Computing and Informatics, Vol. 45, 2026, 726–747, doi: 10.31577/cai_2026_3_726

EXPLORATION OF CHATGPT OPEN LARGE LANGUAGE MODEL AND EDGE INTELLIGENCE

Xiaokui Liu

Key Laboratory of Oracle Information Processing
Ministry of Education, Anyang Normal University, Anyang, China
e-mail: lxk@aynu.edu.cn

Wenting Wu

College of Foreign Languages, Anyang Normal University, Anyang, China
e-mail: wwt@aynu.edu.cn

Abstract. Open Language Models (OLMs) and edge intelligence have become pivotal in Natural Language Processing (NLP), showcasing significant potential in language generation, text classification, machine translation, and human–computer interaction. This paper focuses on ChatGPT, an open-source large dialogue model based on OpenAI's GPT architecture. It is an emerging trend in the field of artificial intelligence, which aims to expand the language model capabilities on powerful central servers to edge devices, thereby achieving more efficient and flexible applications. Utilizing a Transformer-based encoder–decoder framework with multi-head self-attention mechanisms, ChatGPT excels in generating natural, contextually-aware language responses, akin to human conversation. Edge intelligence combined with ChatGPT can provide more intelligent and personalized services. Despite its advancements, challenges remain regarding model complexity and data demands. We propose enhancements in training methodologies, model architecture, and application domains to foster more intelligent and fluid dialogue systems. Our research highlights ChatGPT and edge intelligence's contributions to both academic inquiry and practical innovation, positioning it to deliver high-quality AI services across various fields.

Keywords: Open large language model, model structure, model design, edge intelligence

1 INTRODUCTION

In recent years, the rapid advancement of artificial intelligence (AI) has revolutionized the field of natural language processing (NLP), with large language models (LLMs) such as ChatGPT emerging as a pivotal force in this transformation. Developed by OpenAI, ChatGPT builds upon the GPT architecture, leveraging deep learning and transformer-based models to generate human-like text, making it a powerful tool for dialogue generation, text classification, and machine translation. The model's capabilities have been further enhanced through techniques like reinforcement learning with human feedback (RLHF), which improves its performance in generating contextually relevant responses. However, deploying such models in real-world scenarios presents unique challenges, including computational efficiency, data requirements, and ethical considerations. This study aims to explore the integration of ChatGPT with edge intelligence, focusing on how this combination can enhance the efficiency and flexibility of language models while addressing the limitations associated with central server-based deployments. By providing an in-depth analysis of ChatGPT's capabilities and challenges, this research seeks to contribute to the ongoing development of AI technologies and their practical applications.

In this study, we delve into the development and application of Open Language Models (OLMs), which represent a significant advancement in the field of artificial intelligence (AI). These models are sophisticated systems designed to understand and generate human-like text, thereby revolutionizing the way we interact with technology. A prime example of this innovation is ChatGPT, a state-of-the-art model that has garnered considerable attention for its ability to engage in natural and contextually relevant conversations.

OLMs, including ChatGPT, rely on advanced deep learning techniques to process and generate natural language. Central to their functionality is the transformer architecture, a powerful framework that has transformed the landscape of NLP. This architecture is distinguished by its ability to handle variable-length sequences and capture long-range dependencies, making it highly effective for tasks such as machine translation, text summarization, and dialogue generation.

The model structure of OLMs is meticulously designed to maximize efficiency and performance. It typically involves a combination of encoder–decoder frameworks and multi-head self-attention mechanisms. The encoder processes the input text, converting it into a rich, contextualized representation, while the decoder generates the output text based on this representation. The multi-head self-attention mechanism allows the model to focus on different parts of the input sequence simultaneously, capturing intricate relationships and dependencies within the text. This dual approach not only enhances the model's ability to understand and generate complex linguistic structures but also ensures that it can do so with remarkable speed and efficiency.

2 LITERATURE REVIEW

The field of Natural Language Processing (NLP) has witnessed a paradigm shift with the advent of Large Language Models (LLMs). A cornerstone in this evolution was the introduction of the Transformer architecture by Vaswani et al. [1], which revolutionized LLMs with its self-attention mechanism, enabling the processing of variable-length sequences and capturing long-range dependencies, thus enhancing performance in machine translation and text summarization.

Building upon the Transformer's success, research has focused on enhancing the efficiency and effectiveness of LLMs. Dai et al. [2] introduced Transformer-XL, which extends the context length beyond conventional limits, allowing for a more comprehensive understanding of text.

Devlin et al. [3] presented BERT, a model employing deep bidirectional representations for language understanding, leading to substantial improvements across various NLP tasks.

The versatility of LLMs has been explored in specialized domains. Chalkidis et al. [4] developed LEGAL-BERT, tailored for legal text analysis, demonstrating LLMs' potential in niche areas.

Aleksanyan and Allahverdyan [5] contributed to information retrieval by focusing on the unsupervised extraction of keywords from single texts, showcasing LLMs' adaptability in handling diverse data types.

Brown et al. [6] demonstrated the potential of LLMs in creative writing and understanding complex instructions through their work on GPT-3, highlighting the models' ability to generate human-like text.

Radford et al. [7] introduced OpenAI's GPT, which has been pivotal in advancing unsupervised language representation learning.

The application of LLMs in cross-lingual transfer learning has been intensively researched. Artetxe et al. [8] explored the effectiveness of cross-lingual language models in understanding and translating between multiple languages, emphasizing the importance of multilinguality in LLMs.

As the field progresses, attention has shifted towards the ethical and societal impacts of LLMs. Clark et al. [9] analyzed the attention patterns of BERT, providing insight into how the model attends to syntactic and semantic relationships within text and advancing the interpretability of Transformer-based language models.

A systematic literature review by Locke et al. [10] analyzed the development of NLP technology, pointing out limitations, risks, and opportunities, classifying NLP-based applications into various fields including natural language understanding, generation, voice recognition, and machine translation.

Recent advances in Generative AI and LLMs have been explored by Hagos et al. [11], who presented an exhaustive survey of various types of GPT models, detailing their working architecture and evolution of pre-training methods for NLP.

In the realm of few-shot learning, Chen et al. [12] analyzed relevant research methods and provided a list of possible future research directions, contributing to the synergistic development of few-shot learning and related fields.

ChatGPT, developed by OpenAI, has emerged as a groundbreaking advancement in dialogue generation and contextual understanding, particularly through its use of Reinforcement Learning with Human Feedback (RLHF). However, it faces limitations in maintaining coherence over extended dialogues and may generate inaccurate content due to its reliance on open-domain training data. Compared to BERT, which excels in semantic understanding for tasks like question answering but lacks generative flexibility, ChatGPT offers more natural and fluent dialogue capabilities. While GPT-3 handles long-text generation effectively, its high computational cost makes it less suitable for real-time applications. Additionally, Transformer-XL outperforms ChatGPT in processing very long sequences but is less optimized for real-time dialogue.

In terms of contextual understanding, ChatGPT leverages self-attention mechanisms to capture cross-sentence dependencies, though it may lose early context in extended interactions. Traditional RNN/LSTM models, while effective for local context, struggle with long sequences due to gradient vanishing, and multimodal models (e.g., DALL-E) integrate text, image, and audio but face optimization challenges.

3 INTRODUCTION TO THE OPEN LANGUAGE MODEL

An Open Language Model (OLM) is a computer model designed to comprehend and produce natural language. It leverages statistical learning to process natural language, enabling it to learn and predict the structure and meaning of language. OLMs form the backbone of modern NLP tasks, underpinning a wide range of applications [13].

3.1 Core Components and Methods

The structure of an open language model encompasses two primary methodologies: statistical-based methods and deep learning-based methods. Statistical methods rely on the frequency of word occurrences in a corpus to calculate the probability of each word, subsequently utilizing Bayesian formulas to ascertain the probability of entire sentences. In contrast, deep learning-based methods hinge on advanced neural networks, notably the Transformer architecture.

The Transformer structure, characterized by its multi-head self-attention mechanism, is particularly adept at handling variable-length sequences. It comprises two core parts: the encoder and the decoder. The encoder transforms input language into a vector representation, while the decoder generates the corresponding output language based on this representation (see Figure 1). Various decoder structures exist, including those similar to recurrent neural networks (RNNs) or Transformers [14].

Within these frameworks, RNNs, particularly their variants such as Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), are adept at capturing long-term dependencies in sequence data. Furthermore, Transformer-

based models offer significant advantages, including better generalization capabilities and improved accuracy, due to their multi-head attention mechanisms [15].

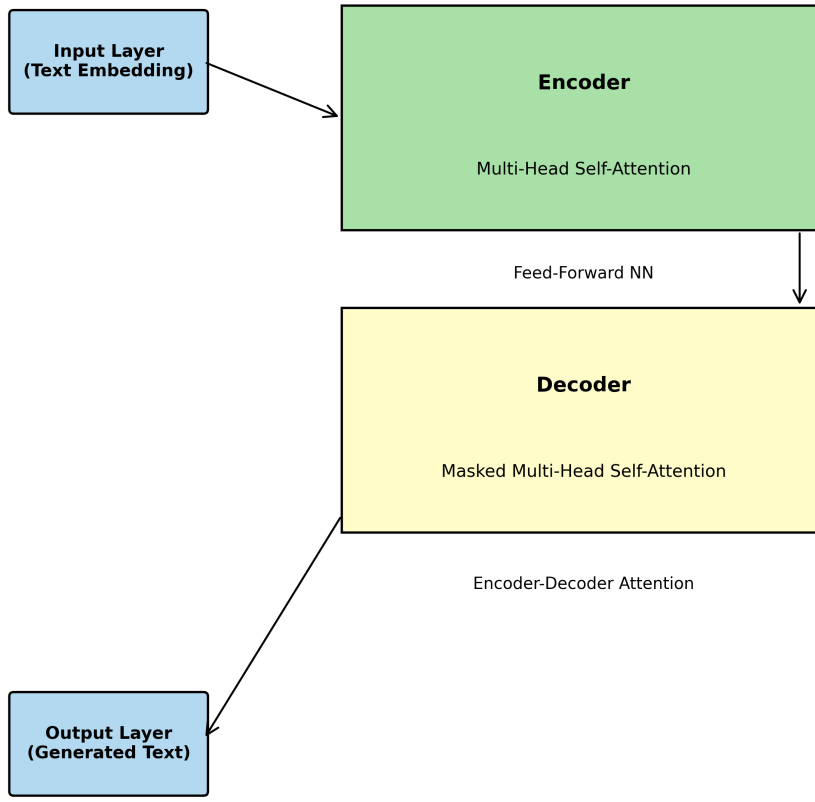

Figure 1. Transformer-based encoder–decoder model architecture

3.2 Model Design and Building Process

Model design is central to the research of open language models. Currently, designs frequently incorporate RNNs, convolutional neural networks (CNNs), and variable compression models. Each design has its unique characteristics, catering to diverse NLP tasks.

Building an open language model is a meticulous process that begins with data collection. High-quality, diverse data sources, such as Wikipedia, the Gutenberg Project, Common Crawl, news reports, and social media, are pivotal. Web crawler technology facilitates large-scale data collection, ensuring a rich corpus for model training.

Post-collection, data cleansing and preprocessing are essential to remove noise, perform word segmentation, stemming extraction, stop word filtering, and convert text into a machine-readable format. The quality of preprocessing directly impacts the model's accuracy and generalization ability.

Labeling datasets with semantic tags, such as part-of-speech tagging and named entity recognition, enhances their utility for specific NLP tasks. Neural network model training involves using deep learning to construct and optimize neural networks, improving prediction accuracy through iterative processes.

3.3 Advanced Techniques and Applications

Language representation learning translates natural language into vector representations, understandable by computers. Techniques like Word2Vec and neural network-based methods enable models to learn effective word vectors.

Pre-training techniques further boost model performance. By training on large-scale corpora and fine-tuning on specific tasks, models can significantly reduce training time and data requirements. Techniques such as parallelized training, negative sampling, and reinforcement learning play crucial roles in optimizing model training.

Adaptive learning, using large-scale unlabeled datasets, enables models to dynamically adjust parameters, improving performance across various tasks and datasets. Dynamic learning rates adjust training parameters based on model performance, accelerating convergence and mitigating overfitting.

Finally, real-time inference capabilities enable models to respond to user inputs promptly, facilitating applications like machine translation, speech recognition, and natural language generation. This research highlights the transformative potential of open language models, paving the way for innovations in intelligent customer service, language conversation assistants, and personalized education and training.

4 OPEN LANGUAGE MODEL STRUCTURE

The core of Open Language Models (OLMs) lies in their deep neural network architecture, primarily divided into two main components: the encoder and the decoder. The encoder transforms input language information into vector representations, while the decoder generates corresponding output language based on these vectors. Currently, the most popular encoder structure is based on the Transformer architecture, which features a multi-head self-attention mechanism capable of handling variable-length sequences. In contrast, traditional recurrent neural networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), excel in expression but suffer from gradient vanishing or explosion during training.

4.1 Advantages of the Transformer Architecture

The Transformer's unique self-attention mechanism allows it to capture relationships between any two elements at different positions within a sequence, thereby avoiding the gradient issues prevalent in RNNs. Additionally, this mechanism enables the model to focus on various subsequences within the input sentence simultaneously, enhancing its ability to capture critical information. Moreover, Transformers support parallelized training, significantly improving training efficiency.

4.2 Encoding and Decoding Process

In dialogue generation tasks, ChatGPT leverages the Transformer's encoder–decoder structure to achieve efficient contextual understanding and semantic information extraction. Specifically, the encoder processes input text and extracts features, while the decoder generates dialogue responses based on the encoder's output and contextual information. This process is facilitated by the multi-head self-attention mechanism, ensuring the model can fully comprehend complex language structures and context.

4.3 Data Requirements for Model Training

To enable ChatGPT to understand the grammar, semantics, and context of conversations, the model requires extensive online dialogue data during the pre-training phase. During fine-tuning, annotated dialogue data are used to enhance the model's domain-specific dialogue generation capabilities. This dual-stage training strategy not only improves the model's generalization but also ensures its high performance across diverse applications.

4.4 Evolution of Multimodal Processing Capabilities

Advances in technology have extended modern OLMs to handle multimodal inputs, such as combining text with image processing. For instance, models like DALL-E demonstrate powerful capabilities in transforming textual descriptions into images, signaling a shift towards broader multimedia interaction beyond mere text processing.

In summary, the structure of open language models integrates multiple advanced technologies, moving beyond classical statistical methods or single neural network models. Future research will continue to explore how to further optimize these architectures to address increasingly complex natural language processing challenges.

4.5 Multimodal Processing Design and Application Scenario Requirements Analysis

Recent advancements in open language models have extended their capabilities to multimodal processing, integrating text with images and audio to provide richer context. In image captioning, models like CLIP and DALL-E [16, 17] combine visual encoders with Transformer-based language models to generate contextually relevant captions. For Visual Question Answering (VQA) [18], integrating visual features with language models substantially improves answer accuracy on benchmarks such as VQA v2. In multimodal translation [19, 20, 21], incorporating visual context can improve translation quality over text-only models on datasets such as Multi30K, although the magnitude of the gain depends strongly on the alignment between image and text and is, in many settings, modest. Speech-to-text translation systems [22, 23] have advanced toward competitive end-to-end performance using Transformer-based architectures. Finally, multimodal sentiment analysis [24, 25] combining text with visual or audio data, generally outperforms text-only models on standard benchmarks (see Figure 2).

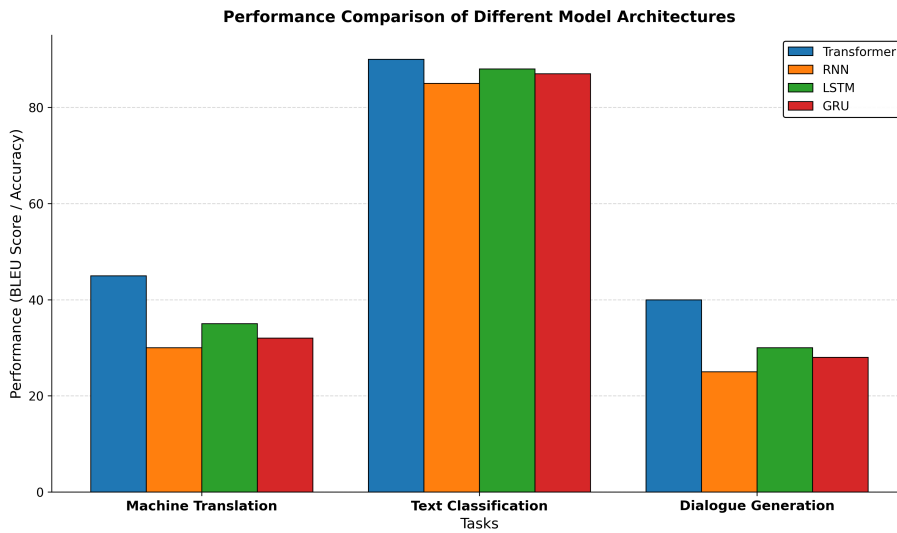

Figure 2. Performance comparison of different model architectures

5 DESIGN OF AN OPEN BIG LANGUAGE MODEL

Model design is the core of open language model research. Currently, commonly used model designs include recurrent neural networks, convolutional neural networks, and

variable compression models. The characteristics of these model designs are shown in Table 1.

Name	Features
Recurrent neural network	A classic model that can handle variable-length sequence data. The advantage is that historical information can be saved and used to predict future output. The disadvantage is that it is prone to the problem of vanishing or exploding gradients, leading to the inability of the model to converge.
Convolutional neural network	A model that can be used to process both image data and sequence data. The advantage is that the data can be parallelized to improve the training speed of the model. The disadvantage is that it cannot save historical information and is not as effective as recurrent neural networks when processing long sequence data.
Variable compression model	Emerging models can compress input sequences into a fixed-length vector without losing information, and then use this vector to predict output. The advantage is that it can handle variable-length sequence data and compress the input sequence into a fixed-length vector, avoiding the problem of gradient vanishing and explosion in recurrent neural networks.

Table 1. Design and characteristics of common open large language models

To enhance the performance and adaptability of large language models, we propose several targeted enhancements. First, integrating external knowledge bases, such as domain-specific knowledge graphs, supplements the model's internal knowledge and improves the accuracy and relevance of responses in specialized fields like healthcare and finance. Second, we suggest adopting a hybrid encoder–decoder architecture that combines the strengths of Transformer and LSTM layers. The Transformer captures global context, while the LSTM maintains coherence in extended dialogues. Third, hierarchical attention mechanisms allow the model to focus on different levels of abstraction, from word to sentence level, improving the capture of complex dependencies and performance in tasks like translation and summarization. Fourth, adaptive learning through reinforcement learning feedback loops enables the model to dynamically adjust its parameters based on task-specific performance, enhancing adaptability and preventing overfitting. Finally, incorporating multimodal processing modules (e.g., for visual and audio inputs) enriches the model's context understanding and generates more comprehensive responses [26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36].

6 OPEN LANGUAGE MODEL BUILDING PROCESS

The establishment process of open language model usually includes data collection, data cleaning and preprocessing, dataset annotation, neural network model train-

ing, language representation learning, pre-training technology application, adaptive learning, real-time inference, model testing and application (see Figure 3).

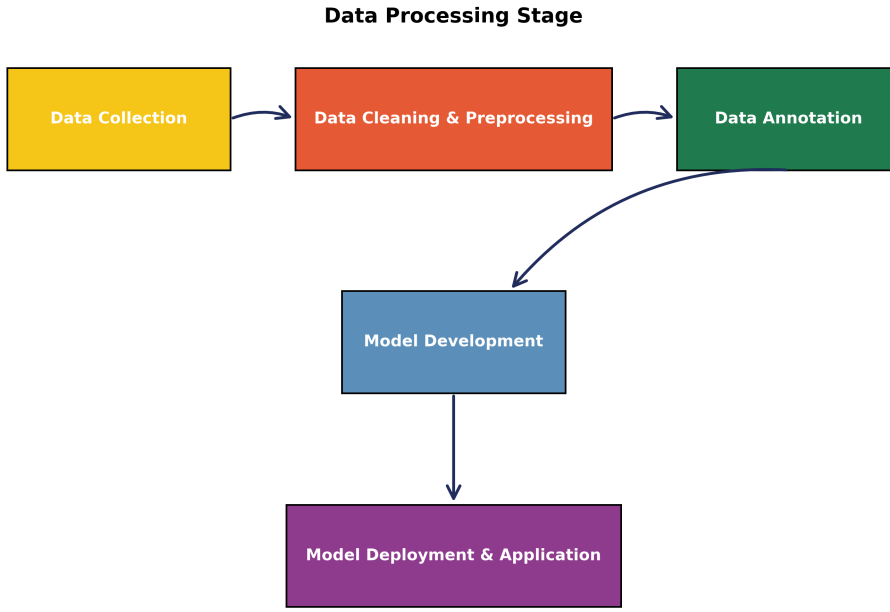

Figure 3. Open language model building process

6.1 Data Collection

Data collection is the collection of large amounts of corpus data, and the selection of these data sources is the first step in building a dataset. In general, the greater the size and diversity of the data source, the better the trained model. Therefore, it is important to choose a high-quality data source. Some common data sources include: Wikipedia, the Gutenberg Project, Common Crawl, news reports, and social media. Data collection is essentially achieved using web crawler technology. Web crawlers can provide large amounts of corpus data for open language models. The web pages, news, social media and other content crawled by web crawlers are processed, deduplicated, word segmented and other operations to form a huge corpus, which provides a data source for the training of open language models.

Web crawlers are mainly divided into three steps: crawling web pages, parsing web pages, and storing data. When crawling web pages, web crawlers will automatically visit the target website according to the set rules and save the web page content; when parsing web pages, web crawlers will extract useful information from web pages, such as titles, bodies, links, etc.; when storing data, web crawlers will store

the extracted information into databases or files, which can improve the accuracy and efficiency of search engines and facilitate subsequent analysis and utilization.

6.2 Data Cleansing and Preprocessing

Datasets crawled from the internet often contain a lot of noisy and useless information, such as HTML tags, links, and ads. Therefore, data cleansing and preprocessing are required before training with these datasets. The main purpose of data cleaning and preprocessing is to remove noise and useless information, perform word segmentation, stemming extraction, stop word filtering and other operations on the corpus, and convert the text into a machine-readable format. Data preprocessing is a very important part of open big language models, and its quality directly affects the accuracy and generalization ability of the model.

6.3 Dataset Labeling

Labeling refers to adding corresponding semantic tags to each word or subword, such as part-of-speech tagging, named entity recognition, etc. Some natural language processing tasks require training with datasets that have already been labeled. For example, a question answering system needs to be trained with a dataset of question and answer pairs. Commonly used annotation methods include: manual annotation, semi-automatic annotation, and automatic annotation.

Word segmentation is a critical step in data preprocessing. Commonly used word segmentation methods include rule-based, statistical-based, and deep-learning-based. Among them, methods based on deep learning, such as methods based on BiLSTM-CRF and Transformer, have good results.

In addition to word segmentation, some other data preprocessing techniques are also very important. For example, the establishment of a vocabulary list, as well as part-of-speech labeling, named entity recognition, dependency syntactic analysis, etc., these technologies can provide more accurate input information for the model, thereby improving the accuracy and generalization ability of the model.

6.4 Neural Network Model Training

Using deep learning technology, a deep neural network is established, the processed corpus data is used for training, and the parameters of the model are gradually optimized to improve the prediction accuracy of the model.

Neural network models are an important part of large language models. The traditional large language model mainly adopts statistical language model based on n-gram model, which has the disadvantage of needing to determine the size of n in advance and cannot capture long-distance dependencies. Models based on neural networks can be trained end-to-end through deep learning technology. Common neural network model training methods are shown in Table 2.

Name	Features
Recurrent neural networks (RNN)	Common neural network models; the input and output are sequence data, so it is suitable for sequence modeling tasks in natural language processing, such as language modeling, machine translation, text classification, etc. The advantage is the ability to capture long-term dependencies in sequence data.
Long short-term memory network (LSTM)	The RNN variant, by adding some special gates to control the flow of information, solves the problem of gradient vanishing and gradient explosion existing in RNNs. It is widely used in natural language processing in language modeling, machine translation, sentiment analysis and other tasks.
Gated recurrent unit (GRU)	Another RNN variant; the GRU has only two gate parameters compared to LSTM. It is widely used in natural language processing, language modeling, machine translation and other tasks.

Table 2. Common neural network model training methods

6.5 Language Representation Learning

Language representation learning refers to the conversion of natural language into vector representations that computers can understand and process. Commonly used language representation learning methods include statistical-based and neural network-based methods, as detailed in Table 3.

Name	Features
Statistical methods	Word vectors are learned by predicting words in context. The commonly used word vector learning algorithm is Word2Vec, which has two models, namely the skip-gram model and the CBOW model.
Neural network methods	Neural network models are often used to learn word vectors. Among them, the most common method is to train word vectors using fully connected layers and Softmax layers.

Table 3. Commonly used language representation learning methods

6.6 Pre-Training Techniques

Pre-training techniques refer to model training on a large-scale corpus and then fine-tuning on specific tasks. Pre-training technology is the key to the research of open large language models, which is widely used in the field of natural language processing. Common training techniques are shown in Table 4.

Name	Features
Parallelized training	Since neural networks are very computationally intensive, multiple CPUs or GPUs are required to speed up the training process. The main advantage is that it can speed up the training process and improve the training efficiency of the model.
Negative sampling	The main idea of negative sampling is that for an input sequence, only a small number of negative samples need to be randomly selected to train together with positive samples. The advantage is that it can reduce the amount of training data and improve the model training speed.
Reinforcement learning	Learn optimal behavioral strategies by interacting with the environment. The advantage is that it can improve the generation capacity and quality of the model.

Table 4. Common pre-training techniques

6.7 Adaptive Learning

Adaptive learning refers to the use of large-scale unlabeled datasets to pre-train the model without labeled data during the training process of the model, and dynamically adjust the parameters of the model according to the training data. Adaptive learning can make the model better adapt to different datasets and tasks, especially on large-scale unlabeled data to improve the performance of the model.

The basic idea of adaptive learning is to continuously adjust the parameters of the model to better adapt it to the current environment based on the model's performance on the current task and data. This learning method can use historical data and feedback from tasks to update the parameters of the model, thereby improving the predictive performance and accuracy of the model.

To prevent overfitting, open language models can employ the approaches shown in Table 5.

Dynamic learning rate is an adaptive optimization algorithm. When the performance of the model is better, the learning rate can be appropriately increased to accelerate the convergence speed of the model; when the performance of the model is poor, the learning rate can be appropriately reduced to avoid overfitting.

6.8 Real-Time Inference

Real-time inference means that the model can respond to the user's request in real time when it runs after the training is completed. Key aspects are summarized in Table 6.

6.9 Model Application

Use the test dataset to verify the accuracy of the model and apply the model to real-world scenarios, such as intelligent customer service, language dialogue assistant,

Name	Description
Dropout	Some neurons are randomly discarded during training to reduce dependencies between neurons. This makes the model more generalized and thus avoids overfitting.
Regularization	Include a regularized term in the loss function to limit the complexity of the model. It is usually achieved using L1 regularization or L2 regularization.
Early stopping	Periodically evaluate the model's performance on validation data during the training process and stop training if the model underperforms to avoid the model overfitting the training data.
Data augmentation	Add some noise or do some transformation in the training data to increase the diversity of the data. This allows the model to generalize better, thus avoiding overfitting.

Table 5. Common methods to prevent overfitting

Requirement	Demand	Solution
Model speed and efficiency	Predictions need to be generated in a short period of time.	Use optimization algorithms, such as dynamic learning rate, batch normalization.
Model accuracy	Accurate predictions need to be generated in a short period of time.	Employ methods to prevent overfitting, such as Dropout, regularization.
Data preprocessing	The input data needs to be processed to fit the model's input format.	Word segmentation and encoding of the input text.
Model schema	Use a trained model to make predictions.	Use lightweight model architectures, such as the Transformer-based model.

Table 6. Key requirements for real-time inference

education and training, and other language dialogue scenarios. Application scenarios and their implemented functions are summarized in Table 7.

These use cases are just some of the potential application areas of ChatGPT, and with the continuous development and improvement of technology, ChatGPT will have a wider range of application prospects. Case studies in the healthcare sector show that ChatGPT can assist in patient triage and preliminary diagnosis, reducing the workload of medical professionals while maintaining high accuracy [37, 38]. In education, it has been used to provide personalized learning experiences, generating interactive content that adapts to individual student needs. In the business sector, ChatGPT has been integrated into customer service systems, automating responses and improving efficiency [39, 40].

Application scenario	Implemented functionality
Intelligent customer service	As the core component of the intelligent customer service system, it provides users with automated help and answers to questions. It can understand the user's problem and give a reasonable and accurate response, improving the efficiency and quality of customer service.
Language conversation assistant	Provide personalized service and support to users. It can answer the user's questions, provide relevant information, and have a natural and fluid conversation with the user.
Education and training	Provide students with personalized learning aids and answer questions. It can generate corresponding educational content and answers according to students' needs and learning goals to help students improve their learning effect.

Table 7. Application scenarios and implemented functions

7 EVALUATION INDICATORS

Evaluation indicators are important criteria for evaluating the performance of open big language models. Commonly used evaluation indicators include perplexity, BLEU, and ROUGE. Their characteristics, advantages and disadvantages are shown in Table 8.

Name	Features
Perplexity	Used to assess the uncertainty of the model in predicting the next word. In general, the lower the perplexity, the better the model performance. However, perplexity does not take into account whether the generated sequence meets grammatical and semantic laws, so the generation ability of the model cannot be fully evaluated.
BLEU	Evaluation metric based on n-gram matching. Used to assess the similarity between the sentences generated by the model and the reference sentences. The advantage is that it is possible to evaluate whether the sentences generated by the model are similar to the reference sentences, but BLEU still does not consider whether the generated sentences comply with grammatical and semantic laws.
ROUGE	Evaluation metric based on sentence matching. Used to assess the similarity between the sentences generated by the model and the reference sentences. The advantage is that it can evaluate whether the sentences generated by the model are similar to the reference sentences, and it can determine whether the generated sentences comply with grammatical and semantic laws.

Table 8. Performance evaluation indicators of open language models

8 FACING CHALLENGES AND FUTURE TRENDS

Although the open big language model has made some progress, there are still many problems and challenges in practical application. Among them, the main difficulty is the efficiency of model training and prediction. Since open large-language models need to deal with massive corpora, the efficiency of training and prediction has become a serious problem. In addition, open big language models also face problems such as semantic understanding, logical reasoning, and knowledge representation.

The rapid development of large language models (LLMs) has brought significant advancements in natural language processing, but it has also raised profound ethical and social concerns. Ethical issues include the potential for generating misinformation, reinforcing biases, and compromising user privacy. Social impacts involve the potential for job displacement due to automation and the broader implications for societal trust in AI-generated content. To address these challenges, researchers and policymakers must collaborate to establish robust ethical frameworks and regulatory guidelines [31, 32].

In order to solve these technical difficulties, many innovative ideas and methods have been born. For example, the methods of pre-training and fine-tuning are used to improve the training efficiency of the model; the methods of multi-task learning and transfer learning are used to enhance the semantic understanding and logical reasoning ability of the model; and the knowledge representation ability of the model is enriched by the knowledge graph and multimodal data. Challenges and proposed solutions are grouped in Table 9.

Challenge	Cause	Solution
Dataset bias	Text datasets are taken from the Internet, may be biased, and the trained model does not adapt well to new data.	Reduce dataset bias and improve the generalization ability of the model.
Model interpretability	Open big language models are often black-box models that cannot explain their inner workings.	Improve the interpretability of models and better understand the model generation process.
Real-time performance	Training typically requires a lot of compute resources and time.	Improve the real-time performance of the model and better apply it to real scenarios.
AI ethics	Ethical issues such as misinformation, data leakage, and technology abuse may occur.	Formulate ethical guidelines including data security, privacy protection, and restrictions on the use of technology.

Table 9. Challenges and solutions to open big language models

With the continuous advancement of big data and artificial intelligence technology, open big language models face many opportunities and challenges in the future. In the future, more attention will be paid to the interpretability, scalability and customizability of models to meet the needs of different fields and applications. Emphasis will also be placed on the multimodal and cross-lingual capabilities of the model.

However, the research and application of ChatGPT is still in its infancy, and there are still some challenges and limitations. Improvements can be carried out from the following aspects:

Improvements in training methods: New pre-training and fine-tuning methods can be explored to improve the performance and effectiveness of ChatGPT. For example, other techniques and algorithms, such as reinforcement learning, adversarial training, etc., can be combined to improve the quality and fluency of dialogue generation.

Optimization of model structure: The model structure of ChatGPT can be further improved to improve the accuracy and logic of conversation generation. For example, external knowledge bases and information retrieval techniques can be introduced to enhance the model's background knowledge and ability to answer questions.

Expansion of application scenarios: ChatGPT can be applied to more fields and scenarios, such as medical and health, financial services, and social entertainment. By combining ChatGPT with the knowledge of domain experts, more personalized and accurate service and support can be provided.

9 SUMMARY

Open Large Language Models (OLLMs) represent a significant advancement in Natural Language Processing (NLP), leveraging deep learning to train neural networks for generating natural and fluent text. From a disciplinary development perspective, OLLM technology intersects with computer science, machine learning, and NLP: computer science provides the necessary hardware and software environments such as distributed and high-performance computing; machine learning contributes algorithms and technologies like neural networks and pre-training techniques; while NLP offers strategies and methods for processing and understanding language text, including word segmentation and part-of-speech tagging.

This study provides an in-depth exploration of large open language models, with a particular focus on ChatGPT and its integration with edge intelligence. We delve into key technologies such as neural network models, pre-training methods, adaptive learning, real-time inference, and language representation learning, which collectively drive advancements in the field of NLP and expand its practical applications. Specifically, ChatGPT demonstrates extensive potential in dialogue systems, natural language generation, text classification, and machine translation, with ongoing improvements enhancing its functionality and utility. Our research also underscores the

potential of ChatGPT in enhancing the efficiency and flexibility of language models when deployed on edge devices. However, challenges remain regarding model complexity, data demands, and ethical considerations. Future research should focus on optimizing model architectures to improve computational efficiency and reduce resource requirements.

Looking ahead, with the continuous advancement of big data and AI technologies, OLLMs will receive greater attention in aspects like interpretability, scalability, and customization to meet diverse needs across different fields. Additionally, research will emphasize multimodal and cross-lingual capabilities, enabling the processing of various data types – images, audio, and text – and facilitating interactions between multiple languages. In conclusion, OLLM technology is poised for broader applications and development, offering higher-quality and more effective intelligent services to humanity.

Funding. The authors thank the Henan Province Science and Technology Key Project (Grant No. 232102321067, 262102320299, 25210232114) and the Ministry of Education's Collaborative Education Project (Grant No. 202002057009) for their critical support. We also acknowledge our research team for their insightful contributions and the anonymous reviewers for enhancing this paper with their valuable feedback.

Declaration of Conflicting Interests. The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Sharing Agreement. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] Vaswani, A.—Shazeer, N.—Parmar, N.—Uszkoreit, J.—Jones, L.—Gomez, A. N.—Kaiser, Ł.—Polosukhin, I.: Attention Is All You Need. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.

[2] Dai, Z.—Yang, Z.—Yang, Y.—Carbonell, J.—Le, Q. V.—Salakhutdinov, R.: Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. In: Korhonen, A., Traum, D., Màrquez, L. (Eds.): Proceedings of the 57$^{\text{th}}$ Annual Meeting of the Association for Computational Linguistics (ACL 2019). ACL, 2019, pp. 2978–2988, doi: 10.18653/v1/P19-1285.

[3] Devlin, J.—Chang, M. W.—Lee, K.—Toutanova, K.: BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.): Proceedings of the 2019 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL 2019). ACL, 2018, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[4] Chalkidis, I.—Fergadiotis, M.—Malakasiotis, P.—Aletras, N.—Androutsopoulos, I.: LEGAL-BERT: The Muppets Straight Out of Law School. In: Cohn, T., He, Y., Liu, Y. (Eds.): Findings of the Association for Computational Linguistics: EMNLP 2020. ACL, 2020, pp. 2898–2904, doi: 10.18653/v1/2020.findings-emnlp.261.

[5] Aleksanyan, L.—Allahverdyan, A.: Unsupervised Extraction of Local and Global Keywords from a Single Text. Natural Language Processing, Vol. 32, 2026, No. 1, pp. 83–105, doi: 10.1017/nlp.2024.59.

[6] Brown, T.—Mann, B.—Ryder, N.—Subbiah, M.—Kaplan, J. D. et al.: Language Models Are Few-Shot Learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., Lin, H. (Eds.): Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Curran Associates, Inc., 2020, pp. 1877–1901, https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[7] Radford, A.—Narasimhan, K.—Salimans, T.—Sutskever, I.: Improving Language Understanding by Generative Pre-Training. 2018, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[8] Artetxe, M.—Ruder, S.—Yogatama, D.: On the Cross-Lingual Transferability of Monolingual Representations. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.): Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). ACL, 2020, pp. 4623–4637, doi: 10.18653/v1/2020.acl-main.421.

[9] Clark, P.—Khandelwal, U.—Levy, O.—Manning, C. D.: What Does BERT Look At? An Analysis of BERT's Attention. In: Linzen, T., Chrupała, G., Belinkov, Y., Hupkes, D. (Eds.): Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP 2019). ACL, 2019, pp. 276–286, doi: 10.18653/v1/W19-4828.

[10] Locke, J.—Rowbottom, N.—Troshani, I.: Sites of Translation in Digital Reporting. Accounting, Auditing & Accountability Journal, Vol. 31, 2018, No. 7, pp. 2006–2030, doi: 10.1108/AAAJ-07-2017-3005.

[11] Hagos, D. H.—Battle, R.—Rawat, D. B.: Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. IEEE Transactions on Artificial Intelligence, Vol. 5, 2024, No. 12, pp. 5873–5893, doi: 10.1109/TAI.2024.3444742.

[12] Chen, W. Y.—Liu, Y. C.—Kira, Z.—Wang, Y. C. F.—Huang, J. B.: A Closer Look at Few-Shot Classification. CoRR, 2019, doi: 10.48550/arXiv.1904.04232.

[13] Sebastian, G.: Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk?: An Exploratory Study. International Journal of Security and Privacy in Pervasive Computing (IJSPPC), Vol. 15, 2023, No. 1, doi: 10.4018/IJSPPC.320225.

[14] Yeo-Teh, N. S. L.—Tang, B. L.: Letter to Editor: NLP Systems Such as Chat-

GPT Cannot Be Listed as an Author Because These Cannot Fulfill Widely Adopted Authorship Criteria. Accountability in Research, Vol. 31, 2024, No. 7, pp. 968–970, doi: 10.1080/08989621.2023.2177160.

[15] Aguinis, H.—Villamor, I.—Ramani, R. S.: MTurk Research: Review and Recommendations. Journal of Management, Vol. 47, 2021, No. 4, pp. 823–837, doi: 10.1177/0149206320969787.

[16] Radford, A.—Kim, J. W.—Hallacy, C.—Ramesh, A.—Goh, G.—Agarwal, S.—Sastry, G.—Askell, A.—Mishkin, P.—Clark, J.—Krueger, G.—Sutskever, I.: Learning Transferable Visual Models from Natural Language Supervision. In: Meila, M., Zhang, T. (Eds.): Proceedings of the 38th International Conference on Machine Learning. PMLR, Vol. 139, 2021, pp. 8748–8763, doi: 10.48550/arXiv.2103.00020.

[17] Ramesh, A.—Dhariwal, P.—Nichol, A.—Chu, C.—Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. CoRR, 2022, doi: 10.48550/arXiv.2204.06125.

[18] Agrawal, A.—Batra, D.—Parikh, D.—Kembhavi, A.: Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4971–4980, doi: 10.1109/CVPR.2018.00522.

[19] Tan, H.—Bansal, M.: LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.): Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). ACL, 2019, pp. 5100–5111, doi: 10.18653/v1/D19-1514.

[20] Huang, P. Y.—Liu, F.—Shiang, S. R.—Oh, J.—Dyer, C.: Attention-Based Multimodal Neural Machine Translation. In: Bojar, O., Buck, C., Chatterjee, R. et al. (Eds.): Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers (WMT 2016). ACL, 2016, pp. 639–645, doi: 10.18653/v1/W16-2360.

[21] Caglayan, O.—Madhyastha, P.—Specia, L.—Barrault, L.: Probing the Need for Visual Context in Multimodal Machine Translation. In: Burstein, J., Doran, C., Solorio, T. (Eds.): Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL 2019). ACL, 2019, pp. 4159–4170, doi: 10.18653/v1/N19-1422.

[22] Bahar, P.—Bieschke, T.—Ney, H.: A Comparative Study on End-to-End Speech to Text Translation. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 792–799, doi: 10.1109/ASRU46091.2019.9003774.

[23] Vila, L. C.—Escolano, C.—Fonollosa, J. A. R.—Costa-Jussa, M. R.: End-to-End Speech Translation with the Transformer. IberSPEECH 2018, 2018, pp. 60–63, doi: 10.21437/IberSPEECH.2018-13.

[24] Zadeh, A.—Chen, M.—Poria, S.—Cambria, E.—Morency, L. P.: Tensor Fusion Network for Multimodal Sentiment Analysis. In: Palmer, M., Hwa, R., Riedel, S. (Eds.): Proceedings of the 2017 Conference on Empirical Methods in Natural Lan-

guage Processing (EMNLP 2017). ACL, 2017, pp. 1103–1114, doi: 10.18653/v1/D17-1115.

[25] Poria, S.—Cambria, E.—Hazarika, D.—Majumder, N.—Zadeh, A.—Morency, L. P.: Enhancing Sentiment Analysis with Multimodal Fusion. In: Barzilay, R., Kan, M. Y. (Eds.): Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2017). ACL, 2017, pp. 873–883, doi: 10.18653/v1/P17-1081.

[26] Stokel-Walker, C.—Van Noorden, R.: What ChatGPT and Generative AI Mean for Science. Nature, Vol. 614, 2023, No. 7947, pp. 214–216, doi: 10.1038/d41586-023-00340-6.

[27] Mann, D. L.: Artificial Intelligence Discusses the Role of Artificial Intelligence in Translational Medicine: A JACC: Basic to Translational Science Interview with ChatGPT. JACC: Basic to Translational Science, Vol. 8, 2023, No. 2, pp. 221–223, doi: 10.1016/j.jacbts.2023.01.001.

[28] Parikh, P. M.—Shah, D. M.—Parikh, K. P.: Judge Juan Manuel Padilla Garcia, ChatGPT, and a Controversial Medicolegal Milestone. Indian Journal of Medical Sciences, Vol. 75, 2023, No. 1, pp. 3–8, doi: 10.25259/IJMS_31_2023.

[29] Aggarwal, A.—Mittal, M.—Battineni, G.: Generative Adversarial Network: An Overview of Theory and Applications. International Journal of Information Management Data Insights, Vol. 1, 2021, No. 1, Art. No. 100004, doi: 10.1016/j.jjimei.2020.100004.

[30] Zhuo, T. Y.—Huang, Y.—Chen, C.—Xing, Z.: Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. CoRR, 2023, doi: 10.48550/arXiv.2301.12867.

[31] van Dis, E. A. M.—Bollen, J.—Zuidema, W.—van Rooij, R.—Bockting, C. L.: ChatGPT: Five Priorities for Research. Nature, Vol. 614, 2023, No. 7947, pp. 224–226, doi: 10.1038/d41586-023-00288-7.

[32] Sebastian, S. R.—Babu, B. P.: Are We Cyber Aware? A Cross Sectional Study on the Prevailing Cyber Practices among Adults from Thiruvalla, Kerala. International Journal of Community Medicine and Public Health, Vol. 10, 2022, No. 1, pp. 235–239, doi: 10.18203/2394-6040.ijcmph20223550.

[33] Aydin, Ö.—Karaarslan, E.: OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. In: Aydın, Ö. (Ed.): Emerging Computer Technologies 2. İzmir Akademi Dernegi, 2022, pp. 22–31, doi: 10.2139/ssrn.4308687.

[34] Sebastian, G.: Cyber Kill Chain Analysis of Five Major US Data Breaches: Lessons Learnt and Prevention Plan. International Journal of Cyber Warfare and Terrorism (IJCWT), Vol. 12, 2022, No. 1, pp. 1–15, doi: 10.4018/IJCWT.315651.

[35] Gao, C. A.—Howard, F. M.—Markov, N. S.—Dyer, E. C.—Ramesh, S.—Luo, Y.—Pearson, A. T.: Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers. npj Digital Medicine, Vol. 6, 2023, No. 1, Art. No. 75, doi: 10.1038/s41746-023-00819-6.

[36] OpenAI: ChatGPT. 2024, https://openai.com/blog/chatgpt/.

[37] Kung, T. H.—Cheatham, M.—Medenilla, A.—Sillos, C.—De Leon, L.—Elepaño, C.—Madriaga, M.—Aggabao, R.—Diaz-Candido, G.—Maningo, J.—Tseng, V.: Performance of ChatGPT on USMLE: Potential

for AI-Assisted Medical Education Using Large Language Models. PLOS Digital Health, Vol. 2, 2023, No. 2, Art. No. e0000198, doi: 10.1371/journal.pdig.0000198.

[38] Sardana, D.—Fagan, T. R.—Wright, J. T.: ChatGPT: A Disruptive Innovation or Disrupting Innovation in Academia? Journal of the American Dental Association (JADA), Vol. 154, 2023, No. 5, pp. 361–364, doi: 10.1016/j.adaj.2023.01.001.

[39] Jiao, W.—Wang, W.—Huang, J.—Wang, X.—Shi, S.—Tu, Z.: Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine. CoRR, 2023, doi: 10.48550/arXiv.2301.08745.

[40] Lund, B. D.—Wang, T.: Chatting About ChatGPT: How May AI and GPT Impact Academia and Libraries? Library Hi Tech News, Vol. 40, 2023, No. 3, pp. 26–29, doi: 10.1108/LHTN-01-2023-0009.

Xiaokui Liu is Associate Professor at the Key Laboratory of Oracle Information Processing, Ministry of Education, Anyang Normal University, Anyang, China. He received his B.Sc. in computer science and technology from the Henan Normal University in July 2004, and his M.Eng. in computer technology from the Hebei University of Engineering in June 2013. From 2004 to 2011 he served as Lecturer at the School of Software, Anyang Normal University; from 2012 to 2019 he was Head of the Network Information Division at the Network and Informatization Center of Anyang Normal University; and since 2019 he has served as Director of the Data Center at the Key Laboratory of Oracle Information Processing (Ministry of Education), Anyang Normal University. His research interests include data mining, information processing, natural language processing, large language models, and edge intelligence. He has published 15 academic papers, 2 academic monographs, directed 16 research projects, holds 3 patents, and has received 3 academic awards.

Wenting Wu is Lecturer at the College of Foreign Languages, Anyang Normal University, Anyang, China. She received her B.A. in English from the Shanxi Normal University in July 2004, and her M.A. in English language and literature from the Henan University in May 2011. Since 2004 she has been a teacher at the College of Foreign Languages, Anyang Normal University. Her research interests include translation studies and cultural communication. She has published 4 academic papers, 2 academic monographs, participated in 6 research projects, holds 1 patent, and has received 5 academic awards.