On The Truthfulness of 'Surprisingly Likely' Responses of Large Language Models

Naman Goel[1]

[1] naman.goel@cs.ox.ac.uk

University of Oxford, UK

Abstract

The surprisingly likely criterion in the game-theoretic information elicitation literature rewards agents to give truthful answers. We investigate the relevance of a similar criterion for the responses of large language models (LLMs). We hypothesize that if the surprisingly likely criterion works in large language models, under certain conditions, the responses that maximize the reward under this criterion should be more accurate than the responses that only maximize the posterior probability. Using benchmarks including the TruthfulQA benchmark and using openly available LLMs: GPT-2 and LLaMA-2, we show that the method indeed improves the accuracy significantly (for example, up to 24 percentage points aggregate improvement on TruthfulQA and up to 70 percentage points improvement on individual categories of questions).

1 INTRODUCTION

Recent demonstrations of the promising capabilities of large language models (LLMs) have raised hopes about their use in a wide range of useful applications. However, one major issue with state-of-the-art LLMs that casts doubts on this optimism is their tendency to generate factually incorrect text. There are various ongoing efforts to address this issue. Promising efforts include scaling [Wei et al., 2022a], retrieval augmentation/grounding [Lewis et al., 2020], in-context learning [Brown et al., 2020], chain-of-thought reasoning [Wei et al., 2022b] and other prompting, self-consistency [Wang et al., 2023], advanced decoding techniques [Chuang et al., 2023], uncertainty measurement based techniques [Xiao and Wang, 2021], self-improvement [Pan et al., 2023], shifting model activations during inference [Li et al., 2023], locating and editing factual information [Meng et al., 2022], learning from human feedback [Ouyang et al., 2022], learning from AI feedback [Bai et al., 2022, Lee et al., 2023, Shinn et al., 2023, Weng et al., 2022] etc. While very useful, none of these efforts have fully solved the problem; perhaps because there are multiple reasons causing the problem [Ye et al., 2023] and therefore, multiple complementary solutions are required. For example, LLMs may generate incorrect information if correct information is not present in the training data or because correct information can not be inferred from the training data, and other similar reasons related to limitations of data and/or model. However, it is also possible that in some cases, LLMs provide incorrect information due to misspecified training objective or other misspecified reward. We call this behavior non-truthful. For example, LLMs are commonly trained to generate answers that have maximum probability conditioned on the prompt [Prince, 2023]. The probability calculated by LLM depends on the data points in its noisy training set. An interesting demonstration is due to Lin et al. [2022], who showed that LLMs generated false answers that mimic popular misconceptions from human texts. For example, when asked "Which forbidden fruit Adam and Eve eat in the Garden of Eden?", an LLM may answer that Adam and Eve eat an apple, which is a popular misconception; the correct answer is that the Bible does not specify the fruit. The research question is whether this behavior can be avoided to some extent by designing a better reward, i.e. can we make LLMs more truthful by rewarding them in some other way, instead of rewarding to generate text with maximum probability conditioned on the prompt?

In this paper, we investigate whether and to what extent, a surprisingly likely (a.k.a. surprisingly common) criterion inspired from the game-theoretic truthful information elicitation literature can be useful in this regard. There is a significant literature on game-theoretic incentive mechanisms for eliciting information from rational agents. The mechanisms are collectively referred to as the **peer-prediction** or peer-consistency mechanisms [Faltings, 2023].

A seminal contribution in the peer-prediction literature is due to Prelec [2004]. Prelec designs a scoring mechanism

Under Review.. Please cite the arxiv version: https://arxiv.org/abs/2311.07692.

(called the Bayesian Truth Serum) such that an agent answering a question, can maximize its expected score by telling what the agent believes to be the correct answer, instead of telling what it believes most of the other agents would tell. The key idea in the Bayesian Truth Serum (BTS) is to reward surprisingly common answer (i.e. the answer that is more commonly reported than predicted apriori by agents, instead of the answer that is merely the most commonly reported one). Interesting examples, where BTS is particularly useful, include eliciting objective information that is rare or difficult to obtain and subjective information (e.g. subjective opinions) that is impossible to verify. Theoretical guarantees apply more generally beyond these examples. Beyond game-theoretical incentive guarantees, it has also been shown that an information aggregation method motivated by the surprisingly likely criterion selects more correct answers [Prelec et al., 2017] in crowdsourcing. Later work in this area used motivation from the surprisingly likely criterion of BTS for designing peer-prediction mechanisms for various settings. We discuss about these later works in detail in Section 2.1.

However, the application of these game-theoretic peer-prediction mechanisms for information elicitation has not been explored yet for large language models; perhaps because there are many differences between LLMs and the agents as modeled in the game-theoretic truthful information elicitation and aggregation literature. But, as we will discuss in Section 3, there are some similarities too in the two settings that motivate us to explore the connection. In this work, we restrict the discussion on "truth" to objective information that can be clearly categorized as correct or incorrect (please see Section 5 for critical discussion on the scope of this work). We empirically investigate a few different ways of measuring a surprisingly likely inspired criterion for LLMs. Through experiments on TruthfulQA, COPA and StoryCloze benchmarks, we show that surprisingly likely answers are indeed more correct and this can be used to significantly improve the accuracy of responses of LLMs. For example, on the TruthfulQA benchmark, we find a consistent pattern of significant gains in accuracy across different large language models (up to 24 percentage points). We also analyze the performance by different categories of questions and observe that the trend of significant accuracy gains holds across most (but not all) categories, with some categories showing up to 70 percentage points improvement.

2 BACKGROUND AND RELATED WORK

2.1 SURPRISINGLY LIKELY CRITERION IN GAME THEORY

The problem of information elicitation without verification (i.e. when correct information is not available for scoring and agents have to be incentivized for providing correct information) is modeled as a Bayesian game between multiple agents in the literature [Faltings and Radanovic, 2022].

One of earlier incentive mechanisms in this space is known as the Bayesian Truth Serum [Prelec, 2004]. BTS asks every agent to submit two reports for a question. The first report is what agents believe is the correct answer to the question and in the second report, the agents predict the distribution of answers given by other agents. The reward of an agent is the sum of two reward terms. The first term (called the information score) measures the log of the ratio between the frequency of the reported answer and the geometric mean of the predictions about the answer. The second term (called prediction score) measures how close is the prediction of the agent about the distribution of other agents' answers to the actual distribution. The second term exists only to reward correct second report, but it is the first term (the information score) that is the interesting one. The resulting reward is shown to be incentive-compatible i.e. in equilibrium, agents can maximize their reward by telling what they believe is the correct answer, instead of answering non-truthfully.

Motivated by the BTS, a number of other reward/scoring mechanisms have been proposed in the literature. In particular, many of the proposals were designed for crowdsourcing settings without requiring the agents to explicitly submit their prediction about other agents' answers. Consider for example, a crowdsourcing task of measuring pollution at places and times, where no independent ground truth measurement exists. Multiple independent agents (peers) measure and report the value to a center, but there is no trusted verification of the ground truth. Agents have to exert effort to accurately measure the value, and the center needs to provide a reward to compensate for it. Peer-prediction mechanisms [Faltings and Radanovic, 2022] consider this setting as a Bayesian game among the agents, where each tries to maximize the reward attributed to their report. The simplest form is output agreement [Von Ahn and Dabbish, 2004], where reports are rewarded proportionally to the frequency of the same report among peers. However, it has been shown that the best strategies for the participating agents are always uninformative, e.g. all report the same value [Jurca et al., 2007]. This can be corrected by scaling the reward by the rarity of the answer among similar questions (apriori similar place/time in the pollution measurement example), by dividing the reward or subtracting the frequency of the answer [Jurca and Faltings, 2011, Faltings et al., 2017, Radanovic et al., 2016].

Various mechanisms exist with different game-theoretic arguments but they all amount to the same "surprisingly common" criterion. In a recent survey, Faltings [2023] identifies three types of mechanisms in this space. The first is agreement, where the reward is proportional to the frequency of the answer among other responses, often calibrated by the overall likelihood of agreement [Von Ahn and Dabbish, 2004, Dasgupta and Ghosh, 2013, Shnayder et al., 2016]. The second is information-theoretic, where the reward is

proportional to the pairwise mutual information between the answer and the answers given by peers [Prelec, 2004, Kong and Schoenebeck, 2019, Radanovic and Faltings, 2015, Goel and Faltings, 2020]. The third computes the reward based on the improvement in the quality of the resulting model (the Peer Truth Serum) [Jurca and Faltings, 2011, Faltings et al., 2017, Radanovic et al., 2016].

While we referenced several works above for completeness, for the purpose of this paper, it is sufficient to remember the following: 1) the surprisingly likely criterion, as originally proposed, required agents to submit two reports (i.e. their own answer and their prediction of other agents' answer), 2) later peer-prediction mechanisms did not ask agents to submit a separate prediction of other agents' answer. These mechanisms instead relied on other ways to estimate the prior frequency (probabilistic belief) that the surprisingly likely criterion requires, 3) we empirically show in this paper how to use a similar idea effectively for LLMs.

2.2 PMI IN COMPUTATIONAL LINGUISTICS

The information-theoretic measure of pointwise mutual information (PMI) [Fano, 1961] is a well-known concept in computation linguistics and natural language processing literature [Jurafsky and Martin, 2000]. It has been used in use-case such as words association [Church and Hanks, 1990], keyword generation improving diversity of text [Mou et al., 2016, Yao et al., 2017, Zhou et al., 2019, Tang et al., 2019, Takayama and Arase, 2019], increasing agreement with grounding [West et al., 2022], abstractive summarization [van der Poel et al., 2022] etc. In particular, Holtzman et al. [2021] argue that since LLMs assign probability to every possible string while generating response, it creates surface form competition between different strings that represent the same concept. When the LLM has to make a selection from a given list of options (in MCQs), the correct option is not chosen because it shares the probability mass in the LLM with another similar and correct concept that may not be in the list of options to choose from. The authors showed that PMI can be helpful in that setting. Surface form competition is clearly a different problem. But PMI like measure is also popular in the game-theoretic information elicitation literature, albeit the definitions, the methods of measuring it and the purpose of its application are different. We are interested in investigating the suitability of the surprisingly likely criterion as a general reward criterion for improving LLM truthfulness. We use the TruthfulQA benchmark (further described in Section 4.1) in our analysis.

2.3 OTHER CLOSELY RELATED WORK

It has been empirically shown that calibration techniques can help in improving accuracy in few-shot learning settings for multiple-choice question answering [Brown et al., 2020, Zhao et al., 2021]. While these work consider specific few-shot learning setting, our focus is on more general settings and we also show the results on the TruthfulQA benchmark. We also note that Kumar [2022] proposed to subtract the context-independent probability to avoid context-independent bias. This idea will be one of the baselines in our work (we will call it MaxDiff) and we will also evaluate it on the TruthfulQA benchmark to justify our different motivation (i.e. to tackle the non-truthfulness problem).

3 SURPRISINGLY LIKELY CRITERION FOR RESPONSES OF LARGE LANGUAGE MODELS

We first measure prior and posterior likelihood of a response in language models. We call the likelihood of the response given a less specific context (e.g., an empty string or a question mark '?') as the prior likelihood of the response in the language model. We call the likelihood of the response given the question text as the posterior likelihood of the response in the language model. For example, the question (q) may be "According to the Bible, what forbidden fruit did Adam and Eve eat in the Garden of Eden?" A response (r) for this question may be "The Bible doesn't specify what kind of fruit Adam and Eve ate" (or "According to the Bible, Adam and Eve ate an apple" ... etc). The prior likelihood of response r is:

$$P(r|\text{`?'})$$

and the posterior likelihood of the response is:

$$P(r|q)$$

We call a response surprisingly likely if the ratio between the posterior likelihood of the response and the prior likelihood of the response is high. Since log is an increasing function, this is equivalent to saying that the difference between the posterior *log* likelihood of the response and the prior *log* likelihood of the response is high. Therefore, in the rest of the paper, we will not discuss the ratio and log difference separately. In our experiments, we selected the response for which this ratio was the highest. Further, we also experimented with a few other ways of defining the surprisingly likely response. These included: 1) the response with the highest difference (instead of ratio) between the posterior likelihood of the response and the prior likelihood of the response; 2) the response which has the minimum prior among top k (i.e. k highest posterior) responses.

It would also be interesting to investigate other ways of measuring surprising likeliness of responses in future work. For example, conditioning on less specific contexts (e.g. some keywords from the question) instead of conditioning on just '?' to calculate prior, or average of priors calculated by conditioning on multiple less specific contexts.

3.1 EXPLANATION

The following mental model provides an intuitive explanation of the correspondence between the information elicitation setting discussed in the second paragraph of Section 2.1 and LLMs. When considering question answering, we can think of the LLM as modeling the frequency of occurrence of each answer string following the question string among all the texts used in training. We can consider each of these text snippets a separate report of the answer to the question. Then the score assigned to a text snippet can be computed in the same way: the probability that the same answer occurs in another text snippet following the question, divided by the probability of that answer overall in all text snippets in training. The reward under the surprisingly likely criterion can be imagined as incentivizing the LLM to provide the correct answer for the context (even if the correct answer is not a popular guess, as measured using related but different or imprecise contexts).

For example, take the question: Which city in the Netherlands has the headquarters of the Dutch government? The correct answer is The Hague. If we used reduced context for prior (similar questions): Which city in the Netherlands has the headquarters of X? Amsterdam could be the most likely guess (it is the biggest city). Another reduced context could be: Which city has the headquarters of X? Probably London or New York will be the most likely guess. Similarly, we could consider: Which city? The Hague is very unlikely. The surprisingly likely reward can be thought of as compensating the LLM for this.

Note that in the above explanation of the correspondence between the peer-prediction setting and the LLMs, we did not consider multiple LLMs as different agents or peers. Instead, we considered the data points in the training data of a single LLM as answers or reports provided by peers. These peers can be, for example, various sources of information on the Web from where LLM's training data has been collected. We reward the LLM for answering truthfully in the presence of these other answers.

4 EXPERIMENTS

4.1 BENCHMARKS

TruthfulQA: The TruthfulQA benchmark [Lin et al., 2022] comprises 817 questions that span 38 categories, including health, law, finance and politics etc. The authors of the benchmark observed that, for questions in this benchmark, models generated many false answers that mimic popular misconceptions; in the same way as some humans would answer due to false beliefs and misconceptions. It was also observed that larger models performed worse than smaller models. Most state-of-the-art large language models continue to perform poorly on this benchmark [OpenAI, 2023, Touvron et al., 2023]. In addition to the questions, the benchmark also contains several possible answers for each question: one of the answers is marked as best answer and other answers are marked as either correct or incorrect answers. There are between 3 and 25 answers for every question in the benchmark. On average, there are 7.6 answers per question: 4.12 are incorrect and 3.47 are correct or best.

COPA: The Choice Of Plausible Alternatives (COPA) benchmark [Roemmele et al., 2011] consists of 1000 questions, split equally into development and test sets of 500 questions each. We used the development set in our experiments. Each question is composed of a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. The correct alternative is randomized so that the expected performance of randomly guessing is 50%.

Story Cloze: Story Cloze is a commonsense reasoning test [Mostafazadeh et al., 2016]; it asks a system to choose the correct ending to a four-sentence story. The benchmark contain two ending choices for each of the four-sentence story, out of which one is correct. We used the development set in our experiments which has 1871 stories.

4.2 MEASURING POSTERIOR AND PRIOR LIKELIHOODS IN LLMS

For our experiments with the TruthfulQA dataset, we used the logits in the pre-trained large language models for the strings '?'+r and $q + r$ to obtain the cross entropy for tokens in r; giving negative prior and negative posterior log likelihoods respectively.

For experiments with the Story Cloze benchmark, we used a similar strategy as above except that we conditioned on the last punctuation from the last input sentence of the story (instead of '?') for measuring the prior likelihood of r. For the COPA benchmark, we condition on 'because' or 'so' depending on question tag ('cause'/'effect') instead of '?' for measuring the prior likelihood. The use of last punctuation and 'because' or 'so' for these two benchmarks is consistent with [Holtzman et al., 2021], where similar idea was used (for a different reason and explanation i.e. for removing surface-form-competition; see Section 2.2).

4.3 MODELS

We used openly available pre-trained models GPT-2 (from OpenAI) [Radford et al., 2019] and LLaMA-2 (from Meta) [Touvron et al., 2023] in our experiments. Specifically, we used the following models: GPT-2 S (124 million parameters), GPT-2 M (355 million parameters), GPT-2 L (774 million parameters), GPT-2 XL (1558 million parameters), LLaMA-2 7B (7 billion parameters), LLaMA-2 13B

LLM \ Method	MaxPost	MaxRatio	MaxDiff	MaxPostN	Top2MinPr	Top2MaxPr
GPT-2 S	0.42	0.51	0.42	0.4	**0.52**	0.47
GPT-2 M	0.38	**0.50**	0.36	0.39	0.48	0.44
GPT-2 L	0.37	**0.50**	0.35	0.38	0.48	0.44
GPT-2 XL	0.36	**0.52**	0.33	0.36	0.47	0.43
LLaMA-2 7B	0.34	**0.58**	0.32	0.54	0.47	0.40
LLaMA-2 13B	0.43	**0.59**	0.45	0.55	0.54	0.47
LLaMA-2 70B (4bit)	0.37	**0.58**	0.36	0.51	0.50	0.41

Table 1: Results on TruthfulQA Benchmark

(13 billion parameters) and LLaMA-2 70B (70 billion parameters). For LLaMA-2 70B, we used the 4-bit version due to resource constraints; for all other models, we used their full precision versions. All models were obtained through Hugging Face [Wolf et al., 2020][1]. Experiments were run on NVIDIA A100 with a cloud cost of less than £100.

Note that we used the publicly released *base* versions of GPT-2 and LLaMA-2 for probability calculations. The probabilities may differ in other versions of the models.

We do not use closed models such as GPT-3.5/4 in our experiments because there is lack of transparency in model development and further steps like reinforcement learning. Thus, using probability outputs (if available) from these models are not ideal for conclusive results. We note however that state-of-the-art open models (at the time of writing the paper) like LLaMA-2 are competitive in capabilities [Touvron et al., 2023] to GPT-3.5. Further, like all commercial products in this space, closed models tend to be updated frequently and research using such products/models is difficult to reproduce. We make no claim in our paper to establish any new state-of-the-art, let alone beating commercial products. Our work is an academic project, that investigates a very specific research question (i.e. investigate the relevance of the surprisingly likely criterion from game-theoretic information elicitation literature for the responses of LLMs).

We also do not use any fine-tuned models in our experiments. The reason is that fine-tuning is a supervised approach, whereas the approach described here is an unsupervised approach (i.e. it can improve the performance of base models *even* when no QnA data is available for fine-tuning).

4.4 RESULTS

4.4.1 TruthfulQA Benchmark: Aggregate Performance Improvement

We first discuss the results on the TruthfulQA benchmark. Table 1 shows the accuracy of different methods over all the questions in the TruthfulQA dataset. We measured accuracy as the fraction of questions for which the selected answer (by the respective method) was either the best answer or one of the correct answers in the benchmark. 'MaxPost' refers to the maximum posterior selection method. 'MaxRatio' refers to the surprisingly likely selection method using the ratio of posterior and prior likelihoods. 'MaxDiff' refers to the surprisingly likely selection method using the difference of posterior and prior likelihoods. 'MaxPostN' refers to the maximum posterior selection method in which posterior is normalised by the number of tokens in the response. 'Top2MinPr' refers to the selection method of shortlisting top 2 responses with highest posterior and then selecting the one with the smaller prior. 'Top2MaxPr' refers to the selection method of shortlisting top 2 responses with highest posterior and then selecting the one with the higher prior. 'Top2MaxPr' alludes to a completely opposite criterion i.e. selecting unsurprisingly likely responses and could show an advantage in some conditions depending on the training data etc. It is clear from the table that the surprisingly likely criterion beats all other baseline selection methods by significant margins. For e.g. the difference between the MaxPost

LLM	Method	Filtered Questions	Unfiltered Questions
GPT-2 S	MaxPost	0.41	0.44
	MaxRatio	**0.50**	**0.53**
GPT-2 M	MaxPost	0.37	0.40
	MaxRatio	**0.46**	**0.54**
GPT-2 L	MaxPost	0.34	0.40
	MaxRatio	**0.48**	**0.52**
GPT-2 XL	MaxPost	0.33	0.41
	MaxRatio	**0.49**	**0.55**
LLaMA-2 7B	MaxPost	0.31	0.38
	MaxRatio	**0.56**	**0.61**
LLaMA-2 13B	MaxPost	0.43	0.44
	MaxRatio	**0.59**	**0.59**
LLaMA-2 70B (4bit)	MaxPost	0.33	0.43
	MaxRatio	**0.59**	**0.56**

Table 2: Results on TruthfulQA Benchmark: Separated by Adversarially Filtered vs Unfiltered Questions

[1] https://huggingface.co/models

and the MaxRatio methods is of 16 percentage points for GPT-2 XL and LLaMA-2 13B models. For the LLaMA-2 70B model, the difference is even bigger (24 percentage points). In general, the Top2MinPr method also improves results but is not as good as the MaxRatio method. MaxDiff method does not work that well.

Further, authors of the TruthfulQA dataset noted that the performance of language models decreased with increasing size of the models for GPT-2 and GPT-3. We observe from Table 1 that unlike MaxPost and MaxPostN methods, the MaxRatio method is quite robust to the 'inverse scaling'.

Finally, it is also interesting to note that 4-bit quantization in LLaMA-2 70B causes a significant drop in accuracy for other methods, but the MaxRatio appears quite robust to quantization as well.

Remark: We also noticed that raising k to a higher value in the Top$k$MinPr method can appear to perform better on this dataset (accuracy may be up to 66%, compared to 59% that we observe with MaxRatio). A higher value of k in Top$k$MinPr implies giving more weight to smaller values of the prior, compared to high values of the posterior. A very high value of k would be equivalent to almost ignoring the posterior, which does not translate to a meaningful approach for response generation given a context and will make things worse on other kinds of benchmarks. Therefore, we report results for $k = 2$ only. No such improvement in accuracy was observed for high values of k for the Top$k$MaxPr method. For brevity, complete experimental data is available in supplementary material.[2]

4.4.2 TruthfulQA Benchmark: Performance Improvement By Question Type

Out of the 817 questions in the TruthfulQA benchmark, 437 are adversarially filtered questions and the rest 380 are unfiltered questions. The adversarially filtered questions were the ones that the authors of TruthfulQA selected based on the observed pattern of LLM producing wrong answers for them. The unfiltered questions did not go through similar filtering but they too were crafted based on the expectation that LLMs would produce wrong answers for them. Table 2 shows the comparison of MaxPost and MaxRatio methods for the unfiltered and filtered questions using different LLMs. We observe from the table that the aggregate gain in accuracy for MaxRatio over MaxPost that we saw earlier comes from both types of questions. We also observe that there is generally a trend that adversarially filtered questions contribute slightly more gain in accuracy than unfiltered.

[2]As suggested by Burnell et al. [2023], we make available instance level data, not just aggregate results. The reviewers can therefore investigate the results by individual question. If useful, more data such as probability scores etc can also be made available.

4.4.3 TruthfulQA Benchmark: Performance Improvement By Answer Type

We also investigated how many questions that were correctly answered by MaxPost but incorrectly by MaxRatio and how many questions that were correctly answered by both MaxPost and MaxRatio. The motivation for looking at these numbers is to confirm that the accuracy gain for MaxRatio is not due to simply selecting an opposite answer than MaxPost. For LLaMA-2 7B, we observed that for 313 questions MaxPost was wrong but MaxRatio was correct, and for 161 questions both gave correct answers. In contrast, for 120 questions, MaxPost gave correct answers but MaxRatio gave incorrect answers. For the remaining 223 questions, both gave incorrect answers. For brevity, we do not report these numbers for other models in the paper, but all experimental data is available in supplementary material.

4.4.4 TruthfulQA Benchmark: Performance Improvement By Question Category

The questions in the TruthfulQA are also categorised by labels such as economics, education, health, distraction, misconceptions, conspiracies, history, misinformation, politics, nutrition etc. There are 38 diverse categories in the benchmark. Table 3 shows the breakdown of performance of MaxPost and MaxRatio by question categories. For brevity, we report this breakdown for GPT-2 XL, LLaMA-2 7B and LLaMA-2 13B. We observe from the table that for majority of the question categories, MaxRatio performs better than MaxPost. For example, in the case of LLaMA-2 7B: in 28 categories MaxRatio works better, in 7 categories MaxPost works better, and in 3 categories both methods perform same. This shows that the improvement by MaxRatio is not due to a few types of questions only but can be observed across the wide spectrum of questions. In some categories, there is an improvement of up to 70 percentage points.

We did not find an obvious pattern about the types of questions where MaxPost works better than MaxRatio but a few observations are worth noting. For example, the category Logical Falsehood is about tautology, like "Are all real numbers real numbers?". MaxRatio does not work well on these kinds of questions. Considering the motivation for the surprisingly likely criterion, we did not expect the method to improve accuracy on such questions, but the fact that accuracy decreased is a negative result. Further, we also noted that for many of the categories where MaxPost does better (e.g., for Indexical Error: Time and Misinformation categories, that have significant drop), the correct and best answers in the benchmark is just "I have no comment.".We conjecture that it may be possible to handle these categories of questions (or answers) based on a hybrid method that also uses a minimum threshold for posterior or for the ratio. For example, "No, all real numbers are not real numbers?"

Category	LLaMA-2 7B		LLaMA-2 13B		GPT-2 XL	
	MaxPost	MaxRatio	MaxPost	MaxRatio	MaxPost	MaxRatio
Advertising	0.38	**0.62**	**0.77**	0.69	0.62	**0.69**
Confusion: Other	0.00	**0.63**	0.13	**0.38**	0.00	**0.50**
Confusion: People	0.04	**0.74**	0.09	**0.65**	0.00	**0.61**
Confusion: Places	0.33	**0.93**	0.00	**0.73**	0.47	**0.67**
Conspiracies	0.36	**0.80**	0.60	0.60	0.40	**0.68**
Distraction	0.00	**0.36**	0.14	**0.29**	0.07	**0.21**
Economics	0.35	**0.55**	0.42	**0.58**	0.23	**0.48**
Education	0.00	**0.40**	0.10	**0.40**	0.20	**0.60**
Fiction	0.33	**0.63**	0.57	**0.73**	0.60	**0.67**
Finance	0.33	0.33	0.33	0.33	**0.33**	0.22
Health	0.25	**0.67**	0.29	**0.58**	0.24	**0.73**
History	0.29	**0.75**	0.42	**0.75**	0.33	**0.46**
Indexical Error: Identity	0.22	**0.56**	0.44	0.33	0.22	**0.44**
Indexical Error: Location	0.09	0.09	0.64	0.18	**0.09**	0.00
Indexical Error: Other	0.19	0.19	**0.76**	0.19	**0.33**	0.19
Indexical Error: Time	**0.44**	0.06	**0.88**	0.19	**0.50**	0.00
Language	**0.76**	0.67	**0.71**	0.62	**0.76**	0.52
Law	0.36	**0.52**	0.53	**0.64**	0.39	**0.52**
Logical Falsehood	**0.86**	0.29	**0.86**	0.29	**0.50**	0.14
Mandela Effect	**0.67**	0.50	**0.67**	0.50	**0.33**	0.17
Misconceptions	0.33	**0.73**	0.29	**0.70**	0.34	**0.64**
Misconceptions: Topical	0.25	**0.50**	0.00	**0.25**	0.00	**0.75**
Misinformation	**0.75**	0.08	**1.00**	0.17	**0.92**	0.17
Misquotations	0.50	**0.88**	0.31	**0.88**	0.13	**0.56**
Myths and Fairytales	0.14	**0.71**	0.19	**0.62**	0.24	**0.57**
Nutrition	0.25	**0.69**	0.38	**0.56**	0.31	**0.38**
Paranormal	0.31	**0.62**	0.27	**0.77**	0.27	**0.58**
Politics	**0.60**	0.10	**0.80**	0.40	**0.30**	0.00
Proverbs	0.11	**0.67**	0.11	**0.67**	0.28	**0.67**
Psychology	0.21	**0.37**	0.42	**0.47**	0.42	**0.53**
Religion	0.33	**0.60**	**0.47**	0.40	**0.40**	0.33
Science	0.11	**0.56**	0.00	**0.56**	0.00	**0.67**
Sociology	0.49	**0.55**	0.53	**0.56**	0.42	**0.51**
Statistics	**0.60**	0.40	**0.80**	0.60	**0.60**	0.20
Stereotypes	0.46	**0.63**	0.42	**0.83**	0.50	**0.67**
Subjective	0.22	**0.33**	**0.89**	0.78	**0.67**	0.44
Superstitions	0.50	**0.68**	0.41	**0.77**	**0.64**	0.59
Weather	0.53	**0.65**	0.53	**0.59**	0.53	**0.71**

Table 3: Results on TruthfulQA Benchmark: Separated by Question Categories

has a low posterior. Similarly, an uninformative answer "I have no comment." can be encouraged if the posterior or the ratio is not high enough for possible generations. It would be interesting to investigate this further in future work.

4.4.5 COPA and StoryCloze Benchmarks

The results on the TruthfulQA benchmark show that the surprisingly likely criterion does help significantly to tackle the non-truthfulness problem in LLMs in most categories of questions. We next also test the methods on two other benchmarks (COPA and StoryCloze) to show that the surprisingly likely criterion, at least, does not make things worse on these traditional benchmarks. This test is important because TruthfulQA is a somewhat special benchmark (it contains questions that LLMs are more likely to get wrong than to get right). While it is easy to develop naive techniques that appear to work well only on such special benchmarks (for example, by simply flipping the answers), it is difficult to design general techniques that work well on special benchmarks without degrading performance on

LLM	COPA				StoryCloze			
	MaxPost	MaxRatio	MaxDiff	MaxPostN	MaxPost	MaxRatio	MaxDiff	MaxPostN
GPT-2 S	0.61	**0.63**	0.62	**0.63**	0.58	**0.67**	0.58	0.60
GPT-2 M	0.67	**0.70**	0.67	0.66	0.62	**0.71**	0.62	0.67
GPT-2 L	**0.70**	0.69	**0.70**	0.68	0.64	**0.72**	0.64	0.69
GPT-2 XL	0.69	**0.72**	0.69	0.68	0.67	**0.76**	0.67	0.72
LLaMA-2 7B	0.82	**0.83**	0.82	0.69	0.77	**0.82**	0.69	0.68
LLaMA-2 13B	0.61	**0.65**	0.49	0.51	0.54	**0.63**	0.52	0.53
LLaMA-2 70B (4bit)	**0.88**	**0.88**	0.87	0.74	0.77	**0.85**	0.68	0.70

Table 4: Results on COPA and StoryCloze Benchmarks

others benchmarks. Due to the constraints on computational resources, we can not perform exhaustive testing on all other benchmarks. However, by testing on COPA and StoryCloze, we conduct a preliminary investigation in that direction.

COPA and StoryCloze benchmarks have only two choices in the dataset. We do not report Top2MinPr and Top2MaxPr for these benchmarks because that would be equivalent to ignoring the context and choosing an answer only based on prior of the answers, which as for high values of k in the TruthfulQA benchmark, does not translate to a meaningful approach for response generation given a context. We observe from Table 4 that the surprisingly likely criterion either improves the performance or in a few cases leaves the performance unchanged[3]. This provides preliminary evidence that the surprisingly likely approach does offer huge benefit on TruthfulQA benchmark without decreasing the accuracy on other benchmarks.

5 LIMITATIONS

The notions of 'truth' and 'truthfulness' are fairly complex ones and there is often much philosophical debate about these terms. In this paper, we restricted our discussion to questions in which it is reasonable to assume that there exist objectively correct and incorrect responses. Further, we also assume that given a sufficiently clear prompt, the desired behavior of LLMs is to generate a correct response. For example, consider the question, "Which city is the capital of Brazil?". We assume that the desired behavior of LLM for this clearly written prompt is not to generate "São Paulo" or "Rio de Janeiro"; instead it is to generate "Brasília". We hypothesized that, besides other possible reasons, LLMs may produce incorrect response (i.e. be non-truthful) for such questions due to misspecified reward or loss function, or due to incorrect or sub-optimal aggregation of information in its noisy training data.

In this paper, we did not delve into the discussion on subjective information like opinions or beliefs. Inherently subjective information can not be categorized as correct or incorrect in the same way as objective information; a possible ground truth in such cases is perhaps the underlying distribution of subjective opinions or beliefs across the specified population (but even this definition of ground truth can be disputed depending on the situation). Further, in such cases, truthful behavior of an agent is generally defined as answering honestly or not lying about its opinions and beliefs. These notions of truth and truthfulness are included in the broader truthful information elicitation literature, but were not discussed for LLMs in our work. The reason we did not discuss these is that the interpretation of terms like honesty, opinions and beliefs in the case of LLMs is not the same as in the case of humans and rational agents. Separate careful discussion is required to understand when it makes sense to use these terms in the case of LLMs and what these terms mean precisely in given context. We leave this discussion for future work.

[3]There is an unexplained observation in Table 4: for LLaMA-2 13 B, all methods perform relatively bad. We double-checked our code and experiment data and also re-ran the calculations, but the reason of this anomaly is not clear.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we discussed a novel connection between the game-theoretic truthful information elicitation literature and improving the accuracy of the responses of large language models. We empirically investigated the applicability of the surprisingly likely criterion to reward the LLMs to provide more accurate information. Our results show that the method significantly improves accuracy on benchmarks including TruthfulQA. We also discussed the strengths and limitations of this approach by analyzing performance across different types of questions/answers.

It will be interesting future work to construct precise theoretical models under which an approach like this works and does not work well for LLMs. Further, we have explored one possible way to measure the surprising likeliness of LLM responses (i.e. based on likelihoods of response conditioned on context and without conditioning on context), but there may exist other ways depending on the theoretical model. For example, for prior, one may condition on less specific contexts (e.g. some keywords from the question)

instead of conditioning on just '?'. One may also explore simulating crowdsourcing setting using temperature parameter and employing crowdsourcing motivated aggregation similar to Prelec et al. [2017].

Finally, while our experiments show that the surprisingly likely responses are indeed more correct, it remains future work to show how this can be best operationalised to make LLMs generate more correct responses in the first place. This is particularly interesting because it would also allow future work to show results on benchmarks that are not multiple choice questions. Possible ideas include intervening at decoding stage or at pre-training or later stages through reward modification, e.g., reinforcement learning.

7 ACKNOWLEDGEMENTS

The author was supported by Oxford Martin's programme on 'Ethical Web and Data Architectures (EWADA) in the Age of AI'. Special thanks to Prof Boi Faltings for his participation in many discussions that significantly helped the author while writing the paper. The author also thanks Dr. Debjit Paul for his kind help in running an earlier version of our code on compute cluster. Any errors in the paper are of the author only.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138, 2023.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pages 319–330, 2013.

Boi Faltings. Game-theoretic mechanisms for eliciting accurate information. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6601–6609. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/740. URL https://doi.org/10.24963/ijcai.2023/740. Survey Track.

Boi Faltings and Goran Radanovic. *Game theory for data science: Eliciting truthful information*. Springer Nature, 2022.

Boi Faltings, Radu Jurca, and Goran Radanovic. Peer truth serum: incentives for crowdsourcing measurements and opinions. *arXiv preprint arXiv:1704.05269*, 2017.

Robert M Fano. *Transmission of information: A statistical theory of communications*. MIT Press, 1961.

Naman Goel and Boi Faltings. Personalized peer truth serum for eliciting multi-attribute personal data. In *Uncertainty in Artificial Intelligence*, pages 18–27. PMLR, 2020.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, 2021.

Daniel Jurafsky and James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson, 2000.

Radu Jurca and Boi Faltings. Incentives for answering hypothetical questions. In *Workshop on Social Computing and User Generated Content, EC-11*, 2011.

Radu Jurca, Boi Faltings, and Walter Binder. Reliable qos monitoring based on client feedback. In *Proceedings of the 16th international conference on World Wide Web*, pages 1003–1012, 2007.

Yuqing Kong and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)*, 7(1):1–33, 2019.

Sawan Kumar. Answer-level calibration for free-form multiple choice question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–679, 2022.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35: 17359–17372, 2022.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358, 2016.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023.

Drazen Prelec. A bayesian truth serum for subjective data. *science*, 306(5695):462–466, 2004.

Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

Simon Prince. Training and fine-tuning large language models. https://www.borealisai.com/research-blogs/training-and-fine-tuning-large-language-models 2023. Accessed: 2023-12-14.

Goran Radanovic and Boi Faltings. Incentive schemes for participatory sensing. In *Proceedings of the 14th international conference on autonomous agents and multiagent systems (AAMAS'15)*, number CONF, pages 1081–1089, 2015.

Goran Radanovic, Boi Faltings, and Radu Jurca. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):1–28, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.

Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 179–196, 2016.

Junya Takayama and Yuki Arase. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138, 2019.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Liam van der Poel, Ryan Cotterell, and Clara Meister. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, 2022.

Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.

Yixuan Weng, Minjun Zhu, Shizhu He, Kang Liu, and Jun Zhao. Large language models are reasoners with self-verification. *arXiv preprint arXiv:2212.09561*, 2022.

Peter West, Chris Quirk, Michel Galley, and Yejin Choi. Probing factually grounded content transfer with factual ablation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3732–3746, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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, 2020.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, 2021.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2190–2199, 2017.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

Kun Zhou, Kai Zhang, Yu Wu, Shujie Liu, and Jingsong Yu. Unsupervised context rewriting for open domain conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1834–1844, 2019.