Enhancing Parameter Efficiency in Large Language Model Fine-tuning Through Hybrid Prompt Tuning and LoRA

Gonge Lehan (G2205030)[1], Ravi Shwetha (G2304191F)[2], Kozhimadam Shahin Shah (G2303660L)[3], Jin Mingwei (G2302766C)[4], Ang Wee Onn (G2304229J)[5]

[1] GONG0072@e.ntu.edu.sg, [2] SHWETHA002@e.ntu.edu.sg, [3] SHAHINSH001@e.ntu.edu.sg, [4] JINM0007@e.ntu.edu.sg, [5] ANGW0111@e.ntu.edu.sg

Abstract

The rapid evolution of large language models (LLMs) has led to significant advancements in natural language processing (NLP). However, fine-tuning these models often demand extensive computational resources, making it impractical for multiple tasks. Our study introduces a new hybrid approach. We combine low-rank adaptations (LoRA) with prompt tuning to fine-tune the FLAN-T5 model. This method aims to improve efficiency and performance across various NLP tasks. We find that it not only preserves the original model's capabilities but also supports sustainable and scalable AI practices in NLP. Through experiments in SQL query generation, dialogue summarization, and sentiment analysis, our results highlight the benefits of parameter-efficient fine-tuning. This approach effectively balances model adaptability with resource efficiency.

Introduction

Recent advancements in large language models have revolutionized the field of NLP, providing the ability to perform a diverse range of tasks with high efficiency. Despite their versatility, these models, when fine-tuned for specific tasks, often incur high computational costs and increased model size, posing a challenge for scalable deployment. In this work, we explore a hybrid fine-tuning approach that integrates the efficiency of LoRA and the adaptability of prompt tuning, applied to the FLAN-T5 model. We delve into the efficacy of this method in enhancing performance across multiple NLP benchmarks, offering insights into its potential for scalable and resource-conscious applications in language understanding.

Related Work

Large Language Model: Large language models (LLMs) like GPT-3, BERT, and T5 represent a paradigm shift in natural language processing, employing vast amounts of data during pretraining to develop a broad understanding of language (Radford et al. 2019; Devlin et al. 2018; Raffel et al. 2020). These models are typically fine-tuned for specific tasks, which involves additional training on a smaller, task-specific dataset to tailor the model's responses to the task at hand. Multitask LLMs and those subjected to instruction tuning, where a single model is trained across multiple tasks or instructed in natural language to perform tasks without task-specific fine-tuning, further exemplify the flexibility and capability of LLMs to generalize across domains (Wei et al. 2021).

Fine-tuning: The full fine-tuning of these models for different tasks can lead to significant increases in computational costs and model size, making it challenging to deploy them in resource-constrained environments. Additionally, creating separate instances of a fine-tuned model for each task is inefficient and limits the scalability of LLM applications (Aghajanyan, Zettlemoyer, and Gupta 2020). Finding the right prompts that effectively guide the model to perform a new task is also a non-trivial endeavor, often requiring substantial trial and error to identify prompts that yield high-quality outputs (Brown et al. 2020).

Parameter-efficient Fine-tuning: In response to the challenges above, research has turned toward parameter-efficient fine-tuning techniques that modify a smaller subset of model parameters while retaining or even enhancing performance on downstream tasks (Houlsby et al. 2019). Techniques like LoRA and prompt tuning have shown promise in this regard. LoRA introduces low-rank matrices to recalibrate the model's attention mechanisms without altering the pre-trained weights directly (Hu et al. 2021). Prompt tuning, on the other hand, involves appending trainable soft prompts to the model's inputs to steer its behavior, leveraging the model's existing knowledge while requiring minimal additional parameters (Lester, Al-Rfou, and Constant 2021). These methods offer more sustainable and scalable approaches to leveraging LLMs across various tasks.

Approach

This section outlines the technical framework and methodologies employed in our project to investigate parameter-efficient fine-tuning methods for Large Language Models (LLMs), focusing on a hybrid approach that combines prompt tuning and Low-rank Adaptation (LoRA) with the FLAN-T5 model.

Model Architecture

At the heart of our approach lies the T5 (Text-to-Text Transfer Transformer) model (Raffel et al. 2020), renowned for its versatility and effectiveness across a broad range of NLP tasks. T5 is designed as a multitasking model, processing input text prefixed with a task-specific tag that guides the model's output generation. This unique architecture makes T5 exceptionally adaptable through prompt tuning, enabling it to excel in task-specific applications by adjusting its input prompt.

However, the original T5 model, pre-trained exclusively on span corruption, encounters limitations due to its pre-training design. It has never processed truly natural input text free of sentinel tokens, nor has it been tasked with generating truly natural targets. Consequently, every target in T5's pre-training environment begins with a sentinel token, an artifact that could potentially hinder its performance on tasks requiring natural language understanding and generation.

To address these challenges and leverage a more natural processing capability, we opted for FLAN-T5 (Chung et al. 2022), an evolution of the T5 model. FLAN-T5 is fine-tuned on a diverse set of tasks presented in natural language, enabling it to interpret and generate more naturally structured text. This choice was also motivated by our interest in comparing fine-tuning effects on a model pre-trained with task-specific capabilities against one without such prior tuning.

Prompt Tuning

Prompt tuning (Lester, Al-Rfou, and Constant 2021) represents a parameter-efficient fine-tuning technique where a small set of trainable parameters (prompts) is optimized to guide the model's attention and processing towards the specific requirements of a task. By crafting and tuning these prompts, we can effectively ”instruct” the FLAN-T5 model to focus on the nuances of a given NLP task without altering the bulk of its pre-trained parameters.

In our project, prompt tuning is utilized to refine the task-specific tags that precede the input text, optimizing these tags to enhance the model's task alignment and performance. This method allows us to leverage the multitasking nature of FLAN-T5, tailoring its formidable general capabilities to our selected NLP tasks through minimal but targeted modifications.

Low-rank Adaptation (LoRA)

LoRA (Hu et al. 2021) introduces a novel approach to model adaptation by inserting trainable low-rank matrices into the transformer layers of the model. This process adjusts the model's weight matrices without the need to retrain the entire network, thus preserving computational efficiency while allowing significant flexibility and customization in model behavior.

In our methodology, LoRA is applied to carefully selected layers of the FLAN-T5 model, aiming to fine-tune its responses and internal representations for improved task-specific performance. The integration of LoRA complements prompt tuning by modifying the model at a structural level, further enhancing its adaptability and efficiency.

Hybrid Fine-Tuning Approach

Our hybrid fine-tuning strategy combines the strengths of prompt tuning and LoRA, optimizing both the input prompt and the model's internal architecture simultaneously. This dual approach aims to maximize the efficiency and effectiveness of the fine-tuning process, tailoring the FLAN-T5 model's capabilities to our targeted NLP tasks with minimal parameter adjustments.

By optimizing the task-specific tag alongside structural fine-tuning through LoRA, we achieve a synergistic effect that enhances the model's performance beyond what could be accomplished by either method alone. This hybrid approach allows us to explore the boundaries of parameter-efficient model adaptation, contributing valuable insights into scalable and sustainable AI practices in the NLP domain.

Experiments

This section delineates the comprehensive set of experiments designed to evaluate the efficacy of our hybrid fine-tuning approach using the FLAN-T5 model across multiple NLP tasks. It details the datasets employed, the evaluation metrics, the model configurations, and the empirical results obtained, providing insights into the performance and implications of our methodology.

Data

The selection of datasets for this study was guided by the objective to encompass a diverse array of NLP tasks, enabling us to assess the hybrid fine-tuning method's versatility and effectiveness comprehensively. The datasets chosen are as follows:

1. **WikiSQL** (Zhong, Xiong, and Socher 2017): A benchmark for the task of SQL query generation from natural language questions, consisting of natural language questions, their corresponding SQL queries, and the tables on which these queries are executed. This dataset is pivotal in our study as it represents a task outside the initial training domain of FLAN-T5, challenging the model to generate accurate SQL queries from given questions and table schemas. The inclusion of WikiSQL allows us to assess the model's capacity to acquire new capabilities, specifically in areas it was not explicitly trained for, such as direct SQL query generation.
2. **SAMSum** (Gliwa et al. 2019): Focused on dialogue summarization, SAMSum includes conversational transcripts alongside human-written summaries, tasking the model with producing succinct and coherent summaries of extended dialogues. Given FLAN-T5's training, which encompasses a broad spectrum of NLP tasks including summarization, this dataset offers an opportunity to explore and potentially enhance the model's inherent summarization capabilities through our hybrid fine-tuning approach.
3. **SST-2** (Socher et al. 2013): The Stanford Sentiment Treebank, aimed at sentiment classification, provides sentences from movie reviews with binary sentiment labels. As sentiment analysis is among the tasks FLAN-T5

was fine-tuned for, SST-2 serves to evaluate how the hybrid fine-tuning method refines and potentially augments the model's pre-existing sentiment analysis proficiency.

For each dataset, the input and output forms are strictly adhered to as per the task definitions, ensuring clarity and consistency in our experimental setup.

Evaluation Method

Our evaluation framework employs task-specific metrics designed to accurately measure model performance:

- For WikiSQL, we use **Logical Form Accuracy** to assess the exact match between the generated and the target SQL queries.
- For SAMSum, we employ the **ROUGE-1/2/L** metrics, evaluating the overlap of unigrams, bigrams, and the longest common sequence between the generated summaries and reference summaries.
- For SST-2, **Accuracy** serves as the metric, quantifying the percentage of correctly classified instances.

These metrics provide a quantitative basis for comparing our approach against established baselines and understanding its effectiveness in diverse NLP contexts.

Model Settings

Experiments were conducted with the following configurations:

- **Base Model**: The foundational model for our experiments is the FLAN-T5 Base model (Google 2024), selected for its optimal balance of performance and computational efficiency.
- **LoRA Settings**: We incorporate Low-rank Adaptation (LoRA) with settings of Rank 4 and Alpha 4, targeting the ”q” and ”v” parameters of the transformer layers. A dropout rate of 0.05 is applied to mitigate overfitting.
- **Prompt Tuning Settings**: Prompt tuning is employed by introducing 30 soft tokens, initialized from the model's vocabulary, to guide the model's focus towards task-specific details.
- **Training Configurations**:
  - *Mixed Precision*: Training efficiency is optimized through the use of mixed precision (bf16).
  - *Batch Size*: A consistent batch size of 16 is used for both the training and testing phases.
  - *Optimizer*: An Adam optimizer with a learning rate of 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 1×10^{-3}, and an epsilon of 1×10^{-8}.
  - *Learning Rate Scheduler*: A cosine learning rate scheduler with 500 warm-up steps is implemented to adjust the learning rate dynamically.
  - *Max Gradient Norm*: Set at 1.0 to prevent the exploding gradient problem, ensuring stable training.
- **Training Duration**: Training is conducted for up to 20 epochs, employing early stopping based on validation performance to avoid overfitting.

This configuration is uniformly applied across all datasets involved in our study—WikiSQL, SAMSum, and SST-2—to facilitate a direct comparison of the hybrid fine-tuning approach's impact on each task.

Results

The outcomes of our hybrid fine-tuning and fully fine-tuned approaches across various NLP tasks are detailed in Table 1. This presentation allows for a direct comparison of our approach's performance enhancements.

Discussion

The comparative analysis detailed in Table 1 showcases the nuanced efficacy of our hybrid fine-tuning approach vs a fully fine-tuned model across a spectrum of NLP tasks. Notably, the hybrid method achieves comparable, if not superior, performance while engaging significantly fewer trainable parameters. This efficiency gain is particularly marked in scenarios where FLAN-T5 inherently excels, such as dialogue summarization and sentiment analysis. The hybrid model's success in these domains underscores the advantage of leveraging pre-existing model strengths while avoiding the potential overfitting that can accompany full model fine-tuning. Specifically, when the base model possesses prior capabilities related to a task, the disparities in the weight matrices are likely of low rank, making LoRA an effective tool for capturing these nuances. Conversely, the fully fine-tuned model, despite being robust, may risk overfitting, thus explaining its occasionally lower performance in familiar tasks.

In contrast, the scenario shifts for tasks like SQL query generation from natural language, which lies outside FLAN-T5's pre-training paradigm. Here, the hybrid model, despite its commendable performance, does not quite match the efficacy of the fully fine-tuned counterpart. This outcome hints at the limitations of parameter-efficient fine-tuning when tasked with instilling entirely new capabilities into LLMs. The fully fine-tuned model, benefiting from its complex architecture and the capacity to learn intricate relationships, outperforms in areas devoid of the base model's prior training. Interestingly, for the fully fine-tuned model, we employed hand-crafted tags specific to each task, a strategy not mirrored in the hybrid fine-tuning setup. Nevertheless, the effectiveness of our approach, particularly the soft tokens in the hybrid model, in implicitly learning these task-specific cues, is evident from the results. This aspect not only highlights the adaptability and learning efficiency of the hybrid fine-tuning method but also suggests that such soft tokens can discern and adapt to task-specific nuances without explicit directive tags.

The insights derived from these observations advocate for the potential of hybrid fine-tuning strategies to optimize LLMs across a diverse array of NLP tasks. They offer a promising avenue towards more resource-efficient and scalable model adaptations, especially in contexts where the base model already aligns with the task at hand. Future endeavors could aim to refine these hybrid fine-tuning techniques further or integrate additional strategies to mitigate

Table 1: Performance Comparison Across Datasets Using Hybrid Fine-Tuning vs. Full Model Fine-Tuning, with the best metrics highlighted in bold.

Model Configuration	**# Trainable Parameters**	**SAMSum**	**WikiSQL**	**SST-2**
		Rouge - 1 / 2 / L	**Logical Form Acc. (%)**	**Accuracy (%)**
Prompt Tuning + LoRA	465408	**50.4 / 25.4 / 41.8**	44.4	**95.0**
Fully Fine-Tuned Model	247577856	48.7 / 24.1 / 40.7	**49.3**	90.5

the performance gap for tasks beyond the base model's original training scope. Such explorations would contribute to the broader narrative of making fine-tuning processes more adaptable, efficient, and universally applicable.

Analysis

The qualitative evaluation of our hybrid fine-tuning approach encompasses both an ablation study to understand the contribution of individual components and an interpretability analysis to inspect the model's behavior on specific tasks.

Ablation Study

In our ablation study, we dissect the hybrid fine-tuning model to understand the individual and collective contributions of prompt tuning and LoRA to overall performance. This decomposition involves evaluating the model in several configurations: using the base model without any fine-tuning, applying only prompt tuning, applying only LoRA, and combining both techniques. The objective is to ascertain the incremental benefits brought about by each method and their synergy when used together. The results of these configurations across our chosen NLP tasks are presented in Table 2.

From the results in Table 2, we draw several inferences. The base model sets the stage with its inherent proficiency in tasks like dialogue summarization and sentiment analysis. When prompt tuning is introduced alone, there's a discernible improvement in performance, particularly for the SST-2 dataset, which demonstrates prompt tuning's capacity to direct the model's focus to salient aspects of the task at hand. However, for the WikiSQL task, where the base model has no pre-training, prompt tuning alone is not sufficient to achieve high accuracy, reflecting its limitation in instilling completely new skills into the model.

In contrast, integrating LoRA brings about marked performance enhancements across all tasks. The significant leap in accuracy for the WikiSQL task with LoRA indicates its potential in imparting new capabilities to the model, especially in areas requiring intricate structural transformations. Combining prompt tuning with LoRA yields the most considerable performance enhancements, suggesting a complementary relationship where prompt tuning refines the model's focus, and LoRA strengthens its structural processing capabilities. This symbiosis is particularly effective for tasks that align with the base model's pre-existing strengths, as seen in the enhanced results for both SAMSum and SST-2. These improvements highlight the potential of our hybrid approach in leveraging the model's innate abilities and extending them through targeted fine-tuning.

Model Interpretation

For a nuanced understanding of our model's decision-making process, we employ the Integrated Gradients method (Sundararajan, Taly, and Yan 2017), visualizing how the model attends to different input features for various NLP tasks. The heatmaps, as shown in Figures 1, 2, and 3, illustrate the influential tokens identified by the model when generating outputs for the WikiSQL, SAMSum, and SST-2 datasets, respectively.

In generating SQL queries, as observed in Figure 1, the model exemplifies a keen understanding of semantic relationships. It assigns attention where it matters most: correlating the token 'death' with the word 'die' and 'Date' with 'day', hence demonstrating a sophisticated grasp of contextual cues. The heatmaps from Figure 1 (a) and (b) lay bare the model's proficiency in not only recognizing explicit vocabulary but also in discerning the implicit connections between tokens and their larger narrative roles within the input.

The summarization task, as evidenced by Figure 2, unfolds the model's narrative intelligence. Here, the focus on specific entities such as 'Mitch', 'Supreme', and 'shirt' in Figure 2 (a), alongside the characters 'Mason' and 'Sophia' in Figure 2 (b), signifies the model's aptitude for distilling the core elements of dialogues into concise summaries. Such interpretive heatmaps illustrate the model's ability to prioritize information that is crucial to the essence of the narrative, a skill that lies at the heart of effective summarization.

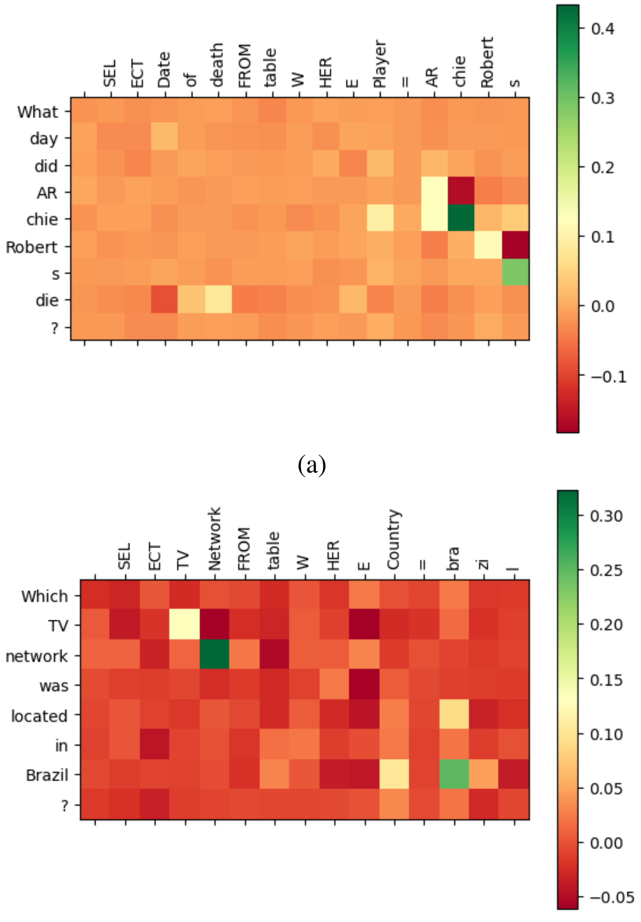
Figure 3's heatmaps illuminate the model's sentiment analysis capabilities within the SST-2 dataset. In Figure 3 (a), positive sentiments are associated with tokens that reflect optimism, like 'life' and 'rich', while Figure 3 (b) highlights the model's alignment with more somber tones, as it draws upon words like 'so' and segments of 'satire'. This contrast in sentiment attribution demonstrates the model's ability to accurately identify positive and negative language cues.

Together, these interpretations provide not just a window into the model's decision-making process, but also confirm the efficacy of the hybrid LoRA and prompt tuning ap-

Table 2: Ablation Study Results: Performance Impact of Prompt Tuning and LoRA on FLAN-T5 Across NLP Tasks, with the best metrics highlighted in bold.

Model Configuration	**SAMSum**	**WikiSQL**	**SST-2**
	Rouge - 1 / 2 / L	**Logical Form Acc. (%)**	**Accuracy (%)**
Base Model	48.5 / 23.6 / 40.3	0.0	90.8
Prompt Tuning Only	48.5 / 23.8 / 40.2	3.6	92.3
LoRA Only	50.2 / 25.1 / 41.5	43.7	94.2
Prompt Tuning + LoRA	**50.4 / 25.4 / 41.8**	**44.4**	**95.0**

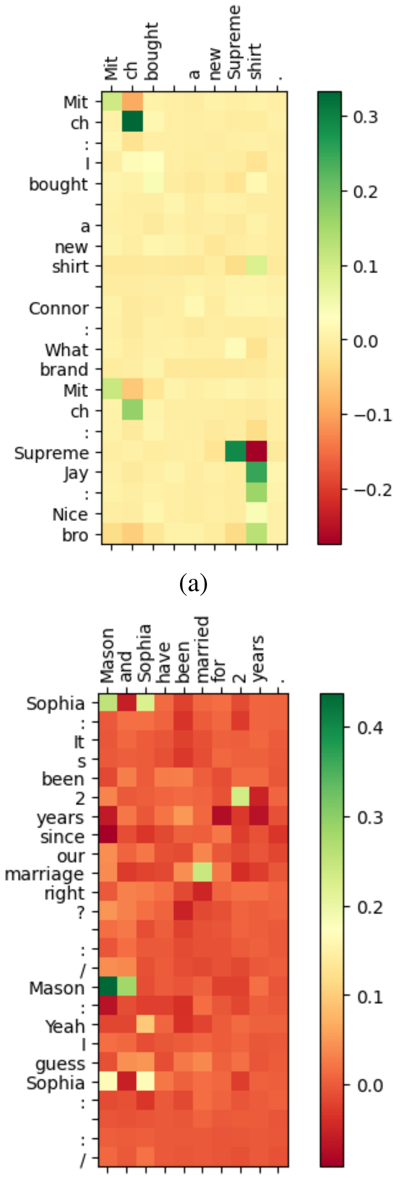
proach. By analyzing these heatmaps, we gain valuable insights into how this hybrid model navigates the complex terrain of language, extracting and leveraging the most pertinent features for task-specific predictions. This reinforces the potential of our model to perform robustly across a variety of NLP tasks, laying a foundation for future models that are both interpretable and efficient.



(a)

(b)

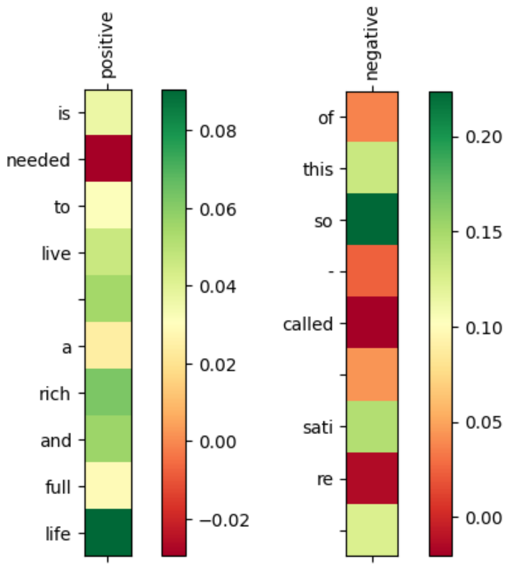
Figure 1: Attribution heatmaps for WikiSQL dataset inputs, visualizing model's focus areas: input text on the left and generated SQL query tokens above.



(a)

(b)

Figure 2: Attribution heatmaps for SAMSum dataset inputs, highlighting areas the model prioritizes for summary generation: conversation text on the left and the corresponding summary tokens on top.



(a) (b)

Figure 3: Attribution heatmaps for SST-2 dataset inputs, revealing words the model emphasizes for positive and negative sentiment classification: input text on the left and predicted sentiment on top.

Conclusion

Our exploration into the hybrid fine-tuning of FLAN-T5 using LoRA and prompt tuning has unveiled a promising avenue for enhancing the performance of LLMs while addressing the constraints of computational efficiency and scalability. The positive outcomes observed across diverse NLP tasks reaffirm the viability of this approach, which harnesses the strengths of both techniques without compromising the model's intrinsic capabilities. Future research may build upon these findings to further refine fine-tuning practices, making them more adaptable and efficient for the ever-growing demands of AI-driven language processing.

References

Aghajanyan, A.; Zettlemoyer, L.; and Gupta, S. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Gliwa, B.; Mochol, I.; Biesek, M.; and Wawer, A. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Google. 2024. FLAN-T5 Base Model. https://huggingface.co/google/flan-t5-base. Accessed: 2024-03-02.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhong, V.; Xiong, C.; and Socher, R. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.