Final Year Dissertation Paper

Using Large Language Models to Match Request for Quotations to Suppliers in a Business-to-Business Supply Chain Procurement Platform

Matthew Lewis
University of Exeter

Abstract

Despite the significant technological advancements which have been introduced to the consumer sales sector, the Business-to-Business (B2B) sector has lagged behind in adopting new technologies. This has resulted in the majority of these processes being slow and reliant on human expertise. This paper examines the implementation of Large Language Models (LLMs) to enhance the efficiency and quality of matching Request for Quotations with suitable suppliers in a B2B supply chain procurement platform. By leveraging LLM vector embeddings and similarity metrics such as cosine similarity, we significantly improve the efficiency of the entity matching processes taking place in these procurement platforms. Through comprehensive experimentation, with the use of various models and approaches, we demonstrate the effectiveness of LLMs within the domain. We develop a system capable of matching RFQs to appropriate suppliers within seconds from a pool of hundreds of companies. With the top-performing model, on average ranking the correct supplier as the second most probable and achieving a success rate of 47.9% in identifying the supplier selected by a procurement agent, we clearly demonstrate the real-world suitability of LLMs to this task. Furthermore, we provide a comparison of the use of open versus closed source LLMs and assess the advantages and disadvantages that such models present. Overall, this paper provides an in-depth analysis of the successful integration of LLMs into the procurement sector, providing insights into improving the efficiency and quality of B2B supply chain procurement platforms.

I certify that all material in this dissertation which is not my own work has been identified.

Contents

1 Introduction

In recent years there has been a fresh wave of interest placed on AI research, with particular interest in the field of Large Language Models (LLMs). The release of the Generative Pre-Trained 3 (GPT-3) model by OpenAI (Brown et al. 2020), and their consumer-oriented products, ChatGPT (OpenAI 2023*b*) and DALL-E (Ramesh et al. 2021), marked a pivotal turning point in the popularisation of such models. The popularity of ChatGPT made it the fastest growing consumer application in history, with it getting over one hundred million users in the 2 months following its release (Hu 2023).

Breakthroughs in the area of LLMs were made possible through recent advances in Neural Networks (NN), a significant increase in computing power and an unprecedented availability of data. The combination of these three factors has ushered in a new age of Data Science and Machine Learning (ML). Private companies have been quick to capitalise on the versatility of LLMs by adapting them to sector specific tasks. Microsoft invested over ten billion dollars in OpenAI in 2023 giving it exclusive rights to the newer GPT model configurations (Times 2023), while it also plays an important role as the cloud supplier for OpenAI's products. It has been quick to integrate the models into its own products, with Microsoft Copilot, a GPT powered assistant, being integrated directly into the Windows 11 taskbar and the Bing search engine. Its subsidiary GitHub has also made its own chatbot assistant designed to help speed up coding tasks, showing the capabilities of such models to be specialised to industries.

Research is now being undertaken in many different areas as researchers seek to discover the extent to which these models can be applied to different use cases. This is enabled by the wide range of functionalities LLMs offer, from calculating vector word embeddings (a numerical representation of a word's semantic meaning) to enabling real-time translation. This project seeks to make use of the entity matching capabilities of LLMs, most notably through the use of the models' vector embedding generation functionalities. It seeks to discover the extent to which the matching of buyers' Request for Quotations (RFQs) with appropriate suppliers can be automated in a Business-to-Business (B2B) supply chain procurement platform using LLMs.

This is an important project to undertake as global trade relies on the speed and reliability of such procurement platforms. Crises such as the Covid-19 pandemic highlighted the difficulties which can arise in product procurement when wide-spread shortages develop, or trade restrictions are put in place (Attinasi et al. 2022). Efficiently and effectively matching RFQs to suppliers may therefore help alleviate potential difficulties faced by businesses in such scenarios and in day-to-day activities with regards to procurement.

1.1 Overview of Large Language Models

The models on which this project relies, LLMs, are a type of "deep learning algorithm that's equipped to summarise, translate, predict, and generate text to convey ideas and concepts" (Lake 2023). These models are based on multi-layered NNs composed, most often, of billions of parameters. They are able to mimic human-like writing and query answering with high quality as a result of their high complexity and the large amount of data they have been trained on.

Their textual vectorisation capabilities rely on their capacity to place similar words, and subsequently sentences, close to each other in high dimensional vector space. For instance, the word "man" would be placed closer to the word king than the word queen. In order to place these words into vector space, they are first of all converted to tokens which can be thought of as a numerical representation of an element of a word, or the word as a whole, which is deemed informative. Once tokenised these word tokens can be passed to an algorithm which uses them in order to create a vector embedding.

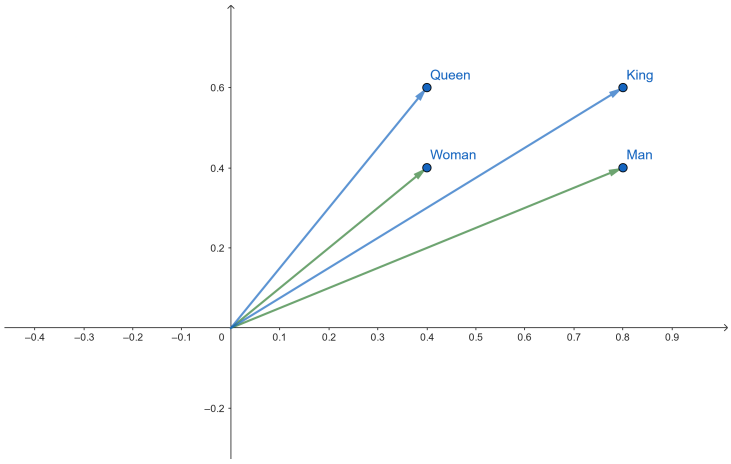


Figure 1: Vector Representation of Words in 2D

These capabilities have translated to impressive human-like performance on certain tasks such as linguistic exams (OpenAI 2023*a*). The integration of LLMs into procurement platforms is therefore an interesting area of research due to the high level of textual data involved in the sector's matching processes. LLMs may hold useful capabilities which can help to improve the efficiency and quality of the entity matchings taking place in procurement platforms.

1.2 Introduction to Supply Chain Procurement Platforms

A B2B supply chain procurement platform is a system which seeks to connect professional buyers and suppliers. It allows them to communicate their respective demands and capabilities. These platforms are based on traditional auctions with key modifications made for the nature of the online process. Buyers typically set their demands out in a RFQ; this is a document which specifies their requirements and so enables bidding from suppliers. The documents can vary in detail but will often specify requirements such as product type, quantity and delivery locations.

It is important that all details in an RFQ are considered before a company decides to bid, as all the request details have to be thoroughly analysed and checked against the ability to deliver the requirements. This need to analyse many RFQs by a company, to see which ones are good matches, can take a lot of time and therefore delay the process as a whole. This inevitably causes dissatisfaction among buyers who have to wait longer for their requests to be fulfilled, as well as among sellers who lose out on potential added revenue. Therefore, there is room for efficiency improvements in these processes.

Procurement platforms have gained in popularity over the years as they have slowly become instrumental to B2B trade. Certain platforms have even gained global recognition. For example, the Chinese tech giant Alibaba, which is at its core a B2B marketplace, reached a market capitalisation of over $850 billion dollars in October of 2020. The success of Alibaba lies, in large part, to its role as the point of contact between western buyers and Chinese suppliers. Its platforms allow these western buyers to submit RFQs, negotiate prices and fulfil transactions on the platform. Such platforms have provided a vital tool in enabling global trade in the 21st century. They have improved the matching processes by providing a centralised platform where they can take place. They have also allowed for more trust to develop between buyers and sellers, who can now leave reviews on the quality of a trade. This has improved the satisfaction among both buyers and sellers. However, there is still room for improvement in the processes, most notably with regards to efficiency.

1.3 Challenges in Current Procurement Matching Processes

Despite their success, the current procurement matching processes in B2B procurement platforms remain very human centred. In most cases, once a new RFQ is submitted on the platform a human agent has to manually match it to potential suppliers. Another option is to have RFQs sent off to all suppliers and have the suppliers manually read through them and decide whether to bid on it or not. This type of system adds significant overhead, especially in times of high demand or of crisis. This therefore greatly slows down the trade processes as a whole and leaves room for improvement.

Furthermore, human matching may lead to user error, or bias being present in the selection process. For instance, crucial details of the RFQs may be overlooked, or a certain company may be unfairly prioritised over others. Language may also be a barrier when dealing with cross-border trade. There is therefore room for improvements to be made in such platforms in order to reduce the time it takes for matchings to occur, and to limit the potential of human errors or biases.

1.4 Potential Advantages of Using Large Language Models in Supply Chain Optimisation

Therefore, there is potential for improvement in the matching processes of procurement platforms through the use of LLMs. (Andersson et al. 2023) highlighted that the most requested aspect of procurement which procurement managers requested be improved was the RFQ process. As the majority of RFQs are generated by humans and LLMs have be shown to perform well on semantic understanding, there is potential for LLMs to be applied to improve the quality of the matchings. By improving both the speed and quality of these matchings, LLMs could increase the competitiveness of procurement matching platforms and improve both the customer and supplier satisfaction.

This dissertation project implements an innovative methodology to assist in this task. The final aim will be to have an LLM which, when given an RFQ, is able to accurately suggest an appropriate supplier which would be capable of fulfilling the demands of the RFQ. This implementation will rely on data provided by the Devon based procurement matching platform Applegate Marketplace Ltd, which has over 140 000 suppliers (Applegate n.d.). The data provided includes past customer RFQs, supplier descriptions and the responses given by suppliers to RFQs. This project will seek to explore in detail the capabilities of different LLMs with regards to the B2B procurement sector, giving an in-depth analysis of possible implementations.

1.5 Outline of Dissertation

This first part of this paper analyses the past work which has been conducted in relation to the topic in the form of a literature review. It then describes the project's aims, setting out in more detail the potential benefits of the work, as well as the baselines and performance measures which will be used. The methodology of the work undertaken is then set out, with explanations of the training process. Finally, the results are presented and analysis of these results is conducted with regards to different criteria. The paper ends with further discussion of the impact of the results and a conclusion. By the end, the reader will have been presented with a clear overview of some of the opportunities LLMs provide with regards to entity matching in B2B procurement platforms.

2 Literature Review

2.1 Large Language Model Evolution and Key Models

The emergence of the field of Natural Language Processing (NLP) coincided with the development of the Turing Test back in the early 1950s (Turing 1950), which set out a framework for assessing artificial intelligence. Over the last decade NLP has grown in importance and popularity. One of the key sectors of NLP is textual representation, which includes most notably word and sentence vectorisation. The first of such methodologies to gain popularity, named Word2Vec, was developed by researchers at Google in 2013 (Mikolov et al. 2013). This system used a Continuous Bag-of-Words technique and a Skip-gram model along with a NN. It trained the models on Google News data to compute semantically meaningful word vector representations. It would do so by learning which words would appear most often in proximity to each other and subsequently place these words close to each other in vector space. The system, however, presented several biases, which were often reflective of the data it was trained on. An example of this was found when giving the model the input "doctor minus man plus woman" which would yield the word "nurse" instead of the more appropriate response of "doctor" (Lee & Trott 2023). This offered an early warning to the potential dangers of training such language model systems on biased data.

A year later, researchers at Stanford University published their paper (Pennington et al. 2014) which introduced the **Glo**bal **Ve**ctor for Word Representation (**GloVe**) model. This algorithm consisted of a global log-bilinear regression model which improved on Word2Vec by using a "weighted least squares model that trains on global word-word co-occurrence" (Pennington et al. 2014). It was trained on a large variety of data, including Wikipedia and Common Crawl data. It achieved greater results than older models on "word analogy, word similarity, and named entity recognition tasks" (Pennington et al. 2014). To this day, GloVe is considered one of the staples of NLP and until the emergence of LLMs was one of the best methodologies for creating meaningful vector embeddings.

A key turning point in the field of NLP came in 2017, when researchers at Google released a paper presenting the Transformer architecture (Vaswani et al. 2017). This was followed by the release of the **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**) LLM in 2018 by (Devlin et al. 2018). This model achieved state-of-the-art performance on many NLP tasks. The Transformer architecture, in particular its attention mechanism, allowed the model to capture long-range dependencies in the input text. This mechanism also allows it to capture deeper contextual information about the input. The embedding associated with a word could therefore consider the context of previous words in order to more accurately represent the word's embedding in vector space. BERT provided the benefit of allowing vector embeddings to be extended to represent entire sentences or a whole corpus of text. This was a significant advancement in the field of NLP and provided the basis for the emergence of future LLMs and subsequently the field of Generative AI. Since its release, the model has been used in various applications, including in the Google Search algorithm (Singh 2021) to improve search results. Overall, the introduction of the Transformer architecture and of BERT mark one of, if not the, most important shift in the history of the field of NLP.

Building upon the foundations of (Vaswani et al. 2017), OpenAI released their first LLM in the 2018 paper "Improving Language Understanding by **G**enerative **P**re-**T**raining" (Radford et al. 2018). While BERT was designed as a bidirectional model, meaning it captured semantic information given the entire sentence and analysing the sentence starting from both the beginning and end of the sentence, the GPT model was designed as unidirectional. This means it is only provided the context of the sentence up until the preceding word. This in particular allowed it to be used on tasks such as text generation, while BERT was more focused on tasks such as text classification. GPT's text generation relies on next word generation given all the preceding words. This made GPT an interesting model as it could be integrated into tools such as chatbots.

Researchers at Facebook released the RoBERTa model in 2019 (Liu et al. 2019), which improved upon the BERT model in many key areas such as sentiment analysis and named entity recognition. Most notably it increased the size of the training data to 160GB, more than ten times that of the original model. It also made use of dynamic masking and larger mini-batch sizes during training. It achieved state-of-the-art performance and replaced BERT as the new baseline for many tasks. Additionally, RoBERTa provided insight into the fact that the performance of LLMs could be drastically improved by training them on a lot more data and through hyperparameter tuning.

These early models, however, had some difficulties when used for semantic matching through vectorisation. The 2019 paper by (Reimers & Gurevych 2019) highlighted these difficulties and sought to discover the reasons behind them. One of the main difficulties highlighted was the slowness of the process, with it taking around 65 hours to find the "most similar pair in a collection of 10,000 sentences" using the BERT model. They sought to improve the performance of LLMs for semantic matching by designing **S**entence-**BERT** (SBERT), a model designed solely for the purpose of generating vector embeddings. They hypothesised that the reason that models such as BERT and RoBERTa were not suited to the task of semantic matching was that the embeddings produced by the model were inconsistent. These vectors were usually obtained by averaging the weights of the output layer or by cutting off a vector to a certain length regardless of input length. This led to inconsistent vectors, which were not necessarily placed correctly in vector space according to their semantic meanings. SBERT changed the approach by having the output of the model be a fixed size vector embedding which strived to placed semantically similar sentences close together. The model showed promising results with it greatly reducing the time and resources needed to compute embeddings while keeping the same level of performance as the more demanding models.

By far the most popular LLMs to have been introduced over the last few years has been the newer iterations of the GPT models released by OpenAI. These include the GPT-3 (Brown et al. 2020) and GPT-4 (OpenAI 2023*a*) models. Having been trained on an estimated couple of hundred to thousand gigabytes of data, these models have achieved state-of-the-art performance on many NLP tasks. Their popularisation, including in the mainstream through the release of ChatGPT, has been one of the most important shift in the field of LLMs. They have rapidly gained in popularity since 2022 and have been integrated into hundreds of sector specific applications. Today their capabilities include text generation and summarisation, as well as the ability to provide users with high-quality embeddings. Through their subscription-based API, businesses and researchers can use these embeddings for tasks such entity matching and semantic analysis. Use cases for the models' other capabilities include integration into personalised chatbots and NLP intensive tasks. One key difference between the latest GPT models and models such as BERT and RoBERTa is the closed-source nature of the models. They therefore require a paid subscription to their API in order to make use of their models' capabilities. Despite this, they currently hold the place as the most popular models in the field.

The legacy of SBERT is now greater than the model which was produced by the researchers in (Reimers & Gurevych 2019). The development of SBERT led to the release of the Sentence-Transformers library. The Sentence-Transformers library, along with Hugging Face's Transformers library (Wolf et al. 2020), today represent the two major libraries in Python for the use of open-source LLMs. These libraries can be used to obtain easy to implement code and models to produce vector embeddings and fine-tune models to specific tasks.

The open-source space has been rapidly expanding over the last few years with an increasing number of well performing models emerging. The MTEB leaderboard (Muennighoff et al. 2022) provides a comparison of these models on various benchmarks. Currently OpenAI's top embedding model text-embedding-3-large (OpenAI 2024) ranks 10th in this leaderboard with it being outperformed by various open-source models. Over the coming years these types of performance benchmarks will become increasingly important as companies seek to choose between using open-source models run locally and using paid for models provided through APIs.

Overall, the field of NLP and subsequently LLMs has dramatically changed over the last decade. LLMs are currently being used in a number of sectors and are to be deployed in many more. One of these sectors being that of procurement, in particular B2B procurement platforms.

2.2 Past uses of Machine Learning and Large Language Models in Buyer/Supplier Matching

Various implementations have previously looked into using ML and LLMs in the matching process of supply chain procurement. (Trappey et al. 2022) focused their research on summarisation for RFQs focused on power transformer purchases. They used a statistical N-gram TF-IDF model to build a word-embedding model which could extract key information such as the voltage, capacity and impedance from the RFQs. They highlighted several benefits to their implementation including a significant decrease in the time needed to analyse lengthy RFQs, which can often be composed of dozens of pages for products such as power transformers. They also highlighted the ability of such a system to minimise the risk of humans missing out important information vital to the RFQ. This paper therefore highlights the semantic extraction capabilities present in LLMs which could be a vital tool for a B2B marketplace.

In (Nia et al. 2019) researchers set out a framework for developing a recommendation system to match RFQs to items in a large scale B2B marketplace. They set out the potential for such a system by highlighting the past implementations of similar recommendation systems in customer-oriented marketplaces. They attribute the lack of similar implementations in the B2B space due to the higher complexity in its processes. They highlight the dangers of possible "information overload" in a B2B platform if the choice of items and suppliers is too large and subsequently difficult to choose from. As matchings are done by human operators, they explain how a lot of strain is potentially placed on these individuals if the system becomes overloaded. They highlight the difficulty such individuals can encounter when providing matching considering the widespread use of abbreviations and naming variations in RFQs. These difficulties are accentuated if individuals are not experts in the sectors where they are asked to match products in. They therefore highlight the potential benefits of a recommendation system which could match requests to the correct items. This type of system would thus reduce the need for heavily human dependent intervention. They set out that the similarity between items could be calculated using a combination of Word2Vec (Mikolov et al. 2013), NNs and cosine similarity measurement. Unfortunately, the work following the framework set out has not been published to this date, so no further information is made available. This framework does however provide clear insight into some of the benefits such as system could have if implemented in a B2B marketplace.

The paper by (Dolle et al. 2020) highlights the benefits that an automated sourcing system could have for small and mid-sized enterprises (SMEs). They highlight the difficulty that these SMEs often face when trying to source product as suppliers often prioritise making partnerships with larger suppliers. They also state that non-automated procurement platforms may also present this bias, as they are trying to please their suppliers by giving them partnerships with larger customers. They therefore propose that automation could reduce the bias present in the process by providing unbiased matching between buyers and sellers regardless of status. This paper therefore highlights the potential benefits LLMs could have in a B2B marketplace in reducing potential biases.

Work by (Beason et al. 2021) looked into using open-source models like BERT in order to perform named-entity recognition (NER) on RFQs. They highlighted the difficulty associated with this task due to a lack of consistency between RFQs which can vary in length and detail. By fine-tuning both SpaCy and BERT models they were able to obtain decent results in NER. Their paper therefore highlights some of the further interesting capabilities that LLMs may hold for semantic analysis in B2B marketplaces.

The papers above highlight some of the key past implementations which have been made in the sector. We can see that multiple different techniques have been explored. There is therefore potential that LLMs be used not only for entity matching but also for tasks such as summarisation or feature extraction. While these papers highlight that work has been done in the sector, work remains fairly limited as seen subsequently.

2.3 Limitations of Past Implementations

While advances have been made to incorporate ML techniques into procurement processes, the procurement sector's level of integration remains far behind other similar sectors (Nia et al. 2019) (Andersson et al. 2023). Implementations have not gone as far as fully automating the matching of buyers and sellers. Instead, research has mainly been focused on summarisation and feature extraction. Reasons for the lack of development have included the lack of reliable methods, the high complexity of the task and the ease of using human operators. Given the importance of procurement processes, with them being an underlying necessity for global trade, it is vital that they catch up to the same level of technological advances as other sectors. LLMs could therefore provide the key to making the necessary advances. Given the broad scope of capabilities they offer, they could not only be used to automate the matching of buyers and sellers but could also be used for things such as NER. While LLMs have shown promising results, they are still in their infancy and have therefore not been researched in great detail yet. Their implementation in the context of a procurement platform could therefore provide a market advantage to the platform which improves its service first. It could potentially improve the service for both buyers and sellers by providing quality matching within seconds, as well as reducing the costs needed to conduct the matchings.

While many of the past projects have had to conduct research using publicly available data, this project has the advantage that it uses data provided by Applegate, containing real-world RFQs and matchings. Success in this project would not only provide a benefit to Applegate but would also show the procurement industry as a whole that automation can be easily and effectively implemented. It would also demonstrate, in a real-world application, the role that LLMs can play in automation, not only in the procurement sector but in platforms as a whole. Using private data may also hold advantages for performance as opposed to training on openly available data, due to the possible higher quality found in the private sector's data (Ahmed et al. 2024). As LLMs are a fast-growing field this implementation could also only be the start of a long possible road of implementations as models are improved over time.

2.4 Ethical Considerations of Large Language Models

While LLMs have been praised as great innovations which have the potential to revolutionise multiple sectors, there also some ethical concerns which have been raised in relation to their usage. These concerns are multi-faceted and of different significance levels (Weidinger et al. 2021). One of the major concerns is that the models could be used in a way which causes harm. In the context of procurement platforms this could present itself in several ways. One could imagine that sensitive data relating to customers or suppliers, which the model was trained on, could be leaked either unintentionally or through model vulnerability exploitation. This would be a severe breach of confidentiality and could cause significant harm to the platform's reputation. In order to mitigate these risks several safeguards are put in place. In the case of using LLMs provided by OpenAI, the sensitive data is transferred to OpenAI through the OpenAI API. OpenAI has certified that such data remains confidential and secure. In the case of open-source models provided by Hugging Face, all data and model training procedures are done locally or through Google Colaboratory and are therefore secure. Access to the data and models will be granted only to the authorised parties as set out by the NDA agreement signed with Applegate.

In addition to fears of potential harm, an implementation of LLMs for automating the matching of buyers and suppliers in a procurement platform could also entail risks to the business sector as a whole. While an LLM has the potential to greatly enhance the matchings, it could also be used to the detriment of buyers in order to implement methods such as dynamic pricing based on the RFQs being inputted. This could have large consequences on buyers, particularly smaller buyers, who could have difficulty adjusting to rapid price fluctuations. The matching processes could also be used to aid unethical, or even unauthorised, matchings to occur if a human-being does not verify the matchings being processed. This could happen for instance if a customer is matched with a rival business, or a supplier is matched with a sanctioned customer. It is therefore vital that human beings remain a part of, even if in a limited capacity, the matching process.

Finally, LLMs also have ethical concerns relating to the resources needed to train and use the models. While countries and businesses around the world are trying to reduce their carbon emissions, LLMs prove to be very energy consuming tasks. Their increased popularity could therefore harbour negative consequences for the environment (Weidinger et al. 2021). The necessity to train these models for several hours using high-powered GPUs is also cause for concern. However, in the case of this project we can see that while power usage can be high during training, the usage thereafter remains minimal. Given the fact that incorporating these models into procurement could greatly enhance matchings so that they are higher quality, and in future iterations could even take into account factors such as emission restrictions, we argue that the benefits of incorporating the automation largely outweighs the potential downsides.

3 Project Aims and Methodology

3.1 Research Outline

This project sets out to improve the RFQ matching processes of procurement platforms through the use of LLMs. By incorporating LLMs into these processes, the project seeks to enhance the efficiency of supply chain procurement platforms and improve the quality of matchings being made.

Several areas of research will be explored, including:

- An investigation into whether LLMs can accurately capture semantic information in RFQs and supplier descriptions for improved matchings.
- An assessment of the performance of non-LLM models compared to LLMs.
- A comparison of the performance of open-source and closed-source LLM models.
- An exploration of the potential advantages of combining multiple models.

Through these research areas the project will provide a well-rounded overview of the potential integration of LLMs into the matching processes of B2B procurement platforms.

The project will investigate the benefits that the Transformer architecture, in particular its attention mechanism, may hold when applied in the context of entity matching through the use of vector embeddings. We hypothesise that the attention mechanism should be well suited to the task, as it enables the representation to take into account the most important parts of a text based on the surrounding context. This should be particularly useful in the case of both RFQs and supplier descriptions, where details present in a small subset of words, or context dependent words, may be the difference between a good and poor match.

Through the use of LLMs, we seek to find highly effective models which can be setup at a limited cost. We aim to show how these models can then be used for cheap and efficient matchings to take

place in an easy to implement environment. We then present different models along with various potential benefits and downsides associated with each. Through this we hope to provide a good overview of the different options available for implementation.

3.2 Baselines

Several baselines are established in order to provide a comparison of the capabilities of different models in the matching of RFQs with supplier descriptions. These include Word2Vec (Mikolov et al. 2013), spaCy's en-core-web-md (Honnibal & Montani 2017), GloVe (Pennington et al. 2014), TF-IDF (Sparck Jones 1972) and BERT (Devlin et al. 2018). While all of these are popular methods, only the BERT model is considered a LLM. However, all of these models enable us to extract vector embeddings to represent a piece of text, in our case the inputted RFQs and supplier descriptions. Given the increased cost of using LLMs compared to other methodologies, it is of interest to determine whether their potential increased performance outweighs the downside of their greater cost. LLMs also have the downside of having token limits, which can limit the amount of information fed into the network. We show, however, that the baseline models, while popular, are poorly adapted to the task of entity matching in procurement platforms, and therefore that newer generation LLMs are better suited.

Model	**Token Limit**	**Embedding Length**	**Time**	**Availability**
Word2Vec	-	300	Low	Open-Source
en-core-web-md	-	300	Low	Open-Source
GloVe	-	300	Low	Open-Source
TF-IDF	-	1505	Low	Open-Source
BERT	512	768	High	Open-Source

Table 1: Baseline Model Specification Comparison

3.3 Performance Measures

The project's model performances will be measured by two different measures. Firstly, we use the cosine similarity score given by: Cosine Similarity (A, B) $= \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}}$ where A and B are vectors. This measure will be used to calculate the similarity between our different vector embeddings (Rahutomo et al. 2012). We can then use the similarity scores in order to set a threshold for differentiating good and poor matches (Rekabsaz et al. 2017)(Thongtan & Phienthrakul 2019). This can notably allow us to find the most similar pair of embeddings for a particular RFQ when comparing to all the supplier descriptions. In order to ensure consistency, all metrics will be calculated on the same testing data. This will allow us to set in place our second metric, which is used in two different ways.

Firstly, given an input RFQ, we focus on calculating the similarity scores between the input and all the supplier descriptions present in the training data. This will allow us to turn our classification task into a ranking task by sorting the matches based on the similarity scores, ordering the pairs from most to least similar. We will then evaluate the performance of models by seeing where the matches labelled as good come up in the list of rankings. The lower the ranking the better the model has performed at matching similar entities together. We repeat this process for each RFQ, comparing to all possible suppliers and ranking. We then look at the mean, median, mean reciprocal rank as well as the percentage of good matches ranked as the top, in the top 5 and in the top 10 most similar entities.

The second method will focus on comparing the inputted RFQ to previous RFQs in the training data and calculating the similarity scores between them. In this method we rely on the RFQs which the model has previously seen, and has learned from, in order to provide matches. We can then extract the suppliers which had been matched to these previous RFQs in order to rank which suppliers may be good matches for a newly inputted RFQ. Similar to the method above we then see how often good matches are ranked highly by the model in order to quantify model performance. This method may hold better performance due to the fact RFQs are being matched to past RFQs, which may be formatted more similarly than when comparing to supplier descriptions. It will however require more resources due to the increased number of calculations needed for the process.

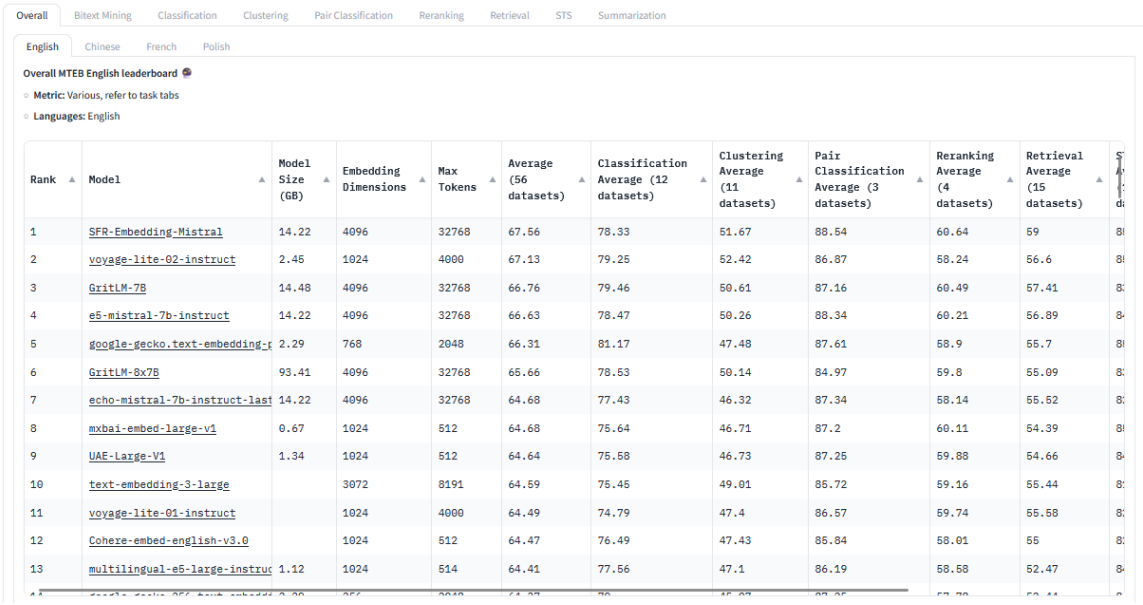
By comparing inputted RFQs to both the supplier descriptions and past RFQs we provide two different methods of ranking. This will allow for insights to be drawn of the different capabilities of LLMs and what aspects of data are best used to perform matchings. One final method which will be briefly explored will be to combine the predictions of multiple separate models into one. We do this in order to see if such a method could provide greater precision capabilities than using individual models. Furthermore, we will also quantify the time needed to run the different methods as this is a necessary metric in order to establish the viability of a system when implemented in real-world scenarios.

3.4 Evaluation of Open-Source vs. Paid Large Language Models

As previously mentioned, this study will not only focus on providing a good model to perform matchings but will also seek to evaluate the benefits and downsides of using open vs. closed source LLMs. This will mainly involve comparing the performance of OpenAI's models, which are currently the most popular LLM models being utilised and are closed source, against the performance of several open-source models available on the Hugging Face platform.

The MTEB leaderboard, as shown in Figure 2, compares the performance of embedding models on several baseline tasks. We can see that OpenAI's text-embedding-3-large model (OpenAI 2024), currently the company's best embedding model, ranks 10th in the leaderboard. Its performance on the MTEB baselines is therefore being exceeded by several open-source models. However, while OpenAI's model is usable through an API, allowing the user to use it without needing significant resources, open-source models require the user to use their own resources to run models. For instance, the size of the top model, SFR-Embedding Mistral (Rui Meng 2024), is over 14 gigabytes. The model therefore requires significant resources to run and would require several state-of-the-art GPU's in order to fine-tune. API based LLMs such as OpenAI's, while costing more in terms of direct cost, may therefore be simpler to use for a user with less resources, while offering competitive performance.

This project will focus on smaller open-source models for its research due to its limited resources. All open-source models will be accessed through the Sentence-Transformers library (Reimers & Gurevych 2019). These include the GIST Embedding v0 and GIST Large Embedding v0 models by (Solatorio 2024), as well as the all-MiniLM-L6-v2 (Reimers & Gurevych n.d.*b*) and all-MiniLM-L12-v2 (Reimers & Gurevych n.d.*a*) models by the SBERT team (Reimers & Gurevych 2019). These models vary in size from 0.09 gigabytes for the all-MiniLM-L6-v2 model, all the way to 1.34 gigabytes for the GIST Large Embedding v0 model. They also vary in their MTEB ranking with the GIST models currently ranking 15th and 18th respectively, while the all-MiniLM models currently rank 77th and 78th. Through these models we seek to explore whether the rankings provided by the MTEB leaderboard show similar patterns in terms of performance when applied to procurement matching. Due to the dynamic nature of the LLM sector, these rankings are currently evolving as newer models are released, therefore conclusions drawn may evolve over time.



Rank	Model	Model Size (GB)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)
1	SFR-Embedding-Mistral	14.22	4096	32768	67.56	78.33	51.67	88.54	60.64	59
2	voyage-lite-02-instruct	2.45	1024	4000	67.13	79.25	52.42	86.87	58.24	56.6
3	GritLM-7B	14.48	4096	32768	66.76	79.46	50.61	87.16	60.49	57.41
4	e5-mistral-7b-instruct	14.22	4096	32768	66.63	78.47	50.26	88.34	60.21	56.89
5	google-gecko.text-embedding-p	2.29	768	2048	66.31	81.17	47.48	87.61	58.9	55.7
6	GritLM-8x7B	93.41	4096	32768	65.66	78.53	50.14	84.97	59.8	55.09
7	echo-mistral-7b-instruct-last	14.22	4096	32768	64.68	77.43	46.32	87.34	58.14	55.52
8	mxbai-embed-large-v1	0.67	1024	512	64.68	75.64	46.71	87.2	60.11	54.39
9	UAE-Large-V1	1.34	1024	512	64.64	75.58	46.73	87.25	59.88	54.66
10	text-embedding-3-large		3072	8191	64.59	75.45	49.01	85.72	59.16	55.44
11	voyage-lite-01-instruct		1024	4000	64.49	74.79	47.4	86.57	59.74	55.58
12	Cohere-embed-english-v3.0		1024	512	64.47	76.49	47.43	85.84	58.01	55
13	multilingual-e5-large-instruc	1.12	1024	514	64.41	77.56	47.1	86.19	58.58	52.47

Figure 2: MTEB Leaderboard Extract

3.5 Project Timeline

Date	Event
22nd November 2023	Literary review and project plan submission
15th December 2023	Further data analysis and baseline set up
15th January 2024	First model iterations using all-MiniLM-L6-v2
5th February 2024	Second model iterations including fine-tuning of all-MiniLM-L6-v2 + first OpenAI model
19th February 2024	Third model iterations including fine-tuning of GIST models
4th March 2024	First draft write-up of model results and final model iterations
1st April 2024	Second draft write-up including model results, literary review, code, bibliography
1st May 2024	Submission of paper and video

Table 2: Timeline of Events

4 Implementation

4.1 Model Presentation and Selection

In this section we go through the implementation that was carried out. Table 3 presents some of the key variations present in our different LLMs. These differences justify the need to study multiple LLMs in our implementation in order to draw a full understanding of the capabilities such models hold. We can see that, while longer embeddings may be beneficial in allowing for a larger vector space to be represented, they can also lead to considerable increases in the time needed to obtain such embeddings. This can be attributed, in large part, to the higher complexity of the models which produce these longer embeddings.

We can also see that the open-source models tend to have far smaller token limits and embedding lengths than their closed-source counterparts. This limitation may diminish the relationships that the models are able to represent from the texts. While the large majority of our data remains within the models' token limits, a few edge cases may therefore be negatively impacted by this aspect of the open-source LLMs selected for this project.

We observe that, while there is nothing the user can do the speed up the obtention of embeddings from the closed-source OpenAI models, due to their acquisition relying on an API, the obtention of open-source embeddings can be dramatically increased through the use of GPUs. For example, the process of obtaining 4000 embeddings can go from taking several hours to around 20 minutes through the use of high-performance GPUs. The use of GPUs is therefore highly recommended for running open-source LLMs, especially larger models such as GIST Large Embedding v0. The increased flexibility of running open-source models therefore provides an advantage in terms of time over API based models.

Model	**Token Limit**	**Embedding Length**	**Time**	**Availability**
all-MiniLM-L6-v2	512	384	Low	Open-Source
all-MiniLM-L12-v2	512	384	Low	Open-Source
GIST Embedding v0	512	768	Medium	Open-Source
GIST Large Embedding v0	512	1024	High	Open-Source
text-embedding-3-small	8191	1536	High	Subscription-Based
text-embedding-3-large	8191	3072	High	Subscription-Based

Table 3: LLM Specification Comparison

4.2 Data and Resources Utilised

The data provided by Applegate is used to train and evaluate all the models. We use data containing past RFQs, supplier descriptions and past RFQ responses that suppliers have given. We limit our training data to cases where the texts have over 100 characters in order to ensure that the representations learned are informative. We also limit the data to cases with responses by paying suppliers only. We do this in order to reduce potential data noise and mislabelling problems, such as those discussed in section 5.5. The preprocessing of the texts is limited to the tokenisation provided by the individual models. This choice is made in order to allow the LLMs to have access to the full context window of the inputs and for the implementation to be as close as possible to a real-world scenario.

The positive class data points are taken to be the RFQs to which suppliers responded they wanted to bid. For our negative classes we found that we were unable to simply take cases where suppliers turned down the opportunity to bid on a RFQ. This is due to the fact that a negative response does not necessarily correlate with a matching being poor but could instead be due to other factors such as limits on capacity that a supplier could be facing. We therefore use categorisation nodes which Applegate had previously labelled RFQs and supplier descriptions with. Using these nodes we randomly shuffle RFQs to suppliers with separate categorisation nodes in common. This results in RFQs being matched to suppliers from inappropriate sectors reflecting potential bad matches. While this methodology is not flawless, especially as it relies on categorisations which may not fully encapsulate sectors, we can see that it provides a solid foundation for generating negative classes.

When performing our ranking tasks, we also make the decision to drop cases where multiple suppliers have requested to bid on the same RFQ. As RFQs are often matched to multiple suppliers, our models may match an RFQ to a supplier whose matching is not in the testing data, leading to out of scope matchings. This choice therefore means that our models may perform better than is

shown in the results. As some of our models perform very well despite this choice, the project will not investigate the impact that it may have on performance any further.

Request for Quotation	**Good Supplier Match**	**Bad Supplier Match**
"We are seeking proposals for the supply of high-quality face masks suitable for medical use, meeting all necessary safety standards and regulations."	"Protective Gear Inc. specialises in manufacturing and supplying top-quality face masks designed for medical use, ensuring compliance with all safety standards and regulations."	"Direct Cargo Services UK Ltd is a freight delivery company specialising in efficient transportation solutions for businesses."

Table 4: Example of Potential Good and Bad Matching

4.3 Design Choices

With our training data created, we proceed by splitting the data into a 80/20 training and testing dataset split. This leaves us with 3720 training matches and 932 testing matches, along with balanced positive and negative classes. We then obtain the embeddings for the RFQs and supplier descriptions in both these datasets using our baseline models and the base LLMs. Using these embeddings, we can then analyse the distribution of the positive and negative classes with respect to the similarity scores between the RFQs and their matched suppliers' descriptions. This is done in order to evaluate the potential for class separation based off the thresholding of the similarity score.

In Figure 3 we can observe the boxplot of the similarity scores for the training data's classes using the GIST Large Embedding v0 model. We can clearly see that the positive class data points in the training data tend to have higher similarity scores than the negative class. This gives initial indication that LLMs can be used for matchings and provides indication that using thresholding on similarity scores may provide a good way to separate positive and negative classes, a tool necessary for fine-tuning the models.

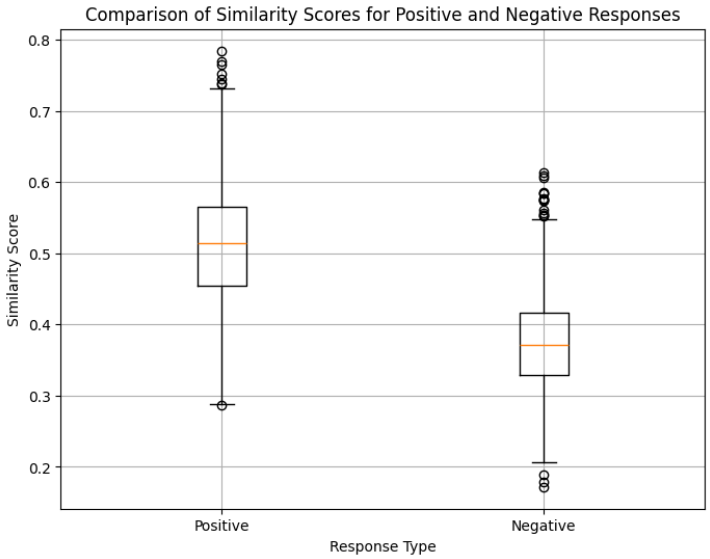


Figure 3: GIST Large Embedding v0 Training Data Class Separation

In order to store the embeddings associated with the RFQs and supplier descriptions we make use of SQLite databases. These databases provide efficient and easy access to the stored embeddings. Databases may be preferred over storing in simple CSV files due to potential storage errors when using

CSVs, especially with longer embeddings such as OpenAI's which may be cut off. In a real-world implementation it would be necessary to implement such databases for efficient retrieval of embeddings in order to provide fast and accurate predictions while reducing the need for redundant embedding calculations.

4.4 Fine-tuning Process of Sentence Transformer Models

In order to accentuate the class separation seen in Figure 3 we then proceed to fine-tuning the LLMs. While OpenAI does allow users to fine-tune models through their API, this does not currently apply to their embedding models. As a result, we will focus on fine-tuning the open-source LLMs only. Using contrastive loss, we provide our fine tuner with both good and poor matches for the same RFQ and seek to accentuate the similarity between the good matches while decreasing the similarity for the poor matches (Hadsell et al. 2006).

The Sentence-Transformers library (Reimers & Gurevych 2019) provides all the necessary tools for easy fine-tuning. We provide the fine tuner with an initial threshold for class separation based off the insights obtained by running the base LLM models on the training data. We find that using 8 epochs provides adequate fine-tuning while limiting overfitting. Due to computational resource limitations, we are restricted to using small batch sizes, most often of 16 but going down to 8 for the GIST Large Embedding v0 model. Fine-tuning this model even with such a small batch size uses the full capacity of the top of the range 40 gigabyte NVIDIA A100 GPU.

Due to these computational limitations, we cannot run larger batch sizes for many of the models. This may therefore limit the models' learning capabilities. We also do not use cross-validation during the fine-tuning process due to the high resource demands involved. We can hypothesise that given more resources the models' performance may be further improved. However, as results seen in Section 5 are adequate we do not investigate this further.

5 Results and Analysis

In this section we present the results obtained. Table 5 sets out a summary of the evaluation performed using a variety of models including traditional models, LLMs and their fine-tuned counterparts. We also evaluate the potential use of ensemble models, which combine the embeddings of multiple LLMs in order to make predictions. The accuracy and F1 score are calculated using simple thresholding of similarity scores into positive and negative classes. While the other metrics are calculated with respect to the ranking methods described in Section 3 and the mean reciprocal rank. Through this evaluation we aim to give the reader a clear assessment of the performance of the various models on the task of entity matching in procurement platforms.

5.1 Assessment of Model Performances

In the top half of Table 5 we evaluate the performance using direct matching between the RFQs and the supplier descriptions. We see that LLMs can indeed be used to provide high quality embeddings which can be used for matchings. Using simple thresholding of class separation, we can observe that LLMs greatly outperform our baselines. In particular, the fine-tuned models, which have been adjusted specifically for this task, perform the best. The fine-tuned GIST Large model obtains an accuracy and F1 score of 94% using a threshold of 0.65. Amongst the non-fine-tuned LLMs the OpenAI Large model performs the best, with it getting a thresholding accuracy of 83% using a threshold of 0.28.

With regards to our ranking evaluation approach, we find that comparing RFQs to past RFQs and then taking the suppliers previously matched is more effective than comparing directly to the supplier descriptions. Using comparison of RFQs to past RFQs we obtain several models whose median ranking is 2. This means that the median classification performed by the models ranked the matching performed by Applegate as the second most likely matching to happen out of 346 possible matches.

When comparing directly to supplier descriptions, the top performing fine-tuned models have a median ranking of 4 out of 451 possible matchings. We can also see that around 80% of the correct matchings will be classified in top 10 most similar pairings when using the fine-tuned models. This therefore highlights the capabilities LLMs have of making high-quality matchings. It also highlights a potential requirement for these LLMs to be provided with similarly structured texts when extracting embeddings for similarity comparison in order to obtain the best results. Matching to RFQs may also overcome the limitations in the descriptions provided by the suppliers which may not fully encapsulate the products they provide.

We can also see that the difference in embedding length of the models, while having a significant effect for the base LLMs, has a less pronounced impact once models are fine-tuned. Additionally, the fine-tune models tuned specifically for the task of matching RFQs directly to supplier descriptions perform worse than the base LLMs on the task of RFQ to RFQ matching. This shows the limitations of fine-tuning which may make the models be task specific and have difficulty generalising. Caution must therefore be used when using fine-tuned models on tasks they were not designed for, whereas the base LLMs are more generalisable. The fine-tuned models do however have a considerably smaller spread in the quality of their predictions as shown by their mean ranking values, which are lower than all the other models. Fine-tuned models may therefore be useful in order to reduce outlier predictions being made.

5.2 Comparison of LLM to Baseline Performance

We can see in Table 5 that all the LLM models, except for BERT, outperform the traditional non-LLM models. This demonstrates that the attention mechanism of LLMs is useful in extracting meaningful vector representations for both RFQs and supplier descriptions. BERT meanwhile is outperformed in some respects by TF-IDF, particularly when comparing RFQs directly to supplier descriptions. The limitations of BERT, which have previously been highlighted, are clearly shown when compared to newer LLMs. The higher performance of the LLMs, over the traditional methods, is particularly noted when doing threshold differentiation between classes. While the traditional methods present little pattern of difference between the classes using thresholding, all the LLMs present some separation. This shows that the pattern is not due to randomness but is truly due to the representations correctly capturing deeper semantic information.

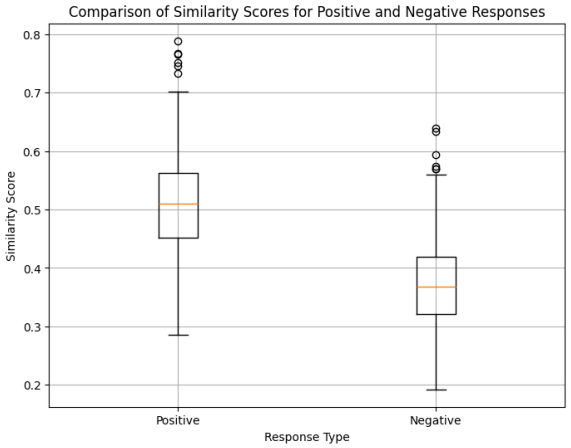
5.3 Impact of Fine-Tuning on Class Separation

We can see in Figure 4 the impact that the fine-tuning process has on class separation for the GISTv0 Large model. We can see that for the base GISTv0 Large model the class separation is present but with overlap between the interquartile ranges. Once fine-tuned we can see that this overlap is no longer present, with only a few outliers being present. This greater separation allows for thresholding of classes to be more viable as an approach and leads to better ranking results when comparing RFQs directly to supplier descriptions. Fine-tuning can therefore be useful if the method implemented relies on this class separation. However, as previously noted the models' generalisation capabilities may be diminished by such a process.

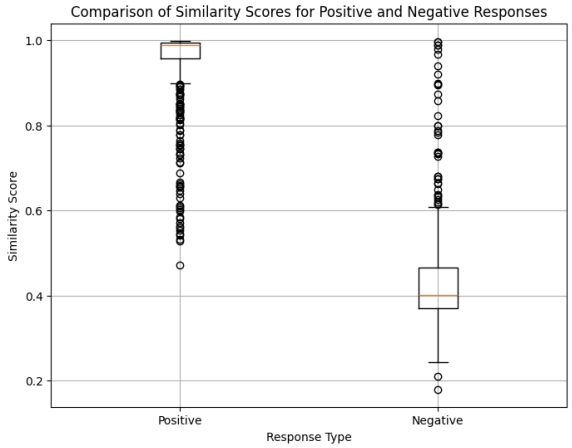
Table 5: Model Performance Metrics

Task	Model	Accuracy	F1 Score	Threshold	Mean Rank	Median Rank	MRR	% ranked 1	%ranked 1-5	%ranked 1-10
RFQ2COMP										
Traditional	Word2Vec	0.5	0.67	0.3	44	40	0.06	0.9	6	13.6
	en_core_web_md	0.5	0.62	0.5	38.5	39	0.09	3.1	12.2	19.1
	GloVe	0.5	0.6	0.75	35.2	32	0.1	4.4	12.6	21.3
	TF-IDF	0.64	0.62	0.01	27.4	21	0.19	10.2	24.6	33.9
LLMs	BERT	0.55	0.62	0.7	31.3	24	0.12	5.8	13.5	24.4
	all-MiniLM-L6-v2	0.74	0.72	0.25	16	10	0.25	13.5	36.1	51
	all-MiniLM-L12-v2	0.71	0.69	0.25	16.7	10	0.23	11.3	35.3	50.6
	GISTv0	0.78	0.78	0.65	12.9	6	0.32	18	46.8	65.6
	GISTv0 Large	0.8	0.79	0.45	11.3	6	0.32	16.6	49	67
	OpenAI text_emb 3 small	0.78	0.79	0.28	12.9	6	0.33	19	48.1	65.6
	OpenAI text_emb 3 large	0.83	0.82	0.28	9.5	5	0.37	22	55.9	71.8
Fine-tuned LLMs	all-MiniLM-L6-v2 fine-tuned	0.93	0.93	0.6	6.87	4	0.4	23.9	58.8	78
	GISTv0 fine-tuned	0.94	0.94	0.7	6.3	4	0.4	23.9	61.9	80.5
	GISTv0 Large fine-tuned	0.94	0.94	0.65	6.3	4	0.41	23.5	62	81.2
Ensemble Models	OpenAI Small + L6v2 fine-tuned	0.94	0.94	0.45	5.82	3	0.43	26.4	63.6	82.5
	OpenAI Small + GISTv0 fine-tuned	0.95	0.95	0.5	5.6	3	0.45	28.8	66.1	83.4
	OpenAI Small + GISTv0 Large fine-tuned	0.95	0.95	0.5	5.52	3	0.45	28.8	63.9	83.4
	OpenAI Large + L6v2 fine-tuned	0.94	0.94	0.45	5.58	3	0.45	28.6	66.5	84.5
	OpenAI Large + GISTv0 fine-tuned	0.95	0.95	0.48	5.53	3	0.45	28.4	67.6	83.6
	OpenAI Large + GISTv0 Large fine-tuned	0.95	0.95	0.5	5.32	3	0.46	28.2	69	84
RFQ2RFQ										
Traditional	Word2Vec	0.5	0.67	0.3	127.7	44	0.08	0	18.2	18.8
	en_core_web_md	0.5	0.62	0.5	60.4	6.5	0.33	20.3	45.9	57
	GloVe	0.5	0.6	0.75	91.45	9.5	0.3	20	42.4	50.9
	TF-IDF	0.63	0.62	0.01	126.75	5	0.4	27.9	52.4	62.6
LLMs	BERT	0.55	0.62	0.7	44.8	4	0.39	28.2	54.1	63.5
	all-MiniLM-L6-v2	0.74	0.72	0.25	17.57	2.5	0.51	39.4	68.2	80
	all-MiniLM-L12-v2	0.71	0.69	0.25	19	2	0.53	42.4	68.2	76.2
	GISTv0	0.78	0.78	0.65	13.7	2	0.55	44.1	69.11	78.2
	GISTv0 Large	0.8	0.79	0.45	13.7	2	0.55	42	68.8	78.2
	OpenAI text_emb 3 small	0.78	0.79	0.28	11.78	2	0.56	42.9	71.5	79.7
	OpenAI text_emb 3 large	0.83	0.82	0.28	10.28	2	0.58	45.9	70.9	80.3
Fine-tuned LLMs	all-MiniLM-L6-v2 fine-tuned	0.93	0.93	0.6	20.4	3	0.49	38.2	61.2	70.9
	GISTv0 fine-tuned	0.94	0.94	0.7	20.2	2	0.51	39.7	63.8	75
	GISTv0 Large fine-tuned	0.94	0.94	0.65	26.4	2	0.52	40	67	75.9
Ensemble Models	OpenAI Large + OpenAI Small	0.82	0.81	0.28	11.7	2	0.57	45.6	70	79.7
	OpenAI Large + GISTv0 Large	0.82	0.82	0.35	11	2	0.57	45	70.6	79.7
	OpenAI Large + L6v2 fine-tuned	0.94	0.94	0.45	13.2	2	0.58	46.2	70.6	80.9
	OpenAI Large + GISTv0 Large fine-tuned	0.95	0.95	0.5	16.5	2	0.59	47.9	72.4	79.4

Note: Mean and median rank values are out of 451 possible matchings for the RFQ2COMP matchings and out of 340 possible matchings for the RFQ2RFQ matchings.



(a) GIST Large Embedding v0 Testing Data Class Separation



(b) GIST Large Embedding v0 Fine-Tuned Testing Data Class Separation

Figure 4: Comparison of GIST Large Embedding Before and After Fine-Tuning

5.4 Comparison of Open vs. Closed Source Model Performances

We can see in Table 5 that both the open and closed source LLMs have performed well on the task of entity matching in procurement. The increased embedding length of the OpenAI embeddings does tend to correlate with slightly better results than the smaller open-source embeddings. This highlights the potential benefit than these longer embeddings may therefore have in capturing more meaningful representations. Of the base LLMs the best performing model is the OpenAI text-embedding-3-large model. Both its thresholding and ranking scores exceed that of the other base LLMs. With over 45.9 percent of the rankings when matching the RFQ to past RFQs correctly reproducing the matching produced by Applegate, the OpenAI large model shows remarkably high quality performance. Its implementation in a B2B platform would therefore yield high quality results.

As previously highlighted, there are benefits to using API based LLMs over locally run open-source models, particularly for users with less computing power. However, these embeddings do take a significant amount of time to obtain for larger batches. It takes around 2-3 hours to obtain 4000 embeddings using the OpenAI Large model, while taking approximately 20 minutes to calculate these embeddings using the GISTv0 Large model on a NVIDIA A100 GPU. These delays in embedding obtention present a potential problem for procurement platforms. If the load suddenly increases on the platform and the embeddings are slow to obtain, then this may cause significant overhead for the company and lead to backlogs. It could therefore be envisioned that a combination of different approaches may be used, with performance potentially being sacrificed for speed in times of exceptionally high demand. However, the time needed to obtain matchings can be significantly improved by storing the supplier descriptions and past RFQ embeddings in databases in order to reduce the need for redundant calculations. Further work would therefore be needed to establish how much of an impact using closed-source models has on the speed of the processes.

5.5 Potential of Combining Models for Better Prediction

Building on the results obtained, we investigated the possibility of combining the embeddings of multiple models in order to improve predictions. In Figure 5 we compare the similarity scores of embeddings produced for identical supplier descriptions in the testing data using the all-MiniLM-L6-v2 and the all-MiniLM-L12-v2 models. We do this in order to investigate whether the representations learned by the models differ from each other. We can see that the similarity scores are approximately

normally distributed, centred around 0.51. Considering these embeddings represent the exact same pieces of text, these similarity scores are relatively low. We can therefore hypothesise that the models are extracting separate meaningful representations. This highlights the potential of combining the embeddings of multiple models to make better predictions. In order to do this, the elements of the embeddings are added together and then averaged. For cases where embedding lengths differ, the smaller embeddings are padded to the length of the longer embeddings beforehand.

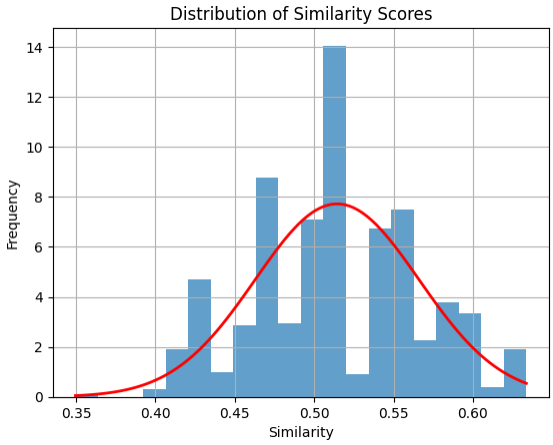


Figure 5: Comparison of all-MiniLM-L6-v2 and all-MiniLM-L12-v2
Supplier Description Embedding Similarity for Testing Data

In Table 5 we can see that our ensemble models, which combine the embeddings of various models, outperform all the other models, including the fine-tuned ones. We can see that combining models such as OpenAI's models, which had lower class thresholding separation, with models which have higher class separation, such as the fine-tuned models, leads to both greater class separation and ranking performance. The combination of our two best models, the OpenAI Large and the fine-tuned GIST Large, leads to the best results. It gets an accuracy and F1 score of 95% when using a threshold of 0.5 and 47.9% of the matchings made by the model are identical to the ones present in the Applegate matchings. The combination of these models therefore offers impressive performance and highlights the potential benefits of combining the predictions of multiple models. Such an implementation would therefore provide very high quality predictions if integrated into a procurement platform.

5.6 Model Potential

With our models showing good performance we can proceed to use them for making predictions. We propose two approaches for doing this. Firstly, we propose a method which makes predictions based solely on similarity scores. These scores can be obtained using one or multiple models. It calculates the similarity between an inputted RFQ and either supplier descriptions or past RFQs and ranks the matchings based off of the similarity score. The match with the highest similarity score can then be chosen as a potential good matching. In the case of using multiple models, we can either average the similarity scores across the various predictions and take the top candidates or make the individual predictions and see which matchings the predictions have in common.

The second method focuses on combining the embeddings of multiple models and then calculating the similarity scores similar to the method described in section 5.5 and then taking the highest similarity scores. In a real-world application, multiple methods may be implemented in order to maximise the quality of the predictions. Additionally, we could foresee that different models may be given different importance weightings in the predictions in order to give more importance to the better performing models. Overall, both the methods set out above allow for good suppliers to be matched

within seconds to an inputted RFQ.

In Figure 6 we present an analysis of the generalisation capabilities of the fine-tuned GIST Large model. We test the performance of the model fine-tuned on the approximately 4000 paying supplier matches when applied to over 21000 non-paying supplier matches. We can see that the threshold separation worsens a little bit, particularly for the negative class, compared to testing on the smaller paying supplier dataset. However, the quality remains high despite the large amount of data it is tested on. This gives initial insight into the potential generalisation capabilities of the models, even when fine-tuned. This therefore solidifies the prospect of their integration into a real-world platform, where generalisation capability is key.

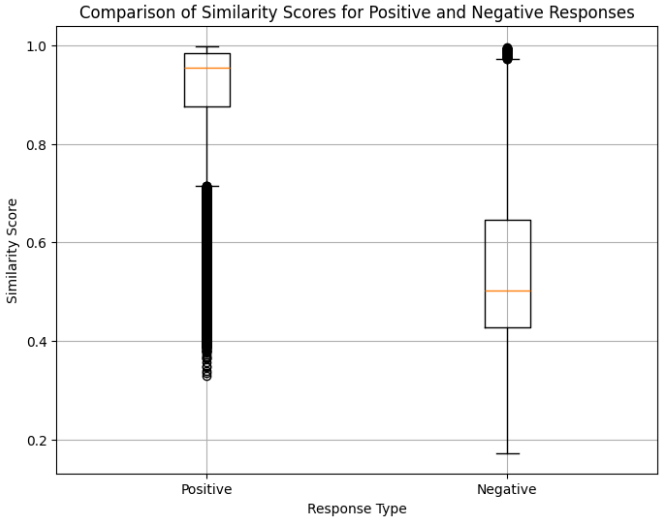


Figure 6: Model Generalisation of Fine-Tuned GISTv0 Large on Non-Paying Supplier RFQs

Furthermore, there is potential for models to be further improved over time. This improvement can happen in two respects. Firstly, newer LLMs are currently emerging every month, with these models often offering higher quality performance. Therefore, the quality of matchings may be improved over time with the release of better LLMs. Secondly, as Applegate produces more data of good quality matchings, the fine-tuned models can be updated with more data. This can be done simply by running the fine tuner again on the new data. This may allow the fine-tuned models to generalise further. The findings above, while of good quality, therefore represent only the beginning of a long road of possible improvements which can be made to the procurement matching process.

5.7 Limitations of Implemented Methodology

While the project has achieved exceptionally good results, there remains some limitations to the implemented methodology. These include limitations with regards to the data used, the training method and LLMs in general.

With regards to limitations in the data used there are multiple potential downsides. Firstly, the data had to be significantly limited, with most models using data from paying suppliers only. This is due to mislabelling problems present in large part in the non-paying supplier matching data including matchings which were said to be good but were composed of suppliers whose industries were separate from the products requested. These problems were most notably present in the data from 2020, at the height of Covid Pandemic. These problems may therefore be due to an overloaded system during that time and may therefore be less present as new data is generated.

Using LLMs we found that such mislabelling could be identified, to a certain extent, by analysing the similarity scores. Through manual analysis of matchings which were said to be good but had very low similarity scores we found several instances of mislabelled data. These cases were mostly clustered in the non-paying supplier data. This therefore highlights how an LLM based matching system may provide benefit during times of overload both by providing matches but also by identifying potential bad matches. Caution must however be carried out as supplier descriptions may not fully encapsulate the range of products a company offers and therefore matches which appear wrong may be correct. Due to these issues, it may therefore be in the interest of a company wanting to implement a system such as the one in this paper to use verified labelled data in order to train their models.

Furthermore, the data used in this project is composed solely of RFQs and supplier descriptions in English. This may therefore limit the generalisability capabilities of the models if the system were to be deployed internationally. As LLMs are improved for better cross-language capabilities, these limitations may become less pronounced over time. A truly multilingual LLM would provide a powerful tool for enabling cross-border trade.

With respect to the training method, limitations include the lack of cross-validation of the fine-tuning process due to resource limitations. It was not feasible to conduct such cross-validation, particularly on larger models such as GIST Large Embedding v0 which use up the full capacity of the top of the range GPU available on Google Colaboratory. We can, however, hypothesise that the generalisation capabilities shown in Figure 6 provide a good indication that the model generalises well.

One further limitation lies in the evaluation methodology, due to the fact that we only makes use of cosine similarity as the comparison metric. While this is the metric most commonly associated with LLM vector embeddings, a growing literature is showing the limitations of the metric (Steck et al. 2024). Before being deployed it may therefore be worth investigating other comparison metrics such as the unnormalised dot product to see if they may provide better matching capabilities.

Finally, as LLMs are NNs, they are modelled as black boxes. This means that their inner components, in particular their weightings, are difficult to interpret. It is therefore complicated to understand how a model arrives at a certain vector representation. This may limit the capability of users to gather metric such as the uncertainty surrounding the representations and predictions. Future LLMs may therefore wish to implement Bayesian methods for quantifying uncertainty when making predictions.

Overall, these limitations remain limited and the implementation of LLMs remains a good source of predictions. Their implementation would provide a powerful tool for a procurement platform to perform the necessary matchings. However, further research to address the limitations set out above may be advised.

6 Further Discussion

6.1 Global Impact of AI Integration in Procurement

The system implemented has shown the viability of using LLMs to match RFQs to adequate suppliers in a B2B procurement platform. A real-world implementation would be capable of providing high quality matchings which could be used by a platform. Such an implementation could significantly improve the procurement matching processes, both in terms of quality and speed. It could solidify the attractiveness of a platform and mitigate potential problems linked to crises. It also shows that LLMs provide more informative vector embeddings than traditional method. LLMs, therefore, truly do represent a turning point in the field of NLP and have the power to revolutionise the procurement sector.

6.2 Feasibility of Integration

The integration of such as system would not only be beneficial but would also be fairly simple to implement. The models shown provide a solid foundation for the backend of a matching system, so its integration would solely necessitate an appropriate frontend interface. The integration would also come at a relatively small cost, with the fine tuning of models being possible for under 10 dollars, while the obtention of embeddings from OpenAI's API are continually getting cheaper as models are optimised. Its integration would also allow the procurement platform to advertise its use of LLMs and AI, a characteristic which customers are increasingly looking for in services they use. Overall, the integration would be simple, cost effective and of high quality.

6.3 Further Research Opportunities

In addition to addressing the limitations set out with regards to the implemented methodology, there are multiple threads for further potential research. One is to include NER methods into the procurement process, either within the matchings themselves, or as a supplement. These methods are useful for extracting key information such as those shown in Table 6, which was extracted using OpenAI's GPT 3.5. NER may also provide an extra way of speeding up the matching process, in particular if human agents are still involved in the process, by allowing for faster extraction of the vital points of long RFQs.

RFQ	**Extracted NER Info**
Please provide a quote for 100 units of FFP2 face masks with CE markings to be delivered by March 1st to London.	• Quantity: 100 units • Regulatory Marking: CE markings • Location: London • Date: March 1st • Product: FFP2 face masks

Table 6: RFQ and Extracted NER Information

Further work may also look into making matchings take into account specific requirements such as delivery times, location, order size and carbon emissions for its matchings. Integrating these extra features would move the system closer to being fully automated and would provide further quality improvements. Overall, many threads of further research are possible with regards to the integration of LLMs into procurement platforms, with the sector likely to considerably evolve over the coming years.

7 Conclusion

In conclusion, this paper has shown that the integration of LLMs to match RFQs with appropriate suppliers in a B2B procurement platforms is both highly effective and efficient. The research has shown that there are multiple different LLMs and techniques which may be used for the task, each with their own advantages and limitations. Furthermore, we showcased how the use of ensemble models can significantly enhance performance, leading to models with a quality very close to that of human procurement agents. These findings suggest that there is immediate potential for the integration of LLMs into the matching processes of procurement platforms and that this can come at a relatively minor cost. We estimate that such an integration would increase satisfaction amongst both buyers and sellers while minimising the risk of system overload. Moreover, we estimate that the scope of

integration of LLMs extends beyond that of matchings and that further work in the field is necessary to explore the viability of further implementations. Overall, we have has shown that the vector embedding capabilities offered by LLMs far outperform those produced by older methods underlining the clear opportunities for their widespread integration into procurement platforms.

References

Ahmed, T., Bird, C., Devanbu, P. & Chakraborty, S. (2024), 'Studying llm performance on closed-and open-source data', *arXiv preprint arXiv:2402.15100* .

Andersson, P., Cramner, I., Nadeem, H. & Rosenqvist, C. (2023), Implementing ai in source to contract operations: How procurement managers in a global organization make sense of ai opportunities and inhibitors, Technical report, EasyChair.

Applegate (n.d.).
URL: *https://www.applegate.co.uk/*

Attinasi, M. G., Balatti, M., Mancini, M. & Metelli, L. (2022), 'Supply chain disruptions and the effects on the global economy', *Economic Bulletin Boxes* **8**.

Beason, S., Hinton, W., Salamah, Y. A. & Salsman, J. (2021), 'Automated analysis of rfps using natural language processing (nlp) for the technology domain', *SMU Data Science Review* **5**(1), 1.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. & Askell, A. (2020), 'Language models are few-shot learners', *Advances in neural information processing systems* **33**, 1877–1901.

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

Dolle, N., Wilhelm, C., Wergunow, A., Rössle, M., Fernandes, M. & Glißmann, L. (2020), An artificial intelligence based sourcing automation concept for smaller and mid-sized enterprises in the metal industry, *in* 'International Conference on Reliability and Statistics in Transportation and Communication', Springer, pp. 94–103.

Hadsell, R., Chopra, S. & LeCun, Y. (2006), Dimensionality reduction by learning an invariant mapping, *in* '2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)', Vol. 2, IEEE, pp. 1735–1742.

Honnibal, M. & Montani, I. (2017), 'spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing', *To appear* **7**(1), 411–420.

Hu, K. (2023), 'Chatgpt sets record for fastest-growing user base - analyst note'.
URL: *https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/*

Lake, R. (2023), 'What is a large language model (llm)?'.
URL: *https://www.investopedia.com/large-language-model-7563532*

Lee, T. & Trott, S. (2023), 'A jargon-free explanation of how ai large language models work'.
URL: *https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/*

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

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Muennighoff, N., Tazi, N., Magne, L. & Reimers, N. (2022), 'Mteb: Massive text embedding benchmark', *arXiv preprint arXiv:2210.07316* .

Nia, A. G., Lu, J., Zhang, Q. & Ribeiro, M. (2019), A framework for a large-scale b2b recommender system, *in* '2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)', IEEE, pp. 337–343.

OpenAI (2023*a*), 'Gpt-4 technical report', *ArXiv* **abs/2303.08774**.
URL: *https://api.semanticscholar.org/CorpusID:257532815*

OpenAI (2023*b*), 'Introducing chatgpt.'.
URL: *https://openai.com/blog/chatgpt*

OpenAI (2024), 'New embedding models and api updates'.
URL: *https://openai.com/blog/new-embedding-models-and-api-updates*

Pennington, J., Socher, R. & Manning, C. D. (2014), Glove: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018), 'Improving language understanding by generative pre-training'.

Rahutomo, F., Kitasuka, T., Aritsugi, M. et al. (2012), Semantic cosine similarity, *in* 'The 7th international student conference on advanced science and technology ICAST', Vol. 4, University of Seoul South Korea, p. 1.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. & Sutskever, I. (2021), Zero-shot text-to-image generation, *in* 'International conference on machine learning', Pmlr, pp. 8821–8831.

Reimers, N. & Gurevych, I. (2019), 'Sentence-bert: Sentence embeddings using siamese bert-networks', *arXiv preprint arXiv:1908.10084* .

Reimers, N. & Gurevych, I. (n.d.*a*), 'sentence-transformers/all-minilm-l12-v2', https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2. Accessed: 2024-04-24.

Reimers, N. & Gurevych, I. (n.d.*b*), 'sentence-transformers/all-minilm-l6-v2', https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2024-04-24.

Rekabsaz, N., Lupu, M. & Hanbury, A. (2017), Exploration of a threshold for similarity based on uncertainty in word embedding, *in* 'Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39', Springer, pp. 396–409.

Rui Meng, Ye Liu, S. R. J. C. X. Y. Z. S. Y. (2024), 'Sfr-embedding-mistral:enhance text retrieval with transfer learning', Salesforce AI Research Blog.
URL: *https://blog.salesforceairesearch.com/sfr-embedded-mistral/*

Singh, S. (2021), 'Bert algorithm used in google search', *Mathematical Statistician and Engineering Applications* **70**(2), 1641–1650.

Solatorio, A. V. (2024), 'Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning', *arXiv preprint arXiv:2402.16829* .
URL: *https://arxiv.org/abs/2402.16829*

Sparck Jones, K. (1972), 'A statistical interpretation of term specificity and its application in retrieval', *Journal of documentation* **28**(1), 11–21.

Steck, H., Ekanadham, C. & Kallus, N. (2024), 'Is cosine-similarity of embeddings really about similarity?', *arXiv preprint arXiv:2403.05440* .

Thongtan, T. & Phienthrakul, T. (2019), Sentiment classification using document embeddings trained with cosine similarity, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop', pp. 407–414.

Times, F. (2023), 'Microsoft confirms 'multibillion-dollar investment' in chatgpt maker openai'.
URL: *https://www.ft.com/content/298db34e-b550-4f80-a27b-a0cf7148f5f6*

Trappey, A. J., Chang, A.-C., Trappey, C. V. & Chien, J. Y. C. (2022), 'Intelligent rfq summarization using natural language processing, text mining, and machine learning techniques', *Journal of Global Information Management (JGIM)* **30**(1), 1–26.

Turing, A. M. (1950), *Computing machinery and intelligence*, Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, & Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B. & Kasirzadeh, A. (2021), 'Ethical and social risks of harm from language models', *arXiv preprint arXiv:2112.04359* .

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. M. (2020), Transformers: State-of-the-art natural language processing, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Online, pp. 38–45.
URL: *https://www.aclweb.org/anthology/2020.emnlp-demos.6*