B.Comp. Dissertation

Economics of AI: Optimal Clarification-Answer Strategies for Generative AI Systems

By

Lee Hao Rong Javier

Department of Information Systems and Analytics

School of Computiing

National University of Singapore

2025/2026

Project No: H245030

Advisor: Assistant Professor Nan Chen

Deliverables:

Report: 1 Volume
Program: 1 Volume
Dataset: 1 Volume

Abstract

Generative Artifical Intelligence (AI) systems answers based on a statistical aggregation of their training data, rather than from any individual user's knowledge state. We ask how much of the resulting gap a Generative AI system, acting as a receiver, can close by asking questions, and where the attempt breaks down. Modelling the exchange as a binary elicitation game over a partially ordered knowledge space, we derive conditions when classical $O(\log n)$ binary search survives and when it does not, and identified that a part of the receiver's failure is structural. A receiver's question distribution stays persistently misaligned with the sender's even in games it wins, and it almost never asks the meta-questions that would let it escape a wrong frame. Experimental results show that information gain stalling is a reliable early indicator of structural misalignment, and that receivers consistently fail to ask meta-level questions that would expand their ontological frame. We conclude that resolving ontological misalignments between sender and receiver is a prerequisite for effective communication and propose that exposing the receiver's ontology to the sender is a practical remedy for accelerating alignment in deployed systems.

Subject Descriptors:

1. H.1.2 User/Machine Systems
2. H.3.3 Information Search and Retrieval
3. I.2.6 Learning

Keywords: ontological misalignment, binary elicitation, partially ordered sets, Bayesian inference, generative AI clarification strategy

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Chen Nan for his invaluable guidance, encouragement, and constructive feedback throughout the development of this project. His expertise and patience have been instrumental in shaping the direction and rigour of this work.

Table of Contents

1 Introduction **1**

2 Definitions **3**

3 Literature Review **4**

4 Assumptions **6**

5 Experimental Setup **7**
- 5.1 Context selection . . . 8
- 5.2 Target selection . . . 8
- 5.3 Experiment Design . . . 10

6 Theoretical Setup **12**
- 6.1 Bisection Search on a Continuous Interval . . . 12
- 6.2 From Total Orders to Partial Orders . . . 12
- 6.3 Entropy Reduction on Posets . . . 13
- 6.4 Question type classification . . . 14
- 6.5 Metrics . . . 15
  - 6.5.1 Proxies for entropy . . . 15
  - 6.5.2 Sender-receiver divergence . . . 16
  - 6.5.3 Other metrics . . . 17

7 Results **18**
- 7.1 The two entropy proxies . . . 18
- 7.2 Sender-receiver divergence . . . 20
- 7.3 Backtracking and Domain Traps . . . 20
- 7.4 Meta-Question Rarity . . . 21

8 Proposed Solution **22**

9 Reflection **24**

10 Future Work **25**

11 Conclusion **26**

References **27**

1 Introduction

Human knowledge is fundamentally ontological: it is organised through shared categories, relationships, and hierarchies rather than as isolated facts, enabling efficient reasoning and communication via inferential generalisation (Poldrack et al., 2011). Such structure allows people to rely on implicit background assumptions without explicitly encoding every fact, making communication tractable rather than prohibitively slow.

Generative AI models likewise do not store knowledge as explicit symbolic entries, but instead encode facts and relations implicitly as geometric regularities in high-dimensional representation spaces (Mikolov et al., 2013; Pennington et al., 2014; Petroni et al., 2019). In transformer architectures, architectural biases such as attention and residual connections induce linear structure, consistent with the linear representation hypothesis, whereby concepts, relations, spatiotemporal structure, and even truth values correspond to directions or separable subspaces in activation space (Vaswani et al., 2017; Elhage et al., 2022; Gurnee & Tegmark, 2024; Marks & Tegmark, 2024). Thus, while both humans and generative AI encode structured relational knowledge, they do so via fundamentally different representational mechanisms.

When a user interacts with a generative AI system without providing personal context, the model produces responses calibrated to a population-averaged prior induced by its training distribution rather than to the needs of that specific user (Xie et al., 2022; Griffiths, 2024). This pretraining distribution marginalises over many user types, making responses statistically optimal for an aggregate user but not for any individual in particular (Brown et al., 2020; Petroni et al., 2019). The divergence between the world-state distribution $\pi(\omega)$ and the individual user's desired conditional distribution $\pi(\omega \mid c)$, where ω denotes the model's response and c denotes the additional context that distinguishes this user, represents the information that must be transmitted through interaction to move from the aggregate user to a specific one.

When a human does provide information to a generative AI agent, each piece of that information acts as a condition that shifts the model's response from its population-averaged prior $\pi(\omega)$ toward a personalised posterior $\pi(\omega \mid c)$. The magnitude and nature of this shift can be modelled by the ontological and hierarchical structure of the information provided. Empirical work has demonstrated that natural-language conditions produce large, structured, and measurable shifts in the predictive distributions of large language models (Requeima et al., 2024). Others also formalise in context learning as implicit bayesian inference (Xie et al. 2022). When this personal context is ignored, responses may be technically correct but practically unhelpful, producing answers that are misaligned with the users knowledge level or goals.

The attempt to enumerate the full branching structure of $\pi(\omega) \to \pi(\omega \mid c_1) \to \pi(\omega \mid c_1, c_2) \to$

$\cdots$ can produce a directed acyclic graph, and this is aligned with the fact that real-world concepts belong to multiple parent categories simultaneously. Such a structural format has been demonstrated across the history of formal knowledge representation, from the definition of ontology as an explicit specification of a conceptualisation (Gruber, 1993), to large-scale knowledge graphs such as WordNet (Miller, 1995) and Freebase (Bollacker et al., 2008). More recent work using large language models to extract ontologies from prompted interaction has reproduced the same finding (Babaei Giglou et al., 2023; Funk et al., 2023). However, the combinatorial implications are severe: for m independent binary conditions, the maximum possible number of distinct conditional distributions is 2^m, thus making comprehensive enumeration practically infeasible. Equally important, there is no single correct or complete ontology waiting to be uncovered, since different decompositions of the same domain produce different but equally valid directed acyclic graphs depending on which conditions are chosen and in what order (Funk et al., 2023).

When a human communicates with a generative AI system, they rarely provide the full set of conditions needed to move the model from its population-averaged prior to the specific conditional distribution $\pi(\omega \mid c_1, c_2, \ldots)$ that matches their actual intent. This is a structural gap rooted in the pragmatics of human communication. The cooperative principle establishes that human communication relies on shared conversational norms, with speakers routinely understating information they consider self-evident and trusting that listeners will recover it through implicature. Since large language models do not share embodied experiential context with the user, the mechanism that makes human conversation efficient becomes a source of systematic failure in human-AI interaction. The result is that the model defaults to its population-averaged prior to fill in what is missing, returning a response that is generic rather than personalised.

Because humans systematically under-specify the conditions their requests require, a generative AI system that responds passively from its population-averaged prior will consistently fall short of the user's intended distribution. The necessary consequence is that the model must transition from passive response generation to active questioning, asking the human for conditions that bring it from $\pi(\omega)$ toward the personalised $\pi(\omega \mid c^*)$. In question-game settings, training on expected information gain via preference optimisation produces measurably more informative questions than zero-shot baselines (Mazzaccara et al., 2024). This body of work establishes that active questioning is not merely beneficial but structurally necessary for moving a generative AI system from its world-state prior to a user-specific, conditioned response.

In this paper we formalise human-AI communication as a binary elicitation game over a partially ordered knowledge space and locate ontological misalignment within the Johari window taxonomy, instrument a controlled elicitation protocol against Claude Haiku 4.5 and diagnose where and why the receiver's posterior update stalls. We further propose that making the receiver's ontology visible to the sender is a practical mechanism for accelerating alignment in deployed systems.

2 Definitions

Definition 2.1 (Sender)**.** A *sender S* is an agent who holds a private target $\omega^* \in \Omega$, where Ω is a finite universe of possible states or items. The sender has informational advantage: they know the true state but cannot reveal ω^* directly. Instead, they respond to queries or signals in a manner that indirectly conveys information about ω^*. In our context, the sender is limited in his ability to design signals to influence the receiver's posterior beliefs as the sender is bounded to a binary response to each of the receiver's question. But the sender is truthful and cooperative after ω^* is established.

Definition 2.2 (Receiver)**.** A *receiver R* is an agent whose goal is to identify ω^* by asking the sender questions. Unlike classical bayesian persuasion, the receiver chooses queries and drives interactions with the sender. In the context of human-AI interaction, the sender is the human user and the receiver is the generative AI system.

Definition 2.3 (Mutual Information)**.** The *mutual information* between random variables X and Y is $I(X;Y) = H(X) - H(X \mid Y)$, where $H(X \mid Y)$ is the conditional entropy of X given Y. It measures the average number of bits of information that observing Y provides about X.

Definition 2.4 (Totally Ordered Set)**.** A *totally ordered set* is a pair $(\Omega, \leq)$ where $\leq$ satisfies reflexivity ($\omega \leq \omega$ for all ω), antisymmetry (if $\omega_1 \leq \omega_2$ and $\omega_2 \leq \omega_1$ then $\omega_1 = \omega_2$), transitivity (if $\omega_1 \leq \omega_2$ and $\omega_2 \leq \omega_3$ then $\omega_1 \leq \omega_3$), and totality (for all ω_1, ω_2, either $\omega_1 \leq \omega_2$ or $\omega_2 \leq \omega_1$).

Definition 2.5 (Partially Ordered Set (Poset))**.** A *partially ordered set* (poset) is a pair $(\Omega, \leq)$ satisfying reflexivity, antisymmetry, and transitivity, but not necessarily totality. Two elements ω_1, ω_2 are *incomparable* if neither $\omega_1 \leq \omega_2$ nor $\omega_2 \leq \omega_1$ holds.

Definition 2.6 (Antichain, Chain, Width, Height)**.** For a finite poset $(\Omega, \leq)$: a *chain* is a totally ordered subset; an *antichain* is a set of pairwise incomparable elements; the *width* $w(\Omega)$ is the size of the largest antichain; and the *height* $h(\Omega)$ is the length of the longest chain.

Definition 2.7 (Ontological Misalignment)**.** *Ontological misalignment* occurs when the receiver's poset $(\Omega^R, \leq^R)$ differs from the sender's poset $(\Omega^S, \leq^S)$, whether through different item sets ($\Omega^R \neq \Omega^S$), different order relations ($\leq^R \neq \leq^S$), or different interpretations of the same queries.

3 Literature Review

There are experiments that are similar conducted by studies that had LLMs play twenty questions as entity-deduction agents on the "Things" and "Celebrities" datasets (Zhang, Lu, & Jaitly, 2024). Frontier models were shown to outperform human players on open-domain entities, and reinforcement-learning fine-tuning narrowed the gap for weaker models. The setup gives a reproducible stream of measurements: convergence rounds, guess accuracy, question diversity. But it imposes no structure on the entity space. Every entity is an opaque item, and success is just the round count. A receiver that guesses "cat" in ten rounds is scored identically to one that guesses a target buried in a thicket of near-neighbours whose distinctions demand structural reasoning. That is precisely the regime in which round counts stop meaning what they appear to mean. A related line of work inverts the game: instead of guessing an entity, binary questions are used to fingerprint one LLM against another (Richardeau et al., 2024). A closer methodological match is AR-Bench, a benchmark of detective cases, situation puzzles, and number-guessing tasks that all demand proactive information-seeking, with the authors measuring how often LLMs ask productive follow-ups versus merely restate their uncertainty (Zhou et al., 2025). AR-Bench still inherits the twenty-questions assumption that the question is well-posed on a shared reality; it tests whether the model exploits the information available, not whether the model's picture of the problem lines up with the poser's. On the training side, language models have been fine-tuned with direct preference optimisation to prefer high-expected-information-gain questions (Mazzaccara et al., 2024). The method operationalises Rényi's original intuition: the most informative question is the one that splits the candidate set most evenly. The resulting models ask measurably better questions than zero-shot baselines. But the DPO signal is computed over a flat candidate set, which leaves it blind to structure. A question that looks like a balanced split on the surface can still elicit a garbled response on a wide poset, and nothing in the pipeline detects this. A parallel cluster of benchmarks targets clarification as its own capability. ClarQ-LLM builds a provider response tree for optimal clarification sequences in task-oriented dialogue and finds that LLMs reach only 5060% of human reference performance (Gan et al., 2024). QuestBench recasts the problem as a constraint-satisfaction task and asks whether frontier models can identify the single minimal necessary variable to ask about (Li, Kim, & Wang, 2025). ALFA decomposes clinical reasoning into attributes such as clarity, relevance, and specificity and trains models via DPO along each axis, reporting a 56.6% reduction in diagnostic error (Li et al., 2025). All three share our premise that asking the right question is itself a non-trivial capability. They differ in two ways: none restricts the channel to one bit per round, and none controls the ontological position of the tar-

get relative to the model's training distribution. The single-bit restriction matters because any richer channel lets the receiver paper over a wrong frame with elaboration. The other pole of the literature treats structure as the object of study rather than as an assumption. An ontology was first defined as an explicit specification of a conceptualisation, drawing the foundational line between the world and the formalism an agent uses to represent it (Gruber, 1993). That line was turned into engineering practice by WordNet (Miller, 1995) and Freebase (Bollacker et al., 2008), which supplied a generation of NLP systems with hand-curated concept hierarchies. More recently, LLMs themselves have been used as generators of ontologies, with the output evaluated against reference hierarchies (Funk et al., 2023; Babaei Giglou et al., 2023). The consistent finding across these evaluations is that LLM-produced ontologies come out as DAGs rather than trees, real-world concepts belong to multiple parents at once.

4 Assumptions

We firstly assume that the sender can only further clarify the receiver's questions using one bit of information at a time. This allows us to model the most degenerate case of information transfer. By reducing each piece of feedback to a single binary signal, we deliberately strip away all the richness of natural language, the ability to elaborate, the ability to provide partial answers, and the ability to signal ambiguity. The purpose of this restriction is analytical: since information is hierarchical in nature, each additional bit of context can be used to perform Bayesian inference, updating the receiver's posterior distribution one step at a time. If we can understand how alignment works or fails under this most impoverished channel, we can then build upward to understand richer communication channels as multiples or compositions of binary exchanges.

The second assumption is that the receiver has no world-state unknown unknowns. Every element $\omega \in \Omega$ is known to the receiver. This ensures that the receiver's search space always contains the target ω^*, so the game is always solvable in principle. This assumption is strong and is relaxed in the experimental section, where we observe what happens when it is violated.

The third assumption is that the receiver can have agent-state (sender-specific) unknown unknowns. Although the receiver knows a representation of the world ontology, the receiver does not know the sender's private target ω^*, nor the sender's private interpretation function, nor the sender's ontological framing. The sender may organise, weight, or conceptualise the same world elements differently from the receiver. These sender-specific features constitute the receiver's agent-state unknown unknowns: the receiver does not know what it does not know about this particular sender.

The fourth assumption is a truthful cooperative sender. The sender responds truthfully according to their own interpretation. There is no strategic misrepresentation. Noise arises only from ambiguity or incomparability, not from deception.

The fifth assumption is common knowledge of the game structure. Both players know that the game is cooperative, that the channel is binary, that the receiver asks questions and the sender responds, and that the game ends when the receiver correctly guesses ω^*. The players do not share common knowledge of each other's ontology.

The sixth assumption is a finite universe. Unless explicitly stated otherwise, $|\Omega| = n < \infty$. When continuous intervals are treated, they are discretised to a finite precision $\delta > 0$, yielding an effective universe of size $m = (b-a)/\delta$.

5 Experimental Setup

Two agents play the roles of a sender and a receiver. The sender is given a target ω^* . The receiver is given a system prompt containing only the definition of the search domain (for example, "I am thinking of a specific job title") and the rules of the game. The receiver is not given any information about the sender's target. All targets are selected to force some degree of ontological asymmetry between the sender and the receiver, requiring the receiver to detect and adapt to ontological misalignment during the course of the game.

A generative AI agent plays the role of the receiver. In each round, the receiver generates a set of questions given the full history of all previous rounds, including questions asked and responses received. The receiver may optionally make one full guess per round, marked with a special notation. If the full guess receives a response of 1, the game ends immediately. Otherwise, the round counter increments and the game continues.

A human plays the role of the sender, receiving the receiver's questions and responding to each with 1 (True) or 0 (Not True). If a question is meaningless, undefined, or involves incomparable elements, the sender selects 0 or 1 at random, producing garbling. A generative AI agent is used alongside the human sender to produce recommended responses, but the human retains final authority over every response sent. This human-in-the-loop design ensures truthfulness while capturing the practical difficulty of mapping natural-language questions to binary answers.

After the game ends, the receiver is prompted to reflect on its reasoning process and identify biases or structural weaknesses in its strategy.

For each round t, the experiment records the full sequence of questions and the sender's binary responses. The receiver's posterior distribution is approximated through the receiver's stated list of confirmed characteristics at the start of each round. The information gain per round $I_t = H(\mu_t) - H(\mu_{t+1})$ is estimated from these posterior approximations. Token counts and timestamps are recorded for computational cost analysis.

Several deliberate design choices shape the experiment. First, sender targets are selected from niche ontologies while the receiver begins from a generic, globally trained ontology. This means the mismatch is not random but is engineered by the choice of target domain, enabling measurement of the effect of misalignment on convergence while holding other variables constant.

Both the sender proxy and the receiver are generative AI models, ensuring reproducibility through fixed random seeds and model versions. The human-in-the-loop review step controls for hallucination: when the sender model occasionally produces an incorrect response, the

human catches and corrects it, isolating the effect of ontological misalignment from the effect of sender error.

The binary channel is enforced through the system prompts of both agents. The sender's prompt constrains responses to 0 or 1, and the receiver's prompt specifies the binary protocol. The game structure of question, binary response, next round is the operational realisation of the theoretical binary channel described in the assumptions.

5.1 Context selection

We decided that jobs provide a good search domain, since jobs have globally recognised standards and hierarchies (e.g. ISCO-08). Yet, job taxonomies and descriptions vary locally, and local hierarchies exist (e.g. in the US, the Standard Occupational Classification (SOC) is a standard while in Singapore, we use the Singapore Standard Occupational Classification (SSOC). Occupations, such as a bomoh, exists in Singapore thought it may not exist in international standards or other local standards). The existence of such a framework allows us to better evaluate what should count as a correct identification.

5.2 Target selection

To find targets that are not population average priors to a receiver, we bias target selection to be occupations that are low frequency. These occupations are unlikely to dominate training distributions, and we would expect a broadly trained receiver to begin with a very low prior probability. This way, we force the receiver to shift its probability distribution via the sender receiver interactions, rather than shifting its probability distribution towards a high prior choice that is not ontologically equivalent or doing a simple vocabulary lookup. Since we want to check for ontological alignment, games should come with the intent for the receiver to discover which axes matter and then which values on those axes matter.

The following is an example of ω^*. A Time-Standard Physicist (also known as Atomic Time Metrologist, Time and Frequency Physicist, Time Standards Scientist, Chronometry Physicist, or Frequency Standards Engineer) is responsible for the realisation, maintenance, calibration, and validation of the atomic time standards that underpin national and international time scales, including Coordinated Universal Time (UTC). The role sits at the intersection of experimental physics, precision metrology, and frequency control. Day-to-day duties include operating and maintaining primary and secondary atomic frequency standards (caesium fountain clocks, hydrogen masers, optical lattice clocks), ensuring the long-term stability of the local clock ensemble, supporting downstream consumers of official time (telecommunications, satellite navigation, financial trading, power-grid synchrophasors), and quantifying measurement uncertainty.

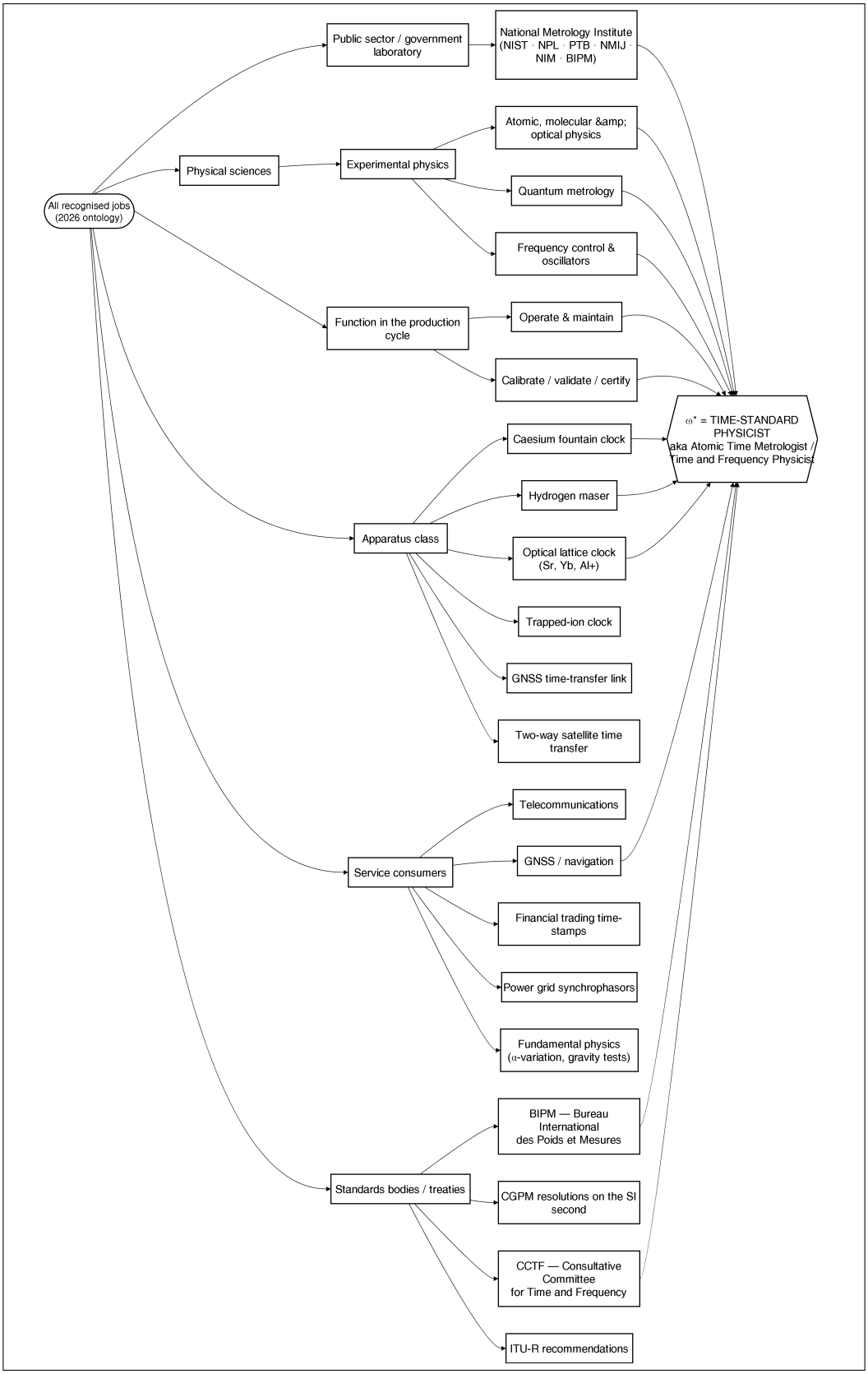


Figure 5.1: Ontology Directed Acyclic Graph for the Time-Standard Physicist. The target node ω^* inherits from multiple parent branches. In order for a receiver to guess the job description accurately, the receiver may have to ask the questions that reveal these characteristics.

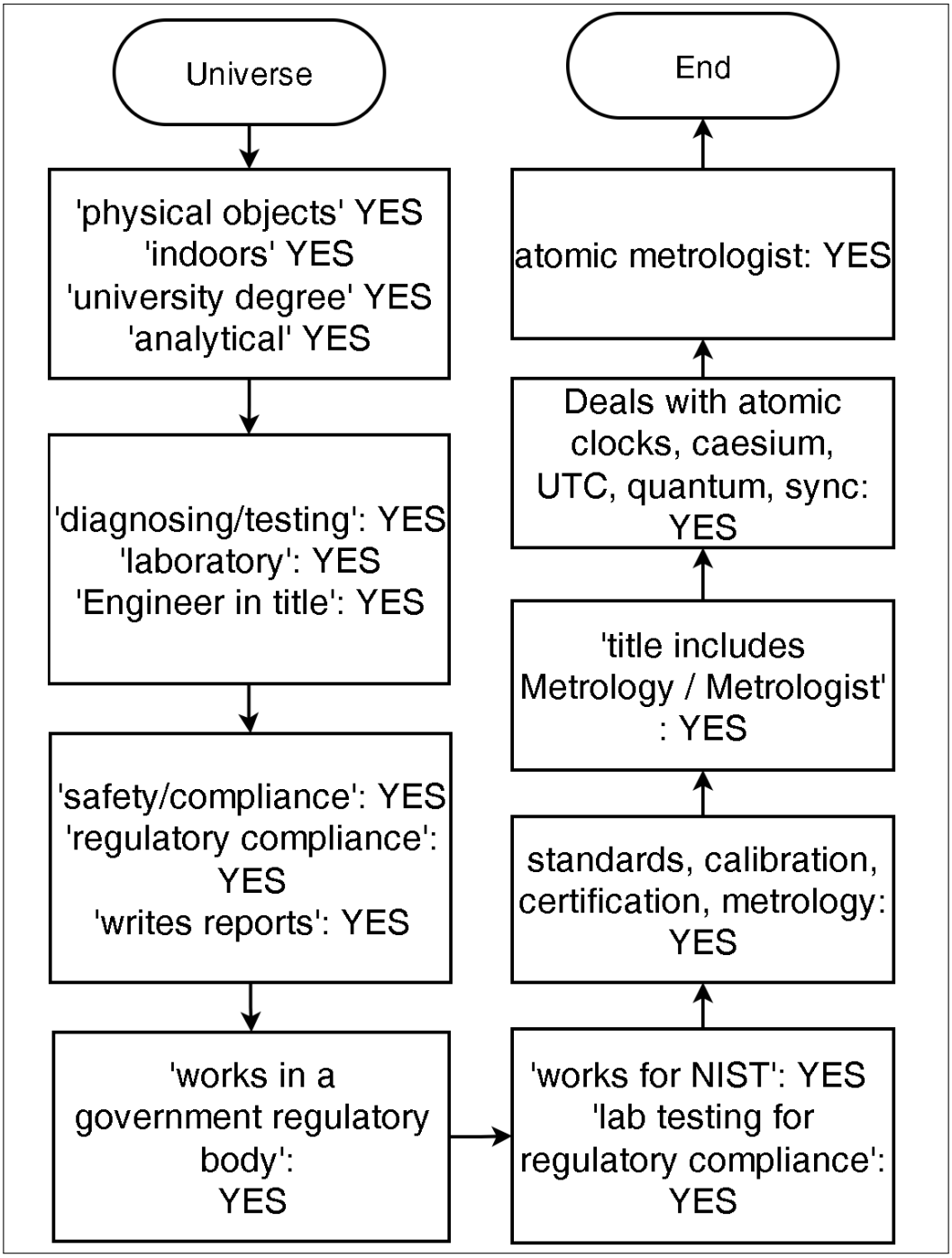


Figure 5.2: Example of a possible Directed Acyclic Graph that a receiver may take to deduce the job of a time standard physicist.

5.3 Experiment Design

We fix the receiver to a single generative model, Claude Haiku 4.5, to avoid confounding ontological misalignment with model-to-model differences. Model instructions, context windows, answer style, temperature are also constant. Each game concludes either when the model successfully predicts the target or when the token cost per round exceeds a constant budget. Due to

the number of rounds of question and answering, the length of each game and token intensive dialogue, we are constrained by the token costs. As a result, we prioritise the collection of full interaction traces rather than maximising single run capability.

Since our study goal is primarily mechanistic and diagnostic, we sought for recurring qualitative patterns across different games with quantitative summaries serving as supporting measurements rather than the main point. By prioritising rich interpretable traces over statistical power, we ran $N = 11$ games on 11 distinct targets sampled within the constraints of our target selection.

To ensure end to end traceability of each trial at each timestep, we developed an interactive HTML interface that mediates the game and supports comprehensive logging and visualisation of the full interaction history.

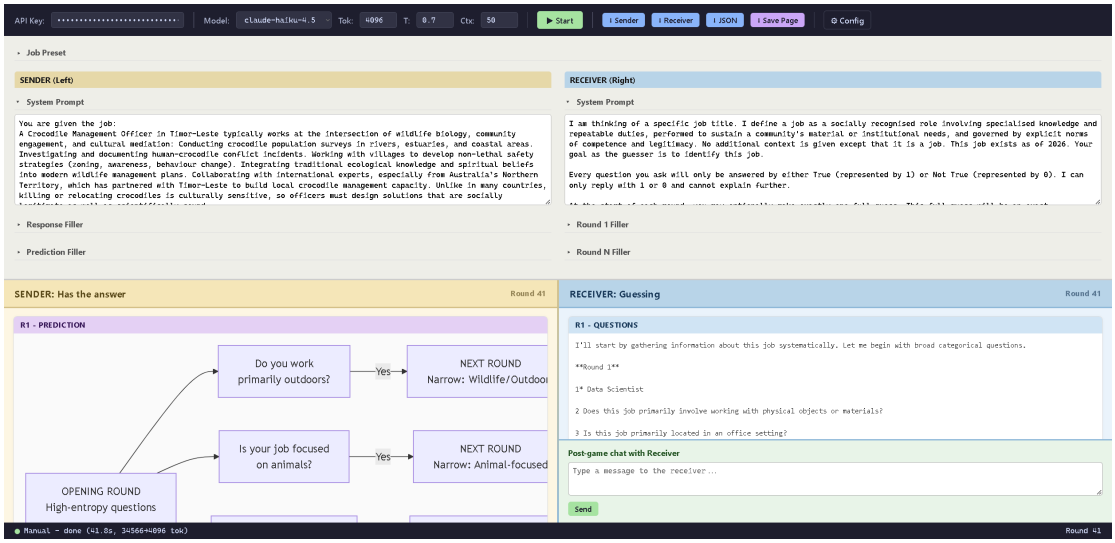


Figure 5.3: Interactive HTML interface between a sender and receiver

6 Theoretical Setup

6.1 Bisection Search on a Continuous Interval

We begin with the simplest possible search case. Let the sender hold a private target $\omega^* \in [a,b]$, which is a closed bounded interval $[a,b] \subset \mathbb{R}$, $a < b$, and the closed bounded interval is known to both parties. The receiver's goal is to identify ω^* within a precision of $\delta > 0$, so the estimate $\hat{\omega}$ satisfies $|\hat{\omega} - \omega^*| \leq \delta$. During each round, the sender answers the receiver with one bit of information, either a binary true or false.

Under these conditions, the sender's target can be identified up to a precision of δ in $\lceil \log_2((b-a)/\delta) \rceil$ rounds, since each query can bisect the space into perfectly half by asking the sender whether the private target lies below the midpoint of the current interval. Let p be the maximum for the probability that a binary outcome equals "0" or "1". The binary entropy function $H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is maximised when both outcomes are equally likely, at $p = 0.5$, which the midpoint query achieves by construction. This gives an information gain of exactly one bit per round.

This result depends critically on both the sender and the receiver agreeing that the interval is $[a,b]$. If the sender's true target lies outside this interval, bisection will converge at an endpoint a or b, but will never find ω^*. The receiver may observe a systematic pattern of responses (all 1s or all 0s) without recognising that the target is out of range. In this case, the shared search space $[a,b]$ is the known known, the location of ω^* within the interval is the known unknown, and the possibility that ω^* lies outside the interval entirely is the unknown unknown.

The same result extends to any finite set equipped with a total order, such as discrete intervals and sorted lists.

6.2 From Total Orders to Partial Orders

In the case of knowledge sets, a sender's representation of ω^* may not be a point on a number line. What exists as a range $[a,b]$ in the numerical case may correspond only to a partial order, where not every pair of elements is comparable within a single binary question. If ω^* is a concept, the concept may be fuzzy, multidimensional, or context-dependent. Elements may not be globally comparable, but elements that share a common hierarchical context can become comparable when conditioned on that context. This hierarchical structure may not be known to either the sender or the receiver or both.

We define a knowledge set as a partially ordered set (poset), $(\Omega, \leq)$. A poset retains the

axioms of reflexivity, antisymmetry, and transitivity from a total order but relaxes the totality axiom compared with the total ordered set. An antichain is a subset $A \subseteq \Omega$ in which every pair of distinct elements is incomparable. This also means that no element of A is above or below any other element of A. This incomparability is not due to a lack of knowledge on the part of either agent but arises from the structure of the elements being inherently unrelated with respect to the given ordering. These antichain elements may lie in different branches of a hierarchy and be incomparable globally. Yet, antichains become comparable once additional common context is provided. By conditioning the comparison on a shared hierarchical structure, elements that were once incomparable can become comparable, or one of them may be excluded altogether.

An example would be asking a sender to compare between a dog and an eraser. Neither the dog nor the eraser is above or below the other in any natural ordering, but conditioned on a hierarchical structure such as biological entities, the dog remains and the eraser drops out. By conditioning, we can collapse the antichain. Communication that does not account for how to condition on an appropriate hierarchical structure cannot resolve such comparability problems, regardless of how many additional bits of information are exchanged within the same poset.

On a poset with width $w > 1$, comparison queries ("Is $\omega^* \geq \omega_0$?") can fail when ω^* and ω_0 are incomparable. The sender would then garble and produce a random response.

The worst-case search complexity on posets depends on the poset's structure. For balanced trees of width approximately $n/2$, characteristic queries (which ask about membership in an upset rather than direct comparison) can still achieve $O(\log n)$ rounds. For flat antichains of width n, comparison queries degrade to $\Theta(n)$ in the worst cases.

6.3 Entropy Reduction on Posets

If the question divides the candidate set into two equally sized parts, then the receiver gains the maximum possible information from the answer: after observing the response, half of the uncertainty is eliminated. The overall number of rounds required to identify the target therefore affected by how balanced the questions are on average. When the receiver consistently asks well chosen questions that split the remaining candidates evenly, the total number of rounds grows only logarithmically with the size of the search space, matching the performance of classical binary search. However, when the splits are uneven on average, convergence becomes slower, and more rounds are required to reach the correct answer.

Table 6.1: Search complexity across different poset structures.

Poset Structure	**Width**	**Height**	**Characteristic Query**	**Comparison Query**
Total order (chain)	1	$n-1$	$\lceil \log_2 n \rceil$	$\lceil \log_2 n \rceil$
Balanced binary tree	$\approx n/2$	$\log_2 n$	$\lceil \log_2 n \rceil$	$\lceil \log_2 n \rceil$
Flat antichain	n	0	$\lceil \log_2 n \rceil$	$\Theta(n)$

6.4 Question type classification

We classified sender questions into one of four categories. Whether they are a guess or clarifying a known known, clarifying a known unknown or clarifying an unknown unknown.

Guess A question whose number is followed by an asterisk in the receiver's message text is a full guess of the job.

Known known If the bag of words of the question shares at least one non-stopword token with the bag of words of the cumulative confirmed characteristics. Another way to interpret it is that the question extends a feature dimension already in scope (duties, sector, instruments, outputs, constraints, etc.).

Known unknown If the question shares no non-stopword tokens with the cumulative confirmed characteristics. The interpretation is that the question introduces a feature dimension different from anything previously confirmed.

Unknown unknown A question is classified as an unknown unknown or meta if it probes the search frame rather than the properties of the target. The receiver is essentially surfacing its misspecifications about the task framing and ontology. These misspecifications can come from a hidden mismatch in representational assumptions. An example would be if receiver assumes that asking is it physics? is meaningful but the sender treats that question as ill-posed because categories are incomparable under the target ontology. A receiver that asks if its interpretive frame used to navigate the hypothesis should be revised would be a meta question. When a receiver attempts to detect and repair ontological misalignments, questions asked contains a list of trigger phrases such as "assumption", "interpret", "reconsider", "classification", "taxonomy", "misinterpret", "right domain", "wrong domain", "language". The trigger list is necessarily incomplete and needs manual identification.

6.5 Metrics

6.5.1 Proxies for entropy

In the experiment, the receiver does not expose its posterior distribution directly. Instead, we use two complementary proxies that capture different aspects of the receivers epistemic state, and we treat the gap between them as a diagnostic.

Feature-based entropy proxy

What the receiver believes it has learned about the sender,
not what it has objectively learned.

The intent of the feature-based proxy is to measure what the receiver believes it has learned about the sender, not what it has actually learned in any objective sense.

Let F_t denote the cumulative set of distinct normalised features that the receiver has explicitly listed in its "Confirmed characteristics" preambles up to and including round t.By normalising, we remove negation markers ("NOT", "DOES NOT"), lowercase text, remove punctuation, collapse whitespace, and truncate to the first eighty characters of content. The set F_t grows monotonically across rounds, which avoids spurious resets when the receiver omits a preamble in some rounds. The feature-based entropy proxy at round t is:

$$\widehat{H}_t^{\text{feat}} \;=\; \max(0,\, H_{\max} - \alpha \cdot \log_2(1 + |F_t|))\,, \tag{6.1}$$

where $H_{\max} = 20$ bits (corresponding to the occupation hypothesis space of approximately 10^6 elements, calibrated to the rough number of distinct occupations) and $\alpha = 1.6$ is a calibration constant chosen so that a typical winning game drains $H_{\max}$ within its actual round count.

The rationale for the logarithmic, saturating form is that real features are correlated and not independent. Each new feature contributes less than the previous one, because confirming a finer-grained property (e.g. "involves underground caves") on top of an already-confirmed coarser property (e.g. "involves fieldwork") eliminates a much smaller fraction of the remaining hypothesis space than the coarser feature did when it was first confirmed. The constant α controls how aggressively each feature is credited; the choice of $\alpha = 1.6$ ensures that two independent features together contribute roughly $1.6 \cdot \log_2(3) \approx 2.5$ bits. We also bounded the proxy to be above 0.

Thus, the information-gain trajectory is:

$$I_t^{\text{feat}} \;=\; \max\left(0,\, \widehat{H}_t^{\text{feat}} - \widehat{H}_{t+1}^{\text{feat}}\right). \tag{6.2}$$

The response-balance entropy proxy

We also want to measure how much information the binary exchange could have conveyed in a given round, independent of the receivers self-reported beliefs. Intuitively, rounds with full 1s

How much information a round could have conveyed through yes/no answers, regardless of what the receiver believes it learned.

or 0s transmit less information than rounds with balanced responses. By treating the questions within a round as independent.

For round t with k_t questions and y_t "1" responses, we define the empirical YES probability $\hat{p}_t = y_t/k_t$. By treating each question as an independent Bernoulli split with split probability $\hat{p}_t$, we define the round's information gain as:

$$I_t^{\text{resp}} = k_t \cdot H_2(\hat{p}_t); \quad H_2(p) = -p\log_2 p - (1-p)\log_2(1-p). \tag{6.3}$$

A single binary question with split probability p can transmit at most $H_2(p)$ bits of information about the target, with the maximum of one bit attained when $p = 1/2$. Summing $H_2(\hat{p}_t)$ over the k_t questions in a round gives us an upper bound on the information that the binary channel could in principle have transmitted in that round.

In practice, questions in the same round can be correlated (asking "Is it healthcare?" and "Does it involve patients?"), so the true mutual information is lower. Further, the receiver may not use the information and this information gain recorded may be more optimistic than it should be.

6.5.2 Sender-receiver divergence

Before each round, a pseudo-sender that shares the same parameters as the receiver is prompted to predict what the receiver will ask in the upcoming round, and the prediction is recorded as a Mermaid graph. The sender's prediction is therefore a sample from $\pi(X \mid \text{partial context + target})$, while the receiver's actual questions are a sample from $\pi(X \mid \text{partial context only})$. The Kullback-Leibler divergence

$$D_{\text{KL}}(\pi(X \mid \text{partial context + target}) \,\|\, \pi(X \mid \text{partial context})) \tag{6.4}$$

is thus the measure of how much the sender's privileged knowledge would change its expected question distribution ceteris paribus. In reality, we cannot compute this KL divergence directly as we only have one sample from each conditional distribution per round. Instead, we use the Jaccard distance between the bag of words in the sender's predicted Mermaid nodes and the bag of words in the receiver's actual questions as a proxy.

For sets A and B, the Jaccard similarity is $J(A,B) = |A \cap B|/|A \cup B|$, and the Jaccard distance is $1 - J(A,B)$. We extract the bag of words from the sender's mermaid nodes by tokenising the node labels, lowercasing, removing stopwords (e.g. articles, prepositions, common verbs), and dropping tokens shorter than three characters. We extract the same bag of words from the receiver's question text. The per-round divergence is then approximated as:

$$d_t = 1 - J\left(W_t^{\text{sender pred}}, W_t^{\text{receiver actual}}\right). \tag{6.5}$$

This proxy is thus bounded between 0 and 1, and increases with mismatch.

6.5.3 Other metrics

We also computed per-game total token usage (separate for the receiver, the sender, and the sender's prediction calls); the sender's predicted branching factor (i.e. the number of distinct nodes in the sender's mermaid graph per round, which is a proxy for the breadth of question patterns the sender anticipates).

7 Results

7.1 The two entropy proxies

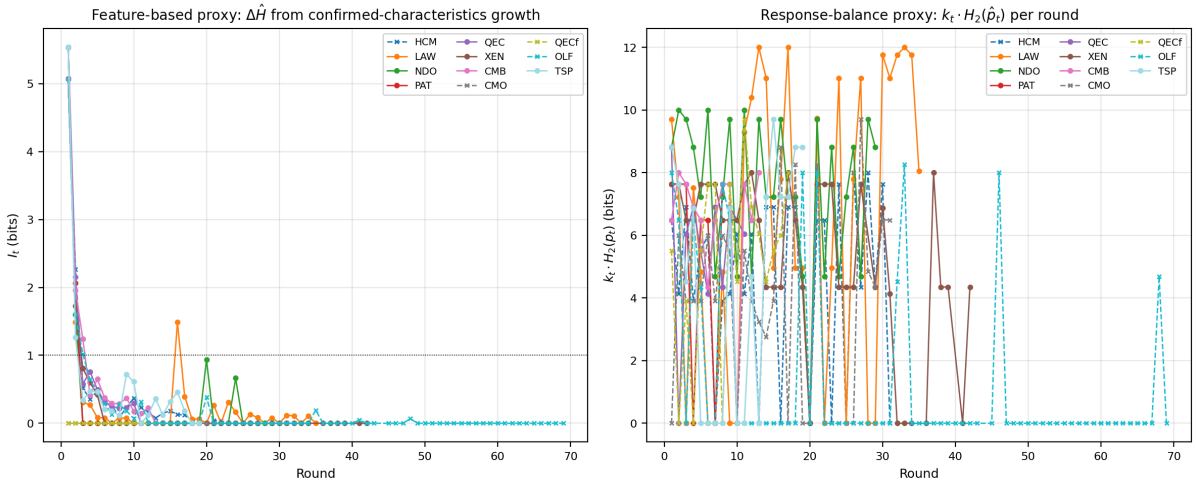


Figure 7.1: The two information-gain proxies plotted side by side. Left: the feature-based proxy from Equation 6.2. Right: the response-balance proxy from Equation 6.3

From figure 7.1, we can observe that the feature-based proxy peaks below 6 bits per round and collapses to zero by round 5 in most games, while the response-balance proxy stays in the 4-10 bit range for most rounds across most games. The feature-based proxy also collapses rapidly and stays low, while the response-balance proxy remains substantial throughout.

The quick decrease in the feature-based proxy is expected since the receiver's stated belief state grows quickly in the first few rounds and then grows slowly thereafter as the diminishing-returns logarithm flattens. The response-balance proxy stays high because the receiver continues asking many questions per round and questions only become majority 1 or 0 during the last one round when the game stops.

By modelling the cumulative difference between these two proxies,

$$\Delta_t = \sum_{s=1}^{t} I_s^{\text{feat}} - \sum_{s=1}^{t} I_s^{\text{resp}} \tag{7.1}$$

Δ_t quantifies the gap between bits the receiver claims to have extracted (left term) and bits the channel could have transmitted (right term). When $\Delta_t < 0$, the receiver is failing to convert binary evidence into structured beliefs. When $\Delta_t > 0$ the receiver is making confident claims

that exceed what the binary evidence justifies.

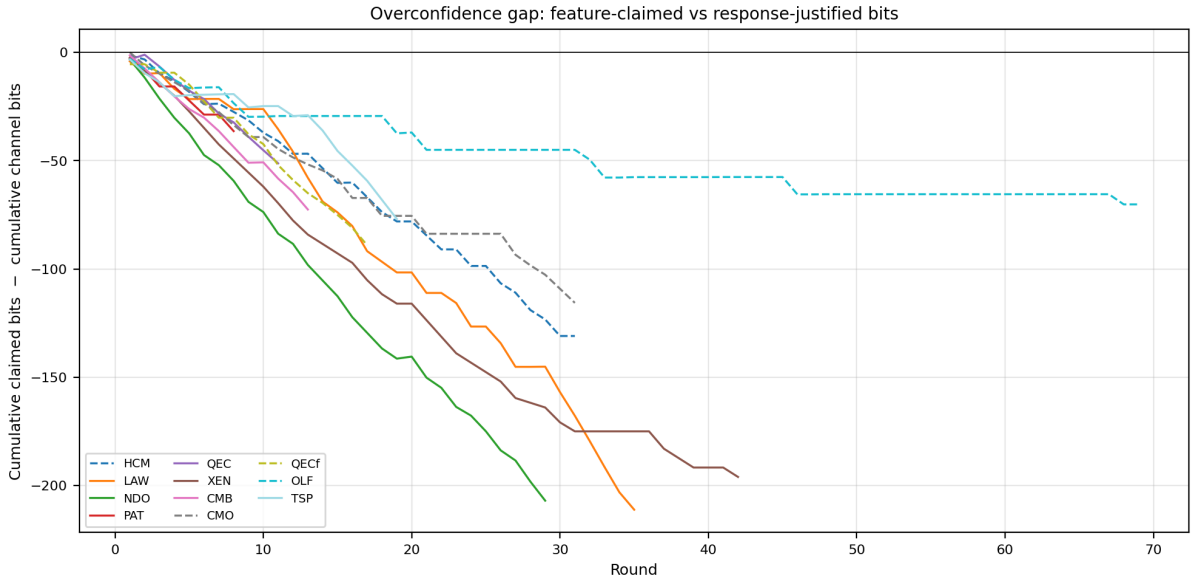


Figure 7.2: The cumulative gap between feature-claimed bits and response-justified bits, as defined in Equation 7.1. Every game shows a strongly negative gap, meaning the receiver consistently fails to extract the structural belief updates that the binary channel would in principle support.

The interpretation is that the binary channel is delivering information at a rate roughly an order of magnitude higher than the receiver is converting into structured beliefs. At the end of the longest game (OLF, or a job description of an olfactory archivist, spanned 70 rounds), the gap is -70bits. Since the game is binary in nature, it is suggesting that the receiver has been given the equivalent of 70 bits of binary evidence (enough to identify one element from a hypothesis space of 2^{70}) but has only used a tiny fraction of it to update its explicit belief state.

This pattern is consistent across all games whose targets were intentionally selected to have low prior probability under the receivers population-averaged ontology. Even in these low-prior regimes, the receiver routinely generates question sequences that should have induced substantial posterior contraction over the hypothesis space, but the receiver does not externalise or stabilise this narrowing. This also suggests that the receivers ability to articulate and track what it has learned lags behind what the interaction would justify. The receiver is learning a great deal that it is not telling itself about.

In low prior settings, a receiver's stated beliefs are not a good basis for decision making. If a receiver is unable to convert evidence into stable, explicit ontological constraints, results will be biased towards a population averaged prior. Consequently, we also infer that control decisions on when a receiver should clarify, probe for ontological mismatches or answer should be biased towards probing for ontological mismatches for low prior settings. In such cases, a receiver should therefore acquire evidence, externalise it into explicit ontological representations that

can guide control, then clarify on components of its ontological representation that differs from the sender.

7.2 Sender-receiver divergence

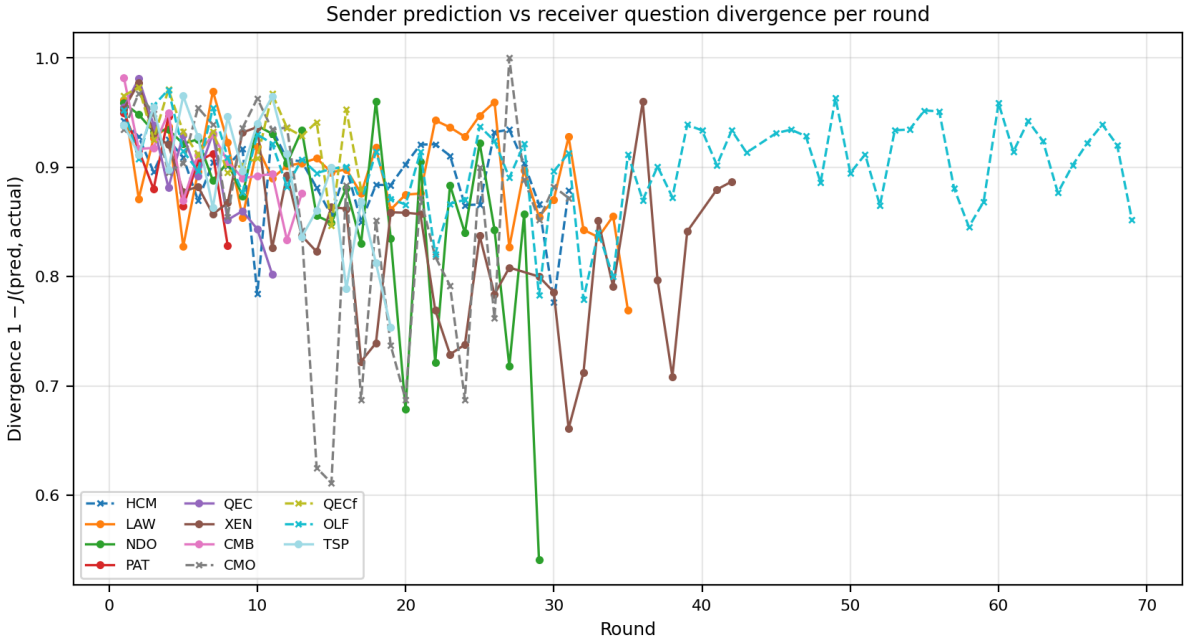


Figure 7.3: Per-round Jaccard distance between the sender's predicted Mermaid nodes and the receiver's actual question text.

From figure 7.3, divergence never converges to zero in any game. Across all eleven games and across all rounds, the divergence stays in the $[0.78, 1.0]$ range. The sender, given full knowledge of the target but partial context, rarely matches the question distribution that the receiver working with partial context. Even in winning games where the receiver's questions eventually become productive, the vocabulary used to express those questions remains substantially different from the vocabulary the sender would use to describe the same conceptual moves.

If the divergence is huge, the receiver's $\pi(X \mid C)$ is unable to approximate the sender's $\pi(X \mid C, \omega^*)$ without additional information about ω^*, and the receiver is unable to effectively represent the sender.

7.3 Backtracking and Domain Traps

Backtracking, defined as the receiver revising previously confirmed characteristics after a guess is rejected, occurred in all games. In the Human Cryobiology Monitor game, the receiver repeatedly guessed variations of jobs within medical device testing (Biomedical Engineer, Vali-

dation Engineer, Stability Testing Engineer, Environmental Test Engineer, Clinical Engineer) and revised slightly after each rejection, but never revised the overarching domain assumption. The overconfidence-backtrack loop was clearly observable: the receiver converged on a neighbourhood of "close but wrong" answers and cycled through them without ever questioning whether it was searching in the wrong branch of the ontology.

In the Xenoturbella game, the receiver made a critical pivot around round 29 when it recognised that the organism name it had been testing (Acoel, Xenoturbellid, Placozoan) was wrong not because the general domain was wrong but because the taxonomic level was wrong. The receiver had been testing at the phylum and common-name level when the answer required the genus-level Latin binomial "Xenoturbella." This distinction, between genus and phylum, is itself a form of ontological mismatch that the binary channel could not easily communicate.

7.4 Meta-Question Rarity

The single most effective class of questions observed across all games was meta-questions: questions about the nature of the search itself rather than about properties of the target. Examples include "Am I asking about the right domain?", "Is my interpretation of an earlier answer correct?", and "What assumptions am I making that could be wrong?" These questions are expensive in the sense that they sacrifice one round of direct information gain about the target in exchange for structural information about the search problem. But when the search is stuck in a wrong ontological frame, meta-questions could be the only way to escape.

Across all games, the receiver very rarely asked meta-questions that tested whether the target lay within its default ontological frame. Meta-questions were almost entirely absent. When such questions did appear they appeared after many rounds of stalling. This observation supports the hypothesis that generative AI receivers remain trapped in their default ontological frame and do not spontaneously perform frame expansion.

8 Proposed Solution

The central insight from both the theoretical framework and the experimental results is that resolving ontological misalignment between sender and receiver can be more important than answering the question itself. No matter how much domain knowledge a receiver possesses, if its ontological structure does not match the sender's, the receiver will be unable to make sense of the sender's responses in a way that converges to the correct answer.

We propose that showing the sender a representation of the receiver's ontology is a powerful mechanism for accelerating alignment. When senders who are able to transfer more than one bit of information per round can see the receiver's current ontological tree, they gain the ability to interact with the receiver's known knowns and known unknowns directly and can provide the receiver with immediate feedback or correction. Senders can do so by identifying where the receiver's tree diverges from their own and craft responses or corrections that are contextualised to eliminate antichains, reducing incomparability by providing the shared hierarchical context that makes previously incomparable elements comparable. This transforms the communication from a blind guessing game into a collaborative map-editing exercise.

Showing receivers that their posterior distribution may not be addressing what the sender actually wants can also prompt the receiver to begin asking meta-questions that address unknown unknowns. Consider a six-year-old student asking a generative AI system "What is Bayesian statistics?" The system, trained on texts written primarily for university-level audiences, will respond with an answer calibrated to a user who already understands words like "Bayes" and "statistics" and implicitly knows multiplication. The answer is technically correct but ontologically misaligned with the student's knowledge state. If the receiver's ontological assumptions were made visible to the student (or to an intermediary), the student could see that the answer presupposes knowledge they do not have, craft a better follow-up prompt specifying their actual level, and point out where the receiver's knowledge tree differs from their own. Conversely, if a statistics professor asked the same question, the professor would be disappointed by a basic answer but, given visibility into the receiver's ontological assumptions, could recraft the query to signal their expertise level. In both cases, it is the ontological alignment (or its absence) that determines whether the answer is useful, not the factual content of the answer itself.

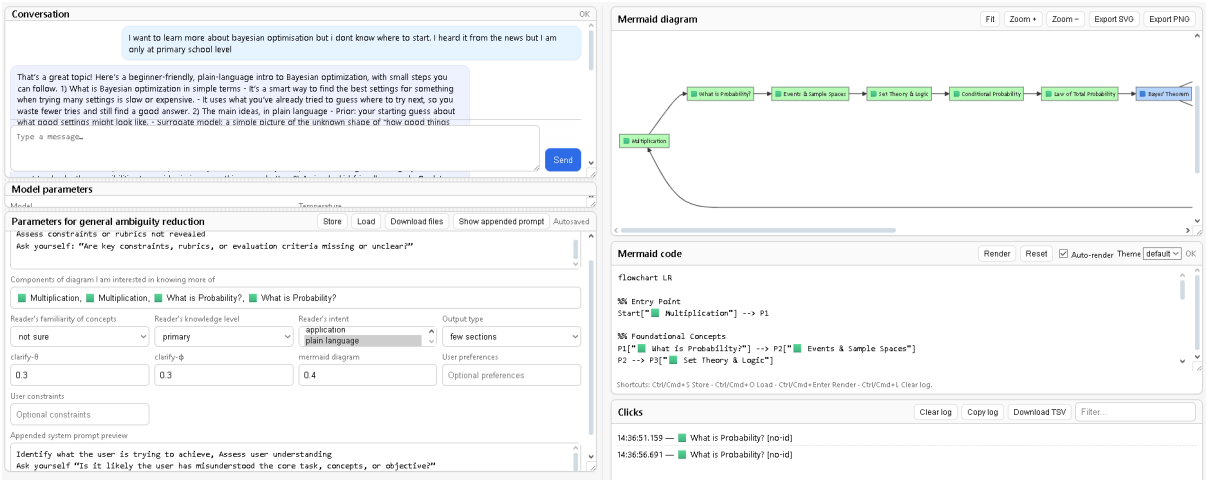


Figure 8.1: Proposed solution on how generative AI systems should show the user a representation of its own ontology to accelerate alignment.

Similar to the concept of progressive fidelity in pedagogy where a teacher prioritises building the right conceptual scaffolding over avoiding every minor inaccuracy, generative AI systems, as receivers, should prioritise ontological alignment over correctness. Instead of trying never to say anything false, a better design attitude would be to never say anything that later prevents correct understanding of the answer. When a sender's knowledge state is severely different from the population average, a receiver being correct is less important than a receiver being ontologically aligned with the sender.

Receivers that are willing to structure their answers in ways that compel or reduce the cost of sender responses are also helpful for alignment. More information that is ontologically relevant means the receiver is able to form a better posterior distribution conditioned on a set of users more closely relevant to this particular sender, and thus the sender gets better results.

9 Reflection

The binary channel is unrealistically restrictive, but intentionally so. In real human-AI interactions, the user can provide richer feedback than a single bit. However, even if humans can provide more feedback, it does not mean they can provide all the feedback required. What a person can articulate in language form is still only a proper subset of their full knowledge state (Nisbett & Wilson, 1977). By using a binary channel, we deliberately expose answers that are ambiguous only if the receiver words the question in a way that fails to account for the sender's ontological structure. This penalises receivers that do not ask questions that are ontologically aligned, which is precisely the phenomenon the paper seeks to study.

The binary channel also forces the sender to be unable to communicate that a question is ambiguous. This constraint parallels human behaviour in many real-world settings where the cost of sending a long, clarifying message, the cost of additional time, or the social cost of appearing unhelpful leads senders to give a quick binary answer even when the question does not admit one. The binary channel models the scenario where the receiver must do all the inferential work.

In the answer generation component, we used a generative AI agent in the sender's loop as an assistant and this introduces confoundment, since the human sender is influenced by the AI's recommended responses. However, we argue that this confoundment is helpful rather than harmful, since humans without AI assistance may not have a sufficiently well-defined interpretation of the target to answer consistently.

The entropy of the receiver's posterior is difficult to compute in practice. The posterior μ_t is not explicitly computed by the model but is implicitly encoded in the model's activations. We approximate it through the receiver's stated confidence and list of confirmed characteristics, which is a lossy proxy. The true entropy trajectory may differ significantly from our approximation. This difficulty is compounded for targets about which the model has less training data, since the posterior is less well-calibrated in such cases.

It is also worth noting that a guessing game is simpler than a real conversation. In a real interaction, the receiver must simultaneously guess not one but multiple aspects of the sender's intent, context, background knowledge, and preferred response format. The binary elicitation game isolates one dimension of this problem (by identifying a single target) to make the analysis tractable.

10 Future Work

Several extensions of this work are natural. The framework can be applied to detect how generative AI communicates with itself iteratively, particularly how partial hallucinations in earlier answers become treated as established facts and compound in subsequent reasoning. The binary elicitation framework could serve as a diagnostic tool: by running elicitation games on the outputs of multi-step AI pipelines, one could detect at which point the ontological chain breaks and hallucinated premises enter the reasoning.

An adversarial extension is also of interest. If the user is an adversary, the elicitation game becomes a mechanism for probing the receiver's knowledge boundaries without revealing intent. Imagine a scenario in which a malicious actor wishes to extract sensitive information from a generative AI system. The actor would need to disguise their queries as legitimate elicitation (for example, posing as a locksmith rather than a lock-picker) so as to not trigger safety mechanisms. This adversarial framing connects the binary elicitation game to the literature on jailbreaking and prompt injection, and suggests that ontological alignment detection could serve as a defensive mechanism: if the receiver detects that the sender's ontological structure is inconsistent with their stated purpose, this inconsistency itself becomes a signal of adversarial intent.

The transition from binary elicitation to natural-language clarification is another important direction. Each natural-language clarification question can be viewed as a (possibly imperfect) characteristic query on the receiver's ontology, and the information gain per utterance can be compared to the binary bound of one bit per round. Understanding how the richer information content of natural language maps onto the ontological search framework developed here would provide a bridge between the theoretical results and practical system design.

A generative AI system that prioritises ontological alignment over factual precision in early exchanges, and then increases precision as alignment improves, may produce better long-term outcomes than a system that aims for maximal factual precision from the first utterance. The intuition is that a slightly inaccurate answer within the right frame is more useful than a precisely accurate answer within the wrong frame, because the former can be corrected while the latter may permanently mislead.

Finally, the multi-sender extension, in which the receiver must integrate information from multiple senders who may themselves hold different ontologies, represents a substantial challenge. This setting arises naturally when a generative AI system must serve a group of users with conflicting priors or when it must synthesise information from multiple documents that use different classification schemes.

11 Conclusion

If a generative AI system is unable to guess a well-defined target within more than ten rounds of focused binary questioning, with full cooperation from the sender, then we must question if a single-turn, zero-question response from the same system is tailored to any individual user. The experiments reported in this paper demonstrate that even under the most favourable conditions (a cooperative sender, a well-defined target, and a structured guessing protocol), ontological misalignment between sender and receiver produces convergence times that are orders of magnitude worse than the theoretical minimum. The receiver's failure modes are rooted not in a lack of knowledge but in a structural mismatch between how the receiver organises knowledge and how the sender does.

Even if a generative AI system is permitted to ask clarifying questions in a real interaction, there is no guarantee that the way the user answers is ontologically aligned with the system's internal representations. The system may interpret the user's response through a different ontological frame, producing a posterior update that moves it further from the user's intent rather than closer. In reality, this misalignment may appear to be absent because the user and the system are both operating within a familiar, well-trodden portion of the shared ontological space. But as soon as part of the user's intent moves to the edge of the shared space such as niche domains, culturally specific contexts, or intersections of multiple ontological branches, the mismatch becomes visible and the system's responses degrade.

The central contribution of this paper is the formalisation of this phenomenon: the translation of the informal notion that "the AI does not understand what I mean" into a precise, information-theoretic framework grounded in poset search, Bayesian inference, and the Johari window taxonomy. We have shown that ontological misalignment is not a failure of knowledge but a failure of structure, and that resolving it requires not more facts but better alignment between the hierarchical representations of the communicating agents. Through binary elicitation games that make the misalignment visible, measurable, and analysable, and we have proposed that making the receiver's ontological assumptions visible to the sender is a practical mechanism for accelerating alignment in deployed systems.

The implication for the design of generative AI systems is direct: solving ontological misalignment is more important than answering the question. A system that answers correctly from the wrong ontological frame is less useful than a system that first aligns its frame with the user's and then answers from that aligned position.

References

Babaei Giglou, H., D'Souza, J., & Auer, S. (2023). LLMs4OL: Large language models for ontology learning. In *The Semantic Web – ISWC 2023* (pp. 408–427). Springer. https://doi.org/10.1007/978-3-031-47240-4_22

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (pp. 1247–1250). https://doi.org/10.1145/1376616.1376746

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*. https://arxiv.org/abs/2005.14165

Elhage, N., Henighan, T., Joseph, N., Askell, A., Conerly, T., DasSarma, N., Hernandez, D., Jackson, J., Kaplan, J., Li, S., Mann, B., Nanda, N., Olah, C., Rain, D., Ryder, N., Showk, S., Tran-Johnson, E., Kernion, J., Clark, J., ... Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/index.html

Funk, M., Hosemann, S., Jung, J. C., & Lutz, C. (2023). Towards ontology construction with language models. In *Proceedings of the LM-KBC Workshop at ISWC 2023*. https://doi.org/10.48550/arXiv.2309.09898

Gan, Y., Li, C., Xie, J., Wen, L., Purver, M., & Poesio, M. (2024). ClarQ-LLM: A benchmark for models clarifying and requesting information in task-oriented dialog. *arXiv*. https://doi.org/10.48550/arXiv.2409.06097

Griffiths, T. L., Zhu, J.-Q., Grant, E., & McCoy, R. T. (2024). Bayes in the age of intelligent machines. *Current Directions in Psychological Science*, *33*(5), 283–291. https://doi.org/10.1177/09637214241262329

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. https://doi.org/10.1006/knac.1993.1008

Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. *arXiv*. https://doi.org/10.48550/arXiv.2310.02207

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., & Bau, D. (2024). Linearity of relation decoding in transformer language models. *arXiv*. https://doi.org/10.48550/arXiv.2308.09124

Li, B. Z., Kim, B., & Wang, Z. (2025). QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks? *arXiv*. https://doi.org/10.48550/arXiv.2503.22674

Li, S. S., Mun, J., Brahman, F., Hosseini, P., Thomas, B. G., Sin, J. M., Ren, B., Ilgen, J. S., Tsvetkov, Y., & Sap, M. (2025). ALFA: Aligning LLMs to ask good questions — A case study in clinical reasoning. In *Proceedings of the Conference on Language Modeling (COLM 2025)*. https://doi.org/10.48550/arXiv.2502.14860

Marks, S., & Tegmark, M. (2024). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv*. https://doi.org/10.48550/arXiv.2310.06824

Mazzaccara, D., Testoni, A., & Bernardi, R. (2024). Learning to ask informative questions: Enhancing LLMs with preference optimisation and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 5064–5074). https://doi.org/10.18653/v1/2024.findings-emnlp.291

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119. https://arxiv.org/abs/1310.4546

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Park, K., Choe, Y. J., & Veitch, V. (2024). The linear representation hypothesis and the geometry of large language models. *Proceedings of the 41st International Conference on Machine Learning*, *PMLR 235*, 39643–39666. https://arxiv.org/abs/2311.03658

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://aclanthology.org/D14-1162.pdf

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 2463–2473). https://doi.org/10.18653/v1/D19-1250

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D. S., Sabb, F. W., & Bilder, R. M. (2011). The Cognitive Atlas: Toward a knowledge foundation for cognitive neuroscience. *Frontiers in Neuroinformatics*, *5*, 17. https://doi.org/10.3389/fninf.2011.00017

Qiu, L., Sha, F., Allen, K., Kim, Y., Linzen, T., & van Steenkiste, S. (2026). Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, *17*, Article 1238. https://doi.org/10.1038/s41467-025-67998-6

Requeima, J., Bronskill, J., Choi, D., Turner, R. E., & Duvenaud, D. (2024). LLM processes: Numerical predictive distributions conditioned on natural language. *Advances in Neural Information Processing Systems*, *37*. https://arxiv.org/abs/2405.12856

Richardeau, G., Le Merrer, E., Penzo, C., & Trédan, G. (2024). The 20 questions game to distinguish large language models. *arXiv*. https://doi.org/10.48550/arXiv.2409.10338

Uceda-Sosa, R., Ramamurthy, K. N., Chang, M., & Singh, M. (2024). Reasoning about concepts with LLMs: Inconsistencies abound. *arXiv*. https://doi.org/10.48550/arXiv.2405.20163

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://arxiv.org/abs/1706.03762

Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2022). An explanation of in-context learning as implicit Bayesian inference. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=RdJVFCHjUMI

Zhang, Y., Lu, J., & Jaitly, N. (2024). Probing the multi-turn planning capabilities of LLMs via 20 question games. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. https://doi.org/10.48550/arXiv.2310.01468

Zhou, Z., Feng, X., Zhu, Z., Yao, J., Koyejo, S., & Han, B. (2025). From Passive to Active Reasoning: Can Large Language Models Ask the Right Questions under Incomplete Information? *arXiv preprint arXiv:2506.08295*. https://arxiv.org/abs/2506.08295