UNIVERSITY of WASHINGTON

Fact-Retrieval and Ensemble Models on OpenBookQA

Chung Yik Edward Yeung, Andrew Nakamoto

Benchmark Introduction

OpenBookQA is a benchmark consisting of 4-option multiple-choice questions involving common sense reasoning about science facts. A set of 1,326 core science facts related to the questions are provided. Crowd workers have 92% accuracy on the benchmark.

Example question:

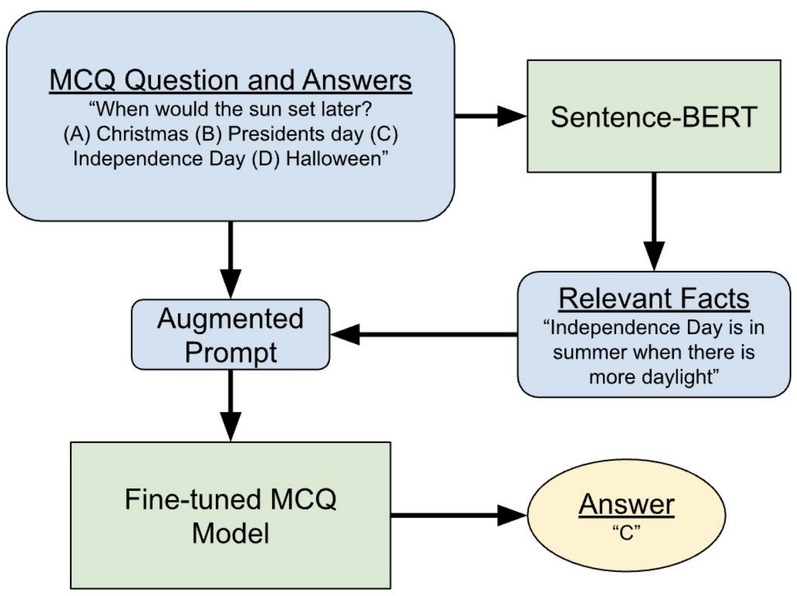
> Some plants are easy for hummingbirds to sip nectar from. While the hummingbirds are taking nectar, the plants also gain something, which is
> A. bird friends
> B. more nectar
> C. extra food
> D. a pollinator

We chose this task because the combination of facts with multiple choice questions encourages models to both use latent knowledge as well as search for specific pieces of helpful information.

This task is important because it evaluates critical common-sense and fact-based reasoning skills that language models have historically struggled at. If this problem can be solved well, it can potentially lead to models that produce false or erroneous text at much lower frequency.

Method

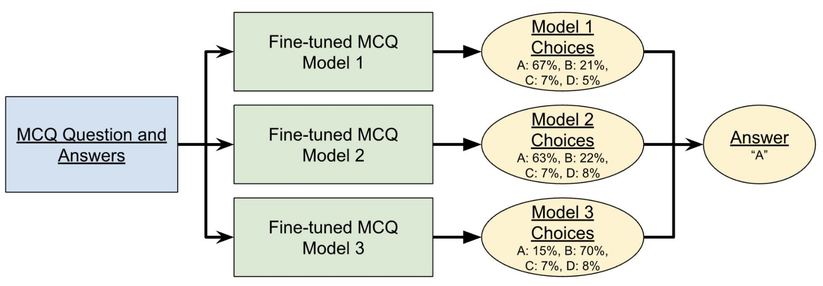
Fact-retrieval Model Pipeline



Fact-retrieval is a two-model system that produces a superior overall model

Ensemble leverages multiple separately-fine-tuned models to reduce variance

Ensemble Model Pipeline



References

Yongfeng Huang, Yanyang Li, Yichong Xu, Lin Zhang, Ruyi Gan, Jiaxing Zhang, and Liwei Wang. MVP-tuning: Multi-view knowledge retrieval with prompt tuning for commonsense reasoning.

Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao, and Gong Cheng. Clues before answers: Generation-enhanced multiple-choice qa, 2022.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system,2020.

Zhifeng Li, Yifan Fan, Bowei Zou, and Yu Hong. Ufo: Unified fact obtaining for commonsense question answering, 2023.

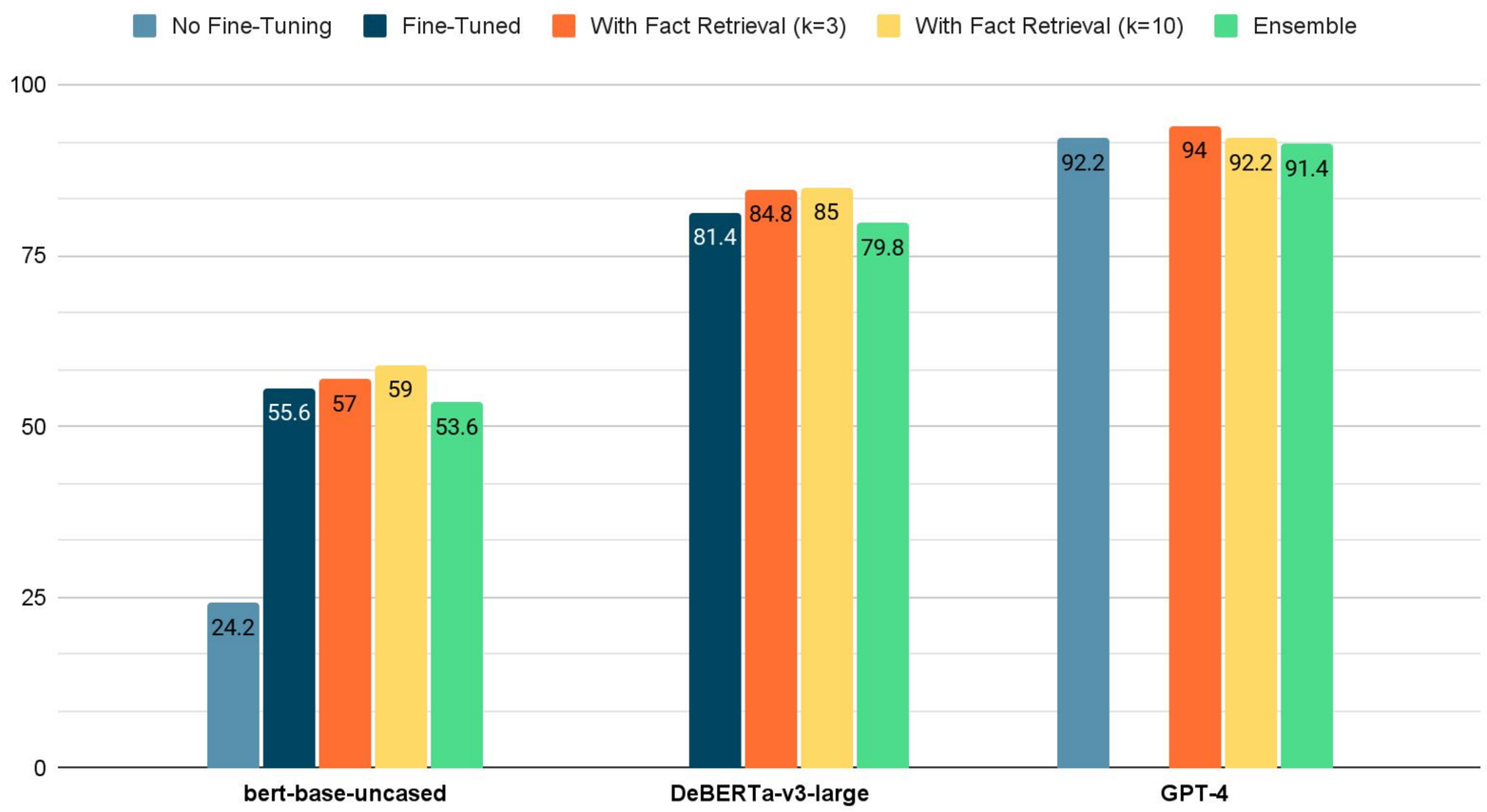
Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

Kuan Wang, Yuyu Zhang, Diyi Yang, Le Song, and Tao Qin. Gnn is a counter? Revisiting gnn for question answering, 2021.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining, 2022.

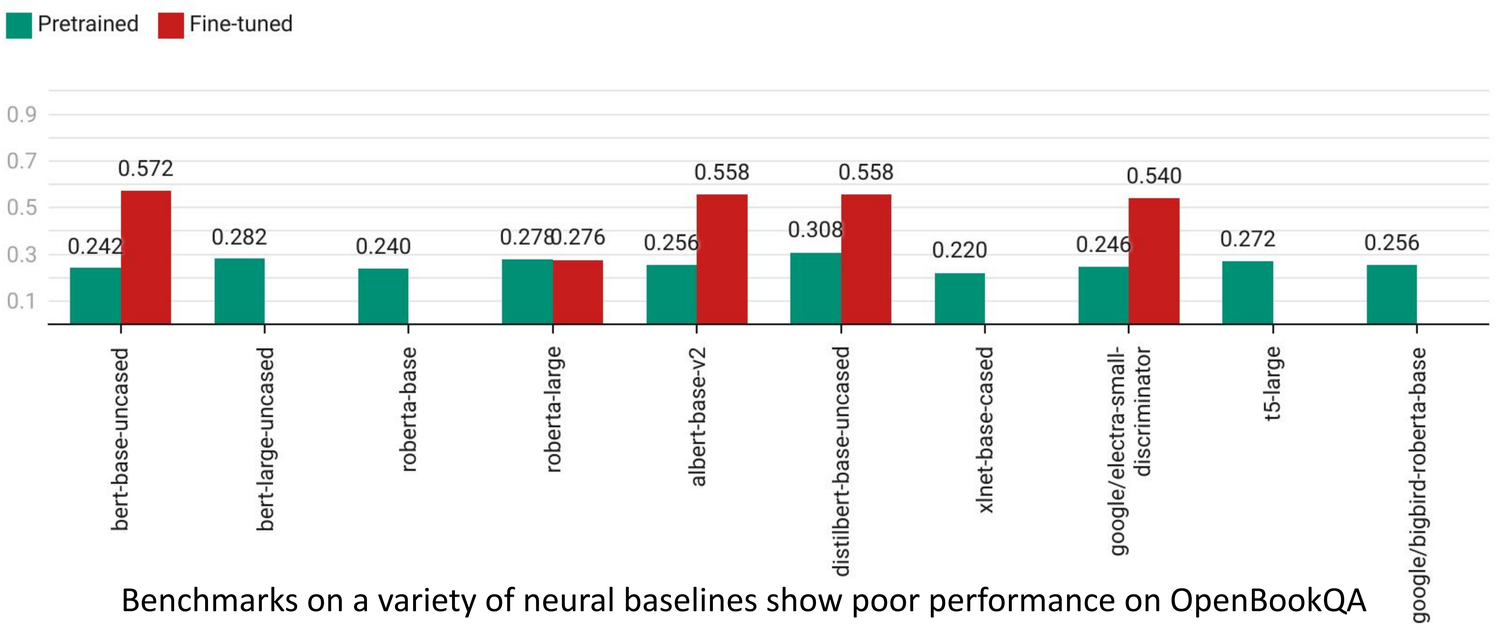
Experiments and Results

Accuracy Improvement of our Methods



Fact-retrieval gains several points of performance, even with models that already perform exceptionally well

Baseline Models Performance on 500 Questions



Benchmarks on a variety of neural baselines show poor performance on OpenBookQA

- We fine tuned google-bert/bert-base-uncased and microsoft/deberta-v3-large with no fact retrieval, with 3 retrieved facts, and 10 retrieved facts. Both showed meaningful accuracy improvements with fact-retrieval.
- We called the GPT-4 API with prompts that had no facts, 3 facts, and 10 facts. The 3-fact prompts resulted in meaningful accuracy improvement.
- We fine-tuned 3 separate models without fact retrieval for both bert-base-uncased and deberta-v3-large. In an ensemble setup, both performed worse than the individual models. GPT-4 also performed worse when prompted 3 separate times in an ensemble setup.
- This may be due to the fact only 3 models were used, not more. It may also be that ensembles are more beneficial for clue-generating models due to their higher variance.

Conclusion

- Fact-retrieval improves accuracy on OpenBookQA by around 3.5% with fine-tuning.
- Excessive facts may harm the performance of top models.
- Pretrained models perform incredibly poorly on the benchmark, but minimal (1 epoch) fine-tuning can quickly increase performance.
- We suggest that our findings imply that science-based common sense is reliant on strong fact retrieval brought out
- Although clue-generation models show improvement when used in ensemble (Huang 2022), we do not see the same boost when an ensemble strategy is applied to other types of models.