WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 1, Number 1, Year 2023
DOI: https://doi.org/10.62051/ijcsit.v1n1.17
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Optimizing Science Question Ranking through Model and Retrieval-Augmented Generation

Ye Zhang[1,*], Mengran Zhu[2], Yulu Gong[3], Rui Ding[4]

[1] Independent Researcher, University of Pittsburgh, Pittsburgh, USA

[2] Independent Researcher, Miami University, Oxford, USA

[3] College of Engineering, Informatics, and Applied Sciences, Northern Arizona University, Flagstaff, USA

[4] Independent Researcher, San Francisco Bay University, San Francisco, USA

*Corresponding Author: Ye Zhang

ABSTRACT

This paper delves into the challenges of discerning optimal answers from science-based questions generated by large language models (LLM), particularly emphasizing the intricate task of ranking. Employing the MAP@3 evaluation metric and drawing from the OpenBookQA dataset, the study explores modeling strategies and highlights the exceptional performance of the Platypus2-70B model. Equipped with a state-of-the-art text encoder, Platypus2-70B achieves an impressive score of 0.909904, setting a benchmark for excellence in future large language model competitions. The paper goes beyond a mere description of model architectures and experimental results, offering a comprehensive journey that envisions the transformative impact of large-scale language models on the landscape of natural language understanding, especially within the intricate domains of scientific exploration.

KEYWORDS

Large language model; OpenBookQA, ranking; Science-based questions; Platypus2-70B.

1. INTRODUCTION

In the rapidly advancing field of natural language processing (NLP), the burgeoning scale and intricacies of language models necessitate a comprehensive examination of their capabilities and methodologies. This study addresses the intricate challenge of ranking science-based questions, shedding light on the nuanced competencies required for effective performance.

Within the contemporary epoch marked by the meteoric growth of language models, this research stands as a discerning inquiry into their efficacy, particularly within the intricate landscapes of science, technology, engineering, and mathematics (STEM) disciplines. Rooted in the meticulously curated OpenBookQA dataset, the research explores not only linguistic intricacies but also pioneers through the multifaceted terrains of STEM knowledge.

This exposition commences with a meticulous dissection of the multifaceted challenges inherent in science-based queries, underscoring the paramount role played by external knowledge integration in refining model responses. It elucidates the fusion of linguistic acuity and comprehensive STEM knowledge required for superior performance.

The subsequent sections conduct a thorough analysis of diverse modeling methodologies, with a particular emphasis on the distinguished Platypus2-70B model—a beacon adeptly navigating the labyrinthine complexities of intricate scientific inquiries. The model's hierarchical architecture, encompassing embeddings, attention mechanisms, self-attention heads, and output layers, is dissected to highlight its transformative capabilities in understanding and generating language.

Beyond the confines of mere evaluation, this exploration transcends its evaluative parameters to emerge as a herald of transformative shifts within the realm of NLP. It accentuates the confluence of contextual comprehension and ranking capabilities, signifying the maturation of language models in addressing the sophisticated nuances intrinsic to scientific inquiries.

As the subsequent sections unfold, they provide more than an exposition of model architectures and experimental results; they offer a scholarly exploration extending beyond the competitive framework. The ensuing reflective discussion envisions trajectories that not only reshape the frontiers of natural language understanding but also redefine their applications in the intricate landscapes of scientific exploration, contributing to the ongoing discourse in NLP research.

2. RELATED WORK

This section highlights key contributions in Natural Language Processing (NLP) that have profoundly influenced language model development. From foundational principles of statistical machine translation to transformative innovations like attention mechanisms, bidirectional transformers, and state-of-the-art pre-training techniques, these works collectively shape the evolving landscape of NLP. The discussed papers set the stage for our current study, providing insights into diverse methodologies and paradigms within the language model research domain.

Brown work [1] introduces a foundational framework for statistical machine translation, providing insights into parameter estimation methodologies. Vaswani [2] presents the Transformer model, pioneering the attention mechanism, and reshaping the landscape of natural language processing. BERT revolutionizes language representation learning through bidirectional pre-training, achieving state-of-the-art results in various NLP tasks [3].

BART introduces denoising sequence-to-sequence pre-training, demonstrating its effectiveness across a spectrum of language tasks [4]. Radford [5] explores generative pretraining as a means to enhance language understanding, paving the way for subsequent large-scale language models. Google's [6] Neural Machine Translation System represents a significant leap in bridging the gap between human and machine translation, leveraging neural networks for improved language understanding.

Karpukhin [7] introduces Dense Passage Retrieval, a method that significantly advances open-domain question answering through enhanced passage retrieval. The Text-to-Text Transformer represents a unified framework for transfer learning, pushing the boundaries of what can be achieved through comprehensive text understanding [8]. Le [9] introduces distributed representations for sentences and documents, laying the groundwork for comprehensive language understanding and document analysis. This groundbreaking paper introduces the Transformer model, revolutionizing NLP with the attention mechanism[10].

Vaswani [11] introduces the Transformer model, a paradigm-shifting architecture in natural language processing (NLP). The authors propose a self-attention mechanism that allows the model to capture global dependencies in input sequences, revolutionizing the field of machine translation and serving as the foundation for various NLP applications. National Academies of Sciences [12] provides a comprehensive overview of the progress and future prospects of quantum computing. It addresses scientific, technological, and societal aspects, offering a roadmap for understanding the potential impact of quantum computing across various domains. Villar [13] explores the intersection of smart city initiatives and sustainability. The authors analyze existing literature to identify key themes,

challenges, and opportunities, providing a holistic view of the efforts to build smart and sustainable urban environments.

X Zhao's article [14] integrates 3D-DenseNet's AI with a gallbladder cancer diagnosis model, improving medical imaging accuracy and providing insights for tech advancement. Li, S.[15] investigates the integration of AI and spiral CT for early lung cancer screening in "Frontiers in Computing and Intelligent Systems," offering insights for medical imaging and cancer detection advancement.

The related work encompasses a diverse spectrum of contributions, spanning statistical machine translation, attention mechanisms, bidirectional transformers, denoising sequence-to-sequence pre-training, generative pretraining, redundancy enhancement for BERT, neural machine translation, dense passage retrieval, transfer learning with unified text-to-text transformers, and distributed representations of sentences and documents. These seminal works collectively form the backdrop against which the current research unfolds, providing insights and inspiration for advancements in the field of natural language processing.

3. METHODOLOGY

This section meticulously delineates the algorithmic framework and models employed in the study, integrating insights and methodologies from the referenced literature. The study's core lies in harnessing the capabilities of large-scale language models (LLMs), with a primary focus on the sophisticated Platypus2-70B architecture and RAG intergration. The incorporation of external knowledge, specifically the RAG method, is identified as a key factor in enhancing the model's upper limit. Micro-adjustments through fine-tuning and the combination of RAG and fine-tuning are crucial in addressing high external knowledge requirements.

3.1. Platypus2-70B Architecture Overview

Platypus-70B is an instruction fine-tuned model based on the LLaMa2-70B transformer architecture, is chosen for its remarkable ability to encapsulate nuanced contextual information within textual data. The model comprises numerous layers, including embeddings, attention mechanisms, self-attention heads, and output layers, exemplifying a transformative architecture for understanding and generating language.

Embedding Layer:

The foundational layer of Platypus2-70B is the embedding layer, responsible for transforming input tokens (X) into continuous vector representations E(X) using the token embedding function:

$$H^{(l)} = Attention\left(LayerNorm\left(Feedforward\left(H^{(l-1)}\right)\right)\right) \tag{2}$$

The Attention mechanism within each layer is defined by:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Here, Q, K, and V represent the query, key, and value matrices, and d_k is the dimension of the key vectors.

Context Pooler:

To distill relevant information from the entire passage retrieved through Retrieval-Augmented Generation (RAG), Platypus2-70B employs a context pooler. The context C is obtained by pooling over all positions i in the output sequence of the last encoder layer:

$$C = Pool\left(H^{(L)}\right) \quad (4)$$

This aggregated context enhances the model's understanding of the broader context surrounding each input sequence. These architectural components collectively contribute to Platypus2-70B's ability to comprehend and process complex language structures, making it a formidable choice for language understanding tasks in the LLM Science Exam competition.

3.2. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) stands as a sophisticated methodology employed to enhance Platypus2-70B's performance by integrating external knowledge extracted from Wikipedia. This section provides an in-depth exploration of the intricate components and processes integral to Retrieval-Augmented Generation.

3.2.1. Wikipedia Retrieval with CirrusSearch

The retriever component efficiently searches and retrieves relevant information from Wikipedia using the CirrusSearch engine. This process involves querying the Wikipedia corpus to obtain top-k passages that are most relevant to the user's input.

3.2.2. Querying Process

Input Query: The user's input question is embedded using Platypus2-70B.

CirrusSearch: The embedded query is used to search the Wikipedia corpus using CirrusSearch.

Top-K Passages: The search results provide a collection of top-k passages that are deemed most relevant to the input query.

This approach ensures that the retriever selects passages that align closely with the user's question, providing valuable context for generating accurate responses.

3.2.3. Conditional Probability Estimation

The generator component, an integral part of the Retrieval-Augmented Generation process, undertakes the task of estimating the conditional probability of potential answers. This estimation relies on the interplay between the input query Q and the retrieved top-k passages P_{top-k}.

The generator's role is to gauge the likelihood of potential answers, ensuring that the subsequent response generation process is influenced by both the model's existing knowledge and the dynamically retrieved information from Wikipedia.

$$GeneratedAnswer = argmax_{answer} P(answer|InputQuery, Top-KPassages) \quad (5)$$

Where:

P(*) denotes the conditional probability.

Input Query is the user’s question.

Top-K Passages includes the relecant information retrieved from Wikipedia.

This approach ensures that the generated answers are influenced by both the model's pre-existing knowledge and the dynamically retrieved information from Wikipedia.

3.2.4. Passage Integration with Platypus2-70B Generator

The generator component seamlessly incorporates the retrieved top-k passages into the response generation process. This integration refines Platypus2-70B's understanding of the user's query and facilitates the generation of contextually rich and informed answers.

3.3. Evaluation Metric

The evaluation metric employed in this study is MAP@3 (Mean Average Precision at 3). MAP@3 is a widely used metric for information retrieval tasks, particularly in scenarios where multiple candidate answers are ranked. It calculates the average precision across the top 3 ranked answers for each question. The formula for MAP@3 is given by:

$$MAP@3 = \frac{1}{N}\sum_{i=1}^{N}\frac{AP@3}{min(3, M_i)} \tag{6}$$

Here, N is the total number of questions, M_i is the total number of correct answers for question i, and $AP@3_i$ is the Average Precision at 3 for question i. The MAP@3 metric provides a comprehensive measure of the model's ability to rank the correct answer choices higher in the list.

3.4. Dataset

The dataset used in this study is tailored for the LLM Science Exam competition, comprising a total of 4000 questions. The questions are science-based and sourced from a variety of topics extracted from Wikipedia.

3.4.1. Data Collection Process

The dataset creation involved two main steps:

GPT-3.5 Turbo Data Generation:

The initial 200 questions were generated using the GPT-3.5 Turbo model, which was tasked with composing multiple-choice questions based on science-related passages extracted from Wikipedia.

Additional Data Augmentation:

To enhance the training set, an additional 500 high-quality questions were created using the GPT-3.5 Turbo model. This resulted in a total of 700 training examples.

3.4.2. Data Split

The dataset is split into two parts: A (20% of the total questions) and B (80% of the total questions). This split allows for model development on a smaller subset while maintaining the majority of questions for final evaluation.

This dataset split is crucial for evaluating the model's performance on unseen data, simulating real-world scenarios where the model encounters novel questions beyond its training set.

4. EXPERIMENT RESULTS

This section presents the comprehensive results of the experiments conducted, shedding light on the performance metrics and comparative analyses of different model configurations. The primary models under scrutiny include DeBERTa v3 without context, DeBERTa v3 with Wikipedia RAG, DeBERTa with STEM (270k) as context, Platypus2-70B without context, and Platypus2-70B with Wikipedia RAG.

The results clearly demonstrate that the Platypus2-70B model, both with and without Wikipedia RAG, outperforms other configurations. The incorporation of Platypus2-70B as the backbone showcases its superiority in handling science-based queries, achieving an impressive MAP@3 score of 0.909904. This highlights the efficacy of large-scale language models, particularly the Platypus2-70B variant, in addressing the challenges posed by the LLM Science Exam.

The following table presents the MAP@3 scores for each model configuration:

Model Configuration	MAP@3 Score
DeBERTa v3 without context	0.704324
DeBERTa v3 with Wikipedia RAG	0.8196
DeBERTa with STEM (270k) as context	0.862047
Platypus2-70B without context	0.858094
Platypus2-70B with Wikipedia RAG	0.909904

Additionally, the introduction of retrieval-augmented generation (RAG) techniques, especially leveraging external knowledge from Wikipedia, further enhances the model's performance. The synergy between advanced language models and external knowledge retrieval mechanisms is pivotal in achieving superior results in science-related question answering tasks.

5. CONCLUSION

This study addresses the challenges of ranking science-based questions in the LLM Science Exam. The Platypus2-70B model, with its sophisticated architecture, achieves an impressive MAP@3 score of 0.909904, setting a benchmark for excellence. The research emphasizes the fusion of linguistic acuity and STEM knowledge, envisioning the transformative impact of large-scale language models on natural language understanding within scientific domains.

Furthermore, the integration of Retrieval-Augmented Generation (RAG) with Platypus2-70B, leveraging external knowledge from Wikipedia, proves crucial in enhancing model performance. The results highlight the efficacy of this approach, showcasing the superiority of Platypus2-70B, especially with RAG, in addressing the challenges posed by science-related questions. This research not only contributes to the competitive framework of the LLM Science Exam but also advances the discourse in NLP, showcasing the evolving landscape of language models in scientific exploration.

REFERENCES

[1] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. https://bit.ly/3RNCRSY.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://bit.ly/3Sk8XUV.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805.

[4] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461. https://arxiv.org/abs/1910.13461.

[5] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[6] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. https://arxiv.org/abs/1609.08144.

[7] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906. https://arxiv.org/abs/2004.04906.

[8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485-5551. https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

[9] Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR. https://bit.ly/3UjkSoQ.

[10] Ashish, V. (2017). Attention is all you need. arXiv preprint arXiv: 1706.03762. https://bit.ly/48THJeZ.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://bit.ly/48TpJBB.

[12] National Academies of Sciences, Engineering, and Medicine. (2019). Quantum computing: progress and prospects. https://bit.ly/3SzTilU.

[13] De la Hoz, R., Villar, S., Castillo, K., Castellón, J., & Coronado, K. (2022). Factores clave para el éxito de ciudades inteligentes y sostenibles: una revisión sistemática de la literatura. INVENTUM, 17(33), 44-54. https://revistas.uniminuto.edu/index.php/Inventum/article/view/3141.

[14] Xinyu Zhao, et al. "Effective Combination of 3D-DenseNet's Artificial Intelligence Technology and Gallbladder Cancer Diagnosis Model". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 81-84, https://doi.org/10.54097/iMKyFavE.

[15] Shulin Li, et al. "Application Analysis of AI Technology Combined With Spiral CT Scanning in Early Lung Cancer Screening". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 52-55, https://doi.org/10.54097/LAwfJzEA.