Loose LIPS Sink Ships: Asking Questions in *Battleship* with Language-Informed Program Sampling

Gabriel Grand[1,2] **Valerio Pepe**[2,3] **Jacob Andreas**[1] **Joshua B. Tenenbaum**[1,2]
[1]MIT CSAIL [2]MIT BCS [3]Harvard SEAS

Abstract

Questions combine our mastery of language with our remarkable facility for reasoning about uncertainty. How do people navigate vast hypothesis spaces to pose informative questions given limited cognitive resources? We study these tradeoffs in a classic grounded question-asking task based on the board game *Battleship*. Our language-informed program sampling (LIPS) model uses large language models (LLMs) to generate natural language questions, translate them into symbolic programs, and evaluate their expected information gain. We find that with a surprisingly modest resource budget, this simple Monte Carlo optimization strategy yields informative questions that mirror human performance across varied *Battleship* board scenarios. In contrast, LLM-only baselines struggle to ground questions in the board state; notably, GPT-4V provides no improvement over non-visual baselines. Our results illustrate how Bayesian models of question-asking can leverage the statistics of language to capture human priors, while highlighting some shortcomings of pure LLMs as grounded reasoners.

Keywords: Question-asking; grounded reasoning; language of thought; resource rationality; Bayesian modeling; LLMs

Introduction

Human beings are question-generating machines. From early childhood, we are driven to ask what, where, when, how and why. But out of all the (infinitely many) grammatically-valid questions we could pose in a given situation, how do we decide which ones to ask? And how do we find good questions efficiently, given such a large search space?

Questions can serve many functions, but a core goal is to gain information: reducing the speaker's uncertainty about the state of the world (Graesser et al., 1993; Hawkins et al., 2015; D. Markant & Gureckis, 2012). Informational value must be *grounded* in a shared speaker-listener environment and is highly context-dependent. For instance, "Are you the guy in the red hat?" is a natural question for Alice to text Bob in a crowded airport—but less so in a face-to-face interaction, or on a day when the local firefighter convention is in town. Asking *informative* questions therefore requires integrating linguistic competence with the ability to represent and reason about possible worlds.

In addition to grounding and context, question-asking is shaped by cognitive resource constraints (Anderson, 1990; Chater & Oaksford, 1999; Lieder & Griffiths, 2019). For instance, we know that both children and adults are "greedy" information-seekers in active learning and may consider only very few hypotheses at a time (Klayman & Ha, 1989; D. B. Markant et al., 2016; Meder et al., 2019; Ruggeri et al., 2016; Vul et al., 2014). Faced with a cognitively demanding search task, people also prefer queries that yield simple answers that are easy to interpret (Cheyette et al., 2023).

In this paper, our goal is to model how people efficiently generate informative questions in a grounded environment, subject to resource constraints. We explore several models in the context of the *Battleship Game* (Rothe et al., 2017, 2018, 2019): an adaptation of the classic board game to an open-ended question-asking task. Rothe et al. cast question-asking as program synthesis, where questions are expressed as symbolic programs in a domain-specific language (DSL). They show that sampling programs and scoring them according to features learned from many human questions can approximate the distribution of questions people ask.

We aim to extend this approach in several ways. Humans express questions via language—not code—and we would like models capable of the same. Nevertheless, symbolic programs are a useful format for expressing and evaluating questions; here, we bridge this gap by modeling meaning-making as the process of *translating* from natural language to a language of thought (LoT). Finally, questions can be cued by the situation; while previous models used the board strictly for top-down utility computation, here, the board also informs question generation in a bottom-up way.

We build on a new approach for modeling language-informed thinking (Wong & Grand et al., 2023) that integrates two powerful computational tools: large language models (LLMs) and probabilistic programs. LLMs allow our models to pose questions in free-form everyday language and translate those questions into symbolic representations. Probabilistic programs formalize the question-asker's world models and support coherent reasoning about the expected informativity of questions. Our work thus contributes to both Bayesian models of cognition and LLM accounts: We highlight how traditional models of active learning over structured hypothesis spaces can be extended to natural language settings, and how LLMs—which are increasingly being tuned to answer queries (e.g., Ouyang et al., 2022)—can also be used to *pose* questions that are coherent and grounded.

Our model (Fig. 1) is formulated as a simple Monte Carlo search that samples k candidate questions stochastically from a prior distribution and estimates their informativity via simulation in an internal world model. LLMs play two distinct roles: (1) as a prior over questions, and (2) as a conditional distribution that maps questions from language into LoT programs. The translation step allows us to symbolically compute the Expected Information Gain (EIG) of candidate questions and choose the highest-value one. By varying k, we can control how much mental computation the model performs before producing a question. We call this overall framework "Langugage-Informed Program Sampling" (LIPS).

Correspondence to gg@mit.edu. Code for this paper is available at: github.com/gabegrand/battleship.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

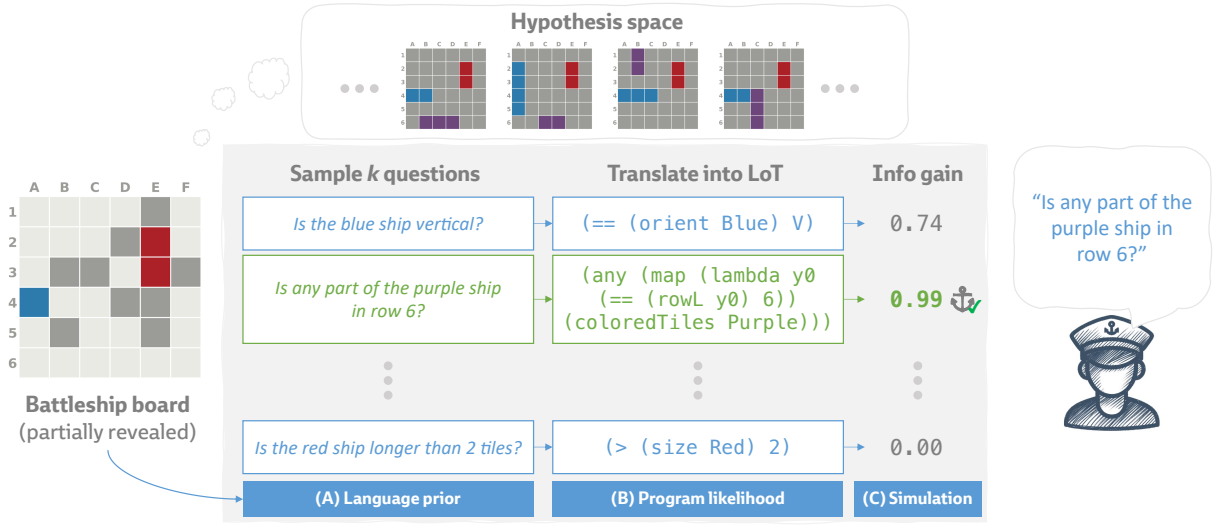


Figure 1: How do people formulate information-seeking questions in a grounded task such as the game *Battleship*? Given a partially-revealed board, our LIPS model (A) samples k questions from a language model prior and (B) translates these into LoT programs. (C) The utility of a question is computed by simulating the program against a hypothesis space of boards compatible with the observation. Here, the best question achieves Expected Information Gain (EIG) of 0.99, meaning the answer would rule out nearly half the boards in the hypothesis space. Our model is well-suited to filtering out samples from a noisy prior that are redundant (e.g., "Is the red ship longer than 2 tiles?") or inconsistent due to lack of grounding.

In our experiments, we compare question priors based on two different LLMs (CodeLlama-7b and GPT-4) as well as a probabilistic context-free grammar (PCFG) hand-engineered for the *Battleship* domain. We find LLM priors yield informative questions that are well-calibrated to human data for surprisingly small values of k. In comparison, the PCFG requires slightly more samples to match mean human performance and yields a higher proportion of unnaturally complex questions. We also explore using LLMs as perceptual pattern learners to propose questions in a bottom-up manner. While a textual encoding of the world state does offer a moderate improvement in efficiency of question-asking, we find that a state-of-the-art multimodal LLM, GPT-4V, provides no improvement over non-visual baselines. Thus, the LLM-based models are also far from complete: they still struggle with grounding, producing many redundant or uninformative questions. In short, our results illustrate how cognitive models of informative question-asking can leverage LLMs to capture human-like priors, while highlighting some of the shortcomings of these models as grounded reasoners.

The Battleship Game

We adopt the *Battleship* task developed by Rothe et al. (2017, 2018), a grid-based environment that evaluates participants' ability to ask goal-directed questions. In this task, participants are presented a partially-revealed board (Fig. 1) and asked to come up with a question that would help to reveal the location of the hidden ships. The task consists of 18 unique 6x6 board contexts, each containing three ships (red, blue, and purple) of varying length (2-4 tiles), orientation (horizontal or vertical), and placement. While later variants extended the paradigm to study multi-turn interactions (Rothe et al., 2019), here we consider the original, single-turn task.

Models

Following prior work, we begin by considering an ideal observer model of a player that starts with a uniform prior $p(s)$ over possible boards consistent with the observed initial state. After asking a question x and receiving an answer y, the player performs a Bayesian update to their belief distribution

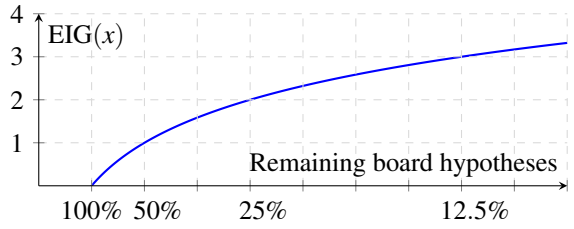
$$p(s \mid y;x) = \frac{p(y \mid s;x)p(s)}{\sum_{s' \in S} p(y \mid s';x)p(s')} \tag{1}$$

where the likelihood $p(y \mid s;x)$ is 1 if y is consistent with s and 0 otherwise. The marginal likelihood can be computed by enumeration or approximated by sampling over a hypothesis space of boards S.

The player's uncertainty about the hidden state of the game board can be measured by the Shannon entropy $H(s)$ (Shannon, 1948), and the value of a question x can be defined as its Expected Information Gain (EIG):

$$\text{EIG}(x) = H(s) - \sum_{y \in Y_x} p(y \mid x) H(s \mid x,y) \tag{2}$$

Intuitively, EIG provides a log-space measure of the number of candidate boards that the player can rule out with question x. For instance, an ideal yes/no question that rules out 50% of possible boards would achieve $\text{EIG}(x) = 1$. (Throughout, we use $\log_2$, so EIG is measured in bits.)



In *Battleship*, questions that admit a large set of possible answers, denoted Y_x, can achieve $\text{EIG}(x) \gg 1$ (e.g., "What is the top-left corner of the red ship?"). However, some answers may be more informative than others; this uncertainty gives rise to the expectation over possible answers in Eq. 2.

In prior work, EIG was considered as one among several heuristic features (complexity, answer type, etc.) in a Boltzmann energy model that was fit to maximize the likeihood of the collected human questions (Rothe et al., 2017). Here we take a complementary approach: instead of fitting our model to human data collected from *Battleship*, we instead aim to sample directly from a distribution of maximally-informative questions—*without* positing the space of features these questions might have. We hypothesize that human-like questions will fall out naturally from a Bayesian model with a very generic prior that is subject to cognitive resource constraints.

We formulate our model as a probabilistic sample-based search with a parameter k that controls the amount of internal computation the model performs. (We are in part inspired by the bounded space model of Ullman et al., 2016 for creative language generation.) Given some proposal distribution over questions, we sample k questions and choose the one that maximizes EIG:

$$\{x_1, \ldots, x_k\} \sim p(x \mid s) \tag{3}$$

$$x^* = \arg\max_{x_i} \text{EIG}(x_i) \tag{4}$$

A central challenge of this approach is choosing a suitable proposal distribution $p(x \mid s)$ that admits efficient sampling. Moreover, as the notation implies, this distribution should ideally be *board-conditional* so as to generate targeted questions about the particular board at hand. To facilitate computation of EIG, it is also desirable to have a proposal distribution that is capable of expressing questions as LoT-like programs that can be deterministically executed against the board following some denotational semantics; i.e., $y = [\![x]\!]_s$. We consider two kinds of question-proposal distribution that allow us to instantiate our LIPS model.

Grammar proposal distribution

We begin by considering a probabilistic context-free grammar(Johnson, 1998) as a proposal distribution over questions. We adopt the grammar of Rothe et al. (2017)[1] whose rules and terminals correspond to key concepts in *Battleship*: ships vary in *color*, *size*, *orientation*, *location*, etc. The grammar also encodes numeric and set-theoretic operations to support comparisons; e.g., "How many of the blue ship's tiles are in column B?"

Answer → Bool | Num. | Color | Orient. | Loc.
Bool → 'T' | 'F' | (and B B) | (touch Ship Ship)...
Num. → 0 | 1 | ... | 9 | (+ N N) | ...
Num. → (size Ship) | (row L) | (col L)...
Color → Ship |
Ship → 'Blue' | 'Red' | 'Purple'
Orient. → 'Horizontal' | 'Vertical' | (orient Ship)
Loc. → 1*A* | 1*B* | ... | 6*F* | (topleft Set)...
Set → (tiles Color) | (∩ Set Set) | 'AllColors'...

As a computational instantiation of a cognitive theory, the PCFG proposal distribution follows in the (probabilistic) language of thought tradition (Fodor, 1975; Goodman & Lassiter, 2015; Goodman et al., 2014). One advantage of this formalization is that it comes with clear-cut denotational semantics: questions are programs that can be executed against the board state to yield an answer. Additionally, the PCFG imparts an inductive bias towards shorter programs that naturally implements Bayesian Occam's Razor (Gelman et al., 1995; Henderson et al., 2010). However, the PCFG also yields a combinatorially-large number of trivial statements (e.g., mathematical propositions that do not make reference to the board). These issues could be addressed by making the PCFG board-conditional—e.g., by training a neural "recognition network" (Ellis et al., 2021) to map board inputs to weights. Nevertheless, achieving a good fit with this model would require collecting a labeled dataset containing many thousands of (s,x) pairs. In the absence of such data, we follow the prior work in treating the PCFG as an unconditional prior with uniform probability over the production rules.

Language model proposal distribution

A recent line of work in probabilistic programming explores using Large Language Models (LLMs) as instantiations of humanlike priors in Bayesian models (Dohan et al., 2022; Ellis, 2023; Lew et al., 2020, 2023). For the purpose of constructing a cognitive model of human question-asking, LLMs represent an attractive proposal distribution for several reasons. First, since they are trained on vast corpora of natural text, LLMs directly encode a prior over plausible questions. Moreover, LLMs are strong in-context learners (Brown et al., 2020) and are increasingly amenable to instruction from the experimenter (Ouyang et al., 2022; Rafailov et al., 2024). Consequently, by constructing an appropriate prompt, we can transform a generic LLM into a proposal distribution over questions in the *Battleship* domain. This approach faces two main challenges, which we detail below.

Grounding generation in the state of the world Ideally, we would like our model to be "stimulus computable" (Yamins & DiCarlo, 2016), accepting the same images and task instructions as a human participant. While multimodal

[1] See Table SI-1 in Rothe et al. (2017) for the full grammar. We omit λ-abstractions, which rarely yield well-formed questions during sampling, and we filter out trivial expressions of depth 1.

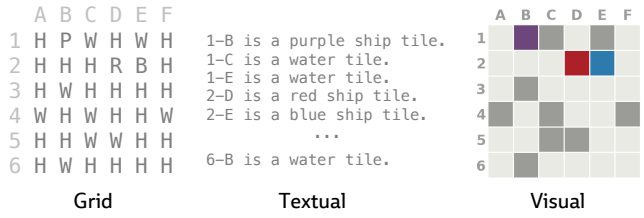


Figure 2: We experiment with 3 different board representations: an ASCII-style grid, a textual serialization, and a visual prompt encoded as an image.

LLMs are growing in popularity and availability (Driess et al., 2023; OpenAI, 2023b), it remains unclear to what extent they are capable of extracting structured visual information—such as a *Battleship* board—into an appropriate computational representation. We experiment with three different types of board representation (Fig. 2) in order to evaluate the degree to which our LLM proposal distributions are able to leverage board-conditional information.

Translating from natural language to the LoT Our LIPS model posits that the question-asker mentally draws and evaluates k samples and chooses the most informative one. This is straightforward in the case of the PCFG, which directly generates programs, but not for the case of LLMs, which output natural language. To address this, we follow the approach of the *Rational Meaning Construction* framework (Wong & Grand et al., 2023), which uses LLMs to implement a "meaning function" that translates from natural language into the LoT. Concretely, we decompose the LLM proposal into separate **linguistic question generation** $p(l \mid s)$ and **language-to-program translation** $p(x \mid l)$ distributions, which we approximate via sampling.

$$p(x \mid s) = \sum_{l} p(x \mid l) p(l \mid s) \tag{5}$$

This formalization admits many possible denotational semantics—$[\![\cdot]\!]_s$ could be implemented by a LISP interpreter, a Python program, or even a LLM. For convenience, we use the same *Battleship* DSL from Rothe et al. (2017), which allows us to take advantage of the fast C++ implementation of the EIG function developed for that work.[2]

Experiment

Participants, materials, and methods

Human data We use the human dataset collected by Rothe et al. 2017, which consists of 26-39 questions for each board composed by a single pool of N=40 participants, for a total of 605 question-board pairs. Participants were not "prompted" with any example questions; they were only given the constraint that the question should admit a single-word answer. As the program annotations in this dataset used an earlier version of the DSL, we manually translated a representative subset of the questions into the latest DSL and used a LLM to annotate the remaining programs.

LLMs We queried GPT-4 (OpenAI, 2023a, 2023b) via API, using `gpt-4-0613` for the textual and grid board formats, and `gpt-4-vision-preview` for the visual format. To compare against a reproducible, open-source LLM, we used CodeLlama (Roziere et al., 2023), a member of the Llama 2 family of models that was finetuned for code generation. We obtained the model weights from HuggingFace (`CodeLlama-7b-hf`) and used the smallest variant of the model, which contains 7B parameters. We performed local inference on a single GPU, taking advantage of the `hfppl` library (Lew et al., 2023) to speed up inference via caching.

Prompting We fed both LLMs identical sets of algorithmically-constructed prompts. For question generation $p(l \mid s)$, each prompt consisted of instructions describing the task setup ("You are playing the board game Battleship. There are three ships on the board..."). In the **zero-shot** condition, the prompt concluded with a target game board (Fig. 2) and text to elicit a question. In the **few-shot** condition, the prompt additionally included 3 example boards, each with 10 questions from the human data. The example boards and questions were sampled without replacement in a leave-one-out manner so as to exclude human data collected for the target board. For translation $p(x \mid l)$, the prompt consisted of a similar task instruction, followed by 12 (l, x) pairs randomly sampled from the human data in the same manner.

Sampling For each LLM condition, we sampled 100 questions/board $\times$ 18 boards. To explore the effects of prompt and board formats, we repeated this process for each combination of {zero-shot, few-shot} $\times$ {textual, grid, visual, no board} using GPT-4(V). For the PCFG, which is not board-conditional, we sampled a single set of 100K questions and computed their EIG values for each board. Following Ullman et al. (2016), to avoid expensive re-collection of data, samples were grouped *post-hoc* into buckets of size k. Since the underlying samples are i.i.d., this provides an unbiased estimate of the true sampler, with the caveat that the effective sample size diminishes with k. Throughout, null hypothesis testing was conducted between conditions using Welch's t-test.

Results and Discussion

Informativity How informative are the questions collected from humans? And to what extent do our models capture the information-seeking quality of human questions? We computed EIG values for all human and model-generated questions (Table 1). Across the 18 boards, the average human question scored EIG $= 1.27$, while the best human question achieved considerably higher EIG $= 3.61$. Despite this large range, *virtually all* (97%) of the human questions were informative (EIG > 0), revealing that participants were highly sensitive to the board state.

In contrast, the underlying proposal distributions ($k = 1$) were substantially noisier than people: questions from CodeLlama and GPT-4 averaged EIG $= 0.65$-0.66, respectively, while questions from the grammar averaged EIG $=$

[2] https://github.com/anselmrothe/EIG

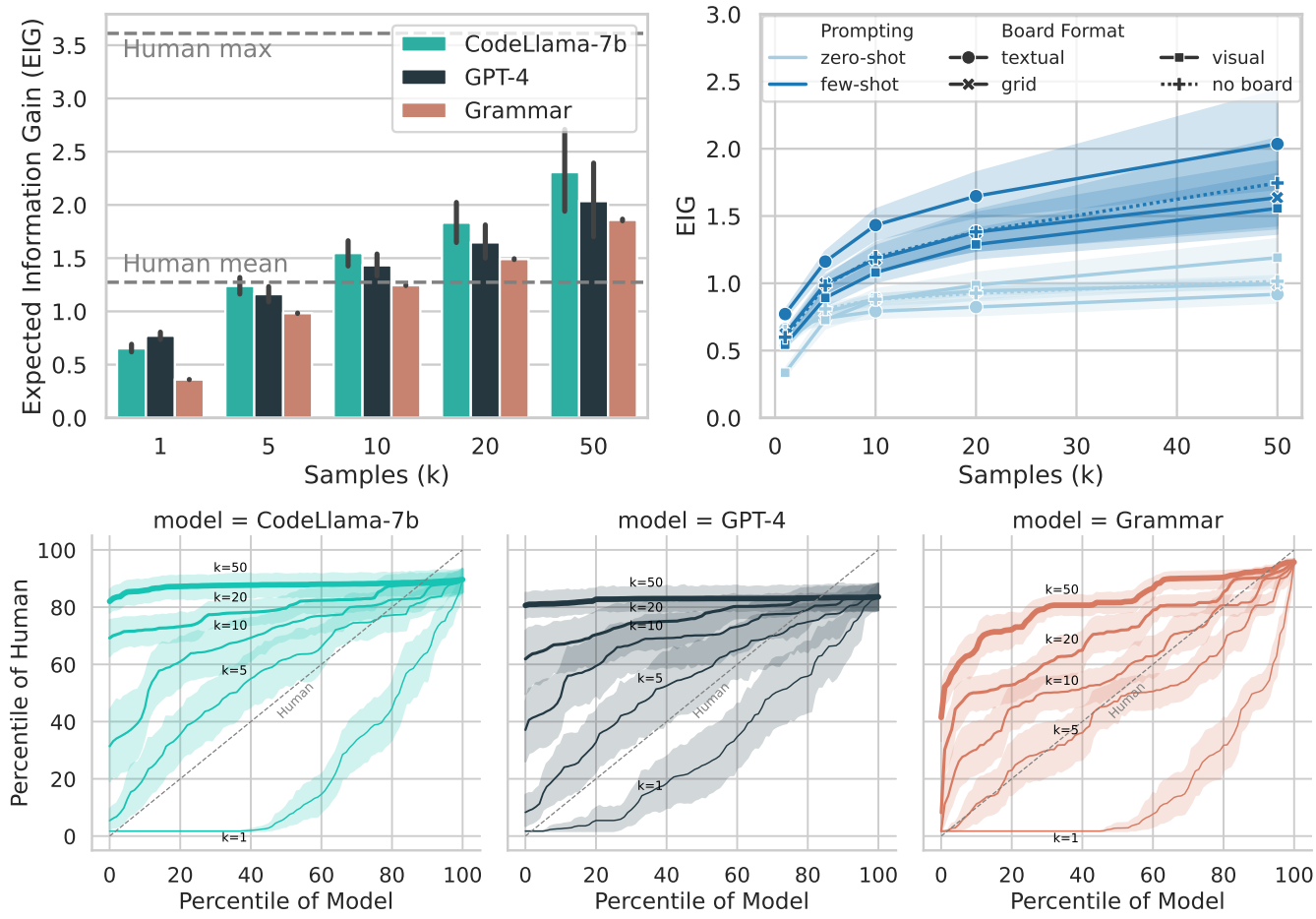


Figure 3: Comparing the informativity of model-generated questions against human data. **(Top left)** LIPS with two LLMs and a hand-engineered grammar as proposal distributions over questions. As k increases, all three models reach mean-human performance, though they fall short of the best human-generated questions. **(Top right)** Evaluating GPT-4's performance with different prompt formats and board representations. Including few-shot examples universally boosts EIG. However, performance varies depending on the board format. Notably, GPT-4(V) was unable to utilize the board's structure in text (grid) or images (visual), implying a failure of grounding. **(Bottom)** Q-Q plots comparing model vs. human EIG values at varying sample sizes. At $k = 5$, all three models are generally well-calibrated to humans, though they fall short of the top 10-20% of human questions. Throughout, error bars and shaded regions indicate 95% bootstrapped confidence intervals. GPT-4 and CodeLlama-7b refer to the few-shot, textual condition unless otherwise noted.

Model	EIG		% Valid		% Informative		Program Depth		Program Size		Question Words	
	μ	σ_M	μ	σ_M	μ	σ_M	μ	σ_M	μ	σ_M	μ	σ_M
Human	1.27	0.04	1.00	0.00	0.97	0.01	3.22	0.07	4.51	0.14	7.12	0.08
Grammar	0.36	0.00	1.00	0.00	0.38	0.00	3.01	0.00	5.13	0.01	–	–
CodeLlama-7b	0.65	0.02	0.75	0.01	0.45	0.01	2.64	0.02	3.24	0.04	6.66	0.04
GPT-4 (few-shot)	0.77	0.02	0.88	0.01	0.59	0.01	2.61	0.02	3.22	0.04	6.23	0.03
GPT-4 (zero-shot)	0.66	0.01	0.40	0.01	0.35	0.01	3.73	0.04	5.04	0.09	5.19	0.02
GPT-4 (no board)	0.60	0.02	0.68	0.01	0.43	0.01	3.08	0.03	4.12	0.07	6.28	0.03

Table 1: Summary statistics of the underlying samples ($k = 1$) across all board contexts. Questions that translated to a parseable program are considered Valid, and those that achieved EIG > 0 are considered Informative. Program Depth and Size refer to the depth and number of nodes of the program abstract syntax tree. Question Words measures the number of words in the natural language question. μ and σ_M denote sample mean and standard error, respectively.

0.36. However, as Fig. 3 (top left) reveals, LIPS allows for a significant boost in performance: with just $k = 5$ samples, both LLMs approached human mean performance; and at $k = 10$, both models significantly outperformed the human mean, with $p < 0.001$ for CodeLlama, and $p = 0.01$ for GPT-4 (textual, few-shot). This trend continues for sample sizes $k = 20$ and $k = 50$, though all models still fall short of the best human-generated questions.

Sample efficiency What represents a cognitively-plausible amount of mental sampling? Fig. 3 (bottom) compares the full distribution of model vs. human EIG values for varying values of k. At $k = 5$, both LLMs were closely calibrated to the human distribution, performing on par with the grammar, which was hand-engineered to capture this distribution. In other words, the Nth percentile of human question-askers wrote questions that were of comparable informativity to the Nth percentile of samples from the model. However, the top human questions (approx. 85-90th percentile) outperformed the top model-generated questions.

Translation fidelity One restriction of our evaluation is that, in order for a question to be considered informative, it needs to be expressable in the *Battleship* DSL. But how effective is the model at translating questions into programs? As Table 1 (% Valid) shows, a high percentage of samples from CodeLlama (75%) and GPT-4 (88%) were successfully translated. Only in the GPT-4 (zero-shot) case did the translation model achieve low fidelity (40%). Since the model does not receive any examples in the zero-shot case, it is not surprising that many of the questions from this distribution were not translatable.

Groundedness To what extent did the LLM-generated questions take the board state into account? Of the valid programs sampled from each model, 40% (CodeLlama) and 33% (GPT-4) were *uninformative* (EIG = 0). This occurs when a question is redundant with respect to information already revealed in the board. (For instance, "Is the red ship vertical?" is uninformative for 3/18 boards in the stimulus set.) The high proportion of uninformative programs highlights a potential failure of grounding. Our evaluation of different board formats, shown in Fig. 3 (top right), provides further evidence of this issue. Of the four board formats (Fig. 2), the "textual" representation was the only one that significantly outperformed the "no board" condition ($p < 0.05$ for k = 1-20). Notably, across k, the "visual" board format performed either significantly worse ($p < 0.05$ for $k = 1, 5$) or was not significantly different than the "no board" condition ($p > 0.05$ for k = 10-50). These results show that that GPT-4V was unable to utilize the board's structure to formulate informative questions relative to a board-agnostic baseline.

Question type What *kinds* of information do humans ask about, and do the models reflect this distribution? As illustrated in Fig. 4, humans ask a diverse range of question types, with a preference for boolean and numeric answers. Owing to its structure, the grammar generates an approximately uniform distribution over types. Meanwhile, both of the few-shot

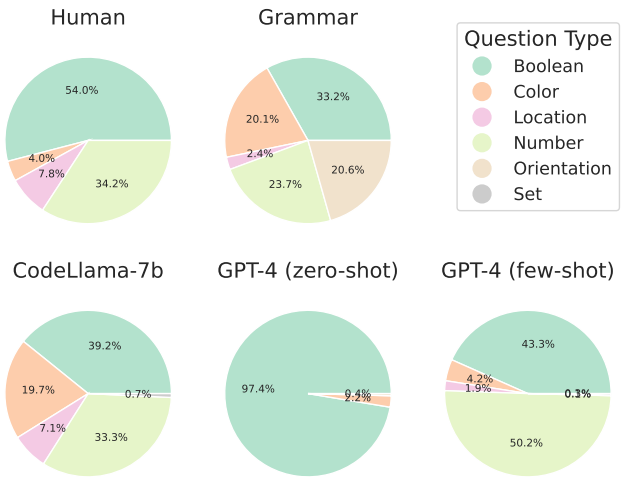


Figure 4: Proportion of top-level question types generated by each proposal distribution at $k = 1$.

prompted LLMs approximate the human distribution, though CodeLlama mirrors it more closely than GPT-4. Without access to examples, GPT-4 (zero-shot) defaults to boolean questions that echo traditional *Battleship* moves; e.g., "Is there a ship at 2-C?" Thus, the different choices of prior encode different inductive biases—and LLMs provide an especially flexible way of encoding both human general knowledge and domain-specific priors into cognitive models.

Conclusion

As more and more people interact with language models on a daily basis, understanding how *humans* seek information through language is a truly important scientific question. In this work, we introduced a new approach to modeling informative question-asking by sampling questions from a noisy LLM prior and translating into programs in a LoT. But where does this LoT come from in the first place? Here, we used an existing DSL as initial step, but our approach could be combined with Bayesian program induction techniques to learn a new DSL from data (Ellis et al., 2021; Grand et al., 2024; Piantadosi et al., 2024; C. Wong et al., 2021). Relaxing our assumptions even further, we might eschew a DSL in favor of a domain-general programming language like Python (Ellis, 2023; Wang et al., 2024).

While Monte Carlo sampling is attractive for its simplicity, there exist more sophisticated inference techniques that offer better sample efficiency. Several such approaches have recently been studied in an AI dialogue context for eliciting user preferences (Li et al., 2023; Piriyakulkij et al., 2023) and clarifying ambiguity (Zhang & Choi, 2023). Applying these inference methods to study people's behavior across longer interactions—such as multi-turn *Battleship*—presents a natural direction for future work.

Acknowledgments

We thank Brenden Lake, Maddy Bowers, Lionel Wong, Sam Cheyette, and Guy Davidson for helpful discussions and feedback on this work.

The authors gratefully acknowledge support from the MIT Quest for Intelligence, the MIT-IBM Watson AI Lab, the Intel Corporation, AFOSR, DARPA, and ONR. GG is supported by the National Science Foundation (NSF) under Grant No. 2141064. JA is supported by NSF Grant IIS-2144855. JBT received support from AFOSR Grant #FA9550-19-1-0269, the MIT-IBM Watson AI Lab, ONR Science of AI and the DARPA Machine Common Sense program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of sponsors.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

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 [arXiv: 2005.14165]. *arXiv:2005.14165 [cs]*.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, *3*(2), 57–65.

Cheyette, S. J., Callaway, F., Bramley, N. R., Nelson, J. D., & Tenenbaum, J. (2023). People seek easily interpretable information. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Dohan, D., Xu, W., Lewkowycz, A., Austin, J., Bieber, D., Lopes, R. G., Wu, Y., Michalewski, H., Saurous, R. A., Sohl-Dickstein, J., et al. (2022). Language model cascades. *arXiv preprint arXiv:2207.10342*.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Ellis, K. (2023). Human-like few-shot learning via bayesian reasoning over natural language. *Thirty-seventh Conference on Neural Information Processing Systems*.

Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Morales, L., Hewitt, L., Cary, L., Solar-Lezama, A., & Tenenbaum, J. B. (2021). Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, 835–850.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *The handbook of contemporary semantic theory, 2nd edition. Wiley-Blackwell*.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *Concepts: New Directions*. MIT Press.

Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In *Psychology of learning and motivation* (pp. 411–436, Vol. 29). Elsevier.

Grand, G., Wong, L., Bowers, M., Olausson, T. X., Liu, M., Tenenbaum, J. B., & Andreas, J. (2024). LILO: Learning interpretable libraries by compressing and documenting code. *The Twelfth International Conference on Learning Representations*.

Hawkins, R. X., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? good questions provoke informative answers. *CogSci*.

Henderson, L., Goodman, N. D., Tenenbaum, J. B., & Woodward, J. F. (2010). The structure and dynamics of scientific theories: A hierarchical bayesian perspective. *Philosophy of Science*, *77*(2), 172–200.

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, *24*(4), 613–632.

Klayman, J., & Ha, Y.-w. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596.

Lew, A. K., Tessler, M. H., Mansinghka, V. K., & Tenenbaum, J. B. (2020). Leveraging unstructured statistical knowledge in a probabilistic language of thought. *Proceedings of the Annual Conference of the Cognitive Science Society*.

Lew, A. K., Zhi-Xuan, T., Grand, G., & Mansinghka, V. K. (2023). Sequential monte carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*.

Li, B. Z., Tamkin, A., Goodman, N., & Andreas, J. (2023). Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Markant, D., & Gureckis, T. (2012). Does the utility of information influence sampling behavior? *Proceedings of the annual meeting of the cognitive science society*, *34*(34).

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive science*, *40*(1), 100–120.

Meder, B., Nelson, J. D., Jones, M., & Ruggeri, A. (2019). Stepwise versus globally optimal search in children and adults. *Cognition*, *191*, 103965.

OpenAI. (2023a). GPT-4 Technical Report.

OpenAI. (2023b, September). GPT-4V(ision) System Card.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Piantadosi, S. T., Rule, J. S., & Tenenbaum, J. B. (2024). Learning as Bayesian inference over programs. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Bayesian models of cognition: Reverse-engineering the mind*. MIT Press.

Piriyakulkij, T., Kuleshov, V., & Ellis, K. (2023). Active preference inference using language models and probabilistic reasoning. *arXiv preprint arXiv:2312.12009*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*.

Rothe, A., Lake, B. M., & Gureckis, T. (2017). Question asking as program generation. *Advances in neural information processing systems*, *30*.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*, 69–89.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2019). Asking goal-oriented questions and learning from answers. *CogSci*, 981–986.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al. (2023). Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental psychology*, *52*(12), 2159.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Ullman, T. D., Siegel, M. H., Tenenbaum, J., & Gershman, S. (2016). Coalescing the vapors of human experience into a viable and meaningful comprehension. *CogSci*.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., & Goodman, N. (2024). Hypothesis search: Inductive reasoning with language models. *The Twelfth International Conference on Learning Representations*.

Wong, C., Ellis, K., Tenenbaum, J. B., & Andreas, J. (2021). Leveraging language to learn program abstractions and search heuristics. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (pp. 11193–11204, Vol. 139). PMLR.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Zhang, M. J., & Choi, E. (2023). Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.