The Impact of Knowledge Distillation on the Performance and Energy Consumption of NLP Models

Ye Yuan
2757396
VU Amsterdam
y.yuan3@student.vu.nl

Eloise Zhang
2764376
VU Amsterdam
j.zhang6@student.vu.nl

Zongyao Zhang
2774811
VU Amsterdam
z.zhang14@student.vu.nl

Kaiwei Chen
2761177
VU Amsterdam
k.chen2@student.vu.nl

Jiacheng Shi
2760762
VU Amsterdam
j.shi2@student.vu.nl

ABSTRACT

Context. Our research tackles a crucial challenge in Natural Language Processing (NLP). While models like BERT and GPT are powerful, they require substantial resources. Knowledge distillation can be employed as a technique to enhance their efficiency. Yet, we lack a clear understanding on their performance and energy consumption. This uncertainty is a major concern, especially in practical applications, where these models could strain resources and limit accessibility for developers with limited means. Our drive also comes from the pressing need for environmentally-friendly and sustainable applications in light of growing environmental worries. To address this, it is crucial to accurately measure their energy consumption.
Goal. This study aims to determine how Knowledge Distillation affects the energy consumption and performance of NLP models.
Method. To explore the impact of distillation techniques on NLP models, we benchmark BERT, Distilled-BERT, GPT-2, and Distilled-GPT-2 using three different tasks from three different categories selected from a third-party dataset. During the experiment, the energy consumption, CPU utilization, memory utilization, and inference time of the considered NLP models are measured and statistically analyzed.
Results. The study reveals notable differences between the original and the distilled version of the measured NLP models. Distilled versions generally consume less energy, with distilled GPT-2 having reduced CPU utilization. These results provide evidence and insights about the possible trade-offs in using distilled models for NLP.
Conclusion. The results of this study highlight the critical impact of model choice on performance and energy consumption metrics. Future research should consider a wider range of distilled models, diverse benchmarks, and deployment environments, as well as explore the ecological footprint of these models, particularly in the context of environmental sustainability.

ACM Reference Format:
Ye Yuan, Eloise Zhang, Zongyao Zhang, Kaiwei Chen, and Jiacheng Shi. 2023. The Impact of Knowledge Distillation on the Performance and Energy Consumption of NLP Models. In *Proceedings of 3rd International Conference on AI Engineering — Software Engineering for AI (CAIN 2024).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

"Amidst the rapid evolution of Natural Language Processing (NLP), a pressing question arises: Can we sustain technological advancement while being environmentally conscious?" This inquiry launches our exploration into Knowledge Distillation within NLP models—a technique poised to redefine efficiency in AI.

NLP's landscape has dramatically transformed with the advent of models like BERT and GPT-2, setting new standards for applications ranging from chatbots to content generation. Nevertheless, their extensive resource demands spotlight a critical challenge: balancing computational prowess with environmental responsibility[1]. Knowledge Distillation emerges as a compelling solution, distilling the essence of these behemoths into more compact, efficient models without sacrificing their core capabilities[2].

This study delves into the intricacies of Knowledge Distillation, driven by the urgent need to reconcile the power of NLP models with practical and sustainable usage[3]. By focusing on energy consumption, we aim to illuminate a path for developing more efficient AI models, contributing to a more responsible technological future. In an era where AI's environmental impact is increasingly scrutinized, our research is catalyzed by the imperative to optimize the energy efficiency of state-of-the-art NLP models[4]. We investigate the dichotomy between their remarkable capabilities and substantial resource demands, aiming to bridge this gap with Knowledge Distillation. Our motivation is twofold: advancing AI's frontiers while fostering sustainability, a balance critical for the continued growth and accessibility of NLP technologies[5].

The **main goal** of this paper is to empirically assess whether Knowledge Distillation can significantly enhance the energy efficiency of NLP models without compromising their performance. We critically analyze the potential impact of this technique on models like BERT and GPT-2 and their distilled counterparts to offer insights into sustainable AI development. To do that, *we perform an*

CAIN 2024, April 2024, Lisbon, Portugal

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

experiment in which the NLP models are tested with the GLUE benchmark [6]. In particular, we measure the performance and the energy consumed by the models executing the SST-2, STS-B, and MNLI tasks of the GLUE benchmark. Our findings aim to guide developers towards more environmentally conscious decisions in NLP model selection, shaping a future where AI is robust and sustainable.

The **results** of our study are multi-faceted, but also congruent. We observed a statistically significant difference in energy consumption among the considered NLP models, with distilled models tending to consume less energy compared to their non-distilled counterparts. We also observed statistically significant differences among models in terms of performance metrics (*i.e.,* CPU usage, memory usage, and inference time). Our analysis revealed significant variations in CPU usage across different models. The results indicated that distilled models exhibited lower CPU usage compared to their non-distilled counterparts, suggesting a more efficient utilization of processing resources. We observed that memory consumption patterns varied significantly between models, with distilled models showing higher efficiency in memory utilization than without distilled. This aspect of performance is crucial for applications where memory resources are a limiting factor. Our study also suggests there are notable differences in inference times, with distilled models achieving faster processing speeds, indicating their suitability for real-time applications.

The **main contributions** of this study are (i) an empirical assessment of the impact of Knowledge Distillation on the performance and energy consumption BERT and GPT-2 NLP models, (ii) a discussion of the obtained results, and (iii) the full replication package [7] of the study containing raw data, source code, and scripts for data analysis.

2 RELATED WORK

Cao *et al.* [4] highlight the pressing need for accurate measurement of energy consumption in the development and training of large-scale NLP models. Their study demonstrates the limitations of current software-based energy assessments, which overlook important factors like hardware differences and the impact of resource usage on energy consumption. Through meticulous experiments using a hardware power meter, the authors obtain precise energy measurements and compare them with software-based estimates for four Transformer-based NLP models. The results reveal a significant 20% average difference between software-based estimates and actual hardware measurements, along with larger standard deviations. However, it's worth noting that while the hardware power meter proves reliable, its accessibility to researchers may vary, potentially limiting its widespread use. Additionally, the study primarily focuses on energy measurement, leaving out critical aspects like latency or memory usage, which warrant further investigation into the overall resource footprint of these models. In our research, we will also consider performance metrics such as CPU usage, memory consumption, and inference time (in milliseconds) for a more comprehensive evaluation of large-scale NLP model resource dynamics.

Cao *et al.* [8] further introduce "IrEne," an innovative and robust energy prediction system designed to address the inadequacies of existing software-based energy measurements for Transformer-based NLP models. Recognizing the complex interplay between energy consumption and model execution, IrEne adopts an interpretable and extensible approach. It achieves this by constructing a model tree abstraction that dissects NLP models into model-specific modules and lower-level machine learning primitives. IrEne's predictive capabilities are grounded in its ability to estimate the inference energy consumption of individual components within the model tree, rendering energy predictions highly interpretable. This approach further encompasses a multilevel prediction methodology, employing resource utilization and model description features for energy estimation. However, it is important to note that this study primarily focuses on energy consumption during the inference phase. Aspects like CPU usage, memory consumption, and inference time were not directly addressed. These metrics will be considered in our research to provide a more complete understanding of resource utilization.

Zadeh *et al.* [9] proposed a model quantization technique called GOBO for compressing the size of model parameters without reducing the accuracy. In this study, the authors applied GOBO to the BERT model and compared the GOBO version of BERT and other BERT variants such as DistillBERT in terms of several measurement dimensions. In this study, GLUE was used as a benchmark for model stress testing. According to the statistical analysis, GOBO can improve TPU performance by reducing model footprint while reducing energy consumption by reducing memory traffic. This research mainly focuses on exploring the impact of specific model quantization techniques on accuracy as well as performance compared to other techniques while energy consumption is not a major research objective. However, our study focuses on the impact of knowledge distillation techniques on energy consumption and performance. Therefore, our study is different from it in terms of both the object and goal of the study. In addition, our study covers not only the BERT model but also the study of the GPT-2 model.

Choi *et al.* [10] proposed AttAcc, an accelerator for the attention layer of Transformer-based generative models. In this study, the authors firstly acknowledge the improvement in throughput and reduction in energy consumption of DGX (serving platform) by applying batch processing in GPT. Further, this study points out that increasing the batch size is the key factor in achieving the above effects. However, Choi *et al.* also concluded that increasing batch size also has a negative effect on storage space and execution time. Therefore, the authors designed a distributed heterogeneous computing system using AttAcc to mitigate the above problems. The authors conducted an experiment using GPT-3 as a subject to compare the performance and energy consumption of pure DGX and AttAcc (DGX+AttAcc). The final results show that the throughput of DGX+AttAcc is significantly better than that of the DGX-only platform (3.24 times higher (640G)). In addition, DGX+AttAcc reduces energy consumption in terms of off-chip memory accesses by reducing the number of accesses. This experiment focuses on exploring the impact of the hardware architecture on the energy consumption and performance of TbGMs whereas our experiment focuses on exploring the impact of knowledge distillation, a software technique, on TbGMs' energy consumption and performance.

3 EXPERIMENT DEFINITION

3.1 Goal and Research Questions

We formulate the goal of this study according to the Goal-Question-Metrics framework [11]. By referring to Table 1, the goal of this paper is to analyze distilled transformers for evaluating their energy consumption and performance from the point of view of software developers in the context of NLP models.

Table 1: Goal description following the GQM framework

Analyze	Distilled Transformers
for the purpose of	Evaluation
with respect to their	Energy Consumption and Performance
from the point of view of	Software Developers
in the context of	Natural Language Processing models

We achieve the above-mentioned goal by answering the following two research questions.

RQ1- *How does the Knowledge Distillation affect the energy consumption of NLP models?* We explore if the process of distillation, which typically involves transferring knowledge from a larger model to a smaller one, has any significant influence on the energy footprint. To answer this RQ, we test the NLP models with the GLUE benchmark and measure the energy consumed performing the SST-2, STS-B, and MNLI tasks.

RQ2- *How does the Knowledge Distillation affect the performance of NLP models?* This question investigates if the distilled version of the NLP models maintain, enhance, or potentially compromise their performance compared to the original version. As done for energy consumption, we observe both groups of models, i.e., the original and distilled version, executing the SST-2, STS-B, and MNLI tasks included in the GLUE benchmark.

Our research question focus on energy consumption and performance NLP models. We evaluate the performance of a NLP model using CPU Utilization, Memory Utilization, and Inference Time as metrics. The details of each metric are elaborated in the following listing:

- **Energy Consumption (J):** The energy consumption of performing a run in the experiment is measured in *Joules*.
- **CPU Utilization (%):** The *percentage* of time the CPU is actively executing instructions, as opposed to being idle, is calculated as Execution Time.
- **Memory Utilization (%):** The *percentage* of a computer system's memory that is used by the NLP model during its operation.
- **Inference Time (s):** The time elapsed in *seconds* from the moment a model receives an input until it produces a result.

CPU Utilization gives an indication of the computational resources required to process tasks, high CPU usage can result in system overheating, overloads, or performance degradation. Memory Utilization offers a direct insight into the size and complexity of a model. Distilled models, being typically smaller, should consequently occupy less memory. However, this does not directly infer a performance advantage and needs to be considered in conjunction with the other two metrics. Inference Time directly relates to the speed at which a model processes an input and produces an output. We expect to observe significant differences in processing time for the distilled model. Throughout the experiment, we collect and analyze data based on these metrics to address the specified questions and provide insights for developers regarding the energy efficiency of NLP model choices. These indicators are easily comparable and quantifiable.

4 EXPERIMENT PLANNING

4.1 Subjects Selection

In this section, we present the NLP models we chose for our experiments, as well as the benchmarks we used to evaluate it.

In this study, we compare *GPT-2* and *BERT*, with their corresponding distilled version. GPT-2 has shown exemplary performance across a range of NLP tasks. However, its computational demands are high due to its extensive parameter count. BERT is another transformer-based model that has set new performance benchmarks on several NLP tasks. Similar to GPT-2, BERT has a high computational demand due to its large model size. Distilled GPT-2 and distilled BERT are both lighter versions created through Knowledge Distillation.

The benchmark tasks chosen for this study are derived from the *General Language Understanding Evaluation* (GLUE) dataset[6], specifically the tasks of *SST-2*, *STS-B*, and *MNLI*. GLUE is a collection of resources for training, evaluating, and analyzing natural language understanding systems. The dataset encompasses various tasks that evaluate the capacity of models to understand the nuances and subtleties of human language. The rationale for its selections is that GLUE amalgamates multiple tasks that span a wide range of NLP challenges, also, GLUE tasks derive from a mix of sources, ranging from news articles to fiction, thereby presenting a more holistic challenge for models. SST-2 is a single-sentence categorization task, that contains human annotations of sentences from movie reviews and their sentiment. This task is given the sentiment of a sentence, and the categories are divided into two types of positive sentiment and negative sentiment. STS-B is a benchmark test for evaluating the capabilities of natural language processing models, focusing on the performance of the model in measuring the semantic similarity of two sentences. MNLI requires the model to determine the relationship between one sentence (premise) and another (assumption). It uses texts from a wide range of genres and topics providing a more complex and diverse testing environment for the model to assess its ability to generalize across different contexts.

4.2 Experimental Variables

In this study, we empirically assess the impact of Knowledge Distillation on the energy consumption and performance of NLP models. The dependent variables correspond to the metrics used to quantify energy consumption and performance. As emphasized by Section 3.1, we measure the energy consumption in Joule of an execution of a NLP model, while to analyze performance we consider *CPU usage*,*memory usage*, and *inference time*. Because we are only interested in the consequences of applying Knowledge Distillation to NLP models, the *application of the technique* embodies our only independent variable. This variable has two treatments including the application or not of Knowledge Distillation. Thus, the GPT-2 and BERT are considered as values without the distillation, and vice versa.

4.3 Experimental Hypotheses

In this research, we aim to reason about the impact of implementing knowledge distillation on each independent variable. To accomplish this in a scientifically meaningful way, we will formulate several hypotheses about the potential outcomes of our experiment. We can answer our research question by relating our findings to these hypotheses.

To answer RQ1, we construct the following null hypothesis:

$$H_0^{model,e} : \mu^e_{GPT2} = \mu^e_{D-GPT2} \cap \mu^e_{BERT} = \mu^e_{D-BERT}$$

The word **model** represents the independent variable for the selected NLP models, while **e** stands for the energy consumption of the selected model running the NLP tasks. Consequently,$H^{model,e}$ then determines the effect of the chosen NLP model on energy consumption. Furthermore, μ^e_{model} represents the average measurement result of energy consumption with the selected model as treatment.

The null hypothesis itself states that no meaningful difference in energy consumption can be detected when executing our benchmark with different NLP models. This leads to the following alternative hypothesis, stating that for energy consumption, a statistically relevant difference can be observed between normal NLP models and their distilled version:

$$H_a^{model,e} : \mu^e_{GPT2} \neq \mu^e_{D-GPT2} \cup \mu^e_{BERT} \neq \mu^e_{D-BERT}$$

To answer RQ2, we construct the following null hypothesis:

$$H_0^{model,metric} : \mu^{metric}_{GPT2} = \mu^{metric}_{D-GPT2} \cap \mu^{metric}_{BERT} = \mu^{metric}_{D-BERT}$$
$$\forall metric \in \{\text{CPU usage, memory usage, inference time}\}$$

The word **model** represents the independent variable for the selected NLP models, while **metric** stands for a dependent variable such that $metric \in \{CPU\ usage,\ memory\ usage,\ inference\ time\}$. Consequently,$H^{model,metric}$ then determines the effect of the chosen NLP model on our dependent variable **metric**. Furthermore, μ^{metric}_{model} represents the average measurement result of variable **metric** with the selected model as treatment.

The null hypothesis itself states that no meaningful difference for any of our selected metrics can be detected when executing our benchmark with different NLP models. This leads to the following alternative hypothesis, stating that for performance, a statistically relevant difference can be observed between normal NLP models and their distilled version:

$$H_0^{model,metric} : \mu^{metric}_{GPT2} \neq \mu^{metric}_{D-GPT2} \cup \mu^{metric}_{BERT} \neq \mu^{metric}_{D-BERT}$$
$$\exists metric \in \{\text{CPU usage, memory usage, inference time}\}$$

4.4 Experiment Design

Based on the identified subjects, variables, and hypotheses, we design an experiment focusing on evaluating the impact of Knowledge Distillation on energy consumption and performance of NLP models. The design adopted will be of a *one factor multiple treatments* type (1F-MT).

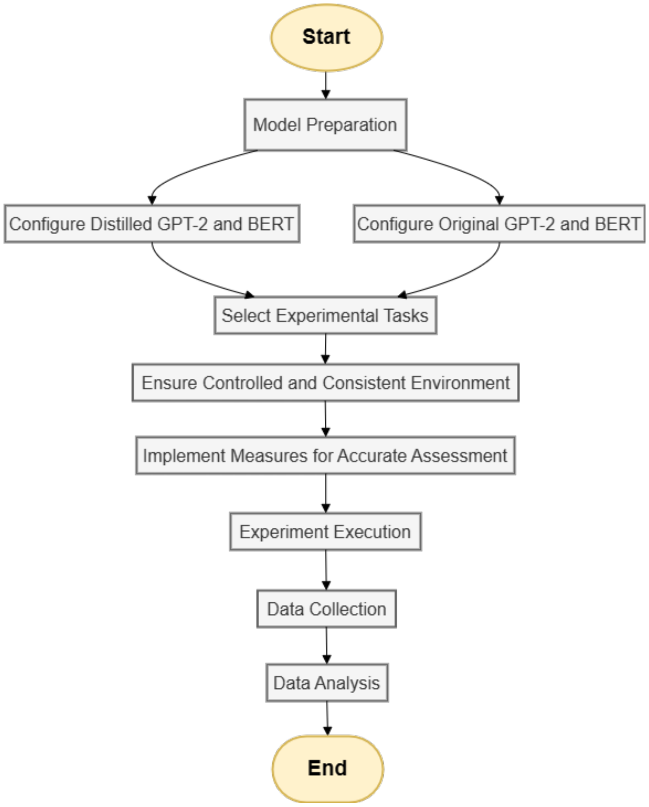
Start
Model Preparation
Configure Distilled GPT-2 and BERT
Configure Original GPT-2 and BERT
Select Experimental Tasks
Ensure Controlled and Consistent Environment
Implement Measures for Accurate Assessment
Experiment Execution
Data Collection
Data Analysis
End

Figure 1: Main phases of the performed experiment

Figure 1 illustrates the comprehensive process of evaluating the impact of Knowledge Distillation on energy consumption and performance of NLP models. It begins with the *Model Preparation* phase, where both distilled and original versions of GPT-2 and BERT models are prepared. The process then progresses through selecting experimental tasks, ensuring a controlled environment, and implementing specific measures for an accurate assessment. The experiment is executed in two parts, focusing on energy consumption and performance metrics, followed by systematic data collection, analysis, and interpretation of results. The process culminates in drawing conclusions and making recommendations.

The Model Preparation phase, the Distilled GPT-2 and Distilled BERT models, along with their original counterparts, GPT-2 and BERT, are meticulously prepared and configured for the selected experimental tasks. Ensuring a controlled and consistent environment is pivotal. To ensure our experimental design accurately assesses the impact of different models on energy consumption and performance, we have implemented the following measures:

- *Hardware and Software Configuration:* We ensure that each model is initialized on identical hardware and software configurations. This includes using processors, memory, and other critical hardware components of the same model and configuration and ensuring all models run on the same version of operating systems and libraries.
- *Consistency in Hyperparameters:* To ensure uniformity in experimental conditions, we use consistent hyperparameter settings for each model, such as *learning rate*, *batch size*, and *number of iterations*.

- *Data Preprocessing and Normalization:* All models undergo the same data preprocessing steps and input normalization methods. This includes *data cleaning*, *formatting*, and *applying the same transformation and normalization techniques.*
- *Minimizing External Interference and Ensuring Stable Power Supply:* To ensure the accuracy and reliability of our data collection, we conduct our experiments in a controlled environment. This includes using an Uninterruptible Power Supply (UPS) system to ensure the stability of the power supply. Additionally, the laboratory environment is monitored to minimize external disturbances, such as fluctuations in temperature and humidity.

A setup with minimal external interference and a stable power supply will be established to foster accurate and reliable data collection.

In the *Benchmark Task Preparation* phase, we have chosen SST-2, STS-B, and MNLI from the GLUE dataset for benchmarking. The extraction process involves filtering the GLUE dataset to retain only the data relevant to these three tasks. For SST-2 (Stanford Sentiment Treebank), this task involves single-sentence categorization, focusing on sentiment analysis. We selected it due to its relevance in understanding models' ability to interpret subjective information in text; For STS-B (Semantic Textual Similarity Benchmark), this task evaluates the model's performance in measuring the semantic similarity of two sentences. It is chosen to assess the models' nuanced understanding of language semantics; For MNLI (Multi-Genre Natural Language Inference), this task requires models to determine the relationship between a pair of sentences. Its inclusion is due to its complexity and diversity, testing the models' ability to generalize across different contexts. We utilize a data processing script that identifies and segregates entries corresponding to SST-2, STS-B, and MNLI based on their metadata labels. This ensures that our benchmarking tasks are focused exclusively on the datasets of interest, allowing for a targeted and consistent evaluation.

As we transition to the *Task Execution and Data Collection* phase, each model undergoes a predefined protocol for task execution. Each model, i.e., BERT, GPT-2, Distilled BERT, and Distilled GPT-2, goes through to three rounds of execution of each selected task: SST-2, STS-B, and MNLI. Therefore, we execute each model for ten runs, each one including the execution of each task. This process ensures consistency and account for any potential variations in individual runs. Moreover, the repetition of runs helps to mitigate any interference or data skewness that might arise from one-off anomalies or concurrent multitasking operations.

As mentioned in Section 3.1, we measure energy consumption, CPU Utilization, Memory Utilization, and Inference Time for each execution. In the *Preliminary Data Analysis* phase, we compute descriptive statistics, which offers initial insights into the data. This process entails calculating the mean, median, mode, range, and standard deviation for each set of data points gathered. Additionally, we clean the data to identify and remove outliers and address any noise or errors that might have infiltrated the data during the collection phase. The collected dataset if organized in the *Data Storage* phase. In this phase, we gathered raw data and the data synthesized in the Preliminary Data Analysis phase and organized it in the replication package [7]. This step is critical to allow practitioners to inspect the data more thoroughly, and replicate the study using different NLP models.

5 EXPERIMENT EXECUTION

In this section, we describe the environment in which we execute our experiment, how we set it up, and how we collect the metrics. The experiment is configured and orchestrated via Experiment Runner[12], a tool for the automatic orchestration of measurement-based experiments.

5.1 Setup

The experiment is executed on a server equipped with a 1TB hard drive, 32GB of RAM, and a 3.4GHz x 8 core CPU. The server operates on Linux and is equipped with Experiment Runner for seamless trial control. For secure connection to the server, we utilize Secure Shell (SSH), a cryptographic network protocol that ensures secure communication over an unsecured network. SSH works by establishing an encrypted connection between a local computer and a remote server, allowing for secure data transmission. This is achieved by using the IP address of the server along with appropriate authentication credentials. Once connected via SSH, we execute the benchmarks on the server. This involves initiating the necessary commands through the terminal interface to trigger the execution of the NLP Models on the designated benchmarks. During these runs, we closely monitor energy consumption and performance. Each run encompasses testing an NLP Model on three benchmarks, performed ten times consecutively.

5.2 Measurements

To gather the metrics we introduced in Section 3.1, we integrate Experiment Runner with *PowerJoular* and Linux's *ps* command. PowerJoular [13] is a monitoring tool that can be used to monitor the CPU usage and energy consumption of a particular process. The *ps* command, which is one of the most commonly used commands in Linux, can capture the CPU usage as well as memory usage of processes. Capturing inference time is relatively easy because Python's *time* module provides timing capabilities. This allows us to count the time it takes from the start and end of the measurement. It is worth noting that we measure the inference time from start to end of each run which means that we calculate the time of all three inference tasks together.

5.3 Experiment Execution Order

For each run, an NLP model is tested on three benchmarks. Since the goal of our study is not the performance of the model on a specific benchmark therefore we do not consider it necessary to collect data from the three benchmark tests separately. In order to minimize the statistical bias or other unknown errors introduced by the data collection processes and to control the overall time consumption of the experiment, each run is executed ten times consecutively. We introduce a 5-minute interval after each run to take into account the problem of CPU performance degradation due to temperature increase when the NLP model performs inference tasks. In addition, the order of the experiments is randomly selected by using a script, available in the replication package [7], in order to minimize possible inter-experiment disturbance. The program we use is a customized random selection program developed by ourselves.

6 RESULTS

In this section, we present the results obtained from our experiment, where four different models, namely BERT, GPT-2, Distilled BERT, and Distilled GPT-2, were subjected to 10 runs each. For each run, three benchmarks were employed: SST-2, STSB, and MNLI. The metrics collected during these runs include CPU Utilization, CPU Power, average CPU, average memory, total energy, and total inference time.

6.1 Descriptive Statistics

Before our statistical analysis, we carefully check the correctness of the collected measures, *e.g.,* in case the energy profiler produced values not matching the technical specifications of the machine. We identified two problematic runs, which involved re-running those runs to guarantee that each model had precisely 10 valid runs.

We initiate our analysis by exploring the data via basic descriptive statistics and visualization. Table 2 presents the average values of all our measured metrics across all models. The analysis reveals intriguing differences between the models. For both GPT-2 and BERT, the distilled versions exhibited lower CPU and memory usage, reduced energy consumption, and faster inference times.

Table 2: *BERT, DistilBERT, GPT2,* and *DistilGPT2* models' performance in the experiment. *CPU* represents average CPU usage, *Memory* denotes average memory usage, *Energy* is the total energy consumption, and *Inference Time* indicates total runtime for inference.

Model	CPU (%)	Memory (%)	Inference Time (s)	Energy (J)
BERT	36.7	5.68	888.22	28902.170
Distilled-BERT	26.4	5.55	674.77	16234.187
GPT-2	49.6	3.95	797.90	31228.713
Distilled-GPT-2	48.3	3.17	557.61	22464.590

Specifically, by referring to Table 2 and the box plots in Figure 2 we make the following observations:

- **Energy Consumption**: the distilled models, *i.e.,* Distilled-BERT and Distilled-GPT-2, consume 16,234.187J and 22,464.590J, respectively. Distilled-BERT exhibits a remarkable 43.96% reduction in energy consumption compared to its non-distilled counterpart, demonstrating its higher energy efficiency. Distilled-GPT-2 exhibits a notable 28.1% decrease in energy consumption compared to its non-distilled counterpart, GPT-2.
- **Inference Time**: distilled models tend to have shorter inference time. Distilled-BERT takes an average of 674.770ms, which is faster than BERT's 891.072ms. Similarly, Distilled-GPT-2 took 557.614ms, a noticeable reduction from GPT-2's 798.988ms.
- **CPU Utilization**: Distilled-BERT shows a sensible reduction in CPU utilization (0.264%) when compared to BERT (0.367%). In the case of GPT models, Distilled-GPT-2 (0.483%) exhibits a sensible lower CPU utilization than GPT-2 (0.496%).

  Memory Utilization: for BERT models, the difference in memory utilization is marginal, with Distilled-BERT using 5.546% and BERT consuming 5.686%. However, for GPT models, Distilled-GPT-2 tends to use considerably less memory (3.167%) compared to GPT-2 (3.951%).

In Figure 2 we can also observe that the data displays an asymmetrical pattern in some cases, with a few outliers present. These deviations are indicative of a distribution that deviates from perfect normality. This aspect will be the subject of formal investigation in the subsequent phases of our analysis. Additionally, the boxplots reinforce our initial assumption that distilled models consume less energy and exhibit enhanced performance. Testing for normality is a crucial step before hypothesis testing. Depending on whether the collected data is assumed to be normally distributed or not, either paired t-tests or Wilcoxon Signed Rank tests are employed. Figure 3 presents density plots for each metric, considering different models. Within each subplot, one metric is plotted for the four models. These plots illustrate that most data adheres to a bell curve, indicating a tendency towards normal distribution. To delve deeper into normality assumptions, Q-Q plots were generated (see our replication package [7]). While most figures closely follow the diagonal line (highlighted in red), it is important to note that some exhibit slight deviations. Due to our relatively modest sample size, we could expect that the Q-Q plots would display more dispersion and variability. In such instances, discerning a clear linear pattern can be challenging, and deviations from linearity might occur due to sampling variability. Consequently, in the subsequent analysis, we employ quantile-quantile (Q-Q) plots and the Shapiro-Wilk test to rigorously assess if our data adheres to a normal distribution.

Highlights – The distilled versions of the measured NLP models tend to exhibit better energy efficiency, faster inference times, and in some cases, reduced CPU utilization and memory usage. The collected data suggests that Distilled-BERT is especially more energy-efficient compared to BERT, while Distilled-GPT-2 offers significant improvements in inference time and memory usage over its original counterpart, GPT-2.

6.2 Hypothesis Testing

In our study, the process of hypothesis testing is conducted in two primary stages. Initially, we apply an Analysis of Variance (ANOVA) to discern if there are any significant differences in our metrics across the various models. ANOVA is a statistical method used to test differences between two or more means, and it is particularly useful when comparing multiple groups simultaneously [14]. This approach provides a holistic view and tests differences across multiple groups. Furthermore, to ensure the validity of our ANOVA results, we conduct Mauchly's Test for Sphericity. This test is crucial as it checks the assumption of sphericity in repeated measures ANOVA, which is essential for the accuracy of the F-ratio in the ANOVA test. Sphericity refers to the condition where the variances of the differences between all combinations of related groups are equal [15]. Following the ANOVA, we apply a t-test for pairwise comparisons to understand the differences between the specific models. The t-test is a statistical test that allows us to compare the means of two groups and determine if they are significantly different from each other. This step is particularly important

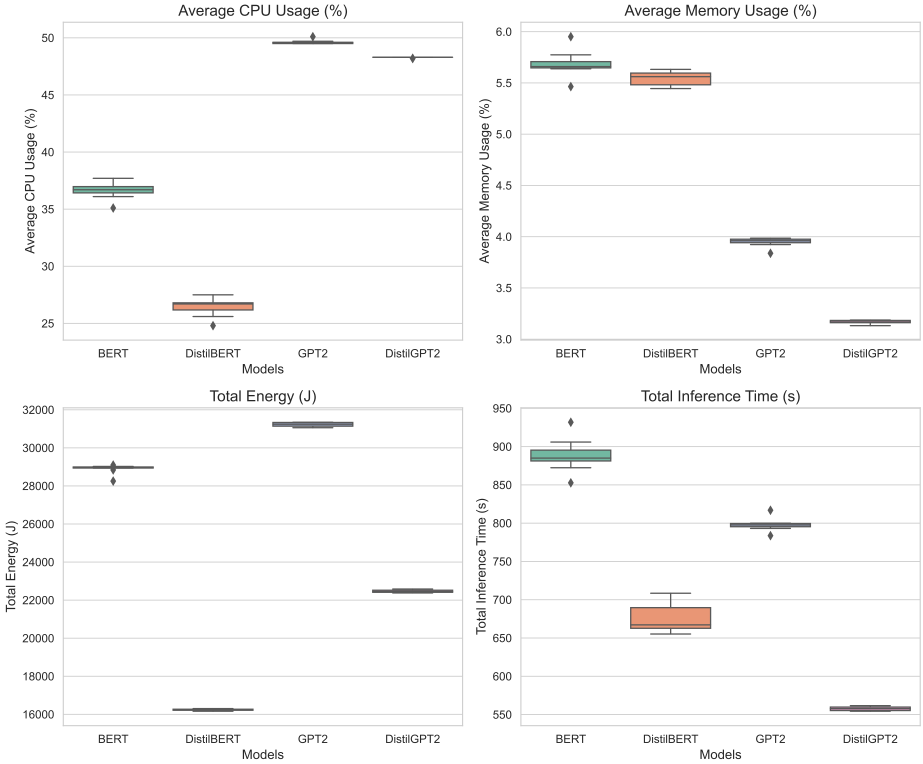

Figure 2: CPU usage, memory usage, energy consumption, and total inference time across the four language models.

to pinpoint where the significant differences lie between individual models [16].

6.2.1 CPU Utilization (RQ1). Our analysis shows a profound impact of the model type on average CPU utilization, with an $F(3, 27) = 4358.705$ and $p < .001$. The observed generalized eta squared (ges) was 0.9969644, indicating a substantial effect size, thus suggesting that the model type significantly influences the CPU utilization. This is particularly noteworthy because it suggests that different NLP models have markedly different computational demands. Sphericity, an important assumption in our ANOVA test, was not met ($W = 0.1398375$, $p = 0.01017544$). This prompted us to apply the Greenhouse-Geisser correction ($\epsilon = 0.69218$), which confirmed the results remained statistically significant at $p < .001$.

6.2.2 Memory Usage (RQ1). The analysis shows a statistically significant difference when analyzing memory usage, $F(3, 27) = 2666.646$, $p < .001$. The substantial effect size (ges = 0.9956913) demonstrates that the type of NLP model in use can have a significant impact on the memory resources consumed. The test indicated that the assumption of sphericity was violated ($W = 0.1418206$, $p = 0.01063451$). Even after using the Greenhouse-Geisser correction ($\epsilon = 0.4912353$), the findings remained robust and statistically significant, $p < .001$.

6.2.3 Total Inference Time (RQ1). The results pointed towards a significant difference among the models in terms of total inference time, $F(3, 27) = 1335.418$, $p < .001$. The robust effect size (ges = 0.9899065) indicates that model choice can substantially affect the time taken for inferences, an essential factor in real-time applications. The sphericity assumption is satisfactorily met ($W = 0.284992$, $p = 0.08669588$).

6.2.4 Energy Consumption (RQ2). The analysis shows a statistically significant difference when analyzing the energy consumed by the models, with $F(3, 27) = 64777.89$, $p < .001$. The extremely large effect size (ges = 0.9998128) underscores the importance of selecting the right NLP model, especially in scenarios where energy efficiency is paramount. Our data met the assumption of sphericity, as confirmed by $W = 0.3804188$, $p = 0.1915562$.

Highlights – The distilled models exhibit differences in performance metrics when compared to their original counterparts. The extent and direction of these differences vary depending on the specific metric being considered. Our results indicate that while GPT2 has a slightly higher CPU utilization on average, distilledGPT2 consumes more power. This suggests that while distillation might optimize certain aspects of a model, it could lead to higher power consumption in certain scenarios. This highlights the importance of carefully considering trade-offs when employing model distillation techniques.

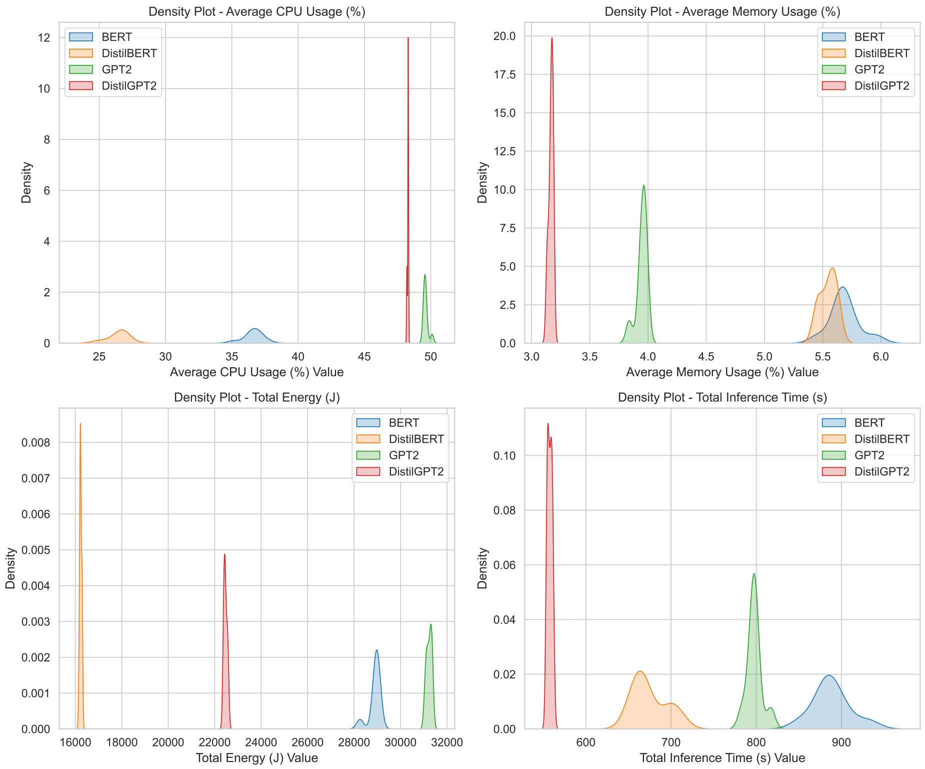

Figure 3: Density plots for CPU usage, memory usage, energy consumption, and total inference time across the four language models.

7 DISCUSSION

In this section we summarize our results for each research question of the study, and then we expand on our interpretation by taking the perspective of software developers working in the domain of the NLP models.

Energy Consumption – Our primary research question (RQ1) focusses on the energy consumption of distilled models compared to standard NLP models. The results of our hypothesis testing demonstrate a significant difference between the two, with distilled models consuming less energy. This finding is crucial in scenarios where energy efficiency is a priority. However, it is important to note that the choice of model, benchmark tasks, and the computing environment can influence these results. Further research is needed to assess the broader applicability of these findings.

Performance – The second research question (RQ2) concerns the performance of distilled models and standard ones. Both Distilled BERT and Distilled GPT-2 have significantly lower average inference time compared to their original counterparts, BERT and GPT-2. This indicates that Knowledge Distillation indeed has a positive impact on model performance, making them more competitive for real-time applications. For memory usage, all models in the experiment performed well. The analysis reveals that the choice of the NLP model significantly impacts memory resources consumed. Distilled-GPT-2, in particular, consumes considerably less memory than the non-distilled GPT-2, while the memory difference between Distilled-BERT and BERT is marginal. In terms of CPU usage, both distilled versions of models have a slightly lower number than their counterparts. Additionally, different models influence CPU usage. To summarize, the distillation model performs better compared to the standard model, however, different experimental environments may have an impact on the results.

Practical insights for developers – Our results offer practical insights for developers in the NLP field. Distilled models are advantageous in scenarios where energy efficiency is paramount, such as in mobile applications or devices with limited processing power. However, for tasks that demand high accuracy or complex language processing, the original, non-distilled models might be more suitable. This decision-making is critical in balancing the trade-offs between efficiency and computational requirements. Our study highlights the critical impact of model selection in the realm of NLP, particularly in terms of energy efficiency and performance. The findings from this research not only contribute to the technical understanding of distilled NLP models but also offer a perspective on their practical applications and environmental implications. As the field of AI continues to evolve, these insights will be invaluable in guiding the development of sustainable and efficient AI technologies.

Other remarks – It is important to note that the choice of benchmarks from the GLUE dataset played a pivotal role in our analysis.

These benchmarks, encompassing diverse tasks, provided a comprehensive ground for testing. However, it is crucial to consider that certain tasks in these benchmarks, which demand intensive memory or computational power, might inherently favor the distilled models due to their optimized architecture. The extent to which our results can be generalized may depend on the nature of the tasks and the specific requirements of different applications. Finally, we acknowledge certain limitations in our study, such as the specific scope of benchmarks and the range of models examined. Future research should expand to include a broader spectrum of NLP tasks, encompassing real-world applications, and should also consider the long-term environmental impacts of these models. Such studies would contribute significantly to the field of sustainable AI development, providing deeper insights into the ecological footprint of NLP technologies.

8 THREATS TO VALIDITY

The following section discusses the threats that could potentially have an impact on this experiment from four different perspectives. The threats discussed below are categorized according to the taxonomy proposed by Cook and Campbell [17].

8.1 Internal Validity

After the complete execution of the BERT and GPT-2 related experiments, similar data errors were found in one run of each of the two experiments. In one run of the BERT as well as the GPT-2 experiments, the data collected at one second showed that the energy was a huge negative number. However, we are still unable to know the reason for this error. This resulted in the total energy for both runs being a large negative number. This means that for both experiments we have only ten rounds of valid run data. To ensure the validity of the data used for the analysis we performed two new runs to ensure that all models had ten rounds of valid data. These two new tests were done separately from the original tests completed so that may have affected the results.

During the runtime, we downloaded the models before the experiments were executed because the first use of the models would take some time to download the models, thus affecting the inference time, CPU usage, and memory usage. Further, the parameters of all models were not changed during the experiment to ensure that the behavior of the models remained unchanged. Another threat to changes in CPU performance is CPU temperature. Excessive CPU temperatures can lead to performance degradation. To mitigate this we introduce a cooling time between each run.

8.2 External Validity

There are three main threats to the external validity of this study: the benchmarks used, the chosen model, and the version of the Python module. For this study, we chose some of the tasks from the representative GLUE dataset, which is widely used for tasks like understanding the nuances and subtleties of human language and therefore fits well with the BERT family of models. However, GLUE is not a very suitable but workable benchmark for GPT-2 because the GPT series of models is mainly used for content generation tasks. Although we focused on the effect of distillation techniques on energy consumption and controlled for variables, we cannot rule out the possibility that the opposite result could be obtained when performing the NLG task.

For model selection both BERT and GPT2 are representative models of their respective domains (NLU and NLG) thus mitigating the threat to generalisability. It is noteworthy that the results of the experiments may not be applicable to other variants of the two models except for the distilled version as this is beyond the scope of our study but is worth investigating in depth in the future.

In addition, different Python libraries may have different optimizations for the model even for the initial model so we ensured that all modules in the four implementation files were kept the same version.

8.3 Construct Validity

Due to the simplicity of our experimental theory, i.e., just a comparison of different versions of the NLP model, the structure of our construction is not complex and can be easily translated into correspondence measures. Furthermore, the theory related to the performance of the distillation model has also been proven by Sanh et al. [2] and the models we used are from the same sources as in the original paper. These mitigate the threat threats to the construct validity. Moreover, in order to mitigate Mono-method bias we use multiple measures such as CPU usage and energy consumption. Some of the measures are redundant such as CPU utilization because the results suggest that models with high CPU utilization also tend to consume more energy.

8.4 Conclusion Validity

From the results of Section 6, all the null hypotheses can be rejected which also shows that the experimental results are statistically significant. In addition, the sample size used for hypothesis testing is reasonable for a t-test but it also cannot produce a very robust result and even cause a violation of test assumptions since the sample size only just reached the lower limit of the test used. For example, the distribution of DistilGPT2 average CPU usage data is affected. To further mitigate the threat to the conclusion validity, the prerequisites of the corresponding hypothesis tests were also examined to ensure that the assumptions for conducting the hypothesis tests were met, e.g., we have tested the distribution of the data. An obvious threat to the conclusion's validity during the analysis of our data was the low statistical power presented by Levene's Test which could also be caused by the small sample size mentioned above. Moreover, the energy data we obtained was collected through PowerJoular which means that if its measurement is not accurate then it will pose a big challenge to the results of the experiment. The negative values in our experimental data suggest that there may be some problems with the PowerJoular.

9 CONCLUSIONS

Our comparative analysis brought to light the pronounced differences between distilled and non-distilled NLP models across various metrics: energy consumption, inference time, CPU utilization, and memory usage. Distilled models, notably Distilled-BERT and Distilled-GPT-2, generally showcased superior energy efficiency and swifter inference. Particularly, Distilled-BERT emerged as notably more energy-efficient compared to its original counterpart, BERT. Conversely, Distilled-GPT-2 demonstrated marked improvements over GPT-2 in both inference time and memory usage. The results from our hypothesis testing further corroborated these observations, emphasizing the paramount influence of the choice of NLP model on both performance and energy consumption metrics.

With the insights garnered from this study, there lies a plethora of possibilities for further investigation. An immediate and apparent extension would be to compare a wider array of distilled models, expanding the horizons to incorporate other cutting-edge architectures and their distilled versions. A pivotal extension would be to assess the performance of models across diverse benchmarks, ascertaining if the observed benefits of distillation remain consistent across various challenges and tasks. Teams could also explore the performance trade-offs inherent in model distillation techniques in different deployment environments, especially on edge devices where computational constraints are accentuated. Furthermore, with multi-modal data becoming increasingly ubiquitous, a study into whether the advantages of distillation transcend data types would be invaluable. Lastly, in our era that is progressively emphasizing environmental sustainability, which is a deep dive into the ecological footprint of these models might reveal if distilled models truly serve as a more eco-friendly alternative in large-scale deployments.

REFERENCES

[1] Y. Ding, J. Ma, and X. Luo, "Applications of natural language processing in construction," *Automation in Construction*, vol. 136, p. 104169, 2022.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[3] A. Al Karkouri, F. Ghanimi, and S. Bourekkadi, "Unveiling the environmental implications of automatic text generation and the role of detection systems," in *E3S Web of Conferences*, vol. 412. EDP Sciences, 2023, p. 01102.

[4] Q. Cao, A. Balasubramanian, and N. Balasubramanian, "Towards accurate and reliable energy measurement of nlp models," *arXiv preprint arXiv:2010.05248*, 2020.

[5] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446

[7] "Replication package of this study," https://anonymous.4open.science/r/cain-2024-distilled-models-energy-perf-rep-pkg-38BF, online.

[8] Q. Cao, Y. K. Lal, H. Trivedi, A. Balasubramanian, and N. Balasubramanian, "Irene: Interpretable energy prediction for transformers," *arXiv preprint arXiv:2106.01199*, 2021.

[9] A. H. Zadeh, I. Edo, O. M. Awad, and A. Moshovos, "Gobo: Quantizing attention-based nlp models for low latency and energy efficient inference," in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2020, pp. 811–824.

[10] J. Choi, J. Park, K. Kyung, N. S. Kim, and J. H. Ahn, "Unleashing the potential of pim: Accelerating large batched inference of transformer-based generative models," *IEEE Computer Architecture Letters*, 2023.

[11] V. R. Basili, G. Caldiera, and D. H. Rombach, *The Goal Question Metric Approach*. John Wiley & Sons, 1994, vol. I.

[12] "Experiment runner," https://github.com/S2-group/experiment-runner, online.

[13] A. Noureddine, "Powerjoular and joularjx: Multi-platform software power monitoring tools," in *2022 18th International Conference on Intelligent Environments (IE)*. IEEE, 2022, pp. 1–4.

[14] A. Field, *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[15] J. F. Hair, "Multivariate data analysis," 2009.

[16] N. Salkind and B. Frey, *Statistics for People Who (Think They) Hate Statistics*. SAGE Publications, 2019. [Online]. Available: https://books.google.nl/books?id=GgeeDwAAQBAJ

[17] D. T. Campbell and T. D. Cook, "Quasi-experimentation," *Chicago, IL: Rand McNally*, 1979.