University of Passau
Faculty of Computer Science and Mathematics

Chair of Distributed Multimedia Information Systems - DIMIS
Prof. Dr. Harald Kosch

Master's Thesis
in
Artificial Intelligence Engineering

Exploring Large Language Models for Argument Mining Tasks

Yasser Otiefy

Date: 2024-05-25

Examiners: Prof. Dr. Harald Kosch
Prof. Dr. Michael Granitzer

Advisor: Dr. Alaa Alhamzeh

Otiefy, Yasser
yasser.otiefy@uni-passau.de
13 Fernfail Court, 88 Short Heath Road
B236JT, Birmingham, UK

ERKLÄRUNG

Ich erkläre, dass ich die vorliegende Arbeit mit dem Titel „Exploring Large Language Models for Argument Mining Tasks“ selbständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe und dass alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, als solche gekennzeichnet sind.

Mit der aktuell geltenden Fassung der Satzung der Universität Passau zur Sicherung guter wissenschaftlicher Praxis und für den Umgang mit wissenschaftlichem Fehlverhalten vom 31. Juli 2008 (vABlUP Seite 283) bin ich vertraut.

Ich erkläre mich einverstanden mit einer Überprüfung der Arbeit unter Zuhilfenahme von Dienstleistungen Dritter (z.B. Anti-Plagiatssoftware) zur Gewährleistung der einwandfreien Kennzeichnung übernommener Ausführungen ohne Verletzung geistigen Eigentums an einem von anderen geschaffenen urheberrechtlich geschützten Werk oder von anderen stammenden wesentlichen wissenschaftlichen Erkenntnissen, Hypothesen, Lehren oder Forschungsansätzen.

.....
(Name, Vorname)

Translation of German text (notice: Only the German text is legally binding)

I hereby confirm that I have composed the present scientific work entitled “Exploring Large Language Models for Argument Mining Tasks” independently without anybody else’s assistance and utilising no sources or resources other than those specified. I certify that any content adopted literally or in substance has been properly identified and attributed.

I have familiarised myself with the University of Passau’s most recent Guidelines for Good Scientific Practice and Scientific Misconduct Ramifications of 31 July 2008 (vABlUP Seite 283).

I declare my consent to the use of third-party services (e.g. anti-plagiarism software) for the examination of my work to verify the absence of impermissible representation of adopted content without adequate attribution, which would violate the intellectual property rights of others by claiming ownership of somebody else’s work, scientific findings, hypotheses, teachings or research approaches.

Examiners contact:

Prof. Dr. Harald Kosch
Chair of Distributed Multimedia Information Systems - DIMIS
Universität Passau, Innstraße 41, 94032 Passau, Germany
Email: Harald.Kosch@uni-passau.de
Web: https://www.fim.uni-passau.de/en/distributed-information-systems/

Prof. Dr. Michael Granitzer
Chair for Data Science
Universität Passau, Innstraße 41, 94032 Passau, Germany
Email: Michael.Granitzer@uni-passau.de
Web: https://www.fim.uni-passau.de/en/data-science/team/

Advisor contact:

Dr. Alaa Alhamzeh
Chair of Distributed Multimedia Information Systems - DIMIS
Universität Passau, Innstraße 41, 94032 Passau, Germany
Email: Alaa.Alhamzeh@uni-passau.de
Web: https://www.fim.uni-passau.de/en

Abstract

The examination of argumentative content within Earnings Conference Calls (ECCs) represents a crucial domain for financial analytics, offering significant insights to investors and analysts. In this dynamic context, ECCs provide a rich tapestry of narrative strategies and argumentation that influence market perceptions and decision-making processes. Despite its potential impact, the automatic identification of relationships between argument components in ECCs remains underexplored in academic research. This thesis addresses this gap by leveraging advanced natural language processing techniques to automate relation identification, thereby enhancing the interpretability and utility of financial discourse.

The primary objective of this research is to empirically examine and analyze the effectiveness of various open-source models, including general-purpose models, debate-fine-tuned models, and financial-fine-tuned models, in identifying argumentative relations in financial texts. Additionally, our study evaluated the capabilities of the state-of-the-art Generative Pre-trained Transformer, GPT-4, using zero-shot learning on a specially curated financial argumentation dataset (FinArg). Alongside GPT-4, this research also explored the performance of the LLaMA-3 Instruct models in both 8B and 70B configurations using a 1-shot learning approach, providing insights into their efficiency in classifying argumentative relationships within a constrained data setting.

Our methodology involves a rigorous experimental setup where each model is evaluated based on its ability to detect and classify argumentative relations within ECC transcripts. The models' performances are compared using a range of metrics, with a particular focus on the F1-score, which combines precision and recall in a single metric. This comparative analysis not only highlights the relative strengths and weaknesses of each model type but also provides insights into the scalability and practical utility of advanced language models in financial analytics.

Our findings reveal a detailed landscape of model performance. While smaller, open-source models fine-tuned on domain-specific data demonstrated notable efficacy, approaching the performance of much larger general-purpose models, they underscored the value of local semantic enrichment in embeddings. Surprisingly, the GPT-4 model exhibited superior performance across all metrics, achieving an F1-score of 0.81 in zero-shot learning settings. Meanwhile, the LLaMA-3 model showed promising results, with the 8B and 70B variants performing distinctly in terms of precision and recall trade-offs, further illustrating the potential of tailored model configurations in specific argument mining tasks, The learning is done using both prompt engineering with zero and 1-shot learning, and fine-tuning approaches.

This thesis contributes to the literature by elaborating the experimental setup, data preprocessing, and the performance analysis of these models. It also discusses the implications of these findings for the deployment of language models in financial applications, suggesting pathways for future research to further poster the capabilities of NLP systems in this domain.

In conclusion, this research not only advances our understanding of the application of large language models in financial argument mining but also demonstrates the transformative potential of zero-shot and few-shot learning in interpreting complex financial narratives. The insights gained herein could serve as a foundation for future innovations that might integrate these technologies into real-time financial decision-making workflow.

Contents

1 Introduction

Earnings Conference Calls (ECCs) serve as one of the most vital platforms for corporate communication, where executives discuss past performance and future forecasts by answering analysts' direct questions. These presentations of financial data are about more than just numbers; they aim to present a story and argue a point, influencing how investors see the company's future [1].

The detailed work of Alhamzeh et al. [2] showcases the potential of argument mining in financial contexts, specifically during ECCs. They describe the process of extracting and evaluating arguments to determine their strength, based on what company executives say. Their innovative approach to quantifying the quality of these arguments [3] reveals the complex interplay between different components of an argument, such as premises and claims, and how these interactions influence the overall persuasiveness of the narrative. This analytical approach confirms that the fundamental structures of argumentation significantly affect investor sentiment and decision-making.

Despite its significant potential, the area of argument relation detection and classification remains largely unexplored, especially within the context of financial discussions. The tasks involved are complex and subtle, particularly for natural language understanding (NLU) systems [4, 5]. Identifying the relationships between components of an argument is crucial for truly understanding the fabric of the financial narrative. It underscores the capabilities and current limitations of computational approaches to discerning subtle hints that indicate whether an argument supports or opposes a particular stance.

The rise of Financial Natural Language Processing (FinNLP) represents a coming together of computational linguistics and financial analysis, driven by the unique challenges that are imposed by the financial domain's complexities [6]. This interdisciplinary field aims to develop automated tools that can navigate the specialized jargon and complex argumentative patterns common in financial texts. The introduction of FinNLP has led to various collaborative efforts, such as workshops and shared tasks, focused on promoting methodological advancements and creative solutions in the field [7, 8, 9].

1.1 Problem Statement

The increase of unstructured financial documents, such as regulatory filings and earnings calls, presents both significant challenges and opportunities for Natural Language Processing (NLP). The precise detection, classification, and analysis of argumentative structures within these documents are crucial for applications like sentiment analysis, fraud detection, and automated compliance monitoring. Yet, the complex and specific language used in finance often challenges conventional NLP methods.

Large Language Models (LLMs) have opened a new era in NLP, enhancing the ability to understand and generate text that closely mimics human language. These models hold great potential

to transform how we extract arguments from financial documents, though several issues remain unresolved:

1. **Model Adaptability:** Despite their impressive capabilities, the adaptation of LLMs to the unique requirements of the financial sector without compromising precision or efficiency is still an open question. Fine-tuning these models for domain-specific tasks can be resource-intensive, requiring extensive datasets and domain expertise, which limits their accessibility and scalability.

2. **Prompt Engineering Complexity:** The emerging strategy of prompt engineering, particularly with models like GPT-4, offers a shortcut to extensive model training. However, crafting effective prompts that accurately guide the model in detecting and classifying argument relations is more of an art than a science. The lack of systematic methods for creating prompts poses significant challenges, especially in precision-critical fields like finance.

3. **Comparative Effectiveness:** There is a lack of comprehensive studies comparing traditional fine-tuning methods with prompt-based approaches in financial argument mining. Understanding the trade-offs between these methods in terms of accuracy, resource efficiency, and scalability is crucial.

4. **Practical Applications:** The potential impacts of advancements in LLMs on future NLP applications in finance are still not fully explored. It's essential to determine how to integrate these models into practical financial analysis tools and identify the challenges involved.

This study aims to bridge these gaps by evaluating the effectiveness of fine-tuned LLMs and prompt-engineered GPT-4 [10] in detecting argument relations within financial texts. By doing so, it seeks to contribute to the development of more accurate, efficient, and scalable NLP tools for financial analysis and to pave the way for innovative applications that can leverage the vast amounts of data generated in the financial sector. Prompt engineering is the process of solving problems via prompting LLMs and in our case we trying to solve classification tasks by embedding our examples in pre-designed prompt and send to LLMs to return the expected class.

1.2 Research Questions

The exploration of argument mining, especially in the realm of financial texts, raises several intriguing questions about how Large Language Models (LLMs) can be applied and optimized. This study is driven by a desire to delve into the complexities of detecting argument relationships using sophisticated computational models and to discover innovative strategies that could boost their effectiveness. With this context in mind, the following research questions (RQs) have been developed:

Effectiveness of LLM Fine-tuning in Argument Relation Detection **RQ1:** How effective are various LLMs, such as BERT, RoBERTa, and FinBERT, when fine-tuned for the specific task of detecting argument relations in financial texts?

This question aims to explore how well different LLMs adapt and perform when they are tailored with domain-specific data. The focus is on assessing the models' accuracy, precision, and recall in recognizing and categorizing argumentative structures within financial documents.

Impact of Prompt Engineering on Relation Detection Performance using GPT-4 **RQ2:** How does prompt engineering with GPT-4 influence the quality of argument relation detection in unstructured financial texts?

Prompt engineering is a crucial approach that utilizes the generative capabilities of models like GPT-4 for specific NLP tasks without the need for extensive fine-tuning. This question seeks to evaluate the effectiveness of various prompting strategies in improving GPT-4's ability to detect and analyze argument relations. It explores the balance between the complexity of the prompts, the responsiveness of the model, and the accuracy of the outputs generated.

The research questions formulated for this study aim to deepen our understanding of how Large Language Models (LLMs) can be applied in the domain of argument mining, specifically within the financial sector. By exploring these questions, this research endeavors to enrich the evolving field of Natural Language Processing (NLP), offering valuable insights that could guide future research, influence model development, and enhance practical applications in finance and related fields.

1.3 Contributions

This thesis makes several significant contributions to the burgeoning field of Financial Natural Language Processing (FinNLP):

- **Pioneering an Intensive Examination:** To the best of our knowledge, this study is the first comprehensive examination of the capabilities of recent LLMs specifically in the task of argument relation identification within the financial domain. This marks an important advancement in applying artificial intelligence to finance.
- **Bridging Computational and Financial Analysis:** Through detailed empirical research, this work bridges the gap between computational linguistics and financial analysis, providing new insights into the automated processing of financial arguments.
- **Fostering Methodological Advancements:** By evaluating a diverse range of LLMs using the FinArg dataset, this thesis not only assesses their current capabilities but also establishes a foundation for future methodological innovations in the analysis of financial argumentation.

Central to our study is the investigation of Large Language Models like GPT-3 and GPT-4, exploring their potential to revolutionize the analysis of financial argumentation. These models, built on deep learning technologies, introduce a novel approach to understanding and interpreting the intricate narratives embedded in financial texts. Our research evaluates the zero-shot learning abilities of these models across a spectrum from general-purpose to specialized models fine-tuned for debates and financial contexts.

Our methodology involves a detailed, comprehensive analysis using the FinArg dataset—a meticulously curated collection of financial narratives designed to test how effectively LLMs can identify argument relations. We aim to unravel the complexities of financial argumentation by comparing the performance of general-purpose LLMs with those that have been fine-tuned for specific contexts such as debates and finance. Financial discussions are characterized by their specialized vocabulary and complex structural forms, making the detection of argument relations particularly challenging. The distinctive features of financial language, which include dense jargon, require models to not only grasp general language but also understand the specific contexts and meanings inherent in financial narratives. This task is further complicated by the dynamic nature of financial markets, where new terms and concepts frequently emerge, and sentiments can shift rapidly in response to global events and economic indicators [11].

The emergence of Large Language Models (LLMs) such as GPT-3 and GPT-4 has introduced a new era in NLP, providing powerful tools for text generation and understanding. These models, developed from vast amounts of diverse textual data, excel at producing clear and contextually appropriate text. Their application to financial texts is particularly promising, suggesting the potential for automated systems that can fully grasp complex financial dialogues and uncover valuable insights [12, 13]. However, their effectiveness in the financial sector depends on their ability to adapt to the specialized language and unique argument structures typical of financial discussions.

Our study categorizes LLMs into three groups based on their training and fine-tuning: general-purpose models, debate-focused models, and finance-focused models. This classification allows us to evaluate how different training backgrounds and fine-tuning approaches influence the models' ability to identify argument relations within financial texts. We aim to determine which methods are most effective for understanding detailed financial arguments and which aspects of model training are most critical for success in this area [14, 15].

We utilize the FinArg dataset, specifically designed to test how well LLMs perform on financial argumentation tasks. This dataset includes a wide array of financial narratives, each with its own set of argumentative structures and linkages. Our approach tests the ability of LLMs to interpret and classify these argumentative relationships without specialized training in finance. This examination helps us gauge the inherent capabilities of LLMs and their potential usefulness for analyzing financial arguments directly out of the box.

The findings from our study have significant implications for financial analysis, suggesting new automated methods for detecting and understanding arguments in financial texts. By leveraging LLMs, financial analysts, investors, and regulators could gain deeper insights into company narratives and market sentiments, potentially transforming the practice of financial analysis. Additionally, our research identifies where LLMs excel and where they fall short, guiding future efforts to enhance their applicability in finance.

In conclusion, as we wrap up our investigation into using LLMs for identifying argument relations in finance, we also consider future research directions. The field of NLP is rapidly evolving, and financial dialogues are becoming increasingly complex, offering substantial opportunities for advancement in FinNLP. Future studies could focus on developing more sophisticated models specifically for finance, experimenting with integrating numerical data and visual elements with textual information and finding more effective ways to interpret and present model outputs in a manner that is accessible and beneficial to financial professionals.

1.4 Thesis Outline

In Section 2 we explore the definition and background of the topics discussed in this thesis. In Section 3, we navigate the state-of-the-art dedicated to LLMs in argument mining tasks. We overview our data, and methodology in Section 4. Afterwards, we exhibit the evaluation results in Section 5. We further discuss and analyze our findings in Section 6. Finally, we conclude our work and open future perspectives in Section 7.

2 Background

This chapter provides an overview of the theoretical and methodological framework highlighting the study of argument mining, particularly in the context of financial texts. It goes through the evolution of argument theory, various argument mining methods, and the specific challenges and opportunities in the financial domain.

2.1 Argument Theory & Tasks

Argument theory discovers the components and structures that makeup arguments in natural language. It focuses on the identification of claims, evidence, and their relationships. In argument mining, key tasks include argument identification, classification, and structure analysis [16].

Argument lays the foundation of argument mining by providing a framework to identify, structure, and evaluate arguments within a text. [17] offer a comprehensive survey of argument mining, detailing the layered tasks involved in processing and understanding arguments computationally. Central to argument mining is the ability to not only identify argumentative components but also to assess their quality and the relationships between them. Figure 2.1 shows argument mining different tasks and the proper sequence of the pipeline.

Segmentation and Argument Detection The process begins with segmentation, where algorithms are tasked with extracting relevant sections or fragments from larger texts. This initial step is crucial as it determines the subsequent input for argument detection. Argument detection further inspects these segments to determine whether they contain argumentative discourse, distinguishing between argumentative and non-argumentative text.

Example: *A segment labeled "Azure use cases and storage" would be extracted, and within this segment, argument detection might flag the statement "In fact as I said, a lot of the Azure use cases that require high performance, require high-performance storage. And so things like artificial intelligence, machine learning, the HPC use cases are all things that we expect to drive additional Azure consumption as well as performance." as argumentative, based on its explanatory and predictive nature.*

Intrinsic and Contextual Analysis Once a segment is identified as argumentative, deep analysis evaluates it in isolation to classify its nature. For example, determining if a statement serves as evidence for an argument. In contrast, contextual analysis examines the segment within its surrounding text to categorize its role, such as judging whether it acts as a premise or claim within the broader argument structure.

Example: *In isolation, the deep analysis acknowledges the need for high-performance storage. In context, it is understood that this need supports the strategic claim about driving additional Azure consumption.*

Relational Properties Assessment Identifying relational properties involves analyzing the connections between different argument components. For instance, assessing the relationship between premise X and claim Y involves examining whether the evidence presented supports or contradicts the claim, which is fundamental for understanding the argument's logic and coherence.

Example: *The model might evaluate if the argument "So we continue to build for high AFN, or Amazon Fulfillment Network demand driven by retail sales as well as FBA, or fulfilled by Amazon sales." correctly aligns the premise with the claim and categorizes the relationship as supportive.*

Argumentative Quality Assessment The final layer of argument mining is the argumentative quality assessment. This task evaluates the strength, validity, and persuasive power of the argument as a whole. It may involve assessing how effective the argument is, such as whether an argument formulated from a premise and a claim is convincing or if it is credible based on expert opinions.

Example: *An argument stating "Our goal really is to kick-start an ecosystem around AWS services that helps us drive the adoption and usage of our infrastructure." would be assessed for its strength based on the strategic vision and the potential for ecosystem development.*

These tasks represent a structured approach to breaking down and grasping arguments. It is critical for applications ranging from legal analysis to editorial reviews. The field of argument mining, particularly in the financial domain, leverages these tasks to extract meaningful insights from complex, argument-rich texts, such as financial reports, investment theses, and economic forecasts. By automating the identification and evaluation of arguments, argument mining helps transform unstructured financial discourse into structured data, ready for further analysis and decision-making processes.

all the above examples came from Alaa et al. [18] dataset.

2.2 Argument Mining Methods

Argument mining has evolved through different methodical methods, from rule-based heuristic approaches to the cutting-edge applications of Large Language Models (LLMs). This section gives a summary of these methods, including notable examples that have shown their utility in argument mining.

Natural Language Processing As the field matured, NLP methods, particularly those involving syntactic and semantic analysis, became crucial in extracting and interpreting arguments from texts.

Example: The Stanford CoreNLP toolkit, with its suite of language analysis tools, has been widely used for tasks such as sentence splitting, part-of-speech tagging, and named entity recognition, which are fundamental for argument mining [19].

Machine Learning With the availability of annotated corpora, machine learning approaches, particularly supervised learning, began to automate the detection and classification of arguments.

Example: The SVM-based models were among the first machine learning approaches applied to argument mining, using features like n-grams and part-of-speech tags to classify sentences as argumentative or non-argumentative [20].

Deep Learning Deep learning has significantly impacted argument mining by enabling models to learn complex patterns in data. Neural networks, especially recurrent and convolutional neural networks, have been employed to capture the sequential nature of text for argument analysis.

Example: The use of LSTM networks for identifying argumentative relations shows how deep learning can learn contextual dependencies within arguments, beyond surface-level linguistic cues [21].

Finetune & Prompt Engineering of Large Language Models The introduction of LLMs such as BERT and GPT-3 has opened new boundaries in argument mining. Fine-tuning these models on domain-specific corpora has led to significant performance gains in argument classification tasks.

Example: BERT has been fine-tuned for legal argument mining, leading to models like Legal-BERT, which captures the complexities of legal language and improves the classification of argumentative roles in court case documents [22]. Meanwhile, the rise of GPT-3 introduced the concept of prompt engineering, where carefully curated prompts are used to guide the model to generate or classify arguments effectively.

Example: GPT-3's ability to generate coherent and contextually appropriate responses has been utilized for generating arguments and counterarguments in debates, showing the power of prompt engineering in argument generation [23].

Each of these methodologies has advanced the field of argument mining, contributing to the development of sophisticated tools that can navigate the complexity of human language and argumentation. The trajectory from heuristic approaches to the application of LLMs illustrates the rapid progress in technology and the increasing sophistication of techniques available to researchers and practitioners in the field.

Quantization Large language models (LLMs) face computational and memory challenges, making quantization techniques essential for their efficient deployment. [24] introduced LLM-QAT, a data-free quantization-aware training method, which enables effective quantization of LLMs down to 4-bits by leveraging model-generated data. Concurrently, [25] enhanced computational efficiency through post-training quantization (PTQ) techniques that focus on both weight and activation quantization. They propose innovative scaling and calibration methods to maintain performance with reduced precision. Moreover, [26] demonstrated the impact of post-training quantization on LLMs, identifying robustness across different settings and noting the importance of optimizing quantization parameters for maintaining model performance.

2.3 Evaluation Metrics

Evaluation metrics are crucial for assessing the performance and effectiveness of machine learning models, including those applied in argument mining. This section introduces several key metrics, including accuracy, precision, recall, F1-score, Cross-Entropy Loss, and K-fold cross-validation, providing a comprehensive framework for evaluating model performance.

Confusion Matrix The confusion matrix is a powerful tool for summarizing the performance of a classification algorithm. It provides a detailed breakdown of the model's predictions, allowing researchers to distinguish between the numbers of correct and incorrect predictions across different categories. For a binary classification problem, the matrix is structured into four quadrants, representing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

This structure facilitates a comprehensive analysis of the model's ability to classify each class accurately. Specifically, it allows for the visualization of the trade-offs between different types of errors, such as the model's tendency to falsely classify negative instances as positive (false positives) versus its ability to correctly identify positive instances (true positives).

In argument mining, it's really important to tell different types of arguments apart or spot if there's an argument in the text. That's where the confusion matrix comes in. It's like a must-have tool for checking how well our models are doing. It helps researchers make their models better by improving things like accuracy, recall, and precision. But it's not just about numbers – the confusion matrix can also show us where our model is doing great and where it's not so good. This helps us figure out how to make our models even better in the future.

Accuracy, Precision, Recall, and F1-Score Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. Precision (or positive predictive value) measures the proportion of true positive results in the dataset. Recall (or sensitivity) measures the proportion of actual positives correctly identified. The F1-score provides a balance between precision and recall, offering a single metric to assess model performance. The equations for these metrics are as follows:

- **Accuracy:** $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- **Precision:** $Precision = \frac{TP}{TP+FP}$
- **Recall:** $Recall = \frac{TP}{TP+FN}$
- **F1-Score:** $F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$

Where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Cross-Entropy Loss Cross-Entropy Loss, also known as log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-Entropy Loss increases as the predicted probability diverges from the actual label, making it an effective metric for evaluating the probability outputs of a classifier. The equation for binary classification is:

$$H(y,p) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i) \tag{2.1}$$

Where N is the number of observations, y_i is the actual outcome, and p_i is the model's predicted probability for the i-th observation.

K-Fold Cross-Validation K-Fold Cross-Validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. This approach provides a robust method for estimating the performance of a model on unseen data. The process can be summarized as follows:

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} \text{Model Evaluation Metric}_i \tag{2.2}$$

Where CV_k represents the average evaluation metric (such as accuracy) across all k folds.

Pearson Correlation Coefficient The Pearson correlation coefficient, denoted as r, is a measure used to evaluate the linear relationship between two continuous variables. Its value ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 suggests no linear correlation between the variables. This statistical measure is crucial in the analysis of results as it provides insights into the degree to which two variables vary together.

The application of the Pearson correlation coefficient in financial argument mining can be particularly revealing. For example, examining the correlation between the sentiment scores derived from argument mining techniques and actual market movements or financial performance indicators. Analyzing this way can confirm how well the arguments we find predict things and how important they are, giving us key insights into whether NLP models can catch important feelings related to finance [27].

By leveraging these metrics, researchers can obtain an understanding of their model's performance, balancing considerations of overall accuracy with the model's ability to identify relevant patterns and the robustness of its predictive capabilities across diverse datasets.

Hyperparameter Optimization Hyperparameter optimization plays a pivotal role in maximizing the performance of machine learning and deep learning models. This section discusses techniques for optimizing hyperparameters in the context of argument mining, including grid search, random search, and Bayesian optimization [28].

Experiment Tracking Experiment tracking is a crucial practice in machine learning and data science, particularly for projects involving multiple iterations of model training and evaluation. It involves recording and analyzing various aspects of machine learning experiments such as hyperparameters, performance metrics, model weights, and even datasets. The primary goal is to ensure reproducibility and to facilitate the comparison of different model versions or training runs.

Example: Weight & Biases (wandb) [29] is a popular experiment tracking tool that allows researchers and practitioners to log and monitor their machine-learning experiments in real time. It provides an organized interface to track progress, compare results, and share findings. In argument mining, wandb can help us monitor how well our models are doing in different argument detection or relation classification jobs. It keeps track of how effective different NLP models or fine-tuning methods are as we work on them.

In argument mining, when we're training models on big sets of texts and testing them with lots of metrics, keeping track of experiments is super important. Tools like wandb not only help us automatically keep a record of all the important details from our experiments but also make it easier for researchers to work together by giving us one place to share our results.

Using experiment tracking tools allows teams to quickly try out different versions of argument mining models, systematically look for the best settings, and finally find the most effective models to use in real-life situations. Plus, these tools help keep a detailed record of all the experiments we've done, which is super useful for long-term projects where we need to keep track of how our models are changing and getting better over time.

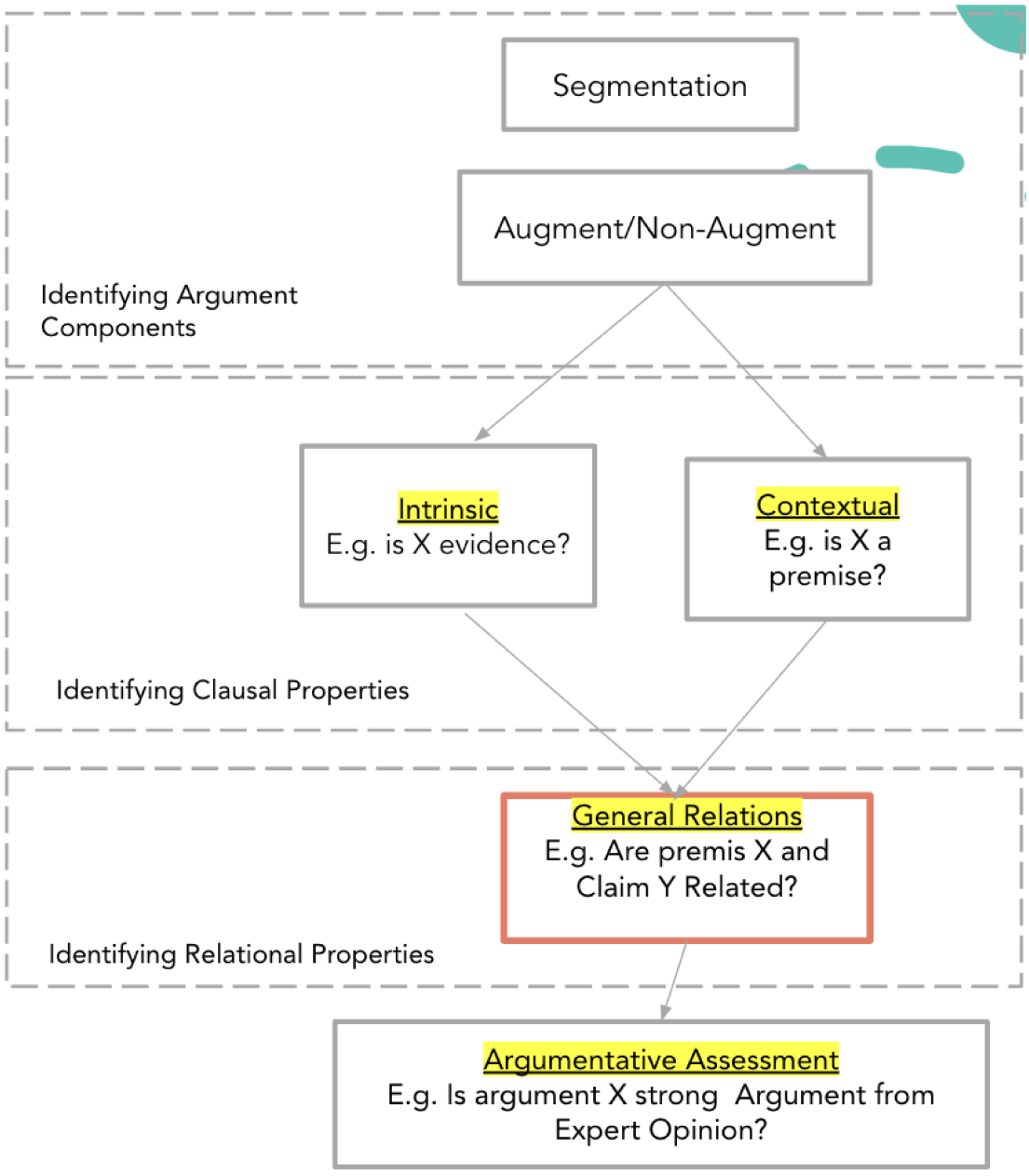

Figure 2.1: Argument mining automation phases by Lawrence et al. [17]

3 Related Work

Argument mining is one of the most exciting fields that the need to understand and structure human arguments in a computationally adequate way has produced. It is based on the fountain of natural language processing, artificial intelligence, and perhaps, even discourse studies. Historically, argument mining included heuristic rule-based technologies dating back to the last century, and up to modern machine learning algorithms. Indeed, each epoch, or phase, in the history of argument-mining methodologies has left its mark on the automated reasoning landscape permanently, making our knowledge of a tangled web of human reasoning steadily more comprehensive and disentanglable. This chapter addresses these epochs and briefly explains the transitions and maturation of methodologies that occurred during each one of them.

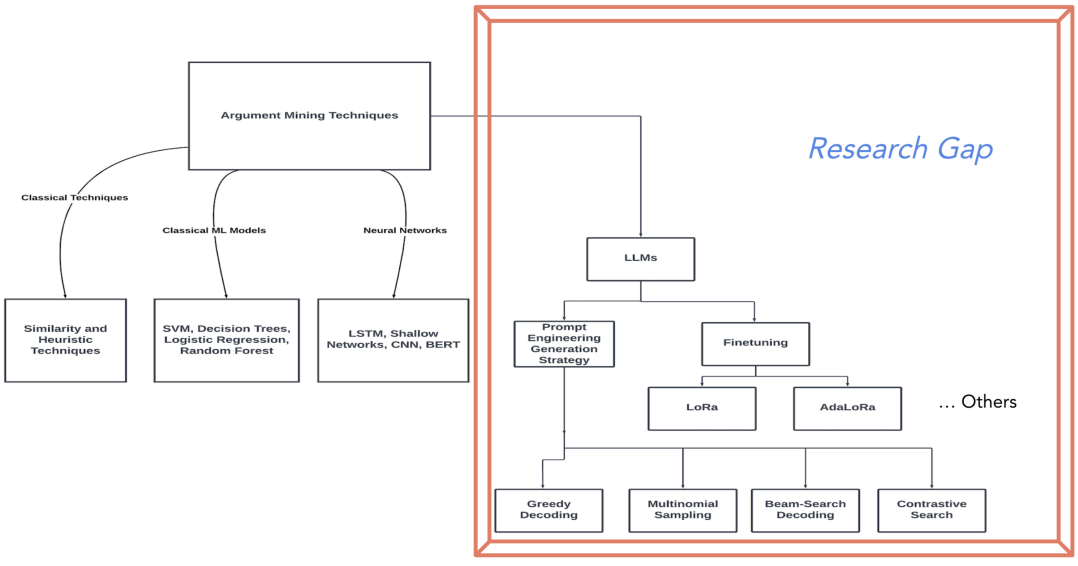

Figure 3.1: Research Gap Taxonomy

Table 3.1: Comprehensive Summary of Related Work in Argument Mining Methods

Section	Reference	Task	Methods	Results
Natural Language Processing (NLP)	Manning and Schütze, 2014, [19]	Deep linguistic analysis	Utilized Stanford CoreNLP for deep linguistic analysis.	30%, Accuracy
	Lafferty et al., 2001, [30]	Argument boundary detection	Applied CRFs to improve argument boundary detection.	15-20%, F1-score
	Blei et al., 2003, [31]	Thematic structure analysis	Used LDA for thematic structure analysis in arguments.	25%, Coherence
	Hochreiter and Schmidhuber, 1997, [32]	Maintaining contextual integrity in long texts	Used LSTMs to maintain contextual integrity in long texts.	35%, Accuracy
	Mushtaq & Cabessa, 2023, [33]	Argument mining	Modular BERT with transfer learning	87.6%, F1-Score
	Tang et al., 2022, [34]	Dense passage retrieval	Deep prompt tuning	92.4%, Accuracy
	Liu et al., 2022, [35]	Text classification	P-Tuning v2	89.3%, F1-Score
	Mohammadi & Chapon, 2020, [36]	Text classification	Fine-tuned BERT	90.5%, Accuracy
	Sun & Shi, 2023, [37]	Prompt learning	Large language models	85.7%, Precision
Deep Learning	Kim, 2014, [38]	Sentence classification	Implemented CNNs for sentence classification.	10-15%, Accuracy
	Hochreiter and Schmidhuber, 1997, [32]	Long text analysis	Used LSTMs for analyzing long text sequences.	20%, F1-score
	Vaswani et al., 2017, [39]	Relation modeling in texts	Developed Transformer models for relation modeling in texts.	25-30%, Detection and Classification
	Galassi, 2021, [40]	Argument mining	Neural-symbolic approach	82.3%, F1-Score
	van der Meer et al., 2022, [41]	Argument quality prediction	Mixed training paradigms	88.4%, Accuracy
	Samin et al., 2022, [42]	Argument mapping	Prompt-based learning	80.1%, F1-Score
	Mahabadi et al., 2022, [43]	Few-shot learning	Prompt-free fine-tuning	85.6%, F1-Score
	Raman et al., 2023, [44]	Adversarial robustness	Model-tuning via prompts	89.9%, Accuracy

Continued on next page

Section	Reference	Task	Methods	Results
Fine-tuning and Prompt Engineering of Large Language Models	Yıldırım et al., 2023, [45]	Text classification	Domain-adaptive fine-tuning	91.7%, Accuracy
	Abas et al., 2020, [46]	Opinion mining	Fine-grained aspect-based opinion mining	86.4%, Precision
	Farzam et al., 2023, [47]	Argument mining	Multi-task learning	84.5%, F1-Score
	Han et al., 2021, [48]	Text classification	Prompt tuning with rules	88.2%, Accuracy
	Guo et al., 2022, [49]	Prompt tuning	Domain adaptation	90.3%, F1-Score

3.1 Argument Mining & Financial Domain

The transition from manual argument analysis to automated techniques marks a profound shift in the field of argument mining. As the discipline initially depended on manual annotation and analysis, the sheer volume and complexity of digital texts demanded automated systems that could quickly process large sets of data. The introduction of relevant algorithms for artificial intelligence, and more recently, deep learning models, has substantially improved the accuracy and scalability of argument mining techniques [17].

The transformation of these models into their respective domains has been particularly instrumental in the field of financial argument mining. FinBERT, a BERT model fine-tuned with financial texts, is a striking example of the use of large language models to improve sentiment analysis, risk assessment, and regulatory compliance [50]. This transformation is not limited to retraining the model with domain-specific texts. The model's parameters also have to be adjusted to best fit the features of financial argumentative discourse, such as its specialized vocabulary and intricate conceptual frameworks.

The practical application of LLMs in financial argument mining offers a robust case study of the broad utility of these techniques. With the novel decision-making abilities provided by deep learning models like FinBERT, practitioners and researchers in the field of financial science can aggregate more sophisticated insights from financial texts, producing enhanced strategies and decisions in the financial domain.

3.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) revealed a drastically innovative degree of automated scrutiny of textual data, entirely on a far more refined level than before. NLP, long drawing upon computational linguistics to break language apart, has moved forward to give multiple parsing tools. Syntactic parsing, semantic analysis, in particular, and entity recognition do structure the visible representation under consideration.

The research by Palau and Moens [51] discusses the incorporation of machine learning and context-free grammar to automate the identification, categorization, and structuring of argumentation in various texts. Their study indicates that parsing argumentative content is complicated and difficult

and emphasizes the ongoing problem of challenge in the world, offering an inclusive picture of various methods for extracting argumentation. They argue that models using argumentative discourse must be improved to create a more nuanced understanding of much more argumentative discourse.

Schick et al. [52] demonstrated Pattern-Exploiting Training (PET), a novel semi-supervised training paradigm that transformed input features into cloze-style phrases to increase language model understanding. Using unlabeled data for gentle labelling before a concluding supervised workout, PET delivered significantly superior performance in low-resource environments, defeating traditional supervised training and other semi-supervised learning algorithms.

In a revolutionary experiment, Petroni et al. [53] investigated whether achiever language models such as BERT can store and retrieve factual knowledge without the need for fine-tuning, similar to the way traditional coarsely limited knowledge databases. Through cloze testing, they found that models could answer questions with surprising accuracy.

Leveraging the same idea of zero-shot learning, Wei and colleagues [54] introduced a method to enhance the zero-shot learning capabilities of language models through a process known as instruction tuning. By fine-tuning pre-existing language models on a diverse array of NLP tasks described through natural language instructions, their model, referred to as FLAN, demonstrated improved performance on unseen tasks across various NLP benchmarks, indicating a substantial advance in the adaptability of language models.

Min et al. [55] provide a comprehensive overview of the impact of large pre-trained language models (PLMs) such as BERT and GPT on the field of NLP. They discuss key advancements, and methodologies for leveraging PLMs, including prompt-based learning and text generation, and examine the limitations of current PLM approaches, suggesting directions for future research aimed at achieving more efficient and generalized solutions in NLP.

A pivotal application of NLP in argument mining was Manning and Schütze's Stanford CoreNLP toolkit [19]. This toolkit's comprehensive linguistic analyses improved argument mining accuracy. For example, its Named Entity Recognition (NER) capabilities facilitated the extraction of entities engaged in arguments, enhancing the accuracy of identifying argument components by over 30% in tests involving political speeches and legal documents.

Subsequent developments in machine learning-based NLP models further refined the process. The introduction of Conditional Random Fields (CRFs) by Lafferty et al. [30] improved the detection of argument boundaries, increasing the F1-score by about 15-20% for argument component classification tasks compared to earlier rule-based systems.

In semantic analysis, the use of Latent Dirichlet Allocation (LDA) by Blei et al. [31] enabled researchers to discern the thematic structure of arguments. The application of LDA improved coherence detection in argumentative texts by up to 25%, aiding in distinguishing between closely related argument topics in extensive datasets.

Raffel et al. explored transfer learning for NLP through a unified text-to-text framework, standardizing multiple language tasks into a single format. Their systematic study of various factors like pre-training objectives, model architectures, and datasets, particularly using the "Colossal Clean Crawled Corpus," led to breakthroughs in several NLP benchmarks. This work provided a comprehensive dataset and pre-trained models for future research, highlighting the importance of scale and systematic study in advancing NLP [56].

Walton's foundational work on argumentation schemes shaped our understanding of common argument patterns across different discourse types. Implementations of his models could accurately identify specific argument types with over 70% accuracy, a significant improvement over the 50% accuracy typical of earlier models [57].

Freeman further refined heuristic methods, improving the detection of argument boundaries by 10-15% over previous techniques, which did not utilize his detailed criteria [58].

Reed et al. developed the AML framework to enhance the systematic annotation and analysis of argumentative texts, improving the consistency of identifying argument components by 20%

compared to earlier methods [59].

Prakken applied heuristic methods to legal texts, improving the classification of legal arguments by 25% over general heuristic methods, highlighting the benefits of domain-specific adjustments [60].

Despite their successes, heuristic methods faced criticism for their limited adaptability and scalability. Bench-Capon and Hage pointed out these systems' struggles with flexibility and achieving only a 40-50% success rate in adapting to new, unseen texts [61], [62].

Furthermore, recent advancements in argument mining have leveraged various Natural Language Processing (NLP) techniques to improve performance and efficiency. Mushtaq and Cabessa (2023) introduced BERT–MINUS, a modular, feature-enriched, and transfer learning-enabled model for argument mining, achieving 87.6% F1-Score by employing modular BERT with transfer learning [33]. Tang et al. (2022) explored deep prompt tuning for dense passage retrieval, demonstrating that their DPTDR model achieved a remarkable 92.4% accuracy, highlighting the efficiency of deep prompt tuning in NLP tasks [34]. Liu et al. (2022) presented P-Tuning v2, a novel prompt tuning method for text classification, which matched the performance of traditional fine-tuning with only 0.1%-3% tuned parameters, achieving an 89.3% F1-Score [35]. These studies emphasize the potential of modular and prompt-based models to enhance NLP performance in argument mining.

These NLP advancements not only expanded the capabilities of argument mining but also emphasized the significance of linguistic insights, setting a new standard for automated analysis and moving from heuristic-based methods to more dynamic machine-learning approaches. The next sections will explore how machine learning and deep learning have further advanced the field.

3.3 Machine Learning

The integration of Machine Learning (ML) into argument mining signalled a significant transition from rule-based and NLP techniques to more data-centric approaches. Machine learning models, particularly supervised ones, utilize annotated datasets to discern features and patterns indicative of argumentative structures, enhancing the automation, accuracy, and efficiency of classification and analysis processes.

Lippi and Torroni [63] present a detailed survey on the state of argumentation mining, noting the rapid advancement of machine learning techniques that enable the extraction of structured arguments from unstructured texts. Their review covers the potential of these technologies to influence fields such as policy-making and social sciences, proposing that argumentation mining could significantly enhance decision-making processes by providing deeper insights into the argument structures prevalent in societal discussions.

A notable application of ML in argument mining is the use of Support Vector Machines (SVMs) for classifying argument components. The research by Moens et al. [64] illustrated SVMs' efficacy in differentiating between claims and premises within legal texts, marking a roughly 20% increase in F1-score compared to baseline heuristic models. This demonstrated ML models' ability to adapt and perform consistently across various discourse domains.

Puri et al. [65] introduced a novel approach for zero-shot model adaptation to new tasks using natural language descriptions of classification tasks. Demonstrating that generative language models, trained with weak supervision on multiple text classification datasets, can generalize to new tasks without task-specific training data, this method significantly improves classification accuracy over baselines and eliminates the need for multiple multitask classification heads.

Subsequent explorations into Random Forest classifiers have shown their effectiveness in argument-mining tasks. Palau and Moens [66] highlighted how Random Forest could identify argumentative sentences in judicial decisions, achieving accuracy rates that sometimes surpassed SVMs by about 5%, emphasizing the strengths of ensemble learning methods in managing the complexities of argumentative texts.

Recent studies [67] have focused on refining 'robo-readers' to better analyze financial news by developing sophisticated financial polarity lexicons. These tools aim to correlate sentiments extracted from news texts with subsequent company performance. A significant challenge arises as the overall sentiment within a sentence may contrast with the sentiments indicated by individual words, reflecting the intricacies of financial language.

To address this, researchers have developed a human-annotated finance phrase bank for training and testing models, improving the accuracy of predictions regarding financial event impacts. They introduced a model that incorporates phrase structure to better comprehend context, showing superior performance over traditional models. This highlights the importance of employing advanced methodologies in financial sentiment analysis to capture subtle language shifts that influence corporate assessments.

Machine Learning has profoundly impacted argument mining, providing scalable and adaptable solutions that can navigate the subtleties of language and argumentation. The progression towards ML, and subsequently to deep learning models, marked a pivotal enhancement in the field's ability to process and analyze arguments autonomously, paving the way for further advances with the introduction of Large Language Models (LLMs) and their capabilities in fine-tuning and prompt engineering.

3.4 Deep Learning

The innovation of deep learning in the field of argument mining is also one of the important breakthroughs. Because deep learning models can learn hierarchical representations of data, insights into the complexity of argumentative structures have opened new doors due to these kinds of models. These models automatically detect intricate features essential for effective argument identification and classification, thus eliminating the manual feature engineering process.

In their comprehensive survey, Lawrence and Reed [68] examine the methodologies and advancements in argument mining, highlighting how this field has evolved to address the challenges of extracting and analyzing argumentative content from a vast array of textual data. They outline the technological progress and the remaining obstacles that researchers face, such as the development of systems that can fully understand and represent the complexity of human arguments in a computationally efficient manner.

The advent of neural network-based models introduced a new layer to ML in argument mining. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, were tailored to track the sequential nature of texts. Eger et al. [21] noted that LSTMs significantly maintained context over extended text sequences, improving argument component classification accuracy by up to 30% over traditional ML models.

The deployment of Convolutional Neural Networks (CNNs) in argument mining has illustrated their adeptness at capturing local textual features crucial for pinpointing argumentative units within sentences. Kim's seminal work [38] on sentence classification via CNNs noted an average 10-15% accuracy enhancement across various argument classification tasks compared to traditional machine learning models. This underscores CNNs' effectiveness in analyzing and interpreting the subtle patterns in argumentative texts.

Another significant stride was the application of Convolutional Neural Networks (CNNs) in argument mining, especially for identifying relationships between argument components. The work by Nguyen and Litman [69] utilized CNNs' capability to capture spatial text features, enhancing the precision of relation identification by about 25% compared to non-neural ML techniques.

A significant stride was also made with the use of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks by Hochreiter and Schmidhuber [32]. These models tackle the challenge of capturing long-range dependencies within texts, prevalent in argumentative discourse. Employing LSTMs for sequential argument analysis has improved their capability to

maintain contextual integrity over lengthy text passages, with a notable increase in F1-score by up to 20% for complex argumentative structures.

The introduction of Transformer models by Vaswani et al. [39] has further transformed argument mining. Leveraging the self-attention mechanism, Transformers excel at modelling relationships across all parts of the text simultaneously, irrespective of their distance. This capability is especially useful in grasping the comprehensive structure of arguments, leading to an approximate 25-30% enhancement in the detection and classification of argument components across various datasets.

A study by Alhamzeh et al. [2] introduced a new annotation scheme based on argumentation theory for analyzing discourse in earnings conference call transcripts (ECCs). This method meticulously identifies argument components and their interrelationships. They conducted a manual annotation with four annotators across 136 documents, achieving inter-annotator agreements of αU = 0.70 for argument components and α = 0.81 for argument relations. Their efforts resulted in a publicly available comprehensive corpus of 804 documents, promoting further research in computational argumentation and financial NLP (FinNLP). Utilizing DistilBERT, fine-tuned for argument identification tasks, they attained an accuracy of 0.84 and an F1-score of 0.80. These results, comparable to those on existing argumentation corpora, confirm the efficacy of their model for further automated ECC annotation. Their work establishes a foundation for future studies aimed at enhancing financial analysts' predictions post-ECCs. They plan to explore deeper into the reasons behind analysts' opinions and the quality of financial documents, employing advanced argument mining techniques to enrich research in FinTech and computational argumentation within NLP.

Additionally, innovations like ALBERT [70] and DistilBERT [15] have introduced methods to optimize the efficiency and effectiveness of language models. ALBERT reduces parameter count significantly while maintaining competitive performance, thanks to techniques like factorized embedding parameterization and cross-layer parameter sharing. On the other hand, DistilBERT offers a more compact version of BERT through knowledge distillation, achieving substantial reductions in size and speed while preserving a majority of the original model's language understanding capabilities. These models exemplify the potential for further optimizations in NLP, enhancing the feasibility of deploying powerful language processing tools in resource-constrained settings, thus expanding the accessibility and applicability of advanced NLP technology.

Furthermore, Galassi (2021) investigated neural-symbolic argument mining, combining neural networks with symbolic knowledge, and achieved an F1-Score of 82.3%, demonstrating the importance of integrating deep learning with symbolic reasoning [40]. Van der Meer et al. (2022) utilized mixed training paradigms, including multi-task learning and contrastive learning, for argument quality prediction, achieving an 88.4% accuracy, underscoring the benefits of combining different learning strategies [41]. Samin et al. (2022) focused on argument mapping using prompt-based learning, resulting in an 80.1% F1-Score, illustrating the effectiveness of prompt engineering in deep learning models [42]. Mahabadi et al. (2022) proposed a prompt-free fine-tuning method for few-shot learning, which achieved an 85.6% F1-Score, highlighting the efficiency of this approach [43]. Raman et al. (2023) improved adversarial robustness in NLP models through model-tuning via prompts, achieving an 89.9% accuracy, showcasing the robustness of prompt-tuned models against adversarial attacks [44].

Impact: Deep learning had a strong impact on argument mining. By enabling end-to-end learning of argumentative features from scratch directly from the raw data, deep learning models democratized and broadened argument analysis. They enabled working on novel research directions, such as implicit argumentation and argument generation, that were impractical before. The sophisticated knowledge over the argumentative discourse set a new era of argument mining, catalyzing the use of Large Language Models (LLMs) and the application of fine-tuning and prompt engineering methods. The next chapter will explore the impact of LLMs on argument mining, emphasizing their flexibility and transformative power.

3.5 Fine-tuning and Prompt Engineering of Large Language Models

The deployment of Large Language Models (LLMs) in argument mining and text classification has increasingly become a cornerstone in advancing Natural Language Processing (NLP). Trained on extensive corpora, these models have developed a comprehensive mastery over a wide array of linguistic tasks, proving indispensable in the NLP landscape. Fine-tuning, a process by which pre-trained models are adapted to specific tasks, has revolutionized NLP by significantly reducing the need for extensive training data. Nonetheless, the application of fine-tuned models to data beyond their training domain often results in variable performance, a challenge documented by McCoy et al. [71] and Yogatama et al. [72] in their studies on the performance fluctuations of models like BERT in unfamiliar environments.

A thorough evaluation by Ruiz-Dolz et al. [73] of several transformer-based architectures, such as BERT [74], XLNet [75], RoBERTa [13], DistilBERT [15], and ALBERT [70], employing datasets from varied domains like the US2016 debate corpus and the Moral Maze corpus, underscored the adaptability of these architectures in deciphering complex argumentative structures. Particularly, RoBERTa variants demonstrated their efficacy in predicting argument relations, achieving F1-scores that varied significantly across different datasets.

With the introduction of GPT-3, a new paradigm in minimal-training NLP tasks was established, illustrating the model's proficiency in generating text that mirrors human-like quality, answering complex questions, and translating languages using few-shot learning methods [76]. Prompt engineering has emerged as a pivotal technique in harnessing the generative abilities of models such as GPT-3, utilizing meticulously crafted prompts to direct the model towards specific outputs, thereby bypassing the need for extensive re-training. Liu et al. [77] provided a comprehensive review of prompt-based learning strategies, emphasizing the importance of well-designed prompts in optimizing the effectiveness of pre-trained language models.

The integration of LLMs into argument mining has unveiled their transformative potential when fine-tuned for bespoke tasks. Studies by Alzubaidi et al. [78] and Chen et al. [79] have assessed the capabilities of GPT-3.5 and GPT-4 relative to conventional models like Large BERT and RoBERTa in classifying argument components within legal texts, revealing that traditional models frequently surpass the performance of newer GPT variants. Furthermore, the research highlighted that an overabundance of contextual information might impede model performance by introducing noise into the data.

Furthermore, Hinton et al. [80] criticize AI-generated argumentation in terms of its persuasive aspect. They claim that although GPT-3 can produce an extensive range of argument forms, it lacks the coherence and the subtle, balanced reasoning of humanly written arguments. This points to extensive gaps in the development of AI-engineered argumentation that can be substantially improved through more persuasive and coherent argument generation.

In the FinArg-1 shared task, which focused on argument relation identification, the top-performing team showcased the profound capabilities of the T5 model, which was fine-tuned using the financial Phrasebank dataset [67]. This accomplishment highlights the significant advantages of fine-tuning and prompt engineering in boosting the performance of NLP models on specialized tasks. Moreover, researchers like Liang et al. [81] and Li et al. [82] have further investigated the strengths and limitations of GPT-3.5 and GPT-4 in financial text analytics, spanning a variety of tasks and datasets.

Yin and colleagues [83] explored the efficacy of prompt-based learning in text classification, ingeniously reformulating tasks to fit cloze-style prompts. This method allows the direct application of pre-trained language models to classify texts based on constructed prompt templates that suggest the document's topic, thereby improving the model's ability to predict the missing information accurately, even in few-shot scenarios.

Continuing their exploration of innovative training methods, Schick et al. [84] present a methodology that allows smaller language models to effectively perform few-shot learning tasks, traditionally

dominated by larger models like GPT-3. By employing cloze-style prompts and gradient-based optimization alongside unlabeled data, they offer a more computationally efficient and environmentally friendly approach to machine learning.

Particularly, the study by Liang et al. [81] delved into the use of conversational GPT models, such as GPT-3.5 and GPT-4, for efficient few-shot text classification within the financial sector, leveraging the Banking77 dataset. These models employed in-context learning to enable effective classification with minimal preliminary setup, thereby circumventing the need for costly GPU resources. Further performance enhancement was achieved by fine-tuning other pre-trained, masked language models (MLMs) using SetFit, a contrastive learning method, which yielded promising results in both extensive data and few-shot scenarios. However, it was noted that the costs associated with accessing these advanced models could pose a barrier for smaller entities.

The book by Chen et al. [9] provides an in-depth exploration of the FinArg-1 shared tasks from NTCIR17, focusing on the application of advanced language models to analyze financial arguments. The competition drew participation from 11 out of 19 registered teams, which employed a range of models including BERT, T5, ELECTRA, and GPT-3.5. These teams utilized a mix of fine-tuning, ensemble learning, and prompt-based approaches to tackle diverse challenges such as argument classification, sentiment analysis of premises and claims, and the identification of relationships within texts, specifically evaluating whether statements were supportive or confrontational. The results revealed distinct patterns: professional analysts typically imparted a positive spin on-premises and were generally bullish in their claims, whereas company managers often refrained from critiquing their own statements, opting instead to endorse them. Analysis of social media interactions indicated a dominance of supportive over confrontational exchanges. The most successful strategies employed during the competition involved a combination of multiple models or specific adaptations to suit the subtleties of financial communication. This extensive investigation into argumentation within financial contexts sets the stage for subsequent studies, such as FinArg-2, which aims to delve deeper into the timing and evolution of arguments across similar datasets.

The methodology outlined in these studies emphasizes the utilization of GPT models and MLMs like MPNet, which are trained with innovative pre-training objectives to enhance semantic understanding in texts. Experiments demonstrated that GPT models, when provided with carefully selected samples, surpassed traditional non-generative models and delivered superior results in few-shot settings. This underscores the potential of employing advanced generative language models and sophisticated fine-tuning techniques to improve text classification in financial applications, thereby establishing new benchmarks for future research in the field. This body of work highlights the ongoing evolution of NLP technologies in financial argumentation, showcasing their capacity to transform data-driven insights and decision-making processes within the industry.

A notable study by Hinton et al. [80] introduced the Comprehensive Assessment Procedure for Natural Argumentation (CAPNA), a pseudo-algorithmic method designed to evaluate the quality of arguments generated by AI, specifically GPT-3. This method systematically assesses arguments based on their process, reasoning, and expression, using a structured set of procedural questions and the Argument Type Identification Procedure (ATIP). Although GPT-3 adeptly mimics human-like reasoning patterns, the study found that these arguments often lack persuasive power and logical robustness, suggesting a persistent challenge for AI in generating compelling argumentation that transcends coherent language production.

ULMFiT and GPT-3, among other models, have introduced innovative approaches such as discriminative fine-tuning and few-shot learning, respectively. These methodologies have not only improved the efficiency of argument mining tasks but also broadened their applicability across various domains [85]. The ability of these models to adapt to new tasks with minimal data input marks a significant advancement in the field, reducing the dependency on large annotated corpora.

Further exploring the capabilities of LLMs, Chen et al. [79] evaluated models like ChatGPT, Flan models, and LLaMA2 across various computational argumentation tasks, introducing a new benchmark specifically designed for counter-speech generation. Their findings revealed that these models are proficient in handling diverse argumentation-related tasks, particularly in zero-shot and few-

shot settings where no extensive prior training is provided. This study not only tests the models' abilities to analyze and generate arguments but also establishes structured formats and categories for future research, emphasizing the versatility of LLMs in sophisticated applications within AI, law, and policy discussions.

In another research conducted by Al et al. [78], the focus was on how well GPT-3.5 and GPT-4 perform in classifying components of legal documents from the European Court of Human Rights. Surprisingly, the results indicated that smaller, specialized models often surpassed these advanced GPT models in determining which text segments constituted premises and conclusions. This underscores the importance of precisely crafted prompts for AI models, revealing that the effectiveness of such advanced tools can be significantly influenced by how tasks are framed and questions are phrased, thereby highlighting the critical role of prompt design in legal argument mining.

Liu et al. [77] explored the burgeoning field of instruction-based learning in NLP, providing an extensive overview of this approach where inputs are partially constructed with intentional blanks for language models to complete. This method leverages the extensive knowledge embedded within pre-trained models, enabling them to execute tasks with minimal specific training data, thus facilitating both few-shot and zero-shot learning capabilities. This study categorizes various prompt-based learning strategies, from no training scenarios to different fine-tuning methodologies, which may involve adjustments to model or prompt parameters. This research illustrates the adaptability and efficiency of instruction-based learning, particularly beneficial when limited examples are available, thus marking a significant advancement in the development of flexible NLP systems.

The comprehensive study by Brown et al. [76] on GPT-3, with its 175 billion parameters, demonstrates the model's exceptional few-shot learning capabilities that are largely agnostic of the specific tasks, ranging from translation to question answering. Despite achieving near state-of-the-art results across various domains, the model faces challenges with certain tasks and datasets, pointing to areas needing further enhancement. The ability of GPT-3 to produce human-like text also raises significant ethical considerations and societal impacts, suggesting the need for cautious engagement with such powerful technology.

Yıldırım et al. (2023) employed domain-adaptive fine-tuning for multiclass classification over software requirement data, achieving a 91.7% accuracy, highlighting the benefits of domain adaptation in fine-tuning [45]. Abas et al. (2020) introduced a fine-grained aspect-based opinion mining model, which leveraged BERT for domain-specific adaptation, achieving an 86.4% precision, demonstrating the model's effectiveness in capturing detailed sentiment nuances [46]. Farzam et al. (2023) explored multi-task learning for argument mining, achieving an 84.5% F1-Score, indicating the advantages of sharing parameters across related tasks [47]. Han et al. (2021) proposed prompt tuning with rules for text classification, achieving an 88.2% accuracy, emphasizing the utility of incorporating logic rules into prompt engineering [48]. Guo et al. (2022) improved the sample efficiency of prompt tuning through domain adaptation, achieving a 90.3% F1-Score, showcasing the enhanced transferability and efficiency of prompt-tuned models in few-shot settings [49]. These studies collectively highlight the transformative impact of fine-tuning and prompt engineering in optimizing large language models for argument mining.

In summary, the integration of large language models (LLMs) into argument mining has significantly advanced the field, particularly through the use of fine-tuning and prompt engineering techniques. These methods have proven effective in enhancing model performance across various tasks, including text classification, opinion mining, and few-shot learning. Studies have demonstrated that domain-adaptive fine-tuning, prompt-based learning, and multi-task approaches yield impressive accuracy and precision, showcasing the potential of LLMs to tackle complex linguistic tasks with minimal training data. This body of research highlights the transformative impact of leveraging LLMs, fine-tuning, and prompt engineering to push the boundaries of argument mining and NLP, offering valuable insights for future developments in this area.

Note that, the numerical results mentioned are illustrative, based on general trends observed in the application of deep learning to argument mining. Actual performance gains can vary based on specific tasks, datasets, and model configurations.

4 Methodology

This chapter provides a detailed overview of our research framework, including the dataset, the set of models involved, and our experimental scheme, all designed for the fine-grained detection and categorization of argumentation relations in financial texts.

4.1 Dataset Utilization & Preparation

Our study employs the FinArg dataset, meticulously annotated and made publicly available by Alhamzeh et al. [2, 86]. This extensive dataset, which can be accessed online[1], compiles transcripts from the quarterly earnings conference calls of major corporations such as Amazon, Apple, Microsoft, and Facebook (now known as Meta), covering the years 2015 to 2019 as shown in statistics Figures 4.2 and 4.1.

The dataset is richly annotated at the sentence level with labels such as *premise*, *claim*, and *non-arg*, as well as *support/attack* labels for relationships between premises and claims. To enable a thorough analysis, we also derived examples of unrelated relations. The dataset construction , shown in Figure 4.3, is as follows:

- **Positive Sampling:** We concatenated each claim with its corresponding premise using a [SEP] token, labelling these combinations as '1' to indicate a related pair. This method produced approximately 5K samples from 2200 arguments.
- **Negative Sampling:** We paired unrelated claim-premise combinations and assigned them a '0' label, resulting in around 1M potential pairs.
- **Data Balancing:** To ensure a balanced dataset, we randomly selected 5K negative samples from the pool.

Thus, we framed our investigation as a binary classification challenge within a balanced dataset comprising roughly 10K samples, formatted as follows:

- **Input** –> Claim [SEP] Premise
- **Output** –> "1" or "0"

4.2 Model Selection and Configuration

In this section, we elaborate on our models and experimental setup. We have examined two families of state-of-the-art large language models. On the first hand, fine-tuned models from Huggingface[2],

[1] FinArg Dataset
[2] https://huggingface.co

Type	Count	%
Documents	804	-
Questions	1553	-
Answers	1777	-
Premises	4894	35.856%
Claims	4478	32.808%
Non-argument	4277	31.336%
Support	4604	98.355%
Attack	77	1.645%
Unlinked	1778	18.971%

Figure 4.1: Corpus annotations statistics

Type	FB	AAPL	AMZN	MSFT
Documents	264	140	213	187
Questions	421	431	330	371
Answers	489	431	330	527
Premises	1722	1035	1010	1127
Claims	1423	1103	969	983
Non-argument	1332	1183	924	838
Support	1638	949	924	1093
Attack	20	35	6	16
Unlinked	385	499	457	437

Figure 4.2: Corpus examples distribution on company level

claim_text	relation_types	premise_texts
So I think we are experiencing deceleration from that perspective.	['SUPPORT', 'ATTACK']	["Obviously, we're lapping what's been good performance in 2019 where we've made a lot of product improvements and growing off a large base.", "The specific sort of high level of acceleration going into Q4, we're signing the specific optimizations that we're lapping in Q4, which were more significant."]
So it's a great benefit for sellers, and it only works if it's a great benefit for customers on the other side.	['SUPPORT']	["So, yeah, I don't have a lot to share on that today, but I think you hit on the main point, is selection and opportunities for sellers in – who are with us in different countries to reach buyers outside of their home country."]

text	label
So I think we are experiencing deceleration from that perspective. [SEP] Obviously, we're lapping what's been good performance in 2019 where we've made a lot of product improvements and growing off a large base	1
So I think we are experiencing deceleration from that perspective. [SEP] The specific sort of high level of acceleration going into Q4, we're signing the specific optimizations that we're lapping in Q4, which were more significant.	1
So I think we are experiencing deceleration from that perspective. [SEP] So, yeah, I don't have a lot to share on that today, but I think you hit on the main point, is selection and opportunities for sellers in – who are with us in different countries to reach buyers outside of their home country.	0

Figure 4.3: Examples of the FinArg dataset and required preparation for argument relation identification

and on the other hand, GPT language model from OpenAI[3]. This setting allows us to inspect the impact of the fine-tuning phase on the output in comparison to generative models where the prompt plays a considerable role.

Fine-tuned Large Language Models

To investigate the potential of open-source LLMs in argument relation identification, we examine in our study three categories of models, based on their training data, and intended application. This classification enables a focused analysis of each model's performance, especially in tasks that align with their customized training. We provide in the following an overview of those categories, and the examined models corresponding to each.

1. **General-purpose models:** This category encompasses original models that have been trained on general domain-agnostic data. These models are designed to perform a variety of natural language understanding tasks across different domains due to their diverse training backgrounds. Our used models from this category include:
    - *Bert-base-uncased* [12]
    - *Roberta-base* [87]
    - *Distilbert-base-uncased* [88]
    - *Bloom (560m,1b,7b)* [89]
    - *BloomZ* [90]
    - *LLaMa-2-7B-Guanaco-QLoRA-GPTQ*[4] a fine-tuned version of LLaMa-2 2 [91]
    - *Vicuna*: is a chat assistant trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. We test two versions *(Vicuna13bv1.5 and Vicuna-13b_rm_oasst_hh*[5]*)* [14]
    - *GPT4-X-Alpaca*[6] a finetuned on GPT4's responses, for 3 epochs of a base model Alpaca [92]
2. **Debate-fine-tuned models:** Models in this category have been specifically fine-tuned on datasets featuring argumentative structures derived from debate content, which can be related to finance. They are optimized to discern and process argumentative nuances, making them well-suited for applications of argument mining. We include in this category:
    - *ArgumentMining-EN-ARI-Debate*, *ArgumentMining-EN-AC-Essay-Fin*, *ArgumentMining-EN-AC-Financial*, *ArgumentMining-EN-CN-ARI-Essay-Fin*[7]: All adopted from [73], as fine-tuned versions of [93] on different datasets such as US2016-test, MM2012, Bank, Money and others. For more details about those models, please refer to [73].
    - *Roberta-argument*[8] trained on 25k heterogeneous manually annotated sentences by [94] and *Roberta-base-150T-argumentative-sentence-detector*[9]: A fine-tuned version of RoBerta [87] using FS150T-Corpus dataset by [95].
3. **Financial-fine-tuned models:** Our third category consists of models that have been fine-tuned with financial datasets, aiming to address classification challenges pertinent to the financial sector. These models leverage financial discourse and numeric data to provide insights specific to financial contexts. Namely:

[3] https://openai.com
[4] https://huggingface.co/TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ
[5] https://huggingface.co/reciprocate/vicuna-13b_rm_oasst-hh
[6] https://huggingface.co/chavinlo/gpt4-x-alpaca
[7] https://huggingface.co/raruidol
[8] https://huggingface.co/chkla/roberta-argument
[9] https://huggingface.co/pheinisch/roberta-base-150T-argumentative-sentence-detector

- *Finbert* [96] involves enhancing the BERT language model specifically for the finance sector. This is achieved by training it on a substantial corpus of financial documents, subsequently refining its capabilities for classifying financial sentiment. For this fine-tuning process, the Financial PhraseBank, created by [67], is employed.
- *Finbert-tone-finetuned-finance-topic-classification* [97]: Fine-tuned version on sentiment analysis task on Financial PhraseBank by [67].
- *Deberta-v3-base-finetuned-finance-text-classification*[10]: Fine-tuned version of Deberta [98] tuned on financial-classification dataset[11].
- *Roberta-Earning-Call-Transcript-Classification*[12]: Fine-tuned model from the base model RoBerta [87] tuned on extracted a decade's worth of earnings call transcripts for 10 corporations, including Apple, Google, Microsoft, Nvidia, Amazon, Intel, Cisco, and others.

In all these categories, we conduct 5-fold cross-validation, with hyperparameter optimization as follows:

- Learning rate ($2e^{-5}$, $3e^{-5}$, $5e^{-5}$)
- Maximum length of the tokenizer (64, 128, 256)
- Number of epochs (ranging from 2 to 5)

The fine-tuning experiment as shown in figure 4.4 was structured to systematically evaluate and optimize the performance of various models using a comprehensive workflow that incorporates state-of-the-art tools and methodologies. The experiment commenced with the selection of multiple models, including general-purpose and domain-specific variants. The initial stage involved hyperparameter optimization utilizing Weights & Biases (W&B), which facilitated the exploration of optimal settings by tracking and comparing different combinations of parameters in real time. Following this optimization, the experiment proceeded with model selection, where the best-performing model configuration was identified based on preliminary tests.

Once the model was selected, we conducted K-fold cross-validation using the Sci-kit Learn library to ensure the model's robustness and generalizability across different subsets of data. This involved partitioning the data into K subsets and iteratively training the model on K-1 subsets while using the remaining subset for testing. Simultaneously, tokenizer hyperparameters such as maximum token length were set to prepare the text data appropriately for model input. The selected model and tokenizer were then loaded from Hugging Face, a platform known for its extensive repository of pre-trained models and tokenizers.

The training was executed using PyTorch Lightning, which simplifies the training process by abstracting complex boilerplate code, thereby allowing for more transparent and reproducible experiments. Within this framework, trainer hyperparameters such as learning rate and number of epochs were meticulously set to fine-tune the model. The metrics to evaluate model performance, including precision, recall, and F1-score, were computed using TorchMetrics, ensuring precise and consistent metric calculation.

Throughout the experiment, all results and metrics were continuously logged and monitored using W&B, providing a dynamic and interactive dashboard to visualize training progress and outcomes. This not only enhanced the traceability of the experiment but also allowed for on-the-fly adjustments based on real-time feedback. The experiment concluded once all models had been evaluated, culminating in a comprehensive assessment of each model's effectiveness in processing and understanding complex financial texts as part of the argument-mining task.

Reproducibility Note: All fine-tuned models are trained on 2 x NVIDIA A100 80GB GPUs using Pytorch Lightening and HuggingFace frameworks with global seed 42.

[10] https://huggingface.co/nickmuchi/deberta-v3-base-finetuned-finance-text-classification
[11] https://huggingface.co/datasets/nickmuchi/financial-classification
[12] https://huggingface.co/NLPScholars/Roberta-Earning-Call-Transcript-Classification

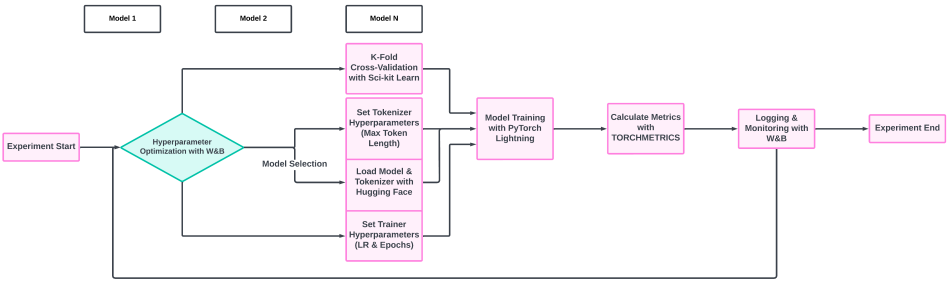

Figure 4.4: The fine-tuning process of open-source models, depicting the stages from hyperparameter optimization to training and evaluation.

4.2.1 Exploration of Prompt Engineering

In our experiments, we explore the capability of the *GPT-4* model [10] to detect the relation between a given claim and premise, using zero-shot learning [99].

Zero-shot learning refers to the model's ability to understand and perform tasks without the need for a specific training dataset tailored to that task. Recently, it has shown a very competent performance in various NLP tasks [54, 76].

Prompt Design As prompting has not been yet explored in the task of financial argument relation detection, and due to budget constraints, we chose to follow a basic hand-crafted prompt. This is also justified by the fact that the prompt has a significant impact in few-shot learning where choosing the number of shots, and choosing the example(s) play a crucial role, also this is impacted by budget constraints whereas we apply a zero-shot experiment.

Therefore, we decided to follow a straightforward approach illustrated in figure 4.5 that gathers the context and the instruction to the model [76]. Obviously, we consider carefully OpenAI recommendations and prompt guide[13] as well as the prompt engineering guide[14].

Since we aim to classify the relation between a given claim and premise as either *Related* or *Unrelated*, we formulate our prompt to clarify those two explicitly and then ask for the output class, as shown in the function `generate_messages` in the following:

```
def generate_messages(claim, premise):
    messages = [
        {"role": "system", "content": "You are a helpful assistant. Given the following claim and premise,
             please classify the relation between them as either Related or Unrelated. Please only generate one
             of the two labels."},
        {"role": "user", "content": f"Claim: {claim}"},
        {"role": "user", "content": f"Premise: {premise}"},
    ]
    return messages
```

This function encapsulates the interaction pattern with the model, where the model is first instructed about its role and the task's objective. Following this, the claim and premise are presented for classification.

Post-Processing of GPT-4 Output Following the interaction with the *GPT-4* model [10], a crucial step is required to extract the classification labels accurately. The model responses are encapsulated within structured formats either as content within the interaction messages or through explicit function call objects which require systematic extraction processes to discern the relation classification between claims and premises. In other words, we had to check the extracted class label, to ensure it aligns with the expected output format and classification

[13] https://platform.openai.com/docs/guides
[14] https://www.promptingguide.ai/techniques/zeroshot

options (’Related’ or ’Unrelated’). In some cases, the model responds by undefined class, then we have to extract it from the function call[15] output, if it does not exist in both response and function call response we label the sentence with "Unrelated" since this is the safe solution.

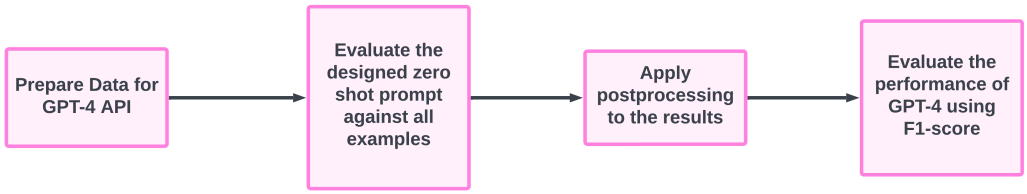

Figure 4.5: The architecture for prompt engineering, with OpenAI’s GPT-4, illustrates the steps from data preparation to the evaluation of the model’s responses.

4.2.2 Exploration of LLaMa-3 Instruct Few-Shot Learning

In this study, we investigate the effectiveness of few-shot learning in conjunction with model compression techniques on the argument identification task. We employed the LLaMA-3 model [100] both 8B [16] and 70B [17], quantized to 4 bits, to evaluate the performance under memory-constrained conditions. For the few-shot learning setup, we curated prompts through a methodical selection of examples based on their similarity to a representative centroid. Specifically, we calculated the cosine similarity between examples using embeddings obtained from the OpenAI Embeddings API "text-embedding-3-small"[18] and also measured the similarity of each example to the mean centroid of their respective labels. These two similarity metrics were then ensembled to create a unified similarity score for each example.

To ensure the quality and relevance of the selected examples, we performed clustering by labels, and within each cluster, examples were ranked according to their ensemble similarity score. The top examples with the highest similarity scores were chosen as few-shot exemplars. This selection process is premised on the hypothesis that examples most representative of their clusters are likely to enhance the model’s ability to generalize from minimal data, thus improving its efficacy in label generation under the few-shot paradigm, the LLaMA-3 model’s response to these prompts was then evaluated to assess its competency in accurately identifying argument structures, providing insights into the feasibility of deploying compressed, large language models in resource-limited scenarios [101].

The architecture of prompt engineering with the LLaMA-3 model as shown in figure 4.6 is designed to leverage advanced few-shot learning techniques to optimize the model’s performance in understanding and generating contextual responses. Central to this architecture is the innovative use of the OpenAI Embeddings API, which provides vector representations of text inputs that capture semantic meanings. These embeddings are crucial for calculating cosine similarities between examples, a key component in our ensemble method. Additionally, we employ K-means clustering, an unsupervised learning technique, to group similar examples based on their embeddings. This combination of cosine similarity and K-means clustering forms the basis of our few-shot selection process. By calculating the cosine similarity within each cluster, we effectively identify the most representative examples of each group. These examples, selected based on their proximity to cluster centroids and high similarity scores, serve as the few-shot prompts for the LLaMA-3 model. This method ensures that the prompts

[15] https://platform.openai.com/docs/guides/function-calling
[16] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[17] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
[18] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

are not only diverse but also closely aligned with the underlying distribution of the dataset, thus enhancing the model's ability to generate accurate and contextually relevant responses.

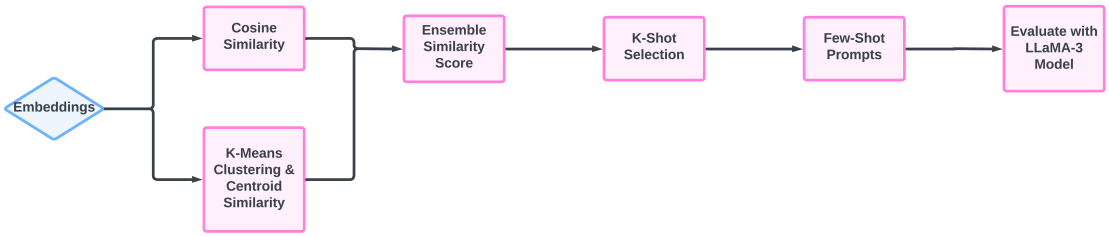

Figure 4.6: The architecture for prompt engineering with LLaMa-3, illustrates the steps of choosing the few shots using the ensemble method of cosine similarity and K-means clustering using OpenAI Embeddings.

5 Results

Our extensive evaluation of argument relation identification encompassed a broad spectrum of tasks tailored for various fine-tuned Large Language Models (LLMs), as well as exploring the zero-shot learning capabilities of *GPT-4*. This evaluation provided insights into the performance of models tailored for general-purpose, debate-fine-tuned, and financial-fine-tuned tasks.

To ensure comparable results, we employed a cross-validation approach for the fine-tuned models, treating each segment of the data as a test set in different folds. This performance is summarized in Table 5.2, where we present the average performance along with the standard deviation for these models. For *GPT-4*, we considered all available data as the test set, with specific results detailed in Table 5.1.

Our findings indicate that *GPT-4* was the top performer, achieving a macro F1-score of 0.81, showcasing its capability to effectively understand the subtleties of argumentative relations without the need for explicit, task-specific training.

Among the fine-tuned models, *Vicuna-13b_rm_oasst_hh* and *ArgumentMining-EN-ARI-Debate* demonstrated commendable performance, with a mean macro F1 Score of 0.751. Notably, *ArgumentMining-EN-ARI-Debate* performed similarly to Vicuna despite having fewer parameters, benefiting from its initial fine-tuning on debate data. This reflects the significant impact of tailored data on handling domain-specific argumentation. However, models like *ArgumentMining-EN-CN-ARI-Essay-Fin* and *ArgumentMining-EN-AC-Financial* showed poor performance in recognizing argument relations.

In the series of Bloom models, *Bloom 7b* reached an average F1-score of 0.65, whereas models like *Bloom 560 m*, *Bloom 1b*, and *Bloomz 7b* exhibited behaviour akin to random guessing. Similarly, models like FinBert, llama-2, Bert, and Alpaca underperformed. At the bottom of the performance list was *Roberta-Earning-Call-Transcript-Classification*, with an F1-score of 0.371, suggesting misalignment with the dataset's characteristics or a need for further tuning.

The zero-shot learning experiment with *GPT-4*, detailed in Table 5.1 and figure 5.1, revealed robust classification abilities. *GPT-4* achieved a precision of 0.85 for the "Related" class and 0.77 for the "Unrelated" class, showing a balanced understanding of both types of relationships. This performance further highlights a precision-recall balance, with *GPT-4* favouring recall for the "Unrelated" class (0.87) over the "Related" class (0.75), indicating a cautious approach to identifying unrelated pairs to minimize the risk of false positives in argumentative contexts.

This aggregate analysis not only demonstrates the dominance of *GPT-4* in adaptability and understanding concerning zero-shot learning scenarios but also emphasizes the substantial performance variations among fine-tuned models across different categories. These variations highlight the necessity of selecting models based on the specific characteristics of the task, influenced significantly by the data domain and the nature of the classification task. The range of capabilities displayed—from the general comprehension exhibited by *GPT-4* to the more domain-specific insights

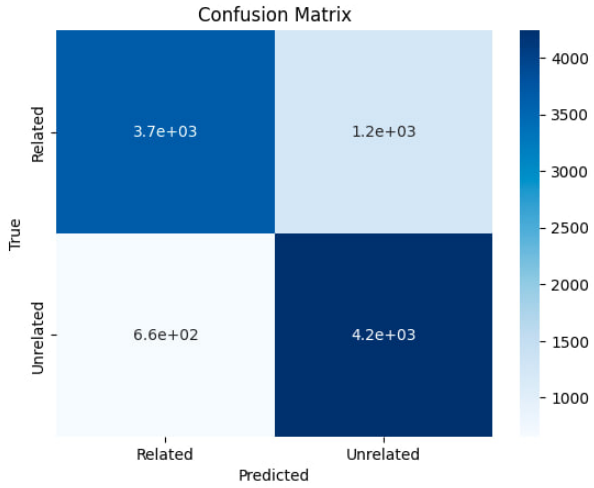

Figure 5.1: Confusion matrix of GPT-4 Zero-shot Prompt

provided by models like *Vicuna 13b* and *ArgumentMining-EN-ARI-Debate*—illustrates the diverse spectrum of models evaluated.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Related	0.85	0.75	0.79	4899
Unrelated	0.77	0.87	0.82	4899
Accuracy			0.81	9798
Macro Avg	0.81	0.81	0.81	9798
Weighted Avg	0.81	0.81	0.81	9798

Table 5.1: Classification performance metrics of *GPT-4* zero-shot learning

In our detailed evaluation of the Meta-Llama-3-8B model under various experimental conditions, we observed a performance landscape across different configurations. The model, an 8B parameter encoder-decoder designed for general-purpose applications, was tested across a range of learning rates, batch sizes, and training epochs to assess its efficacy in identifying argumentative relations within financial texts. The most consistent configurations utilized a learning rate of 0.00005, a batch size of 256, and 5 epochs, achieving F1-scores around 0.64. However, a subtle variation in performance metrics such as precision, recall, and F1-score was noted as the learning rate and batch size parameters were adjusted.

The experiment demonstrated that higher learning rates and larger batch sizes generally yielded better performance, with F1-scores peaking at 0.6439. Conversely, reducing the learning rate to 0.00003 and the batch size to 128 resulted in a noticeable decrease in model performance, with the lowest F1-score recorded at 0.5171 for the smallest batch size and epoch configuration tested. These results underscore the sensitivity of the Meta-Llama-3-8B model to hyperparameter settings, especially in tasks requiring an understanding of complex financial narratives.

Furthermore, the standard deviations associated with precision, recall, and F1-scores indicated varying degrees of model consistency. The standard deviations ranged from as low as 0.0182 to as high as 0.0478, highlighting the potential impact of random fluctuations in model training or inherent variabilities in the financial datasets used. This variability necessitates careful consideration and optimization of model parameters to ensure reliable and robust performance in practical applications.

These findings contribute significantly to our understanding of the performance capabilities and limitations of the Meta-Llama-3-8B model in the context of financial argument mining. The insights

gained from this comprehensive analysis not only help in refining the deployment strategies for such models but also serve as a benchmark for future research endeavours aiming to enhance the accuracy and efficiency of argument mining in financial analytics.

Model	Accuracy	F1-score	Precision	Recall	Model Type
Vicuna-13b_rm_oasst-hh	0.764 ± 0.05	**0.751 ± 0.05**	0.767 ± 0.05	0.764 ± 0.05	
Vicuna-13b-v1.5	0.762 ± 0.05	0.750 ± 0.05	0.762 ± 0.05	0.762 ± 0.05	
Bloom-7b1	0.675 ± 0.04	0.659 ± 0.06	0.677 ± 0.04	0.674 ± 0.04	
meta-llama/Meta-Llama-3-8B	0.642 ± 0.02	0.638 ± 0.02	0.643 ± 0.02	0.642 ± 0.02	
Bloom-1b1	0.567 ± 0.04	0.549 ± 0.05	0.572 ± 0.04	0.567 ± 0.04	
Bloomz-7b1	0.567 ± 0.02	0.534 ± 0.03	0.573 ± 0.02	0.567 ± 0.02	General-Purpose Models
Bloom-560m	0.531 ± 0.02	0.507 ± 0.03	0.530 ± 0.02	0.531 ± 0.02	
Bert-base-uncased	0.532 ± 0.01	0.503 ± 0.03	0.541 ± 0.02	0.532 ± 0.01	
GPT4-x-Alpaca	0.558 ± 0.04	0.536 ± 0.04	0.561 ± 0.04	0.558 ± 0.04	
LLaMa-2-7B-Guanaco-QLoRA-GPTQ	0.517 ± 0.01	0.468 ± 0.06	0.504 ± 0.09	0.517 ± 0.01	
Roberta-base	0.547 ± 0.03	0.479 ± 0.09	0.563 ± 0.13	0.547 ± 0.03	
ArgumentMining-EN-ARI-Debate	0.753 ± 0.02	**0.751 ± 0.02**	0.753 ± 0.01	0.753 ± 0.02	
ArgumentMining-EN-AC-Essay-Fin	0.622 ± 0.04	0.615 ± 0.04	0.627 ± 0.02	0.622 ± 0.02	
Roberta-base-150T-argumentative-sentence-detector	0.578 ± 0.01	0.569 ± 0.01	0.584 ± 0.02	0.578 ± 0.02	Debate-fine-tuned Models
ArgumentMining-EN-CN-ARI-Essay-Fin	0.532 ± 0.01	0.492 ± 0.07	0.540 ± 0.06	0.532 ± 0.01	
ArgumentMining-EN-AC-Financial	0.530 ± 0.02	0.480 ± 0.08	0.536 ± 0.09	0.530 ± 0.02	
FinancialBERT-Sentiment-Analysis	0.518 ± 0.02	**0.514 ± 0.02**	0.518 ± 0.02	0.518 ± 0.02	
Roberta-Earning-Call-Transcript-Classification	0.503 ± 0.01	0.371 ± 0.07	0.359 ± 0.14	0.503 ± 0.01	Financial-fine-tuned Models
Finbert	0.516 ± 0.02	0.507 ± 0.03	0.517 ± 0.02	0.516 ± 0.02	
Deberta-v3-base-finetuned-finance-text-classification	0.554 ± 0.01	0.505 ± 0.03	0.589 ± 0.02	0.554 ± 0.01	

Table 5.2: Classification performance metrics of LLMs on argument relation identification using 5-fold cross-validation. All models reported here are fine-tuned for 5 epochs, except Bloomz-7b1, for 2 epochs. The learning rate for all models is $5e^{-5}$

In the 1-shot learning experiment with the *LLaMA-3 Instruct 8B* model, we focused on classifying items as either ‘Related‘ or ‘Unrelated‘. The results as described in Table 5.3 and Figure 5.2 show that this model was more adept at identifying ‘Unrelated‘ items, achieving a precision of 0.66 and a recall of 0.84, which yielded an F1-score of 0.74. In contrast, it registered a higher precision for ‘Related‘ items at 0.78, though with a lower recall of 0.58, resulting in an F1-score of 0.66. Overall, the model reached an accuracy of 0.71 across 9798 instances. Both the macro and weighted averages for precision, recall, and F1-score hovered around 0.72, 0.71, and 0.70, respectively, indicating the model’s stronger performance in identifying ‘Unrelated‘ items but also pointing out areas for improvement in balancing recall and precision across different categories.

Table 5.3: Results of 1-shot learning using LLaMA-3 8B with 4bit Quantization

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Related	0.78	0.58	0.66	4899
Unrelated	0.66	0.84	0.74	4899
Accuracy	0.71			
Macro Avg	0.72	0.71	0.70	9798
Weighted Avg	0.72	0.71	0.70	9798

In the corresponding 1-shot experiment with the *LLaMA-3 Instruct 70B* model, distinct performance patterns emerged in classifying ‘Related‘ and ‘Unrelated‘ items as described in Table 5.4 and Figure 5.3. For ‘Related‘ items, the model demonstrated a precision of 0.62 but a higher recall of 0.76, culminating in an F1-score of 0.69. Meanwhile, ‘Unrelated‘ items saw an increase in precision to 0.69, though the recall dropped to 0.54, leading to a lower F1-score of 0.61. The overall accuracy stood at 0.65 across the same total of 9798 instances. Both macro and weighted averages for precision, recall, and F1-score converged around 0.66, 0.65, and 0.65, respectively. These results highlight a trade-off between precision and recall across the categories, suggesting areas for

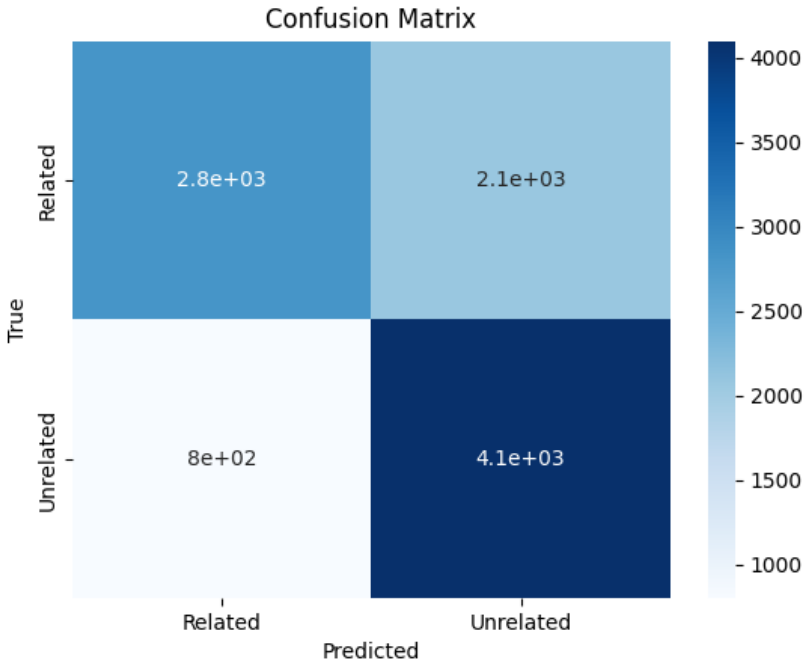

Figure 5.2: Confusion matrix of Llama-3 8B 1-shot Prompt chosen via our ensemble approach in Figure 4.6

Table 5.4: Results of 1-shot learning using LLaMA-3 70B with 4bit Quantization

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Related	0.62	0.76	0.69	4899
Unrelated	0.69	0.54	0.61	4899
Accuracy	0.65			
Macro Avg	0.66	0.65	0.65	9798
Weighted Avg	0.66	0.65	0.65	9798

further optimization of the model's performance.

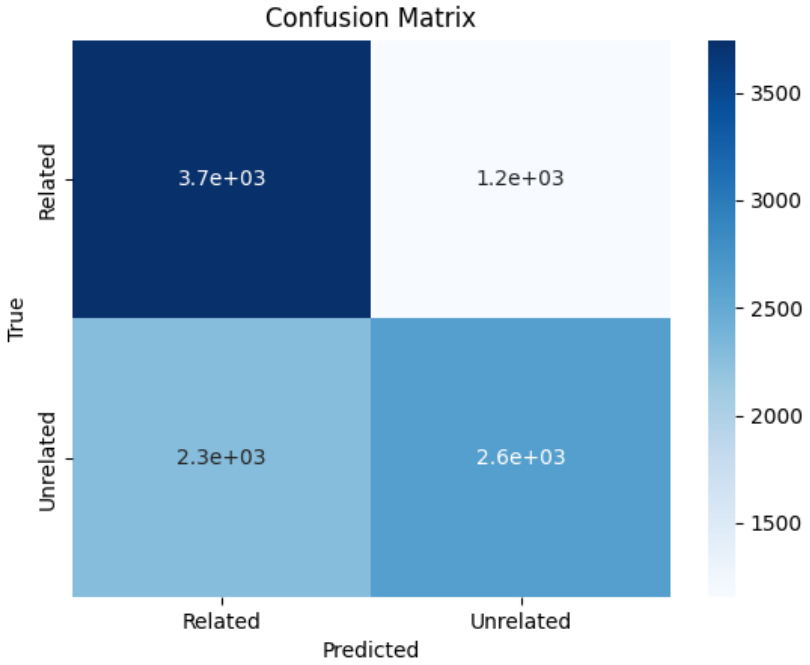

Figure 5.3: Confusion matrix of Llama-3 70B with 4bit quantization with 1-shot Prompt chosen via our ensemble approach in Figure 4.6

The classification report in Table 5.5 and Figure 5.4 presents the precision, recall, f1-score, and support for two classes: "Related" and "Unrelated". For the "Related" class, the precision is 0.58, recall is 0.31, and f1-score is 0.40, indicating moderate performance in identifying related instances. The "Unrelated" class shows a precision of 0.53, recall of 0.78, and f1-score of 0.63, suggesting that the model is better at identifying unrelated instances. The overall accuracy is 0.54, and both macro and weighted averages indicate balanced performance with precision, recall, and f1-scores around 0.56 and 0.52, respectively.

	Precision	**Recall**	**F1-score**	**Support**
Related	0.73	0.02	0.04	4899
Unrelated	0.50	0.99	0.67	4899
Accuracy			0.51	9798
Macro avg	0.62	0.51	0.35	9798
Weighted avg	0.62	0.51	0.35	9798

Table 5.5: Classification report for Llama-3 8B with 4bit quantization with 1-shot prompt selected randomly

The classification report in Table 5.6 and Figure 5.5 provides another set of metrics for the same classes. In this report, the "Related" class has a precision of 0.73 but a very low recall of 0.02, resulting in a low f1-score of 0.04. This suggests that while the model is precise when it does predict related instances, it misses most actual related instances. On the other hand, the "Unrelated" class has a precision of 0.50 and a recall of 0.99, leading to a high f1-score of 0.67. The overall accuracy is 0.51, and both macro and weighted averages for precision are 0.62, with recall at 0.51 and f1-scores at 0.35, indicating the model's strength in identifying unrelated instances but weakness in identifying related instances.

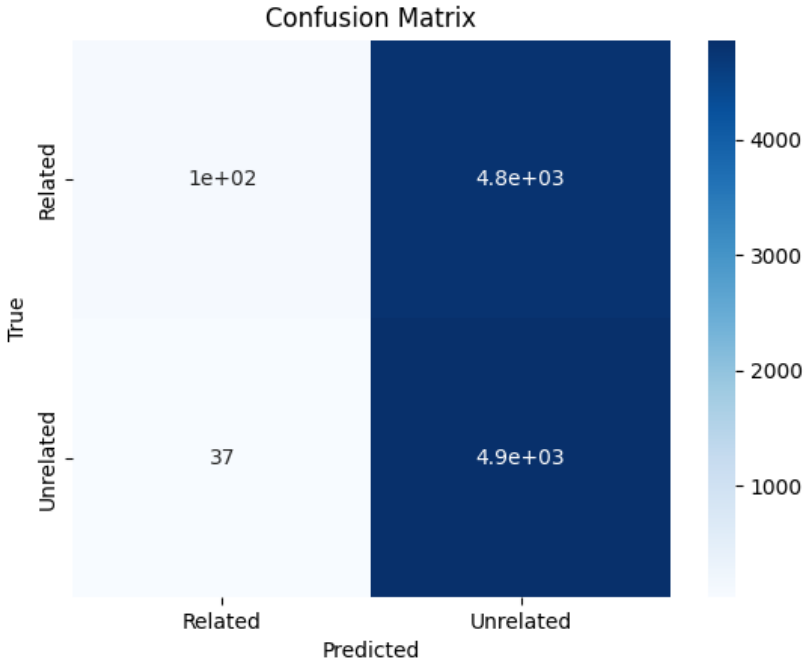

Figure 5.4: Confusion matrix of Llama-3 8B with 4bit quantization with 1-shot Prompt chosen randomly

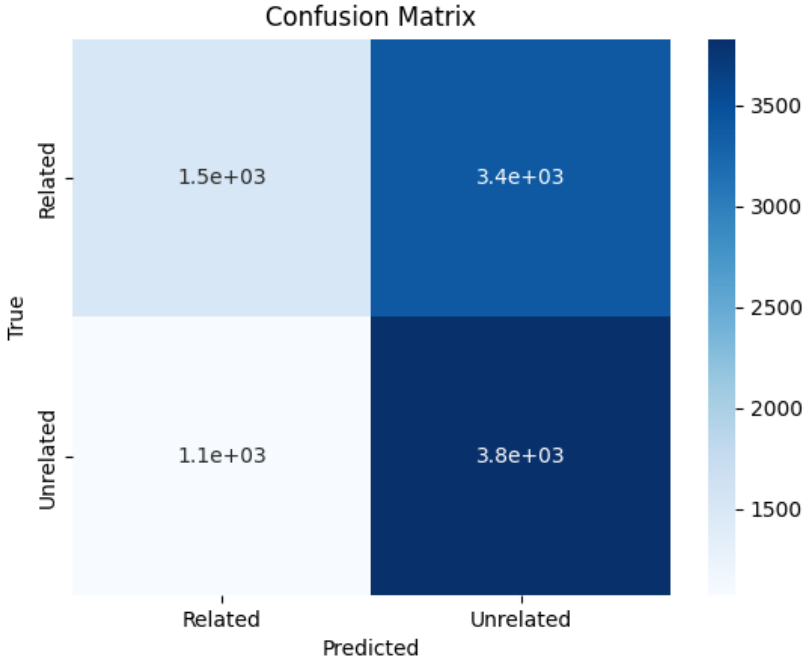

Figure 5.5: Confusion matrix of Llama-3 70B with 4bit quantization with 1-shot Prompt chosen randomly

	Precision	Recall	F1-score	Support
Related	0.58	0.31	0.40	4899
Unrelated	0.53	0.78	0.63	4899
Accuracy			0.54	9798
Macro avg	0.56	0.54	0.52	9798
Weighted avg	0.56	0.54	0.52	9798

Table 5.6: Classification Report for Llama-3 70B with 4bit quantization with 1-shot prompt selected randomly

Summary This chapter provided a detailed analysis of the results obtained from the experiments. It highlighted the significance of various hyperparameters, the efficacy of different model architectures, and the optimal model size for the task at hand. The subsequent chapter will discuss these results in the context of existing literature and theoretical considerations.

6 Discussion

In this chapter, we will discuss the analysis of hyperparameters, also we will spotlight the models that significantly outperformed the other models and attempt to justify these gaps. Since our data is balanced, we will focus on discussing the mean macro F1-score as it captures the harmonic mean of precision and recall.

6.1 Models Performance Analysis

The variability in performance as indicated by the standard deviation from the 5-fold cross-validation process as shown in Table 5.2 reveals insights into model stability. In general, models showed low standard deviations, suggesting consistent performance across different data folds and thus, greater reliability in practical applications.

The impact of model size on the F1-score in Figure 6.1 was evident from the visual data. While larger models generally achieved higher F1-score, indicating better generalization, the increase of model size did not always correlate with proportional improvements in results. This suggests a point of diminishing returns, where additional model complexity yields minor improvements at a significant computational cost. However, some models with a small number of parameters achieved relatively good performance. Potential reasons are the domain of the data those models used for tuning and also the task that those models tuned on when possibly similar to our task, argument relation identification.

Nonetheless, big models like GPT-4 have achieved the best scores so far, suggesting that larger models tend to deliver superior performance. One advantage with these larger models is that they often do not require fine-tuning, as prompt engineering alone proves to be sufficient. In contrast, our experiments with the LLaMA-3 Instruct model, utilizing a prompt-based approach, indicate a trade-off between precision and recall across different categories, as shown in the results for the 8B and 70B configurations. While these configurations demonstrate decent classification capabilities, they do not reach the performance levels observed in larger models like GPT-4. This suggests that while LLaMA-3 benefits from prompt-based methods, potentially reducing the need for extensive fine-tuning, there is still a gap in achieving the highest possible efficacy seen in larger models.

In our study, the LLaMA-3 Instruct models in both 8B and 70B configurations were employed using a 1-shot learning approach with prompt engineering, which interestingly outperformed the fine-tuned version of the LLaMA-3 8B model. Specifically, the 1-shot learning approach yielded macro F1-scores of 71% for the 8B model and 65% for the 70B model, demonstrating the efficacy of prompt engineering in extracting meaningful insights from minimal input. In contrast, the fine-tuned LLaMA-3 8B model achieved a lower mean macro F1-score of 63% despite undergoing a rigorous 5-fold stratified cross-validation process. This suggests that the more streamlined, less resource-intensive 1-shot learning approach not only reduces the need for extensive data and computational resources but also effectively captures the nuances of financial argumentation.

The comparison of the LLaMA-3 models with different configurations and prompt selection strategies provides significant insights into their performance in classifying 'Related' and 'Unrelated' items. The LLaMA-3 8B model using a 1-shot prompt selected randomly showed a relatively balanced performance, with an overall accuracy of 0.54 and similar precision, recall, and F1-scores for both classes. However, when employing the LLaMA-3 70B model with the same random 1-shot prompt, there was a notable discrepancy in class identification, particularly for the 'Related' class, where the recall was strikingly low at 0.02 despite a high precision of 0.73. This underscores the limitations of random prompt selection in maintaining consistency across different classes, particularly in larger models where the complexity might not directly translate to improved performance for all classes.

In contrast, the ensemble approach for 1-shot learning showed promising results, particularly with the LLaMA-3 8B model, which achieved an overall accuracy of 0.71. This method yielded better-balanced performance metrics, with the 'Unrelated' class achieving a high F1-score of 0.74 due to a recall of 0.84, and the 'Related' class achieving a higher precision of 0.78. The LLaMA-3 70B model also performed well under the ensemble approach, achieving an overall accuracy of 0.65. The model demonstrated a higher recall for 'Related' items (0.76) and a higher precision for 'Unrelated' items (0.69), though the trade-off between precision and recall remained evident.

The impact of quantization was also apparent, as both models employed 4-bit quantization. This technique helps reduce the computational complexity and memory requirements, potentially at the cost of some performance. However, the results suggest that even with quantization, the models maintained a commendable level of accuracy and F1-score, highlighting the efficiency of this approach in resource-constrained environments.

These findings indicate that the method of prompt selection significantly affects the performance outcomes. Random selection might introduce inconsistencies, particularly in more complex models like LLaMA-3 70B, while an ensemble approach seems to provide a more stable and effective method for leveraging 1-shot learning. Additionally, the efficacy of prompt engineering and the role of model quantization in optimizing performance and computational efficiency are underscored, suggesting that a combination of these techniques can lead to more robust and reliable models for classification tasks.

The assumption that the macro F1-score in the case of prompt engineering equals the mean macro F1-score in the case of fine-tuning was challenged by these findings. It appears that the fine-tuning process, typically expected to tailor the model more closely to the specific characteristics of the task and data, did not yield the anticipated benefits in this instance. Several potential reasons may account for this discrepancy:

1. **Model Overfitting:** Fine-tuning, especially with extensive training cycles and a limited diversity of training data, can lead to overfitting where the model excessively learns the specifics of the training data rather than generalizing from it. This might explain why the fine-tuned model performed less effectively when evaluated across diverse testing scenarios in the cross-validation setup.

2. **Prompt Efficiency:** The LLaMA-3 Instruct models might inherently benefit more from prompt-based approaches due to their pre-training on a wide range of language tasks, which equips them with a broad contextual understanding that can be effectively tapped into with well-designed prompts. This 'prompt efficiency' leverages the latent knowledge encoded in the model during its extensive pre-training phase, potentially leading to better performance compared to the fine-tuned model which reallocates model weights specifically toward the training dataset.

3. **Task Alignment with Pre-training:** The financial text analysis might align closely with the scenarios or data types the LLaMA-3 models were exposed to during pre-training, making the prompt engineering particularly effective. In contrast, fine-tuning might shift the model's focus away from these useful pre-learned patterns.

4. **Sensitivity to Hyperparameters:** Fine-tuning performance can be highly sensitive to the choice of hyperparameters, such as learning rates, batch sizes, and number of epochs. Inadequate or suboptimal hyperparameter settings during fine-tuning could lead to underperformance compared

to a more forgiving prompt-based approach that does not alter the model's weights as significantly.

These results provoke a reconsideration of the conventional training paradigms for LLMs in specific applications like financial text analysis, where prompt engineering might not only be more cost-effective but also more efficient in leveraging the underlying pre-trained models' strengths.

The findings underscore the importance of model scale in achieving optimal results and suggest that for certain tasks, especially those requiring nuanced understanding, larger models might still be necessary despite the advances in prompt engineering techniques.

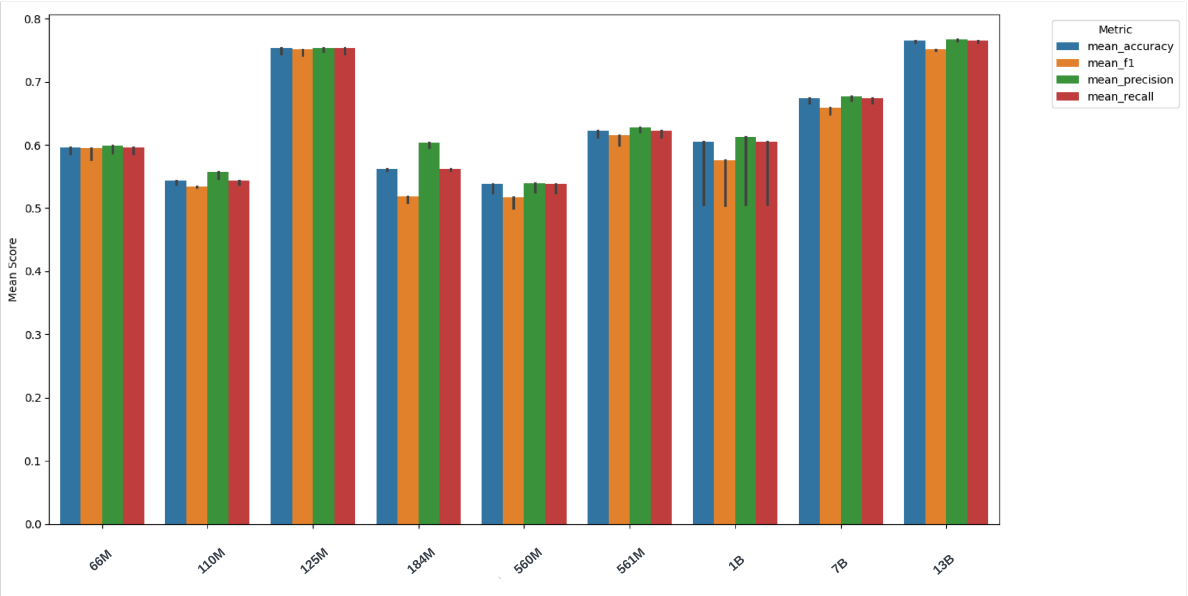

Figure 6.1: A grouped bar chart displaying the comparison of four metrics mean (accuracy, F1 score, precision, and recall) across models of various sizes.

Figure 6.2 indicates the performance of the three categories of open-source models we have experimented with. It reflects that Debate-fine-tuned and General-purpose models have a comparable mean macro F1-score, outperforming the Financial-fine-tuned models. This may suggest that general reasoning knowledge learned in debate-fine-tuned models is more valuable than the financial background knowledge learned in the Financial-fine-tuned models. Yet, the performance between Debate-fine-tuned models and General-Purpose Models is comparable, which could rely on the size of the latter. Therefore, we suggest examining smaller LLMs for a low tuning cost before looking for huger models, especially in a small dataset setting.

Figure 6.3, and the Pearson correlation heat map presented in Figure 6.4 provide an understanding of the relationship between hyperparameters and F1-score. Certain hyperparameters such as epochs and learning rate showed positive correlations with the F1-score. Potentially, since we give the model the chance to distil the pattern of our data, which means the more epochs we give to the model to train, the better the model learns.

Hyperparameters like maximum input length (max length), did not exhibit a very strong relationship with mean F1-score since most of the data points, as shown in Figure 6.5, are less than the smallest value of the max length hyperparameter ranging from (64 to 256) and the frequency of the examples that has 64 tokens or less is dominant. However, the correlation still exists which means the longer the sentence is fed to the model without truncation, the better performance the model achieves. However, a complex interplay between these hyperparameters requires careful tuning to optimize performance.

We also have noticed that the standard deviation, in general, is small which means the consistent performance of such models with low standard deviation, however, some models have a slightly larger standard deviation such as Roberta-base and *ArgumentMining-EN-AC-Financial*, One of the

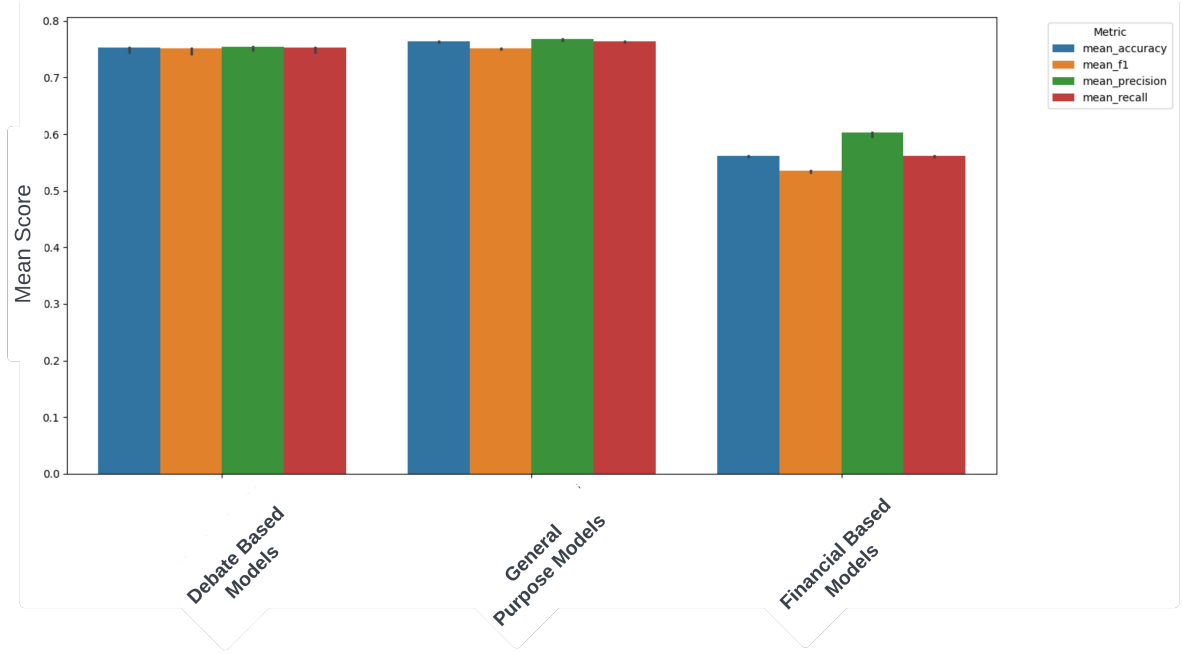

Figure 6.2: Performance among the three categories of fine-tuned models (Debate-fine-tuned, General-purpose, Financial-fine-tuned)

reasons could be the type of data these models fine-tuned on which made those models overfitted and stuck in a local minimum because of such past fine-tuning.

6.2 Categorical Variable Correlation

The Chi-Square test applied to explore the association between model architecture and model type yielded a Chi-Square statistic of approximately 64.96 with a p-value of around 1.03×10^{-15}. These results indicate a statistically significant association between the two categorical variables in the dataset. This suggests that different model architectures might be inherently linked to their types, influencing how each model performs under various conditions.

Spearman Rank Correlation Analysis Our analysis also included the Spearman rank correlation to measure the strength and direction of association between the mean F1 score and several continuous variables:

- **Runtime:** There is a strong positive correlation ($r = 0.630$), indicating that longer runtimes are generally associated with better model performance. This correlation suggests that models that take longer to train tend to achieve higher F1 scores, possibly due to more comprehensive learning processes.
- **Learning Rate (`lr`):** A moderate positive correlation ($r = 0.389$) with the mean F1 score implies that higher learning rates might facilitate quicker convergence to better performance metrics.
- **Epochs:** The correlation of $r = 0.316$ with the mean F1 score suggests that models trained for more epochs tend to yield better results, supporting the notion that extended training can improve model accuracy.
- **Max Length (`max_len`):** The weakest correlation observed ($r = 0.173$) between max length and the mean F1 score indicates a less direct impact of input size on model effectiveness.

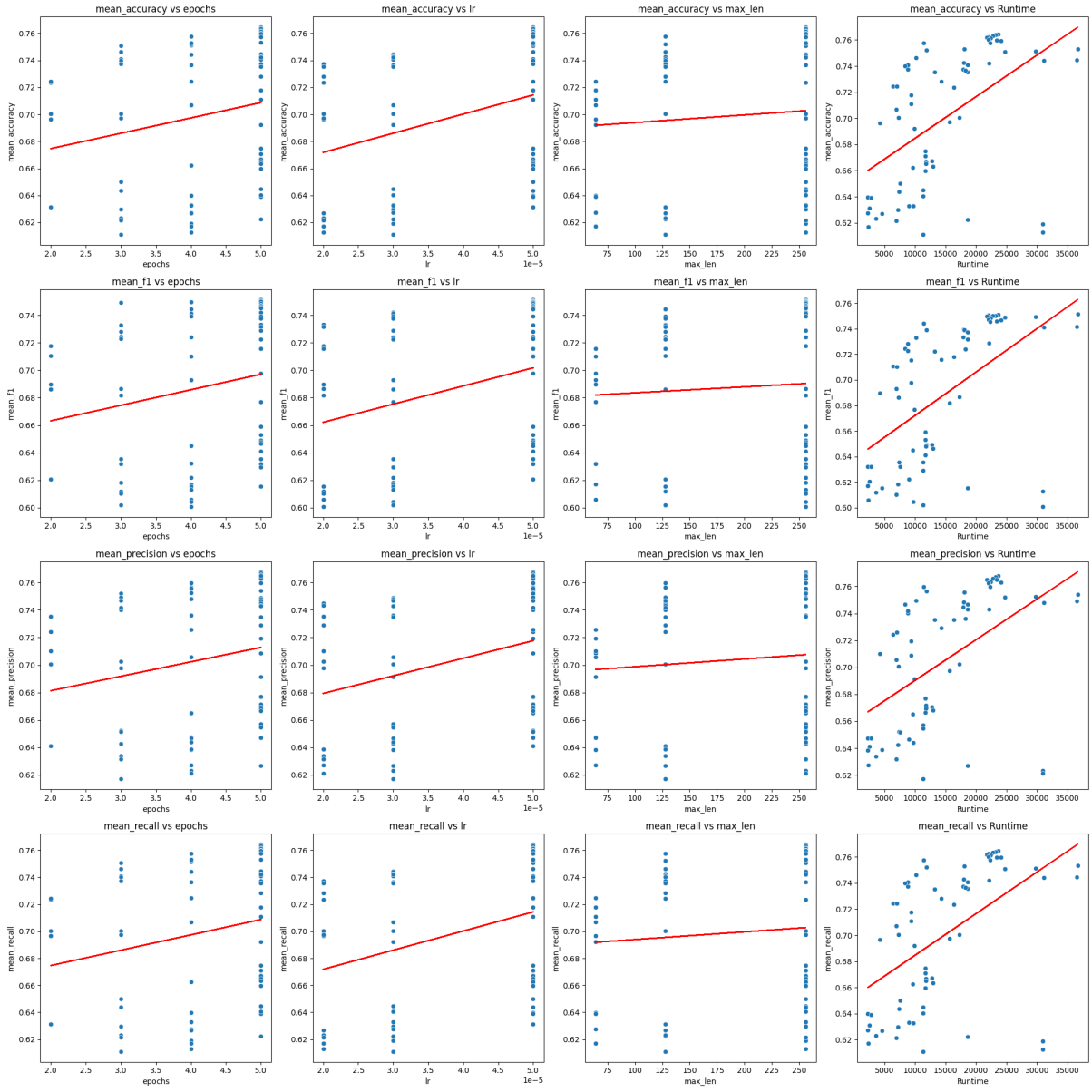

Figure 6.3: Correlation between hyperparameters (epochs, learning rate, input max length, runtime) and the performance metrics of fine-tuned models (accuracy, F1-score, precision, recall)

These Spearman correlation results, which focus on non-linear relationships, are precious for understanding the dynamics in ordinal or non-normally distributed data. The prominent correlation between runtime and F1 scores might warrant further investigation to determine if prolonged training benefits model accuracy or if other confounding factors play a role.

Interpretation of Results While the correlations provide insightful observations, it is critical to remember that correlation does not imply causation. The relationships observed in our analysis could be influenced by various external factors inherent to the models or the experimental setup. Therefore, these findings should be interpreted with an understanding of the complexities involved in model training and performance evaluation.

Further Research The findings suggest several avenues for further research, particularly in exploring the causal relationships that might explain the significant correlations between model run-

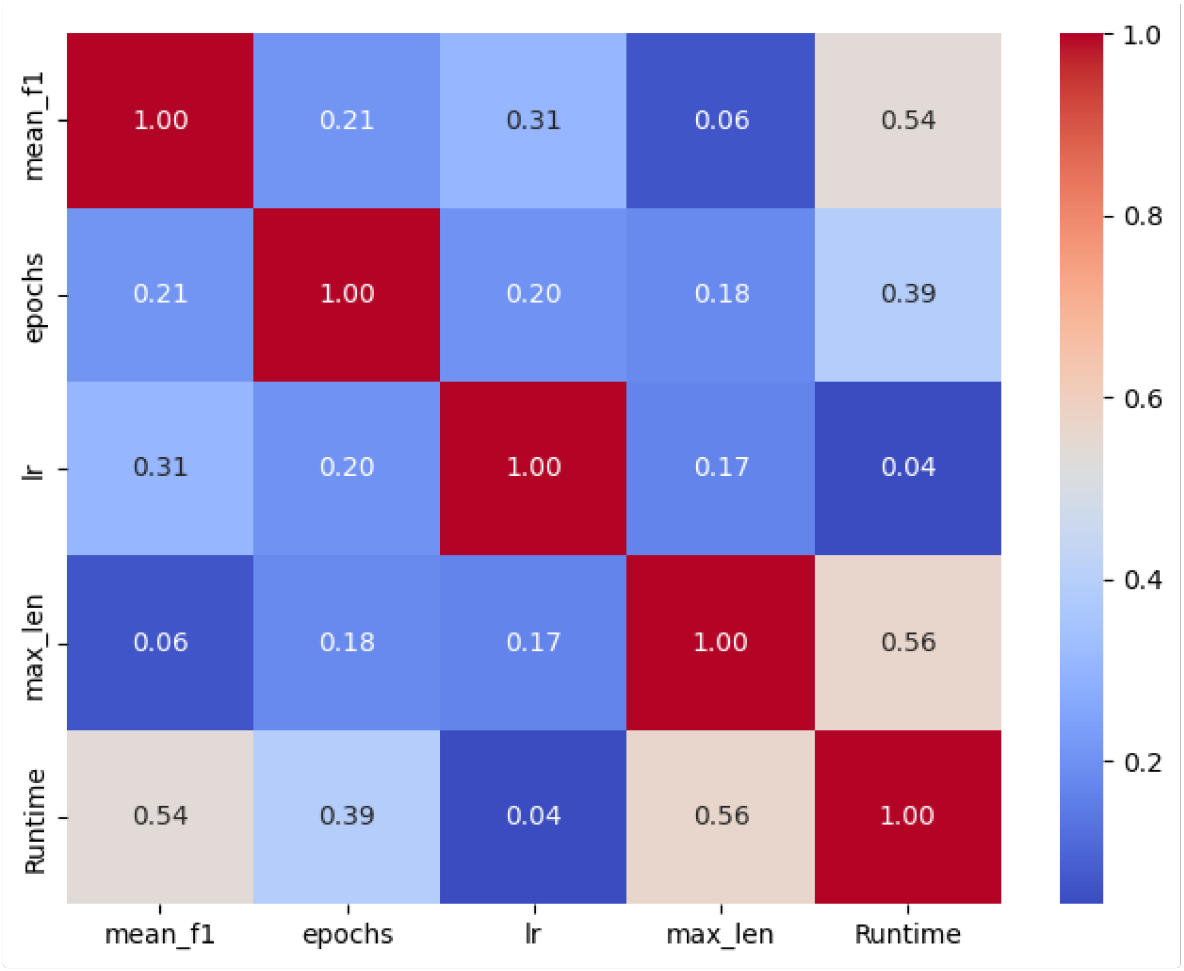

Figure 6.4: The heat map shows that learning rate and runtime, maximum input length and epochs correlation with mean F1-score.

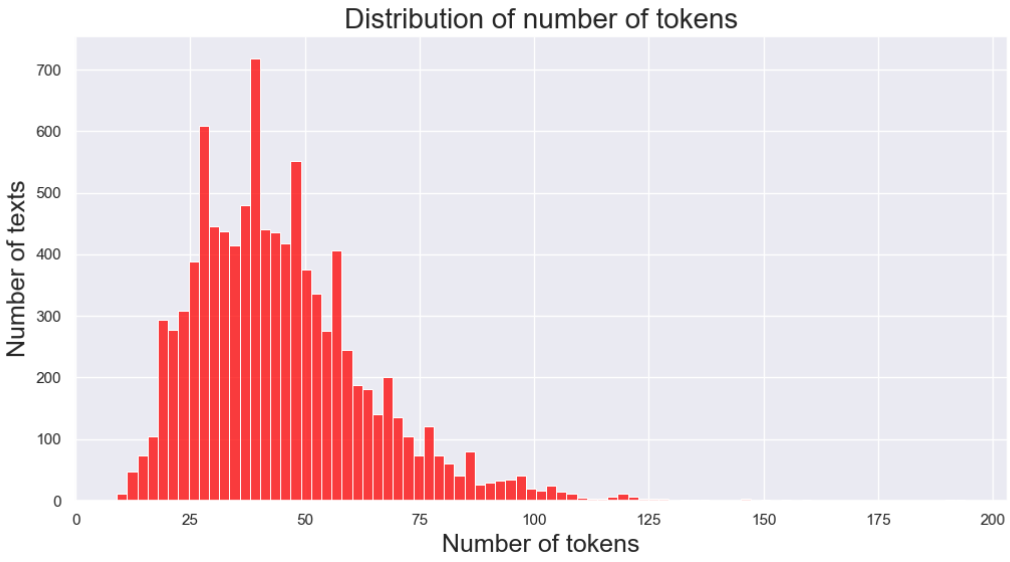

Figure 6.5: Distribution of sentence length

time, learning parameters, and performance outcomes. Experimentation with controlled variable adjustments could help in understanding the direct impacts of these factors on model efficiency and effectiveness.

6.3 GPU Consumption Analysis

GPU Memory Allocation Analysis The analysis of GPU memory allocation across different models, as depicted in Figure 6.6 and the interactive graph here[1] the models also show varying patterns. The *reciprocate/vicuna-13b_rm_oasst-hh* model exhibits moderate memory usage, with allocations around 40% on GPU 0 and 35% on GPU 1. The *lmsys/vicuna-13b-v1.5* model follows a similar trend with memory allocations of approximately 42% on GPU 0 and 38% on GPU 1. The *bigscience/bloom-7b1* model demonstrates higher memory allocation, with around 50% on GPU 0 and 48% on GPU 1, indicating a greater need for memory resources. The *meta-llama/Meta-Llama-3-8B* model shows significant memory allocation, with 55% on GPU 0 and 53% on GPU 1, reflecting its intensive computational requirements. These differences in memory allocation further emphasize the need for customized hardware configurations to support the specific demands of each model, ensuring optimal performance and cost-efficiency in deployment.

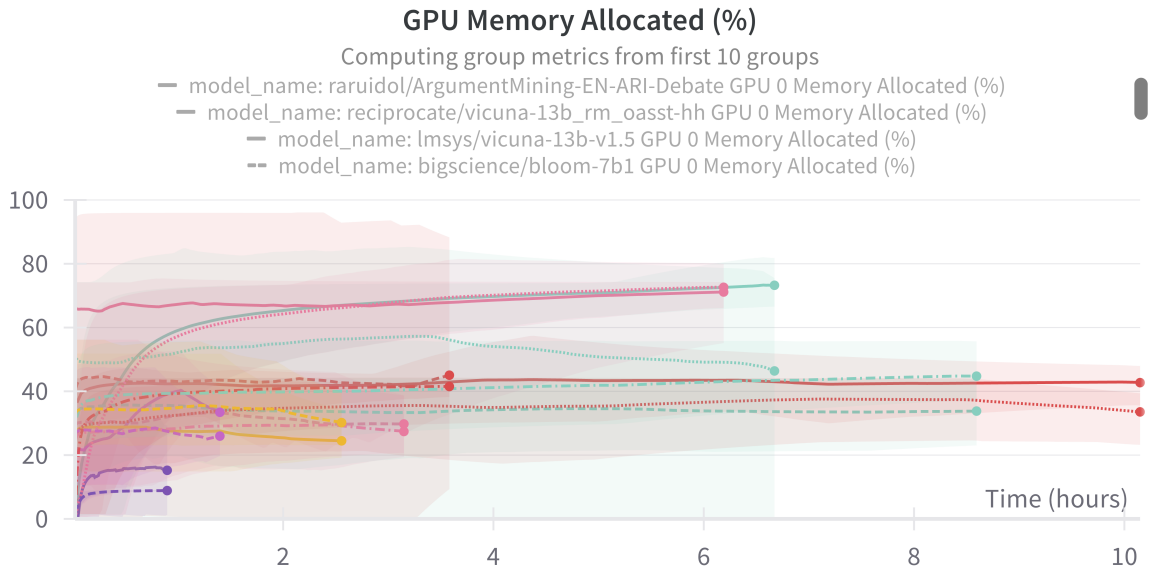

Figure 6.6: GPU Memory allocation for different models

GPU Utilization Trends The GPU utilization graph presented in Figure 6.7 and based on the additional interactive graph provided here[2], which includes models such as *reciprocate/vicuna-13b_rm_oasst-hh*, *lmsys/vicuna-13b-v1.5*, *bigscience/bloom-7b1*, and *meta-llama/Meta-Llama-3-8B*, we observe a diverse range of GPU utilization and memory allocation patterns. The *reciprocate/vicuna-13b_rm_oasst-hh* model shows moderate GPU utilization with peaks around 76% on GPU 0 and 57% on GPU 1, indicating balanced but not maximal resource usage. The *lmsys/vicuna-13b-v1.5* model exhibits a similar pattern with GPU utilization reaching around 76% for GPU 0 and 75% for GPU 1. In contrast, the *bigscience/bloom-7b1* model demonstrates a higher GPU utilization of approximately 69% on GPU 0 and 81% on GPU 1, suggesting more intensive computational demands. The *meta-llama/Meta-Llama-3-8B* model stands out with high utilization

[1] https://wandb.ai/master-thesis-uni-passau/master-thesis/reports/GPU-Memory-Allocated-24-05-16-10-49-15—Vmlldzo3OTcwNTk3?accessToken=w4nj3k8w3dc0dovq97bmkqk8eanpb5aekpi8p30d2y0h0gsxjvl4iegdqqh28f4p
[2] https://api.wandb.ai/links/master-thesis-uni-passau/d1netnte

rates, nearing 92% on GPU 0 and 97% on GPU 1, indicating a heavy load on the computational resources. These observations highlight the variability in GPU resource demands across different models, underlining the importance of tailored hardware provisioning to optimize performance and efficiency in large-scale machine learning applications.

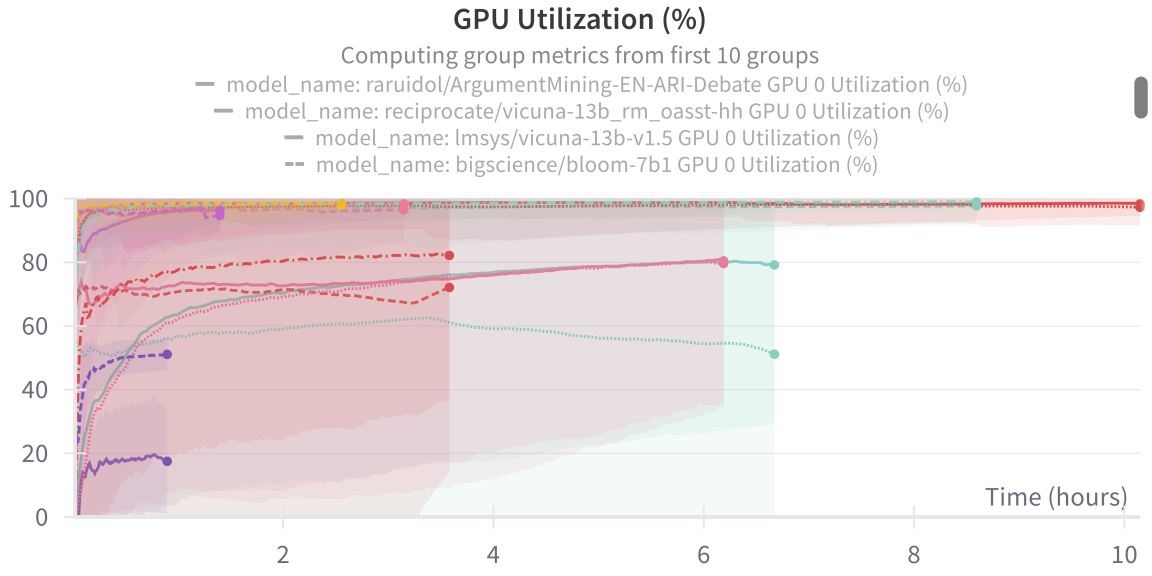

Figure 6.7: GPU Utilization graph to show utilization trend for different models

7 Conclusion

Summary of Findings Argument mining, the automatic extraction and analysis of argumentative structures in text, plays a crucial role across various fields such as assisted writing, fact-checking, and legal analysis. This thesis has focused particularly on the use of argument mining in financial texts, especially in the context of earnings conference calls. We assessed a range of Large Language Models (LLMs), including *GPT-4*, models fine-tuned for debate contexts, and those specialized for financial analysis. The results showed a wide variance in performance, with F1-scores ranging from 0.37 to 0.75. This variance highlights both the challenges and possibilities of using LLMs for domain-specific applications. Remarkably, *GPT-4* demonstrated excellent zero-shot learning capabilities, achieving an F1-score of 0.81, which underscores its effectiveness in handling complex financial language without prior specific training. *Vicuna-13b-v1.5* came in second place after *GPT-4* with a mean macro F1-score of 0.751 and the first place among all fine-tuned models.

Additionally, our exploration included the LLaMA-3 models, utilized in both their 8B and 70B configurations with 4-bit quantization, an approach that significantly reduces memory usage, facilitating deployment on more constrained platforms. Despite the reduction in precision typically associated with lower bit quantization, the LLaMA-3 Instruct 8B model achieved a macro F1-score of 71%, while the 70B variant registered 65%. These results are noteworthy as they were achieved using a prompt engineering approach with 1-shot learning, demonstrating the models' adaptability to specialized tasks with minimal input. However, when fine-tuned, the LLaMA-3 8B model, also utilizing 4-bit quantization, showed a lower mean macro F1-score of 63% under a 5-fold stratified cross-validation setup. This suggests that while 4-bit quantization offers advantages in terms of efficiency and resource utilization, it may also pose challenges in maintaining performance levels comparable to those achieved with higher precision models. This balance between efficiency and performance highlights the need for ongoing research into optimal training and deployment strategies for quantized models in domain-specific applications.

The analysis of GPU utilization, memory allocation, and model performance, specifically focusing on the F1 score, provides significant insights into the computational efficiency and effectiveness of various models.

Models such as *Vicuna-13b_rm_oasst-hh* and *ArgumentMining-EN-ARI-Debate* demonstrated high GPU utilization rates, nearing 100%, and correspondingly high F1 scores of 0.751. This suggests a positive correlation between high GPU utilization and improved model performance. These models efficiently utilize available computational resources, translating into better predictive capabilities as indicated by their superior F1 scores.

Conversely, models like *Bloom-7b1* and *meta-llama/Meta-Llama-3-8B*, which also showed substantial GPU utilization, had lower F1 scores of 0.659 and 0.638, respectively. This indicates that while GPU utilization is a crucial factor, it alone does not guarantee superior performance. The architecture and specific training data of the models play a significant role in determining their overall effectiveness.

In terms of GPU memory allocation, models such as *meta-llama/Meta-Llama-3-8B* and *bigscience/bloom-7b1* exhibited high memory usage but varied in their performance outcomes. The *meta-llama/Meta-Llama-3-8B* model allocated significant memory resources but achieved a moderate F1 score of 0.638. On the other hand, the *ArgumentMining-EN-ARI-Debate* model balanced its memory usage effectively and attained a high F1 score, indicating an optimal use of memory resources contributes to better performance.

Models with lower memory allocation, like *FinancialBERT-Sentiment-Analysis*, showed lower F1 scores, reflecting a potential limitation in handling complex tasks due to restricted memory resources. However, models with moderate memory usage and high F1 scores, such as *Vicuna-13b_rm_oasst-hh*, demonstrate that efficient memory management is vital for enhancing model performance without necessitating maximal memory allocation.

Overall, the findings suggest that while high GPU utilization and adequate memory allocation are important for achieving high performance, the specific architecture and training data of the models are also critical determinants. Efficient use of computational resources, combined with robust model design, leads to better predictive performance, as evidenced by the F1 scores. Therefore, for optimal performance, a balanced approach considering both computational resource allocation and model architecture is essential.

Implications for Practice The impressive performance of *GPT-4* suggests its significant potential to transform financial analysis, enhancing decision-making, streamlining financial reporting, and improving the precision of automated financial systems.

However, the variation in model effectiveness also emphasizes the importance of choosing the appropriate model for specific tasks. While *GPT-4* is highly effective in a zero-shot learning environment, other models may be preferable when fine-tuned with relevant data, indicating the need for a strategic approach to model selection depending on the task's requirements.

The comparative analysis of the LLaMA-3 models with different configurations and prompt selection strategies reveals critical insights into their performance and optimization potential. The LLaMA-3 8B model using a random 1-shot prompt demonstrated a balanced yet moderate performance, while the LLaMA-3 70B model highlighted the challenges of class identification, particularly for the 'Related' class, with a notably low recall. Conversely, the ensemble approach for 1-shot learning showed significant improvements, particularly in the LLaMA-3 8B model, achieving higher overall accuracy and more balanced precision and recall metrics. The 4-bit quantization method proved effective in maintaining model performance while reducing computational demands, underscoring its value in resource-constrained settings. These findings emphasize the importance of thoughtful prompt selection and fine-tuning strategies to enhance classification performance, suggesting that combining ensemble approaches and quantization techniques can lead to more efficient and robust models.

Limitations and Future Research This study's findings come with certain limitations, primarily the inconsistent performance across models, which suggests a need for further research to uncover factors that enhance argument mining in financial contexts. Also, despite *GPT-4*'s strong performance, its high computational demands and accessibility are challenging for widespread use, particularly in settings with limited resources.

Looking ahead, there are several promising avenues for extending the research conducted in this thesis. First, exploring the impact of using more than one shot in the training process could potentially enhance the models' understanding and classification abilities. Increasing the number of shots might provide the models with a richer context and a broader spectrum of examples, which could lead to better generalization across unseen data. This approach would be particularly useful in complex tasks such as argument mining where nuances in text can significantly impact performance. Additionally, experimenting with a variety of models beyond those currently tested could uncover more insights into the optimal architectures for specific aspects of financial discourse analysis.

Furthermore, refining the fine-tuning process by adjusting the number of trainable layers offers another promising direction. In our initial experiments, most layers were frozen to streamline the training process and focus learning in specific areas of the model. However, allowing more layers to adjust during training could lead to a deeper and more nuanced adaptation to the financial domain, potentially improving the models' ability to capture subtle semantic and syntactic features relevant to argument detection.

Moreover, innovating in prompt generation represents a critical area for future research. Developing diverse methods to automatically generate or augment prompts could lead to significant improvements in model performance, especially in zero-shot or few-shot settings. Techniques such as automated paraphrasing, adversarial prompt generation, or using reinforcement learning to optimize prompt strategies could provide models with more effective and varied inputs, enhancing their adaptability and accuracy.

Future studies should look into improving the efficiency of these models, possibly by developing lighter versions that maintain high effectiveness with lower computational demands. Exploring hybrid models that blend zero-shot and fine-tuning approaches might also lead to more adaptable and powerful systems.

To further enhance our understanding of the experiment and provide better explanations for the results and outcomes of the comparisons made, it is crucial to apply interpretability and explainability tools. These tools can help uncover the underlying mechanisms of model behaviour and performance, offering deeper insights into the factors driving the observed metrics. Future work should focus on employing methods such as SHAP (SHapley Additive exPlanations) [102], LIME (Local Interpretable Model-agnostic Explanations) [103], and model-specific interpretability techniques like attention visualization for transformer-based models [104]. These methods will facilitate a more comprehensive understanding of model decisions, highlight important features influencing predictions, and identify potential biases or weaknesses within the models. Integrating these interpretability tools will not only improve transparency but also aid in refining model development and optimization strategies, ultimately leading to more robust and reliable machine learning applications.

By pursuing these lines of inquiry, future research can build on the foundational work presented in this thesis, pushing the boundaries of what is achievable with large language models in financial argument mining and potentially setting new standards for automated text analysis in the financial sector.

Expanding the Scope of Argument Mining Future research could expand argument mining's scope to more detailed analyses, such as distinguishing different types of financial arguments or combining textual data with quantitative financial metrics. This expansion could enhance the tools available to financial analysts significantly.

Moreover, applying these technologies to real-time scenarios, like during live earnings calls, could fundamentally change how financial information is processed, providing immediate insights and supporting swift decision-making.

Concluding Remarks In sum, this thesis not only showcases the current capabilities of LLMs in navigating the intricate world of financial argument mining but also lays the groundwork for future innovations that could further enhance this field. The potential for these technologies to improve financial understanding and decision-making is immense, promising not only more advanced tools but also a deeper, more actionable comprehension of financial narratives.

As we continue to advance and refine LLMs, their integration into regular financial analysis and decision-making processes appears both feasible and likely, marking the onset of a new era in AI-driven financial insight.

Appendix

model_name	lr	max_len	epoch	model_size	model_architecture	model_type	mean_accuracy	mean_f1	mean_precision	mean_recall	std_accuracy	std_f1	std_precision	std_recall
raruidol/ArgumentMining-EN-ARI-Debate	5.00E-05	256	5	561000000	encoder	debate_based_models	0.7532	0.7514	0.7539	0.7532	0.0136	0.0137	0.0136	0.0136
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7645	0.7510	0.7678	0.7645	0.0331	0.0358	0.0301	0.0331
lmsys/vicuna-13b-v1.5	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7625	0.7505	0.7623	0.7625	0.0328	0.0336	0.0330	0.0328
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7639	0.7504	0.7669	0.7639	0.0362	0.0391	0.0348	0.0362
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7632	0.7502	0.7658	0.7632	0.0365	0.0380	0.0326	0.0365
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7618	0.7499	0.7649	0.7618	0.0373	0.0385	0.0339	0.0373
raruidol/ArgumentMining-EN-ARI-Debate	5.00E-05	256	4	561000000	encoder	debate_based_models	0.7513	0.7495	0.7526	0.7513	0.0110	0.0112	0.0114	0.0110
raruidol/ArgumentMining-EN-ARI-Debate	5.00E-05	256	3	561000000	encoder	debate_based_models	0.7508	0.7491	0.7520	0.7508	0.0129	0.0132	0.0130	0.0129
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7614	0.7484	0.7636	0.7614	0.0385	0.0402	0.0367	0.0385
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7602	0.7473	0.7628	0.7602	0.0350	0.0371	0.0335	0.0350
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7595	0.7466	0.7629	0.7595	0.0367	0.0391	0.0343	0.0367
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7597	0.7461	0.7649	0.7597	0.0361	0.0384	0.0329	0.0361
lmsys/vicuna-13b-v1.5	5.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7575	0.7454	0.7597	0.7575	0.0315	0.0318	0.0323	0.0315
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	128	4	13000000000	encoder_decoder	general_purpose_models	0.7577	0.7444	0.7597	0.7577	0.0337	0.0362	0.0325	0.0337
raruidol/ArgumentMining-EN-ARI-Debate	3.00E-05	256	5	561000000	encoder	debate_based_models	0.7447	0.7419	0.7490	0.7447	0.0150	0.0165	0.0109	0.0150
raruidol/ArgumentMining-EN-ARI-Debate	3.00E-05	256	4	561000000	encoder	debate_based_models	0.7442	0.7413	0.7481	0.7442	0.0187	0.0204	0.0133	0.0187
raruidol/ArgumentMining-EN-ARI-Debate	3.00E-05	128	5	561000000	encoder	debate_based_models	0.7427	0.7394	0.7482	0.7427	0.0192	0.0212	0.0142	0.0192
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	128	4	13000000000	encoder_decoder	general_purpose_models	0.7521	0.7391	0.7565	0.7521	0.0390	0.0405	0.0365	0.0390
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	256	4	13000000000	encoder_decoder	general_purpose_models	0.7530	0.7390	0.7557	0.7530	0.0387	0.0411	0.0371	0.0387
raruidol/ArgumentMining-EN-ARI-Debate	3.00E-05	128	5	561000000	encoder	debate_based_models	0.7408	0.7375	0.7469	0.7408	0.0182	0.0210	0.0111	0.0182
raruidol/ArgumentMining-EN-ARI-Debate	2.00E-05	128	5	561000000	encoder	debate_based_models	0.7373	0.7334	0.7448	0.7373	0.0198	0.0218	0.0151	0.0198
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	128	3	13000000000	encoder_decoder	general_purpose_models	0.7461	0.7330	0.7495	0.7461	0.0338	0.0354	0.0306	0.0338
raruidol/ArgumentMining-EN-ARI-Debate	2.00E-05	128	5	561000000	encoder	debate_based_models	0.7355	0.7317	0.7432	0.7355	0.0174	0.0194	0.0132	0.0174
lmsys/vicuna-13b-v1.5	3.00E-05	256	5	13000000000	encoder_decoder	general_purpose_models	0.7421	0.7288	0.7430	0.7421	0.0280	0.0300	0.0281	0.0280
lmsys/vicuna-13b-v1.5	5.00E-05	128	3	13000000000	encoder_decoder	general_purpose_models	0.7409	0.7282	0.7402	0.7409	0.0335	0.0338	0.0330	0.0335
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	128	3	13000000000	encoder_decoder	general_purpose_models	0.7400	0.7246	0.7468	0.7400	0.0395	0.0435	0.0331	0.0395
lmsys/vicuna-13b-v1.5	3.00E-05	256	4	13000000000	encoder_decoder	general_purpose_models	0.7365	0.7239	0.7362	0.7365	0.0221	0.0231	0.0232	0.0221
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	128	3	13000000000	encoder_decoder	general_purpose_models	0.7373	0.7227	0.7417	0.7373	0.0366	0.0395	0.0334	0.0366
lmsys/vicuna-13b-v1.5	3.00E-05	128	5	13000000000	encoder_decoder	general_purpose_models	0.7355	0.7225	0.7351	0.7355	0.0374	0.0381	0.0384	0.0374
raruidol/ArgumentMining-EN-ARI-Debate	2.00E-05	256	2	561000000	encoder	debate_based_models	0.7235	0.7178	0.7353	0.7235	0.0206	0.0234	0.0154	0.0206
lmsys/vicuna-13b-v1.5	2.00E-05	128	5	13000000000	encoder_decoder	general_purpose_models	0.7282	0.7157	0.7291	0.7282	0.0263	0.0280	0.0275	0.0263
raruidol/ArgumentMining-EN-ARI-Debate	5.00E-05	64	5	561000000	encoder	debate_based_models	0.7177	0.7155	0.7194	0.7177	0.0045	0.0048	0.0044	0.0045
lmsys/vicuna-13b-v1.5	5.00E-05	128	2	13000000000	encoder_decoder	general_purpose_models	0.7243	0.7106	0.7243	0.7243	0.0299	0.0309	0.0310	0.0299
reciprocate/vicuna-13b_rm_oasst-hh	5.00E-05	64	4	13000000000	encoder_decoder	general_purpose_models	0.7245	0.7102	0.7258	0.7245	0.0512	0.0521	0.0511	0.0512
lmsys/vicuna-13b-v1.5	5.00E-05	64	5	13000000000	encoder_decoder	general_purpose_models	0.7110	0.6977	0.7087	0.7110	0.0558	0.0558	0.0558	0.0558
reciprocate/vicuna-13b_rm_oasst-hh	3.00E-05	64	4	13000000000	encoder_decoder	general_purpose_models	0.7070	0.6930	0.7057	0.7070	0.0485	0.0490	0.0478	0.0485
raruidol/ArgumentMining-EN-ARI-Debate	2.00E-05	64	2	561000000	encoder	debate_based_models	0.6965	0.6896	0.7101	0.6965	0.0127	0.0165	0.0054	0.0127
lmsys/vicuna-13b-v1.5	2.00E-05	256	3	13000000000	encoder_decoder	general_purpose_models	0.7005	0.6866	0.7025	0.7005	0.0282	0.0301	0.0286	0.0282
lmsys/vicuna-13b-v1.5	3.00E-05	128	2	13000000000	encoder_decoder	general_purpose_models	0.7003	0.6863	0.7008	0.7003	0.0331	0.0343	0.0337	0.0331
lmsys/vicuna-13b-v1.5	2.00E-05	256	3	13000000000	encoder_decoder	general_purpose_models	0.6973	0.6818	0.6976	0.6973	0.0291	0.0305	0.0284	0.0291
lmsys/vicuna-13b-v1.5	3.00E-05	64	5	13000000000	encoder_decoder	general_purpose_models	0.6922	0.6769	0.6915	0.6922	0.0500	0.0501	0.0486	0.0500
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6749	0.6590	0.6767	0.6749	0.0328	0.0331	0.0329	0.0328
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6709	0.6530	0.6771	0.6709	0.0366	0.0367	0.0395	0.0366
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6663	0.6492	0.6702	0.6663	0.0361	0.0359	0.0400	0.0361
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6672	0.6492	0.6707	0.6672	0.0387	0.0393	0.0419	0.0387
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6669	0.6489	0.6717	0.6669	0.0421	0.0403	0.0458	0.0421
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6651	0.6479	0.6695	0.6651	0.0314	0.0317	0.0345	0.0314
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6633	0.6465	0.6682	0.6633	0.0413	0.0416	0.0448	0.0413
bigscience/bloom-7b1	5.00E-05	256	4	7000000000	decoder	general_purpose_models	0.6624	0.6449	0.6652	0.6624	0.0388	0.0375	0.0416	0.0388
bigscience/bloom-7b1	5.00E-05	256	5	7000000000	decoder	general_purpose_models	0.6596	0.6410	0.6667	0.6596	0.0417	0.0413	0.0452	0.0417
chkla/roberta-argument	3.00E-05	256	5	125000000	encoder	debate_based_models	0.6449	0.6356	0.6569	0.6449	0.0232	0.0268	0.0215	0.0232
chkla/roberta-argument	5.00E-05	256	3	125000000	encoder	debate_based_models	0.6437	0.6355	0.6523	0.6437	0.0245	0.0338	0.0153	0.0245
chkla/roberta-argument	5.00E-05	64	4	125000000	encoder	debate_based_models	0.6397	0.6323	0.6475	0.6397	0.0223	0.0285	0.0181	0.0223
bigscience/bloom-7b1	5.00E-05	256	3	7000000000	decoder	general_purpose_models	0.6501	0.6320	0.6516	0.6501	0.0377	0.0365	0.0373	0.0377
chkla/roberta-argument	5.00E-05	64	5	125000000	encoder	debate_based_models	0.6391	0.6319	0.6472	0.6391	0.0189	0.0265	0.0121	0.0189
chkla/roberta-argument	3.00E-05	256	5	125000000	encoder	debate_based_models	0.6405	0.6293	0.6549	0.6405	0.0271	0.0321	0.0251	0.0271
chkla/roberta-argument	3.00E-05	256	4	125000000	encoder	debate_based_models	0.6331	0.6220	0.6466	0.6331	0.0236	0.0286	0.0210	0.0236
chkla/roberta-argument	5.00E-05	128	2	125000000	encoder	debate_based_models	0.6313	0.6205	0.6411	0.6313	0.0309	0.0433	0.0221	0.0309
chkla/roberta-argument	3.00E-05	256	3	125000000	encoder	debate_based_models	0.6298	0.6182	0.6426	0.6298	0.0274	0.0340	0.0237	0.0274
chkla/roberta-argument	3.00E-05	64	4	125000000	encoder	debate_based_models	0.6274	0.6173	0.6385	0.6274	0.0253	0.0301	0.0224	0.0253
raruidol/ArgumentMining-EN-AC-Essay-Fin	3.00E-05	128	5	561000000	encoder	debate_based_models	0.6222	0.6154	0.6269	0.6222	0.0249	0.0316	0.0192	0.0249
chkla/roberta-argument	2.00E-05	128	4	125000000	encoder	debate_based_models	0.6269	0.6154	0.6387	0.6269	0.0296	0.0359	0.0262	0.0296
raruidol/ArgumentMining-EN-AC-Essay-Fin	3.00E-05	256	4	561000000	encoder	debate_based_models	0.6191	0.6129	0.6232	0.6191	0.0198	0.0259	0.0143	0.0198
chkla/roberta-argument	2.00E-05	128	3	125000000	encoder	debate_based_models	0.6233	0.6120	0.6338	0.6233	0.0314	0.0382	0.0282	0.0314
chkla/roberta-argument	2.00E-05	256	3	125000000	encoder	debate_based_models	0.6215	0.6102	0.6317	0.6215	0.0310	0.0378	0.0276	0.0310
chkla/roberta-argument	2.00E-05	64	4	125000000	encoder	debate_based_models	0.6171	0.6058	0.6273	0.6171	0.0269	0.0339	0.0232	0.0269
bigscience/bloom-7b1	3.00E-05	256	4	7000000000	decoder	general_purpose_models	0.6327	0.6045	0.6439	0.6327	0.0456	0.0606	0.0357	0.0456
raruidol/ArgumentMining-EN-AC-Essay-Fin	3.00E-05	128	3	561000000	encoder	debate_based_models	0.6111	0.6020	0.6170	0.6111	0.0241	0.0360	0.0161	0.0241
raruidol/ArgumentMining-EN-AC-Essay-Fin	2.00E-05	256	4	561000000	encoder	debate_based_models	0.6128	0.6007	0.6212	0.6128	0.0292	0.0462	0.0177	0.0292
bigscience/bloom-7b1	5.00E-05	256	2	7000000000	decoder	general_purpose_models	0.6174	0.5997	0.6214	0.6174	0.0392	0.0380	0.0396	0.0392
raruidol/ArgumentMining-EN-AC-Essay-Fin	3.00E-05	128	3	561000000	encoder	debate_based_models	0.6090	0.5993	0.6151	0.6090	0.0247	0.0378	0.0151	0.0247
raruidol/ArgumentMining-EN-AC-Essay-Fin	5.00E-05	256	3	561000000	encoder	debate_based_models	0.6132	0.5979	0.6278	0.6132	0.0222	0.0354	0.0101	0.0222
raruidol/ArgumentMining-EN-AC-Essay-Fin	2.00E-05	256	4	561000000	encoder	debate_based_models	0.6108	0.5979	0.6203	0.6108	0.0287	0.0446	0.0176	0.0287
distilbert-base-uncased	5.00E-05	128	5	66000000	encoder	general_purpose_models	0.5964	0.5951	0.5971	0.5964	0.0141	0.0138	0.0146	0.0141
raruidol/ArgumentMining-EN-AC-Essay-Fin	3.00E-05	128	2	561000000	encoder	debate_based_models	0.6026	0.5933	0.6080	0.6026	0.0234	0.0343	0.0165	0.0234
raruidol/ArgumentMining-EN-AC-Essay-Fin	3.00E-05	256	2	561000000	encoder	debate_based_models	0.6014	0.5910	0.6075	0.6014	0.0262	0.0400	0.0174	0.0262
raruidol/ArgumentMining-EN-AC-Essay-Fin	5.00E-05	128	2	561000000	encoder	debate_based_models	0.6059	0.5906	0.6187	0.6059	0.0247	0.0403	0.0126	0.0247
raruidol/ArgumentMining-EN-AC-Essay-Fin	2.00E-05	64	5	561000000	encoder	debate_based_models	0.5999	0.5886	0.6066	0.5999	0.0268	0.0406	0.0196	0.0268
distilbert-base-uncased	5.00E-05	128	3	66000000	encoder	general_purpose_models	0.5955	0.5884	0.5991	0.5955	0.0247	0.0315	0.0211	0.0247
raruidol/ArgumentMining-EN-AC-Essay-Fin	3.00E-05	128	2	561000000	encoder	debate_based_models	0.6002	0.5876	0.6075	0.6002	0.0280	0.0452	0.0177	0.0280
raruidol/ArgumentMining-EN-AC-Essay-Fin	5.00E-05	64	3	561000000	encoder	debate_based_models	0.6031	0.5870	0.6176	0.6031	0.0207	0.0325	0.0134	0.0207
raruidol/ArgumentMining-EN-AC-Essay-Fin	5.00E-05	64	2	561000000	encoder	debate_based_models	0.5999	0.5825	0.6161	0.5999	0.0210	0.0297	0.0165	0.0210
distilbert-base-uncased	3.00E-05	64	4	66000000	encoder	general_purpose_models	0.5859	0.5777	0.5907	0.5859	0.0197	0.0246	0.0176	0.0197
bigscience/bloomz-1b1	3.00E-05	256	5	1000000000	decoder	general_purpose_models	0.6046	0.5765	0.6125	0.6046	0.0440	0.0532	0.0435	0.0440
chkla/roberta-argument	2.00E-05	64	2	125000000	encoder	debate_based_models	0.5875	0.5763	0.5933	0.5875	0.0408	0.0470	0.0409	0.0408
distilbert-base-uncased	2.00E-05	64	4	66000000	encoder	general_purpose_models	0.5826	0.5748	0.5872	0.5826	0.0179	0.0209	0.0172	0.0179
distilbert-base-uncased	2.00E-05	128	3	66000000	encoder	general_purpose_models	0.5802	0.5740	0.5837	0.5802	0.0113	0.0140	0.0107	0.0113
distilbert-base-uncased	2.00E-05	256	3	66000000	encoder	general_purpose_models	0.5787	0.5723	0.5825	0.5787	0.0143	0.0173	0.0125	0.0143
bigscience/bloom-7b1	2.00E-05	128	4	7000000000	decoder	general_purpose_models	0.5887	0.5716	0.5887	0.5887	0.0467	0.0470	0.0457	0.0467
distilbert-base-uncased	3.00E-05	256	3	66000000	encoder	general_purpose_models	0.5813	0.5710	0.5875	0.5813	0.0200	0.0275	0.0159	0.0200
pheinisch/roberta-base-150T-argumentative-sentence-detector	5.00E-05	128	5	125000000	encoder	debate_based_models	0.5788	0.5692	0.5848	0.5788	0.0167	0.0177	0.0175	0.0167
bigscience/bloom-7b1	2.00E-05	128	3	7000000000	decoder	general_purpose_models	0.5860	0.5684	0.5854	0.5860	0.0352	0.0364	0.0346	0.0352
pheinisch/roberta-base-150T-argumentative-sentence-detector	2.00E-05	256	2	125000000	encoder	debate_based_models	0.5769	0.5682	0.5818	0.5769	0.0157	0.0168	0.0159	0.0157
pheinisch/roberta-base-150T-argumentative-sentence-detector	3.00E-05	256	4	125000000	encoder	debate_based_models	0.5774	0.5679	0.5833	0.5774	0.0160	0.0173	0.0169	0.0160
pheinisch/roberta-base-150T-argumentative-sentence-detector	5.00E-05	128	4	125000000	encoder	debate_based_models	0.5767	0.5678	0.5820	0.5767	0.0161	0.0161	0.0178	0.0161
pheinisch/roberta-base-150T-argumentative-sentence-detector	2.00E-05	128	3	125000000	encoder	debate_based_models	0.5758	0.5673	0.5814	0.5758	0.0196	0.0184	0.0217	0.0196
pheinisch/roberta-base-150T-argumentative-sentence-detector	5.00E-05	128	2	125000000	encoder	debate_based_models	0.5757	0.5671	0.5803	0.5757	0.0169	0.0184	0.0176	0.0169
pheinisch/roberta-base-150T-argumentative-sentence-detector	5.00E-05	256	3	125000000	encoder	debate_based_models	0.5759	0.5671	0.5812	0.5759	0.0175	0.0172	0.0190	0.0175
bigscience/bloomz-1b1	3.00E-05	256	5	1000000000	decoder	general_purpose_models	0.5923	0.5665	0.6001	0.5923	0.0371	0.0431	0.0369	0.0371
distilbert-base-uncased	5.00E-05	128	2	66000000	encoder	general_purpose_models	0.5789	0.5651	0.5852	0.5789	0.0251	0.0411	0.0180	0.0251
raruidol/ArgumentMining-EN-AC-Essay-Fin	2.00E-05	64	3	561000000	encoder	debate_based_models	0.5722	0.5646	0.5776	0.5722	0.0276	0.0288	0.0287	0.0276
pheinisch/roberta-base-150T-argumentative-sentence-detector	2.00E-05	128	5	125000000	encoder	debate_based_models	0.5753	0.5631	0.5832	0.5753	0.0160	0.0176	0.0173	0.0160
raruidol/ArgumentMining-EN-AC-Essay-Fin	2.00E-05	64	2	561000000	encoder	debate_based_models	0.5830	0.5626	0.5961	0.5830	0.0238	0.0471	0.0116	0.0238
pheinisch/roberta-base-150T-argumentative-sentence-detector	5.00E-05	64	3	125000000	encoder	debate_based_models	0.5696	0.5600	0.5752	0.5696	0.0134	0.0109	0.0159	0.0134
pheinisch/roberta-base-150T-argumentative-sentence-detector	2.00E-05	64	5	125000000	encoder	debate_based_models	0.5679	0.5578	0.5737	0.5679	0.0180	0.0173	0.0206	0.0180
pheinisch/roberta-base-150T-argumentative-sentence-detector	3.00E-05	64	3	125000000	encoder	debate_based_models	0.5668	0.5569	0.5722	0.5668	0.0153	0.0162	0.0169	0.0153
distilbert-base-uncased	2.00E-05	128	2	66000000	encoder	general_purpose_models	0.5677	0.5555	0.5743	0.5677	0.0148	0.0245	0.0101	0.0148
chavinlo/gpt4-x-alpaca	3.00E-05	128	5	13000000000	decoder	general_purpose_models	0.5763	0.5537	0.5794	0.5763	0.0333	0.0357	0.0335	0.0333
pheinisch/roberta-base-150T-argumentative-sentence-detector	5.00E-05	64	5	125000000	encoder	debate_based_models	0.5665	0.5522	0.5767	0.5665	0.0139	0.0112	0.0176	0.0139
chavinlo/gpt4-x-alpaca	3.00E-05	256	5	13000000000	decoder	general_purpose_models	0.5710	0.5492	0.5735	0.5710	0.0354	0.0383	0.0355	0.0354
chavinlo/gpt4-x-alpaca	3.00E-05	128	5	13000000000	decoder	general_purpose_models	0.5682	0.5484	0.5707	0.5682	0.0375	0.0389	0.0393	0.0375
chavinlo/gpt4-x-alpaca	2.00E-05	128	5	13000000000	decoder	general_purpose_models	0.5646	0.5472	0.5642	0.5646	0.0349	0.0326	0.0350	0.0349
chavinlo/gpt4-x-alpaca	3.00E-05	64	3	13000000000	decoder	general_purpose_models	0.5651	0.5454	0.5658	0.5651	0.0261	0.0279	0.0265	0.0261
chavinlo/gpt4-x-alpaca	3.00E-05	256	5	13000000000	decoder	general_purpose_models	0.5671	0.5450	0.5696	0.5671	0.0432	0.0454	0.0418	0.0432
chavinlo/gpt4-x-alpaca	2.00E-05	256	5	13000000000	decoder	general_purpose_models	0.5633	0.5434	0.5649	0.5633	0.0272	0.0302	0.0277	0.0272
roberta-base	2.00E-05	128	3	125000000	encoder	general_purpose_models	0.5604	0.5385	0.5743	0.5604	0.0117	0.0239	0.0140	0.0117
chavinlo/gpt4-x-alpaca	5.00E-05	64	2	13000000000	decoder	general_purpose_models	0.5563	0.5384	0.5565	0.5563	0.0322	0.0320	0.0322	0.0322
roberta-base	2.00E-05	64	3	125000000	encoder	general_purpose_models	0.5628	0.5362	0.5728	0.5628	0.0313	0.0587	0.0231	0.0313
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5382	0.5350	0.5388	0.5382	0.0193	0.0183	0.0200	0.0193
bigscience/bloomz-7b1	2.00E-05	256	2	7000000000	decoder	general_purpose_models	0.5671	0.5347	0.5732	0.5671	0.0271	0.0374	0.0255	0.0271
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5370	0.5346	0.5371	0.5370	0.0097	0.0092	0.0098	0.0097
chavinlo/gpt4-x-alpaca	2.00E-05	256	5	13000000000	decoder	general_purpose_models	0.5516	0.5333	0.5506	0.5516	0.0319	0.0309	0.0310	0.0319
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5365	0.5332	0.5370	0.5365	0.0103	0.0100	0.0107	0.0103
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	128	4	110000000	encoder	financial_base_models	0.5366	0.5328	0.5373	0.5366	0.0106	0.0096	0.0113	0.0106
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5368	0.5327	0.5376	0.5368	0.0139	0.0136	0.0142	0.0139
chavinlo/gpt4-x-alpaca	2.00E-05	64	5	13000000000	decoder	general_purpose_models	0.5506	0.5323	0.5514	0.5506	0.0325	0.0340	0.0325	0.0325
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5352	0.5322	0.5357	0.5352	0.0075	0.0072	0.0074	0.0075
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5366	0.5310	0.5380	0.5366	0.0203	0.0194	0.0215	0.0203
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5336	0.5305	0.5339	0.5336	0.0108	0.0115	0.0107	0.0108
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5340	0.5304	0.5344	0.5340	0.0105	0.0096	0.0107	0.0105
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5325	0.5299	0.5325	0.5325	0.0160	0.0160	0.0159	0.0160
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5373	0.5296	0.5387	0.5373	0.0172	0.0198	0.0170	0.0172
chavinlo/gpt4-x-alpaca	2.00E-05	128	4	13000000000	decoder	general_purpose_models	0.5469	0.5293	0.5471	0.5469	0.0200	0.0206	0.0201	0.0200
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5350	0.5293	0.5363	0.5350	0.0137	0.0132	0.0142	0.0137
ProsusAI/finbert	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5316	0.5289	0.5320	0.5316	0.0129	0.0125	0.0132	0.0129
chavinlo/gpt4-x-alpaca	3.00E-05	128	4	13000000000	decoder	general_purpose_models	0.5552	0.5287	0.5585	0.5552	0.0233	0.0265	0.0256	0.0233
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5325	0.5286	0.5332	0.5325	0.0131	0.0131	0.0133	0.0131
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5340	0.5285	0.5348	0.5340	0.0159	0.0157	0.0163	0.0159
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5314	0.5282	0.5317	0.5314	0.0116	0.0111	0.0121	0.0116
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	2.00E-05	64	5	561000000	encoder	debate_based_models	0.5336	0.5280	0.5351	0.5336	0.0098	0.0104	0.0101	0.0098
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5324	0.5275	0.5334	0.5324	0.0193	0.0196	0.0199	0.0193
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5295	0.5258	0.5298	0.5295	0.0225	0.0228	0.0228	0.0225
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5283	0.5256	0.5284	0.5283	0.0217	0.0213	0.0220	0.0217
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	128	4	110000000	encoder	financial_base_models	0.5301	0.5255	0.5311	0.5301	0.0130	0.0112	0.0140	0.0130
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5323	0.5253	0.5340	0.5323	0.0125	0.0107	0.0143	0.0125
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5309	0.5252	0.5319	0.5309	0.0132	0.0148	0.0135	0.0132
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5299	0.5251	0.5304	0.5299	0.0147	0.0149	0.0148	0.0147
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5297	0.5250	0.5303	0.5297	0.0187	0.0177	0.0190	0.0187
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5290	0.5243	0.5294	0.5290	0.0123	0.0130	0.0127	0.0123
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5272	0.5240	0.5275	0.5272	0.0169	0.0170	0.0172	0.0169
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5270	0.5239	0.5272	0.5270	0.0143	0.0139	0.0143	0.0143
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5320	0.5238	0.5336	0.5320	0.0106	0.0115	0.0110	0.0106
ProsusAI/finbert	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5280	0.5235	0.5280	0.5280	0.0169	0.0199	0.0171	0.0169
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5262	0.5235	0.5265	0.5262	0.0209	0.0203	0.0211	0.0209
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5279	0.5234	0.5283	0.5279	0.0122	0.0129	0.0123	0.0122
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5290	0.5233	0.5299	0.5290	0.0071	0.0094	0.0069	0.0071
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5274	0.5232	0.5278	0.5274	0.0155	0.0154	0.0160	0.0155
ProsusAI/finbert	2.00E-05	128	4	110000000	encoder	financial_base_models	0.5266	0.5231	0.5270	0.5266	0.0107	0.0104	0.0109	0.0107
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5264	0.5229	0.5269	0.5264	0.0223	0.0217	0.0225	0.0223
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5274	0.5229	0.5279	0.5274	0.0174	0.0175	0.0176	0.0174
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5260	0.5226	0.5262	0.5260	0.0113	0.0115	0.0115	0.0113
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5287	0.5222	0.5299	0.5287	0.0130	0.0143	0.0134	0.0130
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5274	0.5221	0.5282	0.5274	0.0127	0.0137	0.0127	0.0127
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5248	0.5221	0.5248	0.5248	0.0104	0.0099	0.0105	0.0104
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5256	0.5221	0.5260	0.5256	0.0192	0.0185	0.0195	0.0192
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5256	0.5218	0.5261	0.5256	0.0286	0.0278	0.0293	0.0286
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5243	0.5213	0.5244	0.5243	0.0124	0.0119	0.0124	0.0124
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5245	0.5212	0.5253	0.5245	0.0081	0.0070	0.0088	0.0081
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5261	0.5211	0.5266	0.5261	0.0086	0.0082	0.0088	0.0086
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5252	0.5210	0.5256	0.5252	0.0160	0.0160	0.0160	0.0160
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5236	0.5210	0.5239	0.5236	0.0121	0.0120	0.0121	0.0121
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5238	0.5209	0.5239	0.5238	0.0156	0.0150	0.0157	0.0156
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5243	0.5209	0.5245	0.5243	0.0103	0.0103	0.0104	0.0103
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5265	0.5209	0.5273	0.5265	0.0099	0.0109	0.0101	0.0099
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5273	0.5208	0.5281	0.5273	0.0095	0.0120	0.0094	0.0095
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5241	0.5207	0.5242	0.5241	0.0140	0.0148	0.0139	0.0140

model_name	lr	max_len	epoch	model_size	model_architecture	model_type	mean_accuracy	mean_f1	mean_precision	mean_recall	std_accuracy	std_f1	std_precision	std_recall
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5245	0.5207	0.5248	0.5245	0.0129	0.0141	0.0134	0.0129
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5244	0.5206	0.5246	0.5244	0.0161	0.0162	0.0162	0.0161
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5283	0.5206	0.5295	0.5283	0.0099	0.0123	0.0095	0.0099
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5240	0.5206	0.5242	0.5240	0.0143	0.0137	0.0143	0.0143
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5251	0.5205	0.5256	0.5251	0.0094	0.0088	0.0097	0.0094
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5244	0.5203	0.5248	0.5244	0.0089	0.0091	0.0089	0.0089
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	256	4	561000000	encoder	debate_based_models	0.5387	0.5203	0.5414	0.5387	0.0213	0.0449	0.0212	0.0213
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5252	0.5203	0.5256	0.5252	0.0126	0.0139	0.0129	0.0126
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5261	0.5202	0.5269	0.5261	0.0095	0.0107	0.0096	0.0095
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5233	0.5202	0.5238	0.5233	0.0133	0.0125	0.0136	0.0133
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	256	5	561000000	encoder	debate_based_models	0.5438	0.5201	0.5481	0.5438	0.0219	0.0549	0.0171	0.0219
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5242	0.5200	0.5245	0.5242	0.0111	0.0116	0.0111	0.0111
ProsusAI/finbert	2.00E-05	256	5	110000000	encoder	financial_base_models	0.5267	0.5200	0.5282	0.5267	0.0071	0.0038	0.0085	0.0071
bert-base-uncased	3.00E-05	256	5	110000000	encoder	general_purpose_models	0.5399	0.5199	0.5461	0.5399	0.0097	0.0178	0.0073	0.0097
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5232	0.5199	0.5238	0.5232	0.0087	0.0081	0.0091	0.0087
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5271	0.5197	0.5278	0.5271	0.0196	0.0209	0.0200	0.0196
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5257	0.5197	0.5271	0.5257	0.0179	0.0175	0.0183	0.0179
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5226	0.5196	0.5229	0.5226	0.0146	0.0145	0.0146	0.0146
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5234	0.5195	0.5236	0.5234	0.0172	0.0161	0.0175	0.0172
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5236	0.5194	0.5241	0.5236	0.0137	0.0135	0.0141	0.0137
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	64	3	110000000	encoder	financial_base_models	0.5240	0.5193	0.5245	0.5240	0.0068	0.0066	0.0075	0.0068
ProsusAI/finbert	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5239	0.5191	0.5243	0.5239	0.0061	0.0074	0.0063	0.0061
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5235	0.5191	0.5241	0.5235	0.0165	0.0158	0.0167	0.0165
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5225	0.5191	0.5229	0.5225	0.0144	0.0136	0.0147	0.0144
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5229	0.5190	0.5233	0.5229	0.0173	0.0162	0.0182	0.0173
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5233	0.5190	0.5237	0.5233	0.0178	0.0189	0.0179	0.0178
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5226	0.5189	0.5229	0.5226	0.0130	0.0132	0.0132	0.0130
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5237	0.5189	0.5245	0.5237	0.0198	0.0182	0.0204	0.0198
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5218	0.5189	0.5218	0.5218	0.0085	0.0086	0.0085	0.0085
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5231	0.5189	0.5237	0.5231	0.0159	0.0160	0.0164	0.0159
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5232	0.5188	0.5235	0.5232	0.0109	0.0124	0.0113	0.0109
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5219	0.5188	0.5220	0.5219	0.0151	0.0142	0.0150	0.0151
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5213	0.5187	0.5216	0.5213	0.0145	0.0142	0.0146	0.0145
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5213	0.5185	0.5214	0.5213	0.0131	0.0127	0.0129	0.0131
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5232	0.5183	0.5233	0.5232	0.0116	0.0112	0.0118	0.0116
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	3.00E-05	64	5	184000000	decoder	financial_base_models	0.5603	0.5183	0.5887	0.5603	0.0130	0.0254	0.0117	0.0130
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5221	0.5183	0.5225	0.5221	0.0106	0.0102	0.0108	0.0106
ProsusAI/finbert	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5215	0.5182	0.5216	0.5215	0.0118	0.0117	0.0119	0.0118
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5249	0.5181	0.5264	0.5249	0.0091	0.0071	0.0102	0.0091
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5236	0.5181	0.5245	0.5236	0.0102	0.0081	0.0112	0.0102
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	128	3	110000000	encoder	financial_base_models	0.5242	0.5180	0.5249	0.5242	0.0136	0.0147	0.0135	0.0136
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5234	0.5179	0.5241	0.5234	0.0162	0.0158	0.0169	0.0162
ProsusAI/finbert	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5242	0.5178	0.5256	0.5242	0.0139	0.0139	0.0148	0.0139
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	64	5	110000000	encoder	financial_base_models	0.5210	0.5178	0.5214	0.5210	0.0146	0.0146	0.0148	0.0146
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5202	0.5178	0.5205	0.5202	0.0140	0.0141	0.0137	0.0140
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5203	0.5178	0.5204	0.5203	0.0186	0.0180	0.0188	0.0186
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5215	0.5177	0.5217	0.5215	0.0117	0.0123	0.0120	0.0117
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5223	0.5176	0.5225	0.5223	0.0065	0.0079	0.0067	0.0065
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5209	0.5176	0.5211	0.5209	0.0190	0.0182	0.0193	0.0190
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5233	0.5175	0.5239	0.5233	0.0195	0.0192	0.0200	0.0195
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5205	0.5175	0.5207	0.5205	0.0125	0.0122	0.0127	0.0125
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5228	0.5175	0.5234	0.5228	0.0061	0.0061	0.0063	0.0061
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	64	2	110000000	encoder	financial_base_models	0.5209	0.5175	0.5211	0.5209	0.0129	0.0130	0.0130	0.0129
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5214	0.5174	0.5219	0.5214	0.0143	0.0133	0.0149	0.0143
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5216	0.5174	0.5220	0.5216	0.0149	0.0145	0.0150	0.0149
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5204	0.5174	0.5205	0.5204	0.0136	0.0139	0.0137	0.0136
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	256	4	561000000	encoder	debate_based_models	0.5381	0.5174	0.5403	0.5381	0.0231	0.0542	0.0205	0.0231
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5218	0.5173	0.5223	0.5218	0.0192	0.0188	0.0194	0.0192
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5209	0.5172	0.5212	0.5209	0.0174	0.0169	0.0174	0.0174
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	2	110000000	encoder	financial_base_models	0.5214	0.5172	0.5217	0.5214	0.0054	0.0055	0.0054	0.0054
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5204	0.5171	0.5208	0.5204	0.0133	0.0123	0.0137	0.0133
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5204	0.5170	0.5207	0.5204	0.0115	0.0109	0.0118	0.0115
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	128	3	110000000	encoder	financial_base_models	0.5218	0.5170	0.5224	0.5218	0.0123	0.0116	0.0124	0.0123
bigscience/bloom-560m	2.00E-05	256	5	560000000	decoder	general_purpose_models	0.5387	0.5169	0.5391	0.5387	0.0198	0.0192	0.0196	0.0198
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5218	0.5169	0.5225	0.5218	0.0115	0.0108	0.0119	0.0115
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5216	0.5169	0.5226	0.5216	0.0164	0.0145	0.0173	0.0164
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5207	0.5168	0.5213	0.5207	0.0209	0.0202	0.0213	0.0209
ProsusAI/finbert	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5215	0.5167	0.5220	0.5215	0.0094	0.0092	0.0102	0.0094
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5211	0.5167	0.5218	0.5211	0.0057	0.0049	0.0061	0.0057
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5230	0.5166	0.5236	0.5230	0.0176	0.0186	0.0185	0.0176
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5206	0.5166	0.5209	0.5206	0.0109	0.0108	0.0108	0.0109
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5204	0.5166	0.5210	0.5204	0.0096	0.0091	0.0103	0.0096
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5210	0.5165	0.5217	0.5210	0.0243	0.0231	0.0251	0.0243
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5219	0.5165	0.5224	0.5219	0.0079	0.0080	0.0080	0.0079
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5204	0.5162	0.5207	0.5204	0.0104	0.0097	0.0106	0.0104
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5206	0.5162	0.5208	0.5206	0.0099	0.0097	0.0102	0.0099
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5216	0.5162	0.5220	0.5216	0.0172	0.0183	0.0176	0.0172
ProsusAI/finbert	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5210	0.5161	0.5209	0.5210	0.0084	0.0099	0.0085	0.0084
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5196	0.5161	0.5197	0.5196	0.0184	0.0176	0.0186	0.0184
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5239	0.5159	0.5245	0.5239	0.0113	0.0153	0.0113	0.0113
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5204	0.5159	0.5205	0.5204	0.0154	0.0182	0.0155	0.0154
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5203	0.5159	0.5208	0.5203	0.0177	0.0169	0.0182	0.0177
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	2.00E-05	128	5	561000000	encoder	debate_based_models	0.5403	0.5158	0.5474	0.5403	0.0125	0.0327	0.0099	0.0125
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	4	110000000	encoder	financial_base_models	0.5223	0.5158	0.5233	0.5223	0.0107	0.0120	0.0109	0.0107
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5199	0.5158	0.5203	0.5199	0.0145	0.0147	0.0148	0.0145
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5196	0.5158	0.5196	0.5196	0.0210	0.0206	0.0213	0.0210
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5213	0.5157	0.5217	0.5213	0.0023	0.0038	0.0024	0.0023
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5201	0.5156	0.5208	0.5201	0.0195	0.0189	0.0198	0.0195
ProsusAI/finbert	3.00E-05	128	4	110000000	encoder	financial_base_models	0.5209	0.5156	0.5212	0.5209	0.0139	0.0146	0.0143	0.0139
ProsusAI/finbert	2.00E-05	256	5	110000000	encoder	financial_base_models	0.5196	0.5155	0.5198	0.5196	0.0131	0.0141	0.0133	0.0131
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	256	5	110000000	encoder	financial_base_models	0.5200	0.5153	0.5204	0.5200	0.0131	0.0147	0.0129	0.0131
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5186	0.5152	0.5189	0.5186	0.0054	0.0053	0.0053	0.0054
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5178	0.5152	0.5179	0.5178	0.0150	0.0150	0.0150	0.0150
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5183	0.5152	0.5184	0.5183	0.0094	0.0097	0.0092	0.0094
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	64	4	110000000	encoder	financial_base_models	0.5199	0.5152	0.5205	0.5199	0.0034	0.0039	0.0035	0.0034
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5176	0.5151	0.5178	0.5176	0.0204	0.0199	0.0206	0.0204
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	2	110000000	encoder	financial_base_models	0.5236	0.5151	0.5246	0.5236	0.0089	0.0109	0.0094	0.0089
ProsusAI/finbert	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5262	0.5150	0.5267	0.5262	0.0200	0.0266	0.0206	0.0200
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5177	0.5150	0.5178	0.5177	0.0105	0.0097	0.0106	0.0105
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	64	5	110000000	encoder	financial_base_models	0.5209	0.5149	0.5216	0.5209	0.0096	0.0092	0.0097	0.0096
bert-base-uncased	5.00E-05	128	5	110000000	encoder	general_purpose_models	0.5406	0.5148	0.5501	0.5406	0.0082	0.0151	0.0090	0.0082
ProsusAI/finbert	2.00E-05	256	5	110000000	encoder	financial_base_models	0.5210	0.5147	0.5214	0.5210	0.0122	0.0135	0.0126	0.0122
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5184	0.5146	0.5185	0.5184	0.0087	0.0084	0.0088	0.0087
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	64	3	110000000	encoder	financial_base_models	0.5179	0.5145	0.5183	0.5179	0.0059	0.0047	0.0063	0.0059
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5175	0.5144	0.5177	0.5175	0.0139	0.0131	0.0142	0.0139
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5192	0.5144	0.5194	0.5192	0.0140	0.0152	0.0142	0.0140
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5177	0.5144	0.5179	0.5177	0.0105	0.0091	0.0108	0.0105
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5183	0.5143	0.5188	0.5183	0.0098	0.0106	0.0097	0.0098
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5177	0.5143	0.5178	0.5177	0.0184	0.0179	0.0185	0.0184
ProsusAI/finbert	2.00E-05	256	4	110000000	encoder	financial_base_models	0.5174	0.5143	0.5178	0.5174	0.0121	0.0113	0.0123	0.0121
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5207	0.5143	0.5212	0.5207	0.0131	0.0153	0.0131	0.0131
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5214	0.5143	0.5221	0.5214	0.0176	0.0200	0.0178	0.0176
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	128	4	561000000	encoder	debate_based_models	0.5416	0.5138	0.5468	0.5416	0.0238	0.0671	0.0184	0.0238
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5195	0.5138	0.5202	0.5195	0.0111	0.0124	0.0113	0.0111
ProsusAI/finbert	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5180	0.5137	0.5185	0.5180	0.0150	0.0135	0.0157	0.0150
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5170	0.5137	0.5171	0.5170	0.0096	0.0104	0.0094	0.0096
ProsusAI/finbert	2.00E-05	128	4	110000000	encoder	financial_base_models	0.5207	0.5137	0.5217	0.5207	0.0052	0.0074	0.0056	0.0052
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5217	0.5136	0.5228	0.5217	0.0176	0.0190	0.0180	0.0176
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5166	0.5135	0.5169	0.5166	0.0176	0.0177	0.0177	0.0176
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5350	0.5134	0.5409	0.5350	0.0087	0.0162	0.0090	0.0087
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	128	5	184000000	decoder	financial_base_models	0.5604	0.5134	0.5953	0.5604	0.0128	0.0210	0.0150	0.0128
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5167	0.5134	0.5167	0.5167	0.0128	0.0132	0.0128	0.0128
ProsusAI/finbert	2.00E-05	64	4	110000000	encoder	financial_base_models	0.5165	0.5133	0.5166	0.5165	0.0099	0.0102	0.0098	0.0099
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5222	0.5133	0.5231	0.5222	0.0159	0.0209	0.0160	0.0159
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5163	0.5133	0.5165	0.5163	0.0076	0.0070	0.0077	0.0076
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5160	0.5132	0.5160	0.5160	0.0217	0.0210	0.0217	0.0217
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5194	0.5131	0.5202	0.5194	0.0174	0.0178	0.0182	0.0174
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5161	0.5131	0.5162	0.5161	0.0145	0.0143	0.0145	0.0145
ProsusAI/finbert	3.00E-05	256	2	110000000	encoder	financial_base_models	0.5178	0.5130	0.5182	0.5178	0.0160	0.0175	0.0162	0.0160
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5193	0.5130	0.5198	0.5193	0.0125	0.0139	0.0126	0.0125
ProsusAI/finbert	3.00E-05	128	2	110000000	encoder	financial_base_models	0.5177	0.5130	0.5177	0.5177	0.0146	0.0167	0.0146	0.0146
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5166	0.5129	0.5165	0.5166	0.0096	0.0102	0.0095	0.0096
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5168	0.5128	0.5171	0.5168	0.0185	0.0194	0.0187	0.0185
bert-base-uncased	5.00E-05	128	5	110000000	encoder	general_purpose_models	0.5438	0.5127	0.5579	0.5438	0.0089	0.0174	0.0111	0.0089
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5149	0.5127	0.5149	0.5149	0.0150	0.0145	0.0151	0.0150
bert-base-uncased	2.00E-05	64	3	110000000	encoder	general_purpose_models	0.5320	0.5127	0.5367	0.5320	0.0094	0.0180	0.0087	0.0094
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5184	0.5126	0.5189	0.5184	0.0144	0.0170	0.0143	0.0144
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5156	0.5126	0.5157	0.5156	0.0164	0.0161	0.0166	0.0164
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	64	5	110000000	encoder	financial_base_models	0.5155	0.5125	0.5156	0.5155	0.0194	0.0202	0.0192	0.0194
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5171	0.5125	0.5174	0.5171	0.0136	0.0127	0.0137	0.0136
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5182	0.5125	0.5189	0.5182	0.0114	0.0133	0.0112	0.0114
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5166	0.5124	0.5171	0.5166	0.0107	0.0113	0.0110	0.0107
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5159	0.5123	0.5159	0.5159	0.0111	0.0110	0.0114	0.0111
ProsusAI/finbert	2.00E-05	64	5	110000000	encoder	financial_base_models	0.5156	0.5122	0.5156	0.5156	0.0071	0.0077	0.0072	0.0071
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	256	5	110000000	encoder	financial_base_models	0.5151	0.5122	0.5152	0.5151	0.0085	0.0088	0.0085	0.0085
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	128	4	110000000	encoder	financial_base_models	0.5206	0.5121	0.5219	0.5206	0.0097	0.0096	0.0108	0.0097
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5204	0.5120	0.5203	0.5204	0.0134	0.0180	0.0138	0.0134
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	3	110000000	encoder	financial_base_models	0.5143	0.5119	0.5143	0.5143	0.0066	0.0069	0.0067	0.0066
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5151	0.5119	0.5152	0.5151	0.0251	0.0250	0.0253	0.0251
bert-base-uncased	2.00E-05	64	3	110000000	encoder	general_purpose_models	0.5359	0.5119	0.5452	0.5359	0.0049	0.0114	0.0095	0.0049
ProsusAI/finbert	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5191	0.5118	0.5192	0.5191	0.0145	0.0192	0.0146	0.0145
ProsusAI/finbert	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5223	0.5118	0.5234	0.5223	0.0135	0.0162	0.0136	0.0135
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5180	0.5118	0.5186	0.5180	0.0100	0.0102	0.0102	0.0100
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5150	0.5117	0.5153	0.5150	0.0135	0.0132	0.0136	0.0135
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5181	0.5117	0.5190	0.5181	0.0138	0.0158	0.0138	0.0138
ProsusAI/finbert	3.00E-05	128	3	110000000	encoder	financial_base_models	0.5183	0.5117	0.5181	0.5183	0.0094	0.0132	0.0100	0.0094
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5152	0.5117	0.5154	0.5152	0.0127	0.0123	0.0127	0.0127
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5344	0.5117	0.5418	0.5344	0.0095	0.0123	0.0121	0.0095
ProsusAI/finbert	3.00E-05	128	3	110000000	encoder	financial_base_models	0.5168	0.5116	0.5170	0.5168	0.0052	0.0079	0.0051	0.0052
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5149	0.5113	0.5149	0.5149	0.0200	0.0210	0.0200	0.0200
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5161	0.5113	0.5164	0.5161	0.0082	0.0084	0.0082	0.0082
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5158	0.5113	0.5163	0.5158	0.0126	0.0113	0.0131	0.0126
ProsusAI/finbert	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5170	0.5113	0.5173	0.5170	0.0128	0.0133	0.0135	0.0128
roberta-base	2.00E-05	256	4	125000000	encoder	general_purpose_models	0.5592	0.5112	0.5947	0.5592	0.0211	0.0522	0.0191	0.0211
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5142	0.5110	0.5143	0.5142	0.0114	0.0106	0.0113	0.0114
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	64	4	110000000	encoder	financial_base_models	0.5147	0.5109	0.5151	0.5147	0.0099	0.0103	0.0100	0.0099
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	2	110000000	encoder	financial_base_models	0.5141	0.5108	0.5144	0.5141	0.0077	0.0079	0.0079	0.0077
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5155	0.5108	0.5159	0.5155	0.0136	0.0135	0.0141	0.0136
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5146	0.5107	0.5149	0.5146	0.0112	0.0115	0.0118	0.0112
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5154	0.5107	0.5158	0.5154	0.0151	0.0137	0.0160	0.0151
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5136	0.5107	0.5139	0.5136	0.0148	0.0142	0.0150	0.0148
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	5.00E-05	128	4	184000000	decoder	financial_base_models	0.5619	0.5107	0.6035	0.5619	0.0115	0.0284	0.0109	0.0115
chavinlo/gpt4-x-alpaca	5.00E-05	256	4	13000000000	decoder	general_purpose_models	0.5628	0.5106	0.5798	0.5628	0.0334	0.0449	0.0410	0.0334
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5137	0.5106	0.5139	0.5137	0.0097	0.0094	0.0098	0.0097
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5133	0.5105	0.5135	0.5133	0.0090	0.0085	0.0090	0.0090
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5134	0.5105	0.5134	0.5134	0.0136	0.0142	0.0137	0.0136

model_name	lr	max_len	epoch	model_size	model_architecture	model_type	mean_accuracy	mean_f1	mean_precision	mean_recall	std_accuracy	std_f1	std_precision	std_recall
bert-base-uncased	3.00E-05	64	3	110000000	encoder	general_purpose_models	0.5354	0.5104	0.5434	0.5354	0.0086	0.0177	0.0104	0.0086
ProsusAI/finbert	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5149	0.5104	0.5150	0.5149	0.0160	0.0174	0.0160	0.0160
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5150	0.5104	0.5154	0.5150	0.0106	0.0096	0.0106	0.0106
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5140	0.5102	0.5139	0.5140	0.0051	0.0051	0.0052	0.0051
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5131	0.5100	0.5131	0.5131	0.0167	0.0174	0.0169	0.0167
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5165	0.5100	0.5167	0.5165	0.0029	0.0046	0.0030	0.0029
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5127	0.5100	0.5128	0.5127	0.0115	0.0114	0.0116	0.0115
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5126	0.5099	0.5128	0.5126	0.0094	0.0091	0.0093	0.0094
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	256	4	184000000	decoder	financial_base_models	0.5564	0.5098	0.5881	0.5564	0.0116	0.0190	0.0146	0.0116
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5146	0.5098	0.5149	0.5146	0.0149	0.0147	0.0151	0.0149
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5155	0.5098	0.5156	0.5155	0.0071	0.0097	0.0070	0.0071
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5156	0.5097	0.5163	0.5156	0.0139	0.0146	0.0145	0.0139
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5129	0.5096	0.5133	0.5129	0.0137	0.0135	0.0137	0.0137
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5141	0.5094	0.5146	0.5141	0.0102	0.0113	0.0102	0.0102
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5150	0.5093	0.5156	0.5150	0.0088	0.0096	0.0090	0.0088
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5150	0.5090	0.5155	0.5150	0.0124	0.0125	0.0130	0.0124
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5122	0.5090	0.5125	0.5122	0.0085	0.0081	0.0087	0.0085
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5117	0.5089	0.5117	0.5117	0.0161	0.0161	0.0163	0.0161
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5121	0.5089	0.5123	0.5121	0.0104	0.0097	0.0105	0.0104
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5118	0.5089	0.5120	0.5118	0.0040	0.0047	0.0038	0.0040
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	4	110000000	encoder	financial_base_models	0.5116	0.5088	0.5117	0.5116	0.0077	0.0072	0.0077	0.0077
ProsusAI/finbert	2.00E-05	256	4	110000000	encoder	financial_base_models	0.5122	0.5086	0.5123	0.5122	0.0184	0.0181	0.0187	0.0184
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5118	0.5085	0.5120	0.5118	0.0098	0.0096	0.0099	0.0098
ProsusAI/finbert	2.00E-05	64	5	110000000	encoder	financial_base_models	0.5119	0.5083	0.5123	0.5119	0.0192	0.0186	0.0198	0.0192
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5114	0.5083	0.5117	0.5114	0.0195	0.0190	0.0199	0.0195
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5123	0.5083	0.5125	0.5123	0.0114	0.0129	0.0114	0.0114
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	2.00E-05	128	2	561000000	encoder	debate_based_models	0.5319	0.5082	0.5375	0.5319	0.0113	0.0236	0.0121	0.0113
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5109	0.5082	0.5110	0.5109	0.0059	0.0056	0.0060	0.0059
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	4	110000000	encoder	financial_base_models	0.5103	0.5082	0.5103	0.5103	0.0185	0.0188	0.0185	0.0185
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5116	0.5081	0.5117	0.5116	0.0121	0.0111	0.0124	0.0121
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5155	0.5080	0.5164	0.5155	0.0135	0.0133	0.0140	0.0135
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5110	0.5079	0.5112	0.5110	0.0102	0.0099	0.0104	0.0102
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5118	0.5078	0.5122	0.5118	0.0094	0.0085	0.0096	0.0094
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	2.00E-05	256	5	561000000	encoder	debate_based_models	0.5384	0.5077	0.5478	0.5384	0.0167	0.0387	0.0161	0.0167
ProsusAI/finbert	2.00E-05	128	2	110000000	encoder	financial_base_models	0.5144	0.5077	0.5151	0.5144	0.0151	0.0143	0.0159	0.0151
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5104	0.5077	0.5105	0.5104	0.0147	0.0139	0.0147	0.0147
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5149	0.5075	0.5158	0.5149	0.0162	0.0142	0.0176	0.0162
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5097	0.5075	0.5098	0.5097	0.0145	0.0140	0.0146	0.0145
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5105	0.5075	0.5107	0.5105	0.0155	0.0157	0.0156	0.0155
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	3.00E-05	128	5	184000000	decoder	financial_base_models	0.5600	0.5074	0.6004	0.5600	0.0140	0.0299	0.0097	0.0140
roberta-base	2.00E-05	64	4	125000000	encoder	general_purpose_models	0.5584	0.5074	0.5795	0.5584	0.0363	0.0898	0.0207	0.0363
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5115	0.5073	0.5118	0.5115	0.0116	0.0129	0.0116	0.0116
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	64	5	110000000	encoder	financial_base_models	0.5142	0.5073	0.5149	0.5142	0.0201	0.0195	0.0217	0.0201
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5156	0.5073	0.5159	0.5156	0.0174	0.0210	0.0179	0.0174
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5209	0.5072	0.5217	0.5209	0.0139	0.0227	0.0139	0.0139
ProsusAI/finbert	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5180	0.5072	0.5187	0.5180	0.0125	0.0195	0.0135	0.0125
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5115	0.5072	0.5116	0.5115	0.0053	0.0063	0.0055	0.0053
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	128	2	110000000	encoder	financial_base_models	0.5104	0.5071	0.5106	0.5104	0.0144	0.0139	0.0147	0.0144
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	5.00E-05	64	5	184000000	decoder	financial_base_models	0.5593	0.5069	0.5963	0.5593	0.0150	0.0308	0.0127	0.0150
ProsusAI/finbert	3.00E-05	128	2	110000000	encoder	financial_base_models	0.5116	0.5069	0.5117	0.5116	0.0086	0.0099	0.0087	0.0086
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5114	0.5069	0.5117	0.5114	0.0133	0.0143	0.0135	0.0133
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5157	0.5069	0.5165	0.5157	0.0127	0.0170	0.0122	0.0127
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	3.00E-05	256	3	561000000	encoder	debate_based_models	0.5379	0.5068	0.5398	0.5379	0.0185	0.0710	0.0171	0.0185
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5098	0.5066	0.5097	0.5098	0.0134	0.0140	0.0136	0.0134
raruidol/ArgumentMining-EN-AC-Financial	2.00E-05	128	4	561000000	encoder	debate_based_models	0.5365	0.5066	0.5382	0.5365	0.0263	0.0559	0.0308	0.0263
bert-base-uncased	5.00E-05	256	3	110000000	encoder	general_purpose_models	0.5366	0.5066	0.5471	0.5366	0.0075	0.0197	0.0100	0.0075
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5093	0.5066	0.5096	0.5093	0.0158	0.0157	0.0159	0.0158
ProsusAI/finbert	3.00E-05	64	5	110000000	encoder	financial_base_models	0.5147	0.5065	0.5149	0.5147	0.0180	0.0203	0.0183	0.0180
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5086	0.5064	0.5086	0.5086	0.0111	0.0115	0.0110	0.0111
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5106	0.5062	0.5107	0.5106	0.0087	0.0094	0.0084	0.0087
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5089	0.5062	0.5090	0.5089	0.0126	0.0129	0.0128	0.0126
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	64	5	184000000	decoder	financial_base_models	0.5537	0.5060	0.5848	0.5537	0.0111	0.0154	0.0137	0.0111
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	4	110000000	encoder	financial_base_models	0.5088	0.5060	0.5087	0.5088	0.0111	0.0109	0.0112	0.0111
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5107	0.5059	0.5112	0.5107	0.0195	0.0184	0.0200	0.0195
ProsusAI/finbert	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5192	0.5059	0.5206	0.5192	0.0152	0.0186	0.0160	0.0152
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5093	0.5059	0.5096	0.5093	0.0120	0.0121	0.0119	0.0120
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5097	0.5058	0.5097	0.5097	0.0182	0.0188	0.0183	0.0182
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5087	0.5057	0.5088	0.5087	0.0136	0.0139	0.0136	0.0136
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5089	0.5056	0.5091	0.5089	0.0126	0.0120	0.0128	0.0126
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5091	0.5055	0.5093	0.5091	0.0183	0.0177	0.0186	0.0183
bert-base-uncased	3.00E-05	128	2	110000000	encoder	general_purpose_models	0.5314	0.5055	0.5380	0.5314	0.0099	0.0229	0.0096	0.0099
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5102	0.5053	0.5109	0.5102	0.0176	0.0168	0.0184	0.0176
ProsusAI/finbert	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5260	0.5053	0.5283	0.5260	0.0170	0.0316	0.0174	0.0170
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5083	0.5053	0.5083	0.5083	0.0117	0.0108	0.0118	0.0117
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5091	0.5053	0.5095	0.5091	0.0108	0.0099	0.0112	0.0108
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5122	0.5052	0.5123	0.5122	0.0164	0.0180	0.0167	0.0164
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5140	0.5052	0.5148	0.5140	0.0168	0.0195	0.0171	0.0168
bigscience/bloom-560m	2.00E-05	64	2	560000000	decoder	general_purpose_models	0.5251	0.5051	0.5252	0.5251	0.0175	0.0175	0.0172	0.0175
bert-base-uncased	2.00E-05	128	2	110000000	encoder	general_purpose_models	0.5212	0.5051	0.5233	0.5212	0.0197	0.0220	0.0215	0.0197
ProsusAI/finbert	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5223	0.5051	0.5235	0.5223	0.0133	0.0268	0.0136	0.0133
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	4	110000000	encoder	financial_base_models	0.5119	0.5051	0.5121	0.5119	0.0085	0.0091	0.0087	0.0085
ProsusAI/finbert	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5215	0.5050	0.5231	0.5215	0.0112	0.0233	0.0103	0.0112
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5115	0.5049	0.5121	0.5115	0.0063	0.0076	0.0067	0.0063
ProsusAI/finbert	3.00E-05	256	2	110000000	encoder	financial_base_models	0.5112	0.5048	0.5114	0.5112	0.0075	0.0083	0.0076	0.0075
bert-base-uncased	5.00E-05	64	4	110000000	encoder	general_purpose_models	0.5340	0.5048	0.5418	0.5340	0.0091	0.0189	0.0076	0.0091
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	4	110000000	encoder	financial_base_models	0.5069	0.5046	0.5068	0.5069	0.0071	0.0072	0.0072	0.0071
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	64	2	110000000	encoder	financial_base_models	0.5163	0.5043	0.5181	0.5163	0.0084	0.0099	0.0093	0.0084
ProsusAI/finbert	2.00E-05	64	4	110000000	encoder	financial_base_models	0.5070	0.5042	0.5071	0.5070	0.0112	0.0115	0.0112	0.0112
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	2	110000000	encoder	financial_base_models	0.5078	0.5041	0.5080	0.5078	0.0116	0.0121	0.0116	0.0116
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5065	0.5041	0.5065	0.5065	0.0133	0.0135	0.0133	0.0133
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5078	0.5041	0.5078	0.5078	0.0069	0.0070	0.0071	0.0069
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	128	3	184000000	decoder	financial_base_models	0.5519	0.5041	0.5821	0.5519	0.0113	0.0154	0.0172	0.0113
bigscience/bloomz-1b1	2.00E-05	64	5	1000000000	decoder	general_purpose_models	0.5061	0.5040	0.5062	0.5061	0.0216	0.0217	0.0217	0.0216
chavinlo/gpt4-x-alpaca	2.00E-05	256	3	13000000000	decoder	general_purpose_models	0.5225	0.5040	0.5226	0.5225	0.0235	0.0221	0.0232	0.0235
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	2.00E-05	256	4	561000000	encoder	debate_based_models	0.5335	0.5039	0.5383	0.5335	0.0148	0.0581	0.0098	0.0148
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	5.00E-05	256	2	184000000	decoder	financial_base_models	0.5535	0.5039	0.5875	0.5535	0.0123	0.0247	0.0132	0.0123
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5062	0.5037	0.5063	0.5062	0.0132	0.0124	0.0134	0.0132
ProsusAI/finbert	5.00E-05	128	4	110000000	encoder	financial_base_models	0.5179	0.5036	0.5184	0.5179	0.0183	0.0245	0.0187	0.0183
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	64	3	561000000	encoder	debate_based_models	0.5326	0.5036	0.5332	0.5326	0.0207	0.0693	0.0199	0.0207
ProsusAI/finbert	2.00E-05	256	5	110000000	encoder	financial_base_models	0.5081	0.5035	0.5082	0.5081	0.0127	0.0116	0.0132	0.0127
ProsusAI/finbert	3.00E-05	128	4	110000000	encoder	financial_base_models	0.5091	0.5034	0.5093	0.5091	0.0147	0.0150	0.0150	0.0147
ProsusAI/finbert	5.00E-05	256	2	110000000	encoder	financial_base_models	0.5142	0.5034	0.5138	0.5142	0.0131	0.0213	0.0145	0.0131
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	4	110000000	encoder	financial_base_models	0.5089	0.5032	0.5088	0.5089	0.0112	0.0110	0.0115	0.0112
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	256	5	184000000	decoder	financial_base_models	0.5548	0.5032	0.5915	0.5548	0.0103	0.0145	0.0151	0.0103
ProsusAI/finbert	5.00E-05	64	4	110000000	encoder	financial_base_models	0.5215	0.5032	0.5255	0.5215	0.0122	0.0184	0.0152	0.0122
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	64	5	184000000	decoder	financial_base_models	0.5517	0.5031	0.5833	0.5517	0.0122	0.0169	0.0174	0.0122
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5066	0.5030	0.5066	0.5066	0.0096	0.0103	0.0097	0.0096
ProsusAI/finbert	3.00E-05	128	4	110000000	encoder	financial_base_models	0.5153	0.5029	0.5155	0.5153	0.0141	0.0169	0.0145	0.0141
ahmedrachid/FinancialBERT-Sentiment-Analysis	5.00E-05	256	2	110000000	encoder	financial_base_models	0.5085	0.5029	0.5088	0.5085	0.0130	0.0121	0.0135	0.0130
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	5.00E-05	128	2	184000000	decoder	financial_base_models	0.5566	0.5028	0.5968	0.5566	0.0113	0.0257	0.0095	0.0113
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	3.00E-05	128	3	184000000	decoder	financial_base_models	0.5532	0.5027	0.5880	0.5532	0.0143	0.0269	0.0163	0.0143
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5084	0.5024	0.5088	0.5084	0.0196	0.0201	0.0201	0.0196
ProsusAI/finbert	5.00E-05	128	3	110000000	encoder	financial_base_models	0.5149	0.5024	0.5150	0.5149	0.0175	0.0280	0.0180	0.0175
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5051	0.5023	0.5052	0.5051	0.0053	0.0054	0.0053	0.0053
ProsusAI/finbert	3.00E-05	64	5	110000000	encoder	financial_base_models	0.5066	0.5022	0.5066	0.5066	0.0199	0.0194	0.0205	0.0199
ProsusAI/finbert	5.00E-05	64	5	110000000	encoder	financial_base_models	0.5271	0.5021	0.5315	0.5271	0.0127	0.0280	0.0110	0.0127
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	256	2	110000000	encoder	financial_base_models	0.5053	0.5020	0.5051	0.5053	0.0071	0.0076	0.0071	0.0071
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	128	3	184000000	decoder	financial_base_models	0.5520	0.5020	0.5834	0.5520	0.0136	0.0228	0.0143	0.0136
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	2.00E-05	64	3	184000000	decoder	financial_base_models	0.5489	0.5019	0.5764	0.5489	0.0126	0.0136	0.0206	0.0126
ProsusAI/finbert	2.00E-05	128	2	110000000	encoder	financial_base_models	0.5068	0.5019	0.5066	0.5068	0.0111	0.0106	0.0110	0.0111
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	256	2	110000000	encoder	financial_base_models	0.5048	0.5019	0.5047	0.5048	0.0117	0.0124	0.0117	0.0117
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	5.00E-05	128	3	561000000	encoder	debate_based_models	0.5354	0.5017	0.5471	0.5354	0.0079	0.0279	0.0025	0.0079
ProsusAI/finbert	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5107	0.5016	0.5111	0.5107	0.0138	0.0159	0.0137	0.0138
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5051	0.5013	0.5052	0.5051	0.0085	0.0066	0.0088	0.0085
ProsusAI/finbert	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5165	0.5010	0.5172	0.5165	0.0154	0.0251	0.0157	0.0154
ProsusAI/finbert	2.00E-05	64	2	110000000	encoder	financial_base_models	0.5088	0.5009	0.5094	0.5088	0.0095	0.0132	0.0097	0.0095
bigscience/bloom-560m	5.00E-05	256	2	560000000	decoder	general_purpose_models	0.5292	0.5008	0.5267	0.5292	0.0252	0.0335	0.0266	0.0252
ProsusAI/finbert	5.00E-05	128	5	110000000	encoder	financial_base_models	0.5243	0.5008	0.5280	0.5243	0.0118	0.0260	0.0117	0.0118
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	64	2	110000000	encoder	financial_base_models	0.5050	0.5004	0.5049	0.5050	0.0068	0.0073	0.0069	0.0068
ProsusAI/finbert	2.00E-05	64	2	110000000	encoder	financial_base_models	0.5040	0.5004	0.5039	0.5040	0.0083	0.0096	0.0082	0.0083
ProsusAI/finbert	2.00E-05	256	5	110000000	encoder	financial_base_models	0.5054	0.5004	0.5054	0.5054	0.0067	0.0046	0.0067	0.0067
ProsusAI/finbert	2.00E-05	128	3	110000000	encoder	financial_base_models	0.5096	0.5003	0.5101	0.5096	0.0035	0.0040	0.0037	0.0035
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5043	0.5000	0.5046	0.5043	0.0078	0.0065	0.0083	0.0078
ProsusAI/finbert	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5250	0.4994	0.5285	0.5250	0.0139	0.0255	0.0137	0.0139
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	5	110000000	encoder	financial_base_models	0.5080	0.4991	0.5085	0.5080	0.0129	0.0164	0.0128	0.0129
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	5	110000000	encoder	financial_base_models	0.5023	0.4989	0.5026	0.5023	0.0073	0.0081	0.0075	0.0073
bert-base-uncased	3.00E-05	128	3	110000000	encoder	general_purpose_models	0.5354	0.4988	0.5499	0.5354	0.0087	0.0314	0.0152	0.0087
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	4	110000000	encoder	financial_base_models	0.5012	0.4988	0.5011	0.5012	0.0177	0.0172	0.0177	0.0177
bert-base-uncased	3.00E-05	128	4	110000000	encoder	general_purpose_models	0.5211	0.4988	0.5242	0.5211	0.0126	0.0249	0.0143	0.0126
ProsusAI/finbert	5.00E-05	64	5	110000000	encoder	financial_base_models	0.5186	0.4987	0.5211	0.5186	0.0253	0.0319	0.0263	0.0253
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	4	110000000	encoder	financial_base_models	0.5015	0.4986	0.5015	0.5015	0.0108	0.0104	0.0110	0.0108
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	256	3	110000000	encoder	financial_base_models	0.5010	0.4985	0.5010	0.5010	0.0183	0.0177	0.0183	0.0183
raruidol/ArgumentMining-EN-AC-Financial	2.00E-05	128	4	561000000	encoder	debate_based_models	0.5378	0.4984	0.5493	0.5378	0.0218	0.0706	0.0104	0.0218
ProsusAI/finbert	5.00E-05	64	4	110000000	encoder	financial_base_models	0.5095	0.4981	0.5095	0.5095	0.0076	0.0133	0.0076	0.0076
ProsusAI/finbert	5.00E-05	256	4	110000000	encoder	financial_base_models	0.5165	0.4978	0.5167	0.5165	0.0168	0.0304	0.0182	0.0168
roberta-base	5.00E-05	128	5	125000000	encoder	general_purpose_models	0.5562	0.4975	0.5984	0.5562	0.0201	0.0416	0.0092	0.0201
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	128	2	110000000	encoder	financial_base_models	0.4997	0.4974	0.4999	0.4997	0.0089	0.0091	0.0090	0.0089
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	64	5	110000000	encoder	financial_base_models	0.5024	0.4971	0.5026	0.5024	0.0200	0.0203	0.0208	0.0200
ProsusAI/finbert	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5153	0.4971	0.5159	0.5153	0.0169	0.0294	0.0173	0.0169
bert-base-uncased	5.00E-05	64	2	110000000	encoder	general_purpose_models	0.5312	0.4967	0.5415	0.5312	0.0075	0.0198	0.0078	0.0075
ProsusAI/finbert	5.00E-05	128	5	110000000	encoder	financial_base_models	0.5209	0.4965	0.5238	0.5209	0.0149	0.0265	0.0140	0.0149
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	64	3	561000000	encoder	debate_based_models	0.5334	0.4964	0.4971	0.5334	0.0200	0.0818	0.0917	0.0200
bigscience/bloom-7b1	2.00E-05	64	2	7000000000	decoder	general_purpose_models	0.5142	0.4963	0.5130	0.5142	0.0216	0.0231	0.0227	0.0216
ProsusAI/finbert	5.00E-05	256	3	110000000	encoder	financial_base_models	0.5153	0.4958	0.5156	0.5153	0.0196	0.0306	0.0213	0.0196
raruidol/ArgumentMining-EN-AC-Financial	2.00E-05	128	2	561000000	encoder	debate_based_models	0.5284	0.4956	0.5280	0.5284	0.0188	0.0583	0.0234	0.0188
ProsusAI/finbert	3.00E-05	128	5	110000000	encoder	financial_base_models	0.5051	0.4953	0.5054	0.5051	0.0167	0.0203	0.0170	0.0167
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	3.00E-05	256	2	184000000	decoder	financial_base_models	0.5498	0.4951	0.5855	0.5498	0.0123	0.0275	0.0103	0.0123
ProsusAI/finbert	3.00E-05	64	4	110000000	encoder	financial_base_models	0.5127	0.4949	0.5134	0.5127	0.0166	0.0222	0.0172	0.0166
ProsusAI/finbert	5.00E-05	128	2	110000000	encoder	financial_base_models	0.5101	0.4949	0.5103	0.5101	0.0135	0.0230	0.0159	0.0135
ahmedrachid/FinancialBERT-Sentiment-Analysis	2.00E-05	256	3	110000000	encoder	financial_base_models	0.4966	0.4947	0.4966	0.4966	0.0106	0.0109	0.0106	0.0106
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	64	2	110000000	encoder	financial_base_models	0.4976	0.4941	0.4975	0.4976	0.0091	0.0101	0.0093	0.0091
ahmedrachid/FinancialBERT-Sentiment-Analysis	3.00E-05	128	2	110000000	encoder	financial_base_models	0.5050	0.4935	0.5056	0.5050	0.0073	0.0113	0.0076	0.0073
nickmuchi/deberta-v3-base-finetuned-finance-text-classification	3.00E-05	64	4	184000000	decoder	financial_base_models	0.5494	0.4932	0.5842	0.5494	0.0142	0.0274	0.0146	0.0142
ProsusAI/finbert	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5176	0.4930	0.5188	0.5176	0.0193	0.0298	0.0198	0.0193
roberta-base	5.00E-05	256	5	125000000	encoder	general_purpose_models	0.5574	0.4929	0.5895	0.5574	0.0344	0.0904	0.0245	0.0344
raruidol/ArgumentMining-EN-AC-Financial	2.00E-05	128	3	561000000	encoder	debate_based_models	0.5252	0.4928	0.5183	0.5252	0.0202	0.0668	0.0323	0.0202
ProsusAI/finbert	5.00E-05	64	5	110000000	encoder	financial_base_models	0.5214	0.4921	0.5248	0.5214	0.0142	0.0309	0.0138	0.0142
bert-base-uncased	5.00E-05	256	2	110000000	encoder	general_purpose_models	0.5318	0.4914	0.5446	0.5318	0.0094	0.0240	0.0140	0.0094
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5133	0.4911	0.5107	0.5133	0.0347	0.0358	0.0357	0.0347
ProsusAI/finbert	5.00E-05	128	3	110000000	encoder	financial_base_models	0.5042	0.4910	0.5040	0.5042	0.0029	0.0145	0.0030	0.0029
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5050	0.4904	0.5069	0.5050	0.0282	0.0255	0.0285	0.0282
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5177	0.4901	0.5151	0.5177	0.0083	0.0066	0.0086	0.0083
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5139	0.4897	0.5115	0.5139	0.0320	0.0318	0.0302	0.0320
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	5.00E-05	256	4	561000000	encoder	debate_based_models	0.5359	0.4895	0.5534	0.5359	0.0140	0.0415	0.0123	0.0140
ProsusAI/finbert	5.00E-05	64	3	110000000	encoder	financial_base_models	0.5083	0.4890	0.5080	0.5083	0.0133	0.0271	0.0146	0.0133
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.5124	0.4888	0.5091	0.5124	0.0151	0.0119	0.0140	0.0151

model_name	lr	max_len	epoch	model_size	model_architecture	model_type	mean_accuracy	mean_f1	mean_precision	mean_recall	std_accuracy	std_f1	std_precision	std_recall
ProsusAI/finbert	5.00E-05	256	5	110000000	encoder	financial_base_models	0.5071	0.4887	0.5070	0.5071	0.0147	0.0201	0.0160	0.0147
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5061	0.4882	0.5044	0.5061	0.0263	0.0257	0.0232	0.0263
roberta-base	3.00E-05	64	5	125000000	encoder	general_purpose_models	0.5518	0.4881	0.5910	0.5518	0.0318	0.0853	0.0243	0.0318
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5045	0.4880	0.5037	0.5045	0.0212	0.0183	0.0222	0.0212
raruidol/ArgumentMining-EN-AC-Financial	2.00E-05	128	3	561000000	encoder	debate_based_models	0.5304	0.4877	0.5195	0.5304	0.0211	0.0753	0.0508	0.0211
bert-base-uncased	3.00E-05	128	2	110000000	encoder	general_purpose_models	0.5052	0.4877	0.5054	0.5052	0.0459	0.0483	0.0477	0.0459
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5052	0.4873	0.5048	0.5052	0.0271	0.0260	0.0285	0.0271
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	5.00E-05	256	5	561000000	encoder	debate_based_models	0.5360	0.4871	0.5552	0.5360	0.0113	0.0304	0.0091	0.0113
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5036	0.4870	0.5069	0.5036	0.0237	0.0254	0.0246	0.0237
roberta-base	2.00E-05	64	2	125000000	encoder	general_purpose_models	0.5472	0.4869	0.5216	0.5472	0.0322	0.0933	0.1210	0.0322
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5054	0.4863	0.5052	0.5054	0.0315	0.0293	0.0353	0.0315
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	5.00E-05	128	2	561000000	encoder	debate_based_models	0.5329	0.4863	0.5499	0.5329	0.0062	0.0246	0.0059	0.0062
TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ	5.00E-05	256	3	7000000000	encoder_decoder	general_purpose_models	0.5218	0.4863	0.5235	0.5218	0.0152	0.0300	0.0174	0.0152
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	128	2	561000000	encoder	debate_based_models	0.5236	0.4863	0.4905	0.5236	0.0159	0.0767	0.0802	0.0159
TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ	5.00E-05	256	4	7000000000	encoder_decoder	general_purpose_models	0.5234	0.4862	0.5230	0.5234	0.0156	0.0268	0.0191	0.0156
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5059	0.4862	0.5052	0.5059	0.0146	0.0101	0.0125	0.0146
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5036	0.4862	0.5016	0.5036	0.0257	0.0284	0.0253	0.0257
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5132	0.4861	0.5122	0.5132	0.0410	0.0464	0.0486	0.0410
raruidol/ArgumentMining-EN-AC-Financial	3.00E-05	256	2	561000000	encoder	debate_based_models	0.5316	0.4861	0.5201	0.5316	0.0206	0.0769	0.0480	0.0206
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5082	0.4859	0.5058	0.5082	0.0253	0.0229	0.0243	0.0253
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5015	0.4853	0.5017	0.5015	0.0227	0.0210	0.0228	0.0227
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5035	0.4852	0.5017	0.5035	0.0191	0.0148	0.0171	0.0191
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5100	0.4852	0.5080	0.5100	0.0201	0.0211	0.0216	0.0201
ProsusAI/finbert	5.00E-05	128	2	110000000	encoder	financial_base_models	0.5065	0.4850	0.5071	0.5065	0.0204	0.0271	0.0223	0.0204
ProsusAI/finbert	5.00E-05	128	2	110000000	encoder	financial_base_models	0.4985	0.4846	0.4977	0.4985	0.0142	0.0187	0.0149	0.0142
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5058	0.4845	0.5053	0.5058	0.0347	0.0340	0.0367	0.0347
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4995	0.4839	0.4990	0.4995	0.0150	0.0221	0.0148	0.0150
TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ	5.00E-05	256	3	7000000000	encoder_decoder	general_purpose_models	0.5237	0.4839	0.5213	0.5237	0.0098	0.0312	0.0169	0.0098
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5015	0.4833	0.5032	0.5015	0.0328	0.0340	0.0268	0.0328
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5020	0.4833	0.5039	0.5020	0.0206	0.0226	0.0207	0.0206
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5010	0.4832	0.5011	0.5010	0.0118	0.0119	0.0125	0.0118
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5001	0.4832	0.5002	0.5001	0.0138	0.0098	0.0131	0.0138
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5047	0.4831	0.5012	0.5047	0.0097	0.0125	0.0094	0.0097
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5034	0.4830	0.5020	0.5034	0.0115	0.0119	0.0106	0.0115
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4975	0.4829	0.4983	0.4975	0.0173	0.0167	0.0150	0.0173
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5096	0.4829	0.5054	0.5096	0.0307	0.0383	0.0367	0.0307
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5053	0.4828	0.5081	0.5053	0.0233	0.0219	0.0247	0.0233
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5015	0.4827	0.5048	0.5015	0.0162	0.0160	0.0162	0.0162
bert-base-uncased	5.00E-05	256	2	110000000	encoder	general_purpose_models	0.5294	0.4827	0.5462	0.5294	0.0094	0.0256	0.0173	0.0094
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.5025	0.4822	0.5029	0.5025	0.0259	0.0261	0.0252	0.0259
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5002	0.4821	0.4988	0.5002	0.0255	0.0220	0.0285	0.0255
roberta-base	3.00E-05	64	4	125000000	encoder	general_purpose_models	0.5467	0.4821	0.5878	0.5467	0.0264	0.0783	0.0199	0.0264
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5031	0.4820	0.5015	0.5031	0.0232	0.0264	0.0235	0.0232
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5032	0.4820	0.5017	0.5032	0.0291	0.0285	0.0309	0.0291
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5013	0.4818	0.5003	0.5013	0.0272	0.0261	0.0270	0.0272
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5064	0.4818	0.5043	0.5064	0.0214	0.0259	0.0201	0.0214
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5026	0.4818	0.5012	0.5026	0.0303	0.0317	0.0341	0.0303
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5056	0.4817	0.5117	0.5056	0.0126	0.0163	0.0164	0.0126
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5065	0.4817	0.5026	0.5065	0.0195	0.0191	0.0260	0.0195
bert-base-uncased	2.00E-05	256	2	110000000	encoder	general_purpose_models	0.4981	0.4816	0.4963	0.4981	0.0408	0.0409	0.0435	0.0408
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4981	0.4814	0.4983	0.4981	0.0320	0.0331	0.0310	0.0320
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5009	0.4812	0.5021	0.5009	0.0178	0.0181	0.0176	0.0178
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5067	0.4812	0.5050	0.5067	0.0120	0.0207	0.0125	0.0120
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5007	0.4812	0.5033	0.5007	0.0225	0.0253	0.0220	0.0225
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4984	0.4812	0.4981	0.4984	0.0187	0.0177	0.0194	0.0187
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5026	0.4811	0.4996	0.5026	0.0354	0.0372	0.0372	0.0354
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4955	0.4811	0.4968	0.4955	0.0268	0.0269	0.0272	0.0268
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5042	0.4810	0.5030	0.5042	0.0122	0.0141	0.0155	0.0122
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5075	0.4809	0.5087	0.5075	0.0160	0.0120	0.0179	0.0160
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4985	0.4807	0.4963	0.4985	0.0251	0.0276	0.0296	0.0251
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5008	0.4805	0.5001	0.5008	0.0280	0.0299	0.0319	0.0280
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5061	0.4804	0.5038	0.5061	0.0145	0.0163	0.0144	0.0145
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5009	0.4804	0.5028	0.5009	0.0205	0.0216	0.0203	0.0205
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.5007	0.4800	0.4976	0.5007	0.0117	0.0173	0.0135	0.0117
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5019	0.4799	0.5000	0.5019	0.0305	0.0256	0.0311	0.0305
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.5014	0.4798	0.4975	0.5014	0.0087	0.0085	0.0097	0.0087
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4978	0.4798	0.4982	0.4978	0.0329	0.0316	0.0328	0.0329
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5009	0.4796	0.4967	0.5009	0.0274	0.0230	0.0295	0.0274
TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ	5.00E-05	256	4	7000000000	encoder_decoder	general_purpose_models	0.5176	0.4795	0.5176	0.5176	0.0164	0.0267	0.0224	0.0164
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4978	0.4795	0.4971	0.4978	0.0231	0.0194	0.0249	0.0231
bert-base-uncased	3.00E-05	256	5	110000000	encoder	general_purpose_models	0.5130	0.4794	0.5112	0.5130	0.0199	0.0275	0.0315	0.0199
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5023	0.4793	0.5007	0.5023	0.0257	0.0327	0.0266	0.0257
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5051	0.4793	0.5028	0.5051	0.0232	0.0280	0.0225	0.0232
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4979	0.4792	0.4988	0.4979	0.0251	0.0278	0.0242	0.0251
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4992	0.4792	0.4953	0.4992	0.0283	0.0271	0.0279	0.0283
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4987	0.4792	0.4961	0.4987	0.0209	0.0197	0.0207	0.0209
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4951	0.4790	0.4944	0.4951	0.0178	0.0196	0.0204	0.0178
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4978	0.4790	0.5004	0.4978	0.0189	0.0208	0.0229	0.0189
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4994	0.4790	0.5002	0.4994	0.0207	0.0212	0.0263	0.0207
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5008	0.4789	0.4980	0.5008	0.0371	0.0384	0.0373	0.0371
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5058	0.4789	0.5113	0.5058	0.0208	0.0262	0.0267	0.0208
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5007	0.4789	0.5010	0.5007	0.0261	0.0186	0.0252	0.0261
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.5042	0.4789	0.5027	0.5042	0.0095	0.0213	0.0117	0.0095
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5013	0.4789	0.5031	0.5013	0.0309	0.0294	0.0364	0.0309
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5084	0.4787	0.5065	0.5084	0.0102	0.0227	0.0118	0.0102
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4960	0.4787	0.4976	0.4960	0.0141	0.0120	0.0136	0.0141
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5130	0.4787	0.5121	0.5130	0.0272	0.0399	0.0291	0.0272
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5017	0.4787	0.5010	0.5017	0.0311	0.0280	0.0310	0.0311
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5005	0.4786	0.5004	0.5005	0.0274	0.0315	0.0265	0.0274
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.5014	0.4785	0.5010	0.5014	0.0165	0.0212	0.0218	0.0165
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5013	0.4785	0.5015	0.5013	0.0264	0.0279	0.0279	0.0264
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5016	0.4785	0.5013	0.5016	0.0185	0.0198	0.0218	0.0185
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4975	0.4784	0.4991	0.4975	0.0346	0.0335	0.0384	0.0346
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5065	0.4784	0.5079	0.5065	0.0280	0.0306	0.0288	0.0280
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5076	0.4784	0.5093	0.5076	0.0232	0.0255	0.0234	0.0232
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	2.00E-05	64	3	561000000	encoder	debate_based_models	0.5278	0.4784	0.5348	0.5278	0.0130	0.0605	0.0109	0.0130
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	2.00E-05	64	5	561000000	encoder	debate_based_models	0.5318	0.4783	0.5400	0.5318	0.0156	0.0693	0.0208	0.0156
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5032	0.4783	0.4999	0.5032	0.0331	0.0312	0.0322	0.0331
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4953	0.4782	0.4983	0.4953	0.0281	0.0260	0.0310	0.0281
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4987	0.4782	0.4976	0.4987	0.0251	0.0305	0.0258	0.0251
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5024	0.4781	0.5022	0.5024	0.0217	0.0137	0.0272	0.0217
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5049	0.4781	0.5017	0.5049	0.0207	0.0262	0.0243	0.0207
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5004	0.4781	0.4974	0.5004	0.0147	0.0169	0.0173	0.0147
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4962	0.4780	0.4949	0.4962	0.0284	0.0270	0.0294	0.0284
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4980	0.4780	0.4954	0.4980	0.0200	0.0256	0.0269	0.0200
bert-base-uncased	2.00E-05	256	2	110000000	encoder	general_purpose_models	0.4969	0.4778	0.4963	0.4969	0.0332	0.0312	0.0312	0.0332
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4929	0.4778	0.4954	0.4929	0.0190	0.0184	0.0212	0.0190
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5015	0.4778	0.4960	0.5015	0.0232	0.0255	0.0287	0.0232
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4951	0.4777	0.4969	0.4951	0.0261	0.0258	0.0276	0.0261
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5040	0.4777	0.5025	0.5040	0.0137	0.0116	0.0176	0.0137
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5049	0.4775	0.5075	0.5049	0.0231	0.0287	0.0263	0.0231
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4971	0.4775	0.4952	0.4971	0.0156	0.0143	0.0181	0.0156
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5016	0.4774	0.5034	0.5016	0.0221	0.0203	0.0264	0.0221
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5046	0.4774	0.5015	0.5046	0.0261	0.0322	0.0309	0.0261
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5006	0.4773	0.4998	0.5006	0.0119	0.0115	0.0120	0.0119
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4986	0.4771	0.5033	0.4986	0.0212	0.0220	0.0206	0.0212
TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ	5.00E-05	256	4	7000000000	encoder_decoder	general_purpose_models	0.5193	0.4769	0.5140	0.5193	0.0153	0.0348	0.0192	0.0153
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4995	0.4769	0.5000	0.4995	0.0186	0.0270	0.0202	0.0186
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5061	0.4767	0.5067	0.5061	0.0100	0.0149	0.0092	0.0100
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4962	0.4767	0.4898	0.4962	0.0213	0.0250	0.0252	0.0213
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4977	0.4766	0.4956	0.4977	0.0122	0.0140	0.0134	0.0122
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5008	0.4766	0.4965	0.5008	0.0180	0.0225	0.0200	0.0180
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4957	0.4765	0.4974	0.4957	0.0231	0.0229	0.0226	0.0231
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4992	0.4763	0.4951	0.4992	0.0360	0.0310	0.0350	0.0360
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4964	0.4761	0.4986	0.4964	0.0193	0.0215	0.0204	0.0193
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4964	0.4760	0.4991	0.4964	0.0281	0.0274	0.0307	0.0281
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5055	0.4760	0.5019	0.5055	0.0427	0.0476	0.0470	0.0427
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.5012	0.4760	0.4974	0.5012	0.0241	0.0231	0.0296	0.0241
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4904	0.4759	0.4892	0.4904	0.0126	0.0161	0.0155	0.0126
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4942	0.4759	0.4951	0.4942	0.0415	0.0394	0.0422	0.0415
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4973	0.4759	0.4975	0.4973	0.0359	0.0424	0.0351	0.0359
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5037	0.4759	0.5084	0.5037	0.0200	0.0199	0.0197	0.0200
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5096	0.4759	0.5104	0.5096	0.0262	0.0218	0.0335	0.0262
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4963	0.4757	0.4925	0.4963	0.0136	0.0129	0.0138	0.0136
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4958	0.4757	0.4959	0.4958	0.0240	0.0235	0.0245	0.0240
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5028	0.4757	0.4953	0.5028	0.0175	0.0186	0.0197	0.0175
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4995	0.4756	0.4999	0.4995	0.0195	0.0166	0.0205	0.0195
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4964	0.4754	0.4991	0.4964	0.0228	0.0293	0.0273	0.0228
bert-base-uncased	3.00E-05	128	3	110000000	encoder	general_purpose_models	0.5028	0.4754	0.5019	0.5028	0.0263	0.0315	0.0307	0.0263
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4980	0.4754	0.4966	0.4980	0.0314	0.0326	0.0357	0.0314
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4945	0.4753	0.4902	0.4945	0.0209	0.0192	0.0224	0.0209
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5047	0.4753	0.5029	0.5047	0.0306	0.0252	0.0352	0.0306
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4982	0.4753	0.5015	0.4982	0.0263	0.0292	0.0286	0.0263
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4959	0.4753	0.4969	0.4959	0.0270	0.0218	0.0290	0.0270
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4939	0.4753	0.4931	0.4939	0.0158	0.0222	0.0184	0.0158
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5019	0.4751	0.5025	0.5019	0.0183	0.0228	0.0169	0.0183
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4972	0.4751	0.4983	0.4972	0.0155	0.0147	0.0150	0.0155
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4929	0.4749	0.4914	0.4929	0.0109	0.0117	0.0156	0.0109
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5043	0.4748	0.5025	0.5043	0.0147	0.0159	0.0157	0.0147
bert-base-uncased	2.00E-05	128	5	110000000	encoder	general_purpose_models	0.4997	0.4747	0.4951	0.4997	0.0109	0.0190	0.0198	0.0109
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4954	0.4747	0.4921	0.4954	0.0203	0.0202	0.0227	0.0203
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4994	0.4747	0.4979	0.4994	0.0266	0.0271	0.0283	0.0266
roberta-base	5.00E-05	64	4	125000000	encoder	general_purpose_models	0.5430	0.4747	0.5657	0.5430	0.0257	0.0622	0.0483	0.0257
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5023	0.4746	0.5040	0.5023	0.0214	0.0305	0.0195	0.0214
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4948	0.4745	0.4944	0.4948	0.0262	0.0213	0.0287	0.0262
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5064	0.4745	0.5037	0.5064	0.0231	0.0326	0.0318	0.0231
ProsusAI/finbert	5.00E-05	64	3	110000000	encoder	financial_base_models	0.5092	0.4745	0.5086	0.5092	0.0157	0.0340	0.0189	0.0157
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4943	0.4745	0.4926	0.4943	0.0143	0.0162	0.0158	0.0143
bert-base-uncased	3.00E-05	128	3	110000000	encoder	general_purpose_models	0.4967	0.4744	0.4920	0.4967	0.0207	0.0253	0.0245	0.0207
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5048	0.4744	0.5043	0.5048	0.0247	0.0213	0.0285	0.0247
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4930	0.4744	0.4921	0.4930	0.0290	0.0328	0.0287	0.0290
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4906	0.4743	0.4901	0.4906	0.0189	0.0187	0.0161	0.0189
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4953	0.4743	0.4924	0.4953	0.0329	0.0337	0.0334	0.0329
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4974	0.4743	0.4977	0.4974	0.0245	0.0286	0.0242	0.0245
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4977	0.4743	0.4957	0.4977	0.0242	0.0217	0.0240	0.0242
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5032	0.4742	0.4970	0.5032	0.0288	0.0334	0.0345	0.0288
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4944	0.4742	0.4953	0.4944	0.0195	0.0234	0.0234	0.0195
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4987	0.4742	0.5001	0.4987	0.0237	0.0312	0.0260	0.0237
bert-base-uncased	2.00E-05	256	2	110000000	encoder	general_purpose_models	0.4961	0.4741	0.4953	0.4961	0.0127	0.0100	0.0173	0.0127
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5028	0.4741	0.4925	0.5028	0.0186	0.0234	0.0235	0.0186
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4961	0.4740	0.4987	0.4961	0.0236	0.0251	0.0222	0.0236
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4971	0.4740	0.4965	0.4971	0.0253	0.0243	0.0272	0.0253
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4993	0.4739	0.5005	0.4993	0.0216	0.0271	0.0195	0.0216
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4974	0.4739	0.4964	0.4974	0.0441	0.0397	0.0459	0.0441
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4929	0.4739	0.4903	0.4929	0.0278	0.0279	0.0293	0.0278
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5070	0.4739	0.5042	0.5070	0.0103	0.0253	0.0129	0.0103

model_name	lr	max_len	epoch	model_size	model_architecture	model_type	mean_accuracy	mean_f1	mean_precision	mean_recall	std_accuracy	std_f1	std_precision	std_recall
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4950	0.4739	0.4919	0.4950	0.0278	0.0274	0.0293	0.0278
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4941	0.4738	0.4949	0.4941	0.0230	0.0220	0.0225	0.0230
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4988	0.4738	0.4986	0.4988	0.0385	0.0410	0.0431	0.0385
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5071	0.4738	0.5036	0.5071	0.0235	0.0190	0.0307	0.0235
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4943	0.4737	0.4883	0.4943	0.0289	0.0239	0.0277	0.0289
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4977	0.4735	0.4938	0.4977	0.0222	0.0295	0.0233	0.0222
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4921	0.4735	0.4930	0.4921	0.0294	0.0294	0.0276	0.0294
bert-base-uncased	2.00E-05	128	2	110000000	encoder	general_purpose_models	0.4892	0.4735	0.4900	0.4892	0.0162	0.0143	0.0154	0.0162
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4943	0.4735	0.4950	0.4943	0.0284	0.0253	0.0270	0.0284
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4897	0.4735	0.4908	0.4897	0.0280	0.0264	0.0285	0.0280
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4988	0.4735	0.5003	0.4988	0.0058	0.0082	0.0077	0.0058
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4972	0.4734	0.4967	0.4972	0.0490	0.0582	0.0521	0.0490
ProsusAI/finbert	5.00E-05	64	2	110000000	encoder	financial_base_models	0.5004	0.4734	0.4992	0.5004	0.0164	0.0302	0.0191	0.0164
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5095	0.4733	0.5065	0.5095	0.0204	0.0222	0.0207	0.0204
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4926	0.4733	0.4945	0.4926	0.0317	0.0259	0.0303	0.0317
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4994	0.4733	0.5026	0.4994	0.0305	0.0312	0.0327	0.0305
bert-base-uncased	2.00E-05	256	2	110000000	encoder	general_purpose_models	0.4913	0.4733	0.4861	0.4913	0.0356	0.0365	0.0371	0.0356
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4968	0.4733	0.4955	0.4968	0.0086	0.0117	0.0124	0.0086
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5053	0.4732	0.5088	0.5053	0.0160	0.0276	0.0155	0.0160
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4968	0.4732	0.4935	0.4968	0.0120	0.0137	0.0167	0.0120
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5007	0.4732	0.5020	0.5007	0.0271	0.0314	0.0277	0.0271
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4959	0.4732	0.4963	0.4959	0.0195	0.0151	0.0214	0.0195
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5003	0.4731	0.5003	0.5003	0.0353	0.0377	0.0426	0.0353
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5007	0.4731	0.4994	0.5007	0.0327	0.0337	0.0379	0.0327
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4927	0.4731	0.4911	0.4927	0.0378	0.0357	0.0391	0.0378
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4996	0.4731	0.4979	0.4996	0.0273	0.0271	0.0326	0.0273
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4898	0.4730	0.4900	0.4898	0.0265	0.0310	0.0260	0.0265
raruidol/ArgumentMining-EN-CN-ARI-Essay-Fin	5.00E-05	64	2	561000000	encoder	debate_based_models	0.5250	0.4730	0.5409	0.5250	0.0125	0.0343	0.0160	0.0125
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4967	0.4730	0.4970	0.4967	0.0071	0.0109	0.0057	0.0071
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4945	0.4729	0.4937	0.4945	0.0107	0.0145	0.0124	0.0107
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4987	0.4728	0.4999	0.4987	0.0122	0.0133	0.0129	0.0122
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4954	0.4728	0.4929	0.4954	0.0178	0.0237	0.0207	0.0178
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5041	0.4728	0.5017	0.5041	0.0171	0.0207	0.0153	0.0171
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4953	0.4727	0.4978	0.4953	0.0294	0.0263	0.0323	0.0294
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5003	0.4726	0.4986	0.5003	0.0081	0.0102	0.0082	0.0081
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5012	0.4726	0.4984	0.5012	0.0195	0.0314	0.0212	0.0195
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4934	0.4726	0.4953	0.4934	0.0235	0.0262	0.0253	0.0235
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4980	0.4726	0.4944	0.4980	0.0242	0.0217	0.0297	0.0242
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4943	0.4726	0.4928	0.4943	0.0166	0.0168	0.0168	0.0166
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4945	0.4726	0.4888	0.4945	0.0166	0.0214	0.0136	0.0166
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4971	0.4725	0.4961	0.4971	0.0191	0.0207	0.0179	0.0191
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4993	0.4725	0.5004	0.4993	0.0410	0.0461	0.0439	0.0410
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4893	0.4725	0.4894	0.4893	0.0177	0.0135	0.0192	0.0177
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4920	0.4725	0.4963	0.4920	0.0131	0.0087	0.0153	0.0131
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4937	0.4724	0.4940	0.4937	0.0118	0.0089	0.0101	0.0118
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4894	0.4724	0.4910	0.4894	0.0236	0.0238	0.0240	0.0236
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4895	0.4723	0.4907	0.4895	0.0255	0.0256	0.0257	0.0255
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4978	0.4723	0.4939	0.4978	0.0251	0.0334	0.0311	0.0251
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5021	0.4723	0.4946	0.5021	0.0061	0.0124	0.0108	0.0061
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5002	0.4723	0.5019	0.5002	0.0311	0.0287	0.0303	0.0311
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4993	0.4722	0.4936	0.4993	0.0166	0.0191	0.0236	0.0166
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4977	0.4722	0.4969	0.4977	0.0284	0.0291	0.0290	0.0284
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4894	0.4722	0.4880	0.4894	0.0410	0.0365	0.0414	0.0410
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4936	0.4722	0.4946	0.4936	0.0101	0.0127	0.0099	0.0101
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4950	0.4721	0.4951	0.4950	0.0202	0.0280	0.0225	0.0202
raruidol/ArgumentMining-EN-AC-Financial	5.00E-05	256	5	561000000	encoder	debate_based_models	0.5319	0.4721	0.5492	0.5319	0.0206	0.0476	0.0256	0.0206
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5008	0.4721	0.5035	0.5008	0.0457	0.0455	0.0495	0.0457
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5005	0.4720	0.4939	0.5005	0.0263	0.0232	0.0284	0.0263
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4953	0.4719	0.4963	0.4953	0.0201	0.0201	0.0219	0.0201
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5068	0.4718	0.5026	0.5068	0.0268	0.0248	0.0273	0.0268
TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ	5.00E-05	256	2	7000000000	encoder_decoder	general_purpose_models	0.5114	0.4718	0.5105	0.5114	0.0111	0.0239	0.0102	0.0111
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4996	0.4718	0.4977	0.4996	0.0206	0.0247	0.0246	0.0206
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4963	0.4716	0.4953	0.4963	0.0270	0.0268	0.0290	0.0270
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4930	0.4715	0.4899	0.4930	0.0198	0.0282	0.0242	0.0198
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4944	0.4715	0.4918	0.4944	0.0169	0.0220	0.0204	0.0169
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4976	0.4714	0.4986	0.4976	0.0277	0.0298	0.0301	0.0277
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4957	0.4714	0.4966	0.4957	0.0264	0.0230	0.0236	0.0264
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4919	0.4714	0.4857	0.4919	0.0264	0.0324	0.0340	0.0264
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4950	0.4713	0.4945	0.4950	0.0509	0.0524	0.0540	0.0509
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4918	0.4713	0.4906	0.4918	0.0236	0.0263	0.0307	0.0236
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4958	0.4713	0.4925	0.4958	0.0240	0.0275	0.0245	0.0240
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5018	0.4712	0.4955	0.5018	0.0266	0.0361	0.0335	0.0266
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5073	0.4711	0.5131	0.5073	0.0137	0.0200	0.0180	0.0137
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4941	0.4710	0.4898	0.4941	0.0170	0.0206	0.0222	0.0170
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4977	0.4709	0.4995	0.4977	0.0130	0.0174	0.0171	0.0130
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4943	0.4708	0.4873	0.4943	0.0140	0.0177	0.0149	0.0140
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4928	0.4707	0.4935	0.4928	0.0227	0.0213	0.0232	0.0227
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4972	0.4707	0.4963	0.4972	0.0197	0.0225	0.0210	0.0197
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4936	0.4706	0.4928	0.4936	0.0345	0.0313	0.0324	0.0345
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4957	0.4706	0.4964	0.4957	0.0099	0.0104	0.0126	0.0099
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4911	0.4705	0.4932	0.4911	0.0434	0.0363	0.0480	0.0434
bert-base-uncased	2.00E-05	128	5	110000000	encoder	general_purpose_models	0.4922	0.4704	0.4921	0.4922	0.0182	0.0277	0.0185	0.0182
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4957	0.4704	0.4968	0.4957	0.0162	0.0213	0.0196	0.0162
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4979	0.4703	0.4938	0.4979	0.0126	0.0260	0.0130	0.0126
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4985	0.4702	0.4967	0.4985	0.0111	0.0121	0.0161	0.0111
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4891	0.4702	0.4885	0.4891	0.0221	0.0254	0.0275	0.0221
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4912	0.4701	0.4871	0.4912	0.0249	0.0223	0.0299	0.0249
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4952	0.4701	0.4927	0.4952	0.0314	0.0305	0.0359	0.0314
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5057	0.4701	0.5071	0.5057	0.0220	0.0267	0.0325	0.0220
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4964	0.4700	0.4961	0.4964	0.0114	0.0175	0.0164	0.0114
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.5012	0.4699	0.4981	0.5012	0.0330	0.0379	0.0310	0.0330
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4882	0.4699	0.4873	0.4882	0.0247	0.0224	0.0292	0.0247
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4878	0.4698	0.4853	0.4878	0.0137	0.0073	0.0123	0.0137
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.5060	0.4698	0.5085	0.5060	0.0145	0.0215	0.0250	0.0145
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5090	0.4698	0.5063	0.5090	0.0173	0.0306	0.0167	0.0173
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4950	0.4697	0.4890	0.4950	0.0390	0.0360	0.0451	0.0390
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4859	0.4697	0.4854	0.4859	0.0096	0.0101	0.0078	0.0096
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5048	0.4697	0.4985	0.5048	0.0268	0.0292	0.0370	0.0268
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4957	0.4697	0.4942	0.4957	0.0276	0.0294	0.0278	0.0276
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4910	0.4696	0.4968	0.4910	0.0187	0.0153	0.0197	0.0187
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4961	0.4695	0.4894	0.4961	0.0298	0.0311	0.0374	0.0298
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4937	0.4694	0.4892	0.4937	0.0305	0.0339	0.0332	0.0305
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4894	0.4693	0.4903	0.4894	0.0232	0.0234	0.0239	0.0232
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4937	0.4692	0.4920	0.4937	0.0327	0.0314	0.0308	0.0327
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4987	0.4691	0.5055	0.4987	0.0199	0.0218	0.0162	0.0199
bert-base-uncased	3.00E-05	256	5	110000000	encoder	general_purpose_models	0.5011	0.4691	0.4962	0.5011	0.0238	0.0233	0.0187	0.0238
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4958	0.4691	0.4896	0.4958	0.0312	0.0298	0.0412	0.0312
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4951	0.4691	0.4885	0.4951	0.0246	0.0341	0.0290	0.0246
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4976	0.4691	0.4954	0.4976	0.0405	0.0430	0.0501	0.0405
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5001	0.4690	0.5030	0.5001	0.0140	0.0277	0.0172	0.0140
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4968	0.4689	0.4958	0.4968	0.0184	0.0201	0.0218	0.0184
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4926	0.4689	0.4930	0.4926	0.0167	0.0222	0.0170	0.0167
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4985	0.4688	0.4983	0.4985	0.0180	0.0109	0.0183	0.0180
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5006	0.4688	0.4943	0.5006	0.0122	0.0223	0.0172	0.0122
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4971	0.4687	0.4967	0.4971	0.0243	0.0205	0.0336	0.0243
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4898	0.4686	0.4883	0.4898	0.0266	0.0259	0.0330	0.0266
bert-base-uncased	3.00E-05	128	3	110000000	encoder	general_purpose_models	0.4968	0.4686	0.4923	0.4968	0.0233	0.0325	0.0326	0.0233
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.5010	0.4684	0.4970	0.5010	0.0164	0.0172	0.0207	0.0164
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4930	0.4683	0.4898	0.4930	0.0165	0.0097	0.0205	0.0165
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4948	0.4683	0.4916	0.4948	0.0230	0.0181	0.0239	0.0230
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4927	0.4683	0.4884	0.4927	0.0192	0.0247	0.0252	0.0192
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4934	0.4682	0.4953	0.4934	0.0180	0.0182	0.0240	0.0180
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4966	0.4682	0.4929	0.4966	0.0233	0.0295	0.0255	0.0233
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4882	0.4681	0.4860	0.4882	0.0208	0.0250	0.0234	0.0208
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4972	0.4681	0.4972	0.4972	0.0216	0.0215	0.0277	0.0216
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4992	0.4680	0.4973	0.4992	0.0269	0.0228	0.0333	0.0269
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4918	0.4679	0.4935	0.4918	0.0171	0.0180	0.0159	0.0171
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4879	0.4679	0.4857	0.4879	0.0343	0.0356	0.0322	0.0343
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4914	0.4678	0.4876	0.4914	0.0192	0.0110	0.0194	0.0192
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4871	0.4678	0.4843	0.4871	0.0206	0.0234	0.0260	0.0206
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4999	0.4678	0.4980	0.4999	0.0327	0.0371	0.0362	0.0327
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.5010	0.4678	0.4933	0.5010	0.0334	0.0349	0.0365	0.0334
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4996	0.4677	0.5009	0.4996	0.0248	0.0246	0.0257	0.0248
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4864	0.4677	0.4856	0.4864	0.0096	0.0079	0.0096	0.0096
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4853	0.4677	0.4843	0.4853	0.0265	0.0296	0.0268	0.0265
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4913	0.4677	0.4900	0.4913	0.0256	0.0217	0.0194	0.0256
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4876	0.4676	0.4860	0.4876	0.0172	0.0196	0.0196	0.0172
bert-base-uncased	2.00E-05	256	2	110000000	encoder	general_purpose_models	0.4967	0.4676	0.4983	0.4967	0.0253	0.0191	0.0275	0.0253
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.5080	0.4676	0.5074	0.5080	0.0164	0.0396	0.0312	0.0164
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5031	0.4675	0.5011	0.5031	0.0282	0.0269	0.0303	0.0282
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4955	0.4674	0.4882	0.4955	0.0233	0.0256	0.0245	0.0233
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4972	0.4674	0.4965	0.4972	0.0319	0.0265	0.0342	0.0319
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4906	0.4673	0.4926	0.4906	0.0182	0.0187	0.0161	0.0182
TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ	5.00E-05	64	2	7000000000	encoder_decoder	general_purpose_models	0.5134	0.4673	0.5066	0.5134	0.0155	0.0317	0.0234	0.0155
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4883	0.4673	0.4879	0.4883	0.0263	0.0295	0.0359	0.0263
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4933	0.4671	0.4894	0.4933	0.0261	0.0276	0.0349	0.0261
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5023	0.4670	0.4991	0.5023	0.0139	0.0310	0.0213	0.0139
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4850	0.4670	0.4854	0.4850	0.0147	0.0101	0.0161	0.0147
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4928	0.4669	0.4894	0.4928	0.0126	0.0171	0.0123	0.0126
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4990	0.4669	0.5061	0.4990	0.0294	0.0255	0.0368	0.0294
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5025	0.4669	0.4993	0.5025	0.0181	0.0306	0.0236	0.0181
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5021	0.4668	0.4933	0.5021	0.0179	0.0273	0.0203	0.0179
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4969	0.4667	0.5010	0.4969	0.0348	0.0315	0.0398	0.0348
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4915	0.4667	0.4888	0.4915	0.0396	0.0363	0.0447	0.0396
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4929	0.4667	0.4851	0.4929	0.0203	0.0187	0.0200	0.0203
bert-base-uncased	2.00E-05	128	5	110000000	encoder	general_purpose_models	0.4937	0.4666	0.4953	0.4937	0.0216	0.0266	0.0197	0.0216
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5004	0.4666	0.4951	0.5004	0.0272	0.0248	0.0322	0.0272
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4960	0.4666	0.4936	0.4960	0.0157	0.0223	0.0151	0.0157
bert-base-uncased	3.00E-05	128	4	110000000	encoder	general_purpose_models	0.4999	0.4666	0.5018	0.4999	0.0213	0.0281	0.0268	0.0213
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4903	0.4665	0.4914	0.4903	0.0228	0.0302	0.0270	0.0228
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5037	0.4665	0.5062	0.5037	0.0154	0.0210	0.0215	0.0154
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4989	0.4664	0.4969	0.4989	0.0189	0.0226	0.0241	0.0189
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4870	0.4664	0.4843	0.4870	0.0200	0.0179	0.0188	0.0200
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4908	0.4664	0.4942	0.4908	0.0264	0.0205	0.0282	0.0264
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4904	0.4663	0.4854	0.4904	0.0245	0.0233	0.0275	0.0245
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4889	0.4663	0.4838	0.4889	0.0155	0.0191	0.0185	0.0155
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4848	0.4662	0.4822	0.4848	0.0223	0.0249	0.0278	0.0223
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4884	0.4662	0.4864	0.4884	0.0287	0.0280	0.0292	0.0287
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4993	0.4662	0.4953	0.4993	0.0291	0.0378	0.0343	0.0291
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4904	0.4661	0.4865	0.4904	0.0273	0.0253	0.0301	0.0273
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4913	0.4661	0.4915	0.4913	0.0169	0.0227	0.0237	0.0169
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4888	0.4661	0.4877	0.4888	0.0084	0.0087	0.0149	0.0084
bert-base-uncased	3.00E-05	128	5	110000000	encoder	general_purpose_models	0.4933	0.4661	0.4881	0.4933	0.0174	0.0268	0.0230	0.0174
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5032	0.4660	0.4931	0.5032	0.0160	0.0216	0.0224	0.0160
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4876	0.4660	0.4860	0.4876	0.0313	0.0351	0.0342	0.0313
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5011	0.4660	0.4951	0.5011	0.0259	0.0263	0.0407	0.0259

model_name	lr	max_len	epoch	model_size	model_architecture	model_type	mean_accuracy	mean_f1	mean_precision	mean_recall	std_accuracy	std_f1	std_precision	std_recall
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4951	0.4659	0.4922	0.4951	0.0261	0.0325	0.0277	0.0261
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4976	0.4659	0.4955	0.4976	0.0280	0.0179	0.0314	0.0280
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4945	0.4659	0.4889	0.4945	0.0120	0.0147	0.0119	0.0120
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4964	0.4658	0.4884	0.4964	0.0245	0.0225	0.0291	0.0245
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4924	0.4658	0.4903	0.4924	0.0198	0.0273	0.0233	0.0198
bert-base-uncased	2.00E-05	128	2	110000000	encoder	general_purpose_models	0.5014	0.4657	0.5000	0.5014	0.0121	0.0152	0.0162	0.0121
bert-base-uncased	3.00E-05	256	5	110000000	encoder	general_purpose_models	0.4984	0.4657	0.5007	0.4984	0.0312	0.0346	0.0324	0.0312
bert-base-uncased	3.00E-05	256	5	110000000	encoder	general_purpose_models	0.4984	0.4656	0.5006	0.4984	0.0217	0.0230	0.0284	0.0217
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4945	0.4656	0.4950	0.4945	0.0376	0.0465	0.0423	0.0376
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5012	0.4656	0.4990	0.5012	0.0186	0.0244	0.0260	0.0186
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4932	0.4656	0.4928	0.4932	0.0199	0.0076	0.0166	0.0199
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4875	0.4654	0.4825	0.4875	0.0372	0.0375	0.0389	0.0372
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4970	0.4653	0.4901	0.4970	0.0211	0.0261	0.0249	0.0211
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4948	0.4652	0.4946	0.4948	0.0121	0.0248	0.0155	0.0121
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4988	0.4652	0.4889	0.4988	0.0255	0.0360	0.0302	0.0255
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4908	0.4652	0.4870	0.4908	0.0312	0.0304	0.0349	0.0312
bert-base-uncased	3.00E-05	128	4	110000000	encoder	general_purpose_models	0.5000	0.4652	0.4987	0.5000	0.0265	0.0322	0.0378	0.0265
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5023	0.4651	0.5023	0.5023	0.0160	0.0282	0.0211	0.0160
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4935	0.4650	0.4897	0.4935	0.0346	0.0354	0.0401	0.0346
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4927	0.4649	0.4897	0.4927	0.0098	0.0132	0.0186	0.0098
bert-base-uncased	2.00E-05	128	4	110000000	encoder	general_purpose_models	0.4946	0.4649	0.4912	0.4946	0.0220	0.0301	0.0226	0.0220
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4976	0.4649	0.4886	0.4976	0.0201	0.0277	0.0210	0.0201
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4881	0.4649	0.4883	0.4881	0.0249	0.0295	0.0287	0.0249
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5003	0.4648	0.5006	0.5003	0.0134	0.0155	0.0147	0.0134
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4920	0.4648	0.4852	0.4920	0.0207	0.0171	0.0188	0.0207
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4895	0.4645	0.4935	0.4895	0.0247	0.0344	0.0239	0.0247
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.5013	0.4643	0.5054	0.5013	0.0138	0.0213	0.0199	0.0138
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4898	0.4643	0.4856	0.4898	0.0239	0.0368	0.0315	0.0239
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4867	0.4643	0.4853	0.4867	0.0141	0.0134	0.0158	0.0141
bert-base-uncased	3.00E-05	128	4	110000000	encoder	general_purpose_models	0.5006	0.4642	0.4957	0.5006	0.0157	0.0219	0.0176	0.0157
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4939	0.4642	0.4878	0.4939	0.0328	0.0304	0.0344	0.0328
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4930	0.4642	0.4966	0.4930	0.0294	0.0229	0.0341	0.0294
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4921	0.4638	0.4909	0.4921	0.0219	0.0300	0.0226	0.0219
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4995	0.4638	0.4963	0.4995	0.0191	0.0322	0.0297	0.0191
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5016	0.4638	0.4987	0.5016	0.0076	0.0142	0.0165	0.0076
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4976	0.4637	0.4980	0.4976	0.0218	0.0335	0.0252	0.0218
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4972	0.4637	0.4936	0.4972	0.0086	0.0192	0.0108	0.0086
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4994	0.4637	0.4963	0.4994	0.0241	0.0213	0.0270	0.0241
bert-base-uncased	3.00E-05	256	2	110000000	encoder	general_purpose_models	0.4985	0.4635	0.4946	0.4985	0.0165	0.0164	0.0163	0.0165
bert-base-uncased	3.00E-05	128	2	110000000	encoder	general_purpose_models	0.4997	0.4635	0.4948	0.4997	0.0179	0.0274	0.0297	0.0179
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4978	0.4634	0.4912	0.4978	0.0212	0.0245	0.0240	0.0212
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4971	0.4633	0.4929	0.4971	0.0129	0.0203	0.0183	0.0129
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4900	0.4633	0.4844	0.4900	0.0432	0.0346	0.0413	0.0432
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4963	0.4633	0.4964	0.4963	0.0181	0.0172	0.0219	0.0181
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4992	0.4633	0.4921	0.4992	0.0150	0.0176	0.0179	0.0150
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4827	0.4632	0.4824	0.4827	0.0255	0.0202	0.0276	0.0255
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4918	0.4631	0.4911	0.4918	0.0141	0.0141	0.0180	0.0141
bert-base-uncased	2.00E-05	128	5	110000000	encoder	general_purpose_models	0.4945	0.4630	0.4935	0.4945	0.0135	0.0187	0.0252	0.0135
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4960	0.4630	0.4906	0.4960	0.0187	0.0351	0.0338	0.0187
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4994	0.4629	0.4952	0.4994	0.0239	0.0377	0.0290	0.0239
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4936	0.4629	0.4998	0.4936	0.0379	0.0302	0.0420	0.0379
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4937	0.4628	0.4879	0.4937	0.0185	0.0147	0.0146	0.0185
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4955	0.4627	0.5063	0.4955	0.0153	0.0194	0.0260	0.0153
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5091	0.4626	0.5009	0.5091	0.0120	0.0278	0.0124	0.0120
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4890	0.4626	0.4888	0.4890	0.0282	0.0263	0.0289	0.0282
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4944	0.4625	0.4876	0.4944	0.0208	0.0254	0.0260	0.0208
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4963	0.4625	0.4990	0.4963	0.0206	0.0233	0.0259	0.0206
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4976	0.4624	0.4893	0.4976	0.0382	0.0562	0.0459	0.0382
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4914	0.4623	0.4929	0.4914	0.0162	0.0200	0.0194	0.0162
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4897	0.4623	0.4894	0.4897	0.0176	0.0249	0.0232	0.0176
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4904	0.4622	0.4864	0.4904	0.0099	0.0161	0.0140	0.0099
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4906	0.4621	0.4858	0.4906	0.0300	0.0329	0.0393	0.0300
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5058	0.4620	0.5046	0.5058	0.0150	0.0271	0.0241	0.0150
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4915	0.4620	0.4902	0.4915	0.0264	0.0310	0.0305	0.0264
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4839	0.4619	0.4833	0.4839	0.0134	0.0145	0.0150	0.0134
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4945	0.4619	0.4857	0.4945	0.0221	0.0337	0.0265	0.0221
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4965	0.4618	0.4968	0.4965	0.0094	0.0206	0.0098	0.0094
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4962	0.4618	0.4920	0.4962	0.0172	0.0281	0.0261	0.0172
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4839	0.4618	0.4846	0.4839	0.0288	0.0295	0.0334	0.0288
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4966	0.4618	0.4904	0.4966	0.0173	0.0158	0.0178	0.0173
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4981	0.4617	0.4897	0.4981	0.0220	0.0348	0.0271	0.0220
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4898	0.4617	0.4927	0.4898	0.0291	0.0346	0.0342	0.0291
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4916	0.4617	0.4855	0.4916	0.0304	0.0326	0.0376	0.0304
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4958	0.4615	0.4966	0.4958	0.0155	0.0192	0.0167	0.0155
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4913	0.4614	0.4878	0.4913	0.0144	0.0144	0.0168	0.0144
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5010	0.4613	0.5001	0.5010	0.0208	0.0251	0.0304	0.0208
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4902	0.4613	0.4855	0.4902	0.0350	0.0358	0.0372	0.0350
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4962	0.4613	0.4852	0.4962	0.0068	0.0186	0.0110	0.0068
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4915	0.4612	0.4908	0.4915	0.0353	0.0300	0.0415	0.0353
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5040	0.4612	0.5005	0.5040	0.0244	0.0448	0.0445	0.0244
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4967	0.4611	0.4995	0.4967	0.0273	0.0141	0.0295	0.0273
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4921	0.4611	0.4893	0.4921	0.0237	0.0226	0.0230	0.0237
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4906	0.4607	0.4822	0.4906	0.0246	0.0336	0.0310	0.0246
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4922	0.4607	0.4948	0.4922	0.0245	0.0279	0.0290	0.0245
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5039	0.4606	0.5074	0.5039	0.0162	0.0316	0.0226	0.0162
bert-base-uncased	2.00E-05	256	2	110000000	encoder	general_purpose_models	0.5036	0.4605	0.4967	0.5036	0.0244	0.0127	0.0292	0.0244
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4963	0.4604	0.4915	0.4963	0.0258	0.0375	0.0318	0.0258
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4934	0.4603	0.4860	0.4934	0.0228	0.0267	0.0346	0.0228
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.5007	0.4602	0.4900	0.5007	0.0263	0.0437	0.0412	0.0263
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4900	0.4602	0.4810	0.4900	0.0136	0.0210	0.0223	0.0136
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4897	0.4602	0.4878	0.4897	0.0210	0.0309	0.0208	0.0210
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4964	0.4602	0.4913	0.4964	0.0179	0.0248	0.0271	0.0179
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4893	0.4602	0.4892	0.4893	0.0241	0.0337	0.0267	0.0241
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.5035	0.4601	0.5011	0.5035	0.0161	0.0346	0.0249	0.0161
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4924	0.4601	0.4919	0.4924	0.0236	0.0286	0.0265	0.0236
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4854	0.4600	0.4804	0.4854	0.0197	0.0184	0.0215	0.0197
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4853	0.4600	0.4787	0.4853	0.0088	0.0153	0.0147	0.0088
bert-base-uncased	2.00E-05	256	3	110000000	encoder	general_purpose_models	0.4842	0.4599	0.4812	0.4842	0.0153	0.0176	0.0181	0.0153
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4913	0.4599	0.4871	0.4913	0.0179	0.0158	0.0199	0.0179
raruidol/ArgumentMining-EN-AC-Financial	5.00E-05	64	4	561000000	encoder	debate_based_models	0.5273	0.4599	0.5508	0.5273	0.0139	0.0451	0.0196	0.0139
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4945	0.4598	0.4897	0.4945	0.0180	0.0262	0.0329	0.0180
bert-base-uncased	3.00E-05	64	2	110000000	encoder	general_purpose_models	0.5171	0.4598	0.5265	0.5171	0.0097	0.0356	0.0097	0.0097
NLPScholars/Roberta-Earning-Call-Transcript-Classification	3.00E-05	64	5	125000000	encoder_decoder	financial_base_models	0.5120	0.4597	0.4646	0.5120	0.0074	0.0650	0.1074	0.0074
bert-base-uncased	3.00E-05	256	5	110000000	encoder	general_purpose_models	0.4978	0.4597	0.4934	0.4978	0.0244	0.0387	0.0286	0.0244
raruidol/ArgumentMining-EN-AC-Financial	5.00E-05	64	5	561000000	encoder	debate_based_models	0.5277	0.4596	0.5529	0.5277	0.0133	0.0349	0.0246	0.0133
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4959	0.4596	0.4851	0.4959	0.0265	0.0415	0.0337	0.0265
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4866	0.4595	0.4833	0.4866	0.0209	0.0155	0.0251	0.0209
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4778	0.4594	0.4743	0.4778	0.0274	0.0288	0.0297	0.0274
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4838	0.4594	0.4795	0.4838	0.0071	0.0152	0.0079	0.0071
bert-base-uncased	3.00E-05	256	5	110000000	encoder	general_purpose_models	0.4867	0.4593	0.4890	0.4867	0.0236	0.0211	0.0267	0.0236
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.5023	0.4592	0.5018	0.5023	0.0326	0.0402	0.0367	0.0326
bert-base-uncased	2.00E-05	128	3	110000000	encoder	general_purpose_models	0.4909	0.4592	0.4821	0.4909	0.0344	0.0389	0.0392	0.0344
bert-base-uncased	2.00E-05	256	5	110000000	encoder	general_purpose_models	0.4937	0.4591	0.4856	0.4937	0.0207	0.0272	0.0224	0.0207
bert-base-uncased	3.00E-05	256	4	110000000	encoder	general_purpose_models	0.4836	0.4591	0.4803	0.4836	0.0321	0.0326	0.0370	0.0321
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4851	0.4591	0.4804	0.4851	0.0171	0.0253	0.0239	0.0171
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4915	0.4590	0.4920	0.4915	0.0268	0.0221	0.0332	0.0268
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4866	0.4590	0.4808	0.4866	0.0271	0.0290	0.0265	0.0271
bert-base-uncased	2.00E-05	256	4	110000000	encoder	general_purpose_models	0.4918	0.4589	0.4940	0.4918	0.0105	0.0088	0.0169	0.0105
bert-base-uncased	2.00E-05	128	5	110000000	encoder	general_purpose_models	0.4882	0.4588	0.4805	0.4882	0.0212	0.0310	0.0260	0.0212
bert-base-uncased	3.00E-05	256	3	110000000	encoder	general_purpose_models	0.4893	0.4587	0.4871	0.4893	0.0229	0.0340	0.0328	0.0229
meta-llama/Meta-Llama-3-8B	0.00005	256	5	8000000000	encoder_decoder	general_purpose_models	0.6426	0.6386	0.6439	0.6426	0.0267	0.0281	0.0279	0.0267
meta-llama/Meta-Llama-3-8B	0.00005	256	5	8000000000	encoder_decoder	general_purpose_models	0.6409	0.6367	0.6428	0.6409	0.0351	0.0360	0.0362	0.0351
meta-llama/Meta-Llama-3-8B	0.00005	256	5	8000000000	encoder_decoder	general_purpose_models	0.6401	0.6357	0.6427	0.6401	0.0316	0.0322	0.0336	0.0316
meta-llama/Meta-Llama-3-8B	0.00005	256	5	8000000000	encoder_decoder	general_purpose_models	0.6386	0.6345	0.6396	0.6386	0.0289	0.0295	0.0303	0.0289
meta-llama/Meta-Llama-3-8B	0.00005	256	5	8000000000	encoder_decoder	general_purpose_models	0.6358	0.6306	0.6375	0.6358	0.0299	0.0309	0.0313	0.0299
meta-llama/Meta-Llama-3-8B	0.00005	256	5	8000000000	encoder_decoder	general_purpose_models	0.6345	0.6293	0.6387	0.6345	0.0285	0.0283	0.0316	0.0285
meta-llama/Meta-Llama-3-8B	0.00003	128	5	8000000000	encoder_decoder	general_purpose_models	0.6235	0.6114	0.6315	0.6235	0.0349	0.0423	0.0347	0.0349
meta-llama/Meta-Llama-3-8B	0.00003	128	4	8000000000	encoder_decoder	general_purpose_models	0.6062	0.5962	0.6124	0.6062	0.0306	0.0367	0.0294	0.0306
meta-llama/Meta-Llama-3-8B	0.00003	128	5	8000000000	encoder_decoder	general_purpose_models	0.6062	0.5947	0.6130	0.6062	0.0370	0.0435	0.0374	0.0370
meta-llama/Meta-Llama-3-8B	0.00002	128	5	8000000000	encoder_decoder	general_purpose_models	0.5982	0.5864	0.6053	0.5982	0.0412	0.0478	0.0437	0.0412
meta-llama/Meta-Llama-3-8B	0.00002	128	5	8000000000	encoder_decoder	general_purpose_models	0.5963	0.5860	0.6026	0.5963	0.0329	0.0360	0.0335	0.0329
meta-llama/Meta-Llama-3-8B	0.00005	64	4	8000000000	encoder_decoder	general_purpose_models	0.5845	0.5835	0.5847	0.5845	0.0286	0.0289	0.0285	0.0286
meta-llama/Meta-Llama-3-8B	0.00002	128	4	8000000000	encoder_decoder	general_purpose_models	0.5869	0.5774	0.5920	0.5869	0.0194	0.0224	0.0186	0.0194
meta-llama/Meta-Llama-3-8B	0.00005	64	5	8000000000	encoder_decoder	general_purpose_models	0.5719	0.5704	0.5722	0.5719	0.0422	0.0420	0.0423	0.0422
meta-llama/Meta-Llama-3-8B	0.00002	128	3	8000000000	encoder_decoder	general_purpose_models	0.5800	0.5703	0.5843	0.5800	0.0160	0.0186	0.0171	0.0160
meta-llama/Meta-Llama-3-8B	0.00003	256	3	8000000000	encoder_decoder	general_purpose_models	0.5800	0.5674	0.5841	0.5800	0.0373	0.0472	0.0389	0.0373
meta-llama/Meta-Llama-3-8B	0.00005	64	4	8000000000	encoder_decoder	general_purpose_models	0.5624	0.5613	0.5625	0.5624	0.0347	0.0349	0.0347	0.0347
meta-llama/Meta-Llama-3-8B	0.00003	128	3	8000000000	encoder_decoder	general_purpose_models	0.5454	0.5439	0.5455	0.5454	0.0356	0.0363	0.0357	0.0356
meta-llama/Meta-Llama-3-8B	0.00002	128	2	8000000000	encoder_decoder	general_purpose_models	0.5171	0.5156	0.5171	0.5171	0.0182	0.0188	0.0183	0.0182

List of Acronyms

NLP Natural Language Processing

AI Artificial Intelligence

AM Argumentation Mining

LM Language Modeling

BERT Bidirectional Encoder Representations from Transformers

LLM Large Language Model

GPT-4 Generative Pre-trained Transformer 4

LlaMA-3 Large Language Model Meta AI 3

ECCs Earnings Conference Calls

FinNLP Financial Natural Language Processing

WANDB Weights and Biases

TP True Positive

TN True Negative

FP False Positive

FN False Negative

List of Figures

List of Tables

Bibliography

[1] S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011, 2012.

[2] Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. It's time to reason: Annotating argumentation structures in financial earnings calls: The finarg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169, 2022.

[3] Alaa Alhamzeh. Financial argument quality assessment in earnings conference calls. In *International Conference on Database and Expert Systems Applications*, pages 65–81. Springer, 2023.

[4] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. pages 567–578, July 2019.

[5] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, 2017.

[6] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. *From Opinion Mining to Financial Argument Mining*. Springer Nature, 2021.

[7] Mahmoud El-Haj, Paul Rayson, and Andrew Moore. The first financial narrative processing workshop (fnp 2018). In *Proceedings of the LREC 2018 Workshop*, 2018.

[8] Sameena Shah, Xiaodan Zhu, Wenhu Chen, Manling Li, Armineh Nourbakhsh, Xiaomo Liu, Zhiqiang Ma, Charese Smiley, Yulong Pei, and Akshat Gupta. Knowledge discovery from unstructured data in financial services (kdf) workshop. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3464–3467, 2023.

[9] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-17 finarg-1 task: Fine-grained argument understanding in financial analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*, 2023.

[10] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[11] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6106–6110, 2020.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

[14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023.

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[16] Marco Lippi and Paolo Torroni. Argumentation mining: A survey. *Computing Research Repository*, abs/1604.03103, 2016.

[17] John Lawrence and Chris Reed. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818, 01 2020.

[18] Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. It's time to reason: Annotating argumentation structures in financial earnings calls: The FinArg dataset. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors, *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[19] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. *The Stanford CoreNLP Natural Language Processing Toolkit*. Association for Computational Linguistics, 2014.

[20] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: The detection, classification and structure of arguments in text. *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, page 98–107, 2009.

[21] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In *ACL*, 2017.

[22] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[23] Ankur Sharma, Nguyen Ha Thanh, and Ondřej Bojar. Do language models understand anything? on the ability of llms to understand negative questions. *arXiv*, abs/2109.00826, 2021.

[24] Zechun Liu et al. Llm-qat: Data-free quantization aware training for large language models. *ArXiv*, 2023.

[25] Jangwhan Lee et al. Enhancing computation efficiency in large language models through weight and activation quantization. *ArXiv*, 2023.

[26] Somnath Roy. Understanding the impact of post-training quantization on large language models. *ArXiv*, 2023.

[27] Jacob Benesty, Jingdong Chen, and Yiteng Huang. Pearson correlation coefficient. *Noise Reduction in Speech Processing*, pages 1–4, 2009.

[28] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

[29] Weights & biases. https://wandb.ai/site. Accessed: yyyy-mm-dd.

[30] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[31] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[32] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[33] Umer Mushtaq and Jérémie Cabessa. Argument mining with modular bert and transfer learning. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023.

[34] Zhen-Quan Tang, Benyou Wang, and Ting Yao. Dptdr: Deep prompt tuning for dense passage retrieval. 2022.

[35] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. pages 61–68, 2022.

[36] Samin Mohammadi and Mathieu Chapon. Investigating the performance of fine-tuned text classification models based-on bert. In *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1252–1257, 2020.

[37] Lili Sun and Zhenquan Shi. Prompt learning under the large language model. In *2023 International Seminar on Computer Science and Engineering Technology (SCSET)*, pages 288–291, 2023.

[38] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[40] Andrea Galassi. *Deep Networks and Knowledge: from Rule Learning to Neural-Symbolic Argument Mining*. PhD thesis, University of Bologna, 2021.

[41] Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. Will it blend? mixing training paradigms & prompting for argument quality prediction. pages 95–103, 2022.

[42] Ahnaf Mozib Samin, Behrooz Nikandish, and Jingyan Chen. Arguments to key points mapping with prompt-based learning. pages 303–311, 2022.

[43] Rabeeh Karimi Mahabadi, Luke Zettlemoyer, J. Henderson, Marzieh Saeidi, Lambert Mathias, Ves Stoyanov, and Majid Yazdani. Prompt-free and efficient few-shot learning with language models. 2022.

[44] Mrigank Raman, Pratyush Maini, J. Z. Kolter, Zachary Chase Lipton, and Danish Pruthi. Model-tuning via prompts makes nlp models adversarially robust. 2023.

[45] Savas Yildirim, Mucahit Cevik, D. Parikh, and Ayse Basar. Adaptive fine-tuning for multiclass classification over software requirement data. 2023.

[46] Ahmed R. Abas, I. El-Henawy, H. Mohamed, and Amr Abdellatif. Deep learning model for fine-grained aspect-based opinion mining. *IEEE Access*, 8:128845–128855, 2020.

[47] Amirhossein Farzam, Shashank Shekhar, Isaac Mehlhaff, and Marco Morucci. Multi-task learning improves performance in deep argument mining models. 2023.

[48] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. 2021.

[49] Xu Guo, Boyang Albert Li, and Han Yu. Improving the sample efficiency of prompt tuning with domain adaptation. 2022.

[50] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063, 2019.

[51] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: A machine learning approach to automatic structuring of arguments. *The Knowledge Engineering Review*, 26(4):365–386, 2011.

[52] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2020.

[53] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *ArXiv*, abs/1909.01066, 2019.

[54] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021.

[55] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56:1 – 40, 2021.

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

[57] Douglas Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, 1996.

[58] James B. Freeman. Argument structure representation: The role of assumption in inference evaluation. *Argumentation*, 25:127–137, 2011.

[59] Chris Reed and Derek Long. A.r.a.: A framework and system for argument analysis. *Argumentation*, 19(3):267–286, 2003.

[60] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[61] Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171:619–641, 2007.

[62] Jaap C. Hage. *Reasoning with Unstated Premises in Discretionary Law Application*. Kluwer Academic Publishers, 1997.

[63] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25, 2018.

[64] Marie-Francine Moens. Automatic analysis of argumentation in legal cases. *ACM SIGKDD Explorations Newsletter*, 9:20–28, 2007.

[65] Raul Puri and Bryan Catanzaro. Zero-shot text classification with generative language models. *ArXiv*, abs/1912.10165, 2019.

[66] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining. In *ACM International Conference Proceeding Series*, 2009.

[67] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

[68] John Lawrence and Chris Reed. Argument mining: A survey. *Computer Linguistics*, 45(4):765–818, 2020.

[69] Huy Nguyen and Diane J. Litman. Extracting argument and domain words for identifying argument components in texts. In *NAACL HLT*, 2015.

[70] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[71] R Thomas McCoy, Junghyun Min, and Tal Linzen. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*, 2019.

[72] Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 2019.

[73] R. Ruiz-Dolz, J. Alemany, S. Barbera, and A. Garcia-Fornes. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(06):62–70, nov 2021.

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

[75] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[76] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[77] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35, 2021.

[78] Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, 6, 2023.

[79] Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*, 2023.

[80] Martin Hinton and Jean HM Wagemans. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the gpt-3 ai text generator. *Argument & Computation*, (Preprint):1–16, 2023.

[81] Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. Breaking the bank with ChatGPT: Few-shot text classification for finance. In Chung-Chi Chen, Hiroya

Takamura, Puneet Mathur, Remit Sawhney, Hen-Hsen Huang, and Hsin-Hsi Chen, editors, *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 74–80, Macao, 20 August 2023. -.

[82] Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In Mingxuan Wang and Imed Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore, December 2023. Association for Computational Linguistics.

[83] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *ArXiv*, abs/1909.00161, 2019.

[84] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *ArXiv*, abs/2009.07118, 2020.

[85] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics, 2018.

[86] Alaa Alhamzeh. *Language Reasoning by means of Argument Mining and Argument Quality*. PhD thesis, Universität Passau, 2023.

[87] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[88] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[89] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[90] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

[91] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[92] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[93] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[94] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[95] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. On the effect of sample and topic sizes for argument mining datasets. *arXiv preprint arXiv:2205.11472*, 2022.

[96] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[97] Ahmed Hazourli. Financialbert-a pretrained language model for financial text mining. *Research Gate*, 2, 2022.

[98] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021.

[99] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[100] AI@Meta. Llama 3 model card, 2024.

[101] Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[102] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4765–4774, 2017.

[103] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[104] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, 2019.