Literature Review: What Makes Good In-Context Examples for GPT-3?

Wesley Elliott Stephen Monteith-Finas | 324745 | wesley.monteith-finas@epfl.ch
Chatbots-R-Us

1 Summary

This paper (Liu et al., 2022) explores the effectiveness of in-context learning in GPT-3, focusing on how the selection of in-context examples can enhance the model's performance across various natural language processing tasks. The authors introduce a method called KATE (KNN And Text Embeddings) which leverages a k-nearest neighbor (kNN) algorithm to select semantically relevant examples to the question being asked. This assists GPT-3 in generating more accurate outputs without the need for fine-tuning. The metric used, to determine similarity is specified to be either the Euclidean distance or the cosine similarity in the latent embedding dimension. The paper evaluates their novel method on tasks such as sentiment analysis, table-to-text generation, and open-domain question answering, using datasets like SST-2, IMDB, ToTTo, and others.

The core claim of the paper is that the strategic selection of in-context examples, using pretrained sentence encoders and fine-tuned models on specific tasks, significantly improves the performance of GPT-3. The authors argue that this method outperforms both random selection and traditional kNN approaches that do not use fine-tuned embeddings.

Problem Identification: The authors identify the problem of optimizing in-context learning in GPT-3, particularly how the selection of in-context examples impacts the model's performance. They argue that while GPT-3 has shown impressive capabilities in few-shot learning, its performance can be inconsistent and highly dependent on the examples provided.

Motivation: The motivation to improve the reliability and effectiveness of GPT-3 through better in-context example selection is well-founded and addresses a real and practical problem in the deployment of AI in natural language tasks. The approach makes sense as it aims to leverage existing capabilities of a promising model while minimizing the need for costly retraining or fine-tuning on specific tasks. Moreover, this problem is relatable and was particularly apparent for the data collection task of milestone I.

2 Strengths

Innovative Approach: The introduction of KATE as a method to enhance in-context learning in GPT-3 is innovative. It addresses the limitations of GPT-3's few-shot learning capabilities by providing a systematic way to select the most relevant examples.

Rigorous Evaluation Process: The paper presents a thorough evaluation across multiple tasks and datasets.

They first selected in-context examples from the SST-2 training set and evaluated KATE on the IMDB test set for sentiment prediction. They then compared KATE using different sentence encoders and used accuracy as the evaluation metric.

Regarding Table-to-Text Generation, they used the ToTTo dataset for evaluation and employed BLEU and PARENT metrics for evaluation. They compared KATE against random selection and kNNroberta baselines.

Regarding Open-Domain Question Answering, they evaluated on Natural Questions (NQ), Web Questions (WQ), and TriviaQA datasets. They used exact match (EM) score as the evaluation metric, compared against state-of-the-art methods like RAG and T5, and evaluated on the test sets of NQ and WQ, and dev set of TriviaQA.

As an ablation study, they analyzed the impact of the number of in-context examples (ranging from 5 to 64).

Finally they also present different case studies, such as qualitative examples showcasing the retrieved in-context examples by KATE, or comparing KATE's outputs with the random baseline to

try and understand failure cases.

This extensive testing not only demonstrates the versatility of KATE but also provides a robust validation of the method's effectiveness.

Clear Improvement Over Baselines: The results clearly show that KATE outperforms baseline methods, including random selection and standard kNN approaches. This is particularly evident in the sentiment analysis task on the IMDB dataset and the question answering tasks, where KATE achieves higher accuracy and Exact Match scores. This supports the authors' claim that careful selection of in-context examples can significantly boost GPT-3's accuracy.

Effectiveness of Fine-Tuned Embeddings: The paper demonstrates that using sentence embeddings fine-tuned on task-specific datasets (e.g., KATEsst-2 for sentiment analysis, KATEnli+sts-b for QA) leads to better performance compared to general pre-trained embeddings like RoBERTa. This validates the authors' claim about the importance of using fine-tuned embeddings for retrieval.

Complementary to GPT-3: The results show that KATE's retrieval module complements GPT-3's few-shot learning capabilities, as evidenced by the poor performance of the kNN baselines without GPT-3. This supports the authors' claim that KATE enhances GPT-3's existing capabilities.

Experimental Reproducibility: The authors have provided the code and identified the datasets used in their experiments, which aids in reproducing their work experimentally. They mention using publicly available datasets like SST-2, IMDB, ToTTo, Natural Questions (NQ), Web Questions (WQ), and TriviaQA for evaluation. The ToTTo code base used for preprocessing and evaluation is cited.

Conceptual Reproducibility: The paper provides sufficient conceptual details and descriptions to reproduce the core ideas and methods proposed. The KATE (kNN and Text Embeddings) method for selecting in-context examples is clearly explained. Experimental setup, baseline methods, and evaluation metrics are well-documented. Ablation studies and analyses on different hyperparameters and settings are included. It was made possible for our group to implement this method for the data collection of milestone I.

Exploring the impact of ordering the examples: The authors conduct an ablation study to analyze the impact of the order in which the retrieved in-context examples are presented to GPT-3. They experiment with different orderings of the in-context examples on the Natural Questions (NQ) dataset using the KATEnli+sts-b method, which uses the CLS embeddings of RoBERTa model fine-tuned on the NLI and STS-b datasets. The authors note that the variation in results across different orderings is quite small compared to the difference between KATE and the random baseline. They conclude that while the choice of order can be data-dependent, the order of in-context examples does not have a significant impact on KATE's overall performance.

3 Weaknesses

Complexity and Scalability: While KATE improves performance, the method's reliance on fine-tuned sentence encoders could introduce additional complexity and computational costs, which may affect scalability, especially for large datasets or real-time applications.

Dependence on Quality of Pre-trained Models: The performance of KATE heavily depends on the quality of the pre-trained sentence encoders used. This could limit its effectiveness if suitable pre-trained models are not available for specific tasks or languages.

Limited Analysis on Failure Cases: The paper could benefit from a more detailed analysis of cases where KATE does not perform well. Understanding these scenarios could help in further refining the approach and extending its applicability.

Missing Analyses: While the paper provides extensive empirical results, it lacks a deeper theoretical analysis of why certain types of in-context examples work better than others. Additionally, an exploration of the impact of the size and diversity of in-context examples could provide further insights into how to optimize the selection process for different applications

Addressing Ethics Issues: The paper does not explicitly discuss or address potential ethical concerns. While the paper does not directly mitigate ethical issues, the authors' focus on improving the reliability and effectiveness of GPT-3 through better in-context example selection could indirectly contribute to mitigating potential misuse or generation of harmful content. Overall, the paper provides good experimental and conceptual reproducibility, but it lacks a dedicated discussion on ethical considerations and mitigation strategies.

References

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.