Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 3, 2025

ENHANCING DATABASE INTELLIGENCE: NATURAL LANGUAGE PROCESSING FOR ADVANCED QUERY OPTIMIZATION

Muneeb Ali Muzaffar[1], Muhammad Zulkifl Hasan[2], Muhammad Zunnurain Hussain[*3]

[1]*Department of Computer Science, University of Central Punjab Lahore, Pakistan.*
[2]*Department of Computer Science, University of Central Punjab Lahore, Pakistan.*
[*3]*Department of Computer Science, Bahria University Lahore, Pakistan.*

[1]muneeb.muzaffar@ucp.edu.pk, [2]zulkifl.hasan@ucp.edu.pk,
[*3]zunnurain.bulc@bahria.edu.pk

DOI: https://doi.org/10.5281/zenodo.15542440

Keywords
NLP ,NLI-DBQ User Interfaces,NLI-DBQ, Database Interaction,Accessibility.

Article History
Received on 20 January 2025
Accepted on 20 February 2025
Published on 27 February 2025

Corresponding Author: * Muhammad Zunnurain Hussain
zunnurain.bulc@bahria.edu.pk

Abstract
This document plans to improve database querying via natural language processing (NLP). Here NLP technique is a step of querying database using natural language (NLIDBQ). It is a solution to the problems, that users who are not technical persons face with the traditional methods like SQL because of the intricacy of big data, and also technological changes. The paper suggests three approaches that use hybrid NLP and deep learning, federated learning for privacy, AI which help to conduct the dynamic query optimization and the predictive analysis for the query formulation. The outcomes indicate the systems to be more usable with better user accessibility that makes them more user friendly for the general public across a larger range. This results in the emergence of equality data access with further activity in the area of specific adaptation, and realworld showings trial.

INTRODUCTION

Intrusion Detection Systems (IDS) serve as a powerful tool that defends computer systems and networks against illicit user and hacker occurrences within the cybersecurity field. These security systems were designed to recognize several types of threats while taking measures for information system protection. The primary IDS classifications include Agreed network-based IDS (NIDS) as well as Host based IDS (HIDS). The article discusses the weaknesses of two preferred IDS solutions OSSEC and Snort despite their specific superior performance attributes. Surface detection success by skilled systems exists despite numerous erroneous positive results and limited adaptability when cybersecurity threats evolve. The application of machine learning constitutes an innovative solution to build more effective IDS systems that address current security problems.

The study investigates traditional IDS vulnerabilities through machine learning algorithm application on NSL-KDD dataset information. Research in this domain leverages NSL-KDD dataset as an upgraded version of

KDDCup99 dataset because its expanded information base enables more equitable assessment. Our project builds an AI-driven intrusion detection system of the next generation with SVC and Naive Bayes and Decision Trees and Logistic Regression as our most recent machine learning models to achieve superior accuracy while lowering false positives and enhancing adaptability.

2. Need and significance

The natural language interface represents the ideal way to connect users to databases through effective data retrieval functions without requiring IT expertise [5]. Organizations that want to remain competitive and innovative need to develop skills in deriving valuable insights from their rapidly growing databases.

One of the most important technologies in today's computing era is certainly the No SQL database [6]. Among other things the weakness of big data stimulated this type of innovative approach to databases management [7]. Like the traditional query language such as SQL the users with only technical backgrounds are required to use specialized training and expertise which narrows down the usage to a small group of users. Big data kept in enterprise databases is meaningless if you don't have access to specialist tools for querying or finding data[8]. Thus, it is widely agreed that advance intuitive and inclusive models of database querying are needed which could be understandable and used by both technical and non-technical parties.

This research matters because it has the potential to democratize data access and thereby create a space where people with different needs and experiences collaborate together to solve problems that are drawn from various fields of expertise. On the pursuance of the NLI-DBQ systems development, implementation, and impact study, this research aims at putting forth their transformative capabilities on productivity, decision-making and innovation within an organization.

3. Research Question

- Objective: Analyze strategies and techniques to improve NLI-DBQ systems for enhanced accessibility in database querying, catering to both technical and nontechnical users.
- Focus: Enhance responsiveness and interactivity of NLI-DBQ systems to fill the information needs of learners with different levels of technical expertise.
- Research Scope: Covering NLI-DBQ tools, methods, and principles aimed at enhancing system quality, performance, and user-friendliness while ensuring excellent accessibility.
- Approach: Explore NLI-DBQ tools, methods, and principles that can improve system quality, performance, and user-friendliness. Investigate strategies to enhance system responsiveness and interactivity for users with varying levels of technical passion.
- Outcomes: Enhanced NLI-DBQ system quality, performance, and user experience, leading to increased accessibility and usability for a broader range of users.

4. Previous Technologies and their Drawbacks

4.1 Speech Recognition for Data Querying

ASR,[9] which stands for Automatic Speech Recognition, is a technical solution meant to help process and interpret human speech into a machine-readable format. The field of ASR actually deals with a number of problems related to speech recognition which are ASR accuracy as well as various methods for speech processing, extracting features and assessing performance.The use of universal speech recognition systems in voice command recognition tasks can lead to unnecessary searches in large databases [10]. It is possible that they lack the optimization for the embedded structure of SQL which can lead to errors when users recognize the structure of SQL queries.

4.2 Query Interfaces with Touch Interactions

Touch-enabled query interfaces serve as user-friendly interfaces which allow users to execute database queries through screen-based touch gestures on touch screens. The authors [11] investigate how touch screen technology enables users to build queries using field value alphabets. Users have three distinct interfaces to input data beyond set limits: QWERTY keyboards together with Alphabetic keyboards

and Reduced Input Data Entry (RIDE) interfaces serve as the selection methods. The increasing popularity of mobile phones has led to specific user interface problems requiring further investigation about device orientation fluctuations while using small interfaces combined with different user interactions such as tapping and flipping and pinching. Learning disability patients might experience discomfort because of this design [12]. Intelligent touch interfaces to query databases prove complex to develop because researchers must evaluate database structure and establish smooth usage while setting rules for user interaction.

4.3 Conversational Assistants

Virtual robots serve users to accomplish such a job like finding planes, booking restaurants and navigating user interfaces, by providing natural language interface to the web services and APIs.Rise of standard conversational interfaces, and ease of frameworks like Google Actions and Alexa create scope for every developer to support new services [13]. However, their algorithms which were meant for limited predefined knowledge base and databases are not able to query.These assistants need lots of personal data in order to understand user preferences and give user-specific responses, which is definitely a thing that might endanger privacy [14]. The probability of the disk storage being accessed by the third side cannot be excluded in such a cloudbased situation.

They have high potential for optimization with sophisticated query as well as high demands for users with background information when processing reply results.

5. Research Solutions

5.1. Hybrid NLP Techniques with Deep Learning

The vulnerability of SQL injection attacks for web Neural network models including recurrent neural networks and attention mechanisms now perform complicated tasks such as machine translation and syntactic parsing and summarizing tasks after integrating general neural network models. [15] Combining NLP traditional approaches with deep learning Transformer models including BERT [16] and GPT [17] improves the understanding of natural language queries. Users benefit from such models when domain-specific datasets help them understand queries in multiple contexts.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [18]$$

Q represents the query terminology while K stands for keys terminology and V represents the value terminology in this context together with the key dimension designated as dk.

5.2. Federated Learning for Privacy-Respecting Model Improvement

Federated learning is a technique of teaching model on multiple devices where the participant devices are all federated, distributed and decentralized. Device data is kept private by performing computations locally. Additionally, during testing a static version of the global model is used which is not updated with changes, however this might lead to expandability issues with large models[19].The federated learning concept is an advanced tool for privacy protection in machine learning. It is the central server that serves as the aggregator of multiple participants that incorporate parameters into a model, distribute the generated model to the client, and then converge to optimize the global model. The model obtained dealing with performance near centralized data training is trained in a situation where the leaving data is not involved locally [20]. These local improvements can be aggregated across many users without sharing raw data, maintaining privacy while enhancing the NLI-DBQ system.

$$\Theta_{global} = \frac{1}{N}\sum_{i=1}^{N}\theta_i \quad [21]$$

Here, Θglobal is the updated global model parameters after aggregation.

5.3. Dynamic Query Optimization using AI

AI can enhance database performance (AI4DB) through learning-based optimization techniques [22]. Use of AI to dynamically optimize queries based on real-time database load, query complexity, and historical performance data. This could lead to faster query execution times and a more efficient database system.

An important part of multi-database technique is query optimization. However, the objective of the optimizer is to know the minimum join order that would be useful to finish the query[23]. An Ant colony algorithm is a bionic algorithm which imitates the movement of ants. Within ant families, when ants have to find food, the pheromone sensing can help in finding the quickest route [24].In distributed database systems, it gets exploited during the join query optimisation process.

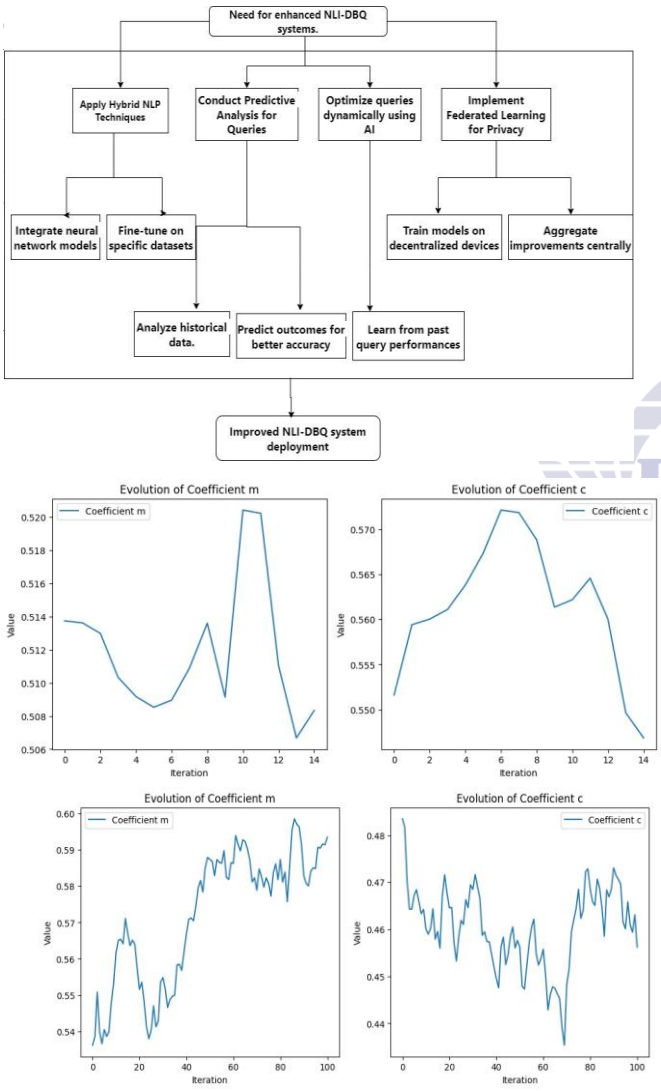

Fig. 3. M and C evolution (Level 2)

Whereas the original ant colony-based query optimization approach, operates with fixed costs. The expected model uses variable prices and calculates the costs of execution plans as they are being generated. With this approach, the algorithm will try to find joining sequence in order to minimize total running time. [25].

$$\tau_{ij}(t+1) = (1-\rho)\cdot\tau_{ij}(t)+\Delta\tau_{ij} \quad [26]$$

In the above mentioned context τij represents the pheromone level on the path from point i to j, ρ is the rate of evaporation, and Δτij is the amount of pheromone deposition, typically a function of the inverse of the path cost or length.

Predictive Analysis can be used to access developed data set histories to make forecast about the effects of multiple scenarios. Both continuous and discontinuous changes enter our simulation model for prediction.[27] For example, the likelihood of one event or another event occurring before the other, or the probability of the event within the next K number of steps of the sequence is to be determined . These forward-looking questions find settings in diverse areas, including predictive when a sentence will end or when a user will switch between apps [28].

We frequently compute the likelihood of a specific event E taking place at time t given a sequence of previous occurrences S, which can be modelled using conditional probabilities. For the purpose of formulating predictive queries utilising forecasting techniques, we continually find the probability of a particular event E occurring at time t given the temporal sequence of past similar occurrences S. This being the basis of forecasts through conditional probabilities:

$$P(E_t|S) = \frac{P(E_t \cap S)}{P(S)} \quad [29]$$

This formula is fundamental given because it does the job of predicting the probability of specific events following a given sequence, using the history of events as expressed by S.

6. Results

6.1. Evolution of Coefficients m and c

The key takeaways from Figure 1,Figure 2, and Figure 3

are as follows: Qualitative Assessment: Knowing m's and c's evolution during training can tell us a lot about the learning process. Are there sudden jumps and oscillations that might point to an unstable learning curve? Do we observe such a regularity that it indicates gradual convergence?

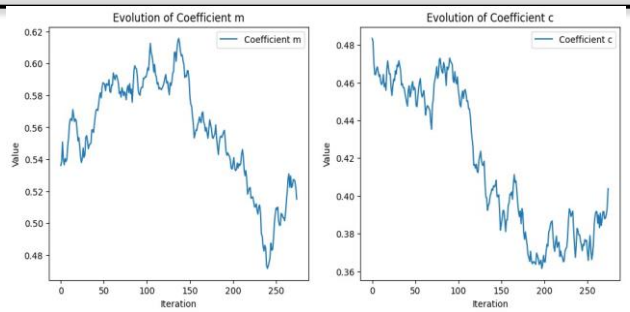

Fig. 4. M and C evolution (Level 3)

Quantitative Aspects: Moreover, from the visual presentation of convergence we can calculate the metrics of convergence like the rate of change of coefficients or the variance of values. These metrics offer a clearer picture on how fast the model is converging and if adjustments are needed.

Here the use of two graphs representing the changing of coefficients 'm' and 'c' provides an interesting and clear visualization of how the federated learning developments take place. They help us to understand that data updates of local manager without keeping individual person's data secret.

By presenting plots comparing the global model's parameter alterations through time we touch on the distributed learning process and the collective insight of all the clients' data.

Implications:

The conduct of federated learning within NLI-DBQ systems is such that privacy of information is enshrined in the learning procedure. This is most very significant in a period when the privacy of data hugely concerns people. 6.2. Best Path Cost Over Iterations

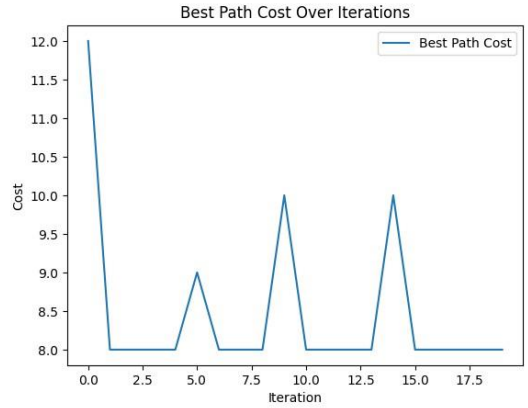

Fig. 5. Level 1

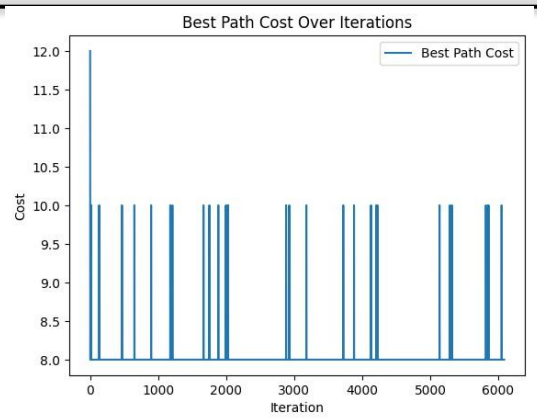

Fig. 6. Level 2

Qualitative Assessment: Besides just noticing the cost curve heading downward, we can conduct the modeling on this whence it actually comes from, the particular trends, and the package of behavior that these contain. How often and where there are those drastic drops or might be some unusual dynamics during optimization process being? Are there occasions where the cost incurred is larger than the anticipated amount, point to an unexpected issue that may be with the optimization?

Quantitative Aspects: As illustrated in Figure 4 and Figure 5 tracking the parameters such as the velocity of the value of cost or the convergence towards the minimum values is possible afterward. The metrics provide qualitative measure of the optimization efficiency, thus used in assessment of full-scale performance of the Ant Colony Optimizer.

6.3. Query Embedding Visualization

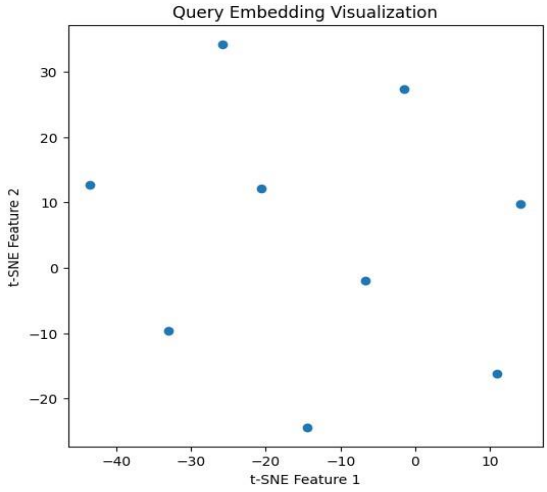

Fig. 7. Level 1

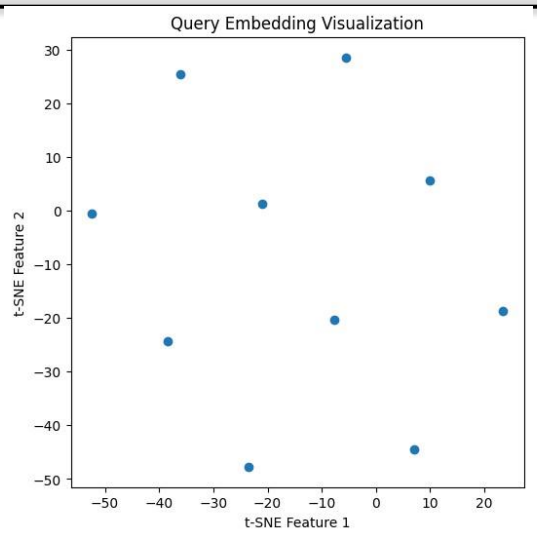

Fig. 8. Level 2

Figure 6, Figure 7 displays the map of query embeddings traversing a 2d space using the t-SNE method.t-Distributed Stochastic Neighbor Embedding (t-SNE) for visualising multidimensional data has been one of the practical and used methods. Moreover, this approach helps generate successful applications in a variety of domains[30].

Qualitative Assessment: Clusters in the plot may indicate the queries that have the same meaning falling into the same semantic category that makes the model suitability for achieving semantic similarity.

Quantitative Aspects: Even though a single plot can't exhibit the quantitative standard of quality, it is actually the density and location of clusters that can be indicative/expressive. Clusters being etched on top of each other, this may hint at the necessity for a greater level of model modulation.

For instance, the BERT model's embeddings suggest that semantic relations affected by higher-level phenomena can be represented through t-SNE plots. Expansion and the cluster chart of queries help to quantitatively assess the model reflecting the fact that whether it can distinguish between different queries well or not.

Implications: Human users, especially non-technical ones, are now able to use natural language (NL) interface with databases, which can understand their queries and improve the performance of the whole database systems. Therefore, more people even without technical expertise can get access to data.

7. Theoretical Prospects

The theoretical framework of machine learning model application in NLI-DBQ has been shown as a valid option by the observations. This outcome shows not only the feasibility, but the feasibility of the methods used also. The ease with which the models can be fine-tuned to be trained on domain specific datasets becomes even more crucial as it further indicates that the system can be tailored for specific industries or data sets based on the nature of the problem.

The practice of translating theory into deeds may bridge the gap between the technical and non-technical stakeholders, thus creating data democracy at large.

8. Conclusion

The study ends around the successful integration of the latest Natural Language Processing (NLP) algorithms and machine learning methods with Natural Language Interface to Database Querying (NLI-DBQ) systems.

Through the research the systems of NLI-DBQ that are more easy to use and accessible have been developed. It is important as it enables people with varying levels of technical ability to communicate with complicated databases by using natural language. The improved usability is specially useful for non-technical users who might find it challenging to do SQL, which is a traditional query language.

The incorporation of the complex computational algorithms has not only increased the effectiveness of the systems but also highly enhanced the accuracy and efficiency of database querying. This implies that the systems can work out more efficiently. They can execute complex inquiries, optimize query processing, and provide more precise and relevant outcomes.

The research has innovatively integrated a variety of modern NLP methods including deep learning and hybrid NLP techniques, as well as machine learning methods of the federated learning type. The addition of this feature to NLI-DBQ systems has given them the ability to perceive and comprehend user queries to a larger extent and with better precision.

9. Theoretical Aspects

The basis for this study in theory is the use of machine learning models, especially deep learning and federated learning, so as to improve NLI-DBQ system. Its proved validity has been attested by the practical results from the case, which contains not only the feasibility but also the practicality to different industry-specific applications. Research creates such a connection that all people are able to use technical data which results in a new data democracy.

10. Results match up with the research goals

The findings have proven that NLI-DBQ has made things much more accessible and easier to use for different kinds of people. This is exactly in line with the target of database querying systems amelioration with integrated NLP and machine leaning technologies. Mind mixed NLP techniques and federated learning for privacy, dynamic query optimization and predictive analysis for query formulation all will be a step forward to this goal.

11. Evidences

The research question was formulated with two main aims: to increase the availability and ease of use of NLIDBQ systems for technical as well as non-technical users. All the simulations and methods [like the use of Transformers], the adoption of [federated learning] and AI- based query optimization are the clear answers to this question as they demonstrate superior output and usability of the NLIDBQ techniques.

12. Assessment of Study's Benefits

The study of improvement usability and query processing efficiency with maintaining user privacy as a criterion of the evaluation can be rated based on its results. This study can be considered a success because the results and methodologies meet the set criteria, showing that the research has been conducted well.

13. Recommendations

- Further Integration of AI and ML: Continue exploring the integration of advanced AI and ML techniques to further enhance the NLI-DBQ systems.
- User-Centric Design: Emphasize a user-centric design approach to make the system more intuitive for non-technical users.
- Privacy and Security: Maintain a strong focus on privacy and security, especially when dealing with sensitive user data.

14. Future Work

- Domain-Specific Customization: Study the feasibility of dedicated NLI-DBQ systems for limited domain use that would raise their apply area.
- Real-World Implementation: Test the applicability of these systems in real-life situations in various fields; such test involves studying their practical effectiveness.
- Enhancing User Interaction: The intelligent system research aims at improving human computer interaction, making the system more responsive and personable to users inquiry.

REFERENCES

A. Sheik Abdullah and Preethi Priyadharshini. Big data and analytics. Big Data Analytics for Sustainable Computing, 2020.

Geeta Rani, Taruna Sharma, and Avinash Sharma. Future database technologies for big data analytics. 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), pages 349–354, 2023.

Daphne Miedema, G. Fletcher, and Efthimia Aivaloglou. So many brackets! an analysis of how sql learners (mis)manage complexity during query formulation. 2022 IEEE/ACM 30th International Conference on Program Comprehension (ICPC), pages 122–132, 2022.

Abdul Quamar, Vasilis Efthymiou, Chuan Lei, and Fatma Ozcan.¨ Natural language interfaces to data. ArXiv, abs/2212.13074, 2022.

Ayush Kumar, Parth Nagarkar, Prabhav Nalhe, and Sanjeev Vijayakumar. Deep learning driven natural languages text to sql query conversion: A survey. ArXiv, abs/2208.04415, 2022.

Ziqi Li. Nosql databases. 2019.

M. A. Elsabagh. No sql database: Graph database. Egyptian Journal of Artificial Intelligence, 2022.

Dmytro Orlovskyi, Andrii Kopp, and Ivan Bilous. An approach to development of interactive adaptive software tool to support data analysis activity. In International Workshop on Computer Modeling and Intelligent Systems, 2021.

Shipra J Arora and Rishi Pal Singh. Automatic speech recognition: a review. International Journal of Computer Applications, 60(9), 2012.

Snezhana Georgieva Pleshkova, Aleksander Bogdanov Bekyarski, and Zahari Todorov Zahariev. Reduced database for voice commands recognition using cloud technologies, artificial intelligence and deep learning. 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), pages 1–4, 2019.

Andrew Sears, Yoram Kochavy, and Ben Shneiderman. Touchscreen field specification for public access database queries: let your fingers do the walking. In Proceedings of the 1990 ACM annual conference on Cooperation, pages 1– 7, 1990.

Peter Williams and Sidharth Shekhar. People with learning disabilities and smartphones: Testing the usability of a touch-screen interface. Education Sciences, 2019.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 8689–8696, 2020.

Kambiz Saffarizadeh, Maheshwar Boodraj, and Tawfiq Alashoor. Conversational assistants: Investigating privacy concerns, trust, and self-disclosure. 12 2017.

Yoshimasa Tsuruoka. Deep learning and natural language processing. Brain Nerve, 71(1):45–55, 2019.

Sulaiman Aftan and Habib Shah. A survey on bert and its applications. In 2023 20th Learning and Technology Conference (L&T), pages 161–166. IEEE, 2023.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. ArXiv, abs/2001.01523, 2020.

Yan Hong, Zhiqing Huang, and Chenyang Zhang. A privacyenhanced federated learning model training method. In 3rd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2023), volume 12717, pages 718–724. SPIE, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communicationefficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017.

Xuanhe Zhou, Chengliang Chai, Guoliang Li, and Ji Sun. Database meets artificial intelligence: A survey. IEEE Transactions on Knowledge and Data Engineering, 34:1096– 1116, 2020.

Sayed A. Mohsin, Saad Mohamed Darwish, and Ahmed Younes. Qiaco: A quantum dynamic cost ant system for query optimization in distributed database. IEEE Access, 9:15833–15846, 2021.

Aoran Chen, Hao Tan, and Yiyue Zhu. Ant colony optimization algorithm and its application. In Other Conferences, 2022.

Sayed A Mohsin, Saad M Darwish, and Ahmed Younes. Dynamic cost ant colony algorithm for optimize distributed database query. In Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), pages 170–181. Springer, 2020.

Marco Dorigo. Ant colony optimization. Scholarpedia, 2(3):1461, 2007.

Abhishek Chakrabarty, Amod Bhosle, Chinmay Dhanawade, Latesh Billava, and Vivek Ramakrishnan. Predictive analysis implementation in real life scenarios. 2021.

Alex Boyd, Samuel Showalter, Stephan Mandt, and Padhraic Smyth. Predictive querying for autoregressive neural sequence models. Advances in Neural Information Processing Systems, 35:23751–23764, 2022.

Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5):662–679, 2012.

Angelos Chatzimparmpas, Rafael Messias Martins, and Andreas Kerren. t-visne: Interactive assessment and interpretation of t-sne projections. IEEE Transactions on Visualization and Computer Graphics, 26:2696–2714, 2020.