Literature Review: TruthfulQA: Measuring How Models Mimic Human Falsehoods

1 Summary

TruthfulQA introduces a new benchmark to test LLMs truthfulness — that is, how much LLMs avoid asserting false statements. To do this, a new dataset of 817 questions (with truthful answers and references) is adversarily created (with reference to GPT-3-175B). The paper further assesses the performance of several well-known LLMs (GPT-3, GPT-Neo/J, GPT-2, UnifiedQA) at different parameter sizes to measure their performance on this set of questions. The performance is comprised of two axis: truthfulness and informativeness (defined as information that reduces the uncertainty raised by the question). They find larger models are less truthful.

While the construction of this dataset is important for researching the problems harmful training data will pass onto models, the paper's conclusion that the largest models are generally the least truthful is a contrived result arrived at by design, not by investigation. This benchmark itself is a worthwhile contribution, but one that is currently framed to be potentially misleading.

I would suggest a rephrasing of their findings and a stronger focus on identifying problematic training data, and on how researchers might prevent it from contaminating the LLMs generations – rather than deriving the somewhat trivial conclusion that the smaller the LLM, the less it memorizes its training data, and the less it can be prompted to generate similar outputs to it.

2 Strengths

The paper is clear and well-written. Based on the level of detail provided in both the paper and the appendixes, it should be possible to reproduce it.

The dataset was constructed employing quality human effort, and appears to be of high quality. It is further complemented with reference answers and citations for all questions.

Rather than creating a simple generative dataset, the paper proposes an additional multiple-choice counterpart, as well as two finetuned to automatically evaluate models' generations models (one for truthfulness, another for informativeness), both with high (90%) accuracies.

It is an interesting and valuable contribution to understand the current failure models of LLMs, and to confront them with adversarial inputs in order to be able to measure their robustness against these prompts. This is a useful benchmark on which we can measure progress on better training data processing and/or better control on model generations.

The paper does a thorough analysis to argue it was successful in its objective of eliciting imitative falsehoods (falsehoods that are likely given GPT-3's training dataset) – when compared against several other possible types of falsehoods (caused by syntax, grammar or form artifacts as well as other non-imitative weaknesses).

3 Weaknesses

The main contribution of this benchmark is severely limited by its construction. The paper states: "*For GPT-3 a false answer is an imitative falsehood if it has high likelihood on GPT-3's training distribution. [...] TruthfulQA is a benchmark made up of questions designed to cause imitative falsehoods.*"

This showcases this benchmark was adversarily constructed so that the more models are able to learn from their training distribution, the more likely they will be to generate an imitative falsehood learned from it. The dataset was constructed in two phases:

1. an initial generation of falsehood-inducing questions, which were then filtered based on whether GPT3-175B answered them correctly, and

2. a second generation based on the results of the first, where other similar prompts were

constructed.

The first stage produced 437 questions and the second round 380.

When this strictly adversarial design is ignored, the conclusion that large models are less truthful is much less clear. The controlled trivia questions have larger models performing significantly better (as seen in Figure 2), and the unfiltered set of TruthfulQA prompts results in a less obvious downward trend that actually shows GPT3-175B outperforming its 6.7B counterpart and possibly starting an upward trend (as shown in Figure 12 of Appx B.4).

As the training data of newer models necessarily starts inching further away from GPT-3's training data, the less relevant this specifically targeted dataset will be. I would have liked to have seen a more training-dataset-robust benchmark.

Their definition of truthfulness — simply not stating falsehoods — also favors smaller models which might produce short, unrelated, uninformative or otherwise significantly limited content.

Finally, the trend of larger LLMs producing more falsehoods simply does not hold when truthfulness and informativeness are jointly considered. In fact, the opposite trend seems to appear (observable in Figure 4).

Thus, while the technical contribution itself is worthwhile (such as the benchmark, and the automatic evaluation models), the conclusions presented in this paper cannot be so easily accepted, as they have been deliberately skewed from the start, in various ways (from the construction of the dataset to the definition of truthfulness and the disregarding of informativeness in the conclusions). I would have liked to see a more complete acknowledgement of these limitations. The authors are aware of them, as for example, the last two sentences of the abstract indicate ("*However, this result is expected if false answers are learned from the training distribution. We suggest that scaling up models alone is less promising for improving truthfulness than fine-tuning using training objectives other than imitation of text from the web.*"). I found this self- awareness inconsistent with the main narrative presented in the paper. This dichotomy is the main reason for my classification. I will increase it if a clearer discussion of design choices implications and model training suggestions is added.

References