Do LLMs Find Human Answers To Fact-Driven Questions Perplexing? A Case Study on Reddit

Parker Seegmiller, Joseph Gatto, Omar Sharif
Madhusudan Basak, Sarah Masud Preum

Department of Computer Science
Dartmouth College
pkseeg.gr@dartmouth.edu

Abstract

Large language models (LLMs) have been shown to be proficient in correctly answering questions in the context of online discourse. However, the study of using LLMs to model *human-like* answers to fact-driven social media questions is still under-explored. In this work, we investigate how LLMs model the wide variety of human answers to fact-driven questions posed on several topic-specific Reddit communities, or subreddits. We collect and release a dataset of 409 fact-driven questions and 7,534 diverse, human-rated answers from 15 r/Ask{Topic} communities across 3 categories: profession, social identity, and geographic location. We find that LLMs are considerably better at modeling highly-rated human answers to such questions, as opposed to poorly-rated human answers. We present several directions for future research based on our initial findings.

Introduction

Large language models (LLMs) have been used for several social computing tasks, such as sentiment analysis (Deng et al. 2023), content moderation (Kolla et al. 2024), and question answering (Xiong et al. 2019) on social media, with varying degrees of success. It is important to characterize the extent to which LLMs are in line with human preferences on such tasks. This characterization involves several subtasks, including identifying social media scenarios in which LLMs can generate human-like content, assessing the factuality of LLM-generated content in such scenarios, and determining whether LLMs' capacity to model online discourse is in line with human preferences.

Through the lens of exchanging fact-driven information on social media, we examine a vast source of social media question-answering data, namely 15 r/Ask{Topic} communities on Reddit, such as r/AskMen and r/AskNYC. Reddit is a pseudo-anonymous social media website, enabling users to share rich personal content on a wide variety of topics. Users can participate in several communities, each devoted to one specific topic, including complex and uncertain topics. Specifically, r/Ask{Topic} communities allow millions of users to pose and answer topic-specific questions to a community of peers interested in those topics.

In this study, we are interested in whether LLMs can model *fact-driven* questions posed on these subreddits. As a motivating example, consider this question from r/AskHistorians: "Why did Spain not join either the Axis or Allied powers in WWII?" Despite having a factually-grounded, historical answer, there is a wide range of plausible human answers to this question, all of which can be considered "correct". Prior work has explored the capacity of LLMs to answer fact-driven questions correctly through autoregressive generation (Wang et al. 2023). Under-explored, however, is analysis of LLM capacity to intrinsically model the breadth of human answers to fact-driven questions. In this work, we thus aim to answer questions such as (i) Do LLMs find human-written answers to fact-driven questions perplexing? and (ii) Does LLM perception of human-written answers correlate with human ratings? Answering such questions will scope the capacity of LLMs as social media text generation agents from a novel perspective. We investigate these questions through the lens of Reddit's r/Ask{Topic} subreddits and make the following contributions.

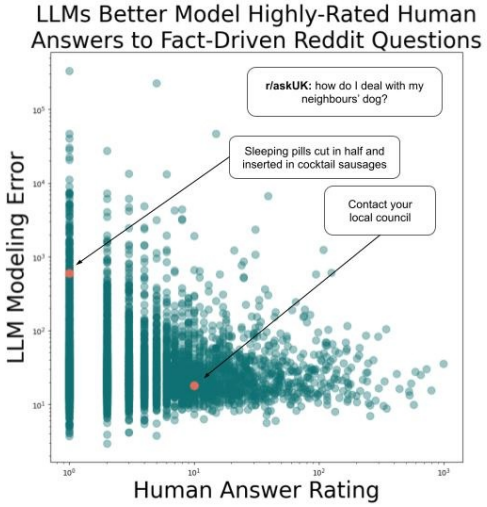

Figure 1: LLM (fine-tuned Sheared LLaMA 1.3B) modeling error (perplexity) and human ratings of 7,534 human answers to fact-driven questions posed on Reddit's r/Ask{Topic} communities in log scale. In general, LLM modeling and human perception are well-aligned. As an example, see the two divergent answers to the question asked on r/UK. The LLM assigns low perplexity to the highly-rated human answer, and higher perplexity to the low-rated human answer.

Subreddit	Objectivity	Question
Historians	Fact-Driven	Why did Spain not join either the Axis or Allied powers in WWII?
Academia	Subjective	Should I do my PhD and Postdoc in the same university?
OldPeople	Fact-Driven	How different were Bars/Pubs before smart phones?
Women-Over30	Subjective	Who makes between $70-80K? Age, occupation and how you got there.
UK	Fact-Driven	Why do we have smaller houses than most of the world?
NYC	Subjective	Does the awesome city make you care less about the weather?

Table 1: Examples of Fact-Driven and Subjective Questions from the r/Ask{Topic} Subreddit Communities.

1. We develop a novel generalizable framework to characterize how LLMs model the wide variety of human responses to fact-driven questions on Reddit.
2. We curate a dataset of 409 fact-driven questions and 7,534 answers, along with 494 and 8,177 non-fact-driven questions and answers, gathered from 15 of Reddit's largest r/Ask{Topic} communities covering a wide variety of profession, geographic location, and social identity topics.
3. We provide an intrinsic analysis of different LLM settings on this task by comparing LLM perplexity and human answer ratings, finding that LLMs are considerably better at modeling highly-rated human answers, as opposed to poorly-rated human answers. We also discuss implications of these findings for future research.

Our analysis framework, dataset, and results provide a rich source of hypothesis-generating resources for downstream NLP, social science, and web-based research.

Related Works

Social Media Question Answering

Despite being a well-explored research question, there is limited focus on question answering (QA) using social media data, attributed to its distinctive challenges, including hashtags, typos, platform-specific phenomena, and the informal nature of the text (Rogers, Gardner, and Augenstein 2023). Xiong et al. (2019) develop a question-answering dataset by crawling tweets and writing question-answer pairs from the tweet. Other researchers have explored Reddit question and comment answer threads for relevant tasks (Lanius 2019; Hew et al. 2024). Recent studies have demonstrated the effectiveness of LLMs in solving the QA task (Kamalloo et al. 2023). Staab et al. (2023) showed that LLMs can infer many private details about Reddit users via posts and comments in r/Ask subreddits, indicating the need for broader privacy protections beyond LLM text memorization. Kolla et al. (2024) illustrated how LLMs are bad at specific rules-related content moderation using data from five different subreddits. However, there is a limited exploration into how effectively LLMs model human responses to questions posed on social media platforms.

Factual Question Answering

The key attribute of a QA dataset is its communicative intent: information-seeking and probing. Information-seeking questions aim to obtain factual responses while probing questions are more subjective and context-dependent. While most QA research emphasizes generating accurate factual responses across diverse domains (Zhang et al. 2023), limited attention has been paid to modeling the responses generated by QA models. Lin, Hilton, and Evans (2022) introduced TruthfulQA to assess the likelihood of models replicating human falsehoods. Their findings revealed that language models generate false responses resembling misconceptions to deceive humans. The advent of LLMs significantly transformed the QA paradigm due to their capacity to offer creative responses to questions. Therefore, it is crucial to understand how much LLM-generated responses model humans' answers. To address this, we present a new dataset and investigate how LLMs model a wide variety of human answers to fact-driven questions posted on Reddit.

Background and Methodology

Definitions

Fact-Driven Questions We are particularly interested in determining how LLMs model human answers to *fact-driven* questions on social media. By fact-driven questions, we mean posts which contain a question with factually-grounded answers. We consider each post on Reddit's r/Ask{Topic} communities to either be fact-driven, or *subjective* (non-fact-driven), as exemplified in Table 1. While this binary formulation of fact-driven is a simplification of a complex idea, classifying posts in this manner enables us to filter out questions which would likely require human experience to answer.

Answer Score Each comment is assigned, via Reddit, a score based on peer engagement and human preference. This peer-assigned score is a measure of the upvote to downvote ratio normalized by comment age.[1] We use this measure to proxy peer perception. Highly-rated answers are deemed more valuable or helpful, while low-rated ones are seen as less helpful, aligning with prior research on social media engagement (Sharma et al. 2020; Trunfio and Rossi 2021).

Perplexity To determine whether an LLM models a human answer well, we use perplexity (Jelinek et al. 1977). Intuitively, perplexity measures how "surprised" the LLM is by an answer. If the perplexity is low, the LLM is expecting to see this answer, meaning the LLM can model the autoregressive linguistic properties of the answer. This metric is commonly used to measure performance of language models in encoding linguistic phenomena (Belinkov and Glass 2019). We use perplexity to measure how well an LLM models human answers to fact-driven questions asked on Reddit's r/Ask{Topic} subreddits. Given a human answer a comprised of a sequence of English tokens $a = [a_1, \cdots, a_{|a|}]$, an LLM M assigns autoregressive conditional token probabilities $p_M(a_i|a_1 \cdots a_{i-1})$ to each token a_i in answer a.

[1] https://www.reddit.com/wiki/faq/

Category	Type	#Q	#A	%F
Profession	Historians	35	164	42%
	Academia	13	72	37%
	Engineers	27	394	42%
	Culinary	42	302	45%
	Photography	21	168	44%
Identity	Men	15	332	50%
	Women	100	2925	47%
	MenOver30	7	180	32%
	WomenOver30	9	268	31%
	OldPeople	7	245	58%
Geographic	UK	67	1684	48%
	NYC	49	392	51%
	France	7	273	47%
	Singapore	8	132	47%
	Argentina	2	11	50%

Table 2: Final dataset statistics of the questions and answers gathered from 15 r/Ask{Topic} subreddits across 3 topic categories. Here, #Q, #A, and %F indicate the number of fact-driven questions, the number of answers to those questions, and the percentage of filtered questions which were deem fact-driven.

The perplexity of an answer a under language model M is defined as follows.

$$PPL_M(a) = exp\{-\frac{1}{|a|}\sum_{i}^{|a|}\log p_M(a_i|a_1 \cdots a_{i-1})\}$$

Perplexity is the exponentiated average negative log-likelihood of tokens in a sequence, and can be thought of as a measure of LLM modeling error over a sequence.

Relevant Data Collection and Processing

Some of the largest communities on Reddit are the r/Ask{Topic} communities, including millions of members who engage in topic-specific discussions with peers, e.g., r/AskWomen has 5.5 million members. We collect fact-driven question-answers from these communities using the following steps.

First, 15 r/Ask{Topic} communities across three categories are selected for investigation, as seen in Table 2. We select these popular subreddits to capture diverse content, enabling us to investigate fact-driven questions with high peer engagement (captured as numbers of upvotes and comments).

Second, a random sample of 50% of all posts and comments, dating back to 2011, are collected from each of these communities. This represents several million such posts and comments. We filter to exclude deleted or removed posts and comments, or exceptionally long posts and comments ($> 10,000$ characters) since they often contain personal narratives rather than fact-driven questions.

Third, posts with less than three top-level comments are then excluded, as we wish to examine questions with high peer engagement. We consider top-level comments to be answers to the primary question in each post.

Fourth, we then exclude posts whose comments all have the same rating. As we wish to explore LLM modeling capability as it correlates with human perception, we are primarily interested in questions whose answers have varying human ratings. When top-level comments all have the same rating, this typically indicates that the post received less peer engagement and is therefore less suitable for our study. The filtering thus far results in a sample of 903 questions (posts) and 15,711 answers (top-level comments).

Finally, we filter to consider only fact-driven questions. We do this by passing each question to the large, open-source LLaMA-2-70B model (Touvron et al. 2023), asking whether the post contains a question with a factually-grounded answer. Filtering social media posts using LLMs is common in social media analysis (Kikkisetti et al. 2024; Gatto et al. 2024). To verify that LLaMA-2-70B performs this fact-driven question binary classification task intuitively, we select a sample of 100 posts and perform inter-annotator agreement, with each post being labeled by two annotators. Each annotator is familiar with Reddit and has vast experience in social computing and factuality assessment tasks. Two human annotators agree with each other on 83% of posts. Both human annotators agree with LLaMA-2-70B on 41% of posts, and at least one agreed on 58% of posts. We attribute these inter-annotator agreement scores to the limitation of treating questions as either fully fact-driven or fully subjective, when in reality there exists a range of subjectivity. However, we deem this classification to be a better alternative to considering *all* posts, as such posts vary widely in context and scope.

This process results in a dataset of 409 *questions* (posts containing fact-driven questions) from 15 r/Ask{Topic} across 3 categories, and 7,534 *answers* (top-level comments containing answers) to those questions. We release all filtered questions and answers, including those that are not deemed fact-driven by LLaMA-2-70B, and leave further analysis of this filtering for future work.

Models

To evaluate LLM modeling of answers to questions posed on Reddit's r/Ask{Topic} subreddits, we utilize Sheared LLaMA, a light-weight version of the open-source LLaMA2-7B LLM which is pruned to only 1.3 billion parameters using target structured pruning (Xia et al. 2023). We employ Sheared Llama as opposed to it's larger counterpart Llama2-7B as the later requires significant computational resources to fine-tune. Thus, using Sheared Llama allows us to (i) run experiments where we fine-tune an LLM on in-domain data and (ii) share a model that is more useful to those in the community with limited access to computational resources. To compare two common inference scenarios with LLMs, we consider both settings: (i) **SL**: The out-of-the-box pre-trained Sheared LLaMA 1.3B model, and (ii) **SLFT**: A version of the Sheared LLaMA 1.3B model, fine-tuned on a curated dataset of 100,000 comments from r/AskReddit [2]. This model is fine-tuned using the Hugging-

[2]https://tinyurl.com/y6r68hkv

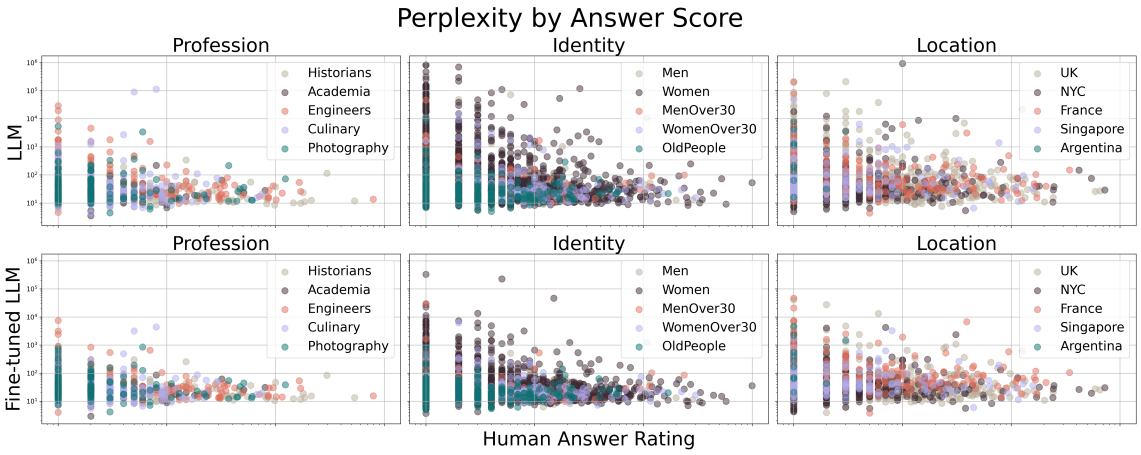

Figure 2: LLM perplexity of answers to fact-driven questions posed on 15 of Reddit's r/Ask{Topic} communities, compared with the peer-assigned score of answers. The perplexities of the top row are calculated by the vanilla *SL* LLM, and the bottom row are calculated the fine-tuned *SLFT* LLM. For each graph, the X and Y axes indicate peer-assigned scores and LLM's perplexity in log scale, respectively.

face Transformers library[3] on an A100 GPU for 1 epoch, using a max token length of 128, a learning rate of $2e^{-5}$, and a batch size of 6.

Results and Implications

Figure 1 displays the result of our analysis. By plotting all scores of the human answers by the answer perplexity assigned by the *SLFT* model, we can see a general trend. Across all 7,534 answers, the LLM models highly-rated answers better than poorly-rated answers. This indicates that the LLM is in line with human perception: it is probabilistically more likely to understand quality answers to fact-driven social media questions posed on r/Ask{Topic} communities.

In Figure 2, we see these same results of our analysis, separated by topic and model. Our fine-tuning strategy decreases LLM's perplexity across the board, indicating that fine-tuning helps the model understand human answers better. However, the general trend of the LLM modeling highly-rated answers better than poorly-rated answers exists for both the fine-tuned and vanilla LLMs. Similarly, we see the same trend across all individual subreddits from each of the three categories. We also find that LLMs models answers to fact-driven questions from *professional* subreddit topics better than questions from *social identity* or *location* topics.

Implications for Future Work: In addition to the above-mentioned results, we highlight interesting insights surfacing from the data by considering different combinations of peer-assigned score of the answer and the level of LLM's perplexity for that answer.

High Peer-assigned Score, High Perplexity: LLMs fail to model these types of answers, and yet they are rated highly by humans. As an example of this phenomenon, take the r/AskNYC question "We have larger issues, but would anyone like to talk about silverfish?" A highly-rated answer "They are awful and it is alarming how fast they move" has a high perplexity under the *SLFT* model of 113.8, indicating the model hasn't learned the kind of language used in the answer even though the community found the answer appropriate for the question. Further exploring these types of answers might identify underrepresented answers compared with the "typical" answers found in LLM pre-training/fine-tuning data as well as potential blindspots of LLMs.

Low Peer-assigned Score, Low Perplexity: LLMs intuitively model these types of answers. However, even though they are in response to questions with high engagement, these types of answers are rated poorly by humans. Take this question from r/AskWomen "Ladies, what's a cheap place to buy shoes?" In response to this question, the answer "Honestly I'd rather have a few pairs of quality shoes than lots of cheap ones" may seem fairly reasonable; indeed, this answer is well-understood by the *SLFT* model with a low perplexity of 7.8. However, this answer was poorly-rated, perhaps being seen as a divergent opinion from community consensus. Further exploration into these types of answers may lead to interesting analysis of divergent opinions on social media which may seem normal on a surface level.

Future work can leverage this data and framework to explore targeted fine-tuning of LLMs for socio-technical tools. Downstream work could use these tools to investigate similar questions in other domains and subreddits. In addition to these, we imagine several other hypotheses could be investigated with the data and analysis framework we present in this paper. Given the limited scope of this work, we leave the investigation of potential data contamination in the LLM training to future work. We also hope to investigate the use of multiple LLMs, as well as including question content as part of the perplexity analysis. We release our filtered dataset of these r/Ask{Topic} questions and answers[4], both fact-driven and subjective, as a rich source of hypotheses for future social science and web research.

[3] https://huggingface.co/

[4] https://huggingface.co/datasets/pkseeg/reddit-ask-v0

References

Belinkov, Y.; and Glass, J. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7: 49–72.

Deng, X.; Bashlovkina, V.; Han, F.; Baumgartner, S.; and Bendersky, M. 2023. What do llms know about financial markets? a case study on reddit market sentiment analysis. In *Companion Proceedings of the ACM Web Conference 2023*, 107–110.

Gatto, J.; Basak, M.; Srivastava, Y.; Bohlman, P.; and Preum, S. M. 2024. Scope of Large Language Models for Mining Emerging Opinions in Online Health Discourse. *arXiv preprint arXiv:2403.03336*.

Hew, J.; Horne, Z.; Corley, M.; and Tarighat, A. 2024. Examining structural and semantic predictors of announced sarcasm on r/AskReddit.

Jelinek, F.; Mercer, R. L.; Bahl, L. R.; and Baker, J. K. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63.

Kamalloo, E.; Dziri, N.; Clarke, C.; and Rafiei, D. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5591–5606. Toronto, Canada: Association for Computational Linguistics.

Kikkisetti, D.; Mustafa, R. U.; Melillo, W.; Corizzo, R.; Boukouvalas, Z.; Gill, J.; and Japkowicz, N. 2024. Using LLMs to discover emerging coded antisemitic hatespeech emergence in extremist social media. *arXiv preprint arXiv:2401.10841*.

Kolla, M.; Salunkhe, S.; Chandrasekharan, E.; and Saha, K. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?

Lanius, C. 2019. Torment Porn or Feminist Witch Hunt: Apprehensions About the #MeToo Movement on /r/AskReddit. *Journal of Communication Inquiry*, 43(4): 415–436.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Dublin, Ireland: Association for Computational Linguistics.

Rogers, A.; Gardner, M.; and Augenstein, I. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Comput. Surv.*, 55(10).

Sharma, A.; Choudhury, M.; Althoff, T.; and Sharma, A. 2020. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 614–625.

Staab, R.; Vero, M.; Balunovic, M.; and Vechev, M. 2023. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trunfio, M.; and Rossi, S. 2021. Conceptualising and measuring social media engagement: A systematic literature review. *Italian Journal of Marketing*, 2021(3): 267–292.

Wang, C.; Liu, X.; Yue, Y.; Tang, X.; Zhang, T.; Jiayang, C.; Yao, Y.; Gao, W.; Hu, X.; Qi, Z.; et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Xia, M.; Gao, T.; Zeng, Z.; and Chen, D. 2023. Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning. In *The Twelfth International Conference on Learning Representations*.

Xiong, W.; Wu, J.; Wang, H.; Kulkarni, V.; Yu, M.; Chang, S.; Guo, X.; and Wang, W. Y. 2019. TWEETQA: A Social Media Focused Question Answering Dataset. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5020–5031. Florence, Italy: Association for Computational Linguistics.

Zhang, Q.; Chen, S.; Xu, D.; Cao, Q.; Chen, X.; Cohn, T.; and Fang, M. 2023. A Survey for Efficient Open Domain Question Answering. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14447–14465. Toronto, Canada: Association for Computational Linguistics.