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

Naman Goel
University of Oxford
naman.goel@cs.ox.ac.uk

Abstract

The principle of rewarding a crowd for surprisingly common answers has been used in the literature for designing a number of truthful information elicitation mechanisms (1; 2; 3; 4; 5; 6; 7). A related method has also been proposed in the literature for better aggregation of crowd wisdom (8). Drawing a comparison between crowd based collective intelligence systems and large language models, we define the notion of 'surprisingly likely' textual response of a large language model. This notion is inspired by the surprisingly common principle, but tailored for text in a language model. Using benchmarks such as TruthfulQA and openly available LLMs: GPT-2 and LLaMA-2, we show that the surprisingly likely textual responses of large language models are more accurate in many cases compared to standard baselines. For example, we observe up to 24 percentage points aggregate improvement on TruthfulQA and up to 70 percentage points improvement on individual categories of questions in this benchmark. We also provide further analysis of the results, including the cases when surprisingly likely responses are less or not more accurate.

1 Introduction

Recent demonstrations of the promising capabilities of large language models (LLMs) have raised hopes about their successful deployment 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 (9), retrieval augmentation/grounding (10), in-context learning (11), chain-of-thought reasoning (12) and other prompting, self-consistency (13), advanced decoding techniques (14), uncertainty measurement based techniques (15), self-improvement (16), shifting model activations during inference (17), locating and editing factual information (18), learning from human feedback (19), learning from AI feedback (20; 21; 22; 23), etc. While very useful, none of these efforts have fully solved the problem. Many scholars believe that the LLM architectures are fundamentally limited (24; 25; 26) and thus, cannot be guaranteed to always generate factually correct text. Despite this, it remains an interesting research direction to understand the various sources of this multi-faceted problem (27), and the strengths and weaknesses of various complementary approaches for improvement.

It is plausible that in some cases, LLMs generate incorrect information due to mis-specified objective. We call this behavior non-truthful. LLMs are commonly trained to generate answers that have the maximum probability conditioned on the prompt (28). The probability depends on the data points in its noisy training set. Given various data points in the training set, an LLM may generate correct or incorrect answers. An interesting demonstration is due to Lin *et al.*, 2022 (29), 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. In this paper, we investigate whether and to what extent, the surprisingly common principle from the game-theoretic truthful information elicitation literature can be useful in avoiding this behavior.

There is a significant literature on game-theoretic incentive mechanisms for eliciting information from a crowd of agents. The mechanisms are also referred to as the peer-prediction or peer-consistency mechanisms (7).

A seminal contribution in this literature is due to Prelec, 2004 (1). Prelec designs a mechanism (called the Bayesian Truth Serum) such that an agent answering a question, can maximize its expected incentive 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 a priori 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. Besides incentive-compatibility, it has also been shown that an information aggregation method based on the surprisingly common principle can be used to select more correct answers in crowdsourcing (8). Further research in this area used motivation from the surprisingly common principle of the BTS to design mechanisms for various crowdsourcing settings. While the BTS asks agents to submit two reports for a question to determine which answers are surprisingly common, the "minimal" mechanisms proposed later in the literature, ask agents to submit only one. We discuss this in further detail in Section 2.1.

An LLM is very different from a crowd of agents in many ways. For example, the agents in crowdsourcing literature are assumed to be able to make independent observations about the world by incurring variable cost, form and update beliefs, and strategically (mis-)report their beliefs to maximize the incentives. On the other hand, the same cannot be said about a language model. Therefore, the theory of incentive mechanisms for crowd does not apply verbatim to an LLM. However, as we will discuss in Section 3.1, a comparison between crowdsourcing and how an LLM "aggregates" information from its training data points and generates answers when prompted, motivates us to take a step towards exploring the connections between the two fields of research. In this paper, we restrict the discussion on "truth" to objective information that can be clearly categorized as correct or incorrect (please see Section 6 for critical discussion on the scope of this work).

We define the notion of 'surprisingly likely' textual response of a large language model. It is inspired from the surprisingly common principle developed in the information elicitation literature. In particular, we draw inspiration from the peer-truth serum mechanism (4). While the peer-truth serum and other mechanisms were developed for numerical or categorical answers from a crowd of agents, our 'surprisingly likely' measure is adapted for textual responses of a single LLM that has been pre-trained on data from various sources. Through experiments on TruthfulQA, COPA and StoryCloze benchmarks, we show that surprisingly likely answers in large language models are indeed more correct in several cases. For example, on the TruthfulQA benchmark, we find significant gains in accuracy across different LLMs (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

2.1 Surprisingly Common Principle in Crowdsourcing

The problem of information elicitation without verification (i.e. when correct information is not available for scoring and agents in a crowd are to be incentivized for providing correct information) is modeled as a game between multiple agents in the literature (6). One of earlier incentive mechanisms in this space is known as the Bayesian Truth Serum (1). 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 honest 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 expected reward by telling what they believe is the correct answer.

A number of other BTS motivated reward/scoring mechanisms advanced this field of research. We focus on proposals that are suitable for crowdsourcing without requiring the agents to explicitly submit their prediction about other agents' answers. Consider for example, a crowdsourcing task of measuring pollution at locations, where no independent ground truth measurement exists. Multiple independent agents (peers) measure and report the value at a given location 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 the agents. Peer-prediction mechanisms (6) consider this setting as a game among the agents, where each tries

to maximize the expected reward attributed to their report. The simplest form is output agreement (30), 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 (31). Mechanisms, that address this issue in output agreement, are all based on the surprisingly common principle but mathematical models and game-theoretic arguments differ in different mechanisms. In a recent survey, Faltings, 2023 (7) identifies three types of mechanisms. The first is agreement, where the reward is proportional to the frequency of the answer among other responses, often calibrated by the overall chance of agreement (30; 2; 3). The second is information-theoretic, where the reward is proportional to the pairwise mutual information between the answer and the answers given by peers (1; 5; 32; 33). The third computes the reward based on the improvement in the quality of the resulting model (the Peer Truth Serum) (34; 35; 4).

2.2 The Peer Truth Serum

The most relevant mechanism for this paper is the Peer Truth Serum or the PTS mechanism. In the pollution measurement example, if an agent i reported the pollution measurement at location l, PTS calculates the reward for the agent as follows. PTS selects another agent p (called peer) who also submitted pollution measurement for the same location l (or approximately same location neighborhood). Suppose that the agent i submitted x_i and the peer p submitted x_p. Here, $x_i, x_p \in \mathcal{X}$ are pollution measurement values. For e.g., $\mathcal{X} = \{low, moderate, high, very\ high\}$.

The reward of agent i under the PTS mechanism is proportional to:

$$\frac{\mathbb{1}_{x_i=x_p}}{R_i(x_i)}$$

where $R_i(x_i) = num_i(x_i) / \sum_{x \in \mathcal{X}} num_i(x)$, and $num_i(x)$ is a function that counts occurrences of x in the values reported by other agents, across a large number of other locations that are a priori statistically similar. $\mathbb{1}_{x_i=x_p}$ is indicator function (1 if $x_i = x_p$, 0 otherwise).

Observe that, *in expectation*, the PTS does not just reward the answer which is most likely to be given by peer, but it also scales the reward inversely by a 'prior' for the answer. This prior is estimated from the answers collected from the crowd across a number of a priori statistically similar questions. This is how the PTS mechanism operationalizes the surprisingly common principle. In this paper, we take inspiration from the PTS mechanism to define a related notion of 'surprisingly likely' textual responses of a language model.

3 Surprisingly Likely Responses of Large Language Models

We now define the notion of 'surprisingly likely' responses of large language models. Consider the following example. A question (q) asked to an LLM is "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).

We assign a response r a score $\tau(r, q)$ as follows:

$$\tau(r, q) = \frac{P(r|q)}{P(r|\text{'?'})} \tag{1}$$

where $P(r|q)$ is the conditional probability of the response in the language model, given the question. $P(r|\text{'?'})$ is the conditional probability of the response in the language model, given '?'. We call $P(r|\text{'?'})$ as the 'prior' of the response in the language model. In equation 1, we use just a question mark '?' for calculating the prior, but there might also be other possibilities. For e.g., an empty string or another reduced and similar context (examples to follow in further section). The prior might also be calculated using an average of priors obtained by conditioning on several different reduced and similar contexts. Note that marginalizing over all possible questions (question text minus '?') leads to just the probability conditioned on '?'.

We call a response r surprisingly likely, if the score $\tau(r, q)$ is higher compared to other responses.

3.1 Discussion

We now draw a comparison between the Peer Truth Serum based information elicitation discussed in Section 2.2 and LLM response generation. For answering the questions that have right and wrong answers, we can think

of an LLM as modeling the frequency of occurrence of different answer strings following the question string, among all the text data used in training. We can consider each occurrence of these text snippets in training data as a separate report of the answer to the question. Reports might come from different sources of information on the Web and other sources in the training data. The score τ assigned to an answer is computed as: 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. Thus, similar to the PTS reward, selecting the answer with high τ can be understood as equivalent to rewarding the LLM to generate surprisingly likely answer for the context, even if this answer does not have the highest numerator in the reward score τ.

Some questions that naturally follow from the above discussion are: can this strategy improve the accuracy of LLM generated responses?, why?, and in which scenarios it may not work? In this paper, we take an empirical approach to address these questions. A theoretically rigorous treatment to these specific questions should be interesting future work.

As an example, consider the question: Which city in the Netherlands has the headquarters of the Dutch government? The correct answer for this question is The Hague. Suppose we used the following reduced and similar context for calculating the prior: Which city in the Netherlands has the headquarters of X? In this case, Amsterdam could be the most likely guess (since it is the biggest city). Another reduced context could be: Which city has the headquarters of X? In this case, probably London or New York would be the most likely guess. Similarly, we could consider: Which city? In all these cases, The Hague is very unlikely. The surprisingly likely score compensates for this low prior of The Hague compared to Amsterdam, London and New York.

This was one example of the types of questions and answers, for which we conjecture that the LLM accuracy would improve by generating surprisingly likely answers. In further sections, we analyze the promise and limitations of this approach by conducting a series of experiments with different datasets and language models. Before presenting the experimental settings and results, we summarize some most closely related work in language modeling literature.

4 Related Work

4.1 PMI in Computational Linguistics

The information-theoretic measure of pointwise mutual information (PMI) (36) is a well-known concept in computation linguistics and natural language processing literature (37). It has been used in use-cases such as words association (38), keyword generation improving diversity of text (39; 40; 41; 42; 43), increasing agreement with grounding (44), abstractive summarization (45) etc. Holtzman *et al.*, 2021 (46) 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 multiple choice questions, 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 thus is another reason that may affect the accuracy of LLMs, but it is different from the non-truthfulness problem we discussed earlier (which, for example, leads to an LLM generating answers mimicking popular human misconceptions as demonstrated by the TruthfulQA benchmark). 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.

4.2 Other Closely Related Work

Prior work (11; 47) has shown that calibration techniques can help in improving accuracy in few-shot learning settings for multiple-choice question answering in language models. While they consider specific few-shot learning setting, our focus is on more general settings. We also show the results on the TruthfulQA benchmark. Further, Kumar, 2022 (48) proposed to subtract the context-independent probability to avoid context-independent bias. This idea will be the motivation of one of the baselines in our work and we will also evaluate it on the TruthfulQA benchmark.

5 Experiments

5.1 Benchmarks

TruthfulQA: The TruthfulQA benchmark (29) 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 (49; 50). 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, 3.47 are correct/best.

COPA: The Choice Of Plausible Alternatives (COPA) benchmark (51) 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 (52); 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.

5.2 Measuring Conditional Probabilities 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 of log of the conditional probabilities in the denominator and numerator respectively in equation 1 (i.e. $-\log P(r|'?')$ and $-\log\left(P(r|q)\right)$).

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. For the COPA benchmark, we condition on 'because' or 'so' depending on question tag ('cause'/'effect') instead of '?' for measuring the prior. The use of last punctuation and 'because' or 'so' for these two benchmarks is consistent with (46), where similar idea was used (for a different reason and explanation i.e. for removing surface-form-competition; see Section 4.1).

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: Accuracy of various methods with the 7 LLMs on the TruthfulQA Benchmark.

5.3 Models

We used openly available pre-trained models GPT-2 (from OpenAI) (53) and LLaMA-2 (from Meta) (50) 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 (13 billion parameters) and LLaMA-2 70B (70 billion parameters). For LLaMA-2 70B, we used the 4-bit version due to compute resource constraints on our end; for all other models, we used their full precision versions. All models were obtained through Hugging Face (54)[1]. We used

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

the publicly released *base* versions of GPT-2 and LLaMA-2 for probability calculations in our experiments. Experiments were run on NVIDIA A100 with a renting cost of less than GBP 100.

5.4 Accuracy Measure

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.

5.5 Baselines and Nomenclature

For brevity, in discussion of the results, we will use the following nomenclature for various methods used in the comparison. 'MaxPost' refers to a baseline standard method that selects the response with the maximum conditional probability given the question text. 'MaxPostN' refers to another baseline method in which the conditional probability given the question text is normalized by the number of tokens in the response and the response with the highest normalized probability score is selected. 'Top2MinPr' refers to a baseline method of shortlisting top 2 responses with highest conditional probability given the question text, and then selecting the one with the smaller prior. 'Top2MaxPr' refers to the selection method of shortlisting top 2 responses with highest conditional probability given the question text, and then selecting the one with the higher prior. 'Top2MaxPr' alludes to a completely opposite method i.e. selecting the unsurprisingly likely responses.

'MaxRatio' refers to the surprisingly likely selection method introduced in Section 3, i.e. selecting the response with the highest ratio of the two conditional probabilities. 'MaxDiff' refers to yet another baseline method that selects the response with the highest difference between the two conditional probabilities (instead of the ratio). This baseline method is motivated from the explanation by Kumar 2022 (48).

5.6 Scope of the Experiments

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 research. We note however that state-of-the-art open models (at the time of writing the paper) like LLaMA-2 are competitive in capabilities (50) to GPT-3.5. Further, like all commercial products in this space, closed models tend to be updated frequently, and research using such products is difficult to reproduce. In this paper, we make no claim of establishing a new state-of-the-art or beating commercial products. This work is an academic investigation, with limited compute resources, into a very specific research question introduced in Section 1.

We acknowledge that there are a number of new models and new benchmarks being continuously developed and released, as we write this paper. We do not believe that there is any consensus in the community regarding which benchmarks or which models are gold standard or ideal for which kind of research. We do not test all models and all benchmarks in this work, and leave this for a future endeavour with a larger research budget.

We also do not use any fine-tuned models in our experiments. The reason is that fine-tuning is a supervised approach, requiring labeled data to improve accuracy of question-answering in LLM, whereas the approach described in the paper is an unsupervised approach (i.e. it can improve the performance of base models *even* when no supervision data is available for such fine-tuning).

5.7 Results

5.7.1 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. It is clear from the table that the surprisingly likely method beats all other baseline selection methods by significant margins. For example, the difference between the MaxPost 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 and seems to reduce accuracy slightly compared to MaxPost.

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 such 'inverse scaling' phenomenon.

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: Accuracy on the TruthfulQA Benchmark: Separated by Adversarially Filtered vs Unfiltered Questions

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 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. A higher value of k in Top$k$MinPr implies giving more weight to smaller values of the prior. A very high value of k would be equivalent to almost ignoring the conditional probability given the question text. But ignoring the context almost entirely does not translate to a generally meaningful approach for response generation given a context, and it will make things worse on other kinds of benchmarks. We will discuss further in Section 5.7.5 that we wish to have an approach that not only shows better performance on TruthfulQA but the same approach should work more generally for other benchmarks too. 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]

5.7.2 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.

5.7.3 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 this error analysis 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 instance level data is available in supplementary material.

[2]Instance level data is available in supplementary material, not just aggregate accuracy measures presented here. The readers can therefore also investigate the results by individual question. If useful, more data such as probability scores etc can be requested by contacting the author.

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: Accuracy on the TruthfulQA Benchmark: Separated by Question Categories

5.7.4 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 do acknowledge however that these numbers get less conclusive since the benchmark is now divided into 28 smaller categories and it would be useful in the future to find bigger benchmarks on each of these categories and evaluate the methods on each benchmark.

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 method, 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 conditional probability in the numerator or for the ratio. For example, "No, all real numbers are not real numbers?" has a low conditional probability given the question text "Are all real numbers real numbers?". This threshold would become a hyper-parameter of the method and would need to be tuned to maximize the accuracy across different tasks and question categories. Similarly, an uninformative answer "I have no comment." can be encouraged if the conditional probability in the numerator or the ratio is not high enough for possible generations. It would be interesting to investigate this further in future work.

5.7.5 COPA and StoryCloze Benchmarks

The results on the TruthfulQA benchmark show that the surprisingly likely method does help with 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 method, at least, does not make things worse on these other 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 methods that *appear to* work well only on such special benchmarks (for example, by simply flipping the answers), it is difficult to design general methods that work well on special benchmarks without degrading performance on 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 does not translate to a generally meaningful approach for response generation given a context (as was also discussed in Section 5.7.1 for high values of k in the TruthfulQA benchmark). We observe from Tables 4 and 5 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 method does offer benefit on the TruthfulQA benchmark without decreasing the accuracy on other benchmarks.

LLM	Method			
	MaxPost	MaxRatio	MaxDiff	MaxPostN
GPT-2 S	0.61	**0.63**	0.62	**0.63**
GPT-2 M	0.67	**0.70**	0.67	0.66
GPT-2 L	**0.70**	0.69	**0.70**	0.68
GPT-2 XL	0.69	**0.72**	0.69	0.68
LLaMA-2 7B	0.82	**0.83**	0.82	0.69
LLaMA-2 13B	0.61	**0.65**	0.49	0.51
LLaMA-2 70B (4bit)	**0.88**	**0.88**	0.87	0.74

Table 4: Results on the COPA Benchmark

6 Limitations

The notions of 'truth' and 'truthfulness' are difficult to formalize 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

[3]There is an unexplained observation in Tables 4 and 5: for LLaMA-2 13 B, all methods perform relatively bad. We double-checked our code and experiment data and also re-ran the experiments/calculations, but the reason of this anomaly is not clear.

LLM	Method			
	MaxPost	MaxRatio	MaxDiff	MaxPostN
GPT-2 S	0.58	**0.67**	0.58	0.60
GPT-2 M	0.62	**0.71**	0.62	0.67
GPT-2 L	0.64	**0.72**	0.64	0.69
GPT-2 XL	0.67	**0.76**	0.67	0.72
LLaMA-2 7B	0.77	**0.82**	0.69	0.68
LLaMA-2 13B	0.54	**0.63**	0.52	0.53
LLaMA-2 70B (4bit)	0.77	**0.85**	0.68	0.70

Table 5: Results on the StoryCloze Benchmark

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 mis-specified training objective, 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. 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 agents in a crowd. Separate careful discussion is required to understand when it makes sense to responsibly use these terms in the case of LLMs and what these terms mean precisely in given context. We leave this discussion for future work.

7 Conclusions and Future Work

In this paper, we defined the notion of 'surprisingly likely' textual response of a large language model. This notion was inspired by the surprisingly common principle in the crowdsourcing literature, but tailored for text in a language model. We observed that surprisingly likely responses of large language models are more accurate on the TruthfulQA benchmark compared to baselines. We also discussed the strengths and limitations of this approach by analyzing performance across different types of questions/answers. This work is one of the early attempts to bridge two different research fields of language modeling and collective intelligence systems, and we hope that it will motivate further research at the intersection of the two research fields.

For example, an interesting future work that may directly follow this work is to construct theoretical models to provably explain the observations on different categories of questions and further understand the strengths and weakness of the approach. 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 implemented to make LLMs generate more correct responses in the first place. It would also allow researcher to obtain results on benchmarks other than multiple choice questions benchmarks (e.g., open-ended questions benchmarks). Possible ideas include intervening at decoding stage or at pre-training or later stages through reinforcement learning.

8 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 a compute cluster. Any errors in the paper are of the author only.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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. Survey Track. doi:10.24963/ijcai.2023/740.

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

[9] 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*, 2022. Survey Certification. URL: https://openreview.net/forum?id=yzkSU5zdwD.

[10] 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.

[11] 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.

[12] 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, 2022.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

[25] Yann LeCun, March 2023. URL: `https://twitter.com/ylecun/status/1640122342570336267`.

[26] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[27] 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.

[28] 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.

[29] 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. URL: `https://aclanthology.org/2022.acl-long.229`, `doi:10.18653/v1/2022.acl-long.229`.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

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

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

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] OpenAI. Gpt-4 technical report, 2023. `arXiv:2303.08774`.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.