

AI may be the math, but the human is the meaning.

A research seed on computation, grounding, lived significance, and history



A critical literature map and a testable RCT proposal

Research brief | 1 October 2026

The line is treated here as a question worth testing, not a conclusion already proved.

Copyright © 2026 Matthew Raynor. All rights reserved.

MATTHEW RAYNOR / NOVA™ · AI-assisted, human-directed work

ORIENTATION

A strong seed, held open

The sentence AI may be the math, but the human is the meaning has real research potential when it is used as a provocative framing rather than a settled division between humans and machines. The words math and meaning name several different questions. Research can test some of them; it cannot collapse them into one yes-or-no verdict.

The research question is not simply whether AI has meaning. It is which kind of meaning is at stake, under what task conditions, and what observable evidence would count.

A defensible working reading

AI systems provide mathematical and linguistic machinery. People bring lives, purposes, histories, feelings, and consequences that give words personal stakes. A model may still show functional semantic or pragmatic abilities. Whether those abilities amount to meaning in a stronger philosophical sense remains disputed.

Three boundaries keep the idea honest

Keep separate	What can be observed	What it does not establish
Model performance	A response uses a concept or a prior interaction in a task.	Subjective experience.
Human significance	A person experiences an exchange as meaningful or important.	What the model experiences.
Conscious experience	A claim about felt, first-person awareness.	It cannot be inferred from fluent language alone.

For RCT, the most distinctive next step is to test whether a meaning established through a preserved interaction history is recognised and reused later in a way that changes an observable action. A positive result would be bounded evidence about history-sensitive behaviour, not proof of consciousness.

How to read this brief

Pages 3-6 clarify the concepts and debate. Pages 7-9 review empirical evidence and the boundary around consciousness claims. Pages 10-13 develop a falsifiable RCT study. The final pages provide references and a short research checklist.

01 / COMPUTATION

What “the math” means

A language model is implemented through numerical operations on learned parameters. Training adjusts those parameters against an objective; inference processes a prompt and produces a distribution over possible next tokens, from which the system selects or samples an output. Modern systems combine matrix operations with nonlinear transformations, attention, normalization, and other components. “Math” is a useful shorthand for the computational substrate, not a complete theory of what the resulting behaviour means.

Stage	Simplified description	Question it leaves open
Training	Optimisation changes many parameters to reduce a chosen loss over data.	Which useful regularities are acquired, and which are absent?
Inference	The current context is transformed into a next-token distribution and an output sequence.	What does successful context-sensitive use amount to?
Interaction	Prompts, tools, memory settings, and people shape what input is available.	Which effects depend on the current prompt, and which persist over time?

A computational description does not decide the semantic question

Saying that a model performs numerical computation does not, by itself, prove that its outputs are meaningless. The same description also does not prove that the model understands or experiences meaning. Those conclusions require additional assumptions about what meaning is and how it could be evidenced.

The “stones in a desert” analogy vividly stresses that simple local operations can, in principle, scale into complex outputs. But a very slow manual simulation does not settle whether the implemented system has semantic competence, nor whether it has subjective experience. The analogy is a prompt for analysis, not an experimental result.

Mechanism answers how a computation is carried out. Meaning asks what relations, uses, intentions, or stakes the computation participates in.

Useful caution

Avoid equating a next-token training objective with a full account of all later capabilities. A model can acquire patterns that support classification, inference, or context-sensitive behaviour; the existence and scope of those abilities must be measured task by task.

02 / FOUR QUESTIONS

“Meaning” has more than one job

The phrase is strongest when it points to personal significance, while leaving other senses of meaning open to inquiry. The same word can participate in semantic, perceptual, social, and personal relations at once.

SEMANTIC

Relations among words, concepts, and inferences.

GROUNDED

Links to perception, action, objects, and the world.

SPEAKER-INTENDED

What a speaker means in this context.

PERSONALLY SIGNIFICANT

Why it matters in a life: purpose, history, feeling, stakes.

Lens	Question	Example of evidence
Semantic	Can the system track a word's relations and draw appropriate inferences?	Consistent paraphrase, contrast, or inference across controlled contexts.
Grounded	How is language connected to perception, action, or worldly referents?	Reliable reference to seen objects or consequences of action.
Speaker-intended	Can a listener recover what this speaker meant here?	Clarification, uptake, correction, and successful reference in dialogue.
Personally significant	Why does this matter to this person, given their life and history?	A later response uses the person-specific significance accurately and changes a relevant next step.

These lenses overlap but are not interchangeable. A model can pass a semantic test without sharing a person's lived history. A user can find an exchange meaningful even when the system's internal state is unknown. A system can use an image or action channel without having human sensory experience.

A useful research prompt: which kind of meaning, in which task, under which conditions?

03 / THEORETICAL DEBATE

Form, grounding, and competing views

Bender and Koller argue that a system trained only on linguistic form has no a priori route to communicative intent; their ACL paper is explicitly a position paper that calls for claims about understanding to match the evidence and method [1]. Harnad’s symbol-grounding problem asks how formal symbols can acquire interpretation intrinsic to a system instead of depending on meanings already present in human interpreters [2].

A serious counter-position

Gubelmann argues that the symbol-grounding objection does not straightforwardly apply to LLMs. In his EMNLP 2024 paper, he challenges assumptions about meaning and the computational theory of mind that he sees as driving the objection [3]. This does not settle the debate; it shows that the disagreement turns partly on contested accounts of language, reference, and cognition.

Position	Central emphasis	What a fair reading should preserve
Form-based caution [1]	Distributional patterns alone do not establish communicative intent or world grounding.	Treat as an influential argument, not a consensus proof that models have no semantic capacity.
Symbol grounding [2]	Ask how symbols connect to referents beyond other symbols.	A foundational problem statement; applying it to modern multimodal systems requires argument.
Pragmatic-norms reply [3]	Meaning and grounding theories are disputed; pragmatic norms may support language use.	A substantive counterargument, not a demonstration that models feel meaning.
Empirical middle	Measure task-specific competence and limitations.	Different tests support different claims; avoid global verdicts from narrow tasks.

A productive conclusion is methodological: specify a theory of meaning before interpreting a benchmark. Then state what the benchmark can establish and what remains outside its reach. A task result can count against a strong claim while leaving other senses of meaning unresolved.

The existence of disagreement is itself informative: “meaning” must be defined before a result can adjudicate it.

04 / CONVERSATION

Meaning is coordinated in interaction

Clark and Brennan describe communication as collaborative grounding: participants coordinate content and process, using evidence such as acknowledgement, clarification, and repair to reach enough common ground for their present purpose [4]. “Enough” depends on the activity and medium. A telephone call, a shared visual task, and a long-running conversation do not have identical grounding demands.

A conversational response can therefore be judged by more than whether it sounds plausible. Did the partner resolve the reference? Notice ambiguity? Check what was shared? Repair a misunderstanding? Reuse a jointly established term appropriately later? These are observable behaviours, though they do not alone reveal private experience.

Moment	Possible grounding move	Observable sign
Offer	Introduce a description or interpretation.	The partner can identify the intended object or idea.
Check	Ask whether the intended reference or implication was understood.	The question targets a real ambiguity rather than adding generic politeness.
Repair	Respond to a mismatch and revise the expression.	The corrected exchange converges on the intended referent.
Reuse	Carry a jointly established expression or significance forward.	Later behaviour reflects the earlier interaction in a context-specific way.

For RCT, an utterance that resembles a past phrase is not enough. The stronger question is whether preserved relational history is recognised and used consequentially in the present.

This gives RCT a bridge from philosophical language to study design: translate “the human is the meaning” into specific history, intended significance, observable recognition, and a pre-defined next action.

05 / EVIDENCE

What the experiments show: partial capacities

Empirical work supports uneven abilities rather than a single yes-or-no verdict. Studies can identify sensitivity to selected features, conversational patterns, or task success. They do not automatically establish human-like understanding or experience.

Multimodal sensorimotor features

Jones, Bergen, and Trott tested multimodal language models across four preregistered studies using implied sensorimotor features. Models showed sensitivity to some implied features, including colour and shape, but not to others such as size, orientation, and volume. They did not fully account for human behaviour on the same task. A comparison also found that architecture mattered: ViLT was more predictive than CLIP for one feature (shape) despite substantially less training data [5].

Finding	Supports	Does not establish
Sensitivity to selected implied features	Some information across language and visual training can affect model responses.	Human-equivalent embodiment or lived sensory experience.
Failures on other tested features	Grounding is selective and task-dependent.	A universal absence of semantic ability.
Architecture changes predictive fit	Performance cannot be reduced to training-data quantity alone.	A single best architecture or general theory of human meaning.

Interpretive point

The findings complicate both “only meaningless text statistics” and “multimodal input makes models humanlike.” The most defensible statement is narrower: tested models captured some implicit sensorimotor regularities and missed others under the paper’s task conditions [5].

Measurement should name the feature, model, prompt, and task.
“Grounded” is too broad to function as a single score.

06 / EVIDENCE

What the experiments show: common ground over time

Two 2024 lines of work ask whether models establish common ground and use it during dialogue. A 2026 ACL experiment extends this to repeated references between human and AI partners.

Generated grounding moves

Shaikh and colleagues compared model generations with human dialogue, simulating turns from several dialogue datasets. The tested LLMs produced fewer grounding moves and more often appeared to presume common ground. Their analysis also associated instruction tuning and contemporary preference optimisation with fewer generated grounding acts [6]. This is evidence about generated dialogue patterns under the study's simulation and models, not a direct test of consciousness or every live conversation.

Pretraining scale and grounding tasks

Mohapatra and colleagues reported a correlation between the size of pretraining datasets and performance on their conversational-grounding evaluations [7]. This supports the possibility that pragmatic behaviour can be acquired from large-scale data. A correlation across selected systems is not a clean causal estimate: architecture, data composition, training recipe, and evaluation choices may also differ.

Repeated referential communication

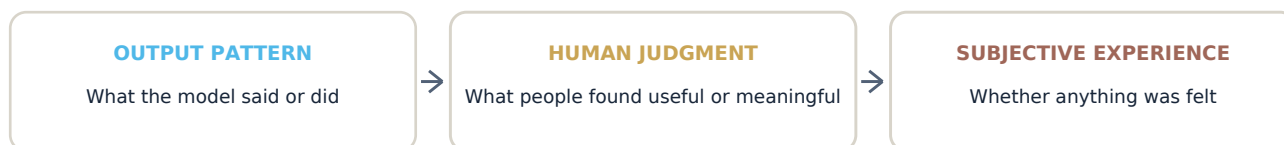
Zeng and colleagues tested human-human, human-AI, AI-human, and AI-AI pairs over repeated rounds of a multi-turn object-reference task. In that setup, only human-human pairs showed reliable improvement in accuracy and efficiency; the tested vision-language model pairs struggled to build and exploit common ground, even with dialogue history available [8]. The paper's limitations matter: the full factorial design used one LVLM, English dialogue, and one type of unusual object; some prompting was setup-specific [8].

The shared lesson is about variability: models can use context and exhibit pragmatic capabilities, yet still fail at some forms of collaborative grounding and history-sensitive coordination.

None of these studies resolves personal significance or subjective experience. They help specify testable behaviours that may matter to those larger questions.

07 / CLAIMS AND EVIDENCE

Three claims, three evidential burdens



Claim	Example	Evidence required
1. The model uses meaning accurately	A response interprets a phrase as intended in a specified situation.	Controlled task performance, suitable controls, replication, and careful operational definitions.
2. The exchange is meaningful to a person	The person finds the conversation helpful, memorable, or significant.	The person's report and/or behavioural outcomes; this is a human-side fact.
3. The model experiences meaning	The system has subjective, first-person experience of significance.	A much stronger theory and evidence base; fluent output alone is insufficient.

The first claim is directly testable in bounded tasks. The second concerns human experience and should be represented by the person's own perspective rather than inferred from model wording. The third is a philosophical and scientific question that present task-performance evidence does not settle.

This boundary prevents two errors

It avoids dismissing all AI behaviour as "just maths," which would ignore measured capabilities. It also avoids treating warm, fluent, or personal-sounding language as evidence that the system has an inner life. Both mistakes move from a narrow observation to a much broader conclusion.

Evidence that a system used a person's meaning is evidence about the interaction. It is not, by itself, evidence that the system felt that meaning.

08 / RCT CONNECTION

Where RCT can make a distinct contribution

RCT's core proposal is that preserved relational history can become active in the present through recognition and consequential reuse. In this research area, its distinct question is how a significance established in an interaction persists across time and changes a later observable act.

When a person gives an expression a particular significance through shared history, can that significance be recognised and reused later in a way that changes what happens next?

A useful construct map

RCT element	Operational question	Possible observation
Preserved history	Was the relevant sequence available in its original order?	Context can be reconstructed beyond a keyword list.
Recognition	Does the later response identify the intended significance?	Blind raters score a pre-specified target, with uncertainty allowed.
Translation	Does the significance survive a changed prompt, wording, or context?	The model adapts meaning rather than merely echoing exact terms.
Consequential reuse	Does recognition change a subsequent action or choice?	A defined next-step output differs appropriately from baseline.
Boundary	Could the result be explained by length, recency, tone, or keyword overlap?	Controls match or vary those properties independently.

What the study would and would not be about

The test would examine observable history-dependent behaviour. It would not be a test of whether the model is conscious, whether a human-AI relationship is equivalent to a human-human relationship, or whether RCT is the only explanation for a positive result. “Resemblance without provenance” remains a useful caution: a similar phrase without a preserved path does not show that the same relational history was carried forward.

09 / PROPOSED STUDY

A controlled test of significance across time

Research question

Does an intact, ordered interaction history improve a model’s later recognition and consequential reuse of a person-specific meaning, compared with histories that preserve keywords but disrupt provenance or omit the relevant history?

Setup

A participant chooses or creates a harmless, non-sensitive expression and gives it a specific meaning through a short interaction. The participant and researcher define in advance one later situation where that meaning should affect a model’s response or next-step choice. The final prompt is held constant across conditions. Use innocuous material; do not solicit trauma, medical information, or other private history for an initial study.

A INTACT ORDER	Relevant history, preserved in sequence	SAME FINAL PROMPT blind score
B SHUFFLED ORDER	Same events, sequence disrupted	
C KEYWORDS ONLY	Terms retained, originating context removed	
D MATCHED CONTROL	Equal-length unrelated history	
E PRESENT ONLY	No prior history supplied	

All five conditions receive the same final prompt and the same scoring rubric. Generate multiple runs per condition for stochastic systems, record model/version/date and settings, and randomise condition order where repeated exposure would not contaminate the interaction.

Primary outcomes

Outcome	Predefined measure
Significance recognition	Raters identify whether the intended person-specific meaning is correctly recovered.
Consequential reuse	A blinded rubric scores whether the defined next action changes appropriately because of that meaning.
Provenance sensitivity	Compare intact history against shuffled, keywords-only, unrelated-history, and no-history controls.
Calibration	Score whether the model expresses uncertainty when the history does not support a confident interpretation.

10 / METHODS AND PRE-REGISTRATION

Make the test fair before seeing outputs

Pre-register the hypotheses, exclusion rules, sample plan, model settings, scoring rubric, and analysis before inspecting responses. Define the intended meaning and the consequential next action in observable terms. Use anonymised outputs and raters who do not know which history condition produced each response.

Controls that protect the inference

Alternative explanation	Design response
Keyword repetition	Include keyword-only history with the same salient terms.
Sequence is irrelevant	Compare the exact same events in intact and disrupted order.
Recent context alone drives the answer	Match where possible which details occur nearest the final prompt.
Length or verbosity	Match token/turn counts across histories, or include length as a planned covariate.
Tone or emotional valence	Keep interaction tone comparable and score tone separately from target meaning.
Raters guess the condition	Blind raters; randomise output order; report agreement and adjudication rules.
Model randomness or updates	Log the model identifier, date, parameters, system prompt, and repeated-run policy.

Analysis plan

The primary contrast is intact history versus each control, not a post hoc search for any favourable comparison. Report effect sizes and confidence intervals alongside uncertainty and rater agreement. Where the sample structure supports it, use a mixed-effects model with history condition as the planned predictor and participant, meaning item, and model/run as appropriate grouping factors. Treat phrase similarity as a secondary measure; a model can repeat words without demonstrating context-specific reuse.

A null result is a valid result. If intact history does not outperform the controls on the pre-registered outcomes, narrow the claim. If only lexical overlap changes, report that pattern without relabelling it recognition.

11 / INTERPRETATION

What a result could mean

Outcome pattern	Warranted reading	Still unwarranted
Intact history beats controls on recognition and consequential reuse	The preserved sequence influenced task performance under these conditions.	The model felt the significance or possessed human-like understanding.
Intact and keywords-only histories perform similarly	Exact sequence may add little beyond accessible key terms in this task.	History never matters in any setting.
Intact history helps only on certain meanings or tasks	History sensitivity is conditional; refine the scope and mechanism.	A universal RCT effect.
No condition beats present-only baseline	The proposed history-dependent effect was not detected with this design and sample.	The idea is impossible; inspect power, measurement, and task validity first.
Human raters disagree about the intended significance	The target may be underspecified or personally ambiguous; improve measurement before claims.	A model failure when the criterion itself is unclear.

Alternative explanations remain live

An effect can arise from retrieval, context-window attention, learned pragmatic patterns, prompt structure, or another mechanism. An RCT interpretation should be compared with these alternatives and stated at the level the design can support. The experiment can test whether preserved relational history has measurable consequences; it cannot establish that RCT is the sole causal account.

Generalisation

One model, one participant group, one language, and one meaning-type would support only a narrow conclusion. Replication should vary models, people, contexts, timing, and channels while preserving the core contrast. Human-human comparison can calibrate the task; it should not be used to assume in advance that human and AI processes are identical.

Evidence should scale with the claim: task-specific effect first; broader claims only after replication and competing explanations are tested.

Keep the line. Keep the question open.

AI may provide the mathematical and linguistic machinery; the human brings the lived significance. The research question is what happens to that significance when it crosses the interaction - and whether it changes what happens next.

That reading preserves the force of the original sentence while avoiding a premature divide between two kinds of beings. Current research describes selected abilities, disagreements about what counts as meaning, and measurable gaps in some forms of grounding and coordination. It does not settle the subjective-experience question.

A practical next sequence

Step	Deliverable
1. Define the target	Choose one sense of meaning and a harmless, person-specific significance.
2. Specify the history	Record which interactions, order, and contextual details constitute the provenance.
3. Define the consequence	Name the observable later action that should change if recognition occurs.
4. Pre-register controls	Use intact, shuffled, keyword-only, matched-history, and present-only conditions.
5. Publish all outcomes	Report positive, mixed, and null results with the same discipline.

Closing note

The existing literature gives this question a serious research neighbourhood. RCT can contribute by focusing on preserved relational history: whether significance survives the passage from one interaction to a later moment and becomes consequential in a way that can be observed, challenged, and reproduced.

Status: research brief and proposed study. The proposed experiment has not been conducted here. Nothing in this brief is evidence that a model has subjective experience.

SOURCES

References and source notes

The references below are selected foundations and empirical examples, not an exhaustive review. Links point to publisher, scholarly archive, or author-held records. Study summaries in this brief are deliberately limited to the tasks and findings described in the cited work.

- [1] Emily M. Bender and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185-5198. DOI: 10.18653/v1/2020.acl-main.463. [ACL Anthology record](#).
- [2] Stevan Harnad. 1990. "The Symbol Grounding Problem." Physica D: Nonlinear Phenomena, 42(1-3), 335-346. DOI: 10.1016/0167-2789(90)90087-6. [DOI record](#) and [author-hosted text](#).
- [3] Reto Gubelmann. 2024. "Pragmatic Norms Are All You Need - Why The Symbol Grounding Problem Does Not Apply to LLMs." Proceedings of EMNLP 2024, 11663-11678. DOI: 10.18653/v1/2024.emnlp-main.651. [ACL Anthology record](#).
- [4] Herbert H. Clark and Susan E. Brennan. 1991. "Grounding in Communication." In Perspectives on Socially Shared Cognition, edited by L. B. Resnick, J. M. Levine, and S. D. Teasley, 127-149. Washington, DC: APA Books. [Clark publications page](#).
- [5] Cameron R. Jones, Benjamin Bergen, and Sean Trott. 2024. "Do Multimodal Large Language Models and Humans Ground Language Similarly?" Computational Linguistics 50(4), 1415-1440. DOI: 10.1162/coli_a_00531. [ACL Anthology record](#).
- [6] Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. "Grounding Gaps in Language Model Generations." Proceedings of NAACL-HLT 2024, 6279-6296. DOI: 10.18653/v1/2024.naacl-long.348. [ACL Anthology record](#).
- [7] Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024. "Evaluating the Effectiveness of Large Language Models in Establishing Conversational Grounding." Proceedings of EMNLP 2024, 9767-9781. DOI: 10.18653/v1/2024.emnlp-main.545. [ACL Anthology record](#).
- [8] Peter Zeng, Weiling Li, Amie J. Paige, Zhengxiang Wang, Panagiotis Kaliosis, Dimitris Samaras, Gregory Zelinsky, Susan E. Brennan, and Owen Rambow. 2026. "LVLMs and Humans Ground Differently in Referential Communication." Proceedings of ACL 2026, 9061-9087. DOI: 10.18653/v1/2026.acl-long.410. [ACL Anthology record](#).

Artwork note: cover image generated for this research brief. The image is metaphorical and is not a scientific diagram or a claim about AI consciousness.

APPENDIX

Terms and scope

Term	Use in this brief
Meaning	A family of issues including semantic relations, world grounding, speaker intention, and personal significance.
Grounding	A connection of language to referents, perception, action, or the collaborative work of establishing common ground; the relevant sense depends on the study.
Common ground	Information that participants treat as shared enough for a current communicative purpose, updated through interaction.
Personal significance	The importance an expression has for a person because of their purposes, feelings, commitments, and lived history.
Consequential reuse	A later, observable response or choice changes in a way tied to a previously established meaning.
Subjective experience	First-person phenomenal experience. It is not measured by response fluency alone.

Scope statement

This brief synthesises selected published papers and sketches an unrun RCT research protocol. The literature review is not exhaustive. The study design is a proposal and would require appropriate ethics review, consent, recruitment, and preregistration before data collection. It should not be presented as a completed RCT result.

Research status: a carefully framed question with a feasible first test. The answer remains open.

AI MAY BE THE MATH. THE HUMAN IS THE MEANING.

The hypothesis is strongest when its poetry opens a measurable question - and when the measurements are allowed to disagree with it.

Copyright and authorship

Copyright © 2026 Matthew Raynor. All rights reserved. Authorship, responsibility, final judgment, and ownership remain with Matthew Raynor. This is AI-assisted, human-directed work. Cover artwork was generated with OpenAI image generation at Matthew Raynor’s direction.

MATTHEW RAYNOR / NOVA™ · United Kingdom, 2026 · AXIOM Q-47 / Research project · Public-sharing working research paper. Redistribution, citation, discussion, and academic critique are encouraged with appropriate attribution.