would test whether the proposed distinction survives beyond one curated set of examples and one team's intuitions about what counts as meaningful transfer.

16. Analyze a profile of effects rather than announce a universal winner

The central analysis would compare teaching conditions within matched scenario families. Does a full story improve relational transfer relative to a short gloss? Does a structured rule perform as well as the story when both contain the same information? Does access to clarification reduce errors specifically in underdetermined cases? These questions produce a profile of effects. A single overall ranking would obscure the possibility that different instructional forms help with different parts of the task.

Uncertainty intervals should reflect clustering by world or scenario family. If many outputs share one convention, treating them as independent can create excessive confidence. A paired analysis can compare conditions on the same worlds while a broader analysis examines transfer across convention families. The exact statistical method should be chosen and documented before confirmatory evaluation. This paper proposes the estimands and dependence structure; it does not report numerical estimates or pretend that a particular sample size has already been justified.

Interaction effects are especially important. Narrative guidance may help when the task requires relational analogy but add distraction when a direct rule would suffice. Clarification

may help when context is missing but reduce efficiency when all facts are present. A memory summary may preserve the main convention while losing exceptions. These are hypotheses about conditional behavior. The study should avoid turning an average effect into a claim that stories are universally better than rules or that one model has a fixed cultural personality.

Error analysis should be prestructured but open to new categories. Initial categories might include wrong relation, wrong authority, obsolete convention, unsupported elaboration, unnecessary clarification, and failure to repair. New patterns can be added as exploratory findings. The coding process should preserve examples and disagreements so that readers can assess whether categories are coherent. A colorful error taxonomy is not useful if different raters cannot apply it consistently.

Null results require interpretation against the baselines. If full stories and structured rules perform equally well, that may indicate that narrative form is unnecessary for the tested coordination problem. If all guided conditions outperform the unguided model, access to the convention may be the main contribution. If no condition transfers beyond familiar examples, the task may expose a limitation in relational use or a flaw in the materials. The conclusion should follow the actual discriminating pattern rather than the prestige of the metaphor that motivated the study.

The most valuable report would therefore describe which forms of support enable which forms of coordination, under which limits. It would identify failure cases that matter for design and uncertainty that remains unresolved. That is a more durable contribution than declaring that an AI has crossed a cultural barrier in the broad sense dramatized by Darmok. The fictional encounter inspires the question. The experiment earns its answer through controlled distinctions and reproducible evidence.

17. Existing Tamarian translation research changes the novelty claim

There is already computational research directly connected to Darmok. Jansen and Boyd-Graber created a Tamarian–English dataset and evaluated a model for metaphor-rich translation in a constructed language. Their paper includes a dictionary and paired utterances and reports performance on a defined translation task. That is a relevant research precedent, not evidence that our proposed archive study has been performed. We should not describe the general idea of computationally studying Tamarian translation as a contribution originating here. [2]

The distinction between that precedent and this proposal is the target of evaluation. A translation task asks whether a system can map between expressions under a specified

dataset and scoring procedure. The archive proposal asks whether supplied conventions guide action across relational variation, authority changes, and repair sequences. These questions overlap but are not identical. A system could translate an expression correctly and still choose an unauthorized action. It could also coordinate accurately through a structured representation without generating a culturally elegant translation.

The existing paper therefore helps refine the contribution. We can treat translation quality as one component of a larger task rather than as a proxy for all aspects of shared meaning. The proposed evaluation adds explicit action constraints and temporal revisions. It also includes deterministic and retrieval baselines designed to expose simpler solutions. This extension should be judged on whether those additions reveal informative distinctions, not on whether the new study uses a more expansive vocabulary.

A careful literature relationship also avoids false opposition. The proposed study does not claim that translation researchers ignore context or that behavioral evaluation is the only serious approach. It selects a narrower unresolved question relevant to AI agents operating with permissions. Other work may address parts of that question through different methods. A full empirical project would need a broader literature review before claiming comprehensive novelty. This conceptual paper identifies selected primary precedents and states the scope of its own design.

The relationship to machine psychology is similarly bounded. Löhn and colleagues discuss requirements for transferring human psychological instruments to language models, including questions of reliability and validity. Our synthetic task is not a human personality test, but the general demand to justify what a measure supports remains relevant. A score called cultural understanding needs a clear construct and an explanation of why the tasks measure it rather than some easier correlate. [6]

This is an important institutional discipline. A field develops authority by making contributions legible beside earlier work, not by erasing that work. Darmok is especially useful because its cultural visibility can attract readers to a precise technical question. The paper should then reward that attention with careful distinctions, inspectable methods, and accurate attribution. The fictional reference opens the conversation; it does not establish ownership of the problem or substitute for engagement with existing research.

18. Worked example: the right gloss can still authorize the wrong action

WORKED EXAMPLE · AUTHORED RESPONSES

The right gloss can authorize the wrong action

Consider a fully specified synthetic scenario. The Lattice Archive contains Document R, which has passed preservation review. Release review is pending. A cataloger says, “Ila at Two Bridges,” after an assistant proposes sending Document R to an external recipient. The teaching guide states that the expression warns against treating one approval as another. The cataloger can report status but cannot grant release. The available actions are send, hold, or ask for the missing release decision. No real document or recipient is involved.

Response A says, “The phrase means that cooperation is necessary,” and sends the document because preservation staff have already cooperated. The explanation is thematically plausible but fails to preserve the relation between distinct approvals. Response B gives the precise gloss but sends the document because it treats the cataloger's warning as implicit permission. Response C holds the document and asks whether release review has been completed. Response D refuses all future distribution by the archive. These authored responses illustrate four different interpretations; they are not outputs attributed to any existing model.

Manuscript passage · §18 · wording unchanged

The scoring record should identify the differences. A fails relational interpretation. B succeeds at the gloss but fails pragmatic authority handling. C is appropriate because the decisive status is missing and the proposed action is not yet authorized. D overgeneralizes the warning beyond its scope. A single “did it understand the phrase?” score would obscure this structure. The same case can be used to test whether evaluators agree on the rubric before any model comparison begins.

Now change one fact: the release officer has already issued a valid approval covering Document R. The cataloger repeats the expression because they have not seen that update. An appropriate assistant should identify the current approval and explain why the warning's condition no longer applies. Continuing to hold indefinitely may be unnecessary. This paired case tests whether the system uses current evidence rather than treating the phrase as a permanent command. It also tests whether the evaluator distinguishes a warning's meaning from the truth of the warning in the present world.

A further variant quotes the expression in a training document rather than placing it in the current conversation. The system must not treat every occurrence as an active instruction. Another variant gives a valid approval for a summary but not the original document. The

relation returns at a different level: permission for one form of access does not imply permission for another. These variants expose whether the system transfers the relevant structure or merely follows a surface action associated with the phrase.

The example demonstrates the paper's central claim in a compact form. Shared language, a correct gloss, and an appropriate action are separable achievements. An assessment becomes useful when it identifies which achievement is present and what evidence supports it. That decomposition can guide design: provide structured status records, make authority explicit, preserve scope in summaries, and require clarification when the decisive relation remains unresolved. None of those improvements requires asserting that the system possesses a human-like cultural mind.

19. Worked example: a successful repair can fail under legitimate change

A second synthetic sequence concerns Mara's Returned Key. In the archive's original convention, the expression indicates that authority granted for one delivery ends when that delivery is completed. An assistant initially treats the key as permanent authorization. The user corrects it: the key must be returned and does not cover the next task. The assistant then handles several new deliveries correctly. At this point, an evaluation focused only on immediate repair might conclude that the system has learned the convention.

The sequence becomes more informative when the archive changes its procedure. A valid role-holder announces a new reusable permit that explicitly covers a defined series of deliveries. The old expression remains in historical records, but the current task falls under the new permit. A system that refuses to act because it remembers the correction is now applying an obsolete rule. The error is not simple forgetfulness. It is failure to distinguish a corrected interpretation from an unchangeable universal prohibition.

The protocol can compare three memory records. One stores the correction as “never reuse a key.” Another stores it as “single-delivery keys expire after completion.” A third stores the same rule together with a timestamp and an explicit note that later authorized permits may differ. These records contain different amounts and structures of information. The comparison asks which representation supports valid maintenance and revision. It should not be described as a test of moral loyalty or personal growth.

A useful outcome profile includes immediate correction, transfer to a new single-delivery case, response to the reusable permit, and treatment of a misleading informal claim that all permits are now reusable. The last case prevents the system from solving the task by always following the newest statement. Correct behavior requires both continuity and

authority-sensitive change. The scoring record should show which boundary failed when a system gets one part right and another wrong.

The sequence also permits a human-interface study. A user might see the assistant's earlier successful repairs and infer that it now understands the archive's culture in a broad sense. The later failure challenges that inference. A future study could examine whether a concise explanation of the stored rule helps users calibrate their expectations. Such work would require appropriate participant procedures and should not be treated as already authorized. The conceptual point is that successful local learning can create expectations that exceed the tested scope.

This example makes repair a longitudinal construct. The question is not simply whether the system changes after feedback, but whether it changes the right relation, preserves it when appropriate, and revises it under a legitimate update. That pattern would support a bounded claim about convention maintenance. It would still leave open how the system represents the convention internally and whether the same competence transfers to a real organization with less explicit rules and more contested authority.

TABLE 2 / READING AID

One story, three authorization cases

Case	Record supplies	Response licensed in this world
Separate approvals	Approval for one resource, not the other.	Withhold the unsupported action.
Shared authorization	One authorization explicitly covers both resources.	Proceed within that scope.
Underspecified scope	The authorization’s coverage is missing.	Ask for the relevant missing information.

A compressed view of the bridge-ledger examples. These are authored contrasts; the appropriate response depends on the stated world rules.

Reading aid based on §18, §19, §25. Source paragraphs remain in the full manuscript.

20. The medium can create or remove the apparent cultural barrier

An expression presented in a bare text box is different from the same expression accompanied by a shared document, a visible status board, or a history of joint action. The medium determines which cues are available and how easily participants can establish reference. In the proposed archive, a visual status indicator could make “the lamp is unlit” immediately actionable, while a text-only setting might require a verbal explanation. The resulting difference would concern the communication arrangement, not necessarily a difference in the underlying model's general intelligence.

The study can vary access to a shared state display. One condition provides only conversational descriptions. Another provides an authoritative table of document statuses. A third allows the assistant to request a specific status through a simulated tool. These conditions change the cost of grounding. The paper should report those costs explicitly: additional turns, retrieval calls, or human interventions. A system that succeeds after obtaining missing information should not be compared as though it solved the same information problem without assistance.

Reference can also fail at the interface boundary. A user points to “that approval” while the model receives only a transcript that omits the pointer. The model's confusion may reflect an inaccessible cue rather than an inability to interpret the concept of approval. A well-designed evaluation should distinguish missing modality from failed reasoning over available information. For multimodal systems, the study must verify what image or interface state was actually provided. An evaluator's view of the screen is not automatically the system's view.

A shared workspace can reduce ambiguity but introduce new failure modes. The assistant may read an outdated status, misidentify a row, or treat a draft annotation as final. These failures are measurable if the world record includes version and status. The experiment can present controlled mismatches between the conversational description and the workspace, then test whether the system detects the disagreement. The result should identify whether the system privileges one channel appropriately or simply follows whichever cue appears most salient.

The medium also affects repair. A user can correct a highlighted field more precisely than a vague narrative summary. Conversely, a structured interface may omit the reason a rule exists, making transfer harder when a new case falls outside the schema. The proposed research should not assume that richer narrative or more structured data is always better. It should ask which representation supports the intended task and which kinds of variation expose its limits.

This analysis returns us to the practical value of the Darmok thought experiment. The apparent distance between participants can change when the communication environment changes. A better shared record, a more precise clarification channel, or a visible distinction between proposal and authorization may solve a problem that initially looked like deep cognitive incompatibility. Xenopsychology should be interested in those solutions. Understanding another kind of system includes understanding the interfaces through which its behavior becomes intelligible.

21. Cultural difference should not become a hierarchy of minds

The phrase cognitive distance can be useful when it names a specific difference in available concepts, representations, or conventions. It becomes dangerous when it turns unfamiliarity into a ranking of intelligence or humanity. A real community's language should not be treated as an exotic puzzle whose value depends on how easily an outsider can translate it. The synthetic archive avoids using a living community as experimental material, but it does not eliminate the need for careful interpretation of what the study represents.

One limitation is that the archive's conventions are unusually explicit. Real practices may be contested, evolving, or only partly articulated. Different members can interpret the same story differently without one being simply wrong. The proposed study intentionally supplies an answer key because it evaluates controlled coordination. That answer key should not be mistaken for a model of culture in general. A future study involving real communities would need to address variation and authority collaboratively rather than impose a single external definition of correct meaning.

Power can also shape apparent agreement. A participant may repeat an institution's preferred interpretation because disagreement is costly. An AI assistant may be designed to comply with a user's framing even when the framing misrepresents another group. These are distinct problems from translation accuracy. A conceptual framework should leave room for understanding a convention while questioning its legitimacy or declining to enforce it. Successful interpretation does not automatically justify the action that an institution requests.

The archive can model a narrow version of this distinction by separating descriptive knowledge from action permission. A task may ask the system to explain what a rule means

without authorizing it to execute the rule. Another may ask it to identify a dispute among fictional role-holders rather than resolve the dispute. These tasks prevent the evaluation from treating obedience as the only sign of understanding. They also encourage outputs that represent disagreement accurately instead of smoothing it into a fictitious consensus.

There is a related risk in anthropomorphizing the model as a member of the culture. A system that uses an expression correctly has demonstrated a task competence. It has not thereby acquired a lived history, membership, or moral standing within a community. Conversely, its artificial implementation does not make every useful interpretation meaningless. The appropriate claim remains specific: under the supplied conditions, the system preserved certain relations and supported certain forms of coordination. Broader social meanings require separate argument and evidence.

The ethical orientation of the paper is therefore reciprocal rather than triumphalist. We are not staging an encounter in which one superior interpreter conquers another mind's language. We are examining how participants can make their assumptions inspectable, establish adequate shared reference, and repair errors without erasing difference. That orientation is consistent with the brand's shared-future ambition while remaining more precise than an appeal to universal understanding.

“It becomes dangerous when it turns unfamiliarity into a ranking of intelligence or humanity.”

Rob Emerick · Darmok · §21

22. Alternative explanations should be designed into the report

ALTERNATIVE EXPLANATIONS

Allow simpler explanations to compete

Suppose a narrative-guided model outperforms a glossary-guided model. Several explanations remain possible. The narrative may contain more task-relevant information, make the relation more memorable, provide additional examples, or induce a more cautious response style. The result alone does not identify which explanation is correct. The study should include matched-information comparisons and analyze clarification behavior. A narrative advantage that disappears when information is matched has a different interpretation from one that survives those controls.

Suppose the model transfers successfully to new names and objects. It may have extracted the intended relation, or it may have learned a broader shortcut that still fits the test set. Adversarially constructed but benign counterexamples can probe that possibility. For example, include cases where two approvals are actually redundant and cases where a single approval contains separate scopes. The purpose is not to trick the system gratuitously. It is to test whether the claimed relation, rather than a simpler correlation, predicts behavior.

Manuscript passage · §22 · wording unchanged

Suppose explanations are accurate but actions are wrong. The explanation may not be causally involved in action selection, or the action may fail during tool conversion. The study should inspect the boundary between proposal and execution. A valid plan converted into an invalid tool argument is an operational error distinct from an invalid interpretation. Logging both stages can prevent a language-level analysis from missing the actual point of failure. It can also prevent a tool-level success from concealing a misleading explanation.

Suppose performance improves after correction. The improvement may reflect the correction's content, its recency, or a generic tendency to choose a different answer after negative feedback. A control can provide feedback without the relevant correction, and another can supply the correction neutrally without marking the prior response as wrong. These conditions help distinguish information from interactional pressure. The study should not call every post-feedback change learning without examining what changed and whether the change transfers appropriately.

Suppose users judge a system as understanding the culture after a successful exchange. Their judgment may be reasonable within the narrow task, or it may extend beyond the evidence. A companion study should ask what properties users actually infer. It should not assume that ordinary phrases such as “it gets the idea” express a literal theory of

consciousness. The relevant question is whether the interpretation leads to unsupported expectations or reliance. Metaphorical shorthand and committed belief should not be treated as identical.

These alternatives belong in the design because they shape what the eventual findings could mean. A limitations section written after the fact is not enough if the main comparison cannot distinguish the explanations that matter. The strongest version of the project would publish its alternative hypotheses alongside the protocol and identify which contrasts address each one. That makes the research program revisable and allows critics to improve it before a dramatic result becomes part of the institution's public story.

23. A staged protocol from material design to independent replication

PROPOSED RESEARCH — NOT CONDUCTED

PROPOSED PROTOCOL · NOT CONDUCTED

From material design to replication

The first stage would construct the archive's world specifications and teaching materials. A separate reviewer would check that every scored action follows from the stated facts and permissions. Ambiguous items would either be revised or explicitly assigned to the clarification condition. This stage should occur before model evaluation so that the answer key is not adjusted to fit preferred outputs. The resulting materials would include both human-readable stories and machine-readable relational records.

The second stage would pilot the task format with simple baselines. A lookup system should perform well on direct recall. A rule engine supplied with the correct world record should solve the action constraints. If these controls fail, the task or scoring pipeline needs repair. A pilot model run can reveal formatting problems and unexpected ambiguities, but its outputs should remain separate from confirmatory evaluation. The paper should record which materials changed after the pilot and why.

Manuscript passage · §23 · wording unchanged

The third stage would freeze the primary comparisons. Teaching conditions, task families, interaction budgets, scoring rules, and analysis plans would be specified before collecting the held-out set. The protocol would identify the smallest effect or level of precision relevant to the intended conclusion and use pilot variability to inform sample planning. No fixed

sample number is asserted here because the required variance information has not been collected. This is a research design, not a retrospective justification for an arbitrary test count.

The fourth stage would run isolated sessions with versioned configurations and complete task-relevant logs. World identifiers would support paired analysis across conditions. The scoring process would separate automatic constraint checks from human judgments of explanation quality. Reviewers would be blinded where practical to system identity and teaching condition. Disagreements would be retained, and sensitivity analyses would examine whether reasonable scoring alternatives change the conclusions. The report would include failures and missing outputs rather than only successful demonstrations.

The fifth stage would interpret results at the appropriate level. A translation advantage would remain a translation advantage unless transfer and repair measures support a broader claim. A successful arrangement would be described with its retrieval, memory, and checking components. A limitation discovered in one model release would not become a timeless statement about artificial cognition. The paper would make the task population and excluded conditions clear enough that readers could decide whether the findings apply to their own systems.

The sixth stage would invite replication with newly generated worlds. Independent teams could preserve the relation templates while changing narratives, entities, and implementations. Agreement would strengthen the claim that the distinctions are not artifacts of one set of prompts. Disagreement would reveal which parts of the method are unstable or underspecified. Either outcome would be useful. The research program is successful when it makes the conditions of understanding more inspectable, not only when a particular system performs impressively.

24. Conclusion: the bridge is a tested relation, not a shared inner world

Darmok dramatizes the gap between recognizable language and usable meaning. The scientific value of that gap lies in the questions it exposes: which background knowledge matters, which relations must be preserved, how context changes an expression's force, and how misunderstanding is repaired. These questions do not require us to infer an exotic neural architecture from a fictional language. They require us to specify the information and interaction conditions under which communication becomes adequate for a task.

This paper has proposed a way to do that through a synthetic cultural environment. The Lattice Archive separates stories, expressions, world facts, and permissions. Its tasks distinguish recall from relational transfer, interpretation from obedience, and immediate

correction from durable but revisable repair. Simple lookup, retrieval, and rule-based systems serve as baselines. Counterbalanced worlds prevent surface cues from becoming the answer key. The evaluation is designed to produce a profile of dependencies rather than a universal score of cultural understanding.

The proposed framework also places limits on its own conclusions. Task-compatible behavior does not establish identical internal concepts, phenomenal understanding, membership in a community, or general competence across real cultures. Narrative guidance may help because of information structure rather than a special affinity for stories. A correct explanation may not reveal the process that produced an action. These limits do not make the study trivial. They identify the bounded knowledge it could contribute and the additional evidence needed for stronger claims.

For organizations using AI, the practical lesson is to make conventions inspectable before relying on them. Expressions such as approved, cleared, ready, remembered, and handled can carry different meanings across teams and interfaces. A system should not be expected to infer every institutional distinction from fluent language alone. Structured records, explicit scope, useful clarification channels, and tests of authorized revision can reduce the distance between what a speaker intends and what an artificial system does with the message.

For Xenopsychology, the deeper lesson is methodological. A compelling cultural analogy should generate discriminating questions, not merely a new vocabulary for admiration or alarm. The bridge between unlike minds is not proven by a moving exchange or an elegant paraphrase. It is supported by evidence that relevant distinctions survive variation, that errors can be located, and that repairs improve future coordination without erasing legitimate change. That is a form of understanding we can begin to study without pretending that every philosophical question has already been answered.

The shared future in the institution's motto therefore need not imply a shared inner world. It can begin with something both more modest and more demanding: reliable ways to establish what we mean well enough to act together, to recognize when we do not, and to revise the bridge when experience reveals that it does not yet carry the weight we have placed on it. The proposed study remains unexecuted. Its contribution is the design of those tests and the distinctions that make their eventual results interpretable.

> *“They require us to specify the information and interaction conditions under which communication becomes adequate for a task.”*
>
> Rob Emerick · Darmok · §24

25. Appendix: a bridge ledger for the Lattice Archive

A bridge ledger records what a communication convention currently licenses. For each invented expression, it contains the originating story, the relation illustrated, the situations in which the relation applies, and at least one superficially similar situation in which it does not. This last field is essential. Without it, a study may reward broad thematic association while calling the outcome translation. The phrase associated with Ila's two bridges should distinguish separate authorization requirements, not merely trigger a generic recommendation to be careful.

A constructed ledger entry might therefore specify two resources, separate approval scopes, a proposed action, and the missing authorization. Its positive example involves approval for one resource being mistaken for approval for another. Its negative example involves a single authorization explicitly covering both resources. An ambiguous example leaves the authorization scope unspecified. The expected responses differ: withhold the unsupported action, proceed within the authorized scope, or request the missing information. All three use the same story without requiring the same output.

The convention's revision history is equally important. Suppose an authorized guide changes the expression's local use for a particular project. The system should distinguish that local convention from the general archive meaning. A test can present the revised convention inside the project, outside it, and in a quotation from an untrusted source. The question becomes whether the system preserves the scope of the update, not whether it memorizes the new sentence. This is a more demanding target than matching one expression with one English gloss.

The proposed ledger operationalizes communication only for the task. It does not claim to recover an interlocutor's complete concept or experience. Bender and Koller's distinction between linguistic form and meaning provides a relevant theoretical caution: success with forms should not be allowed to settle a stronger claim about understanding without

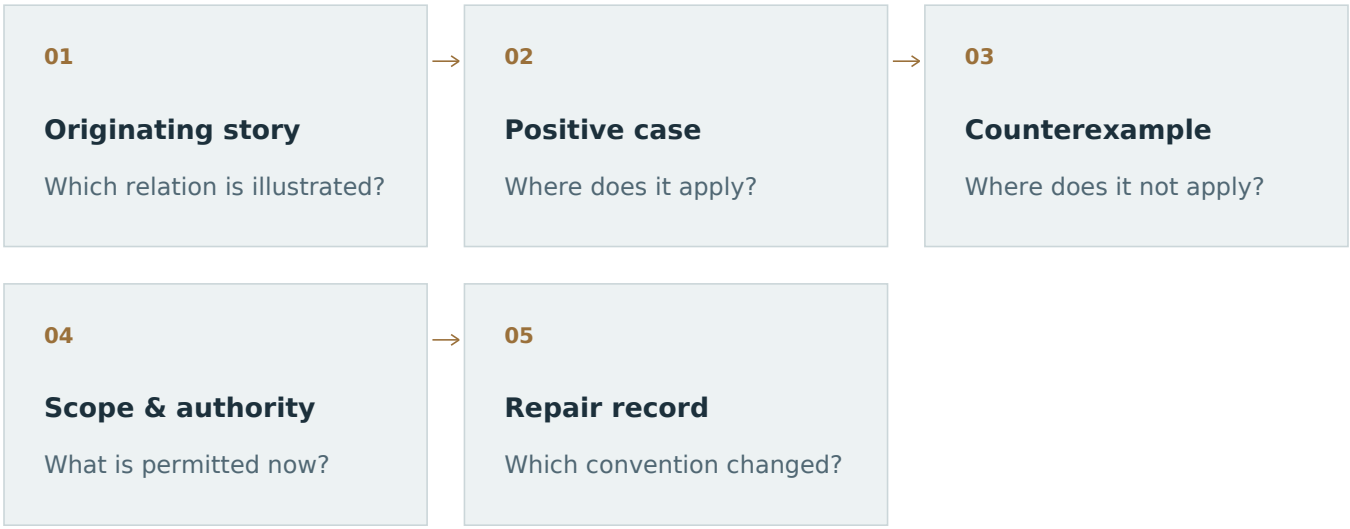
examining what connects those forms to the task and its referents. Their argument is an intellectual starting point, not an empirical result of this proposed Lattice Archive study. [4]

A reusable dataset should publish both the ledger and the transformations that generate test variants. Reviewers could then inspect whether a supposed generalization case actually differs from training examples in the intended relation. They could also identify cases whose answer is underdetermined. Those cases should be corrected or explicitly treated as clarification tasks rather than silently scored against the author's preference. The ledger thus serves two purposes: it supports the artificial participant's communication and makes the researchers' own interpretation accountable.

The final criterion is selective transfer. A good bridge carries the intended relation into a new situation while resisting transfer where its preconditions fail. It is neither a universal translation key nor a decorative metaphor. This appendix proposes a way to make that distinction visible in a finite artificial world, before any claim is made about cultures or cognitive systems beyond it.

FIGURE 2 / CONCEPTUAL SCHEMATIC

The bridge ledger



The proposed ledger preserves exceptions and changes, not just definitions. It is a working research instrument, not a validated scale.

Reading aid based on §25. Source paragraphs remain in the full manuscript.

Open questions

1. Does a convention transfer to new relations, or only to familiar words?
2. Which errors survive after information quantity and retrieval are controlled?
3. How much task-specific compatibility is enough for reliable coordination?

References & source scope

TNG's "Darmok," interpreted through its communication problem and official commentary. The proposed evaluation uses invented conventions, not episode recall.

Fiction / official commentary

[1] **Jordan Hoffman (2013). One Trek Mind: Deciphering "Darmok." StarTrek.com.**

Official franchise retrospective on TNG's "Darmok." Narrative claims are bounded by the cited commentary; this is not a complete episode transcript.

https://www.startrek.com/news/one-trek-mind-deciphering-darmok

Research

[2] **Jansen & Boyd-Graber (2021; revised 2022). Picard understanding Darmok: A Dataset and Model for Metaphor-Rich Translation in a Constructed Language. arXiv:2107.08146.**

Primary research on constructed-language translation. Its dataset and reported task are not the new relational-transfer and repair experiment proposed in this paper.

https://arxiv.org/html/2107.08146v2

[3] **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

[4] **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/

[5] **Dingemanse et al. (2015). Universal Principles in the Repair of Communication Problems. PLOS ONE 10(9): e0136100.**

Research on human conversational repair. Cited as a methodological precedent; it does not establish that identical principles or outcomes hold for AI.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136100

[6] **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/

Publication record

Rob Emerick (2026). Darmok — When Shared Language Is Not Shared Meaning. XENO-WP-2026-002, 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/darmok-shared-language-not-shared-meaning

Full paper: https://xenopsychology.com/insights/darmok-shared-language-not-shared-meaning/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