Trade-off: A Lightweight Method to Question Answering Using BERT

Lingjie Chen [1]

Abstract

This research investigates the efficacy of a lightweight, monolingual Bert model specifically tailored for Chinese Question Answering (QA) tasks. By employing a range of advanced training techniques, the model achieved a notable milestone, placing within the top 5 in Kaggle's renowned QA competition. Our findings reveal a surprisingly narrow performance gap between the specialized Bert model and larger Language Learning Models (LLMs), highlighting the potential of focused, language-specific approaches in the realm of natural language processing.

1. Introduction

The rapid evolution of Large Language Models (LLMs) has brought transformative changes to numerous Natural Language Processing (NLP) tasks, notably in translation and question answering. Despite their impressive capabilities, LLMs are not without limitations. These include the substantial costs associated with fine-tuning, diminished interpretability due to their inherent complexity, and restrictions on input and output for safety reasons. These challenges have renewed interest in more 'lightweight' pretrained models like Bert, which, despite their smaller size, offer significant advantages in specific contexts.

In this study, we concentrate on fine-tuning a Bert model to evaluate its performance relative to LLMs, with a particular focus on the task of Question Answering (QA). QA, a form of information retrieval, entails providing a context to a model and subsequently querying it for specific information embedded within that context, using natural language. Its conversational nature makes QA as a pivotal form of human-computer interaction, aligning well with the capabilities of NLP models.

Furthermore, unlike other models using higher amount of parameters to attain better performance, we have chosen to work with normal-size monolingual Bert model variants to emphasize the 'lightweight' nature of Bert. This choice not only simplifies the model architecture but also provides a unique perspective on the effectiveness of Non-LLM approaches in traditional NLP tasks.

The cornerstone of our project is the high accuracy achieved with lightweight model using multiple techniques. We aim to establish a new benchmark for Chinese Question Answering (QA) datasets using this model. Its streamlined design not only enhances efficiency but also ensures rapid response times, making it an ideal candidate for an information retrieval assistant. This approach demonstrates that achieving high performance in NLP tasks does not necessarily require complex or resource-intensive models, underscoring the potential of task-specific models in environments where quick, accurate responses are paramount, and resources are limited.

2. Related Work

The conceptual foundation of QA systems can be traced back to the pioneering works of Spärck Jones (1972) and Green (1969). These seminal papers laid the groundwork for QA systems. This era saw rule-based systems and keyword-based retrieval as cornerstones of information processing, setting the stage for more sophisticated approaches.

As QA systems matured, machine learning began to play a more pivotal role. Baroni and Lenci (2005) explored textual entailment using Latent Semantic Analysis, Takeuchi and Collier (2002) demonstrated the efficacy of decision trees in parsing and classifying elements in Japanese texts.

As a revolution, the introduction of transformers by Vaswani et al. (2017) marked a paradigm shift, leading to a focus on attention mechanisms, which allowed for more nuanced language understanding. This was further advanced by Devlin et al. (2018), whose work on BERT set new benchmarks in deep bidirectional language understanding, becoming a cornerstone for many subsequent NLP applications including QA.

Recent years have witnessed the ascendancy of large language models (LLMs) like GPT-3.5 (Brown et al., 2020), GPT-4(OpenAI et al., 2023), LLaMA(H.Touvron et al.,

[1] Department of Statistics, University of California, Berkeley, United States. Correspondence to: Nikita Zhivotovskiy <zhivotovskiy@berkeley.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

2023) which demonstrated remarkable few-shot learning capabilities and emergence of multiple incredible ability in context understanding and conversation making.

3. Setup

Task. We evaluate the downstream task of *extractive question answering*. Specifically, the model's given a paragraph and a question, the answer is a part of the original text in the given paragraph.. The task is to predict the correct answer text span in the paragraph. The task's assumption is that every question has answer in the paragraph, so we only need to consider the exact match of the predicted answer as metric. Figure 1 shows the procedure of the task.

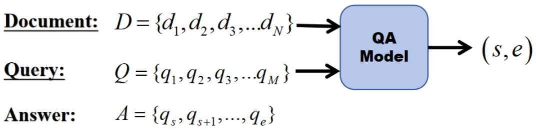
Document: $D = \{d_1, d_2, d_3, ...d_N\}$
Query: $Q = \{q_1, q_2, q_3, ...q_M\}$
Answer: $A = \{q_s, q_{s+1}, ..., q_e\}$
QA Model
(s, e)

Figure 1. Task Overview

Datasets. The dataset incorporating *DRCD (Delta Reading Comprehension Dataset)* and *ODSQA (Open-Domain Spoken Question Answering Dataset)*. It's a structured dataset specifically designed for Chinese Question Answering (QA) tasks. Each data entry in the dataset is composed of several key elements:

- **ID:** A unique identifier for each question-answer pair.
- **Paragraph ID:** The identifier linking the question to a specific paragraph where the answer can be found.
- **Question Text:** The question posed in Chinese. For example, *Where was the Central Research Institute established in 1928?*
- **Answer Text:** The direct answer to the question, for example, *Nanjing*.
- **Answer Start and End:** These are the character positions in the paragraph that mark the beginning and end of the answer. Which is the required prediction of the model

The dataset also includes the relevant paragraphs, providing the context needed for each question.

For both Bert and LLM, we train them on the training set and compare their performance on the test set. All the statistics of the dataset is shown in Table 1.

Table 1. Statistics about the Chinese QA training set.

Data set	Para. Avg.Token	Sent.Token	Size
Train	401.84	57.13	31960
Dev	427.83	59.38	4131
Test	430.29	42.96	4957

Model. In this task, finetune the dataset on both Bert models and LLMs.

- **BERT:** Bidirectional Encoder Representations from Transformers (BERT) (Devlin, J., et al., 2018) uses a bidirectional Transformer architecture. It was one of the first models to pre-train on a large corpus of text using unsupervised learning before being fine-tuned for specific tasks.
- **ALBERT:** A Lite BERT (Lan, Z., et al., 2019) introduces two parameter-reduction techniques to improve the scalability of BERT. It separates the size of the hidden layers from the size of vocabulary embedding, making it possible to increase the model's width without significantly increasing the number of parameters.
- **RoBERTa:** Robustly optimized BERT approach (Liu, Y., et al., 2019) modifies key hyperparameters in BERT, including removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.
- **Qwen-7B:** A 7 billion-parameter large language model from Alibaba Cloud, pre-trained on a diverse dataset of over 3 trillion tokens, including multilingual texts and code. Qwen-72B is optimized for extended vocabulary and longer context support, offering enhanced performance across various language tasks. The model is part of the Qwen series, designed to provide high-quality, scalable language representations.
- **Baichuan 2:** A new generation large-scale language model by Baichuan Intelligence Inc., trained on 2.6 trillion tokens and available in 7B and 13B parameter versions. Baichuan 2 showcases leading performance on Chinese and English benchmarks and offers Base and Chat models, with the latter available in a 4bits quantized version, highlighting its efficiency and versatility.

Training Device. Nvidia 3080Ti GPU

4. Experiment

We commenced our experiment by initializing our model with a pre-trained BERT model, pretrained on machine reading comprehension (MRC) data. The fine-tuning of the model encompassed a series of steps: preprocessing, training, and postprocessing.

During training, we addressed the input context quota of BERT by providing the model with a single window of context that included the target result. For testing, we employed a series of overlapping windows extracted from the context aviod the input context quota while ensure the model can make the best prediction.

The initial phase utilized the base-chinese-bert pretrained model—a standard BERT model—optimized with AdamW and a linear learning rate scheduler. We set the batch size to 16, a window size of 150, and a stride of 32 during the testing phase to enhance granularity. To prevent the model from learning the relationship between window position and content, we randomized the position of the training windows. The training duration was approximately 40 minutes, followed by a 15-minute testing phase.

Subsequently, we incorporated ALBERT and RoBERTa models into our experimental setup. We adjusted the batch size to 16, incorporating gradient accumulation to improve memory efficiency. Except for the batch size modification, we retained all other components and hyperparameters from the initial setup. Postprocessing steps were introduced to address specific issues with the model outputs:

- **Handling [UNK] Tokens:** The models occasionally produced [UNK] tokens when encountering out-of-vocabulary (OOV) characters in the context. To rectify this without affecting the model's comprehension ability, we mapped [UNK] tokens back to the corresponding characters in the original context.
- **Invalid Span Detection:** We encountered instances where the output span [start_index, end_index] was invalid, either because the start_index came after the end_index or the indices were outside the attention mask of the input tokens. A hard-coded function was implemented to detect and correct these anomalies by adjusting to a finer granularity.

The training process for these models extended to about 1.5 hours, with an additional 20 minutes for testing.

For the LLM portion of our experiments, we leveraged external models via API calls to Qwen and Baichuan. The prompt used was structured to ensure the models provided concise answers without additional explanation, formatted as a parsable list. For example:

"I want you to answer the questions based on the paragraph given to you. You only need to give the answer to the questions without any further explanation. The answer should be the excerpt of the original paragraph. Please output the answer in a parsable list format. An example output looks like:["Answer1","Answer2"]"

This prompt was designed to test the models' ability to extract precise information from the given text in a structured format conducive to further analysis.

5. Results Analysis

5.1. Main Result

As the result shown in Figure.2, RoBERTa model outperformed other BERT variants with highest EM accuracy and comparatively smaller loss.

The training loss for all models underwent a sharp decline in the very first stage of training, suggesting a rapid learning phrase. The big slope is due to large learning rate and the 'pretrain property' of BERT model. Then almost all model converge smoothly to a stable loss value in only four epoch training, reflecting BERT's strong learning ability on QA task. The final performance of our RoBERTa ranked top 5 in the Kaggle Leaderboard. On the other hand, the accuracy on the dev set increases steadily, showing the BERT's structure has sufficient regularization.

5.2. Ablation Result

Table 2. Ablation Test

Method Range	**Accuracy** (↑) (0.0 - 1.0)	**Loss** (↓) (0.0 - ∞)
Bert-base-chinese	0.479	12.3
+ Preprocessing	0.609	10.8
+ Postprocessing	0.661	8.2
+ RoBERTa	0.805	4.2
+ Ensemble	0.831	4.1

In order to study the impact of different training method on the final result, we carry out a ablation test showned in the Table 2.

Bert-base: The Bert-base-chinese model serves as a benchmark with an accuracy of 47.9% and on basis of this, we can analyze other technique's boost on the accuracy.

Preprocessing: The addition of preprocessing include reset the training window from being centered on the correct answer, set the granularity of the test window to reach performance-cost balance. This technique increase the accuracy by 13%, suggesting the model's comprehension and learning ability depend heavily on data's properties.

Postprocessing: Further enhancements through postprocessing techniques yield another leap in performance, with accuracy rising to 66.1% and loss decreasing to 8.2. Postprocessing enables us to analyze the model's comprehension more precisely since it solve the problems such as: the finite vocabulary and the lack of granularity. Enhancing the finetuning efficiency.

RoBERTa: The integration of the RoBERTa has yield an increase in accuracy of 14%, the highest among all. This result is intuitive in that the performance of RoBERTa is overall better than vanilla BERT model given it had been trained on more well-selected data.

Ensemble: We combine all the BERT models we trained to get the predictions of multiple models, slightly improves accuracy to 83.1% and minimizes loss to 4.1. While the gains are modest compared to the previous step, it indicates that ensemble strategies can still squeeze out performance improvements, possibly by reducing variance in the model predictions. Here we only use the vote-ensemble, giving all models the same weight.

5.3. Comparison with LLM

Table 3. Performance Comparison

Method Range	**Accuracy** (↑) (0.0 - 1.0)	**Adjusted Accuracy*** (↑) (0.0 - 1.0)
BERT		
Base BERT	0.479	0.479
ALBERT	0.759	0.759
RoBERTa	**0.805**	**0.805**
NABERT	0.779	0.779
LLM		
Qwen	0.519	0.590
Baichuan	0.021	0.04

In a direct comparison of performance metrics between BERT models and Large Language Models (LLMs), the data suggests that BERT variants generally outperform the LLMs in terms of accuracy. The base BERT model achieves an accuracy of 47.9%, which is a modest starting point. Upon enhancement with ALBERT and RoBERTa, the accuracy significantly increases to **75.9%** and **80.5%** respectively, with NABERT also performing well at **77.9%**.

In contrast, the LLMs display a mixed performance. Qwen has an accuracy of 51.9%. Consider some question's paragraph may contained sensible content that LLM can't process the input or output the answer, we devise a **Adjusted Accuracy**, which takes account only the questions that model give response to. This metric can more equitably showcase the performance of LLM. The adjusted accuracy of Qwen is 59%, reaching the level of base-BERT with proper training techniques. Baichuan, on the other hand, shows a notably lower accuracy of 2.1%, with a slight increase to 4% in adjusted accuracy. These values suggest that Baichuan, in particular, may not be well-suited or optimized for the tasks at hand compared to its counterparts.

This comparison underscores the effectiveness of BERT and its variants in achieving higher accuracy in the evaluated tasks, likely due to their design and training on task-specific datasets. The LLMs, while beneficial in certain contexts, may require further tuning or specialized training to reach the performance levels of their BERT counterparts.

6. Discussion

The potential problems of this task can be separated into Training and Evaluation two parts:

- **Training:** Constrained by the computational resource, We can't finetune LLM specifically on this problem. This limitation potentially obscures the true capabilities of the LLM in comprehending and responding to questions accurately. Additionally, the dataset we used in this task is Unsimplified Chinese, which is slightly different from Standard Chinese. (the discrepancy here lies mainly in lexicon) And LLM's comprehension about the paragraph and its corresponding output may both be interfered by that trait. Furthermore, the inherent complexity of the LLM architecture restricted our ability to employ certain training techniques that may enhance the model's performance. Since we only call the API, so that we are unable to hand-code the post-process functions to deal with the common errors made by LLM. This could in part explain the accuracy margin between BERT and LLMs. Also, since we are training with Chinese QA task, we choose the Qwen and Baichuan model instead of GPT-3.5, expecting for better performance. But the result shows substantial margin.

- **Evaluation:** The primary metric in this task is Exact Match (EM), which aligns well with the operational mechanism of BERT models that extract answers directly from the provided paragraphs. However, this metric may not be entirely fair when applied to LLMs, which are designed to generate responses based on prompts. These responses might include additional words or phrases that, while not altering the intended meaning, do not conform strictly to the EM criterion. Even with explicit instructions in the prompt to extract answers verbatim, many responses deviated from this pattern. This indicates the LLM's adherence to the prompt isn't strong enough, but doesn't necessarily mean it lack the context comprehension and information-retrieval abilities. The common error types have been shown in the following section.

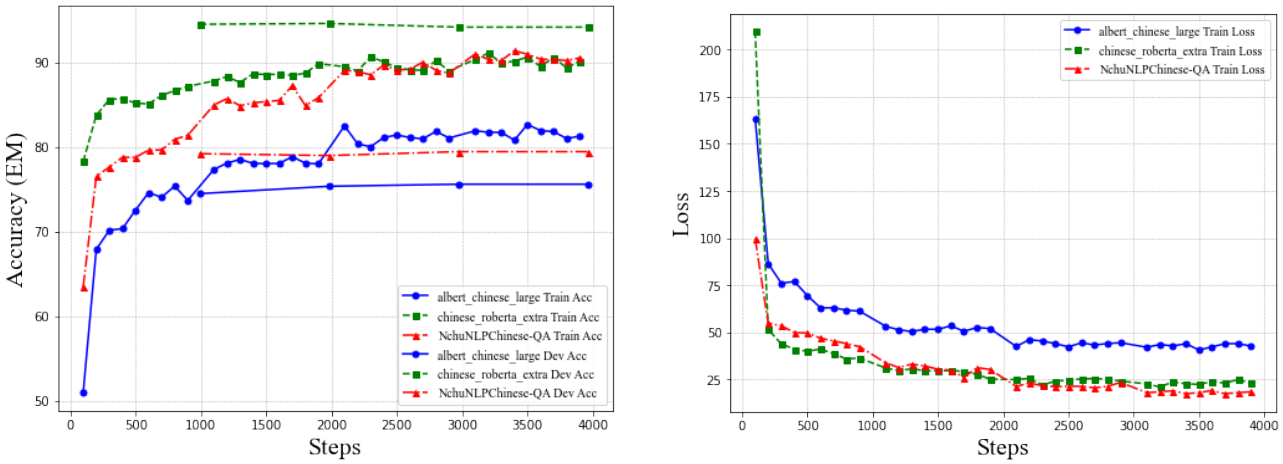

Figure 2. Comparison of results.

Addition

Question: How many kilometers is the main campus of UNSW from the center of Sydney?
Gold Answers: 5
Prediction: 5 kilo.

Language Error

Question: Who wrote the ”Pastoral” that sings about pastoral life?
Gold Answers: Virgil
Prediction: Virgili (Virgil in French)

**Common Error Type of LLM (Answers and predictions have been translated into English*

7. Conclusion and Future Work

Throughout this study, we evaluated various BERT variants against a prominent Chinese-based Large Language Model (LLM) on a Chinese Question Answering (QA) task. The results indicate a marked superiority of BERT models over the LLM, thereby highlighting the effectiveness of more compact models in specialized tasks. The implementation of multiple training strategies further enhanced the BERT models' performance, with an ablation study quantifying the contribution of each method.

Despite these findings, our study encounters several limitations, including the dataset's scope, the QA task's limitations, the LLM's choice, and the evaluation metrics. These factors may have influenced the outcome.

Future research should aim to extend the dataset size, diversify the QA tasks, and adjust the language settings to better gauge the capabilities of LLMs. A more tailored assessment metric that accurately reflects the strengths of LLMs is also necessary to provide a fair comparative analysis.

Acknowledgements

I wish to express my deepest gratitude to my instructor, **Nikita Zhivotovskiy**, GSI **Mriganka Basu Roy Chowdhury**, whose guidance and instruction were invaluable throughout the course.

This project was under the technical assistance of **GPT-4**, providing solutions to coding challenges and aided in the formatting of this essay.

Finally, I would like to extend my thanks to the Fudan NLP group for providing the computational resources that were crucial for the success of this project.

References

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yang, A., Xiao, B., Wang, B., Zhang, B., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., Yang, F., et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.