

**B9DA112 Applied Research Methods**

Research Proposal: Enhancing text and voice(DBS specific) based Chatbot using Transformers and comparing it with traditional models LSTM

**Student Name**: Sandeep Kumar

**Student Number**: 2004975

**Lecturer**: Paul McEvoy

**Module Code**: B9DA112

Contents

[Contents 2](#_Toc202366425)

[**Abstract** 3](#_Toc202366426)

[**Introduction** 3](#_Toc202366427)

[**Research Question** 4](#_Toc202366428)

[**Research Objectives** 4](#_Toc202366429)

[**Literature Review** 5](#_Toc202366430)

[**Early Chatbots and Foundational Theories** 6](#_Toc202366431)

[**NLP and Conversational AI** 9](#_Toc202366432)

[**NLP and Deep Learning(RNN, LSTM, GRU)** 11](#_Toc202366433)

[**Transformers and Beyond(BERT, SSM, Memory-Augmented)** 13](#_Toc202366434)

[**Research Methodology** 14](#_Toc202366435)

[**Research Philosophy** 14](#_Toc202366436)

[**Research Approach** 15](#_Toc202366437)

[**Research Design** 16](#_Toc202366438)

[**Research Method** 16](#_Toc202366439)

[**Data Collection and Analysis** 17](#_Toc202366440)

[**Artefact Design and Presentation** 19](#_Toc202366441)

[**Timeline** 20](#_Toc202366442)

[**Conclusion** 20](#_Toc202366443)

[**References** 20](#_Toc202366444)

Research Proposal: Enhancing text and voice(Dublin Business School specific) based Chatbot

**using Transformers and comparing it with traditional models LSTM**

# **Abstract**

This work looks at the development of a chatbot tailored for Dublin Business School(DBS), using transformer-based models, state-of-the-art architectures in raw language processing that render contextually relevant text **(Behl and Bibhu, 2024)** and comparing their efficiency against “Long Short-Term Memory”(LSTM) network, a recurrent neural network architecture capable of capturing long range dependencies crucial for understanding context **(Hochreiter & Schmidhuber, 1997).** The methodology adopted in this study is quantitative in nature and aims to evaluate metrics such as response accuracy, model training time & comprehension rate in order to answer research questions and objectives. While transformers are expected to surpass the traditional models **(Vaswani et al., 2017)**, limited scope of this work does not allow to  incorporate qualitative aspects such as user experience, emotions or subjective feedback during interactions with chatbot.  It is recognized that these areas would bring a more rounded aspect to the work, and further study would need to be undertaken to further this area. However, this  study may provide easy, scalable and adaptable solutions for other institutions beyond DBS, offering a framework for broader educational chatbot implementations.

# **Introduction**

AI assistants are a technology that has transformed the way businesses interact with their customers by providing self-service power to the users, helping them get what they want, when they want it, the way they want it **(Freed, 2021).** These assistants are saving significant labor and operational costs for businesses by automating routine enquiries and support tasks **(Patil et al., 2024)**. One prominent example of AI assistants is Conversational AI, or chatbots, which fall under this category and use full conversational dialogue to accomplish one or more tasks **(Freed, 2021).**

Beyond business environments, the application of AI-powered chatbots has also seen rapid growth in the education sector in recent years particularly for providing student support and improving communication **(Tapia‑Hoyos, 2021).** This growth is partly due to the fact that traditional communication between prospective students and universities is often handled manually, which can be a time-consuming process and a burden on admissions staff (**Nguyen, Le, Hoang, and Nguyen, 2021**).



These virtual assistants not only answer and manage queries but also offer a spectrum of services like academic tutoring, counselling, career guidance. They can also automate routine tasks like ticket creation for administration and IT, access to academic records, or academic advising **(Mashilo et al., 2024).** This growing reliance stems from the need for 24/7 availability, scalability, and delivery of context‑aware, consistent, and accurate responses—capabilities often lacking in traditional systems **(Patil et al., 2024)**.

To appreciate how conversational AI reached its current capabilities, particularly in education, it is important to examine its evolution over time. The evolution of conversational AI spans several decades. Early systems, inspired by the Turing test **(Turing, 1950)**, relied on rule-based, scripted responses, as seen in pioneers like ELIZA **(Weizenbaum, 1966)** and PARRY **(Colby, K., 1975)**. These systems could simulate conversation but were domain-specific and unable to maintain dialogue over extended interactions. They were rule based and had no contextual intelligence. Similarly, vector-based models such as Word2Vec **(Mikolov et al., 2013)** and GloVe **(Pennington et al., 2014)** improved semantic representation but had limited ability to process multi-turn conversations effectively. More recent advances in deep learning architecture, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, allowed better understanding of sequences and conversational flow. However, these models often suffer from vanishing gradients and are computationally inefficient for longer dependencies(**Hochreiter & Schmidhuber, 1997)**.

These limitations in earlier models prompted researchers to explore more advanced architectures, most notably the Transformer-based model introduced by **(Vaswani et al., 2017)** in their seminal paper “Attention is All You Need.” In this paper they explained how the Transformer architecture uses self-attention mechanisms that allow the model to weigh the relevance of each word in a sentence relative to the others, thereby improving both processing efficiency and the preservation of conversational context. The emergence of the Transformer architecture addressed many of the challenges by introducing parallel processing through self-attention, contextual coherence, and transfer learning**(Behl and Bibhu, 2024).** This progression underscores the significance of using a Transformer-based approach in enhancing chatbot capabilities within a domain-specific academic context like DBS, where student queries responses are grounded in institutional data.

Building on these technological advancements, the present research attempts to apply transformer-based architecture within academic domain, aiming to demonstrate its potential advantages over traditional LSTM models in developing a DBS-specific chatbot. Two chatbot variants, one using Transformers and the other using LSTM, are expected to be developed and evaluated based on quantitative performance metrics, including response accuracy, training time, comprehension rate, and model perplexity, using a customized dataset derived from DBS websites, brochures, and institutional documentation.

Building on the discussion so far and considering the overall intention of this work, the following research questions and commensurate objectives are proposed.

## **Research Question**

How can transformer-based architecture improve efficiency ("Attention is All You Need" by Vaswani et al. in 2017) and accuracy of a DBS-specific chatbot compared to traditional models like LSTM?

## **Research Objectives**

1. **Design and develop a DBS-specific chatbot using transformer-based and LSTM based models** – This will allow implementation and demonstration of comparable chatbot versions leveraging Transformers and LSTM models, respectively for evaluating their capabilities in responding to the student specific queries.
2. **Compare the chatbot’s performance against traditional models like LSTM(training time , comprehension rate, perplexity)** – This comparative analysis will attempt to provide metrics essential for comparison between the chatbot versions, addressing directly the research question.

# **Literature Review**

The Literature Review section looks to examine the current scholarly articles pertaining to Evolution of Conversational AI and rise of NLP and deep learning models. It is divided into four main topics, these are Early chatbots and fundamental theories, NLP and Conversational AI , NLP and Deep Learning(RNN, LSTM, GRU), Transformers and Beyond(BERT, SSM, Memory-Augmented)

A diagram with text and colorful lines

AI-generated content may be incorrect.

Figure 1 Literature map

## **Early Chatbots and Foundational Theories**

1. **"Computing Machinery and Intelligence" (Alan Turing, 1950) -** The conceptual roots of the conversational systems trace back to the release of groundbreaking paper "Computing Machinery and Intelligence" **(Alan Turing, 1950),** where he proposed the now-famous Turing Test as a means to evaluate machine intelligence. In this he addresses and reframed the most fundamental question of artificial intelligence: can machine think? Into an imitation game which determines whether a machine can exhibit human behavior or not. He argued that if computers are given enough memory, machines can learn and simulate human reasoning. This laid the foundation of future developments and machine learning in AI influencing development of conversational agents and chatbots which is also significant in the context of this research.
2. **“A computer program for the study of natural language communication between man and machine” (Weizenbaum, 1966)** - Building on these ideas, The 1960s saw the creation of Eliza one of the earliest and influential development of conversational system. Keywords appearing in the input triggers the decompositions rules which help in analyzing the input sentence. Responses are then generated by reassembly rules associated with selected decomposition rules. These transformative rules and keywords constitute the script for a particular class of conversation which is data for the ELIZA **(Weizenbaum, 1966**). Consider following sentence.

Figure 2 Decomposition ad assembly rule example (Weizenbaum, 1966)

"It seems that you hate me”

In this sentence lets assume that machine understands only two words “You” & “me”

After applying decomposition template sentence looks like

1-It seems that 2-you 3-hate 4-me

Where decomposition template looks like

“0 YOU 0 ME”

Where “0” in the decomposition rule stands for indefinite number of words.

The reassembly rule might be

“What makes you think I hate you”

This assembly rule template threw away the first component(“It seems that”), translated the two known words (2-“you” to “I” and “me” to “you”) and added a phrase “” to front of the sentence “What makes you think” **(Weizenbaum, 1966).**

A close-up of a message

AI-generated content may be incorrect.

Figure 3 A conversation between Eliza and a human (Weizenbaum, 1966)

The program gave the illusion of understanding despite no cognitive capabilities. Through his study, Weinbaum’s not only demonstrated early computational linguistics but also highlighted the limitations of human-computer interaction and expressed concern about humans overinterpreting humans’ behavior.

1. **Artificial Paranoia: A Computer Simulation of Paranoid Processes (****Colby, K., 1975)-** In 1975, Colby introduced PARRY, another conversational chatbot to simulate the conversational behavior of a person with paranoid schizophrenia (overly suspicious, anxious, or fearful person). PARRY incorporated a psychological model that accounted for beliefs, fears, and hostility levels unlike ELIZA, which used pattern matching mechanisms to produce responses. Model’s (known as Paranoid model in the paper) Input output behavior imitates the behavior of human whose information processing is dominated by a mode psychiatrist as “paranoid”. The task of the program of this “Paranoid model” (fig 1 below) is to interpret the input expressions and to produce internal (affective) and external(linguistics) responses **(Colby, K., 1975)**.

**A diagram of a diagram

AI-generated content may be incorrect.**

Figure 4 Paranoid model’s program general tasks(Colby, K., 1975)

In the context of this research, PARRY marks a significant step from simple scripted replies to architecture which was more advanced as it exhibited emotionally reactive dialogue patterns for example Parry will respond with hostility if the anger level is high.

1. **The Elements of AIML Style. ALICE A.I. Foundation (Wallace, 2009) -** Another important technology used in the evolution of chatbots was Artificial Intelligence Markup Language(AIML) developed by Richard Wallace. AIML defines a set of data-structures known as AIML objects which partially defines the computer programs that process them. These units are made up of units called topics and categories and contain parsed or un-parsed data.

**A screenshot of a computer program

AI-generated content may be incorrect.**

Figure 5 A simple AIML structure (Wallace, 2009).

This simplified XML-based language includes tags enabling bots to recursively invoke pattern matchers, thereby streamlining the process of language parsing and response generation(**Shum, He, and Li, 2018**).

## **NLP and Conversational AI**

Integration of machine learning revolutionized the conversational system, particularly in the early 2000s, when system learned patterns from data, instead of relying solely on the rule based or pattern matching criteria. These Machine learning approaches focused on probabilistic intent recognition, using statistical methods to map user inputs to appropriate responses.

1. **The pipeline processing of NLP. In E3S Web of Conferences** **(Elov, Khamroeva, and Xusainova, 2023)** - Step by step processing of the text in NLP is called s NLP pipeline.

Figure 6 Typical NLP pipeline (Elov, Khamroeva, and Xusainova, 2023)

The first phase of the pipeline includes the text cleaning and development of necessary features which are converted into a comprehensible format by modeling algorithms. The conversion process typically analyzes text by segmenting sentences and words, techniques known as sentence segmentation and word tokenization. Then the tokenized words are converted into their basic form by getting suffixes, prefixes removed through techniques like Stemming and Lemmatization. Stop words, punctuations, digits are removed**(Elov, Khamroeva, and Xusainova, 2023).**

Next stage converts tokens into numerical format (text representation or vectorization) for machine learning models. Bag-of-words is one of the earliest techniques and describes the occurrence (frequency) of words in a particular document. In this there is no consideration of the order of the words or their semantic meaning. Another scheme TD-IDF(Term Frequency–Inverse Document Frequency) was developed which not only measured the frequency of the words but their importance as well in a document relative to a corpus. One-hot-encoding represents words as binary vectors and leads to dimensionality issues for larger sentences or vocabulary.

Another class of techniques known as Word embeddings in Natural Language Processing (NLP) are a way to represent words as numerical vectors which capture semantic and syntactic relationships between words, allowing machines to understand the meaning and context of words. The earlier word representation techniques treated words as an atomic unit and there is no notion of similarity between words.

1. **Indexing by Latent Semantic Analysis- Scott (Deerwester et al. ,1990)-** Earlier techniques tried to match words of queries with words of documents. The problem with approach was that users wanted to retrieve on the basis of conceptual content, and individual words provided unreliable evidence about the conceptual topic or meaning of a document. The goal in this research was to find out the relationship between terms and documents (i.e. to uncovering latent semantic structure) to estimate the parameters of the underlying model. In simple terms term-document matrix is created its dimension is reduced using matrix factorization mathematical techniques. The reduced matrix reveals the relation between word and document related to these latent dimensions **(Deerwester et al. ,1990).**
2. **Efficient estimation of word representations in vector space(Mikolov et al., 2013)**

Word2Vec architecture introduced by **(Mikolov et al., 2013)** was a significant advancement in text representation as it captures the semantic relationship between the words. In this paper, several techniques were introduced to learn high quality word vectors from corpus as big as billion words vocabulary. The most word vector methods rely on the distance or angle between pairs of word vectors as the primary method for evaluating the intrinsic quality of such a set of word representations. The vectors representations of the similar word provide multiple degrees of similarity between them. For example, even if the following algebraic operations are performed on the vector’s representations of the words “King”, “Man” & “women” below, resultant vector will be close to the vector representation of the word Queen.

vector(”King”)- vector(”Man”) + vector(”Woman”) ~ vector(“Queen”)

The architectures used in Word2Vec includes “Continuous Bag-of-Words” & “Continuous Skip-gram” log-linear Models to predict words from their surrounding context **(Mikolov et al., 2013).**

1. **GloVe: Global vectors for word representation(Pennington et al., 2014)**

**(Pennington et al., 2014)** introduced an alternative architecture, GloVe (Global Vectors for Word Representation), which integrates the strengths of both global matrix factorization techniques like Latent Semantic Analysis (LSA) **(Deerwester et al. ,1990)** and local context window models such as the skip-gram approach proposed by **(Mikolov et al., 2013)**. While LSA tends to underperform on semantic tasks and skip-gram models often fail to fully leverage global corpus statistics, GloVe addresses these limitations by employing a global log-bilinear regression model. This unsupervised method effectively captures word co-occurrence information across the entire corpus and has demonstrated superior performance on tasks such as word analogy, semantic similarity, and named entity recognition.

## **NLP and Deep Learning(RNN, LSTM, GRU)**

Before deep learning, NLP relied on language models like Bag-of-Words, TF-IDF, and n-gram which struggled with data sparsity, poor generalization to the unseen data and limited context windows. These shortcomings paved the way for neural architecture. Neural networks are built out of neural units, originally inspired by biological neurons but now simply an abstract computational device. A neural network consists of layers of neurons, which are connected to each other in fully connected feedforward configuration. The power of neural networks comes from the ability of early layers to learn representations that can be utilized by later layers in the network **(Jurafsky & Martin, 2009).**

1. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Jurafsky & Martin, 2009) –** Simplest form of Neural network is feedforward network which is a multi-layer network in which units are connected with no cycles. Output flows from one layer to other higher layers in one direction. It represent words in prior context by their embeddings rather than just by their word identity as in as n-gram language models.

**A diagram of a network

AI-generated content may be incorrect.**

Figure 7 A simple 2-layer feedforward network (Jurafsky & Martin, 2009)

Consider a training sentence “I have to make sure that the cat gets fed.” But model has never seen the words “gets fed” after the word “dog”. Now let’s consider a test sentence “I forgot to make sure that the dog gets”. The n-gram will not be able to predict “gets fed” after the word “dog” but neural network will be able to predict correctly as both words “dog” and “cat” have similar embeddings.

RNN(Recurrent Neural Networks) on the other hand contains cycle within its network connections meaning that values of the units can be dependent on their earlier outputs as an input. **(Jurafsky & Martin, 2009).**

A diagram of a complex flowchart

AI-generated content may be incorrect.

Figure 8 Simple recurrent neural network after Elman (1990).

The hidden layer includes a recurrent connection as part of its input in above simple RNN and provides a kind of memory or context. RNN language models **(Mikolov et al., 2010)** process the input sequence of one word at a time, attempting to predict the next word from the current word and the previous hidden state.

There are multiple difficulties with RNN like their limited capability to take forward the critical information and insufficient, decaying error back flow (**Hochreiter & Schmidhuber, 1997)** also known as vanishing gradient(vector signifying the loss function) issues. To address these issues more complex architecture has been designed to maintain the relevant context over time and forget the irrelevant information required for decision. LSTM (Loong Short Term Memory) is one of such architecture.

1. **Long short-term memory - Hochreiter, S. & Schmidhuber –** It is a novel RNN architecture which in conjunction with gradient based algorithm overcome error backflow issue of the RNN architecture.In addition to the Recurrent layer, it adds additional context layer through the use of specialized neural units that use gates to control the flow of information in and out of the networks. The gates use additional weights and consist of a feed forward layer with an activation function. LSTM works well over a broad range of parameters like learning rate, input gate bias and output gate bias.

Advancements in Deep Learning have played a key role in developing systems which can provide contextual and coherent suitable replies. It lets developers train chatbots on massive data sets, which help them understand natural language better than previous conversational agents. Well-known systems like Google’s Meena, OpenAI’s GPT-3.5 and GPT-4, and Facebook’s BlenderBot used complex architectures and large datasets to create advanced conversational agents. Even with these advancements, these models still faced problems like difficult training, lack of reliability, and trouble handling long-term information and led researchers to look for new architectures which could overcome these weaknesses.

## **Transformers and Beyond(BERT, SSM, Memory-Augmented)**

1. **Attention is all you need (Vaswani, A. et al., 2017)** - Introduction of Transformer based architecture (Vaswani et al., 2017), revolutionized the NLP by introducing self-attention mechanism, in which the entire sequence of words\tokens is processed in parallel rather than sequentially, which increased speed and accuracy, especially when working with larger datasets (in contrast to RNN). They are better at understanding context and long-range dependencies in text, crucial in coherent conversations. Transformers also benefit from transfer learning. Pre-trained models can be fine-tuned for specific tasks, making it easier and faster to build strong NLP systems without training from scratch. In Short, RNN laid the foundation for advanced conversational AI, transformers have taken it to the next level, in terms of performance, flexibility and speed, making them preferred choice for modern chatbot development.

As transformer-based models continue to advance, newer architectures, such as the State-space models (SSMs) which offer a powerful framework for dynamical system analysis and simplifies the transformer architecture by replacing complex attention and multi-layer perceptron (MLP) blocks with a single, unified SSM block**(Lin and Michailidis, 2024).**

Another architecture Memory-Augmented Model demonstrates that transformer-based large language models become computationally universal when combined with external memory. Memory-augmented models capture long-range dependencies more uniformly, especially across topic boundaries **(Schuurmans, 2023).**

These advancements indicate a shift towards more efficient, specialized, and adaptable architectures in conversational AI, moving beyond the limitations of traditional transformer models.

# **Research Methodology**

Research methodology can be referred to as theory of how research should be undertaken**(Saunders, Lewis and Thornhill, 2009)**. It can be viewed as a multi-stage process that must be undertaken systematically by researchers to produce valid results from reliable data that are aligned with the objective of the research. The stages include formulation of research topic, literature review, understanding the research philosophy, research approach to be undertaken, formulation of research design, addressing ethical issues, data collection(sampling, secondary, observation, semi-structured, questionnaire), data analysis(quantitative or qualitative or mixed), project deliverables like report and presentations**(Saunders, Lewis and Thornhill, 2009)**. These stages represent a perspective about research that presents information in a progressive way from philosophical framework to more specific and detailed procedures **(Creswell and Creswell, 2018)**.

In this section, research philosophy, approaches, design and methods will be covered.

## **Research Philosophy**

In planning a study researcher adopts a research philosophy that contains researcher’s thoughts and assumptions about the way he looks at the world **(Saunders, Lewis and Thornhill, 2009)**. In other words, it is the basis of research, which involves the choice of research strategy, formulation of problem, data collection, processing and analysis **(Žukauskas et al., 2018)**.

According to **(Creswell, Creswell 2018)** there are 4 main research philosophies, positivism, constructivism, transformative and pragmatism. Positivism is a deterministic philosophy where it is believed that causes likely lead to specific outcome, therefore it often focuses on identifying and evaluating the causal factors that shape the outcome, as seen in the experimental designs. It also aims to reduce the ideas into smaller sets such that it can be tested through hypothesis and research questions. Knowledge generated in this research is built on the measurement of objective reality or human behavior through measurable indicators. Accepted approach to research by positivists is when a formulated theory is either accepted or refuted based on data collection and analysis.

Constructivism, on the other hand, is typically associated with qualitative research. Subjective meanings are formed through personal experiences, which are complex and prompt researchers to explore multiple perspectives. Research questions are often open-ended, allowing researchers to construct broader understanding believed to be influenced by their social or historical background. Instead of starting with a theory (as in positivism) it collects data that either supports or refutes the theory**(Creswell, Creswell 2018)**.

Transformative approach promotes social justice and change. Researchers in this approach go beyond understanding the problem and address participants issues like discrimination, inequality and empowerment **(Mertens, 2017)**.

Another worldview comes from Pragmatism which doesn’t commit to one system of philosophy or reality. Researchers are free to multiple approaches to collect and analyze data rather than subscribing to one(qualitative or quantitative) **(Creswell, Creswell 2018).** This applies to mixed methods of research & supports ideas from both qualitative or quantitative methods and is more productive. It prioritizes research problems and applies whatever methods are most useful for addressing them. “Numbers and words can work together to produce richer and more insightful analyses of complex phenomena than could either one alone” **(Rossman & Wilson, 1985).**

The research philosophy adopted in this study is positivism because the nature of the research analysis is a comparison of Transformers models with classical models, through quantitative observable, measurable facts such as accuracy, training time, and efficiency metrics and not on personal beliefs, or emotions.

## **Research Approach**

According to **(Saunders, Lewis and Thornhill, 2009)** the two main research approaches are deduction and induction. These are logical reasoning strategies and guide about how research theory relates to data. With “Deduction” a theory and hypothesis are developed, and research strategy is designed (usually quantitative) to test the hypothesis. Key features include objectivity, control of variables, operationalism (facts to be measured quantitively), reductionism (breaking down of complex problems.) and generalization. While on the other hand, with induction, data is collected (e.g. survey, observations, interviews etc) to build a theory rather than testing an existing one. It is often associated with qualitative studies. Unlike deduction, it allows flexibility, encourages alternative explanations and generally suits smaller samples and develops deeper insights.

The approach chosen for this DBS Chatbot study is deductive. In this approach, it begins with a theory or hypothesis and tests it through data analysis. In this case established theory is that transformers-based architecture has higher efficiency than traditional models like LSTM. The research then tests it by applying both models to the same dataset (DSB specific) to compare the performance metrics. Then the hypothesis will be rejected or accepted based on the tests making it a deductive approach. This approach is particularly appropriate in this study as there is already established literature providing evidence that Transformers outperforms traditional models and the same can be tested in the context of DBS chatbot. Deductive research approach applied to DBS specific chatbot study can be visualized as

`

Figure 9 Five stages of deductive research(Robson,2002)

## **Research Design**

According to **(Robson,2002)**, Research design can be considered as a process or general plan that turns research question into research project. It consists of three layers, research strategies, research choices and time horizons**(Saunders, Lewis and Thornhill, 2009)**.

The different research strategies are experiments, survey, case study, action research, grounded theory, ethnography & archival research. Time-horizons answers the question of whether research a snapshot taken at a particular time (cross-sectional) or a series of snapshots and representation of events over a period of time(longitudinal) **(Saunders, Lewis and Thornhill, 2009)**.

In this research the best fitting design is Experimental study as it involves study of effect of change in performance of models in terms of accuracy, speed, etc. Furthermore, the research follows a cross-sectional design, meaning that the data is collected at a single point in time, or during a specific period, for each evaluation of dataset. The design is appropriate as study evaluates the performance of both the Transformer and classical models at a particular stage and no longitudinal analysis over time is required.

## **Research Method**

Research approaches in a methodological sense (methodological approaches) are categorized into Qualitative, quantitative and mixed methods. Quantitative approaches can be viewed as examination of variables which can be measured typically on instruments and numerical data that can be analyzed statistically. While qualitative research is an approach where, rather than numbers, it uses non numerical forms of data like words, observations, and explores understanding or meaning from how individuals or groups interpret a social or human issue. Mixed methods leverage the strengths of both qualitative and quantitative methods by integrating numerical data and comprehensive understanding of the data**(Creswell, Creswell 2018).**

This study adopts a mono-method (single data collection technique) quantitative approach focusing exclusively on numerical comparisons and statistical analysis. It does not incorporate qualitative methods such as interviews, surveys or subjective analysis. It involves training and testing different models (Transformer and LSTM) on the same DBS-specific dataset and focuses purely on comparing the performance metrics of the two comparisons. As this is an applied research project, the primary aim is to directly compare how the models perform in a practical, real-world environment rather than exploring theoretical concepts. Therefore, the research method is quantitative, as it seeks to derive insights from data that can be generalized and quantified.

There are various types of research in existence. Some of them mentioned by **(Rossman & Wilson, 1985)** Applied(Real world) research and Basic(Academic) research. The applied research focuses on solving practical problems. It aims to understand lived experiences and generate actionable insights relevant to daily life and practice. It is relatively small-scale research carried out by individuals or small teams. In contracts, basic research is mostly concerned with developing and extending theoretical knowledge. Such research is largely undertaken in universities. Such research is undertaken substantially in universities and largely as the result of an academic agenda**(Saunders, Lewis and Thornhill, 2009)**.

This study is an applied research project that will adopt a positive research philosophy in a deductive manner, testing an established theory that Transformer models outperform LSTM in efficiency and accuracy. The research follows a cross-sectional design and uses a mono-method quantitative approach to compare model metrics such as accuracy, training time, and comprehension rate. A functional chatbot artefact will be developed using Python and tools like Google Colab and HuggingFace, with the aim of improving student support and offering scalable solutions for academic institutions

# **Data Collection and Analysis**

There are two main types of data. Primary and secondary. Primary data is collected directly by researchers for their studies addressing specific research objectives. On the other hand, Secondary data is already collected data for some other purpose and can be used in research purpose like documentary data, survey data etc. **(Saunders, Lewis and Thornhill, 2009)**.

For this research, documentary secondary data is used as it is collected from DBS brochures, documents and DBS web pages.

A diagram of a survey

AI-generated content may be incorrect.

Figure 10 Types of secondary data (Saunders, Lewis and Thornhill, 2009).

The data collection and analysis pipeline for this research look like below.

Data is gathered from various sources, including DBS web site pages, brochures and websites. The brochures can have different formats(words, pdf), therefore the text extracted from them would require parsing, before it can further be utilized in the data processing pipeline. For example, documents might have page layouts, table structures, and images, which might generate data which is irrelevant in further processing. In order to simplify this process, GitHub [Docling](https://github.com/docling-project/docling) library will be utilized, which first converts the documents in different formats into a Docling’s unified fundamental document representation. Then Docling’s serializer will then serialize the content to the desired simplified text format. Steps involved in the process are outlined below. DBS Postgraduate program brochure - [Dbs-postgraduate-programmes-pdf](https://github.com/sandeepkumar-84/DBS/blob/7386b46200e03ebf3818d9855cdf21eebcdd6564/1-Dbs-postgraduate-programmes.pdf) , a 96 page 11.6 MB file obtained from the DBS website. Below screen shot shows layout of a page from the brochure.

A screenshot of a website

AI-generated content may be incorrect.

Below image shows an excerpt from above pdf page converted into the text form using docling library.

A screenshot of a computer

AI-generated content may be incorrect.

A white text on a blue background

AI-generated content may be incorrect.

Figure 11 pdf into text conversion through docling

Web Scraping is a technique which scrapes the data from web pages and can be used to build a corpus. There are various libraries available including BeautifulSoup, which is utilized in this study to scrape the DBS web site pages into simple text data. Below is the screen shot of the data obtained from the <https://www.dbs.ie/about-dbs/contact-us> link.

A close up of text

AI-generated content may be incorrect.

Data collected from above sources will form a corpus (all the words and sentences of the content) for the study. This corpus will be fed further into the data processing pipeline to ultimately train the models. Intent JSON files will also be created specifically for DBS data which will be used to train and test the chatbot. Intent files typically contain

1. Patterns - example user inputs for each intent
2. Responses- Predefined responses that chatbot provides for the intent.
3. Tags - Identifier (Unique) for each intent.

The Corpus will be loaded into the data frame and its shape and information are checked. The rest of the steps are discussed below.

1. **Data Cleaning** – Removal of special characters, bad words & contractions from words.
2. **Stemming** – The process of transforming words into their root form by eliminating prefixes and suffixes is called as Stemming of the words. In this study is Porter Stemmer algorithm developed by Martin Porter in the 1980s will be utilized for stemming to enhance the effectiveness of Natural Language Processing.
3. **Tokenization** – The process of breaking down of text into smaller units like words or phrases into non sensitive tokens. For machines analyzing tokens is a much easier task than words themselves. This reduces overhead of processing complexity. Another purpose is to prevent sensitive data like social security numbers or account numbers getting exposed to and prevent data breach.
4. **Vectorization** – Tokenized data is changed to numerical forms or vectors to be processed by machine learning algorithms. One of the techniques utilized in this process would be Word2vec which helps learn word embeddings from large datasets.
5. **Label and dictionary creation** – based on distinct labels or intents in the file dictionaries will be developed which will be utilized by the models in conversion from ids to labels and vice versa.
6. **Visualizations** – Different visualizations will be plotted to get insights from the distribution and frequencies of the words, patterns, length and rates of the appearance of the words and combinations of the words in conversations.
7. **Model Initialization & Training** – Models used in the study will be initialized and trained with the intent data where intent will be the target variable, and its prediction will generate the required response. Pattern or the user input will be the independent variable in the dataset. Thirty percent of the conversations will be considered for test and rest as training data set.
8. **Model Evaluation** – Models will be evaluated to measure the effect of Transformer vs LSTM models on chatbot performance (accuracy, perplexity, fluency, latency, etc.) .
9. **Draw Conclusions –** Interpretation of results.

**Models used in the study are**

**Transformer models** – Generation of a new representation of a sequence by connecting different positions within same sequence is called as intra attention or self-attention. The type of neural network architecture that processes sequential data, particularly text, using self-attention mechanisms are called Transformer models. Unlike RNNs (Recurrent Neural Networks) which process data sequentially, transformers can process entire sequences in parallel, making them efficient, effective and faster.

**BERT** is a Bidirectional transformer pretrained on unlabeled text to predict masked tokens in a sentence and to predict whether one sentence follows another. It reads text in both directions and thus learns the meaning of words in context from both directions. The BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications **(Devlin et al., 2018).** In this research the BERT model will be fine-tuned on the DBS-specific data for developing a chatbot capable of understanding and responding to queries relevant to Dublin Business School. Fine-tuning will enable the model to align with the institution-specific terminology and context, thereby improving the chatbot’s accuracy and relevance in handling user queries.

**LSTM** – It is a recurrent network architecture which is designed by Hochreiter & Schmidhuber to overcome the error back flows problems. It can learn to bridge time intervals in case of noisy input sequences without loss of short time lag capabilities achieved by efficient gradient based algorithm for an architecture enforcing constant error flow (neither exploding nor vanishing gradients). LSTM has the capability of capturing long term dependencies in sequential data making them ideal for tasks like speech recognition, language translation and time series forecasting**(Hochreiter & Schmidhuber, 1997).**

**GRU -** Gated Recurrent Unit - is a type of recurrent neural network (RNN) architecture that is like LSTM (Long Short-Term Memory), however it has simpler architecture and fewer parameters making it easy to train and computationally faster. At each step, the GRU creates a "candidate activation vector" using the current input and the previous hidden state. This vector helps decide how to update the hidden state for the next step**(Chung et al., 2014).**

# **Artefact Design and Presentation**

1. User friendly Chatbot interface developed in TK Interface library capable of
   1. Accepting both text and voice-based inputs, displaying real time response.
   2. Features like scrolling, buttons to send and stop conversation etc.
   3. Displaying real time responses.
2. Tools
   1. Visual Studio\Code for development of AI and python code.
   2. Python environment.
   3. Libraries including, but not limited to, NLTK, Keras, PyTorch , seaborn, wordcloud, Transformers.
   4. HuggingFace for deployment
   5. Google Speech-to-Text for voice input processing
3. Presentation
   1. Documentations like Report and Presentations.
   2. Visualizations
   3. Graphs
   4. Real time demonstration.

# **Timeline**

|  |  |  |
| --- | --- | --- |
| Phase | Activities | Duration |
| Literature Review | In-depth review of existing approaches | 0.5 month |
| Data Preparation and Analysis | Dataset collection and preprocessing | 0.5 month |
| Model Development & Deployment | Implementing machine learning models | 1.5 months |
| Model Evaluation | Performance testing and comparison | 0.5 month |
| Documentation | Writing and presenting findings | 0.5 month |

# **Conclusion**

This study attempts to demonstrate the potential advantages of transformer-based architecture over traditional LSTM models in developing a DBS specific chatbot that supports both text and voice-based interactions. While focus remains on quantitative evaluation, future studies can explore qualitative aspects like emotions and user experience to further boost the performance and relevance of the chatbot conversation. The study can be extended to ensure it is highly scalable and relevance to schools and institutions beyond DBS, aiming to enrich student engagement, support services, and automate query resolution using cutting-edge AI technologies.

# **References**

1. Turing, A.M. (1950) ‘Computing Machinery and Intelligence’, Mind, 59(236), pp. 433–460. Available at: <http://www.jstor.org/stable/2251299>
2. Weizenbaum, J., 1966. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), pp.36-45.
3. Colby, K., 1975. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. New York: Elsevier.
4. Shum, H.Y., He, X.D. and Li, D., 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*, pp.10-26.
5. Wallace, R.S., 2009. *The Elements of AIML Style*. ALICE A.I. Foundation. Available at: <http://www.alicebot.org>
6. Elov, B.B., Khamroeva, S.M. and Xusainova, Z.Y., 2023. The pipeline processing of NLP. In *E3S Web of Conferences* (Vol. 413, p. 03011). EDP Sciences.
7. Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
8. Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
9. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), pp.391-407.
10. Žukauskas, P., Vveinhardt, J. and Andriukaitienė, R., 2018. Philosophy and paradigm of scientific research. Management culture and corporate social responsibility, 121(13), pp.506-518.
11. Saunders, M., Lewis, P. and Thornhill, A., 2009. Research methods for business students. Pearson education.
12. Mertens, D.M., 2017. Transformative research: Personal and societal. International Journal for Transformative Research, 4(1), pp.18-24.
13. Rossman, G.B. and Wilson, B.L., 1994. Numbers and words revisited: Being “shamelessly eclectic”. Quality and quantity, 28(3), pp.315-327.
14. Behl, V. and Bibhu, V. (2024) 'Leveraging Transformer Networks for Enhanced Conversational AI', in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, Greater Noida, India, pp. 937-943. doi: 10.1109/IC3I61595.2024.10829231. Available at: <https://ieeexplore.ieee.org/abstract/document/10829231>
15. Vaswani, A. et al., 2017. Attention is all you need. In Advances in Neural Information Processing Systems. Curran Associates, Inc. [online] Available at: <https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
16. Creswell, J.W. and Creswell, J.D., 2018. *Research design: Qualitative, quantitative, and mixed methods approaches*. 5th ed. Thousand Oaks, CA: Sage Publications.
17. Robson, C., 2002. Real world research.
18. Lin, J. and Michailidis, G. (2024) Deep Learning-based Approaches for State Space Models: A Selective Review. arXiv. Available at: <https://doi.org/10.48550/arXiv.2412.11211>
19. Schuurmans, D., 2023. *Memory Augmented Large Language Models are Computationally Universal*. arXiv. Available at: <https://doi.org/10.48550/arXiv.2301.04589>
20. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805. Available at: <https://doi.org/10.48550/arXiv.1810.04805>
21. Hochreiter, S. & Schmidhuber, J., 1997. *Long short-term memory*. Neural Computation, 9(8), pp.1735–1780. doi:10.1162/neco.1997.9.8.1735.
22. Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555. Available at: <https://doi.org/10.48550/arXiv.1412.3555>
23. Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 12, 2025. <https://web.stanford.edu/~jurafsky/slp3>.
24. Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue, V., 1995. Multilingual spoken-language understanding in the MIT Voyager system. Speech Communication, 17, pp.1-18.
25. A. Freed, Conversational AI: Chatbots that work, Manning, October 2021.
26. Patil, K., Patil, R., Koyande, V., Thakur, A.S. and Kadam, K., 2024, October. Analyzing Chatbot Architectures Utilising Deep Neural Networks. In 2024 IEEE 6th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA) (pp. 15-19). IEEE.
27. Tapia-Hoyos, J. J (2024). Chatbots in the service of university students: A review. Social Sciences in Brief, 1, 1-7. <https://doi.org/10.47909/ssb.08>.
28. Nguyen, T.T., Le, A.D., Hoang, H.T. and Nguyen, T., 2021. NEU-chatbot: Chatbot for admission of National Economics University. *Computers and Education: Artificial Intelligence*, *2*, p.100036.
29. Modiba, Mashilo & Shekgola, Mahlatse. (2024). Utilising Artificial Intelligence Chatbots for Student Support at Comprehensive Open Distance E-learning Higher Learning Institutions in the Fifth Industrial Revolution. Journal of Education, Society & Multiculturalism. 5. 26-48. 10.2478/jesm-2024-0003.