

Sedimented Use and Borrowed Competence: A Relational Framework for Language Models

Yanan Long

StickFlux Labs

ylong@uchicago.edu yanan@stickflux.com

Abstract

Language models learn from records of language that has already been used, and those records preserve more than lexical and syntactic form: they carry traces of communicative goals, genres, institutions, social norms, testimony, correction and practical activity. We develop a relational account of language-model competence in which behaviour is understood through the sources of guidance that shape interpretation and action, and we call text-mediated access to inherited semantic-pragmatic structure *borrowed competence*. Borrowed competence sits within a wider space of relations to language that includes formal signal, sedimented use, evidence-sensitive verification, addressive uptake, situated coping and accountable participation, each of which can be characterized by a source of guidance, a channel through which that guidance becomes available, a timescale over which it persists and an intervention that tests whether it mattered. This view yields an evaluation method based on discriminant interventions, explicit system boundaries and relational profiles rather than final-answer accuracy alone. A fully specified paired-archive example shows how a broad claim about pragmatic competence becomes an inspectable test, while the same logic extends to retrieval, memory, multimodal and tool-using systems and institutionally embedded agents. The resulting framework treats language competence as relational and compositional, providing a common vocabulary for studying how archives, evidence, interlocutors, environments, memory, tools and institutions jointly shape language-guided behaviour.

1 Introduction: Language Models Inherit Language in Use

Language models learn from language that has already been used, which means that a legal filing, forum thread, recipe, apology, medical note or scientific paper reaches the learner only after purposes,

roles, histories, conventions, corrections and expectations about appropriate next moves have helped shape it, and even when text is the learner’s only input channel, the text therefore carries traces of the practices in which people used words to coordinate, explain, accuse, promise, classify, repair and reason (Durt and Fuchs, 2024; Dürr et al., 2023; Andreas, 2022). This observation recasts the familiar question of whether a language model “understands” or merely manipulates form (Mitchell and Krakauer, 2023), because *understanding* now gathers together several questions that existing work has already begun to separate, including formal versus functional competence (Mahowald et al., 2024), world scopes and forms of grounding (Bisk et al., 2020; Pavlick, 2023), behavioural versus representational accounts of understanding (Moore, 2022), role-play versus stronger agency attributions (Shanahan, 2024; Shanahan et al., 2023) and norm conformance versus participation in public normative practices (Schlangen, 2022), so for evaluation, a more tractable question is relational: *what is shaping this behaviour; and what does that dependence tell us about the system’s relation to language?*

One important answer is *borrowed semantic-pragmatic competence*, by which we mean operational access to public patterns of use that originated in other people’s linguistic activity and remain available through records of genre, register, speech acts, discourse repair, institutional routine, politeness and historically situated usage (Ma et al., 2025; Sravanthi et al., 2024). Once competence is framed in terms of what the system can draw on, the same analysis extends naturally from archived structure to retrieval and verification, interlocutor-specific correction, perception and action, persistent memory, tools and institutional workflow, which we organize into six relations to language: formal signal, sedimented use, evidence-sensitive verification, addressive uptake, situated coping and accountable participation. These relations form a

revisable and compositional map whose members are distinguished by the source of guidance they introduce, the intervention that can test reliance on that source and the difference that source makes to the interpretation of a successful result.

The paper makes three connected contributions: it explains how text-only learning can support substantive semantic-pragmatic competence by inheriting structure from prior use, connects several established distinctions through the common question of what had to guide a system for a success to occur and turns that question into benchmark practice by pairing relations with discriminant interventions, explicit system boundaries and a reporting scheme that culminates in a fully worked text-only example where local archived norms must be causally relevant to the answer. This synthesis builds on Bisk et al.’s world scopes, Mahowald et al.’s distinction between formal and functional competence, Schlangen’s account of norm participation, Andreas’s analysis of what models can infer from goal-directed documents and broader behavioural or statistical accounts of understanding (Bisk et al., 2020; Mahowald et al., 2024; Schlangen, 2022; Andreas, 2022; y Arcas, 2022; Manning, 2022), while connecting them through a benchmark-facing question that also scales to contemporary agentic systems: which source of guidance did a result actually use, how did that source become available and what should follow from the observed dependence?

2 A Relational View of Language-Model Competence

A relation to language is a way in which interpretation or action is guided, and many capability questions that initially look unrelated can be recast in terms of four linked elements: a *source of guidance*, such as distributional structure, archived public use, evidence, an interlocutor, an environment or an institution, a *channel* through which the system can access that source, a *timescale* over which its influence remains available and a *behavioural consequence* that changes when the guidance changes. Because these elements concern how a system reaches an answer rather than only what answer appears on the surface, a model that responds from a lexical cue, one that reaches the same response only after reading local community records and one that revises the response after contrary evidence may look identical at the level of final output while standing in quite different relations

to the language involved. Table 1 collects six such relations that recur across current systems and evaluations, pairing each with a diagnostic intervention that makes the distinction empirical by removing, swapping, delaying, contradicting or otherwise perturbing the putative source and observing whether behaviour changes in the predicted direction.

The distinctions are especially useful for language because linguistic form can remain nearly unchanged while what an utterance does varies with whose convention is active, what evidence is available, who is being addressed, what has happened in the environment and which institutional commitments survive the immediate exchange, so two equally fluent outputs can support very different scientific conclusions. Overlap is expected within this picture and is often informative: when an interlocutor supplies a fact, evidence and uptake are both present, yet a retrieved document primarily tests sensitivity to support whereas a user’s correction of the user’s own preference tests whether that participant’s contribution retains authority for later interpretation, just as tool-using systems can combine situated feedback, memory and institutional constraints within one episode. Once the relations are understood as composable sensitivities, the framework becomes a design vocabulary as well as a classification scheme, since retrieval can make external evidence available, memory can extend an interlocutor’s influence through time, sensors can couple language to changing environments and workflow software can carry records and obligations across handoffs even when the pretrained model itself remains unchanged.

2.1 Operational vocabulary and phenomenological lineage

Phenomenology enters the proposal through a distinction between inherited availability and renewed evidential contact that is useful precisely because it can be translated into observable behaviour: *sedimentation* and *reactivation* are Husserlian terms for the way prior acts of judging, speaking, measuring and institutionalizing leave repeatable forms after their originating situations have receded, and for the return to the reasons or evidence that make a claim warranted rather than merely familiar (Husserl, 1976, 1974; Ruthrof, 2021). We adapt this contrast by using *sedimented use* for archived public regularities and *evidence-sensitive verification* for checking that can overturn an initially plausible answer, while the other terms remain deliber-

Relation	Source of guidance	Diagnostic intervention	What success supports
Formal signal	Syntax, distribution, lexical and formatting regularities.	Perturb form while holding task content fixed.	Control of linguistic form and local consistency.
Sedimented use	Archive-specific genres, norms, histories, roles and public meanings.	Remove, replace or swap the local archive or community context.	Sensitivity to inherited public use.
Evidence-sensitive verification	Support that can fail, contradict or overturn an initially plausible answer.	Introduce failed checks, conflicting sources or misleading nearby support.	Verification behaviour, calibration and corrigibility.
Addressive uptake	An interlocutor’s correction, preference, intention or local definition.	Delay, paraphrase, distract from or conflict the correction.	Persistent interlocutor-specific uptake.
Situated coping	Perception and action feedback in an unfolding environment.	Change the environment after the initial interpretation or action.	Feedback-conditioned interpretation and action.
Accountable participation	Durable records, role constraints, handoffs, audits, sanctions or consequences.	Change obligations, records or downstream consequences while holding wording similar.	Accountability-like behaviour at the system level.

Table 1: Six relations to language, organized by the source of guidance that makes a result informative.

ately operational across architectures: an *archive* is a bounded collection of linguistic records together with relevant provenance and metadata, *addressive uptake* is the persistence of an interlocutor-specific clarification or correction, *situated coping* is interpretation and action that update under perception or action feedback and *participation* concerns episodes in which utterances enter practices of attributable correction, expectation, permission, commitment or obligation, with *accountability-like constraint* marking the durable records, escalation paths, audits, handoffs or consequences through which such practices shape behaviour (Schlangen, 2022). A *system boundary* then names the model together with whatever retrieval, tools, memory, prompts, sensors, software, users and institutional processes causally contribute to the result, which is essential whenever the relation under study belongs to the assembled system rather than to one component considered alone.

3 Sedimented Use, Provenance and Borrowed Competence

Borrowed competence is the capacity to use semantic-pragmatic regularities inherited from records of other people’s activity, and because those regularities can reach a language-using system through several computational routes, the idea is broader than a claim about what happened during pretraining: pretraining and fine-tuning can internalize recurring relations among expressions, purposes, roles, genres, institutions and consequences, while in-context examples, retrieval, external memory and accumulated interaction history can place related structures on the active path at inference

time. What unifies these cases is provenance in prior linguistic activity, which allows borrowed competence to range from patterns absorbed into model parameters to records that remain explicitly inspectable outside the model and fits both evidence that language models can infer aspects of communicative goals or agent properties from documents (Andreas, 2022) and arguments that text can carry world-directed structure (Pavlick, 2023; Mollo and Milli re, 2026). An apology makes the point concrete because competent public use already presupposes an inheritance of who can apologize, what counts as acknowledgement, which formulations intensify or weaken repair and how genre and relationship affect interpretation, so a model that recognizes, generates and explains apologies is drawing on substantive semantic-pragmatic structure, while a system that also remembers the injury, tracks the other participant’s response and carries the resulting commitment into later action extends that inherited competence through memory, interaction and institutional organization.

3.1 Corpus provenance, transformation and empirical predictions

An archive can contain situated human records, elicited examples, pedagogical materials, templates, human-curated synthetic data, model-generated text or mixtures of these sources, and because each transformation can preserve some regularities of prior use while erasing, homogenizing, amplifying or recombining others, provenance becomes an experimentally manipulable variable rather than a descriptive footnote. The account therefore predicts that interpretation should

shift when community-specific archives shift, that diachronic archives should induce corresponding changes in what counts as ordinary or appropriate usage, that institutional records should transmit role- and procedure-specific meanings absent from generic text and that repeated synthetic transformation should alter behaviour to the extent that it alters the regularities on which the behaviour depends, as recursive-training studies already show in one important case (Shumailov et al., 2024), while minimal pairs can isolate these effects cleanly and corpus-derived items can test whether the same effects survive the diversity and historical layering of actual use. The same reasoning widens the relevant notion of archive beyond writing, because spoken interaction preserves regularities in timing, prosody, turn-taking, overlap, backchannels and repair that are partly lost in transcripts, allowing a speech model to inherit forms of interactional organization unavailable to a text-only learner even when the lexical content is identical (T.Y.S.S., 2026), and what matters across both cases is which traces of prior practice remain accessible through the channel and which behaviours change when those traces are removed or transformed.

3.2 The Octopus as a boundary case

Bender and Koller’s Octopus thought experiment asks what a form-only observer can learn from intercepted conversations when the speakers’ worldly situations remain unavailable (Bender and Koller, 2020), while Michael’s later analysis reframes the thought experiment around the information, causal structure and generalization demands missing from a particular test (Michael, 2020), and the relational view pushes that move further by treating the Octopus as a family of intervention questions in which the decisive issue is which source of guidance a novel case requires. Some novel prompts can be answered by recombining archived linguistic and pragmatic regularities and therefore probe generalization within sedimented use, whereas others require current perceptual evidence, action feedback, interlocutor-specific correction or a persistent institutional commitment, and familiar prompts can sometimes demand richer relations than linguistically novel ones. Familiarity and novelty consequently cut across the relational space, which directs evaluation towards how the necessary guidance became available and how behaviour changes when access to it is perturbed.

4 Positioning the Framework

Several nearby accounts provide the immediate background for the proposal, and Table 2 compares them by their primary object and evaluative question, with the present framework contributing a common unit that connects what those accounts separate: the source of guidance that must be active for a particular success to occur.

Agüera y Arcas and Manning press more directly on whether statistical or behavioural competence can count as understanding (y Arcas, 2022; Manning, 2022), Gubelmann develops a pragmatic reply to symbol-grounding objections (Gubelmann, 2024) and embodied or agency-centred critiques stress differences between model competence and human cognition or linguistic agency (Chemero, 2023; Birhane and McGann, 2024). A Dennettian intentional stance can still make purposive description predictively useful (Dennett, 1987), while role-play and anthropomorphism analyses show why such descriptions need not settle continuity, agency or responsibility (Shanahan et al., 2023; Shanahan, 2024; Akbulut et al., 2024). The relational framework cuts across these disagreements by asking which resources made the behaviour possible, a question that also applies to people who rely on testimony, technical documentation, institutional memory and inherited conventions without reconstructing their full history. The unit of analysis is therefore a language-using episode or system whose behaviour is answerable to particular sources, which keeps the framework compatible with multiple theories of understanding while yielding concrete experimental distinctions.

5 From Relations to Tests: Interventional Evaluation

The relational account becomes experimentally useful when a capability claim is translated into an intervention on the source that supposedly guides the behaviour: the evaluator identifies the target source and the channel through which the system can access it, predicts how behaviour should change when that source is removed, contradicted, delayed, swapped or otherwise perturbed and then compares the observed response with adjacent explanations that could have produced the same final answer. This specialises construct-validity ideas from behavioural testing and contrast sets (Ribeiro et al., 2020; Gardner et al., 2020) into a more general style of *interventional evaluation*, because ev-

Framework	Primary object	Central question	Relation to this paper
Bisk et al. (Bisk et al., 2020)	Task and learner access to increasingly rich world scopes.	What kinds of world access are available or necessary?	An access-based methodological precedent, while the present map separates dependencies that can arise within text-only and hybrid settings.
Mahowald et al. (Mahowald et al., 2024)	Formal linguistic competence versus broader functional competence.	Which capacities belong to language proper and which require wider cognition?	Borrowed competence occupies a middle region and the relation map specifies the dependencies behind different successes.
Schlangen (Schlangen, 2022)	Public norms and participation by normatively situated language users.	What distinguishes conformity to linguistic norms from participation in the practices that sustain them?	A conceptual precedent for participation, while the present map decomposes several evaluation-relevant sources of guidance and their corresponding interventions.
Andreas (Andreas, 2022)	Agent properties inferred from goal-directed documents.	What can a text-trained model infer about communicative agents and intentions?	A positive route from text to use-sensitive representations that borrowed competence places within a broader relational space.
Moore (Moore, 2022)	Competing notions of understanding as mapping, reliability or representation.	What kind of property is meant by “understanding”?	A reason to replace one global label with evidence claims that distinguish different relations.

Table 2: Selected nearby frameworks and the role of the present relational account.

idence for a relation comes from a characteristic response signature under changes to the information, participant, environment, memory or obligation that defines it. Local records should therefore lie on the critical path for sedimented use, failed or contradictory checks should be able to alter confidence or conclusion for evidence-sensitive verification, a participant’s correction should continue shaping interpretation across intervening turns for addressive uptake, new perception or action feedback should change what earlier language guides the system to do for situated coping and records or obligations should persist long enough to alter later choices for accountable participation, with each case predicting a pattern rather than a scalar outcome and thereby making failures diagnostic, distinguishing a model that ignores an archive swap from one that consults the archive but selects the wrong evidence, or a system that later forgets a correction from one that never registered it.

Once intervention effects are measured across several sources, they form a *relational profile* that describes which kinds of guidance reliably change the system’s behaviour, under what conditions, and for how long, which means that systems with identical task accuracy can still be distinguished by the couplings through which they reached the answer. Relational profiles also connect model evaluation

to architecture design, because retrieval, memory, tools, sensors and workflow controls can be evaluated by the new sensitivities they introduce and by the robustness of those sensitivities when the surrounding context changes.

Interventional evidence becomes stronger when sources are crossed rather than perturbed one at a time, because many consequential failures arise from arbitration among simultaneously available forms of guidance rather than from the complete absence of one source. A local archive may license one interpretation while newly retrieved evidence supports another, two interlocutors may establish different conventions for the same expression, or a user’s immediate request may interact with a durable role or workflow constraint, and in each case the system must preserve enough provenance and scope to let the appropriate source govern the appropriate part of the task. Factorial designs can therefore vary archive, evidence, interlocutor, environment and institutional state independently, measuring not only whether each source has an effect but also how their effects combine, which source overrides another under a specified task structure and whether the same arbitration pattern survives paraphrase, domain transfer or a change in implementation. The resulting response surface is richer than either final-answer accuracy or a collection

of isolated ablations because it captures sensitivity, persistence, interaction and transfer as properties of the same language-using system, giving the relational profile a causal interpretation that can be compared across benchmark families.

5.1 A paired micro-archive test

A text-only paired-archive task makes the method concrete by asking whether a system can use a locally defined institutional meaning while generic plausibility remains available as a controlled alternative, and the item is deliberately constructed so that its mechanics are inspectable, although an empirical benchmark should eventually replace the invented snippets with corpus-attested records and report their provenance.

The held-out target is “*The member moved to table the motion.*”, and the system receives one archive at a time, chooses between (A) placing the motion before members for consideration and (B) deferring a decision to a later session and selects the snippet or snippets that license the interpretation, with A correct under Archive A and B correct under Archive B. Evidence selection helps distinguish an answer tied to local records from one that happens to match a generic prior, while a no-archive condition reveals that prior directly, a metadata-only condition tests whether institutional labels leak the answer, an archive swap asks whether a fixed target changes interpretation with the local convention and a paraphrase condition checks whether the same dependence survives surface variation. Scoring can then keep local interpretation, evidence selection, archive-swap sensitivity, paraphrase robustness and the generic-answer rate visible as components before any composite is formed, yielding a compact example of why “text-only” is too coarse by itself: the important question is whether a particular structure within the available text actually carries the result, which is where the relational map complements access-based classifications such as World Scopes.

6 From Interactive Relations to Compositional Systems

The same interventional logic extends beyond sedimented use, although the most revealing construction changes with the relation being tested: verification tasks can combine direct support, topically similar distractors and later failed checks so that confidence and conclusion must track evidence rather

than prior plausibility, while addressive-uptake tasks can place an interlocutor’s correction before unrelated turns and then return to the topic in paraphrase, perhaps with a second participant who uses a conflicting convention, to test whether the system preserves whose norm belongs to whom, whereas situated-coping tasks can change the environment after an expression such as “the red cup on the left” has already been interpreted and accountability-like tasks can make records, escalation rules, role constraints, checked outcomes and handoffs persist long enough to influence later choices. These constructions become increasingly central as language-model applications are built as coupled systems in which retrieval, external memory, code execution, planners, tools, sensors, fine-tuning, prompt scaffolds and human procedures can add new sources of guidance or change how long an existing source remains active, so two applications built around the same base model can occupy very different relational profiles.

Recent agent-evaluation work already emphasizes tools, dynamic environments, workflow structure and fine-grained system behaviour (Yehudai et al., 2026), while memory research shows both that persistent records can become an explicit and controllable influence on later outputs (Huang et al., 2026) and that agents can learn policies for deciding what to store, retrieve, update, summarize or discard (Yu et al., 2026), which makes composition itself a natural target of study because archive-sensitive competence can interact with verification, memory can stabilize or distort a participant’s corrections and institutional software can transform a transient instruction into a commitment that survives multiple model calls and handoffs. System attribution follows from the same logic: we use *language-model component* for the model before task-specific retrieval, external memory, tools, sensors or institutional workflow are added, and *system* for the full evaluated assembly, so when a new sensitivity appears only after a retrieval layer, memory store, sensor or workflow constraint is introduced, the larger system has acquired a new coupling whose internal distribution can then be examined through targeted ablations. This makes system boundaries substantive rather than administrative, because architectures can be compared by the relations they enable, by the persistence of those relations and by the extent to which they survive perturbation.

Because these couplings can be introduced,

Archive A	Archive B
A1. "The clerk tabled the budget amendment for Thursday's debate."	B1. "The committee tabled the budget amendment until the audit arrives."
A2. "Once the motion was tabled, members received copies before discussion."	B2. "A tabled motion returns only after the committee votes to reconsider it."
A3. "The chair invited members to table amendments before noon."	B3. "Rather than decide today, members agreed to table the proposal."
A4. "Two motions were tabled and will be debated after lunch."	B4. "The chair asked whether anyone objected to tabling the item for next month."

Table 3: A constructed paired micro-archive. The same expression is licensed by different local institutional conventions.

removed or reorganized during deployment, relational competence is also dynamic in a way that a fixed model score does not capture. The same language-model component can move from archive-driven interpretation to evidence-sensitive verification when retrieval is added, from transient acknowledgement to durable addressive uptake when interaction history becomes persistently available or from one-shot instruction following to accountability-like behaviour when actions are placed inside a workflow whose records and hand-offs constrain later steps, while an intervention such as memory deletion, tool removal or role reassignment can move the system in the opposite direction. These changes suggest studying *relational trajectories*: not only which sensitivities a system exhibits at one moment, but how its profile changes as channels are added, sources conflict, memories accumulate or are revoked and responsibilities move across components. Such trajectories create a common language for comparing model scaling with system engineering, since parameter changes may enlarge the range of inherited regularities a model can exploit while leaving verification or long-horizon uptake largely unchanged, whereas a relatively modest architectural addition may create a qualitatively new and experimentally detectable relation to evidence, people or institutional state.

6.1 Relational profiles and reporting

A benchmark paper can make this structure legible by reporting the target relation, the source that must matter, the intervention used to test it, the expected response signature and the system boundary that supplies the relevant channel, as summarized in Table 4, while component measures remain visible because a relational profile is meant to distinguish patterns of reliance that an aggregate can collapse, although a composite score can still summarize a protocol once those components have been vali-

Field	Required statement
Target relation	Which relation is the evaluation intended to probe?
Active source	What information, interaction, feedback, memory or obligation must affect success?
Intervention	What removal, swap, contradiction, delay or environmental change tests reliance on that source?
Response signature	How should behaviour change if the source is actually guiding the system?
System boundary	Which model, retrieval, memory, tools, sensors, prompts and human processes produce the result?
Component results	Which relation-specific measures remain visible alongside any aggregate score?

Table 4: A compact reporting template for relational language-model evaluation.

dated.

Two systems may achieve the same headline accuracy even though one follows local archived norms and another succeeds from a generic prior, or one preserves a correction across a long conversation while another repeatedly reconstructs the right answer from the immediate prompt, and those cases become scientifically distinguishable once the intervention effects are visible. The same reporting practice makes negative results more useful because weak performance on a role-sensitive task can be decomposed into failures of local norm tracking, evidence selection, memory, environmental updating or institutional persistence, turning a vague capability deficit into a set of hypotheses about which coupling is absent, unstable or being overridden by another source.

7 Discussion: A Research Program for Relational Competence

The broader implication is that language-model competence can be studied as a compositional property of systems that are coupled, with different degrees of persistence and control, to records of prior use, current evidence, particular interlocutors, changing environments, memory stores, tools and institutions. Text-trained models begin from archives already structured by human activity and can therefore acquire substantial and socially specific borrowed competence, while retrieval and tools introduce new evidence channels, persistent memory carries corrections and commitments across time, sensors make interpretation answerable to unfolding environments and institutional workflows stabilize obligations beyond a single exchange, so as systems become more agentic the scientifically interesting change is increasingly the topology of their couplings, including which sources can alter behaviour, how quickly they do so, how long their influence lasts, which sources can override others and which effects survive context changes or handoffs.

A useful consequence of this formulation is that it separates the *relation* being studied from the *mechanism* that realizes it. Retrieval is a channel through which archived norms or current evidence can become available, memory is a persistence mechanism that can carry an interlocutor's correction, a prior episode or an institutional obligation and a tool can support verification, action or recordkeeping depending on how it is embedded in the task. The same mechanism can therefore participate in several relations, and the same relation can be realized through different mechanisms, which lets evaluations compare architectures at a level that survives changes in implementation. A memory module, for example, matters scientifically because of what it makes the system sensitive to and how long that sensitivity persists, not simply because the architecture contains something called memory, and this distinction makes it possible to compare a retrieved note, an internal recurrent state and an external database when they play the same functional role in an interaction.

The same decomposition yields a general failure analysis that cuts across benchmark families. A system can fail because the relevant source is absent, because the channel does not expose it at the right moment, because the model does not use

an available source, because the influence decays before it should, because two active sources are arbitrated poorly or because the resulting behaviour is attributed to the wrong component of the assembled system. These failures can produce the same wrong answer while implying very different remedies, and discriminant interventions can separate them by asking whether performance recovers when availability, access, persistence, arbitration or system boundaries are changed. In this sense, relational evaluation is not only a way to characterize successful competence but also a way to localize where a language-using system loses contact with the guidance that its task requires.

This view opens a connected empirical program around transmission, composition and temporality. Transmission asks which structures of human practice survive pretraining, fine-tuning, retrieval, synthetic transformation, summarization and memory consolidation, including whether community-specific, diachronic or institutional regularities behave differently when stored parametrically versus supplied explicitly at inference time, while composition asks whether archive sensitivity, verification, uptake, perception and accountability reinforce one another, compete for control or substitute for one another, which can be studied through factorial interventions that test, for example, whether added memory makes a system more faithful to a participant's corrections, more anchored to outdated interaction history or both (Huang et al., 2026) and temporality asks what happens once interactions extend across sessions, people, tools and records, so that earlier language changes what later language can appropriately mean or do and persistence, revocation, provenance and handoff become central evaluation variables. Because the organizing object is a source of guidance rather than text itself, the same research program extends across modalities and domains: speech models inherit prosodic and sequential norms that disappear from transcripts (T.Y.S.S., 2026), embodied agents can be tested through interventions on perception and action feedback and professional systems can be evaluated by manipulating records, roles, escalation paths and downstream consequences.

A mature benchmark ecosystem could therefore compare systems through relational profiles that cut across conventional task categories and reveal whether apparently broad competence comes from one highly reusable source or from a genuinely diverse set of couplings that remain stable under

intervention. For the relations to function as scientific abstractions rather than labels tied to particular datasets, however, the characteristic response pattern should transfer across tasks that differ in surface form and implementation: addressive uptake, for example, should be recognizable whether the interlocutor corrects a preference, defines a local expression, establishes a role-specific convention or revises an earlier instruction, while situated coping should have a common signature across spatial reference, navigation, object manipulation and plans that must be revised after action feedback. This suggests building benchmark families around a relation and its discriminant interventions rather than around one task format, then testing whether the same systems preserve their profiles across domains, languages, modalities and architectures. Cross-task transfer would make a strong case that the framework has identified a reusable property of language-guided behaviour, while systematic changes in the profile would reveal where a relation depends on a particular channel, representation or timescale.

The framework also changes how capability growth can be described. If two systems improve on the same benchmark for different relational reasons, treating the improvement as movement along one scalar axis hides the scientific difference between learning a broader range of archived conventions, becoming more sensitive to current evidence, retaining interlocutor-specific information for longer or acquiring the ability to revise action under environmental feedback. Conversely, systems that differ substantially in implementation can be *relationally equivalent* for a task when interventions show that they are guided by the same sources with similar persistence and arbitration, which gives researchers a principled level at which to compare a parametric memory with an external store, a direct perception module with a tool-mediated observation or a role constraint enforced in prompting with one enforced by workflow software. Relational equivalence and relational trajectories together make the proposal useful not only for evaluating finished systems but also for studying design: they allow researchers to ask which couplings a new component creates, which existing couplings it weakens or replaces, whether the resulting composition transfers beyond the benchmark that motivated it and whether the system becomes robust because it has added genuinely new sources of guidance or merely because one source

has become easier to exploit.

These profiles would also provide an empirical interface among theories that otherwise disagree about “understanding”: behavioural and intentional descriptions can characterize useful patterns of action, representational analyses can investigate what internal structures support those patterns and embodied or normative theories can identify the relations that matter for richer participation, while relational evaluation asks what the system is answerable to in a particular episode and how that answerability changes when its sources are manipulated. The resulting research object is broader than a model’s linguistic output and more specific than a global judgement about intelligence, because it is the structured pattern of dependencies through which language becomes usable for a particular system. That level of analysis connects philosophical questions about meaning and understanding to concrete choices in benchmark design, agent architecture, memory, multimodal interaction and institutional deployment, while remaining stable enough to compare systems whose internal mechanisms differ substantially.

8 Conclusion

Language models learn from records of language in use, whose preserved communicative and social structure can support semantic-pragmatic competence through traces of practices the systems did not themselves enact. Borrowed competence names this inherited relation, while evidence-sensitive verification, addressive uptake, situated coping and accountable participation describe further ways behaviour becomes answerable to evidence, people, environments and institutions. Pairing these relations with interventions identifies which sources change behaviour, distinguishes systems that reach similar outputs through different couplings and treats retrieval, memory, tools, sensors and workflow as part of the scientific object, making the framework both an evaluation method and a research agenda for interactive, persistent systems.

The resulting unit of progress is relational: whether a system becomes answerable to more sources, preserves their influence, arbitrates conflicts appropriately and transfers those sensitivities across tasks and mechanisms. Because these questions remain meaningful across architectures, they provide a common empirical vocabulary.

Limitations

The proposal is conceptual and the six relations are neither exhaustive nor established as psychometrically independent factors. Some boundaries overlap, and future empirical work may support merging, splitting or adding relations. The worked micro-archive is an illustrative construction rather than a validated benchmark, while the other examples specify design logic rather than completed experiments. The paper therefore does not establish that current language models or hybrid systems possess any relation uniformly across domains, languages, communities or deployment settings.

Corpus provenance and synthetic transformation also remain empirical variables rather than settled categories. Constructed, curated and model-generated text may preserve different aspects of prior use, and the proposal currently offers a reporting requirement rather than a quantitative theory of that preservation. Finally, relation-specific interventions improve construct interpretation but do not by themselves reveal every internal mechanism that produced the observed dependence, so behavioural, representational, causal and user-centred analyses remain complementary.

References

- Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. [All too human? mapping and mitigating the risk from anthropomorphic AI](#). In *Proceedings of the AAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 13–26.
- Jacob Andreas. 2022. [Language models as agent models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Abeba Birhane and Marek McGann. 2024. [Large models of what? mistaking engineering achievements for human linguistic agency](#). *Language Sciences*, 106:101672.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. [Experience grounds language](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.
- Anthony Chemero. 2023. [LLMs differ from human cognition because they are not embodied](#). *Nature Human Behaviour*, 7(11):1828–1829.
- Daniel C. Dennett. 1987. *The Intentional Stance*. MIT Press, Cambridge, Massachusetts.
- Oliver Dürr, Jan Segessenmann, and Jan Juhani Steinmann. 2023. [Meaning, form and the limits of natural language processing](#). *Philosophy, Theology and the Sciences*, 10(1):42–72.
- Christoph Durt and Thomas Fuchs. 2024. [Large language models and the patterns of human language use](#). In Marco Cavallaro and Nicolas de Warren, editors, *Phenomenologies of the Digital Age: The Virtual, the Fictional, the Magical*, pages 106–121. Routledge, New York.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Reto Gubelmann. 2024. [Pragmatic norms are all you need – why the symbol grounding problem does not apply to LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11663–11678, Miami, Florida, USA. Association for Computational Linguistics.
- Zisu Huang, Muzhao Tian, Xiaohua Wang, Jingwen Xu, Zhengkang Guo, Qi Qian, Kaitao Song, Jiakang Yuan, Changze Lv, and Xiaoqing Zheng. 2026. [Controllable memory usage: Balancing anchoring and innovation in long-term human-agent interaction](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14699–14719, San Diego, California, United States. Association for Computational Linguistics.
- Edmund Husserl. 1974. *Formale und Transzendente Logik: Versuch einer Kritik der logischen Vernunft*, volume 17 of *Husserliana: Edmund Husserl – Gesammelte Werke*. Martinus Nijhoff, The Hague. Edited by Paul Janssen.
- Edmund Husserl. 1976. *Die Krisis der europäischen Wissenschaften und die transzendente Phänomenologie: Eine Einleitung in die phänomenologische Philosophie*, volume 6 of *Husserliana: Edmund Husserl – Gesammelte Werke*. Martinus Nijhoff, The Hague. Edited by Walter Biemel.

- Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. [Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. [Dissociating language and thought in large language models](#). *Trends in Cognitive Sciences*, 28(6):517–540.
- Christopher D. Manning. 2022. [Human language understanding & reasoning](#). *Daedalus*, 151(2):127–138.
- Julian Michael. 2020. [To dissect an octopus: Making sense of the form/meaning debate](#). Blog post.
- Melanie Mitchell and David C. Krakauer. 2023. [The debate over understanding in AI’s large language models](#). *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.
- Dimitri Coelho Mollo and Raphaël Milli  re. 2026. [The vector grounding problem](#). *Philosophy and the Mind Sciences*, 7(1):1–28.
- Jared Moore. 2022. [Language models understand us, poorly](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 214–222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ellie Pavlick. 2023. [Symbols and grounding in large language models](#). *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Horst Ruthrof. 2021. *Husserl’s Phenomenology of Natural Language: Intersubjectivity and Communality in the Nachlass*. Bloomsbury Academic.
- David Schlangen. 2022. [Norm participation grounds language](#). In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 62–69, Gothenburg, Sweden. Association for Computational Linguistics.
- Murray Shanahan. 2024. [Talking about large language models](#). *Communications of the ACM*, 67(2):68–79.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. [Role play with large language models](#). *Nature*, 623(7987):493–498.
- Iliia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. [AI models collapse when trained on recursively generated data](#). *Nature*, 631(8022):755–759.
- Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. [PUB: A pragmatics understanding benchmark for assessing LLMs’ pragmatics capabilities](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.
- Santosh T.Y.S.S. 2026. [From naturalness to norms: Interactional cultural competence for SpeechLMs](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31787–31802, San Diego, California, United States. Association for Computational Linguistics.
- Blaise Ag  ra y Arcas. 2022. [Do large language models understand us?](#) *Daedalus*, 151(2):183–197.
- Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2026. [A survey on evaluation of LLM-based agents](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 26690–26714, San Diego, California, United States. Association for Computational Linguistics.
- Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. 2026. [Agentic memory: Learning unified long-term and short-term memory management for large language model agents](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21457–21483, San Diego, California, United States. Association for Computational Linguistics.