SQuADBrT: Different Baselines and Re-think Models on the SQuAD

Shenghan Wu * **Peike Wang** † **Yancan Chen** ‡ **Xingchen Li** § **Xuefeng Zhang** ¶

Abstract

The paper examines Machine Reading Comprehension (MRC) using the SQuAD dataset, focusing on Extractive and Generative MRC techniques. We introduce a self-correcting Large Language Model (LLM) that uses external feedback and a 'Re-think' strategy within T5 and RoBERTa frameworks to improve response accuracy. Our approach, tested with rigorous metrics, shows marked improvements in LLM accuracy and F1 scores, affirming the value of feedback and re-evaluation in MRC. The T5-3B model excels, achieving 95.04% F1 and 90.06% accuracy with the rethink strategy among 6 models.

1 Introduction

Machine Reading Comprehension (MRC), a challenging work in natural language process, are aimed at building an ability to read text and then answer questions about it. This work not only requires understanding of natural language but also the knowledge about the real world. To answer the question "What does photosynthesis produce?", one could locate the pertinent segment of the text, "oxygen and glucose," understand that "produce" indicates the results, and thereby conclude that the correct answer is "Oxygen and glucose."
In the project, we compared different types of models. There are two distinct types of technologies for MRC, Extractive MRC and Generative MRC respectively. The SQuAD (Rajpurkar et al., 2016) dataset fits both of technologies. Extractive MRC approach (like RoBERTa) involves answering specific questions where the answer explicitly exists within the text. And Generative MRC (like LLaMA 2, T5, Flan-T5) involves understanding the overall context of the text to generate an accurate answer. In this project, Our approach is depending on exist LLM to achieve Machine Reading Comprehension. By the training with SQuAD, T5 model get acceptable accuracy and F1 Score. Our team further increases the parameters, and we get outstanding improvement in result. For further improving the performance, our team use Chain-of-Thought and Re-think Strategy. Chain-of-Thought help us get outstanding improvement in experimental result.However, testing in T5, our data in this project show limited re-think maybe not helpful for the performance. Therefore, We re-train Llama2. By the increasing the times of re-think, the performance trend to increase. Our project performance is totally better than human performance(86.8%)(Rajpurkar et al., 2016).

2 Related Works

2.1 Extractive MRC: RoBERTa

RoBERTa(Liu et al., 2019), an encoder-only model based on Transformers (Vaswani et al., 2017). It extends BERT (Devlin et al., 2018) by training on a larger dataset, omitting the Next Sentence Prediction (NSP) task, dynamically changing the masking pattern during **Masked Language Model** (MLM) training, and employing a larger Byte-Pair Encoding (BPE)(Sennrich et al., 2015). Those optimizations allow RoBERTa to deliver state-of-the-art performance on many NLP tasks. In the Question Answering benchmark, Roberta extracts text segments from the given context via marking the position of answer, which is the key part of the Extractive MRC.

2.2 Generative MRC: T5/Flan-T5

T5 (Raffel et al., 2020a), an encoder-decoder based framework, treats every text processing problem as text-to-text problem. Based on the previous method and large data set, T5 model allows researchers to

*E1101760@u.nus.edu
†E1132289@u.nus.edu
‡E1143641@u.nus.edu
§E0945876@u.nus.edu
¶E1132343@u.nus.edu

use the same model, loss function, hyperparameters across diverse set of tasks. It has a pre-train model and provides a simple way to train our task using the same loss function and decoding procedure. T5 inherits Transformer structure and makes some changes. It removes the layer Norm Bias and places the layer Normalization outside the residual path. **Flan-T5** (Chung et al., 2022) is a resulting of finetuned T5. By scaling the number of tasks, scaling the model size, and finetuning on **chain-of-thought (CoT)** (Wei et al., 2022) data, it achieves strong few-shot (Brown et al., 2020) performance even compared to much larger models.

2.3 Generative MRC: LLaMA 2

LLaMA 2 (Meta, 2023) , a decoder-based framework,focuses on optimizing the decoder part of Transformers (Vaswani et al., 2017). It adopts the **RMS Norm (Root Mean Square layer normalization)** for better training computational performances, which lead to reducing computational time by 7% to 64% (Zhang and Sennrich, 2019). The activate function, SwiGLU, achieves the optimal value in log-perplexity (Shazeer, 2020). Additionally, **Rotary Positional Embedding (RoPE)** is utilized, combining the benefits of both the convenience of absolute position encoding and the representation of relative positional relationships between tokens (Su et al., 2021). Byte-Pair Encoding (BPE) is also empolyed in this model. Importantly, Llama 2 increased **context length** and **Grouped-Query Attention (GQA)** in architectural differences (Meta, 2023). **GQA** achieves quality close to multi-head with less computation (Ainslie et al., 2023).

2.4 Re-think: External Feedback

Mirroring human learning strategies of error and correction, LLMs can be improved through the paradigm of learning from feedback (Huang et al., 2022; Madaan et al., 2023; Huang et al., 2023; Gero et al., 2023; Jiang et al., 2023). External feedback provides an invaluable outside perspective crucial for identifying and correcting errors not self-identifiable by LLMs(Pan et al., 2023). Furthermore, the most effective method for LLMs to correct their outputs is through external feedback, enabling them to verify the accuracy of their responses by interacting with external environments(Gou et al., 2023). This approach ensures a more accurate and reliable performance from LLMs, like a human's ability to learn and adapt from external inputs.

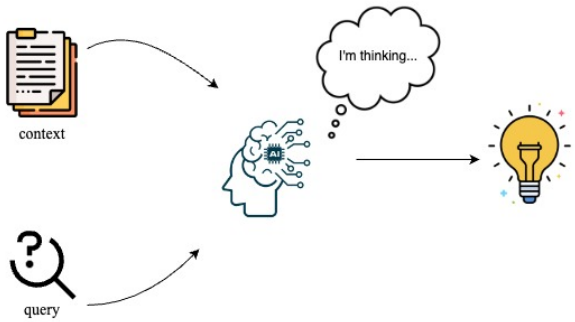

Figure 1: The process of the vanilla model.

2.5 Application areas of rethink

The **re-think** technology has been applied various domains, including Information Extraction (Gero et al., 2023), Code Generation (Jiang et al., 2023; Ni et al., 2023) and Explainable Text Generation Evaluation (Xu et al., 2022). Since LLMs might generate incorrect response, the external feedback is an easier way for a better performance, say, using the same LLM, guided by different prompts (Pan et al., 2023).

3 Method

3.1 Vanilla Model

In the training of our baseline model, we amalgamate the contextual information and queries derived from the SQuAD dataset, utilizing both as inputs in conjunction with their corresponding answers for model training. For queries that encompass multiple alternative answers, each answer is treated as an individual training instance. Throughout the generative phase, the context and query are amalgamated to facilitate the trained model in generating appropriate responses to the queries. The process of vanilla model is shown on the Figure 1.

To discover the foundational performance across different model architectures on the SQuAD dataset, we fine-tuned three types of individual models: RoBERTa (encoder-only model) (Liu et al., 2019), LLaMA (decoder-only model) (Meta, 2023), and T5 (encoder-decoder model) (Raffel et al., 2020b).

RoBERTa represents an evolution and optimization of the earlier BERT (Devlin et al., 2018) model, which itself marked a significant advancement in the field of NLP, also showing one of the top performances on the SQuAD task. In our survey, we

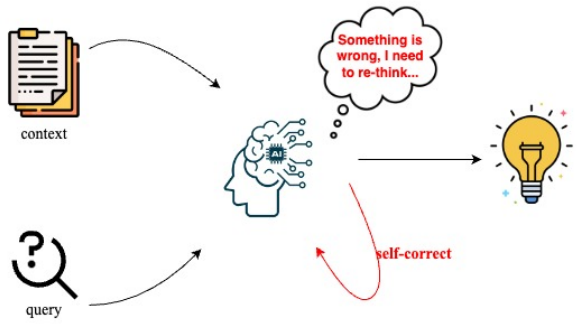

Figure 2: self-correct-model

chose **RoBERTa-base** and **RoBERTa-large** as our candidates.

LLaMA represents a peak in instruction-based models, demonstrating an excellent performance across a wide range of tasks. Considering the computational ability, we chose the most advanced **LLaMA2 7B** as our candidate model.

T5 is esteemed within the domain of transfer learning for its proficiency about redefining a wide range of natural language processing tasks into a text-to-text format. We selected **T5-base**, **Flan-T5-base**, and **T5-3B** as the candidate models for our analysis.

3.2 Self-correct Model

We present the self-correct model on the SQuAD task, which characterize by rethinking the answer to a query based on the external feedback. Recent research (Pan et al., 2023; Huang et al., 2023; Gou et al., 2023) highlights the crucial significance of external feedback in fostering the ongoing self-improvement of LLMs, which proves the external feedback may improve the performance on the question-answering task, like SQuAD.

We propose a framework, shown in the Figure 2, that is designed to assess the accuracy of the generated answers. While the current answer diverges from the ground-truth, we construct a prompt with this incorrect answer to allow the model regenerate the answer of the query. The definition of the refined prompt is shown in the Figure 3.

4 Experiment

4.1 Corpus

We utilized the **Stanford Question Answering Dataset (SQuAD)** for both training and testing phases of our project.

${context}
${query}
${incorrect answer} is the wrong answer, please think again.

Figure 3: prompt

4.2 Data Preprocessing

For our data preprocessing, considering the presence of multiple answers for some questions in the test set, we transformed all questions from **442 topics** in the training set into **87,599 question-answer pairs**. Similarly, from the test set, we generated **10,570 question-answer pairs**, corresponding to all questions across **48 topics**.

4.3 Model Evaluation

To evaluate model performance, we employed two distinct metrics, both disregarding punctuations and articles (a, an, the). The first, **Exact Match**, gauges the proportion of predictions that precisely align with any of the ground truth answers. The second, **F1 Score**, assesses the average overlap between predictions and truth answers.

4.4 Baseline

To establish a baseline, we experimented with two variants of the **RoBERTa** model. The **RoBERTa-base**, with **123 million parameters**, served as our initial benchmark. The **RoBERTa-large**, an enhanced version with **354 million parameters**, was expected to yield a more nuanced understanding of language semantics and structure. As illustrated in Table 1, our results indicate that models with more parameters tend to deliver better performance.

Model Name	Accuracy	F1 Score
RoBERTa-base	85.64	92.10
RoBERTa-large	88.04	94.16

Table 1

4.5 More Works

In pursuit of further advancements, we trained the **T5 (Text-to-Text Transfer Transformer) base** model with 220 million parameters, its larger counterpart, the **T5-3b** with 3 billion parameters, and the **Llama2** model with **7 billion parameters**. As depicted in Table 2, the T5-3b outperformed others,

showcasing superior efficacy despite having fewer parameters than Llama2.

Model Name	Accuracy	F1 Score
T5-base	83.42	90.25
T5-3b	89.93	95.04
Llama2-7B	86.58	93.35

Table 2

4.6 Chain-of-Thought

We incorporated the **chain-of-thought** approach using the **FLAN-T5-base** model. This model, leveraging the **T5-base architecture**, was fine-tuned to foster chain-of-thought reasoning, prompting the generation of intermediate reasoning steps before concluding with a final answer. As shown in Table 3, this approach significantly improved both accuracy and F1 scores compared to the T5-base model.

Model Name	Accuracy	F1 Score
T5-base	83.42	90.25
FLAN-T5-base	86.06	92.21

Table 3

4.7 Re-think Strategy

To further enhance performance, we implemented a **'Re-think'** strategy, enabling models to revisit and potentially correct their initial answers. Table 4 presents the results of applying this strategy to the **T5-3B** and **Llama2-7B** models. We observed a notable increase in accuracy, albeit with a slight reduction in F1 scores. Furthermore, when the Llama2-7B model was prompted to 'rethink' three times in response to incorrect Q&A pairs, we saw additional improvements in accuracy, though F1 scores experienced a minor decline.

Model Name	Accuracy	F1 Score
T5-3b	89.93	95.04
T5-3b-rethink	90.06	93.37
Llama2-7B	86.58	93.35
Llama2-7B-rethink	87.18	92.20
Llama2-7B-rethink-3-times	89.37	93.18

Table 4

5 Conclusion

We have presented a comprehensive exploration of Machine Reading Comprehension (MRC) with a focus on the Stanford Question Answering Dataset (SQuAD). Our work centered on two distinct types of MRC technologies - Extractive and Generative MRC - and emphasized the integration of external feedback mechanisms in language model training. We introduced a novel self-correct model, leveraging recent advancements in encoder-decoder frameworks like T5 and RoBERTa, to improve the quality of answers generated by these models.

Our findings underscore the significant role of external feedback in refining the learning and response-generation process of LLMs. This approach, which simulates human error-correction learning strategies, has proven effective in enhancing the performance of LLMs on question-answering tasks. By implementing a 'Re-think' strategy, our models were able to revisit and adjust their initial answers, leading to notable improvements in accuracy.

The efficacy of our approach was validated through rigorous data preprocessing and model evaluation methods. We transformed a large set of questions from the SQuAD dataset into question-answer pairs, providing a comprehensive dataset for training and testing. Our evaluation employed precise metrics, showing that the T5-3b model performed the most outstandingly in both accuracy and F1 scores, while T5-3b has the most solid "encoder-decoder" architecture and the largest parameters. Our metrics also demonstrate the enhanced performance of our FLAN-T5-base model, which showed significant improvements over the T5-base model in both accuracy and F1 scores by using the chain-of-thought approach. Finally, our detailed evaluation of the re-think strategy shows that it will notably improve the accuracy of the model, but with a minor decline in the F1 scores.

In summary, our study contributes to the field of MRC by demonstrating the effectiveness of incorporating feedback mechanisms and rethinking strategies in language models. Our approach not only enhances the accuracy of these models but also aligns them more closely with human cognitive processes. The insights gained from this research can be instrumental in advancing applications in areas such as automated customer service, education, and information retrieval, where accurate and adaptive response generation is crucial.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. 2023. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Self-evolve: A code evolution framework via large language models. *arXiv preprint arXiv:2306.02907*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Meta. 2023. Llama 2. *https://ai.meta.com/llama/*.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-tau Yih, Sida Wang, and Xi Victoria Lin. 2023. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*, pages 26106–26128. PMLR.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2022. Sescore2: Retrieval augmented pretraining for text generation evaluation. *ArXiv*, abs/2212.09305.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.