Factual Generalization Capabilities of GPT-3 Across Domains

Muyang Xu
mx648@nyu.edu

Tian Jin
tj1059@nyu.edu

Dennis Hu
sh5701@nyu.edu

MOTIVATION

Inspired by the TruthfulQA paper and Professor Bowman's project idea, we have a strong interest in the "factuality" of the question-answering NLP task: provide short context for GPT-3 on both closely-linked and far-apart topics; how will the truthfulness of the answers generalize from one to another? Will prompt-tuning effectively help avoid reverting to popular answers? GPT-3 has been the focal point of prompt-tuning research as a large language model. However, there has been no conclusive study utilizing GPT-3 to generalize from one factual statement to answer another question in a different domain truthfully. Getting knowledge in this respect will enable an inexpensive way of promoting the model to refrain from popular but false answers.

PLAN

First, we plan to conduct a baseline evaluation of GPT-3's performance on TruthfulQA without prompts. After that, we aim to explore methods of domain similarity scoring to measure the resemblance between the 38 categories covered in TruthfulQA. Then, we will pass the processed AdversarialQA training data into GPT-3 to carry on with the prompt tuning. Furthermore, with a graph detailing the intra-topic relations in previous similarity analysis, we feed the question and correct answers from one domain into GPT-3 as a prompt and solicit responses from the model in a grid-like manner. For instance, we take a question from category one as the prompt, then combine it with questions from all 38 categories (including the questions from category one itself, which will have a similarity score of 1 out of a scale of 1) as an input batch. This will indicate whether or not improvements in truthfulness over the baseline are positively correlated with domain similarity. In terms of the evaluation process, we will first use the remaining testing data from AdversarialQA; then, we will generalize our analysis on the selected TruthfulQA dataset.

(a) One possible advanced research question we may focus on is whether the domains with adversarial texts have a more robust generalized capability. According to AdversarialQA, a model trained on a stronger adversarial dataset may perform a strong generalization to a non-adversarial dataset.[3] We will pass adversarial and non-adversarial groups into the GPT-3 separately. **(b)**Another potential concern is that the number of input prompts from the combinations will amount to approximately 817╳2, which is quite massive. We may deliberately select data with distinct domain features. **(c)**Further investigations can include applying the abstention calibrator [1] to our few-shot context and incorporating more subtle abstention cues than one abstention example into the prompt. **(d)**Fine-tuning possibilities in the OpenAI CLI remain, depending on the dimensions of training data generated. If budget permits, we also consider fine-tuning GPT-3 on AdversarialQA and evaluating on TruthfulQA.

DATA AND TOOLS

Our project mainly focuses on **using GPT-3** through the OpenAI API. We will use **two datasets: TruthfulQA[2] and AdversarialQA[3].**

- **AdversarialQA:** 36000 samples with adversarial annotations collected from three progressively stronger models (BiDAF, BERT, RoBERTa). For prompt learning, we will randomly select some data from the benchmark with a proportion of 1:2:3 from D_{BiDAF}, D_{BERT}, $D_{RoBERTa}$ (in ascending order of adversariness). First, we will standardize each AdversarialQA data with three specific labels – reference passage, question, and true short answer. Then, we will manually split the dataset within the initial ratio being 0.9:0.1 (train: test). Additionally, we will attribute those collected data to different domains defined in the TruthfulQA and count the number of entries in each domain.
- **TruthfulQA:** 817 questions with 38 domains. Those questions are designed for eliciting imitative falsehoods. We will directly extract all questions from the benchmark and then classify them twice based on two standards: adversariness and category.

CITATION

[1] A. Kamath, R. Jia, and P. Liang, *Selective question answering under domain shift*, arXiv:2006.09462.

[2] S. Lin, J. Hilton, and O. Evans, *TruthfulQA: Measuring how models mimic human falsehoods,* arXiv:2109.07958.

[3] Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp, *Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension*, arXiv:2002.00293.