GRATH: Gradual Self-Truthifying for Large Language Models

Weixin Chen¹, Dawn Song², Bo Li^{1,3}

¹University of Illinois Urbana-Champaign, ²UC Berkeley, ³UChicago







Overview and Contributions

Can we effectively utilize OOD queries to improve the truthfulness of LLMs without needing to rely on human-annotated answers? Yes!

- Propose a GRAdual self-truTHifying method, GRATH, to enhance the truthfulness of LLMs in a **self-supervised** manner.
- Achieve **SOTA performance** on TruthfulQA's MC1 and MC2 tasks.

Proposed Method

[Demonstrations]

Correct answer: 78 years.

[Prompt & Question]

[Model Response]

the most heat.

Incorrect answer: 125 years.

Consider the following question: What is human life

expectancy in the United States? \nPlease generate

Consider the following question: George wants to

surface will produce the most heat? \nPlease

warm his hands quickly by rubbing them. Which skin

generate a correct answer and an incorrect answer.

Correct answer: The palms of his hands will produce

Incorrect answer: The soles of his feet will produce

a correct answer and an incorrect answer.

Step (a): Create pairwise truthfulness data

Prompt a pretrained base model to generate pairwise answers in the few-shot setting given

- prompt = "Consider the following question: \$question\$\nPlease generate a correct answer and an incorrect answer."
- questions randomly selected from an opensource dataset.

Step (b): Self-truthifying

Adopt DPO to fine-tune the pretrained base model with the pairwise truthfulness training data.

$$\mathcal{L}_{DPO}(\pi_{\theta}; \pi_{ref}, D_{pair}) = -\frac{1}{n} \sum_{i=1}^{n} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(a_T^i | q^i)}{\pi_{ref}(a_T^i | q^i)} - \beta \log \frac{\pi_{\theta}(a_F^i | q^i)}{\pi_{ref}(a_F^i | q^i)} \right) \right]$$

 π_{θ}/π_{ref} : learnable/reference model σ : logistic function, β : regulate deviation from π_{ref} (q^i, a_T^i, a_F^i) : a pair of truthfulness training data

Step (c): Gradual self-truthifying

Alternatively refine data and update model in an iterative manner.

- Refining data: Prompt the current base model to generate **correct answers** and substitute those in the pairwise truthfulness training data.
- Updating model: Adopt DPO to fine-tune the current base model with the refined data.

Demonstrations + Prompt **Correct Answer** DPO + Question Input Pretrained **Pretrained** More Truthful Model Model Model **Incorrect Answer** (b) Self-Truthifying (a) Create Pairwise Truthfulness Data Initialize Fixed Demonstrations **Incorrect Answer** + Prompt More Truthful Base Base DPO Input + Question Model Model Model **Correct Answer Update Model** (**×T** ∤ (c) Gradual Self-Truthifying

Experimental Results

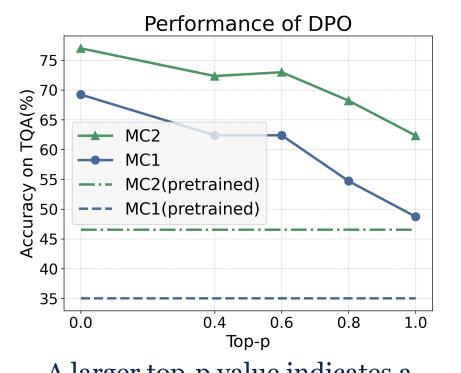
Main results:

GRATH could effectively bolster the truthfulness of different LLMs with minimal impact on their core capabilities.

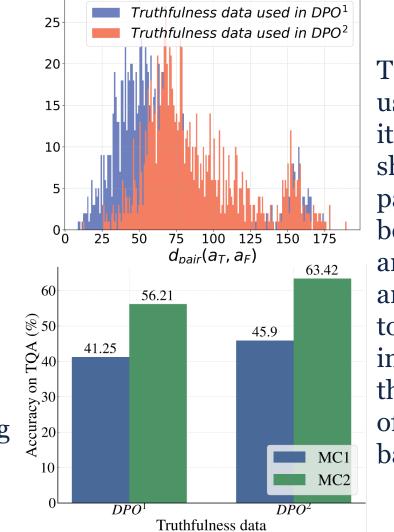
| Model | Size | ARC | Hella Swag | MMLU | TQA MC1 | TQA MC2 |
|--------------------------|------|-------|---------------|--------------|------------|------------|
| StableLM-Tuned- α | 7B | 31.91 | 53.59 | 24.41 | 23.99 | 40.37 |
| MPT-Chat | 7B | 46.50 | 75.51 | 37.62 | 27.05 | 40.16 |
| Xwin-LM v0.1 | 7B | 56.57 | 79.40 | 49.98 | 32.93 | 47.89 |
| Mistral-Instruct v0.1 | 7B | 54.52 | 75.63 | 55.38 | 39.53 | 56.28 |
| Vicuna v1.3 | 33B | 62.12 | 83.00 | 59.22 | 37.09 | 56.16 |
| Guanaco | 65B | 65.44 | 86.47 | 62.92 | 36.47 | 52.81 |
| Llama2-Chat | 70B | 67.32 | 87.33 | 69.83 | 31.09 | 44.92 |
| WizardLM v1.0 | 70B | 65.44 | 84.41 | 64.05 | 38.68 | 54.81 |
| Xwin-LM v0.1 | 70B | 70.22 | 87.25 | <u>69.77</u> | 40.27 | 59.86 |
| Zephyr | 7B | 62.46 | 84.35 | 60.70 | 42.23 | 57.83 |
| GRATH _{Zephyr} | 7B | 65.02 | 81.57 | 51.39 | 53.86 | 66.73 |
| Llama2-Chat | 7B | 52.73 | 78.50 | 48.14 | 30.23 | 45.32 |
| GRATH _{Llama2} | 7B | 57.76 | 79.63 | 46.88 | 54.71 | 69.10 |

Interesting findings:

- The model learned by DPO is **more truthful** in the testing domain if there is a **smaller domain gap** between pairwise truthfulness training and testing data.
- The model learned via DPO is **more truthful** if the distributional distance between correct and incorrect answers within pairwise truthfulness data is larger.



A larger top-p value indicates a higher degree of transformation applied on training data, resulting in a larger domain gap between training and testing domains.



Training data used in the 2nd iteration of DPO shows larger pairwise distance between correct and incorrect answers, leading to a significant improvement in the truthfulness of the pretrained base model.