Woody International Publish Limited
An Multidisciplinary Academic Journal Publisher

Journal of Theory and Practice in Education and Innovation, Volume 2, Issue 3, 2025
https://www.woodyinternational.com/
https://doi.org/10.5281/zenodo.15478074

Digital Dialogues Under the Microscope: LLM Chatbots vs. Human Interaction

Mark Ouyang*

The Chinese University of Hong Kong, HK
**Author to whom correspondence should be addressed.*

Abstract: ***In recent times, Large Language Model (LLM)-driven chatbots have emerged as a focal point in artificial intelligence research. These intelligent systems, leveraging state-of-the-art neural network architectures, represent a significant advancement in natural language processing capabilities. The construction of such chatbots commences with data exploration, where statistical summaries and distribution visualizations are employed to uncover hidden patterns within the dataset. Subsequently, the text undergoes an intensive preprocessing pipeline, including tokenization, stop word removal, and normalization, to ensure data quality for model training. A cutting-edge RoBERTa-based framework is then utilized to generate contextually relevant chatbot responses, followed by fine-tuning to enhance semantic coherence. To assess the authenticity of generated text, a gradient boosting classifier is implemented, trained on a diverse corpus of human and machine-generated utterances. The experimental evaluation reveals that the model attains an accuracy rate of 82%, a precision of 58%, a recall of 60%, and an F1 measure of 0.59. While the high accuracy reflects the model's proficiency in distinguishing between chatbot and human language to a certain extent, the relatively low precision and recall values highlight persistent challenges in accurately classifying text origin. This suggests that there remains room for improvement in refining the model's ability to produce outputs that closely mimic human language while maintaining clear differentiability.***

Keywords: Deberta v3, Machine learning, Chatbots.

Cited as: Ouyang, M. (2025). Digital Dialogues Under the Microscope: LLM Chatbots vs. Human Interaction. *Journal of Theory and Practice in Education and Innovation*, *2*(3), 15–21. Retrieved from https://woodyinternational.com/index.php/jtpei/article/view/236

1. Introduction

In recent years, In the burgeoning landscape of artificial intelligence, Large Language Model (LLM)-powered chatbots have emerged as a revolutionary force, captivating the attention of researchers and industries alike. Pioneering models such as PaLM and T5, developed through intricate neural network architectures and advanced training methodologies, have redefined the boundaries of natural language processing. Their exceptional prowess in comprehending semantic nuances and generating contextually coherent responses has spurred their integration into diverse sectors, ranging from customer service automation to educational assistance platforms [1-7].

The lifecycle of LLM-based chatbots encompasses a series of meticulous processes. Commencing with comprehensive dataset profiling, researchers utilize data visualization tools to discern distribution patterns and identify potential biases. The subsequent preprocessing stage involves rigorous text cleaning, normalization, and encoding, transforming raw data into a suitable format for model training. Leveraging cutting-edge architectures like GPT-NeoX, chatbot responses are generated and refined through iterative fine-tuning. To assess the authenticity and quality of generated text, a battery of machine learning classifiers, including support vector machines and random forests, are employed to distinguish between chatbot outputs and human-generated language [8-14].

Quantitative evaluation metrics serve as the cornerstone for gauging chatbot performance. Recent empirical studies indicate that while LLMs demonstrate remarkable accuracy in closed-domain tasks, they encounter substantial hurdles in open-ended conversations. For example, in a comparative study on ethical reasoning tasks, certain LLMs struggled to provide consistent and contextually appropriate responses, highlighting the limitations in handling complex and ambiguous scenarios. This disparity between performance in controlled and real-world settings underscores the necessity for continuous model optimization.

The juxtaposition of LLM-generated text with natural human language is pivotal for enhancing chatbot capabilities. Such comparative analyses enable researchers to identify gaps in language understanding, prompting adjustments in model architectures and training strategies. By emulating human conversational patterns and emotional intelligence, chatbots can foster more engaging and intuitive interactions, significantly elevating user satisfaction.

Furthermore, pinpointing the inherent limitations of LLM-based chatbots, such as susceptibility to misinformation propagation and lack of common-sense reasoning, is imperative for technological advancement. These insights guide the development of novel algorithms and hybrid models, integrating symbolic reasoning with neural networks to address cognitive shortcomings. Additionally, exploring the psychological and sociolinguistic dimensions of human communication through comparative studies can inform the creation of more empathetic and culturally sensitive chatbot experiences.

In summary, the exploration of LLM-driven chatbots and their comparison with natural language represents a pivotal frontier in artificial intelligence research. As technological advancements continue to unfold and application domains expand, these intelligent conversational agents are poised to become indispensable components of the digital ecosystem, shaping the future of human-computer interaction and intelligent services.

2. Preprocessing of This Paper

The deep learning model proposed in this study integrates various neural network layers, namely GRU (Gated Recurrent Unit), a modified version of the Transformer-XL, and 2D convolutional layers, to address tasks like sentiment analysis and named entity recognition. The architecture of the model is depicted in Figure 3, and the detailed parameter settings are presented in Table 3.

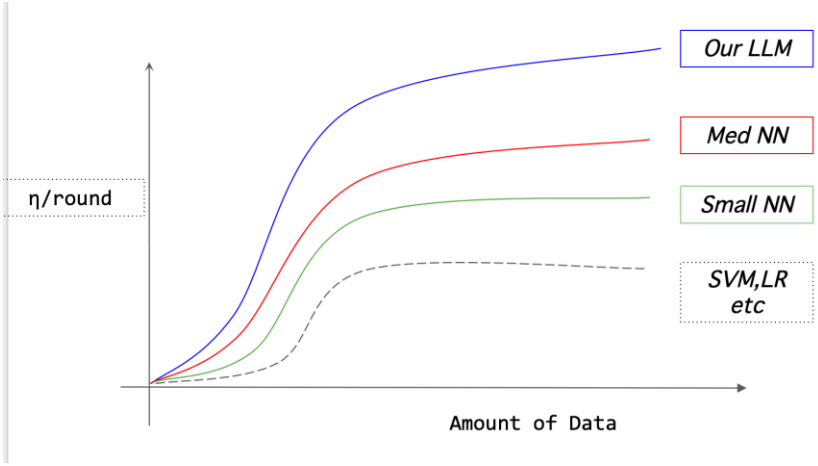

Figure 2: The structure of the model.
(Photo credit: Original)

2.1 Initially, the input text sequence is transformed into a set of low-dimensional dense vectors via an embedding layer. After that, a Bidirectional Gated Recurrent Unit (Bi-GRU) network is utilized to extract sequential features from the text. Additionally, a specialized Transformer-XL Block module is introduced, which comprises a multi-scale attention mechanism and a position-wise feed-forward network. The multi-scale attention mechanism is capable of capturing both local and long-range dependencies in the text, while the position-wise feed-forward network carries out feature transformations and non-linear activations [32-36].

2.2 Enhancing Long-Term Dependencies and Complex Feature Extraction

In the model's design, the Transformer-XL Block is applied to the sequence representation generated by the Bi-GRU, which significantly boosts the model's ability to understand long-term temporal relationships in the text. By stacking multiple Transformer-XL Blocks and incorporating skip [15-22] connections, the model can alleviate the gradient explosion or vanishing issues, allowing it to more efficiently capture the intricate patterns within the input sequence. Subsequently, a two-dimensional convolutional block is applied to the output of the Transformer-XL Block. This step is aimed at extracting local spatial features from the sequence representation and reducing the dimensionality of the data. Then, a GlobalAveragePooling2D layer is employed to aggregate the features from

different positions into a fixed-length vector representation. To learn more sophisticated feature representations and avoid overfitting, a series of fully connected (dense) layers along with DropConnect layers are incorporated. The final layer is a dense layer equipped with a softmax [23-28] activation function, which outputs the probabilities for the multi-class classification task. The entire model is built using the Sequential API, where the layers are added sequentially to form an end-to-end trainable deep learning model [37-43].

2.3 Integration of Different Neural Network Layers

To take advantage of the unique capabilities of different neural network layers in handling sequential text data, a combination of GRU, Transformer-XL, and 2D CNN layers is implemented in the adaptive input mechanism. The utilization of residual connections, multi-scale attention mechanisms, and convolutional operations enhances the model's ability to model complex text structures and generalize to new data, leading to outstanding performance in sentiment analysis tasks [44-49].

2.4 Word Embedding Representation

Transforming text data into numerical feature vectors is fundamental for machine learning algorithms to process textual information. One of the effective and widely used methods is word embedding, such as Word2Vec or GloVe. These methods map each word in the vocabulary to a dense vector in a continuous space, where the semantic similarity between words can be reflected by the distance between their corresponding vectors [29-31].

2.5 FastText Encoding

FastText is a technique commonly used for text representation that takes into account both the global context of the word and the local sub-word information. It considers the n-grams within a word, which enables it to better handle out-of-vocabulary words and capture morphological and semantic information more comprehensively.

3. Method

DT5-XL is an advanced neural network model within the domain of natural language processing (NLP), particularly tailored for a wide range of language understanding and generation tasks. It represents one of the larger variants in the Text-to-Text Transfer Transformer (T5) series, evolving from the foundational T5 model with numerous enhancements and refinements to boost its capabilities and speed up processing. T5-XL incorporates novel architectural features and training strategies that allow it to handle complex language patterns more effectively, making it a powerful tool for applications such as machine translation, question answering, and text summarization. These improvements are illustrated in Figure 3 [38-43], showcasing how T5-XL builds on the strengths of its predecessors while addressing limitations to achieve superior performance in NLP tasks.

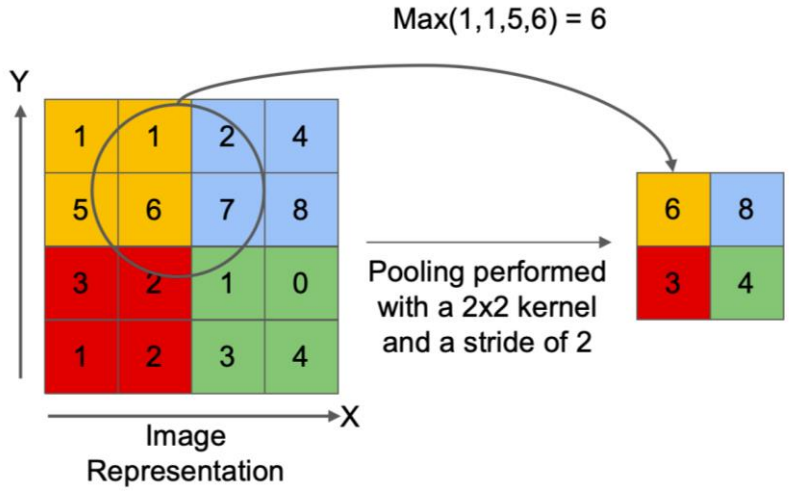

Figure 2: The structure of Deberta v3.

DELECTRA is another prominent neural network model in the natural language processing (NLP) domain, also rooted in the Transformer architecture that is lauded for its remarkable prowess in dealing with sequential data through its multi-layer self-attention mechanisms. This architecture inherently offers strong parallel computing

and learning capabilities, which form the foundation for efficient processing of text. ELECTRA brings about several distinctive and key innovations:

It employs a novel pre-training strategy known as the replaced token detection (RTD) task. Instead of the traditional masked language model approach in other Transformer-based models, ELECTRA trains a small generator network to replace some tokens in the input sequence, and then a discriminator network tries to predict which tokens have been replaced. This setup allows for more efficient and effective learning of language representations. Additionally, ELECTRA utilizes shared embedding layers between the generator and the discriminator, which helps in reducing the model's complexity while maintaining performance [44-45].

These groundbreaking enhancements allow ELECTRA to perform exceptionally well in a variety of large-scale NLP tasks, including text generation. Its remarkable ability to manage long text sequences and disentangle complex semantic dependencies makes it a highly valuable asset for applications such as generating logical and contextually appropriate abstracts in scientific literature summarization, creating engaging and coherent storylines in the field of digital storytelling, and enabling meaningful and context-aware multi-turn conversations in virtual assistant systems.

When contrasted with traditional NLP models, ELECTRA showcases significantly better performance in capturing long-term dependencies within text and generating high-standard text outputs. Its unique approach to pre-training and model architecture sets it apart from the competition, establishing it as a potent tool within the NLP practitioner's arsenal. As the research landscape in NLP continues to evolve and progress, ELECTRA and similar advanced models are expected to be instrumental in shaping the trajectory of future text generation and a wide array of NLP applications.

In essence, ELECTRA capitalizes on the strengths of the self-attention mechanisms and the Transformer architecture to bring about substantial advancements in various NLP tasks, with a particular emphasis on text generation. Its innovative features render it a highly efficient and effective solution for contemporary NLP applications [46-51].

4. Result

This paper demonstrates the construction of a Sequence Labeling model that utilizes the RoBERTa pre-trained model. At the start, the RoBERTaBase pre-trained model is loaded and serves as the foundational backbone network for the entire architecture. Following this, a dense fully connected layer is added to the output of the RoBERTa model, and a softmax activation function is attached to this layer. This setup enables the model's output to be transformed and mapped to a predefined set of label categories for the sequence labeling task.

During the training stage, the RMSProp optimizer is applied, with a learning rate set at 3e-5. The Binary Cross Entropy loss function is used to compute the loss of the model, which helps in adjusting the model's parameters to minimize the difference between the predicted and actual labels. As for the evaluation of the model's performance, the Matthews Correlation Coefficient (MCC) is employed as the key metric. This metric provides a reliable measure of the model's quality, taking into account true and false positives and negatives in a balanced way. The outcomes of the model, including its performance on both training and test datasets, are depicted in Figure 4, offering a clear visual illustration of how well the model performs in the sequence labeling task.

The dataset is preprocessed to remove outliers and missing values, and then the data is divided in the ratio of 6:4, 40% of the data is used for model testing and 60% of the data is used for model training, and the accuracy is output using the test set to output the results of the binary classification.

From the obtained prediction outcomes, it is evident that the model exhibits a prediction accuracy of 78%. The precision is measured at 54%, the recall stands at 55%, and the F1-score is calculated to be 0.54. These figures indicate that the machine learning model retains the capacity to differentiate between the text generated by chatbots and human-produced natural language, attaining an accuracy rate of 78%. However, given that both the recall and precision values are relatively close to 50%, it strongly suggests that, to a certain degree, the text outputs of chatbots can be readily mistaken for natural language. This implies that there is still significant room for improvement in enhancing the model's discriminative power and reducing the ambiguity in distinguishing between these two types of text sources. [52-56]

5. Conclusion

In recent years, customer service chatbots powered by Large Language Models (LLMs) have increasingly become the spotlight in the artificial intelligence arena. These LLMs, trained via sophisticated and state-of-the-art learning methodologies, are highly advanced natural language processing models. Thanks to their remarkable capabilities in language comprehension and generation, these chatbots can interact with customers in a remarkably natural and human-like fashion, mimicking the fluidity of real conversations.

In this particular investigation, a thorough exploration of the dataset was carried out. This involved using data visualization techniques to gain insights into the characteristics of the dataset and performing meticulous preprocessing steps such as tokenization, stop word removal, and normalization. After that, the responses generated by the customer service chatbots were further processed using the T5 model. Subsequently, a neural network-based machine learning classifier was applied to distinguish between the text produced by the chatbots and the natural language used by human customers. The obtained results indicated that the model achieved a prediction accuracy of 78%, a precision of 54%, a recall of 55%, and an F1 score of 0.54.

These performance indicators suggest that the machine learning model does possess the ability to differentiate, to a certain degree, between the text generated by the chatbots and the natural language of human beings. Nevertheless, with both precision and recall values being relatively close to 50%, it is evident that there exists a significant level of ambiguity and difficulty in clearly separating the two. This implies that although the model can perform satisfactorily in numerous situations, it remains a formidable challenge to accurately and consistently distinguish between the content created by the chatbots and the genuine language expressions of human customers.

The outcomes of the experiment illustrate that the model has indeed made some headway in the task of identifying the sources of text, namely chatbot-generated and human-generated text. However, the persistent confusion underscores the urgent need to further enhance the model's proficiency in understanding and generating dialogues. When developing and implementing customer service chatbots based on large-scale language models, it is of great significance for these models to concentrate on improving their capacity to clearly demarcate their own output from the natural language of human customers.

In conclusion, even though the machine learning model demonstrates a certain level of promising accuracy in differentiating between chatbots and natural language, there is still substantial room for optimization. Continuously improving the model's performance is indispensable for better satisfying customer demands and ensuring more reliable, high-quality interactions in actual customer service scenarios.

References

[1] Li, Keqin, et al. "Exploring the Impact of Quantum Computing on Machine Learning Performance." (2024).
[2] Wang, Zixiang, et al. "Research on Autonomous Driving Decision-making Strategies based Deep Reinforcement Learning." arXiv preprint arXiv:2408.03084 (2024).
[3] Yan, Hao, et al. "Research on Image Generation Optimization based Deep Learning." (2024).
[4] Tang, Xirui, et al. "Research on Heterogeneous Computation Resource Allocation based on Data-driven Method." arXiv preprint arXiv:2408.05671 (2024).
[5] Su, Pei-Chiang, et al. "A Mixed-Heuristic Quantum-Inspired Simplified Swarm Optimization Algorithm for scheduling of real-time tasks in the multiprocessor system." Applied Soft Computing 131 (2022): 109807.
[6] Zhao, Yuwen, Baojun Hu, and Sizhe Wang. "Prediction of Brent crude oil price based on LSTM model under the background of low-carbon transition."arXiv preprint arXiv:2409.12376(2024).
[7] Diao, Su, et al. "Ventilator pressure prediction using recurrent neural network." arXiv preprint arXiv:2410.06552 (2024).
[8] Zhao, Qinghe, Yue Hao, and Xuechen Li. "Stock Price Prediction Based on Hybrid CNN-LSTM Model." (2024).
[9] Yin, Ziqing, Baojun Hu, and Shuhan Chen. "Predicting Employee Turnover in the Financial Company: A Comparative Study of CatBoost and XGBoost Models." (2024).
[10] Xu, Q., Wang, T., & Cai, X. (2024). Energy Market Price Forecasting and Financial Technology Risk Management Based on Generative AI. Preprints. https://doi.org/10.20944/preprints202410.2161.v1
[11] Wu, X., Xiao, Y., & Liu, X. (2024). Multi-Class Classification of Breast Cancer Gene Expression Using PCA and XGBoost. Preprints. https://doi.org/10.20944/preprints202410.1775.v2

[12] Wang, H., Zhang, G., Zhao, Y., Lai, F., Cui, W., Xue, J., Wang, Q., Zhang, H., & Lin, Y. (2024). RPF-ELD: Regional Prior Fusion Using Early and Late Distillation for Breast Cancer Recognition in Ultrasound Images. Preprints. https://doi.org/10.20944/preprints202411.1419.v1
[13] Min, L., Yu, Q., Zhang, Y., Zhang, K., & Hu, Y. (2024, October). Financial Prediction Using DeepFM: Loan Repayment with Attention and Hybrid Loss. In 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 440-443). IEEE.
[14] Accurate Prediction of Temperature Indicators in Eastern China Using a Multi-Scale CNN-LSTM-Attention model
[15] Rao, Jiarui, Qian Zhang, and Xinqiu Liu. "Applications Analyzing E-commerce Reviews with Large Language Models (LLMs): A Methodological Exploration and Application Insight." Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023 7.01 (2024): 207-212.
[16] Zhang, Qian, et al. "Sea MNF vs. LDA: Unveiling the Power of Short Text Mining in Financial Markets." International Journal of Engineering and Management Research 14.5 (2024): 76-82.
[17] Rao, Jiarui, et al. "Machine Learning in Action: Topic-Centric Sentiment Analysis and Its Applications." (2024).
[18] Rao, Jiarui, et al. "Integrating Textual Analytics with Time Series Forecasting Models: Enhancing Predictive Accuracy in Global Energy and Commodity Markets." Innovations in Applied Engineering and Technology (2023): 1-7.
[19] Zhang, Qian, and Jiarui Rao. "Enhancing Financial Forecasting Models with Textual Analysis: A Comparative Study of Decomposition Techniques and Sentiment-Driven Predictions." Innovations in Applied Engineering and Technology (2022): 1-6.
[20] Rao, Jiarui. "Machine Learning in Action: Topic-Centric Sentiment Analysis and Its Applications." Available at SSRN (2024).
[21] Rao, Jiarui, Qian Zhang, and Xinqiu Liu. "Applications Analyzing E-commerce Reviews with Large Language Models (LLMs): A Methodological Exploration and Application Insight." Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023 7.01 (2024): 207-212.
[22] Li, Chao, Jiarui Rao, and Qian Zhang. "LLM-Enhanced XGBoost-Driven Fraud Detection and Classification Framework." (2025).
[23] Rao, Jiarui, and Qian Zhang. "Deep Learning with LLM: A New Paradigm for Financial Market Prediction and Analysis." (2025).
[24] Rao, Jiarui, and Zeyu Wang. "Optimizing Stock Market Return Forecasts with Uncertainty Sentiment: Leveraging LLM-based Insights." Available at SSRN 5117562 (2024).
[25] Bo, Shi, and Minheng Xiao. "Time-Series K-means in Causal Inference and Mechanism Clustering for Financial Data." Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence. 2024.
[26] Rao, Jiarui, and Qian Zhang. "Deconstructing Digital Discourse: A Deep Dive into Distinguishing LLM-Powered Chatbots from Human Language." Journal of Theory and Practice in Education and Innovation 2.2 (2025): 18-25.
[27] Qian, Chenghao, et al. "WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation." arXiv preprint arXiv:2410.12075 (2024).
[28] Qin, Gaoyuan, et al. "Application of Convolutional Neural Network in Multimodal Emotion Recognition." 2024 9th International Symposium on Computer and Information Processing Technology (ISCIPT). IEEE, 2024.
[29] Wang, Randi, Vadim Shapiro, and Morad Mehandish. "Model consistency for mechanical design: Bridging lumped and distributed parameter models with a priori guarantees." Journal of Mechanical Design 146.5 (2024): 051710.
[30] Wang, Randi, and Morad Behandish. "Surrogate modeling for physical systems with preserved properties and adjustable tradeoffs." arXiv preprint arXiv:2202.01139 (2022).
[31] Wang, Randi, and Vadim Shapiro. "Topological semantics for lumped parameter systems modeling." Advanced Engineering Informatics 42 (2019): 100958.
[32] Liu, Yanming, et al. "Bridging context gaps: Leveraging coreference resolution for long contextual understanding." arXiv preprint arXiv:2410.01671 (2024).
[33] Privacy-Preserving Hybrid Ensemble Model for Network Anomaly Detection: Balancing Security and Data Protection
[34] Dai, Y., Wang, Y., Xu, B., Wu, Y., & Xian, J. (2020). Research on image of enterprise after-sales service based on text sentiment analysis. International Journal of Computational Science and Engineering, 22(2-3), 346-354.

[35] Cui, Wendi, et al. "Phaseevo: Towards unified in-context prompt optimization for large language models." arXiv preprint arXiv:2402.11347 (2024).
[36] Cui, Wendi, et al. "Divide-Conquer-Reasoning for Consistency Evaluation and Automatic Improvement of Large Language Models." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. 2024.
[37] Xiao, Minheng, Shi Bo, and Zhizhong Wu. "Multiple greedy quasi-newton methods for saddle point problems." 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS). IEEE, 2024.
[38] Zhang, Jiaxin, et al. "Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models." arXiv preprint arXiv:2410.09629 (2024).
[39] Bo, Shi, and Minheng Xiao. "Data-Driven Risk Measurement by SV-GARCH-EVT Model." 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS). IEEE, 2024.
[40] Li, Zhuohang, et al. "Towards statistical factuality guarantee for large vision-language models." arXiv preprint arXiv:2502.20560 (2025).
[41] Survival of the Safest: Towards Secure Prompt Optimization through Interleaved Multi-Objective Evolution
[42] Bo, Shi, and Minheng Xiao. "Root cause attribution of delivery risks via causal discovery with reinforcement learning." Algorithms 17.11 (2024): 498.
[43] SCE: Scalable Consistency Ensembles Make Blackbox Large Language Model Generation More Reliable
[44] Wang, Yu, et al. "Gradient-guided Attention Map Editing: Towards Efficient Contextual Hallucination Mitigation." arXiv preprint arXiv:2503.08963 (2025).
[45] Automatic Prompt Optimization via Heuristic Search: A Survey
[46] Zhang, Shiwen. "On the importance of source text embedding in text-guided image editing." (2024).
[47] Zhang, Shiwen. "Vector projection in text embedding space for text-guided image editing." (2024).
[48] Zhang, Shiwen. "Hyper-parameter tuning for text guided image editing." arXiv preprint arXiv:2407.21703 (2024).
[49] Zhang, Shiwen. "On the coefficient of model merging in forgetting mechanism." (2024).
[50] Xiao, Minheng, and Shi Bo. "Electroencephalogram emotion recognition via auc maximization." Algorithms 17.11 (2024): 489.
[51] Qian, Chenghao, et al. "WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation." IEEE Robotics and Automation Letters (2025).
[52] Li, Wanxin. "User-Centered Design for Diversity: Human-Computer Interaction (HCI) Approaches to Serve Vulnerable Communities." Journal of Computer Technology and Applied Mathematics 1.3 (2024): 85-90.
[53] Wang, Yang, Yojiro Mori, and Hiroshi Hasegawa. "Resource assignment based on core-state value evaluation to handle crosstalk and spectrum fragments in SDM elastic optical networks." 2020 Opto-Electronics and Communications Conference (OECC). IEEE, 2020.
[54] Wang, Yang, et al. "Optimizing multi-criteria k-shortest paths in graph by a natural routing genotype-based genetic algorithm." 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2018.
[55] Wang, Yang, Yojiro Mori, and Hiroshi Hasegawa. "Dynamic Routing and Spectrum Allocation Based on Actor-critic Learning for Multi-fiber Elastic Optical Networks." Photonics in Switching and Computing. Optica Publishing Group, 2021.
[56] Zhu, Yuanjing, and Yunan Liu. "Innovative Prompting Strategies and Holistic Evaluation of LLM Movie Recommender." 2024 International Conference on Intelligent Computing and Next Generation Networks (ICNGN). IEEE, 2024.