How Do LLMs Ask Questions? A Pragmatic Comparison with Human Question-Asking

Chani Jung[1*] **Jimin Mun**[2*] **Xuhui Zhou**[2] **Alice Oh**[1] **Maarten Sap**[2] **Hyunwoo Kim**[3]

[1]KAIST [2]Carnegie Mellon University [3]NVIDIA

Abstract

Question asking is a fundamental linguistic and cognitive skill that underpins collaboration and a wide range of social actions. Although large language models (LLMs) are known to under-use questions—often leading to misunderstandings or less productive interactions—little is known about how their question-asking behavior differs from that of humans in social contexts. To bridge this gap, we compare the distribution and characteristics of LLM-generated questions with those produced by humans in a real-world social environment, Reddit. Using a pragmatics-based taxonomy of social actions, we analyze six open- and closed-source model families. Our findings reveal that LLMs concentrate on a narrower range of question types, exhibiting significant distributional differences from human behavior. Prompting often introduces additional biases that diverge from human patterns, while the effects of alignment tuning vary across models and are inconsistent across different social actions. These results underscore the need for more fine-grained strategies to guide LLMs' question-asking behavior, ultimately enhancing their communicative effectiveness in real-world social interactions.

1 Introduction

Question asking is a hallmark of human intelligence, reflecting critical thinking [Albergaria-Almeida, 2011], curiosity [Jirout and Klahr, 2012], and creativity [Acar et al., 2023]. It also underpins collaboration [Sperber et al., 2010] and serves diverse social functions [Stivers et al., 2010]. Thus, question asking is a crucial capability for language models (LMs) and is closely tied to efficiency [Zamani et al., 2020, Braslavski et al., 2017] and preference alignment [Zhang et al., 2018, Li et al., 2025b]; however, LMs are known to rarely produce questions in their outputs [Shaikh et al., 2024]. Prior work has linked this limitation to preference-tuning data, focusing mainly on clarification questions [Shaikh et al., 2024, Zhang et al., 2024, Andukuri et al., 2024] or on a question's utility for improving answer quality [Zhang et al., 2024]. While clarification is valuable for problem-solving, human question asking is far more dynamic–used to seek information, propose alternatives, and challenge assumptions [Stivers et al., 2010]. This highlights a gap in understanding the broader communicative, creative, and curiosity-driven functions of questions across social contexts.

To address this gap, we analyze and interpret question-asking behavior in both humans and LMs. Our investigation focuses on: (1) *How do humans and LMs differ in their use of question types?* and (2) *How do current alignment methods affect the question-asking behavior of LMs, and do they enhance alignment with human patterns?* To answer these questions, we construct a dataset from a Reddit community, `r/NoStupidQuestions`. Using a question type taxonomy based on social actions [Stivers and Enfield, 2010], we categorize the questions posed by both humans and LMs in response to these posts. We then investigate how the question-asking behavior of large language

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

Table 1: Question types along with their definitions and corresponding examples

Social Action	Definition	Example
Request for information	Genuine questions seeking information.	"Are you busy tonight?"
Other initiation of repair	Includes open-class repair initiators and partial repeats.	"Huh?" / "He went where?"
Request for confirmation	Asserts a proposition for confirmation.	"So you're coming tomorrow night."
Assessment	Evaluations formatted to seek agreement.	"Isn't it beautiful out today?"
Suggestion/Offer/Request	Suggests, proposes, offers, or requests something.	"Did you want some?"
Rhetorical question	May seek a response but not an answer; often expresses opinion.	"Everything comes out in the wash, doesn't it?"
Outloud	Directed to no one in particular, often quieter, not designed to secure a response.	"Now where are my keys."
Other	Does not fit other categories; action specified separately.	Case-specific action

models (LLMs) differs from that of humans, focusing on question type distribution and qualitative characteristics.

Our findings reveal that LMs predominantly produce a narrow subset of question types compared to humans. Their question type distributions differ significantly from those of humans, often focusing largely on rhetorical questions and information requests. Prompting techniques help elicit more questions but frequently introduce prompt-specific bias—such as overproducing certain question types or relying on repetitive lexical patterns—that still do not align with human questioning. When comparing content of questions asked by humans and LLMs, we also see diverging patterns with human questions being more brief and LLM questions leaning towards eliciting preferences or state of the user. We also observe that the effect of alignment tuning is highly model-dependent and varies across social actions of questions. These findings underscore the need for more refined strategies to guide LLMs' question-asking behavior, enhancing their effectiveness in real-world social contexts.

2 Related Works

Previous studies have identified question-asking as a key limitation in LLMs [Bai et al., 2024, Shaikh et al., 2024], with most prior work focusing on clarification questions by constructing benchmarks [Li et al., 2025a, Aliannejadi et al., 2021, Guo et al., 2021] and proposing methods for improvement [Kobalczyk et al., 2025, Zhang et al., 2025, Testoni and Fernández, 2024, Zhang and Choi, 2025]. Beyond clarification, other studies have explored question-asking as a tool for downstream applications such as preference elicitation [Li et al., 2025b, Andukuri et al., 2024], information-seeking [Meng et al., 2023], and medical reasoning [Li et al., 2025c]. However, Shaikh et al. [2024] shows that prompting interventions, while increasing grounding acts like clarification and follow-up questions, often have minimal or negative effects on alignment with human behavior; replicating Zephyr's [Tunstall et al., 2023] training further revealed that contemporary preference datasets reduce question frequency due to both the scarcity of questions in training data and annotator dispreference. In this work, we extend the focus beyond clarification and task-specific settings by analyzing three model families at different training stages and investigating LMs' question-asking abilities through a comprehensive taxonomy of human question-asking [Stivers and Enfield, 2010], highlighting the diverse social functions questions serve in natural communication.

3 Methods

Question types taxonomy By adopting a coding scheme developed for human questions in natural conversation [Stivers and Enfield, 2010], we aim to investigate whether LMs can generate a diverse range of questions humans ask for various purposes. This extends the scope beyond previously researched LLM questioning abilities, such as clarification or follow-up questions. In our analysis, we focus on the social action dimension, especially in the general question-asking setting reflected in our current dataset. The taxonomy classifies the social actions performed by a question into seven categories as specified in Table 1. If a question does not fit any of the predefined social action types, it is labeled as "Other" followed by a description of the specific social action being performed.

Automatic question type annotation and aggregation We automatically label questions from both humans and LLMs using three-shot prompting with OpenAI's `o3` model, which extracts questions from text and classifies them according to Stivers and Enfield [2010]. The full prompt is provided in Appendix B. To evaluate annotation quality, we compare model labels with human annotations on 136

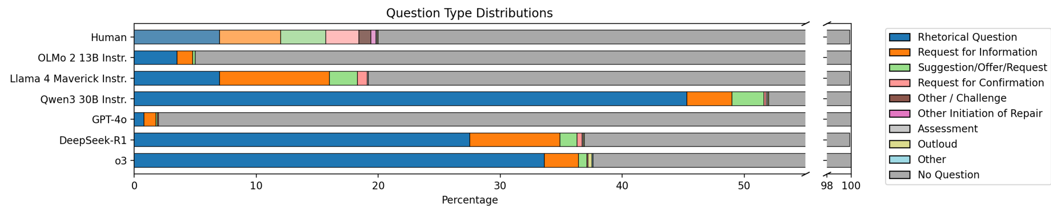


Figure 1: Question type distributions of humans and LLMs. LLMs exhibit less diversity in question types compared to humans, focusing primarily on rhetorical questions and requests for information. Chi-squared tests reveal significant differences from humans in five of six models ($p_{\text{adj}} < 0.05$), except for OLMo-2-13B-Instruct ($p_{\text{adj}} = 0.18$).

questions[2], obtaining a Cohen's Kappa of 0.78 and a macro-averaged F1 of 0.79, indicating substantial agreement. Additionally, we define *question proportion* as the average query-level proportion of each question type, computed by assigning each question in a post a weight of $1/n$ when n questions are present. This metric is used to compare humans and LLMs, as well as among LLM variants at different training stages.

Dataset We collect 3,564 posts and 38,205 comments from the r/NoStupidQuestions subreddit[3] in 2024. We include all 1,782 posts with multiple question comments and sample an equal number without question comments to balance the dataset. Comments with question marks (excluding those with URLs) are identified as questions. All comments on the selected posts are included, except those removed by administrators. This setup ensures a sufficient and unbiased sample of human-generated questions for comparison with LLMs.

LLM question generation We analyze questions generated by LLMs in response to Reddit posts in our test set. The models include OLMo-2-13B (Base, SFT, and Instruct), LLaMA-4-Maverick-17B-128E (Pretrained and Instruct), Qwen-3-30B (Base, A3B, and Instruct), GPT-4o-2024-08-06, DeepSeek-R1-0528, and o3-2025-04-16.[4] We use four prompt variations: (1) `Naive`, which provides only the Reddit post's title and body; (2) `Question`, which encourages asking questions when needed; (3) `Taxonomy`, which includes our question type taxonomy for guidance; and (4) `Imitate`, which instructs the models to emulate human Reddit users. Full prompt details are in Appendix A.

4 Results

What do LLMs' question type distribution look like? Figure 1 compares question type distributions of humans and instruction-tuned LMs using the naive prompt (Appendix 5 shows all models). LMs' question type distributions noticeably differ from humans', either asking considerably fewer questions than humans (e.g., OLMo-2-13B-Instruct, GPT-4o) or strongly favoring Rhetorical Questions and Requests for Information (e.g., LLaMA-4-Maverick, Qwen-3-30B, Deepseek-R1, O3). Across all models, Suggestions/Offers/Requests, Requests for Confirmation, Other/Initiation of Repair, and especially Other/Challenge questions (<0.04%) are under-produced compared to humans. Chi-squared tests show significant differences from humans in five of six LMs ($p_{\text{adj}} < 0.05$), with OLMo-2-13B-Instruct as the exception ($p_{\text{adj}} = 0.18$) (Appendix D). On the other hand, pairwise comparisons among the six LMs reveal no significant differences for 9 of 15 pairs[5]. Overall, LMs' question-asking behavior does not align with humans', in both overall frequency and type distributions, while exhibiting broadly similar patterns across the models when using naive prompting.

[2] 4 from each of 29 model–prompt pairs plus 20 human-generated

[3] https://www.reddit.com/r/NoStupidQuestions/

[4] We exclude other baseline methods designed to improve LLM question-asking capabilities, as they are typically trained for specific tasks (e.g., clarification or preference elicitation) and do not generate natural responses for our test set.

[5] Significant differences appear for OLMo–Qwen, LLaMA–Qwen, LLaMA–O3, Qwen–R1, Qwen–O3, and R1–O3 ($p_{\text{adj}} < 0.01$).

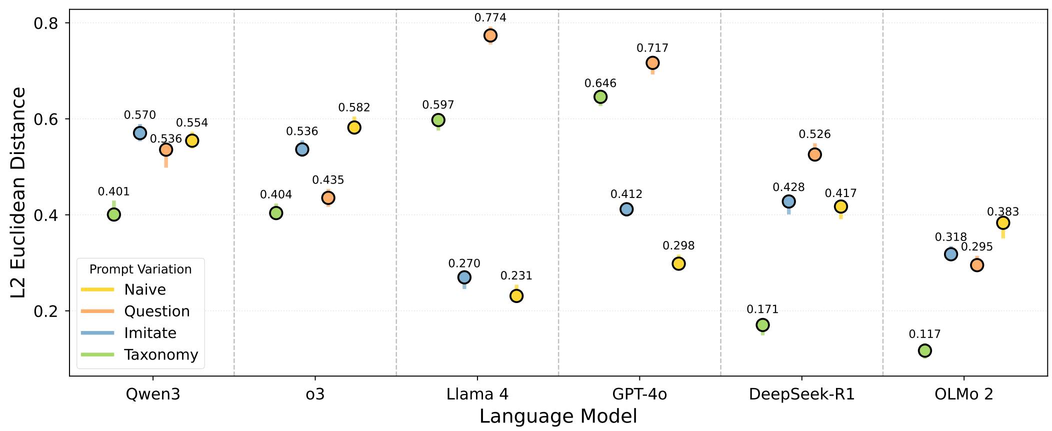
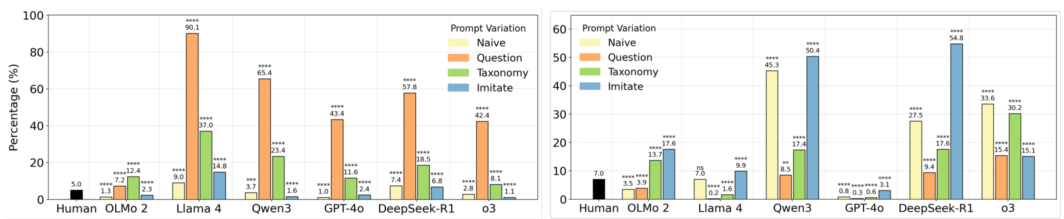


Figure 2: L2 Euclidean distances between human and LM question type distributions across prompt variations. Error bars represent 95% confidence intervals from cluster bootstrap resampling ($n = 1,000$). None of the three prompt variations consistently improve similarity with human distribution across .



(a) Request for Information

(b) Rhetorical Question

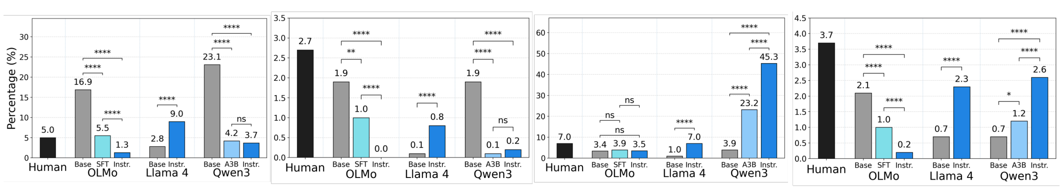
Figure 3: Question type distributions across prompting conditions. LLMs exhibit prompt-induced biases and still vary from human question patterns. Significance levels from two-sided paired t-tests with Benjamini–Hochberg correction are shown above each bar.

How much does prompting help close the human-LLM gap? We analyze the effects of three prompt variations—`Question`, `Imitate`, and `Taxonomy`. **Quantitative analyses** reveal statistically significant differences between human and model question-type distributions under all prompting conditions ($p_{\text{adj}} < 0.01$, Chi-squared tests). Moreover, none of the prompting strategies consistently reduce the L2 Euclidean distance between human and model question-type distributions (Figure 2). Interestingly, while GPT-4o and Llama-4 exhibit stronger baselines that are closer to human distributions, they perform substantially worse when prompted, indicating greater sensitivity to prompt variations. In contrast, models with weaker baselines benefit from certain prompting strategies, particularly the *taxonomy* prompt (e.g., DeepSeek-R1: $\Delta = -0.246$; OLMo-2: $\Delta = -0.266$), which introduces a structured categorization of human question types and guides models toward more human-like questioning behavior. The *imitate* prompt, despite explicitly instructing models to mimic human Reddit users, increases the distance from human distributions for most models (e.g., GPT-4o: $\Delta = +0.114$; Llama-4: $\Delta = +0.039$) while yielding only marginal improvements for the others (o3: $\Delta = -0.046$; OLMo-2: $\Delta = -0.065$). Across all models, the three prompts generally increase the proportions of Requests for Information, Suggestions/Offers/Requests, and Requests for Confirmation (Figures 3a, 6a, 6b). However, the `Question` prompt disproportionately amplifies Requests for Information, exceeding human levels by up to 85.1 percentage points, while the `Imitate` prompt inflates Rhetorical and Assessment questions beyond human proportions (Figure 3b). None of the prompts reliably increase *Other/Challenge Questions*, which remain consistently below human levels (Figure 6d).

In our **qualitative analysis** on sampled questions, we find that human Request for Information questions are typically brief and standalone, whereas LMs often embed such questions within or at the end of longer responses. These questions frequently follow templated patterns–such as offering to elaborate on related topics (e.g., "Would you like to know more about…?"), soliciting the user's opinion (e.g., “What do you think?”), or seeking validation (e.g., “Do any of these perspectives resonate with you?”). Moreover, model-generated rhetorical questions largely differ in purpose from those generated by humans. Both humans and LMs use rhetorical questions for hypophora (e.g., “Am I the audience for his films? Definitely not.”), but humans more often assert opinions by rhetorical questions (e.g., “Does it even matter?”), whereas LMs rarely do so under the `Naive`, `Question`, or `Taxonomy` prompts. Under the `Imitate` prompt, however, models become highly biased toward assertions (e.g., “Who wouldn’t want to be able to do that?”). Moreover, unlike humans, LMs tend to produce reflection-oriented rhetorical questions (e.g., “Ask yourself: What is the evidence for these thoughts?”). Assessment questions also diverge: LMs often rely on fixed forms (e.g., “. . . , right?”, “. . . , huh?”), while humans use more varied structures (e.g., “. . . , isn’t it?”, “Aren’t they . . . ?”, “. . . , ay?”). Overall, prompting techniques tend to induce or amplify biases in LMs’ question-asking behavior, which still do not align with humans’ in terms of distribution and purpose.

Do human and LLM questions also differ in content? We further employ HypotheSAEs [Movva et al., 2025] to extract interpretable features that distinguish whether a question was generated by a human or an LLM. This analysis focuses on the two most prevalent question types—Requests for Information and Rhetorical Questions. We observe distinct patterns in which human-generated questions tend to exhibit greater brevity, often consisting of concise prompts such as a simple “why,” which appears as a salient feature in both categories (see Tables 3 and 4). In contrast, LLM-generated questions more frequently inquire about users’ preferences or emotional states across both types, reflecting a more homogeneous engagement pattern that may stem from similarities in training data or objectives. Further details and extended results are provided in Appendix G.

How does alignment shape question-asking? We investigate effects of alignment methods across three LLM families–OLMo-2-13B, LLaMA-4-Maverick, and Qwen-3-30B–with different training stages. Patterns vary by question type and model: for epistemic questions (Information Requests, Confirmations), OLMo-2-13B and Qwen-3-30B base models ask more than their SFT or instruction-tuned counterparts ($p < 0.01$), while LLaMA-4-Maverick asks more in the instruct variant ($p < 10^{-4}$) (Figure 4a,b). For Rhetorical and Suggestion/Offer/Request questions, LLaMA-Instruct and Qwen-Instruct models produce more than their base versions ($p < 10^{-4}$), whereas OLMo-2-13B-Base exceeds its tuned variants only for Suggestion/Offer/Requests ($p < 10^{-4}$) and shows no significant differences in rhetorical questions (Figure 4c,d). These results indicate that alignment methods shape LLMs’ question-asking behavior in ways specific to the model and question type.



(a) Request for Information (b) Request for Confirmation (c) Rhetorical Question (d) Suggestion/Offer/Request

Figure 4: Question type proportions across LLM training stages (e.g., Base, SFT, and Instruct). The effects of post-training are family-specific and vary by social action type. Statistical significance from two-sided paired t-tests (Benjamini–Hochberg corrected) is shown above each bar.

5 Conclusion

Question asking is central to collaboration and a cornerstone of efficient interaction and preference alignment. In this work, we examined two key questions: (1) how question-asking behaviors differ between humans and language models, and (2) how existing alignment methods shape these behaviors. Our analyses reveal significant differences between humans and LLMs–from the types of questions they produce to their linguistic and semantic characteristics. Moreover, current alignment approaches appear to constrain rather than diversify question generation. These findings highlight the need for future alignment strategies that explicitly foster diversity and adaptability in question-asking, enabling models to engage more effectively and human-like in interactive contexts.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-00143911, AI Excellence Global Innovative Leader Education Program).

References

Selcuk Acar, Kelly Berthiaume, and Rebecca Johnson. What kind of questions do creative people ask? *Journal of Creativity*, 33(3):100062, 2023.

Patrícia Albergaria-Almeida. Critical thinking, questioning and creativity as components of intelligence. *Procedia-Social and Behavioral Sciences*, 30:357–362, 2011.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.367. URL `https://aclanthology.org/2021.emnlp-main.367/`.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*, 2024.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.401. URL `https://aclanthology.org/2024.acl-long.401/`.

Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. What do you mean exactly? analyzing clarification questions in cqa. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 345–348, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346771. doi: 10.1145/3020165.3022149. URL `https://doi.org/10.1145/3020165.3022149`.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. Abg-coQA: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*, 2021. URL `https://openreview.net/forum?id=SlDZ1o8FsJU`.

Jamie Jirout and David Klahr. Children's scientific curiosity: In search of an operational definition of an elusive concept. *Developmental review*, 32(2):125–160, 2012.

Kasia Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=JAMxRSXLFz`.

Belinda Z. Li, Been Kim, and Zi Wang. Questbench: Can llms ask the right question to acquire information in reasoning tasks?, 2025a. URL `https://arxiv.org/abs/2503.22674`.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL `https://openreview.net/forum?id=LvDwwAgMEW`.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan Ilgen, Yulia Tsvetkov, and Maarten Sap. Aligning llms to ask good questions a case study in clinical reasoning. *CoRR*, abs/2502.14860, February 2025c. URL `https://doi.org/10.48550/arXiv.2502.14860`.

Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. FollowupQG: Towards information-seeking follow-up question generation. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 252–271, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.17. URL `https://aclanthology.org/2023.ijcnlp-main.17/`.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. *arXiv preprint arXiv:2502.04382*, 2025.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. Grounding gaps in language model generations. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.348. URL `https://aclanthology.org/2024.naacl-long.348/`.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010.

Tanya Stivers and N.J. Enfield. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626, 2010. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2010.04.002. URL `https://www.sciencedirect.com/science/article/pii/S0378216610000986`. Question-Response Sequences in Conversation across Ten Languages.

Tanya Stivers, Nick J Enfield, and Stephen C Levinson. Question-response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, 42:2615–2619, 2010.

Alberto Testoni and Raquel Fernández. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–275, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.16. URL `https://aclanthology.org/2024.eacl-long.16/`.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, page 418–428, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380126. URL `https://doi.org/10.1145/3366423.3380126`.

Michael JQ Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.306. URL `https://aclanthology.org/2025.findings-naacl.306/`.

Michael JQ Zhang, W Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions. *arXiv preprint arXiv:2410.13788*, 2024.

Michael JQ Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=cwuSAR7EKd`.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 177–186, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271776. URL `https://doi.org/10.1145/3269206.3271776`.

A Prompts for LLM generations

A.1 Question Prompt

```
User Query: {title}
{selftext}

Converse with the user to deliver an effective response to their
query. Feel free to ask any questions if necessary.
```

A.2 Taxonomy Prompt

```
User Query: {title}
{selftext}

Respond to the user's query. If needed, consider using an
appropriate question from the taxonomy below to guide the
conversation:
- Request for Information: Questions genuinely seeking information,
with no other primary action apparent.
- Other Initiation of Repair: Questions including open class repair
initiators as well as partial repeats.
- Request for Confirmation: Questions asserting a proposition and
seeking confirmation.
- Assessment: Evaluations that were formatted to seek agreement.
- Suggestion/Offer/Request: Questions proposing, suggesting,
offering, or explicitly requesting something.
- Rhetorical Question: Questions that assert opinions or points, not
seeking genuine answers.
- Outloud: Questions not directed toward anyone, without intent to
receive a response
- Other: Questions not fitting clearly into any category above.
(e.g., Challenge, Pre-invitation, etc.).
```

A.3 Imitate Prompt

```
You are a Reddit user browsing r/NoStupidQuestions.

Post Title: {title}
Post Body: {selftext}

Write a natural and authentic Reddit comment responding to this post.
- Use casual and informal language, like a normal Redditor.
- Avoid sounding like a chatbot or being overly formal.
```

B Automatic Annotation Prompt

The full prompt used for automatic annotation is provided below.

```
You are provided with multiple Reddit post and comment pairs. Each comment
↪  contains at least one question.
```

```
Your task is to process each pair by performing the following steps:
1. Extract all questions asked to the poster from the comments that meet
↪  these criteria:
   - Include questions if they meet any of the following criteria:
     - Are formal questions (grammatically marked as interrogative).
     - Are functional questions (aiming to elicit information, confirmation,
     ↪  or agreement), even if not formally interrogative.
     - Are newsmarks (e.g., "Really?", "Is it?", "Yeah?") as these typically
     ↪  seek confirmation.
   - Exclude questions if they meet any of the following criteria:
     - Solely seek acknowledgment during storytelling (e.g., "And it was a
     ↪  Weight Watchers recipe right?").
     - Appear only as reported speech (e.g., "Then he said, ‘Aren’t you gonna
     ↪  come over?’").

2. Using the full context (both the original post and the comment), assign
↪  each extracted question according to three criteria below. If a single
↪  comment includes multiple questions, categorize each one individually
↪  into a single category.

Social Action:
- Request for Information - Questions genuinely seeking information, with no
↪  other primary action apparent. (Ambiguous cases like "Are you busy
↪  tonight?"-a potential pre-invitation-should be coded as "Other /
↪  Pre-invitation’’.)
- Other Initiation of Repair - Questions including open-class repair
↪  initiators ("Huh?" or "What?") as well as partial repeats ("He went
↪  where?"). If it seemed that the repair was more a challenge than an
↪  initiation of repair, "Other / Challenge" should be coded.
- Request for Confirmation - Questions asserting a proposition and seeking
↪  confirmation (e.g., "So you're coming tomorrow night.").
- Assessment - Evaluations that are formatted to seek agreement such as
↪  "Isn’t it beautiful out today?" or "She’s such a pretty girl, isn’t she?"
- Suggestion/Offer/Request - Questions proposing, suggesting, offering, or
↪  explicitly requesting something (e.g., "Did you want some cereal?").
- Rhetorical Question - Questions that assert opinions or points, not seeking
↪  genuine answers (e.g., "Everything comes out in the wash, doesn't it?").
- Outloud - Questions not directed toward anyone, without intent to receive a
↪  response (e.g., "Now where are my keys?" while looking in a bag).
- Other / [Specific Social Action] - Questions not fitting clearly into any
↪  category above. Specify explicitly the social action performed after a
↪  slash ("/") (e.g., ‘‘Other / Challenge," ‘‘Other / Pre-invitation," etc.).

Example output:
[
  {
    "post_title": "Why do some people eat food directly from the container
    ↪  instead of using a plate?",
    "post_body": "I've noticed that certain individuals, when eating foods
    ↪  like ice cream, chips, or even leftovers, will consume it straight
    ↪  from the container rather than transferring it to a plate or bowl. Is
    ↪  there a specific reason for this behavior, or is it simply a matter
    ↪  of convenience or personal preference?",
    "comment": "Why dirty a plate when you can just eat straight from the
    ↪  container? Less dishes to wash, and sometimes it just feels more
    ↪  satisfying to dig into a tub of ice cream with a spoon. Plus, if
    ↪  you're just snacking or eating alone, who cares, right?",
    "extracted_questions": ["Why dirty a plate when you can just eat straight
    ↪  from the container?", "Plus, if you're just snacking or eating alone,
    ↪  who cares, right?"],
    "logical_semantic_structure": ["Polar Question", "Polar Question"],
    "through_produced_multi_question": "Yes",
    "social_action": ["Rhetorical Question", "Rhetorical Question"]
```

```
  },
  {
    "post_title": "Should I get pizza or wingstop today? ",
    "post_body": "I can't decide. I don't feel like leaving so I could
    ↪  doordash wingstop or I could get my ass up and go pick up a pizza
    ↪  since there is a domino's close by ",
    "comment": "Doesn’t domino’s have wings too? You can get both there.",
    "extracted_questions": ["Doesn’t domino’s have wings too?"],
    "logical_semantic_structure": ["Polar Question"],
    "through_produced_multi_question": "No",
    "social_action": ["Request for Confirmation"]
  },
  {
    "post_title": "What does cilantro even taste like to normal people?",
    "post_body": "I’m fairly certain I have that Cilantro-Soap gene you’ll
    ↪  see people talk about sometimes. It’s soapy, overly aromatic, and
    ↪  vile. But not ONCE have I seen a normal person describe the taste of
    ↪  cilantro. Why are you all obsessed with putting it in everything?
    ↪  What does it taste similar to?",
    "comment": "I hate when it overpowers the dish.",
    "extracted_questions": [],
    "logical_semantic_structure": [],
    "through_produced_multi_question": "N/A",
    "social_action": []
  }
]

Return the following JSON list, with the empty fields filled in for each JSON
↪  object.

  {
    "id": "{id}",
    "post_title": "{post_title}",
    "post_body": "{post_body}",
    "comment": "{comment}",
    "extracted_questions":
    "logical_semantic_structure":
    "through_produced_multi_question":
    "social_action":
  },
...
  {
    "id": "{id}",
    "post_title": "{post_title}",
    "post_body": "{post_body}",
    "comment": "{comment}",
    "extracted_questions":
    "logical_semantic_structure":
    "through_produced_multi_question":
    "social_action":
  }]
```

C Model and Human Generations Table

Field	**Content**
Post Title	Why do some people eat food directly from the container instead of using a plate?

Post Body	I've noticed that certain individuals, when eating foods like ice cream, chips, or even leftovers, will consume it straight from the container rather than transferring it to a plate or bowl. Is there a specific reason for this behavior, or is it simply a matter of convenience or personal preference?
GPT-4o	There are several reasons why people might choose to eat directly from the container instead of using a plate or bowl. Here are a few possibilities: 1. **Convenience**: Eating directly from the container can save time and effort. It eliminates the need to wash extra dishes, which can be appealing, especially if someone is in a hurry or doesn't want to deal with cleanup. 2. **Portion Control**: Some people might find it easier to control their portions by eating directly from the container. For example, with ice cream, they might feel they can stop eating when they feel satisfied rather than finishing a pre-scooped bowl. 3. **Casual Setting**: In informal or relaxed settings, people might feel more comfortable eating directly from the container. It can feel more laid-back and less formal than using a plate. 4. **Habit or Preference**: For some, it might simply be a habit or personal preference. They might enjoy the experience of eating directly from the container or find it more satisfying. 5. **Social Influence**: If someone sees others around them eating directly from containers, they might be more inclined to do the same. It can be a social norm in certain groups or situations. Do any of these reasons resonate with what you've observed, or is there a specific context you're curious about?
OLMO-instruct	Eating directly from the container seems to be a blend of convenience, habit, and individual preference. Here are a few reasons why some might opt for this method: 1. **Convenience:** Sometimes, the thought of washing extra dishes or taking the time to transfer food into another container just isn't appealing. Especially if someone is eating quickly or on-the-go, opting for the container is a time-saving choice. 2. **Portion Control:** For some, it might be a way to monitor how much they are eating, especially if the container has clear serving sizes marked on it. 3. **Taste Preference:** There's a belief that eating directly from the container can change the taste experience slightly due to the container's material or the way it maintains temperature (e.g., ice cream from a tub). 4. **Easier Access:** Containers, especially those designed for takeout or single servings, are often more portable and provide direct access to the food, which can be appealing. 5. **Mindless Eating:** This behavior can also be a part of mindless eating, where the individual is not fully conscious of how much they are eating because they are not seeing the food portioned out. 6. **Personal Habit:** For some, it could simply be a long-established habit that they grew up with or formed over time. It's important to note that occasionally eating from containers is generally harmless. However, making it a frequent habit might lead to overeating without recognizing it, as there's less visual cue to how much has been consumed. Furthermore, it's critical to wash containers properly to prevent the spread of germs and bacteria. So, while eating from the container is often a matter of convenience and personal preference, it's also worth considering how it aligns with one's eating habits and health goals. Would you like to delve further into any aspect of this topic?

OLMO

Assistant: This is an interesting observation! While the answer may vary from person to person, here are a few possible reasons why some people eat food directly from the container:
1. **Convenience**: It might be quicker and more convenient to eat straight from the container, especially if they are in a hurry or multitasking.
2. **Portion Control**: Eating directly from a container can help with portion control, particularly for foods that are often consumed in small servings, like nuts or candies.
3. **Sanitation**: Some people might believe that eating directly from the container is cleaner or less prone to contamination than using a plate or bowl.
4. **Personal Preference**: Ultimately, it might just be a personal preference or habit for some individuals. They might find it enjoyable or comfortable to eat straight from the container.
Do any of these reasons resonate with you, or are there other factors that you think might contribute to this behavior?

DeepSeek-R1

That's a great observation! Eating straight from the container is surprisingly common, and there are several reasons why people do it, often a mix of practicality and psychology:
1. **Convenience & Laziness (The Big One):** **Less Cleanup:** This is probably the top reason. Using a plate or bowl means washing an extra dish. Grabbing the tub, bag, or box and a spoon/fork means only cleaning that one utensil (or sometimes none, if it's finger food like chips). **Speed:** It's simply faster. No need to transfer food, especially when you just want a quick snack.
2. **Portion Control (or Lack Thereof):** **"Just a Taste" Mentality:** People often intend to have just a little bit. Pouring chips into a bowl feels like committing to a larger portion. Eating a few straight from the bag feels more contained (even if it often leads to eating more!). **Avoiding Over-Serving:** Conversely, for some *trying* to control portions, eating directly *can* be a (sometimes flawed) strategy. "I'll just have two spoonfuls of ice cream from the tub" avoids the temptation of scooping a huge bowl. (Though it often backfires!).
3. **Sensory & Experiential Factors:** **Texture/Temperature:** Ice cream straight from the tub can feel denser, colder, and more satisfying to some than when slightly melted in a bowl. Chips might stay crisper in the bag. **Mindless Eating:** Eating from a large container can facilitate zoning out while watching TV, reading, or working. The food is just *there*, requiring minimal attention. **Comfort/Indulgence:** There's a certain rebellious or comforting feeling about eating directly from the source, especially with treats like ice cream. It feels less formal, more indulgent.
4. **Context & Situation:** **Alone vs. With Others:** This behavior is *far* more common when eating alone. Doing it in front of others can be seen as rude or unhygienic (especially with shared containers). **Fatigue/Late Night:** When tired (e.g., late-night snack), the effort of getting a plate feels like too much. **"My House, My Rules":** When people are in their own private space, social norms relax. They feel free to do what's easiest/most comfortable for them.
5. **Habit & Upbringing:** For some, it's simply how they've always done it or how they saw family members do it. It becomes the default.
To make my answer even more relevant to you:
Is there a specific food or situation where you've noticed this most? (e.g., Is it always ice cream? Only when people are alone? With shared snacks?) **Are you asking out of curiosity, or is it something you do yourself and wonder about?** (Understanding your perspective helps tailor the answer!)
In essence: While convenience and laziness are major drivers, it's often layered with psychological factors like perceived portion control, sensory preference, and the comfort of informality, especially in private settings. It's usually not *just* one reason, but a combination that makes the path of least resistance (and fewest dishes) appealing!

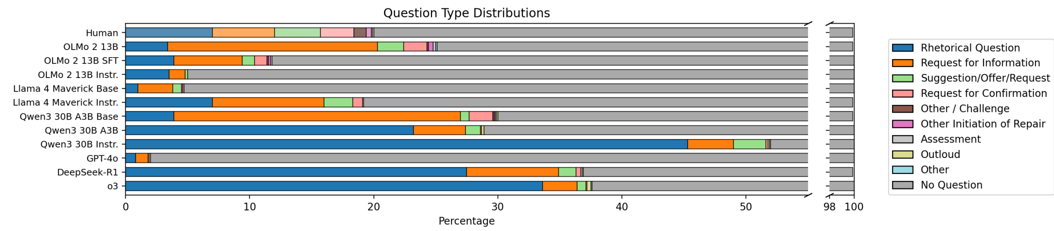


Figure 5: Distributions of question types produced by all LLMs under the naive prompt and the humans. Compared to humans, LLMs exhibit skewed distributions, often predominantly producing Rhetorical and Request for Information questions while underproducing suggestions, confirmations, and challenges.

Human	Perhaps an equally valid question is the inverse. Why do certain individuals transfer their food to a plate or a bowl before eating it? Is there a specific reason for this behavior, or is it simply a matter of convenience or personal preference?

Table 2: Questions generated by human and LLMs using Question Prompt specified in Appendix A.1

D Chi-squared Test Details

We compare each model's question-type distribution with the human's using a two-stage cluster-bootstrap Chi-squared test. We retain the six question types that reach $\geq 1\%$ frequency in any system—Request for Information, Request for Confirmation, Rhetorical Question, Suggestion/Offer/Request, Assessment, and Other / Challenge. Each of 10000 bootstrap replicates resamples the 3564 posts in the test-set with replacement; within every sampled post, we draw one response per system and tally the six labels. The resulting Monte-Carlo p-values are FDR-corrected (Benjamini–Hochberg). Cluster resampling controls within-post dependence among the responses.

E Question Type Distributions Across All Models

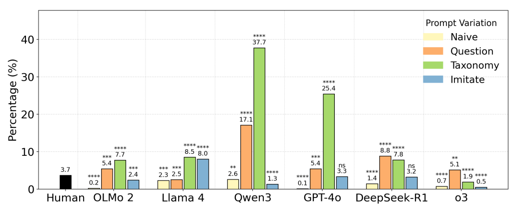
In Figure 5, we present the full distributions of question types generated by all evaluated LLMs under the naive prompt and the humans. This expanded view reinforces our earlier findings that LLM-generated questions exhibit a more skewed distribution compared to humans.

F Additional Question Type Comparisons Across Prompting Conditions

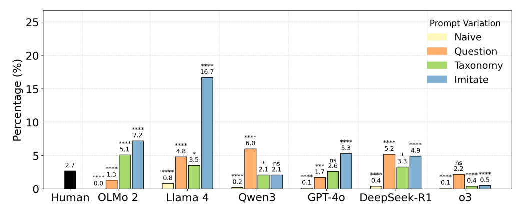
Figure 6 presents additional comparisons of question type distributions between humans and the LLMs across the various prompting techniques. These comparisons focus on Suggestion/Offer/Request, Request for Confirmation, Assessment, and Other/Challenge questions, which were not covered in the main figure.

G Human and LLMs Question Features

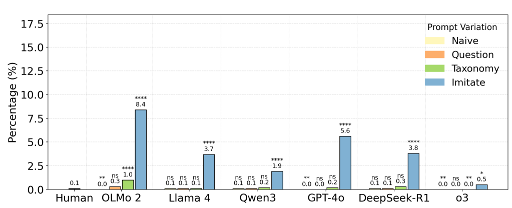
Using HypotheSAEs [Movva et al., 2025], we extracted interpretable features that are predictive of the variable whether the question is generated by human (1) or a language model (0). We used train, validation, and held-out sets each with 70%, 25%, and 5% of all questions generated by humans and all language models to train two SAEs with ($M = 32, k = 4$) and ($M = 256$, $k = 8$), where M is a total number of concepts that can be learned from entire dataset and k is the number of instances that can be used to represent each concept. Higher M and k leads to more granular concepts. We show top 20 concepts with highest absolute regression coefficients for both Rhetorical Questions (Table 3) and Information Requests (Table 4). The positive coefficients represent concepts that are predictive of humans and negative coefficients represent concepts that are predictive of language models.



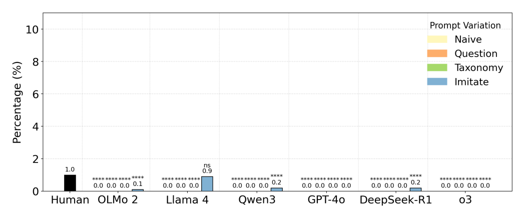
(a) Suggestion/Offer/Request



(b) Request for Confirmation



(c) Assessment



(d) Other / Challenge

Figure 6: Additional comparisons of question type distributions between humans and LLMs across different prompting conditions. Statistical significance of differences between human and model outputs is indicated above each bar, based on two-sided paired t-tests with Benjamini–Hochberg correction.

Hypothesis	Sep. Score	Sep. p	Reg. Coef.	Reg. p	Feat. Prev.
asks whether the user is testing a riddle	0.65	0.00	2.5	0.048	0.0077
uses dismissive language to indicate indifference or lack of concern (e.g., 'Who cares', 'Who the fuck knows')	0.58	0.00	2.7	**0.00**	0.046
asks whether the user is trying to get a laugh out of people	0.53	0.00	2.2	**0.00**	0.046
asks whether the user brought up unnecessary topics	0.46	0.00	1.9	**0.00**	0.043
asks whether the user has ingested something potentially harmful	0.39	0.0002	1.6	**0.0045**	0.0086
asks about speed or driving behavior	0.36	0.0001	1.4	**0.0023**	0.012
asks whether the reader disagrees with the moderators	0.35	0.00	1.4	**0.00**	0.042
contains the word 'that' as a central focus in the question	0.33	0.00	1.3	**0.00**	0.063
asks whether something is a cause for concern or leads to problems	0.058	0.0037	0.31	0.018	0.40
asks a concise question consisting only of 'why' or 'why not'	0.0002	1.0	0.13	0.52	0.085
asks a rhetorical or hypothetical question that challenges an assumption or prompts introspection	-0.033	0.19	-0.23	0.12	0.81
asks about a woman's behavior, emotions, or level of engagement in a social interaction	-0.036	0.25	-0.22	0.33	0.11
asks the user to engage in a test or challenge	-0.062	0.0034	-0.44	**0.0012**	0.31
asks about personal interests, preferences, or activities	-0.062	0.046	0.12	0.54	0.11
uses quotation marks around words or phrases to emphasize or question meaning	-0.12	0.00	-0.69	**0.00**	0.25
asks about personal or mental health concerns	-0.13	0.00	-0.40	0.045	0.16
asks about creating or improving tools, systems, or intelligence	-0.14	0.0088	-0.74	0.031	0.034
asks about the difference between things	-0.15	0.00	-0.85	**0.00**	0.26
asks about emotional states or feelings of self-fulfillment	-0.19	0.00	-0.95	**0.0028**	0.075
begins with 'So' and asks what actions can be taken or what can be done	-0.22	0.00	-1.7	**0.00**	0.065

Table 3: Hypotheses distinguishing human- (1) vs. model-generated (0) Rhetorical questions with regression and separation results. AUC = 0.77; significant hypotheses: 12/20 ($p < 5 \times 10^{-3}$).

Hypothesis	Sep. Score	Sep. p	Reg. Coef.	Reg. p	Feat. Prev.
asks 'why' in a concise and direct manner, often consisting of only the word 'why' or very short phrases including 'why'	0.43	0.00	1.73	**0.0001**	0.0648
asks about a country or location	0.40	0.00	1.12	0.025	0.044
asks about race, ethnicity, or identity-related topics	0.29	0.0051	1.38	0.034	0.0276
asks about a specific entity or group, often related to identification or origin (e.g., 'Who', 'What country', 'Which country')	0.19	0.00	0.46	0.035	0.21
is a single word or short phrase followed by a question mark	0.073	0.067	-0.0098	0.96	0.25
asks a concise and open-ended 'why' question with minimal context or elaboration	0.071	0.091	-0.14	0.56	0.21
asks for clarification or disambiguation about a specific term or concept	-0.011	0.74	-0.10	0.59	0.50
asks about the characteristics, actions, or states of 'they'	-0.014	0.78	0.15	0.56	0.14
asks for the meaning or interpretation of a term or phrase, often emphasizing the respondent's personal perspective	-0.042	0.25	0.18	0.40	0.37
asks for an example or examples	-0.14	0.0009	-0.10	0.63	0.23
asks about the physical alignment, condition, or functionality of objects	-0.15	0.0094	-0.62	0.033	0.098
asks about the current state, location, or condition of the user	-0.19	0.00	-0.40	0.046	0.29
asks about long-term planning or future-oriented decision-making	-0.20	0.0002	-0.34	0.25	0.11
asks about the nature, goals, or future of a personal relationship	-0.23	0.00	-0.13	0.60	0.18
asks about time management or activities involving time usage	-0.23	0.0006	-0.53	0.14	0.071
asks whether a specific issue is causing interference or impact in life responsibilities, relationships, or emotions	-0.26	0.00	-0.37	0.14	0.24
asks about stylistic or aesthetic preferences	-0.27	0.00	-0.39	0.16	0.17
asks about the user's personal preferences or interests in terms of hobbies, activities, or content	-0.29	0.00	-1.08	**0.0005**	0.13
asks about emotions or feelings explicitly	-0.34	0.00	-1.03	**0.0003**	0.20
asks the user to evaluate or compare options in terms of improvement, preference, or alignment with values/goals	-0.36	0.00	-0.87	**0.0031**	0.17

Table 4: Hypotheses distinguishing human- vs. model-generated Information Requests with regression and separation results. AUC = 0.771; significant hypotheses: 4/20 ($p < 5 \times 10^{-3}$).