Proceedings of CONF-CDS 2025 Symposium: Application of Machine Learning in Engineering
DOI: 10.54254/2755-2721/2025.AST26333

Application and Comparative Analysis of Mainstream Large Language Models in Chinese Natural Language Processing Tasks

Zheming Xu

School of Information Engineering, Chang'an University, Xi'an, China
17767196938@163.com

Abstract. The scale leap of large language models has promoted the development of natural language processing these days, but Chinese NLP still faces special challenges, including word segmentation difficulties and cultural complexity like idioms and allusions. Also, systematic comparative analysis of mainstream models' performance in Chinese applications remains limited. This study compares and analyzes five representative large language models—GPT, LLaMA, Claude, ERNIE, and Kimi—across three core Chinese NLP tasks: text generation, machine translation, and text analysis. The research reveals performance and specializations of these models through evaluation of their applications in real-world scenarios. ERNIE demonstrates great cultural understanding and Chinese-specific content generation, while GPT tends to perform better in creative diversity. Claude shows excellent analytical precision and safety standards, LLaMA provides valuable customization flexibility through its open-source nature, and Kimi specializes in long-form document processing. The results reveal that different models have varying performance patterns across different application domains. Model selection requires balancing cultural factors, technical performance, and deployment constraints. These findings advance model selection strategies for Chinese NLP tasks.

Keywords: Large language models, Chinese natural language processing, Text generation, Machine translation, Text analysis

1. Introduction

Over the past few years, the development of deep learning has enabled large language models to scale up dramatically, revolutionizing natural language processing. The Transformer architecture, introduced in 2017 [1], provided the foundation for models like BERT and GPT, which have achieved significant improvements in language understanding and generation capabilities. In particular, the release of ChatGPT has pushed large language models into the public eye, showing their strong abilities in handling diverse natural language processing tasks.

However, unlike English, Chinese natural language processing faces some unique challenges. For example, Chinese lacks clear word boundaries, making word segmentation a fundamental challenge [2]. Also, the polysemous nature, flexible word order, and cultural complexity like idiom and

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.AST26333

allusions further increase semantic understanding difficulty. With the explosive growth of Chinese Internet content and the increasing importance of the Chinese language market, the development of efficient Chinese natural language processing (NLP) systems is of great theoretical significance and application value.

There is a lack of systematic comparative analysis on the performance of these models on Chinese NLP tasks [3]. Existing research mostly focuses on the performance evaluation of individual models or compares them only for specific tasks, and lacks a comprehensive evaluation of mainstream large language models in multiple Chinese NLP tasks. This research status makes it difficult to provide effective guidance for model selection in practical applications.

Based on this, this paper systematically compares and analyzes the application effects of current mainstream large language models in Chinese natural language processing tasks. Specifically, this study selected GPT, LLaMA, Claude, ERNIE, and Kimi as representative models to conduct a comprehensive evaluation across three core Chinese NLP tasks: text generation, machine translation, and text analysis. The models were analyzed across performance, limitation, and applicability scenarios. This study aims to provide empirical evidence and guiding suggestions for model selection and optimization in Chinese NLP applications.

2. Introduction to the mainstream large language models

2.1. GPT (OpenAI)

The GPT family is a generative pre-trained language model developed by OpenAI. GPT-1 first proposed the two-stage training paradigm of "unsupervised pre-training+supervised fine-tuning" in 2018, and laid the foundation for subsequent development with 117 million parameters [4]. And the evolution from GPT-2 (1.5 billion parameters) to GPT-3 (175 billion parameters) marked a significant breakthrough in few-shot learning capabilities [5].

In terms of Chinese processing, the early GPT models were mainly optimized for English, with limited support for Chinese. GPT-4 significantly improves the multilingual processing capability and performs well on tasks such as Chinese-English translation, and the latest version further optimizes the efficiency of Chinese Token processing.

2.2. LLaMA (Meta)

The LLaMA family is an open source base language model developed by Meta, released in 2023 and continuously iterated. LLaMA-1 contains four versions, 7B, 13B, 30B, and 65B, and LLaMA-2 further releases versions 7B, 13B, and 70B, as well as the conversation-optimized LLaMA-2-Chat family [6]. These models are trained entirely on publicly available datasets. Specifically, LLaMA-2 improved with 40% more training data, extended context length to 4096 tokens, and enhanced inference efficiency.

In terms of Chinese processing, the original LLaMA model is mainly optimized for English, with relatively limited Chinese support. For this reason, the open source community has developed several Chinese-optimized versions to improve Chinese processing capability by expanding the Chinese vocabulary and fine-tuning LoRA with Chinese instruction data.

2.3. Claude (Anthropic)

Developed by Anthropic, Claude is a conversational AI assistant. The family of models uses an innovative "Reinforcement Learning with AI Feedback" (RLAIF) technique, which guides model

DOI: 10.54254/2755-2721/2025.AST26333

behavior through a set of ethical principles defined in advance and reduces reliance on human feedback [7]. By adopting this approach, Claude is able to provide detailed reasoning when dealing with sensitive issues like euthanasia and historical problem, instead of simply avoiding them.

When processing Chinese, Claude shows a strong ability to understand context and provide suitable responses, while maintaining its commitment to safety and usefulness.

2.4. ERNIE (Baidu)

Baidu has developed ERNIE, a series of large language models based on its ERNIE (Enhanced Representation through Knowledge Integration) framework. ERNIE 1.0, released in 2019, incorporated knowledge graph entities into pre-training, which improved performance on tasks such as natural language interfaces and named entity recognition [8]. Subsequent ERNIE 2.0 adopted a continuous multi-task learning framework, and ERNIE 3.0 introduced a hundred billion parameter scale, both of which excelled in Chinese comprehension and knowledge inference. Later iterations (ERNIE4.0/4.5, 2023-2025) further scaled parameters and introduced turbo variants.

In March 2023, Baidu formally released ERNIE Bot, the first user-facing conversational AI product of the series, which demonstrated powerful capabilities in literary creation, business copywriting, mathematical reasoning, Chinese comprehension, and multimodal generation.

2.5. Kimi (Moonshot AI)

Kimi is a family of large language models developed by Dark Side of the Moon, known for its ultra-long context processing capabilities. Originally released in October 2023, the model supports a context window of 200,000 Chinese characters and was further expanded to 2 million characters in March 2024.

Kimi k1.5 achieved a score of 77.5 in the AIME math competition and 94th percentile in the Codeforces programming competition, comparable to the performance of OpenAI's o1 model [9]. Kimi is also multimodal, capable of handling multiple input forms such as text, images, and audio.

3. Applications and cases of mainstream large language models in Chinese natural language processing tasks

3.1. Text generation and dialog system

Text generation and dialogue systems represent fundamental applications of large language models in Chinese natural language processing, requiring models to produce coherent, contextually appropriate Chinese content while maintaining conversational flow and cultural sensitivity.

GPT has been widely adopted for Chinese content creation, with platforms like Rednote content creators using it to generate product reviews and lifestyle articles. Many Chinese marketing agencies employ GPT for social media campaigns and advertising copy generation. LLaMA powers several open-source Chinese writing assistants, including customized versions for legal document drafting in Chinese law firms and technical manual creation for manufacturing companies [10]. Claude is extensively used in educational technology, with platforms like VIPKID integrating it for personalized Chinese tutoring sessions and homework assistance. ERNIE demonstrates particular strength in Chinese language processing applications, with research showing its superior performance in Chinese journalism contexts, making it particularly effective for news content generation and editorial tasks [11]. ERNIE 4.0 powers Baidu Wenku’s AI features for instant document summarization and outline generation. Kimi has been used for generating comprehensive

DOI: 10.54254/2755-2721/2025.AST26333

Chinese research reports by consulting firms like Zero One Think Tank. Also, it is deployed by news organizations such as The Paper for creating detailed analytical articles.

In customer service applications, Taobao and JD.com rely on these models to answer millions of customer questions in Chinese each day. They help the customers find the appropriate products, track their orders, and assist in troubleshooting. The dialogue systems based on these models are also used by educational institutions like Tsinghua University for grading and providing feedback on Chinese essays.

3.2. Machine translation (Chinese and English bidirectional translation)

Machine translation between Chinese and English is now a major use case for large language models, and different models each have their own strengths in dealing with cross-language challenges.

GPT has been widely implemented in translation services, with companies like DeepL and Microsoft integrating GPT models for document translation and real-time communication tools. International law firms frequently use GPT for translating Chinese legal contracts and patents into English, while global consulting firms employ it for translating Chinese market research reports and business proposals. LLaMA powers several open-source translation platforms, including customized versions for academic paper translation used by Chinese universities and technical documentation translation for multinational corporations operating in China [12].

Claude has gained popularity in educational settings, with language learning platforms like HelloTalk incorporating it for contextual translation assistance and explanation of cultural nuances. ERNIE shows superior performance in Chinese-to-English translation for culturally-specific content, particularly in translating Chinese idioms, classical literature, and traditional cultural references that require deep cultural understanding. Baidu Translate uses ERNIE 4.5 to enhance its document-level translation capabilities. Kimi excels in translating lengthy technical documents and research papers, with pharmaceutical companies using it for translating Chinese clinical trial reports and regulatory documents into English.

Financial institutions like Bank of China and ICBC utilize these models for translating financial reports and regulatory documents between Chinese and English, enabling better communication with international stakeholders. News organizations such as Xinhua and China Daily employ translation models for real-time news translation, while e-commerce platforms like Alibaba integrate them for product description translation to facilitate international trade [13].

3.3. Text analysis

Text analysis capabilities of large language models have found extensive applications in Chinese content processing, with each model showing distinct strengths in semantic understanding, information extraction, and document analysis tasks.

It is noteworthy that major Chinese media organizations like People's Daily and Xinhua News, are utilizing GPT for generating news digests and categorizing articles. LLaMA has been explored by Chinese academic institutions such as the Evidence-Based Medicine Center at Lanzhou University to assist with systematic literature reviews and meta-analyses, particularly for tasks such as paper screening, data extraction, and comprehensive literature synthesis [14]. Claude demonstrates strong analytical capabilities in legal document review, with Chinese law firms employing it for contract analysis and regulatory compliance checking. ERNIE supports content

DOI: 10.54254/2755-2721/2025.AST26333

analysis features in Baidu's search engine, providing semantic understanding for query processing and content categorization.

E-commerce platforms such as Tmall and Taobao extensively utilize these text analysis capabilities for customer satisfaction prediction, product review sentiment analysis, and multimodal data processing to enhance user experience and business intelligence [15]. Government agencies employ these models for policy document analysis and public opinion monitoring, while healthcare institutions use them for medical record analysis and clinical decision support.

4. Model comparison and analysis

4.1. Performance comparison

Some distinctive strengths and weaknesses can be observed through systematic analysis of these models' applications in text generation, machine translation, and text analysis.

In Chinese text generation capabilities, GPT shows excellent creative writing and content diversity. The model also excels in crafting persuasive marketing copy and engaging social media content. ERNIE has stronger performance in culturally-specific content generation, particularly in traditional Chinese poetry. Kimi stands out in long-form content generation.

For machine translation tasks, generally speaking, GPT and Claude demonstrate comparable accuracy in Chinese-English bidirectional translation. GPT has slight advantages in creative content translation while Claude is better in maintaining logical consistency and safety standards. LLaMA's performance depends on fine-tuning, and custom versions often do well in legal and technical translation.

In Chinese text analysis applications, ERNIE, with its deep integration of Chinese linguistic resources, leads in Chinese semantic understanding and cultural context recognition. Kimi does well with long analytical documents that require consistent coherence. Also, it can address complex research reports effectively. Claude has better performance in thinking logically and extracting key information, while LLaMA's open-source nature allows researchers to adapt it more freely for academic research purposes.

Notably, there are differences in Response speed and computational efficiency. ERNIE and Kimi, both optimized for Chinese, tend to respond faster to Chinese queries, while GPT and Claude are steadier across languages but usually require more computing resources.

4.2. Application scenarios and limitations

GPT is highly effective at creative tasks like social media marketing, advertising, and communicating in multiple languages. However, it tends to fall short when faced with tasks that require a deep understanding of Chinese culture, such as understanding classical literature, idioms, or local humor.

LLaMA performs well in academic research settings, whether it's helping with systematic literature reviews, analyzing academic papers, or being fine-tuned for specialized tasks like processing legal documents. Although its open-source design allows customization, effective use in Chinese often requires technical skills for updating vocabulary and fine-tuning, making it less accessible to non-technical users.

Claude does especially well in professional analysis, including legal contract review, educational content assessment, and structured information extraction. Nevertheless, its conservative safety

DOI: 10.54254/2755-2721/2025.AST26333

mechanisms constrain performance in creative writing tasks, entertainment content generation, and culturally-expressive applications where innovation outweighs safety considerations.

ERNIE achieves optimal performance in Chinese cultural content creation, including Chinese traditional poetry generation, Chinese cultural reference interpretation, and Chinese search engine applications. However, its deployment is restricted to China's technological ecosystem, limiting international business applications and cross-border communication scenarios.

Kimi demonstrates excellence in long-document analysis, comprehensive research report generation, and multi-document synthesis tasks. However, its specialization in extended context processing results in suboptimal performance for general conversational interactions, simple query responses, and real-time dialogue applications where brevity and immediacy are preferred.

5. Conclusion

This paper mainly compares and analyzes the performance of GPT, LLaMA, Claude, ERNIE, and Kimi in Chinese NLP tasks as well as their usage scenarios and limitations. Comparative analysis shows that no single model dominates all application scenarios. Each model demonstrated specialized strengths that matched specific use cases and organizational requirements.

ERNIE demonstrates excellent cultural understanding in Chinese-specific applications, while GPT maintains versatility in a variety of creative tasks. Claude excels in analytical precision and safety standards. LLaMA provides valuable customization flexibility through its open-source nature, and Kimi performs remarkably in processing complex long-form tasks. The evaluation highlights that optimal model selection depends on balancing cultural authenticity, technical performance, and practical deployment considerations. However, the scope of this paper is limited, and it does not include other large language models such as Gemini, Falcon, T5, etc., in the comparative analysis. At the same time, there is no in-depth discussion and analysis on the principles of each model. Future models should integrate advanced Chinese-specific tokenization, culturally-aware training approaches, and adaptive fine-tuning mechanisms to achieve ideal performance. Development trajectories for future large language models point toward multimodal integration and enhanced cross-language capabilities that support natural code-switching between Chinese and other languages. More efficient architectures will democratize access to sophisticated Chinese NLP technologies, enabling broader organizational adoption.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", in Advances in Neural Information Processing Systems, 2017.
[2] Q. Li and S. Hao, "ERAS: Evaluating the Robustness of Chinese NLP Models to Morphological Garden Path Errors", in Proc. NAACL, 2025.
[3] Z. Qiu et al., "CLongEval: A Chinese Benchmark for Evaluating Long-Context Large Language Models", in Findings of EMNLP, 2024.
[4] A. Radford et al., "Improving Language Understanding by Generative Pre-Training", OpenAI, 2018.
[5] T. Brown et al., "Language Models are Few-Shot Learners", in Proc. NeurIPS, 2020.
[6] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models", arXiv preprint arXiv: 2307.09288, 2023.
[7] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback", arXiv preprint arXiv: 2212.08073, 2022.
[8] Z. Zhang et al., "ERNIE: Enhanced Language Representation with Informative Entities", in Proc. 57th Annual Meeting of the Association for Computational Linguistics, 2019.
[9] Kimi Team, "Kimi k1.5: Scaling Reinforcement Learning with LLMs", arXiv preprint arXiv: 2501.12599, 2025.
[10] Q. Huang et al., "Lawyer LLaMA Technical Report", arXiv preprint arXiv: 2305.15062, 2023.

DOI: 10.54254/2755-2721/2025.AST26333

[11] M. Li et al., "NewsBench: A Systematic Evaluation Framework for Assessing Editorial Capabilities of Large Language Models in Chinese Journalism", Proceedings of ACL 2024, 2024.
[12] H. Calzi Kleidermacher et al., "Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers", arXiv preprint arXiv: 2502.17882, 2025.
[13] M. Nunziatini, "Machine Translation in the Financial Services Industry: A Case Study", Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks, 2019.
[14] X. Luo et al., "Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses", J Med Internet Res, 2024.
[15] X. Zhang et al., "Research on Multimodal Prediction of E-Commerce Customer Satisfaction Driven by Big Data", Appl. Sci., 2024.