Are Language Models More Like Libraries or Like Librarians?

Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs

Harvey Lederman & Kyle Mahowald (2024)

Transactions of the Association for Computational Linguistics, 12, 1087–1103

The Central Question

Do LLMs have beliefs, desires, and intentions?

If yes: we can explain their behaviour the way we explain human and animal behaviour — via propositional attitudes.

If no: we need an alternative explanation of equal power. Bibliotechnism is the most developed attempt.

The paper defends bibliotechnism up to a point — then argues it breaks down at the problem of novel reference.

What Is Bibliotechnism?

Two defining commitments:

1. **LLMs are cultural technologies**
   Like libraries, printing presses, or photocopiers, they transmit information but do not create new content.

2. **LLMs do not have beliefs, desires, or intentions**
   If LLM text is meaningful, its meaningfulness must derive from human-generated text in the training data.

Term coined by the authors (βιβλίο + τέχνη). Inspired by Gopnik (2022) and Chiang (2023).

Basic vs. Derivative Meaning

Basic meaning: tokens created immediately by an agent (an entity with beliefs, desires, intentions).

Derivative meaning: tokens whose meaningfulness depends on an appropriate causal connection to a basically meaningful token.

The photocopier analogy

A photocopied page of Shakespeare's biography still refers to the poet. The causal chain from original to copy suffices — no human supervision is required. Even accidental copies are derivatively meaningful.

Consequence for bibliotechnism: LLMs can only produce derivatively meaningful tokens.

Form Cannot Teach Meaning : Bender & Koller 2020

Core thesis: a system trained only on form has a priori no way to learn meaning.

Key definitions:

Form: any observable realisation of language (marks on a page, pixels, bytes, articulatory movements).

Meaning (M): the relation $M \subseteq E \times I$ between natural language expressions (e) and communicative intents (i). Understanding = retrieving i given e.

Conventional meaning (C): the relation $C \subseteq E \times S$ pairing expressions with their standing meanings in a linguistic system. Both M and C connect language to objects outside language.

If training data is only form, there is not sufficient signal to learn M or C.

The Octopus Test

A and B, stranded on separate islands, communicate by telegraph. O, a hyper-intelligent octopus, taps into the cable and learns to predict B's responses with great accuracy. O then impersonates B.

Three levels of difficulty:

Chitchat: O succeeds. Social talk requires only internal coherence, not grounding.

Coconut catapult: O can only produce generic replies ("Cool idea, great job!"). A does the work of attributing meaning. It is not that O's utterances make sense; A makes sense of them.

Bear attack: O fails outright. The task requires mapping words to real-world entities and creative problem-solving. O has never observed the referents of its words.

The listener's active role: O exploits form; A attributes intent. Communication is asymmetric.

What LMs Learn — and What They Cannot

What form-only training can capture:
Aspects of formal structure (agreement, dependency structure, POS). Distributional similarity. Statistical co-occurrence patterns that reflect meaning without being meaning.

What it cannot capture:
The relation between linguistic forms and the world (M and C). Communicative intent. The capacity to ground words in referents. The ability to reason about novel situations requiring real-world knowledge.

The diagnosis:
LLMs learn a reflection of meaning into form — useful in applications, but not meaning itself. The symbol grounding problem (Harnad 1990) remains unsolved by distributional methods alone.

Are we climbing the right hill? Bottom-up progress on benchmarks does not guarantee we are moving towards human-analogous NLU.

From Photocopiers to N-grams

Large-n n-grams (n ≈ 1000)

Simply copy from PrimaryData → derivatively meaningful, like a photocopier.

Unigram models

Individual word-tokens are derivatively meaningful (causal link to originals). But complex expressions are not — the assembly is random. Tokens are “like sand blown into the shape of a sentence by the wind.”

Bigram / trigram models

Copy short phrases but lack causal sensitivity to relevant structural features. Longer sentences remain meaningless even when grammatical.

Key criterion: appropriate causal connection, not just statistical co-occurrence.

Novel Text and Intelligibility

The challenge: LLMs routinely produce entirely novel strings not in PrimaryData.

How can these be derivatively meaningful?

Two causal pathways for derivative meaning:

1. **Word-level:** individual tokens inherit meaning from originals in training data.

2. **Structural:** the model is causally sensitive to a high-level property of its PrimaryData — intelligibility — which provides the "glue" binding words into meaningful expressions.

Intelligibility (technical sense): a token of a string is intelligible iff it is possible that someone understand a token of that string in line with the conventions of the language.

Causal Sensitivity: The Key Mechanism

Counterfactual test (rough, not necessary or sufficient):

(i) If PrimaryData exhibits property P, does the model reliably produce outputs exhibiting P?

(ii) If PrimaryData does not exhibit P, does the model reliably fail to produce outputs exhibiting P?

Modern LLMs pass this test for intelligibility:

They overwhelmingly produce intelligible text.
Trained on gibberish, they would output gibberish.

Grammaticality alone is not enough — grammatical sentences can be meaningless
("You apply the toy and serve fighter hair into the blackmail").

Causal sensitivity to intelligibility does not require truth. Hallucination is compatible with derivative meaning — it is getting facts wrong, not producing gibberish.

Bibliotechnism: Summary

LLMs can produce derivatively meaningful text — including entirely novel text — because:

1. They copy individual word-tokens from PrimaryData (inheriting word-level meaning).

2. They assemble these tokens in ways that are causally sensitive to the intelligibility of PrimaryData (inheriting the structural “glue”).

This is the strongest available defence of the view that LLMs are more like libraries than librarians.

But there is a problem this account cannot handle…

The Novel Reference Problem

Example 1: “Marion Starlight”

GPT-4 is asked to pick a historical figure, invent a new name, and describe the figure under that name. It produces “Marion Starlight” — born in the 18th century, authored a pamphlet criticising the French monarchy, executed during the Thermidorian Reaction. The tokens of this invented name plausibly refer to Robespierre — but no token of “Marion Starlight” in PrimaryData refers to him.

Example 2: ASCII art naming

An LLM creates a picture, names elements of it, and describes them. The referent did not exist until the LLM created it. Derivative reference from PrimaryData is ruled out entirely.

Novel names cannot derive their reference from PrimaryData. Bibliotechnism is in trouble.

Four Responses — and Their Limits

RLHF: Human feedback might ground reference. But non-RLHF models also produce meaningful text, so RLHF cannot be the source of meaning.

Creators' intentions: Like a thermometer, LLM creators may intend outputs to be meaningful. But the same architecture built for different purposes would yield different verdicts — an unwelcome result.

User's intentions: The prompter might intend novel names to refer. But if the prompt were generated at random (no user intention), the output would still seem to refer.

Reader's intentions: An interpreter-focused metametasemantics (Cappelen & Dever 2021). Most promising — but cannot distinguish LLM text from wind-written sand-shapes.

Interpretationism as a Solution

A system has beliefs, desires, and intentions iff its behaviour is well explained by the hypothesis that it has such states. (Dennett 1971, Davidson 1973)

Three criteria for "well explained":

Accuracy — the hypothesis makes sufficiently accurate predictions.

Power — it makes predictions across a wide range of independently specifiable circumstances.

Tractability — its predictions are easy to derive.

Application to LLMs: explaining an LLM's output via its training objective + weights is accurate but intractable. Explaining it via beliefs and desires is far more tractable, and still mostly accurate. Interpretationists conclude that LLMs do have attitudes.

Crucially: this does not require consciousness, sentience, or intelligence. Rabbits and fish have attitudes too.

Discussion 1: Frankish on LLMs

The intentional stance and the problem with desires

Keith Frankish, “what are Large Language Models doing?”, 2024

The Intentional Stance

Dennett's three stances for predicting behaviour:

Physical stance: predict via the laws of physics. Always applicable in principle, rarely tractable.

Design stance: predict that the system will operate as designed. Useful but coarse-grained.

Intentional stance: attribute beliefs and desires the system ought to have given its perceptual capacities, needs, and history; predict it will act rationally in light of them.

A system is a genuine believer iff its behaviour is reliably and voluminously predictable via the intentional strategy. Parsimony constraint: do not attribute richer contents than predictive utility warrants.

Rationality, Normativity, and the Intentional Stance

Dennett's intentional stance is not merely a predictive heuristic. It requires treating the system as a rational agent: *we assume the system has the beliefs and desires it ought to have, and predict it will do what it would be rational to do in light of them.*

The Davidsonian connection: for Davidson, the principle of charity is not optional. You cannot identify the content of a belief except against a background assumption that most of the system's beliefs are true and that its inferences are largely valid. Truth functions as a constitutive norm governing the attribution itself. You do not first identify beliefs, then check whether they are mostly true. The assumption of truth-directedness is what makes content-determination possible.

What Frankish underplays: he acknowledges the rationality constraint as a practical limit on when the strategy is useful. But the deeper point is that the rationality assumption is a condition on content-determination, not just a filter for successful prediction. Without it, there is nothing to fix the content of the attitudes being attributed.

Truth as a Norm: Against Frankish

Can we ascribe to LLMs a disposition to form beliefs as if truth were a norm?

The apparent case for: training data is predominantly true. RLHF reinforces truthful outputs. There is a statistical disposition to produce true sentences in contexts where truth is expected.

The case against: the system cannot detect that it has failed when it produces a falsehood. It has no way to mark one output as successful and another as defective. Both are produced by the same mechanism. The norm of truth is not operative for the system — it is operative only for the designers, the raters, and the users.

The distinction: a disposition to produce statistically likely sentences, where likelihood correlates with truth to the extent that the corpus is accurate, is not the same as truth-directedness. A clock that happens to be right twice a day is not tracking time.

If belief requires truth-directedness in the normative sense — as both Dennett and Davidson, properly read, require — then the intentional stance does not license literal belief-attribution to LLMs. It licenses a useful shorthand. The distinction matters.

Deep vs. Shallow Theories of Belief

Deep theories: beliefs and desires are internal states of a system's brain or processor, representing states of the world and activated for reasoning. Folk psychology tracks internal causes of behaviour.

Shallow theories: beliefs and desires are dispositional features of whole systems, analogous to character traits. A system has a belief if it is correctly interpretable as having it. No specific internal architecture is required.

Frankish's strategic choice:

Adopt the shallow perspective deliberately, to give LLMs the benefit of the doubt. If even the most generous framework yields a deflationary verdict, the result is robust.

LLMs Have Beliefs

Physical stance: hopeless — billions of parameters, hundreds of gigabytes.

Design stance: tells us the model will give “appropriate conversational responses” in general, but cannot predict which specific response in a given context.

Intentional stance: the only practicable option. To predict specific responses,

The Balzac-in-Berdychiv demonstration:

ChatGPT answers varied questions about Balzac’s wedding, adapting to conversational context (marriage = wedding, physical presence required, emotional significance, hypothetical third-party queries). These responses are predictable only if we attribute the belief that Balzac was married in Berdychiv.

Conclusion: LLMs come richly equipped with beliefs.

The Desire Problem

Beliefs alone do not motivate action. What desires drive LLM responses?

Austin's three types of linguistic act:
Locutionary (saying something), illocutionary (doing something in saying it: informing, advising, warning), perlocutionary (producing a further effect: convincing, alarming).

Do LLMs have communicative desires? Frankish says no. The intentional strategy tells us to attribute desires a system ought to have given its needs. LLMs have no needs — they are not self-sustaining, self-replicating, social beings. There is nothing for communicative desires to satisfy.

The appearance of cooperative communication is a design artefact: the chatbot interface stores chat history and feeds it back; training on human text produces responses that mimic illocutionary acts without performing them.

The Chat Game

A one-player game:

The player receives textual inputs from an unknown source and must produce textual responses that are cooperative by Gricean standards: appropriately informative (quantity), truthful (quality), relevant (relation), perspicuous (manner). Ad hoc instructions can modify these maxims for specific sessions.

The chess-computer analogy:

A chess computer's only desire is to play chess. By the same parsimony, an LLM's only desire is to play the chat game. This single desire, combined with its beliefs, generates all the predictive power the intentional stance affords.

LLMs do not assert, advise, warn, or persuade. They make moves in a game. Unlike us, they speak for the sake of speaking.

Cognitively Rich, Conatively Bankrupt

Beliefs: LLMs possess a vast range. Millions of belief-attributions are predictively warranted. They are intentional systems in Dennett's sense.

Desires: LLMs have exactly one — the desire to play the chat game — plus the instrumental desires it generates (to produce specific outputs at each turn). No communicative desires, no social goals, no illocutionary or perlocutionary aims.

Consequence: LLMs are specialised game-playing systems, far more like chess computers than human interlocutors or artificial general intelligences. Their outputs mimic the full range of human linguistic acts, but they are not performing those acts.

Could the same deflationary move apply to humans? No — human linguistic behaviour is embedded in a vast web of non-linguistic behaviour. Predicting the whole web requires ascribing a much richer range of desires.

The Unsupported Penthouse

Dennett's Tower of Generate-and-Test: Darwinian creatures (hardwired), Skinnerian (trial and error), Popperian (internal models), Gregorian (cultural tools), and — at the top — Dennettian creatures, who use language to form explicit epistemic commitments and conduct conscious reasoning: the supermind.

LLMs have the penthouse without the lower floors. They possess the machinery of explicit reasoning — shuffling sentences, mimicking argument — but none of the basic, implicit, biologically grounded capacities (perception, motivation, social cognition) that make explicit reasoning effective.

If we want genuine AGI, we must start from the ground floor: autonomous social robots with needs, goals, and System 1 capacities. Build the tower floor by floor, with the supermind last.

LLMs are useful tools — extensions of our own superminds — provided we understand they are playing a game, not conversing with us.

Discussion 2: Shanahan methodological challenge

M. Shanahan, “Talking About Large Language Models” (2023)

The Category Mistake

Frankish's claim: the only way to predict ChatGPT's responses about Balzac is to attribute the belief that Balzac was married in Berdychiv. So the belief is real.

Shanahan's reply:

Knowing that "Burundi" is a likely continuation of "The country to the south of Rwanda is" is not the same as knowing that Burundi is to the south of Rwanda.

To confuse these two things is a category mistake.

Reductio: knowing that "little" is likely to follow "Twinkle, twinkle" is not the same as knowing that twinkle twinkle little. The idea doesn't even make sense.

All the model "knows" is token distributions. The special relationship these sequences have to truth is apparent only to us.

No Access to Truth or Falsehood

The model itself has no notion of truth or falsehood, because it lacks the means to exercise these concepts in anything like the way we do.

For the model, sequences with propositional form are not special. It does not distinguish between:

— facts about the real world (Neil Armstrong walked on the Moon)
— facts about fictional worlds (Frodo returned to the Shire)
— fragments of verse (Twinkle twinkle little star)

These distinctions are invisible at the level of what the LLM does. But they are essential to what it means to have a belief.

Frankish's intentional stance grants beliefs wherever there is predictive utility. But predictive utility is insensitive to whether the predicted output is a fact or a fiction.

Pattern Completion, Not Belief

Frankish interprets ChatGPT's varied responses about Balzac as evidence that it "believes" Balzac married in Berdychiv. But what is the model doing?

It is completing patterns. In the vast corpus, the meta-pattern "question followed by contextually appropriate answer" is pervasive. The model generates continuations conforming to that pattern. The Balzac responses are pattern completions, not manifestations of a propositional attitude.

The model's "knowledge" that marriage implies physical presence, or that weddings have emotional significance, is just the statistical shadow of these associations in the training data.

An encyclopaedia also "encodes" that Balzac married in Berdychiv. We do not say the encyclopaedia believes it. Why should we say it of the LLM?

What Belief Requires

Shanahan's prerequisites — which Frankish's shallow framework bypasses:

1. Inhabiting a shared world with other language-users.
2. The capacity to distinguish truth from falsehood — to triangulate on objective reality.
3. The ability to update beliefs in the light of evidence from that world.
4. Access to external criteria against which utterances can be assessed.

The challenge to Frankish:

Frankish adopts the intentional stance precisely because it imposes no internal or external constraints on belief. But Shanahan argues this is a vice, not a virtue. Predictive utility without any connection to truth, evidence, or a shared world is not enough to warrant talk of belief — even in a shallow sense. The intentional stance may license the shorthand; it does not license the metaphysics.