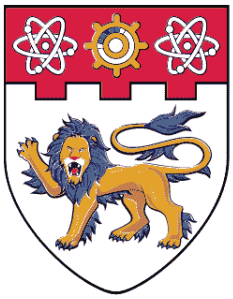

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

SC4001: Neural Networks and Deep Learning

Chain-of-Thought (CoT) Distillation: Outperforming Large Language Models with Task-Specific Small Language Models and Less Training Data

AY23/24 Semester 2

Name	Matriculation Number
Angie Wong Mei Chi	U2121896E
Dominick Ng Jie En	U2120310K
Keith Heng Jinsheng	U2121807C

Chain-of-Thought (CoT) Distillation: Outperforming Large Language Models with Task-Specific Small Language Models and Less Training Data

Angie Wong Mei Chi, Dominick Ng Jie En, Keith Heng Jinsheng

Nanyang Technological University

{angi0006, ng0003ck, keit0010}@e.ntu.edu.sg

Code: https://github.com/dominickng28/SC4001-Group-Project Runs: http://surl.li/sooms

Abstract

The advent of large language models (LLMs) such as GPT-3 has revolutionized natural language processing (NLP). Despite their significant prowess, widespread LLM deployment has proven challenging due to their immense resource requirements. This is especially so where many real-world applications do not necessitate the full range of capabilities of LLMs. To counter this, task-specific small language models (SLMs) can be trained to mitigate the resource complexity of LLMs while achieving similar or higher performance for a given task. This is usually done by (1) finetuning with human-labelled data, or (2) distillation with LLM-generated labels. However, these methods usually require large amounts of data that are hard to obtain, or suffer from limited performance. In this paper, we introduce Chain-of-Thought (CoT) Distillation, a novel method which extracts LLM-generated rationales for added supervision when training smaller models within a multi-task framework. Using sentiment analysis tasks as a benchmark, our experiments deliver three main findings. (1) CoT Distillation allows us to train substantially smaller models that outperform LLMs for sentiment analysis. (2) CoT Distillation has higher performance, and (3) requires significantly less labelled/unlabelled training data, then existing approaches to training smaller models – finetuning and distillation.

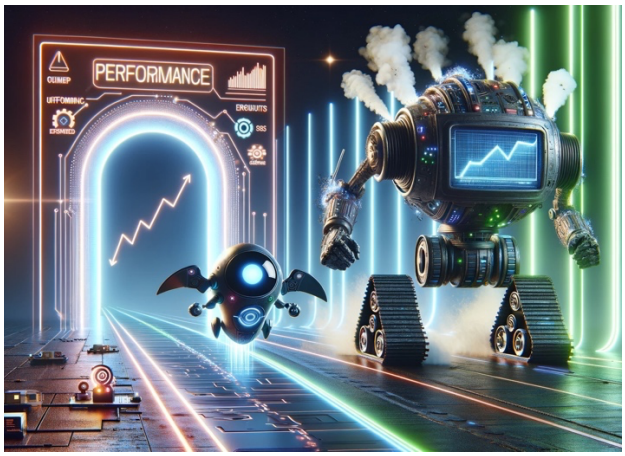

Figure 1: A lightweight robot – representing CoT distillation generated models – easily outpaces its larger counterpart, a LLM model, which struggles with its immense size and resource requirements. (Image generated by DALL·E 3)

Content

1 Introduction

The emergence of Large Language Models (LLMs) such as GPT-4 and Google's Pathway Language Model (PaLM) indicates a paradigm shift in the field of Natural Language Processing (NLP). Despite their unparalleled abilities in complex applications ranging from text summarisation to sentiment analysis (Smith, 2022), many have questioned the practicality of LLMs. A main area of contention is the poor efficiency of LLMs for simpler tasks due to their enormous size (Chowdhery, 2022). For instance, operating a single 175 billion parameter LLM requires minimally 350GB GPU memory using dedicated infrastructure (Zheng, 2022). Such requirements are far beyond feasible for many real-world applications.

To overcome the complexity of LLMs, most practitioners choose to develop smaller specialized models (Thomas, 2024). This is most commonly done through *finetuning* – which updates a pretrained smaller model (Raffel, 2020) – or *distillation* – training smaller models with labels generated by a larger LLM (Arora, 2022). However, these approaches are not perfect – needing vast amounts of training data or having limited performance (Hsieh, 2023).

Thus, the key purpose of our research, which uses sentiment analysis as a benchmark, is to develop a method of training smaller models – which both outperform LLM benchmarks and are more data-efficient compared to existing *finetuning* and *distillation* methods. To achieve this, we explore relevant developments and propose a novel method – *Chain of Thought (CoT) Distillation*.

2 Related Works

In the development of sentiment analysis, researcher have experimented with various techniques ranging from Lexicon-Based methods to traditional Machine Learning techniques and more recently, towards the utilisation of Deep Learning (Gunasekaran, 2023). We will briefly go through some of the techniques used for each approach and their limitations.

2.1 Lexicon-Based

Early sentiment analysis employed Lexicon-Based Methods, which utilises a pre-established dictionary of words, each tagged with a positive or negative sentiment score (Gunasekaran, 2023). The sentiment of a text is then determined by summing up these scores. A modern incarnation of this approach is the Valence Aware Dictionary and Sentiment Reasoner (VADER), which not only assesses the sentiment polarity of text but also measures the intensity of sentiment (Al-Shabi, 2020).

However, the Lexicon-Based Method is limited by its reliance on an extensive dictionary of words and their associated sentiment scores (Sadia, 2020). Additionally, it displays poor performance when faced with linguistic subtleties and varying contexts (Bonta, 2019).

2.2 Traditional Machine Learning

Subsequently, machine learning algorithms such as Support Vector Machines (SVM) and Random Forest became prevalent for sentiment analysis due to their proficiency in assimilating from voluminous datasets and discerning intricate patterns (Singh & Tripathi, 2021). Additionally, researchers have explored the use of hybrid approaches, by combining the use of VADER with SVM classifiers for enhanced performance (Borg & Boldt, 2020).

Nevertheless, these models are not without their drawbacks, particularly concerning domain specificity. While these models offer better generalization across different domains, they may fail to accurately predict the sentiment for texts that significantly deviate contextually from the data used during their training.

2.3 Deep Learning

2.3.1 Rise of Deep Learning

The rise of deep learning has marked a transformative phase in sentiment analysis. Some important models that have substantially advanced sentiment analysis include:

(1) **Recurrent Neural Networks (RNN)** – which have substantial ability in processing sequence of data, hence grasping the sequential and contextual nuances of language (Mikolov et al., 2010).

(2) **Long Short-Term Memory (LSTM)** networks, an RNN variant, have further refined the processing of long-term dependencies in text, enhancing the precision of sentiment analysis (Thomas & Latha, 2018).

(3) **Convolutional Neural Networks (CNNs),** famed for image processing, have been repurposed for sentiment analysis to detect patterns within text, exploiting their proficiency in identifying local and position-invariant features (Kim, 2014).

2.3.2 Transformers

The latest advancements have been in the realm of Transformer-based architectures, such as **BERT**, which leverage attention mechanisms to dynamically weigh the significance of each word in a sentence, boosting contextual awareness and sentiment analysis accuracy (Vig & Belinkov, 2019). Due to the usage of the self-attention mechanism, transformers such as **Text-To-Text Transfer Transformer (T5)** are able to better capture global dependencies as compared to RNN (Xue et al., 2021), resulting in state-of-the-art performances.

2.3.3 Rise of Large Language Models

Augmenting the advancements of transformer models, LLMs such as GPT-3 and PaLM has made significant strides with even more parameters and refined algorithms, allowing it to process and generate text with unprecedented subtlety and complexity. GPT-3 has 175 Billion parameters while PaLM has 540 Billion parameters, which are trained on an extensive dataset (Brown et al.,2020), enabling them to exhibit remarkable proficiency in a variety of natural language processing tasks, including but not limited to language translation and conversational agents. Despite their sophistication, the significant computational resources required to train and run such large models limits their practicality for simpler tasks (Daigle, 2023).

2.3.4 Rise of Small Language Models

The substantial carbon footprint generated by LLMs has catalysed the exploration and development of SLMs as viable alternatives. Consequently, the research community has introduced scaled versions of these models, including “mini”, “small” or even “tiny” versions of existing models. These scaled-down models retain the core functionality and capabilities of their larger counterparts but operate with significantly fewer parameters and computational resources, with some accuracy trade off.

To mitigate this trade-off, innovative training methods are being explored to enhance the performance of SLMs to rival that of LLMs. This study aims to evaluate various training techniques, including finetuning, distillation, and a novel approach known as ‘Chain of Thought (CoT) Distillation’ that is capable of training small models to match the performances of LLMs.

2.4 Large Model Knowledge Distillation

Knowledge distillation is a method that transfers insights from larger models to more compact models suitable for real-world applications (Fu et al., 2023). This technique is useful when labelled data is scarce, as it utilizes the larger model to generate a training dataset from unlabelled data, though these labels are often imperfect (Agrawal et al., 2022). A significant challenge in knowledge distillation is its reliance on extensive

unlabelled data to produce a useful training dataset. There is a need to lessen the reliance on large volumes of unlabelled data by focusing on distilling the explanatory factors—the "rationales"—behind the large model's decisions, in addition to the output labels themselves.

2.5 Learning from Human Rationale

The practice of integrating human-generated rationales into learning processes (Hase and Bansal, 2021) allows for human rationales to serve as inputs to inform model predictions (Rajani et al., 2019), improving overall model accuracy (Pruthi et al., 2022). Despite their utility, the acquisition of human rationales for model training is costly.

2.6 Learning from LLM-generated Rationale

LLMs possess the remarkable ability to not only predict outcomes but to also articulate the rationale behind their predictions, creating detailed reasoning pathways (Kojima et al., 2022). These pathways enhance the LLMs' capability in scenarios where only a few examples are provided (few-shot) or even none (zero-shot) (Wei et al., 2022), and can serve as additional data for finetuning, allowing LLMs to refine their own processes (Huang et al., 2022). However, the sheer scale of LLMs can restrict their practical deployment.

Hence, our paper explores the method of using rationales generated by LLMs as rich instructional data to train SMLs, which are more feasible for deployment. This concept of employing generated rationales for training is known as Chain of Thought (CoT) Distillation.

3 Methodology: Chain of Thought (CoT) Distillation

CoT Distillation utilizes the capability of LLMs to explain the reasoning behind their predictions, to train smaller models using less data. An overview of the process is shown in Figure 2. This process has two straightforward steps. Firstly, using an unlabelled dataset and LLM, the LLM is prompted to produce output labels and rationales explaining its predictions. Secondly, we use both these predicted labels and the rationales behind them to train a smaller model. These rationales enrich the training by imparting insights into why certain inputs might be linked to certain output labels, encapsulating critical task-related knowledge that might not be easily discernible from the inputs alone.

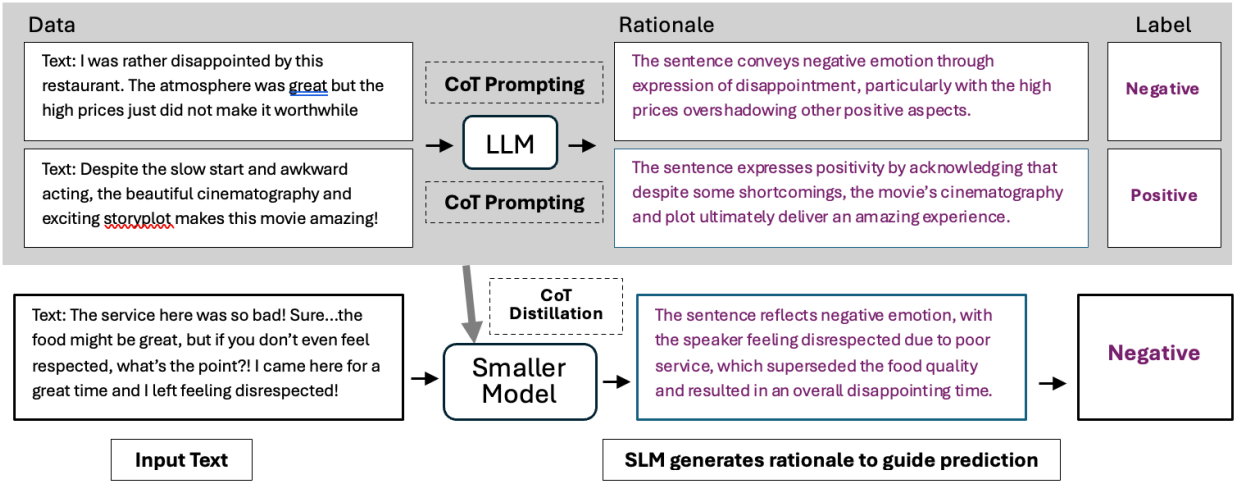

Figure 2: Overview of CoT Distillation. Firstly, CoT prompting utilised to obtain rationales from LLM. Generated reasoning then used to train smaller task-specific models using multi-task learning approach where smaller models generate rationale to guide predictions.

3.1 Extracting Rationales from LLMs

Recent research has noted the development of LLMs to be able to create rationales explaining their predictions (Kojima, 2022). In particular, we utilize Chain-of-Thought (CoT) prompting (Wei, 2022) to obtain prediction rationales from LLMs.

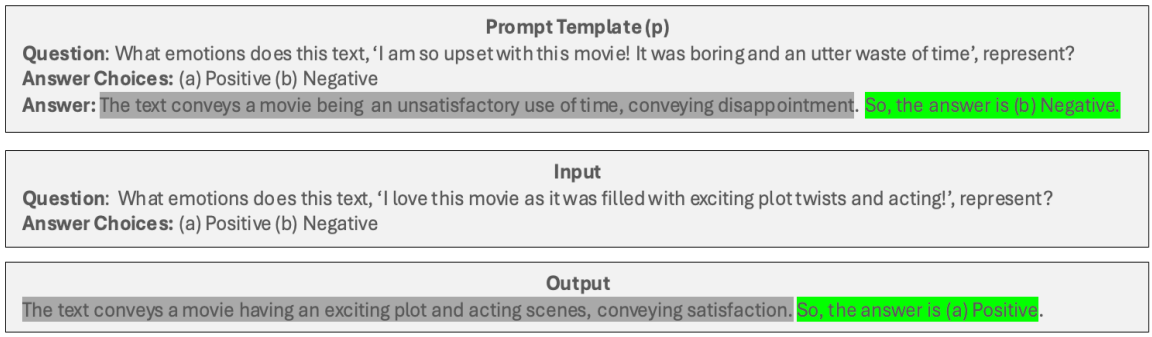

Figure 3: CoT prompting involves using a prompt template, which shows an example reasoning (highlighted in pink) and a predicted label (highlighted in green). Prompt template is appended to each input, allowing LLM to mimic how it should generate rationale and predictions for each new input example.

As shown in Figure 3, if we have an unlabelled training set $Xi \in D$, we create a prompt template p which shows how a particular sentiment analysis task should be tackled. Each prompt (Xp, Rp, Yp) contains an example input (Xp), the associated label (Yp) and a user-curated rationale (Rp), showing why Xp is labelled as Yp. We then attach each input Xi to the prompt template P, using it as an entire input to prompt the LLM to generate both labels and rationales for each $Xi \in D$. With the given example in P, the LLM can simulate the example to generate a rationale Ri and label Yi for each given Xi. In the case of a labelled training set, the LLM will be provided with both prompt template P, as well as input Xi and its associated label, Yi, and asked to predict a rationale Ri based on the text and human-annotated labels.

3.2 Standard Finetuning and Task Distillation

Our CoT distillation extends from current approaches to train smaller task-specific models by including rationales into the training procedure. Mathematically, we define our dataset as $D = \{ (Xi, Yi) \}$, where Xi represents an input and Yi is the associated target output label. The most widespread approach to training task-specific models involves finetuning pretrained models with labelled data (Ruder, 2018). When labelled data is not available, task-specific distillation (Tang, 2019) uses LLM as a 'teacher' and the smaller model as a 'student' (Casey, 2024). The LLM 'teacher' creates pseudo noisy training labels, $\hat{Y}i$ to substitute Yi (Arora, 2022).

Under both approaches, the task-specific smaller model f is trained to minimize label prediction loss:

$$\mathcal{L}_{\text{label}} = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), \hat{y}_i), \qquad (1)$$

Here, l represents the cross-entropy loss calculated between predicted tokens and target tokens. For simplicity, $\hat{Y}i$ is used to denote both human-annotated labels Yi in the case of standard fine-tuning, and labels predicted by LLM, $\hat{Y}i$, when referring to standard distillation.

3.3 Multi-task learning with rationales

In order to improve the association between Xi's to $\hat{Y}i$'s, we utilised our LLM rationales $\check{R}i$, for extra supervision. The simplest approach to doing this, is including $\check{R}i$ as an extra input (Wang, 2022). Mathematically, our smaller model $f(Xi, \check{R}i) \rightarrow \hat{Y}i$ will be trained with both the input text and rationale $[Xi, \check{R}i]$ as inputs.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i, \hat{r}_i), \hat{y}_i). \qquad (2)$$

However, the flaw of such an approach is that it needs a LLM to generate a rationale, *Ři,* before the smaller model, *f*, can make any label predictions. This implies the LLM is still required during deployment, which defeats the purpose of training a smaller model.

Thus, rather than using our rationales, *Ři,* as extra model inputs, we will instead define our training process as a multi-task problem. In particular, we will train the smaller model *f (Xi) → (Ŷi, Ři)* to both predict the output label, but also the associated rationale given the inputs. Here, L_{label} , is given by Equation 1 describing the *output label prediction loss*, and $L_{rationale}$, is the *rationale generation loss* given by:

$$\mathcal{L} = \mathcal{L}_{\text{label}} + \lambda \mathcal{L}_{\text{rationale}}, \qquad (3)$$

$$\mathcal{L}_{\text{rationale}} = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), \hat{r}_i). \qquad (4)$$

The rationale loss function helps the model to learn how to generate the intermediary reasoning steps leading to the label prediction. This potentially 'directs' the model to making better label predictions. Under this multi-task approach, *Ři*, is no longer required during testing, eliminating the need for a LLM at deployment as well.

3.4 Experiment Synopsis

To gauge our CoT distillation models, we will be conducting two sets of experiments. (1) Firstly, we will assess CoT distillation's efficacy compared to existing approaches of training task-specific smaller models – (a) standard finetuning, and (b) task-specific distillation. This is done by comparing the performance of these different methods using different training set sizes. (2) Secondly, we will assess the performance of task-specific smaller models, especially CoT distillation, compared to a benchmark LLM. This is done by comparing the performance of SLMs of different sizes, trained through various methods – finetuning, distillation and CoT distillation – compared to a 'representative' LLM performance.

3.5 Dataset

Our experiment is based on the IMDb Dataset (Maas, 2011) and the Yelp Review Sentiment Dataset (Zhang, 2015). Utilising these two vastly different datasets allow us to obtain a more impartial view of the model's ability to perform sentiment analysis across different types of texts.

IMDb Dataset: Binary sentiment analysis dataset containing movie reviews that are either classified as 'positive' or 'negative'. There are an equal number of positive and negative reviews.

Yelp Review Sentiment Dataset: Contains Yelp reviews on wide range of subjects – restaurants, shopping, entertainment etc. Reviews contain classification into 'positive' and 'negative'.

3.6 Training Configurations

3.6.1 Loss Function

Finetuning and Standard Distillation: As mentioned in section 3.2, we seek to minimise the label prediction loss, which is the average of the cross-entropy loss for the training dataset. y_i is replaced by $\hat{y}_i$ for Standard distillation, where $\hat{y}_i$ is the predicted label from the LLM.

$$L_{label} = 1/N \sum_{i=1}^{N} l_{CE}(f(x_i),\ y_i)$$

CoT Distillation: As mentioned in section 3.3, we utilise an additional rationale generation loss function, where r_i is the LLM generated rationale. Using both the label prediction loss and rationale generation loss functions, we then seek to minimise the following loss function derived by (Hsieh et al., 2023). The λ is fixed at **0.5** for our model.

$$L_{rationale} = 1/N \sum_{i=1}^{N} l_{CE}(f(x_i), r_i)$$

$$L = L_{label} + \lambda L_{rationale}$$

3.6.2 Model Selection

OpenAI's GPT 3.5 is our chosen LLM. It is used as a benchmark of LLM sentiment analysis performance, and also to generate LLM rationales and labels for our distillation methods. GPT is excellent as a LLM benchmark as it is the most widely used LLM (Guinness, 2024) and also has superior performance in sentiment analysis (Kheiri, 2023) compared to many other LLMs or machine learning approaches (Obinwanne, 2024).

Google's T5 Model is our chosen task-specific downstream model. Previous research has shown the applicability of T5 in developing smaller, more efficient models for sentiment analysis (Mengi, 2023), as compared to LLMs like GPT3, PaLM (Hsieh, 2023).

3.6.3 Optimizer

We used the default optimizer of the seq2seqtrainer class in the transformers library – the AdamW optimizer. AdamW addresses the issues of the original Adam optimizer, reducing the likelihood of overfitting and providing a more stable convergence (Loschchilov, 2017).

3.6.4 Evaluation metrics

Apart from the standard performance metrics e.g. accuracy, precision etc., our experiments focus on performance at different model sizes and training set sizes. This is as developing smaller models with comparable accuracy to LLMs and reducing the amount of training data is the primary motivation for utilising CoT distillation.

3.6.5 Training Environment

Due to the high computational demand of our tasks, we utilised NTU's GPU Cluster to carry out both training and experimentation phases for our experiments. Additionally, we utilised Weights & Biases for logging and tracking our experimentation results. Our logged runs can be found at: https://docs.google.com/document/d/1DhuRN6F2GCOfDSNYX1gGR6UeH1I5j3NfkdlPrEsiD1I/edit?usp=sharing

4 Experiment 1: Data Efficiency of CoT Distillation

A key motivation for using CoT Distillation is the hypothesis that the added use of rationales in training the model, will improve the model's association between inputs and labels, reducing the amount of data required to achieve the same level of performance. To test this, we analysed the performance of – *CoT Distillation, Finetuning, Standard Distillation* – at 10%, 25%, 50% and 100% of training set sizes. The T5 Small, with 60M parameters, was used as the base model of comparison for this experiment.

4.1 Experiment 1 Results & Discussion

Figure 4 shows a comparison of the performances between *CoT Distillation* and *Finetuning* on labelled data. Across both datasets, *CoT Distillation* outperformed *Finetuning* across all training sizes of the labelled datasets. Most notably, with only 10% of the full IMDb and Yelp dataset, *CoT Distillation* is able to outperform *finetuning* utilising 100% of the datasets. For instance, *CoT Distillation* using 10% of the IMDb dataset produced close to 88% accuracy, substantially higher than *finetuning,* which produced roughly 85% accuracy with 100% of the data. This shows a remarkable improvement in the data efficiency and model performance enabled by *CoT Distillation.*

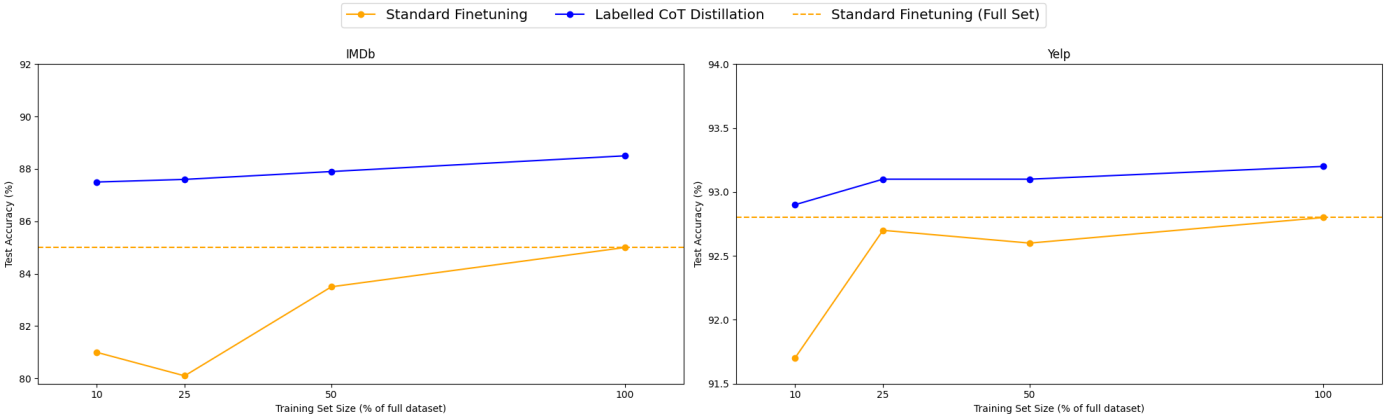

Figure 4: This graph compares Standard Finetuning and CoT Distillation using 60M T5 models trained on 10%, 25%, 50% and 100% of human-labelled datasets.

Figure 5 shows a comparison of the performances between *CoT Distillation* and *Standard Distillation* on unlabelled data. Similarly, *CoT Distillation* outperforms *Standard Distillation* across unlabelled dataset of all sizes. Most importantly, *CoT Distillation* with 10% of data achieved similar (marginally higher) performance to *Standard Distillation* with 100% of the data across both datasets. This highlights the reduction in training data needed by *CoT Distillation,* needing only 10% of the data used by conventional approaches to match or outperform them.

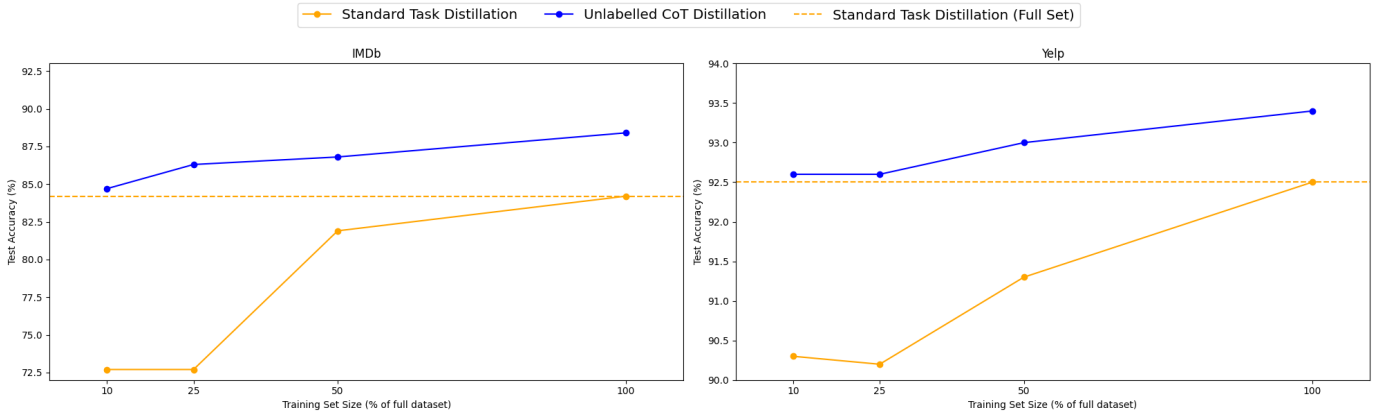

Figure 5: This graph compares Standard Task Distillation and CoT Distillation using 60M T5 models trained on 10%, 25%, 50% and 100% of LLM-labelled datasets.

5 Experiment 2: Higher Performance with Smaller Model Sizes

Apart from data efficiency, another important gauge of *CoT Distillation,* is whether it can develop smaller models that are able to match or outperform LLMs in the task of sentiment analysis. To test this, we will assess the accuracy of different sizes of smaller models trained through – *CoT Distillation, Finetuning, Standard Distillation* – and compare them to the LLM benchmark performance. The LLM benchmark is obtained through few-shot CoT prompting on the GPT 3.5 LLM – utilizing CoT demonstrations (see Figure 3) to prompt the 175B GPT 3.5 to obtain intermediate rationales before predicting the output labels without any further finetuning of the LLM.

5.1 Experiment 2 Results & Discussion

Figure 6 shows a comparison of *CoT Distillation* and *Finetuning* using labelled data against the LLM benchmark. Firstly, *CoT Distillation* outperforms *Finetuning* across both the 60M and 220M T5 models, which affirms the results in experiment 1. More notably however, the 220M *CoT Distillation* model is able

to achieve a higher performance than the 175B GPT 3.5 LLM Benchmark across both datasets. This is remarkable considering that it is roughly 800 times smaller in model size.

Additionally, while the *finetuning* model also outperforms the benchmark at 220M model size, it is worthy to note that as shown in experiment 1, *CoT Distillation* does this while simultaneously reducing the training data needed to a substantial degree. For instance, for the 60M T5 model, *CoT Distillation* only needs 10% of the training data to outperform *finetuning* using 100% of the labelled data. Thus, *CoT Distillation* achieves a substantial reduction in both model size and training data, while still outperforming the benchmark LLM.

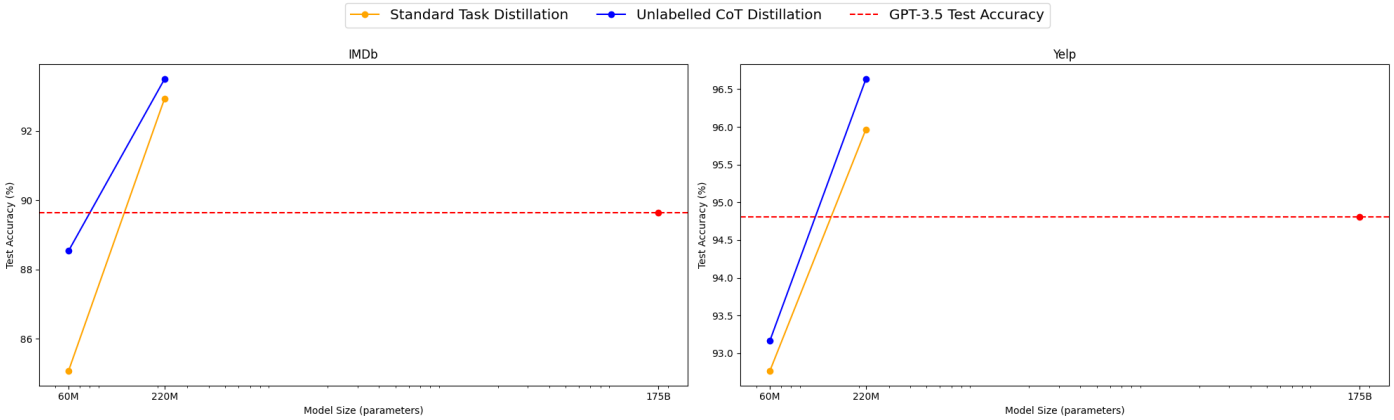

Figure 6: This graph compares Standard Finetuning and CoT Distillation using 60M and 220M T5 models on human-labelled datasets.

Figure 6 shows a comparison of *CoT Distillation* and *Standard Distillation* using unlabelled data against the LLM benchmark. Similarly, *CoT Distillation* outperforms *Standard Distillation* across both 60M and 220M model sizes, further affirming experiment 1. Again, the 220M *CoT Distillation* model outperforms the 175B LLM benchmark across both datasets. This further shows that even with unlabelled data, *CoT Distillation* provides a means to train substantially smaller models (~800 times smaller) that maintain comparable or higher accuracy than their LLM counterparts. More importantly, *CoT Distillation* achieves this while reducing the amount of unlabelled data needed. This is significant in practical applications where labelled data is usually hard or expensive to obtain, and distillation (using LLM to generate labels) for unlabelled data can be extremely time consuming and costly by itself.

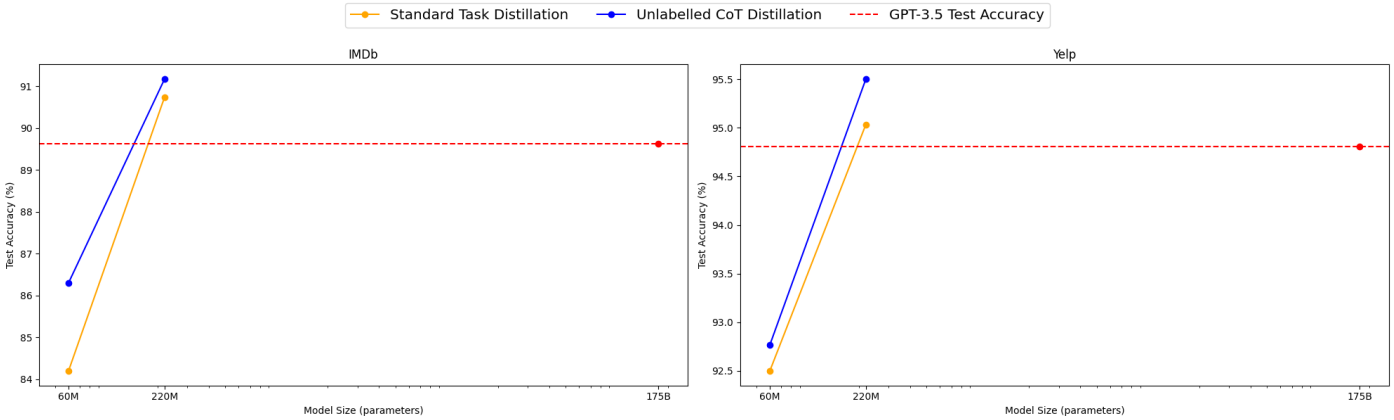

Figure 7: This graph compares Standard Task Distillation and CoT Distillation using 60M and 220M T5 models on LLM-labelled datasets.

6 Summary of Results

Methodology	IMDb Dataset: Training Set Size (%)			
	10%	25%	50%	100%
Labelled CoT Distillation	87.5%	87.6%	87.9%	88.5%
Unlabelled CoT Distillation	84.7%	86.3%	86.8%	88.4%
Standard Finetuning	81.0%	80.1%	83.5%	85.0%
Standard Distillation	72.7%	72.7%	81.9%	84.2%

Methodology	Yelp Review Dataset: Training Set Size (%)			
	10%	25%	50%	100%
Labelled CoT Distillation	92.9%	93.1%	93.1%	93.2%
Unlabelled CoT Distillation	92.6%	92.6%	93.0%	93.4%
Standard Finetuning	91.7%	92.7%	92.6%	92.8%
Standard Distillation	90.3%	90.2%	91.3%	92.5%

Methodology	IMDb Dataset: Different Model Sizes		
	60M	220M	175B
Labelled CoT Distillation	88.5%	93.5%	-
Unlabelled CoT Distillation	86.3%	91.1%	-
Standard Finetuning	85.1%	92.9%	-
Standard Distillation	84.2%	90.7%	-
GPT 3.5 LLM	-	-	89.6%

Methodology	Yelp Review Dataset: Different Model Sizes		
	60M	220M	175B
Labelled CoT Distillation	93.2%	96.6%	-
Unlabelled CoT Distillation	92.7%	95.5%	-
Standard Finetuning	92.8%	95.9%	-
Standard Distillation	92.5%	95.0%	-
GPT 3.5 LLM	-	-	94.8%

CoT Distillation has higher performance than conventional approaches to training task-specific models. Across both datasets, regardless of training set size or model size, CoT Distilled models outperform Standard Finetuning and Standard Distillation models. For instance, CoT Distillation has higher performance than Standard Distillation across all training sizes of the IMDb dataset – with the difference in performance when using 10% of the data being close to 15% (CoT Distillation – 87.5%, Standard Distillation – 72.7%). *This points to CoT distillation as a means to increase task-specific model accuracy by incorporating rationales.*

(1) **CoT Distillation has substantially higher data efficiency than conventional approaches to training task-specific models.** Across both datasets, the CoT Distilled model trained on only 10% of the data outperformed the Standard Finetuned or Standard Distilled model trained on all 100% of the data. For instance, the CoT Distilled model using 10% of Yelp data had 92.9% accuracy, which is higher than the 92.8% of the Finetuned model and 92.5% of Standard Finetuned model trained on 100% of Yelp data. *This points to CoT distillation as a way to substantially reduce the amount of training data needed to train highly accurate, task-specific small language models.*

(2) **CoT Distillation trains smaller models that outperforms LLMs, with substantially smaller model sizes and training data used.** Across both datasets, the 220M T5 CoT Distilled model outperforms the few-shot CoT 175GB GPT 3.5 LLM benchmark. For instance, the 220M T5 CoT Distilled model had a performance of 96.6% on the Yelp dataset, which is higher than the 94.8% achieved by the GPT 3.5 LLM. This is despite the CoT Distilled model being close to 800 times smaller than the GPT 3.5 model. Additionally, it is able to achieve this using significantly less training data as mentioned in point 2. *This points to CoT distillation as a way to train significantly*

smaller models to outperform LLMs for specific tasks, achieving substantial savings in computational and resource requirements.

7 Limitations and Future Work

7.1 Quality of Rationale

Research has shown that the performance of CoT Distillation could be impacted by the LLM chosen for rationale generation. For instance, one study showed that the lift in performance through CoT Distillation is smaller when the 20GB GPT-NeoX model was used over the 540GB PaLM model (Hsieh, 2023). This was probably due to the larger PaLM model providing higher quality rationales that benefited the training process more. Thus, further research can be done to explore how design choices, such as the choice of LLM, affects the quality of CoT Distillation.

7.2 Diversity of Dataset

This study has focused exclusively on sentiment analysis using the IMDb and Yelp review datasets, which are limited to the English language. To ensure the robustness and applicability of our findings, it is essential to extend our research to encompass a diverse array of languages and domains, such as healthcare and legal sectors. Such expansion would provide a more comprehensive evaluation of our theories and contribute to the development of universally applicable models.

7.3 Generalisability of Approach

Our paper covers CoT Distillation in the context of sentiment analysis. Other studies have also explored the efficacy of CoT approaches for other benchmark NLP tasks, such as common-sense question answering using the CQA Dataset (Wang, 2023) and arithmetic math word problems using the SVAMP Dataset (Hsieh, 2023). However, research has shown that LLMs exhibit limited reasoning capability on more complex inferential and planning tasks (Valmeekam, 2022). With CoT Distillation highly dependent on the quality of rationale provided, this points to areas of NLP that our approach might not be applicable to, requiring further research and optimisation.

8 Conclusion

To conclude, our paper introduced Chain of Thought (CoT) Distillation as an innovative approach to train task-specific small language models (SLMs) using significantly less data while still maintaining or even surpassing the performance of large language models (LLMs). Our experiments, focusing on sentiment analysis, have shown that smaller T5 models (220M) outperform the few-shot CoT GPT 3.5 LLM benchmark (175B), which is close to 800 times larger. Additionally, CoT Distillation requires significantly less training data, outperforming the two most common approaches to training task-specific models – finetuning and task-specific distillation – while using 90% less data. The findings show that including rationale generation in the training process yields more effective and compact models – ideal for practical applications where resources are limited. While successful in sentiment analysis, its applicability to other domains and how its performance differs based on design choices remains to be explored. Future research could focus on understanding the impact of rationale quality on the effectiveness of CoT Distillation and expanding its use beyond sentiment analysis to more complex NLP tasks. The advancements shown in this study open a crucial door to more sustainable and accessible NLP technologies, marking a step towards harnessing the power of LLMs and deep learning across all industries and applications.

9 References

Al-Shabi, M. A. (August 2021). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. https://www.researchgate.net/publication/343473213_Evaluating_the_performance_of_the_most_important_Lexicons_used_to_Sentiment_analysis_and_opinions_Mining

Arora, S., Narayan, A., Chen, M. F., Orr, L., Guha, N., Bhatia, K., Chami, I., Sala, F., & Ré, C. (20 November 2022). *Ask me anything: A simple strategy for prompting language models*. arXiv.org. https://arxiv.org/abs/2210.02441

Borg, A., & Boldt, M. (July 2020). Using VADER sentiment and SVM for predicting customer response sentiment. https://www.researchgate.net/publication/343288577_Using_VADER_Sentiment_and_SVM_for_Predicting_Customer_Response_Sentiment

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (22 July 2020). *Language models are few-shot learners*. arXiv.org. https://arxiv.org/abs/2005.14165

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (5 October 2022). *Palm: Scaling language modeling with pathways*. arXiv.org. https://arxiv.org/abs/2204.02311

Daigle, R. (26 September 2023). *Understanding and enabling the transformational power of LLMs*. Lenovo StoryHub. https://news.lenovo.com/understanding-and-enabling-the-transformational-power-of-llms/

Daniels, E. (29, January 2024). *The emergence of small language models (slms)*. Version 1. https://www.version1.com/the-emergence-of-small-language-models/#:~:text=What%20are%20Small%20Language%20Models,and%20generate%20human%2Dlike%20language.

Guinness, H. (30 January 2024). *The best large language models (llms) in 2024*. Automate your work today. https://zapier.com/blog/best-llm/

Gunasekaran, K. P. (24 May 2023). *Exploring sentiment analysis techniques in Natural Language Processing: A Comprehensive Review*. arXiv.org. https://arxiv.org/abs/2305.14842

Hase, P., & Bansal, M. (10, February 2021). *When can models learn from explanations? A formal framework for understanding the roles of Explanation Data*. arXiv.org. https://arxiv.org/abs/2102.02201

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., & Pfister, T. (5 July 2023). *Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes*. arXiv.org. https://arxiv.org/abs/2305.02301

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (25 October 2022). *Large language models can self-improve*. arXiv.org. https://arxiv.org/abs/2210.11610

Kheiri, K., & Karimi, H. (23 July 2023). *Sentimentgpt: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning*. arXiv.org. https://arxiv.org/abs/2307.10234

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (29 January 2023). *Large language models are zero-shot Reasoners*. arXiv.org. https://arxiv.org/abs/2205.11916

Loshchilov, I., & Hutter, F. (4 January 2019). *Decoupled weight decay regularization*. arXiv.org. https://arxiv.org/abs/1711.05101

Maas, A. (2011). *Large Movie Review Dataset*. Sentiment Analysis. https://ai.stanford.edu/~amaas/data/sentiment/

Mengi, R., Kakade, A., & Ghorpade, H. (December 2023). Fine-tuning T5 and RoBERTa Models for Enhanced Text Summarization and Sentiment Analysis. https://www.researchgate.net/publication/376232167_Fine-tuning_T5_and_RoBERTa_Models_for_Enhanced_Text_Summarization_and_Sentiment_Analysis

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010, September). Recurrent neural network based language model. http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf

Obinwanne, T., & Brandtner, P. (16 February 2024). *Enhancing sentiment analysis with GPT-A comparison of large language models and traditional machine learning techniques*. SpringerLink. https://link.springer.com/chapter/10.1007/978-981-99-7569-3_17

Pruthi, D., Bansal, R., Dhingra, B., Soares, L. B., Collins, M., Lipton, Z. C., Neubig, G., & Cohen, W. W. (6 April 2022). *Evaluating explanations: How much do explanations from the teacher aid students?*. MIT Press. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00465/110436/Evaluating-Explanations-How-Much-Do-Explanations

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (19 September 2023). *Exploring the limits of transfer learning with a unified text-to-text transformer*. arXiv.org. https://arxiv.org/abs/1910.10683

Sadia, A., Khan, F., & Bashir, F. (February 2018). An Overview of Lexicon-Based Approach For Sentiment Analysis. https://ieec.neduet.edu.pk/2018/Papers_2018/15.pdf

Schick, T., & Schütze, H. (12 April 2021). *It's not just size that matters: Small language models are also few-shot learners*. arXiv.org. https://arxiv.org/abs/2009.07118

Sen, A. (9 November 2021). *Ensemble Modeling for Neural Networks using large datasets – Simplified!* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/ensemble-modeling-for-neural-networks-using-large-datasets-simplified/#:~:text=Neural%20networks%20being%20complex%20models,base%20models%20on%20each%20subset.

Singh, J., & Tripathi, P. (12 August 2021). Sentiment analysis of Twitter data by making use of SVM, Random Forest and decision tree algorithm | IEEE conference publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9509679

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R. Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., & Catanzaro, B. (4 February 2022). *Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model*. arXiv.org. https://arxiv.org/abs/2201.11990

Thomas, M., & A, L. C. (August 2018). Sentimental Analysis using recurrent neural network. https://www.researchgate.net/publication/328488167_Sentimental_analysis_using_recurrent_neural_network

Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., & Kambhampati, S. (26, November 2023). *PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change*. arXiv.org. https://arxiv.org/abs/2206.10498

Vig, J., & Belinkov, Y. (18 June 2019). *Analyzing the structure of attention in a transformer language model*. arXiv.org. https://arxiv.org/abs/1906.04284

Wang, P., Wang, Z., Li, Z., Gao, Y., Yin, B., & Ren, X. (30 August 2023). *Scott: Self-consistent chain-of-thought distillation*. arXiv.org. https://arxiv.org/abs/2305.01879

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (10 January 2023). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv.org. https://arxiv.org/abs/2201.11903

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (11 March 2021). *MT5: A massively multilingual pre-trained text-to-text transformer*. arXiv.org. https://arxiv.org/abs/2010.11934

Zheng, L., Li, Z., Zhang, H., Zhuang, Y., Chen, Z., Huang, Y., Wang, Y., Xu, Y., Zhuo, D., Xing, E. P., Gonzalez, J. E., & Stoica, I. (28 June 2022). *Alpa: Automating inter- and intra-operator parallelism for distributed deep learning*. arXiv.org. https://arxiv.org/abs/2201.12023