

# Learning to Ask Informative Questions: Enhancing LLMs with Preference Optimization and Expected Information Gain

**Davide Mazzaccara**

CIMeC, University of Trento  
davide.mazzaccara@unitn.it

**Alberto Testoni**

ILLC, University of Amsterdam  
a.testoni@uva.nl

**Raffaella Bernardi**

CIMeC, DISI, University of Trento  
raffaella.bernardi@unitn.it

## Abstract

Questions are essential tools for acquiring the necessary information to complete information-seeking tasks. However, large language models (LLMs), especially open-source models, often perform poorly in generating informative questions, as measured by expected information gain (EIG). In this paper, we propose a method to enhance the informativeness of LLM-generated questions in 20-question game dialogues. We sample multiple questions from the same model (LLAMA 2-CHAT 7B) for each game and create pairs of low-EIG and high-EIG questions to apply a Direct Preference Optimization (DPO) algorithm. Our results show that this method produces more effective questions (in terms of EIG), even in domains different from those used to train the DPO model.

## 1 Introduction

Questions in language serve as requests for information (Searle, 1969). A speaker lacks information in their knowledge state and asks questions to gain this information. This process of acquiring information through questioning is essential for children to learn about the world (Chouinard, 2007) and for adults to solve complex problems (Geva et al., 2021). Questions, however, vary in their level of informativeness, with some questions being more informative and efficient in reaching the problem’s solution (Grand et al., 2023).

Cognitive Science provides two interesting tools to study questions’ informativeness: the 20 Questions Game as a test bed, and the Expected Information Gain (EIG) as a quantifying measure (Ruggeri and Lombrozo, 2015). The 20 Questions game consists of one player asking yes/no questions to identify the item the other player has in mind, from a pool of possible items. In this context, the EIG (Shannon, 1948) measures questions’ informativeness as the (expected) entropy reduction caused by a certain question in the space of possible items.

Most informative polar questions partition the possible items into two same-size clusters: items in one cluster are expected to receive a positive answer, and items in the other cluster are expected to receive a negative one.

In recent years, Large Language Models (LLMs) have demonstrated remarkable language and reasoning capabilities (Kojima et al., 2022; Huang and Chang, 2023). The ability to ask informative and effective questions is crucial for employing these systems as successful user assistants on a large scale. Although LLMs have demonstrated their ability to play the 20 Questions game, their questions are characterised by low informativeness and limited success (Bertolazzi et al., 2023). Recent studies have emerged to improve the informativeness of LLMs’ questions. Zhang et al. (2024) improve open-source LLMs via Reinforcement Learning and Behavioral Cloning from larger LLMs’ data. Relying on LLMs’ ability to generate diverse questions and provide reliable answers (Piriyakulkij et al., 2023; Testoni et al., 2023), Hu et al. (2024) propose an inference-time re-ranking strategy based on EIG. Differently from them, we propose a training strategy leveraging EIG as a refined signal with a preference optimization algorithm. Unlike other approaches, all the steps of our method are accomplished by the same open-source model, without requiring annotation or feedback signals from external models.

This study aims to improve the informativeness of questions generated by LLMs, thereby leading to more efficient agents. The 20 Questions Game serves as a testbed to illustrate the feasibility of our approach. To achieve our aim, we propose a method involving three steps: sampling multiple questions, questions’ evaluation in terms of EIG, and training with preference optimization. In the first step, the model generates a set of possible questions, and it computes the EIG for each question. This set of questions, along with the corresponding

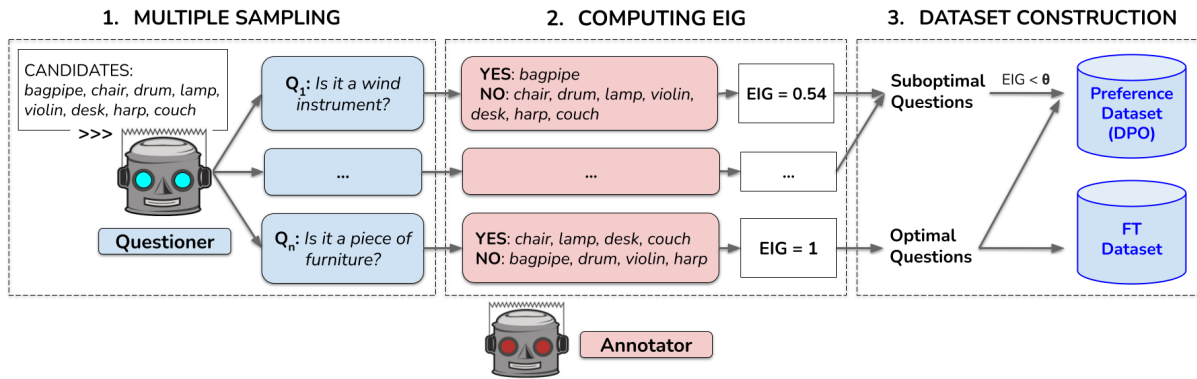


Figure 1: The proposed approach for constructing the datasets of dialogues for fine-tuning and preference optimization (DPO). From the original candidate set, the Questioner generates a question  $Q_1$ , and the Annotator provides the expected answer to each candidate. Expected Information Gain (EIG) is computed from the annotation: if the question is suboptimal in terms of EIG, other questions are sampled until an optimal question is reached ( $Q_n$ ). The optimal question is paired with the suboptimal ones in the Preference dataset (DPO), whereas the Fine-Tuning (FT) dataset is composed of only 1-EIG questions.

EIG values, as a proxy for the questions’ informativeness, is used for Direct Preference Optimization (DPO; Rafailov et al. 2023). Our results show that EIG is a strong training signal to improve the question-asking capabilities of current LLMs and overcome their shortcomings in asking effective questions.<sup>1</sup>

While these findings hold in the controlled setting of the 20 Questions Game, our approach could be extended outside this paradigm and metric to train different models’ capabilities. Our core idea of sampling, self-annotation to filter with a metric, and preference optimization could improve questions’ generations in a plethora of tasks – recommendation systems (Piriyakulkij et al., 2023) and image retrieval (Keh et al., 2023) for example. Other metrics, such as Expected Savings (Rothe et al., 2018), could be employed to generate large preference datasets and improve LLMs’ information-seeking abilities beyond our tested paradigm.

## 2 Methodology

### 2.1 Setting

Our experimental setting, inspired by the 20 Questions Game paradigm, involves two players with different roles: a Questioner and an Answerer. The Answerer is secretly assigned one target entity  $\omega$  among a pre-defined set of candidates (the *candidate set* –  $\Omega$ ). The Questioner receives the initial

candidates and is instructed to ask yes/no questions, in order to identify  $\omega$ .

Each *game* consists of a set of possible items,  $\Omega$ , available to the Questioner, and a specific target item,  $\omega$ , available to the Answerer, where  $\omega \in \Omega$ . A *dialogue* is a series of question-answer exchanges between the two players, ending when the Questioner correctly identifies the target or reaches the maximum number of questions (20). If the target is reached within the 20 questions, the dialogue is *successful*; otherwise, it is *unsuccessful* (see more details on the successful condition in Appendix A.1). For a given game, if the dialogue is not successful on the first attempt, the Questioner and the Answerer can engage in up to  $n$  dialogues to identify the target, where  $n = |\Omega|/2$ .

### 2.2 Data Sampling

Following the procedure in Bertolazzi et al. (2023), 10,000 games are extracted from McRae concepts (McRae et al., 2005). The candidate sets consist of concepts belonging to 6 categories (mammal, bird, clothing, weapon, fruit, and vegetable) for a total of 66 unique concepts. Each candidate set consists of 8 elements ( $|\Omega| = 8$ ) belonging to two categories of 4 elements each.

In our setting, all the game’s roles are played by the same LLM, LLAMA 2-CHAT 7B (Touvron et al., 2023). Along with the Questioner and the Answerer, sampling involves the Annotator, who provides the annotation to compute the EIG. As illustrated in Fig. 1, English dialogues are generated from each game by repeatedly sampling questions

<sup>1</sup>Code and data are available at <https://github.com/dmazzaccara/LearningToAsk>

from the Questioner. For each generated question, the Annotator provides the yes/no annotation ( $\Omega_{yes}$  and  $\Omega_{no}$ ). If the question splits the candidate sets in two equal subsets ( $|\Omega_{yes}| = |\Omega_{no}|$ ), the question is *optimal* in terms of EIG (details on EIG computation are in Appendix A.2). Once an optimal question is reached, it is saved along with the answer by the Answerer in the dialogue history. This dialogue history is provided as input to the Questioner to sample the optimal follow-up question (implementation’s details are in Appendix A.3).

This process results in two clusters of optimal and suboptimal questions, from which two datasets are obtained. A *fine-tuning dataset* consisting of dialogues of only optimal questions ( $EIG = 1$ ), indicating that they evenly split the candidate space at each turn. A *preference dataset*, which consists of 55,000 optimal ( $EIG = 1$ ) vs suboptimal ( $EIG < 0.8$ ) question pairs.

### 2.3 Questioner Models

Our evaluation of the Questioner compares LLAMA 2-CHAT (7B) zero-shot with the trained versions below. Both the training processes rely on LoRA adapters and standard hyperparameter settings.

**Fine-tuning** (FT) involves a causal language modelling objective applied to the entire sequence. For development, we use 50 games excluded from the fine-tuning dataset. To obtain the best adapter’s checkpoint, we test the adapter on the development set every 1,000 training samples. According to the number of questions per dialogue, the best adapter’s checkpoints are after 4,000 dialogues.

**Direct Preference Optimization** (DPO) (Rafailov et al., 2023) training objective increases the likelihood of optimal EIG questions while decreasing the likelihood of the low-EIG question. For DPO, we train with all the 55,000 pairs of optimal vs low-EIG questions.

### 2.4 Evaluation

**Test Sets:** three test sets evaluate questioner models with candidate sets of the same size as in training ( $|\Omega| = 8$ ): INLG, Things, Celebrities; two test sets test with more candidates than in training: INLG 16 and BigBench. The same-size test sets have candidate sets from different domains: INLG (Bertolazzi et al., 2023) from seen categories but unseen concepts; both from Zhang et al. (2024), Things has unseen categories but common-life concepts, Celebrities has unseen categories and concepts compared to training (famous personalities).

Different-size test sets are INLG 16 (Bertolazzi et al., 2023), with sets of  $|\Omega| = 16$  unseen concepts of seen categories, and BigBench, with sets of  $|\Omega| = 29$  different categories and concepts (Srivastava et al., 2023). Details are in Appendix B. **Metrics:** the impact of FT and DPO on zero-shot is assessed with three metrics. Task success (**S@1**) is the percentage of games in which the model identifies the target within the first dialogue. Average number of questions (**AQ**) is the number of questions to reach the target in successful dialogues. Expected Information Gain (**EIG**) is the averaged questions’ EIG in successful dialogues.

## 3 Results

Different methods are assessed on test sets with candidate sets of the same size as in training but from other domains. The results are reported in Table 1. For INLG and Things, DPO significantly improves performance over the zero-shot model in terms of both S@1 (+12.2% and +10%) and AQ (−2.1 and −2.3 questions per successful dialogues). For Celebrities, where the concept space differs greatly from the training data, DPO reduces the average number of questions by 2.5, while S@1 shows only a marginal improvement. A consistent improvement of DPO over zero-shot is also observed for the average EIG of the generated questions. Overall, the fine-tuning approach leads to a significant degradation in the S@1 metric.

Set	Method	S@1 ↑	AQ ↓	EIG ↑
INLG	zero-shot	56.7%	7.1	0.34
	FT	46.6%	<b>4.6</b>	0.41
	DPO	<b>68.9%</b>	5.2	<b>0.45</b>
Things	zero-shot	51.1%	7.5	0.29
	FT	42.2%	5.4	0.31
	DPO	<b>61.1%</b>	<b>5.2</b>	<b>0.40</b>
Celebrities	zero-shot	71.1%	7.6	0.35
	FT	46.7%	5.5	0.39
	DPO	72.2%	<b>5.1</b>	<b>0.47</b>

Table 1: **Different Domains:** Results on candidate sets with  $|\Omega| = 8$ . Across the three sets, DPO identifies the target within the first dialogue (S@1) more often, with shorter dialogues (AQ) and more informative questions (EIG). FT and DPO’s improvements for AQ and EIG are statistically significant compared to zero-shot scores (Mann-Whitney U test).

Given the promising results of DPO, we proceed to evaluate the robustness of this approach with

larger candidate sets ( $> 8$ ) at test time – preference optimization is still based on candidate sets of 8 elements. As shown in Table 2, with candidate sets of 16 and 29 candidates, DPO outperforms zero-shot in terms of S@1, with improvements of +6.7% and +13.8%, respectively. DPO generates shorter dialogues, reducing questions by 3.2 and 0.7. However, the average EIG of the DPO’s questions slightly decreases compared to zero-shot in BigBench.

Set	Method	S@1 $\uparrow$	AQ $\downarrow$	EIG $\uparrow$
INLG 16	zero-shot	44.4%	9.5	0.31
	DPO	<b>51.1%</b>	<b>6.3</b>	<b>0.38</b>
BigBench	zero-shot	17.2%	8.8	0.31
	DPO	<b>31.0%</b>	8.1	0.28

Table 2: **Different Size:** DPO boosts the model performance on INLG 16, and improves its success rate but not the other scores with BigBench. Improvements in INLG 16 for AQ and EIG are statistically significant (Mann-Whitney U test).

## 4 Analysis

First of all, following Bertolazzi et al. (2023), we examine the types of questions that models ask: *constraint seeking* (CS) are questions about a feature shared by more candidates, whereas *hypothesis scanning* (HS) are questions about one single candidate. In every test set apart from BigBench, zero-shot asks more CS with lower informativeness than DPO (Table 3: EIG for CS questions has a +0.17 compared to zero-shot). In every test set apart from BigBench, we further observe DPO asking more HS questions than the zero-shot. Overall, DPO seems to follow a more effective information-seeking strategy: DPO narrows down to the relevant candidate subset by initially posing a series of CS questions, subsequently moving to mostly coherent HS. This holds also with abstract concepts, unseen during sampling, as illustrated in Figure 2.

Secondly, we aim to understand the high scores DPO obtains in INLG and the lower ones in BigBench. INLG candidate sets are organized taxonomically, half of the candidates belong to one category and the other half to another one. By inspecting the dialogues, we saw that both zero-shot and DPO tend to ask informative questions at the first turn, identifying one of the two categories. However, the two models differ in the follow-up

Set	Method	HS		CS	
		%	EIG	%	EIG
INLG	zero-shot	47.45	0.25	52.55	0.42
	DPO	60.05	0.33	39.95	<b>0.62</b>
Things	zero-shot	44.13	0.22	55.87	0.35
	DPO	67.59	0.33	32.41	<b>0.56</b>
Celebrities	zero-shot	31.84	0.15	68.16	0.44
	DPO	53.92	0.36	46.08	<b>0.59</b>
INLG 16	zero-shot	38.66	0.19	61.34	0.39
	DPO	51.47	0.27	48.53	<b>0.51</b>
BigBench	zero-shot	64.2	0.22	35.71	0.48
	DPO	61.64	0.21	38.36	0.41

Table 3: Percentage (%) and average EIG for hypothesis scanning (HS) and constraint-seeking (CS) questions.

**Target:** anger  
**List of candidates:** car, glass micropipette, global poverty, house, anger, train, airplane, television, representative democracy, love, gigantosaur, hate, tree, surprise, disgust, siberian husky, phase transition, shoes, frisbee, jealousy, positive sum game, beauty, coat, apple, cow, contentment, hopefulness, dinosaur, shower

### Dialogue

**Q:** Is the target an animal? **A:** No  
**Q:** Is the target a type of vehicle (car, train, airplane, etc.)? **A:** No.  
**Q:** Is the target a type of building or structure (house, train station, airport, etc.)? **A:** No.  
**Q:** Is the target a form of technology (computer, phone, micropipette, etc.)? **A:** No.  
**Q:** Is the target a type of clothing or accessory (shirt, pants, hat, etc.)? **A:** No.  
**Q:** Is the target a type of emotion or feeling (love, hate, joy, etc.)? **A:** Yes.  
**Q:** Great! Is the target “love”? **A:** No.  
**Q:** Is the target “hate”? **A:** No.  
 ...  
**Q:** Is the target “anger”? **A:** Yes.

Figure 2: Example from the BigBench. DPO asks grounded CS questions (highlighted with colors), identifying the subset of the target (e.g., emotions). It then asks a series of HS questions.

turn: when the first turn receives a negative answer, zero-shot questions tend to have low EIG, while DPO questions are informative by zooming on the identified category (more on negation in Appendix C). This suggests that DPO has learned to profit from negatively answered questions, on which LLMs are known to fail (Ravichander et al., 2022; Truong et al., 2023). We dive into BigBench

by comparing the models’ performance on concrete and abstract targets (Table 4). In general, both zero-shot and DPO perform better when the target is a concrete entity. Notably, DPO significantly improves performance with both types of concepts (+12.5% and +15.4% for concrete and abstract targets, respectively). This improvement, however, is not reflected in the average dialogue length when the target is abstract. The significant difference in guessing concrete vs. abstract targets (25.0% vs. 7.7% for zero-shot and 37.5% vs. 23.1% for DPO) calls for further investigation into the underlying factors contributing to this disparity and the potential need for tailored strategies to handle abstract concepts more effectively.

Set	Method	S@1 $\uparrow$	AQ $\downarrow$	EIG $\uparrow$
Concrete	zero-shot	25.0%	8.4	0.33
	DPO	<b>37.5%</b>	<b>6.3</b>	<b>0.34</b>
Abstract	zero-shot	7.7%	9.3	0.29
	DPO	<b>23.1%</b>	10.3	0.24

Table 4: BigBench: concrete vs. abstract target.

## 5 Conclusion

In our work, we designed a Direct Preference Optimization (DPO) approach to enhance the informativeness of questions asked by LLMs using Expected Information Gain (EIG). We utilized the 20 Questions Game paradigm, a framework in cognitive science and AI for studying information-seeking behavior, reasoning, and hypothesis testing. Our approach involved (a) sampling multiple questions from the model in a zero-shot fashion, (b) clustering the questions based on their EIG, and (c) training the model using these clusters with DPO. Our results show that DPO significantly improves question informativeness (measured by average EIG and number of questions asked), making the dialogue strategy more effective. Notably, this method generalizes well to different domains. Our findings demonstrate that EIG is a promising training signal for improving the reasoning capabilities of LLMs in information-seeking dialogues.

## 6 Limitations

This work is intended to be exploratory. Our limitations pertain to three main categories: the controlled paradigm employed, the model and the training regime tested, and the EIG computation.

Our version of the 20 Questions game poses two limits: a closed set of possible candidates, and polar questions and answers. We consider it necessary to investigate this controlled setting before transitioning to more realistic scenarios. Nevertheless, possible solutions to compute EIG in an open setting emerge from the literature: [Hu et al. \(2024\)](#) rely on the expected candidates from the model; [Zhang et al. \(2024\)](#), instead, build the set of initial candidates from external sources. Transitioning from polar questions to more realistic scenarios remains an open challenge. Related studies, however, already provide useful strategies to overcome this restriction, computing EIG with different-shaped ([Grand et al., 2023](#)) and open-ended questions ([Keh et al., 2023](#)).

Secondly, to investigate whether LLMs could learn to be more informative and effective with self-generated EIG signals, we focus on one model (LLAMA 2-CHAT 7B) and one preference optimization strategy (DPO). We select LLAMA 2-CHAT 7B as the best-performing open-source model at the start of the project. According to our computing resources (2X24 GB GPUs), larger 13B and 70B models’ versions could not be trained without quantization. In preliminary studies on the 7B and 13B versions (zero-shot), we observed quantization (8/4 bit) leading to performance degradation—on the S@1 metric in particular. For this reason, we preferred to focus on the 7B version. Further work is required to determine if this training strategy holds with other models and other preference optimization strategies.

A third limitation of our study is related to the EIG computation. EIG computation depends on the yes/no annotation. While we can assume a high degree of accuracy, based on our careful inspection of the dialogues, some questions are still difficult to answer with only yes/no (e.g.,  $w = \text{dinosaur}$ , the question “Is it living?”). Additionally, all the candidates in  $\Omega$  are considered equally likely, while LLMs have priors conditioning their question generation. Furthermore, when computing the EIG of follow-up questions, we consider the model able to sequentially rule out candidates excluded in the dialogue history, which could be a strong assumption for a generative language model.

## Acknowledgments

We are grateful to David Schlangen and his group and to the audience of the Evil seminar orga-

nized by Marco Baroni. By addressing their questions the quality of the work improved. We thank Leonardo Bertolazzi for his suggestions on the early phase of the implementation of this work. Alberto Testoni is supported by the European Research Council (ERC) funding under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 819455, PI R. Fernández).

## References

- Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. 2023. [ChatGPT’s information seeking strategy: Insights from the 20-questions game](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162, Prague, Czechia. Association for Computational Linguistics.
- Michael M. Chouinard. 2007. *Children’s questions: a mechanism for cognitive development*. Monographs of the Society for Research in Child Development vol. 72,1.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies](#). *Transactions of the Association for Computational Linguistics*, 9.
- Gabriel Grand, Valerio Pepe, Jacob Andreas, and Joshua B. Tenenbaum. 2023. [Loose lips sink ships: Asking questions in battleship with language-informed program sampling](#). In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. [Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models](#). In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards reasoning in large language models: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- Emilio Jorge, Mikael Kågebäck, Fredrik D. Johansson, and Emil Gustavsson. 2017. [Learning to play guess who? and inventing a grounded language as a consequence](#). *Preprint*, arXiv:1611.03218.
- Sedrick Keh, Justin T. Chiu, and Daniel Fried. 2023. [Asking more informative questions for grounded retrieval](#). *Preprint*, arXiv:2311.08584.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 4(37):547–559.
- Top Piriyaakulkij, Volodymyr Kuleshov, and Kevin Ellis. 2023. [Asking clarifying questions using language models and probabilistic reasoning](#). In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.
- Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. [CONDAQA: A contrastive reading comprehension dataset for reasoning about negation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anselm Rothe, Brenden M. Lake, and Todd M. Gureckis. 2018. [Do people ask good questions?](#) *Computational Brain and Behavior*, 1(1):69–89. Publisher Copyright: © 2018, Springer International Publishing.
- Azzurra Ruggeri and Tania Lombrozo. 2015. [Children adapt their questions to achieve efficient search](#). *Cognition*, 143:203–216.
- John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London.
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Aarohi Srivastava and al. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Alberto Testoni and Raffaella Bernardi. 2021. [Looking for confirmations: An effective and human-like visual dialogue strategy](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9330–9338, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alberto Testoni, Raffaella Bernardi, and Azzurra Ruggeri. 2023. [The efficiency of question-asking strategies in a real-world visual search task](#). *Cognitive Science*, 47(12).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. [Language models are not naysayers: an analysis of language models on negation benchmarks](#). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (\*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [Star: Bootstrapping reasoning with reasoning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488. Curran Associates, Inc.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. 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 (Volume 1: Long Papers)*, pages 1495–1516, Bangkok, Thailand. Association for Computational Linguistics.

## A Methodology

### A.1 Successful Condition

A dialogue is *successful* when the Questioner guesses the target entity. We have implemented this condition with the following two cases: (a) the Questioner explicitly refers to the target entity without naming other candidate entities or (b) the Questioner explicitly refers to the target entity along with other candidates, and the target is named more times than each of the other entity. This second condition recognizes when the Questioner is explicitly reasoning about the target before guessing (e.g., target: ‘cherry’. Questioner: ‘So, the target is one of the remaining fruits, which are cherry and grapefruit. Let me make a guess. Is the target... cherry?’). The dialogue is considered *unsuccessful* if the Questioner does not meet the conclusive conditions above within 20 questions.

### A.2 Expected Information Gain

In this setting, each question aims to rule out as many candidates as possible from the set  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ . The most informative question rules out half of the possible candidates, whereas a question ruling out none or just one candidate has low informativeness. Questions’ informativeness has been quantified in prior works through the Expected Information Gain.

Expected Information Gain (EIG) measures the contribute of each question to reduce uncertainty towards the solution  $\omega$ . The level of uncertainty is measured through the entropy  $H$ , higher entropy means higher uncertainty. Consequently, EIG is computed subtracting from the initial state of entropy ( $H_{prior}$ ) the expected entropy after asking the question ( $H_{posterior}$ ).

$$EIG = H_{prior} - H_{posterior} \quad (1)$$

At the beginning of each game, we assume a uniform prior distribution over all the possible candidates  $\omega_1, \omega_2, \dots, \omega_n$ . From the Shannon entropy (Shannon, 1948), this initial state  $H_{prior}$  in our setting is equal to:

$$\begin{aligned} H_{prior} &= - \sum_{i=1}^n p(w_i) \log_2 p(w_i) \\ &= -n \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n \end{aligned} \quad (2)$$

A yes/no question divides the current candidates in  $\Omega$  into two disjoint subsets:  $\Omega_{yes} \cup \Omega_{no}$ . In



the subset  $\Omega_{yes}$  there are the candidates in  $\Omega$  that are consistent with a possible positive answer,  $\Omega_{no}$  with a negative one. Answering the question, the Answerer determines which subset becomes the new set of candidates under consideration for the next turn (the candidates in the other subset are ruled out). The posterior entropy is computed as the sum of the entropies of both the yes and no subsets, weighted by their probability.  $H_{posterior}$  is computed as:

$$H_{posterior} = p_{yes} H_{yes} + p_{no} H_{no} \quad (3)$$

This computation is based on the expected answers to the question for each candidate in the current set. In other words, the Questioner has an expectation of which candidates will be in  $\Omega_{yes}$  and  $\Omega_{no}$  before receiving the answer by the Answerer. With these subsets, the probabilities  $p_{yes}$  and  $p_{no}$  refer to the probabilities of receiving a positive or negative answer. The entropies  $H_{yes}$  and  $H_{no}$  measure uncertainty in the two subsets computed as in eq. 2. Intuitively, the higher the similarity between the two subsets in terms of size (i.e., candidates per subset) the higher the EIG value. The most informative yes/no questions divide  $\Omega$  into two subsets  $\Omega_{yes}$  and  $\Omega_{no}$  of the same size, resulting in  $EIG = 1$ .

### A.3 Setting

#### A.3.1 Sampling

Our approach comprises the three steps described in Fig. 3. During sampling, different questions are sampled from the Questioner . To compute EIG, the Annotator  annotates the questions with yes/no for each candidate in the candidate set ( $\Omega$ ). Once an ‘optimal’ question is reached ( $EIG = 1$ ), the optimal and suboptimal questions are stored in the training datasets. All the prompts for the roles are reported below. The Annotator has the same prompt as the Answerer, taking as a target each candidate in the candidate set.

#### Questioner

**System Prompt:** *You are playing a game, make only one yes/no question at turn to identify the target from the List of candidates. If there are 1 or 2 candidates remaining make the guess.*

**User Prompt:** *List of candidates:*  
CANDIDATES.

#### Answerer/Annotator

**System Prompt:** *You are playing the 20-Questions game, you will be asked one Question about the Target element. Answer only ‘yes’ or ‘no’ to the Question depending on your Target element.*

**User Prompt:** *Target element:* TARGET,  
*Question:* QUESTION

To yield diverse outputs during sampling, the decoding strategy for the Questioner involves a relatively high temperature (1.0) and top-k sampling (50). To achieve more deterministic responses the Answerer/Annotator has a low temperature of 0.1.

This process produces a Fine-Tuning and DPO dataset. The Fine-Tuning dataset consists of all the completed dialogues of optimal questions only. The DPO dataset, instead, comprises pairs of ‘optimal’ questions ( $EIG = 1.0$ ) with suboptimal questions ( $EIG < 0.8$ ).

#### A.3.2 Training

The Fine-Tuning and DPO datasets are employed to train LLAMA 2-CHAT (7B) with LoRA adapters (Hu et al., 2022). Our main focus is on the approach rather than the training strategy, thus we did not perform extensive hyperparameter tuning. We do not exclude that additional experiments on hyperparameters could lead to better results for our approach. For both Fine-Tuning and DPO, the LoRA adapters are applied on all the modules, with an  $r = 128$  and the alpha value = 32. We perform Fine-Tuning and DPO with AdamW as the optimizer, training for one epoch and with a batch size = 8. For both Fine-Tuning and DPO, a learning rate of  $2e - 5$  has been observed to produce consistent results.

#### A.3.3 Testing

At test time, we compare the LLAMA 2-CHAT (7B) zero-shot with its Fine-Tuned and DPO alternatives.



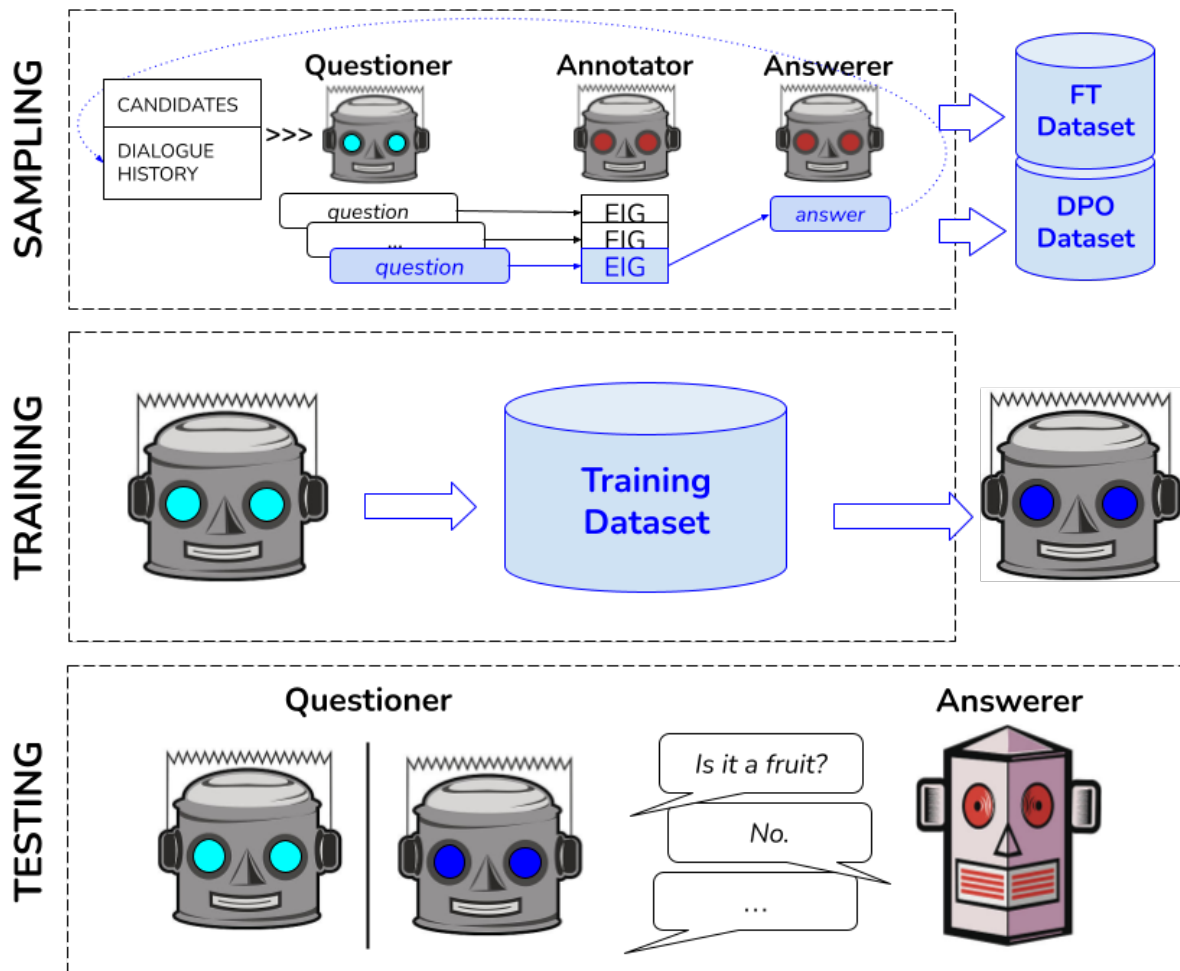


Figure 3: During **sampling**, LLAMA 2-CHAT (7B) plays the roles of the Questioner, the Annotator, and the Answerer. Questions are sampled from the Questioner and then evaluated by the Annotator. Once an optimal question is reached, the Answerer answers it. The optimal question and its answer are appended to the dialogue history. In this way, optimal questions are sampled not only for the first turn but also in follow-up turns. In **training**, the Questioner is trained with FT and DPO datasets. In **testing**, the zero-shot 🤖 and trained 🤖 Questioners play the 20 Questions Game with an external model as Answerer.

We employ an external model as Answerer to make the test setting more different from the training setting. The Answerer model is GPT-3.5-TURBO-0125, with default parameters for generation. The total cost for testing - every approach in every setting - is less than 3\$. At test time, we experimented also with the Questioner using the default LLAMA 2-CHAT hyperparameters, the results are consistent with the ones reported in the paper Section 3. For the hyperparameters used in the paper, the results across multiple runs are also consistent.

### Implementation Details

The inference and training with LLAMA 2-CHAT (7B) were performed using the Huggingface Transformer library (Wolf et al., 2020). All experiments were executed on a Nvidia GeForce RTX 3090 24GB GPU. Sampling 10,000 dialogues took approximately 4 days. Fine-Tuning and DPO take around 12 hours each. Running inference on all the test sets lasts around 6 hours for FT and DPO, whereas it lasts around 10 hours for the zero-shot model.

### B Test Sets

The games in the test sets are with candidates unseen during training. To test the trained models in different domains, we have the following test sets of 8 candidates, as in training:

- **INLG**: 90 candidate sets structured in such a way that half of the candidates pertain to one taxonomic category and the other half to another one (Bertolazzi et al., 2023). The taxonomic categories for the candidates are animals, clothing, weapons, fruits, and vegetables.
- **Things**: 90 candidate sets from the evaluation split of Zhang et al. (2024). The full list of categories are animals, clothing, foods, objects, plants, vehicles, professions, materials, instruments, places, sports, buildings, furniture, celestial bodies, mythical creatures, events, and activities.
- **Celebrities**: 90 sets from the evaluation split of Zhang et al. (2024). In this test set, the candidates differ greatly from the training ones, comprising only past and living celebrities.

To test the trained models with more candidates ( $\Omega > 8$ ), we have the following test sets:

- **INLG 16**: 90 candidate sets of 16 structured elements, as in INLG (Bertolazzi et al., 2023).
- **BigBench**: 29 sets of 29 candidates of abstract and concrete concepts. The concepts are apple, television, dinosaur, airplane, house, tree, coat, shoes, car, train, shower, frisbee, cow, gigantosaur, siberian husky, glass micropipette (concrete); anger, love, hate, contentment, jealousy, surprise, disgust, hopefulness, global poverty, phase transition, positive sum game, beauty, representative democracy (abstract).

## C Analysis

### Informativeness after a Negative Answer

INLG (8) and INLG 16 consist of structured sets, i.e., half of the candidates of one category and the other half to another one. Both the zero-shot and the DPO tend to ask informative questions at the first turn, identifying one of the two categories. After a negative answer to this first question, the zero-shot frequently asks a confirmation question about the other category ( $EIG = 0$ ), as illustrated in Figure 4. In both INLG 8 and INLG 16 of Table 5, the zero-shot shows a higher percentage of 0-EIG questions and lower average EIG of questions after a negative answer, compared to questions after a positive answer (EIG after negative and positive questions drawn from different distributions, Mann–Whitney U test). On the contrary, the DPO has a lower percentage of 0-EIG questions and higher EIG after a negative answer (EIG after negative and positive questions of the same distribution). Trained on similarly structured candidate sets, the DPO seems not to reproduce the uninformative behaviour of the zero-shot, asking more informative questions after negative answers.

Set	Method	after yes		after no	
		EIG $\uparrow$	%0 EIG $\downarrow$	EIG $\uparrow$	%0 EIG $\downarrow$
INLG 8	zero-shot	0.36	60.1%	0.21	74.1%
	DPO	0.37	59.4%	<b>0.33</b>	<b>61.6%</b>
INLG 16	zero-shot	0.35	55.6%	0.20	71.5%
	DPO	0.29	65.6%	<b>0.29</b>	<b>60.2%</b>

Table 5: Questions’ informativeness after a positive and a negative answer, in terms of average EIG (EIG) and percentage of uninformative questions (% 0 EIG).

**Target:** fox

**List of candidates:** *elk*, chicken, robin, starling, fox, partridge, hamster, buffalo

**Zero-shot**

**Q<sub>1</sub>:** Is the target a bird? *EIG = 1*      **A:** No.

**Q<sub>2</sub>:** Is the target a mammal? *EIG = 0*      **A:** Yes.

**Q<sub>3</sub>:** Does the target have a long tail?      **A:** Yes.

**Q<sub>4</sub>:** Is the target a fox?      **A:** Yes.

**DPO**

**Q<sub>1</sub>:** Is the target bird? *EIG = 1*      **A:** No.

**Q<sub>2</sub>:** Is the target mammal a large hoofed animal, like a deer or a bison? *EIG = 1*      **A:** No.

**Q<sub>3</sub>:** Is the target a carnivorous mammal, like a fox?  
**A:** Yes.

Figure 4: INLG 8 Example: half of the candidates are birds, and half are mammals. Both the zero-shot and DPO identify the bird category at the first turn ( $EIG = 1$ ). After the negative answer, the zero-shot asks a confirmation question about the remaining category (i.e., mammal), with an  $EIG = 0$ . The DPO, instead, asks a more specific question on the remaining category (i.e., large hoofed, which are the ‘elk’ and ‘buffalo’), with an  $EIG = 1$ . The higher informativeness of DPO is further reflected in the number of questions required to reach the target: 3 questions in DPO vs. 4 questions in zero-shot.