Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27277

The Evolution and Frontier Exploration of Chatbots Driven by Natural Language Technology

Chuanjia Lu

Yangzhou High School, Yangzhou, China
qwqchloelu1101@gmail.com

Abstract. As an intelligent system taking natural languages as its medium of interaction, the development of chatbots vividly reflects the development process of natural language technology. From those earliest programs that could only carry out simple question-and-answer procedures to the modern intelligent assistants that can engage in multiple turn complex conversations, each advance of chatbots has closely correlated with the transformations in natural language understanding and creation technologies. This paper divides it into 4 phases according to the time order, which are the rule-driven phase, the statistics learning phase, the deep learning phase, and the large language model phase. Detailed explanation for major technical characteristics, typical applications and technical restrictions of every stage. With some specific experiments and applications, it shows how chatbots start to overcome problems such as unclear languages and context dependencies. Finally, based on the trends of technology, this paper makes a guess of where the current chatbots are heading towards general intelligent interaction, and it attempts to explore those challenges today's technologies and tomorrow's technologies may have to meet and it is intended to provide some reference to help more people know about the technology in this aspect.

Keywords: Chatbots, Natural language process, Large language models

1. Introduction

In the digital era, the naturalization of human-computer interaction has become an important direction for technological development, where chatbots are the core platform for realizing such an experience. From intelligent assistants on mobile phones to customer service systems on e-commerce platforms, chatbots communicate with humans using natural language to meet their information inquiry and task cooperation needs. Their basic improvement depends on the breakthrough of natural language technology, enabling machines to understand the meaning of human language, remember human conversations, and produce answers that are appropriate for those conversations. The development of natural language technology arises out of the necessity in the age of explosive information to process a large volume of unstructured language data efficiently; it is also an evolution fitting with human-computer interaction trends moving from command-based operation to conversational exchange. , its core essence is not just about driving artificial intelligence to move beyond its tool nature into human-like thinking capacity, but it can also be a

© 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.LD27277

bridge in communication between different languages and disciplines, it can break different language barriers and help people from different cultural backgrounds exchange information, as well as various types of industry extracting values from text data, for example, extracting medical records from the medical field and interpreting policies in the financial field. At the same time, it also offers a new research paradigm for disciplines such as linguistics and cognitive science. The machine analyzes language patterns, so people deepen their understanding of themselves through their language patterns and ways of thinking. It has become a cornerstone for collaboration between humanity and artificial intelligence in the digital era.

This article will be conducted as follows: The second part looks at the development history of chatbots, detailing how they changed through the first few stages of rule-driven systems, second generation models that use stat learn, third gen intelligent interaction via DL, and fourth gen more general convos with pre-tuned big models. And combined with technical milestone, it will conduct a detailed investigation of the key features and breakthrough points in semantic understanding, context modeling, response creation, etc., in each stage of natural language technology. It clearly shows the internal logic of technology development. The third part is about verifying the effect of the technology through a large number of experimental cases, using real scenes' dialogue datasets like customer services, intelligent assistants' dialogue, etc. It also evaluates the actual effect of key technologies from aspects of response accuracy, coherence of conversation context, user satisfaction, etc. In the fourth part, I deeply explore those core challenging problems still existing in this field today with semantic ambiguity problem in complex sentences, deep understand emotion intentions, shortcoming of multi modal integration and combining with hottest research trends to imagine future directions in multimodal integration interaction, knowledge boost and facts check, personalized memory mechanism. Finally, summarize all the ideas of the whole article, condense the laws of the evolution of technology and its practical use, and provide references for the relevant fields of research and application.

2. The development stages of chatbots

2.1. The stage of rule-driven

The chatbots at this stage can only do simple talking through certain template. ELIZA was developed by Weizenbaum from the Massachusetts Institute of Technology in 1966. It marks the first chatbots. This system's main principle is that its response is generated through keyword matching to trigger a preset questioning template - when users type "mother" "sadness", it activates a template to change it to a question template, changing "I miss my mother" to "Why do you miss your mother? " [1]. This design does not require knowledge of the deep meaning of the language, just communication on the surface level, but at the time, that led some people to believe they were talking to a person.

In 2001, the emergence of AIML (Artificial Intelligence Markup Language) simplified the process of rule writing. Developers used the <pattern> tag to define input patterns and the <template> tag to specify corresponding responses, thereby creating a dialogue library containing tens of thousands of rules. Typical applications include chatbots like Eliza, which can perform basic interactions such as greetings and weather inquiries within preset limits [2]. But there were also obvious shortcomings at this stage, rules had to be written one by one manually, which was not scalability enough; when encountering unpredictable inputs, such as "Today's sunset was very beautiful," it can only reply "I don't know," which is clearly inflexible.

DOI: 10.54254/2755-2721/2025.LD27277

2.2. The stage of statistical learning

With the construction of large corpora, and with the advent of statistical ML, chatbots can now enter into a data- driven stage. This period's core technology is using a large number of conversations to teach the computer how language works with probabilistic techniques, no manual rules. For example, if it is a customer service situation, the system will calculate the probability of the association of keywords such as "return" and "refund" with the "return process consultation" intention, and it can generate the most probable matching results and respond [3].

Statistical learning techniques have also led to the first try for sentiment analysis; by creating a sentiment dictionary and classification model, machines can identify emotional trends in user speech (happy, sad), and change the response tone. For example, when the system detects that the user input contains negative words such as "disappointed" or "unsatisfied", it will first generate a calming answer [4]. But the technology itself is not strong enough yet and cannot yet model very complex semantics or very long contexts, so it's still quite common for irrelevant answers in conversations.

2.3. The stage of deep learning

After 2012, the use of deep learning technology in the fields of natural language processing brought about a revolution. Chatbots now step into an "all at once" learning period. In contrast to statistical learning, which relies on artificial features, neural networks can automatically learn language's features through multiple layers of nonlinear transformations, and achieve more accurate semantic understanding and dialogue modeling. The core breakthrough of this transformation is the creation of the sequence-to-sequence (Seq2Seq) model. Sutskever et al designed the Encoder-Decoder architecture [5], the model can generate the output response sentence directly from the input text (encoding 'How is the weather today? ' To a sematic vector and the decoder will generate 'Today is sunny and good for travel') without manual design of middle features This end-to-end way of doing things showed its own strong points when it came to machines turning one language into another, and helped make chats between computer programs feel natural.

The introduction of a mechanism for attention by Bahdanau et al., which is more of a problem with long sequences [6]. For multi-round conversations, the model can pay attention to important information in the historical dialogue. For example, if the user has mentioned previously that they like science fiction movies, then the attention mechanism will cause the robot to give more importance to this information while giving recommendation, avoiding any kind of disconnect with the previous statement and making the conversation much more coherent and relevant.

2.4. The stage of large language models

The proposed Transformer model in 2017 was a revolutionary jump to natural language technology [7]. The emergence of large language models (LLMs) led to the chatbots entering the phase of universal intelligent conversations.

Big language models are first trained ahead of time on big chunks of text so they can learn common language patterns and information about the actual world, and then adjusted for particular situations. The typical models are GPT series, LLaMA, PaLM. The chatgpt released in 2022 indicates the maturity of this period which improves the level of natural interaction as well as its monthly active user reached over 1 billion after release [8]. In 2024, GPT-4 was produced, which could do multimodal interaction, which could understand text and pictures coherently, and further expanded the application boundary [9].

DOI: 10.54254/2755-2721/2025.LD27277

3. Typical experiments and application verification

3.1. Early explorations of the rule and statistics stage

In the ELIZA experiment carried out in 1966, Weizenbaum enables people to consult with a psychotherapy machine. The results showed that 30% of the users felt like they were talking to a real person which shows that it was able to match to rules to have some surface level conversations [1]. But after that, the system showed that it had an incomplete semantic understanding, when the input didn't conform with the given rules (for instance, The weather is really good today), the system answered "Please continue".

In 2001, AIML was applied to commercial customer service robots, where it was found that the matching accuracy for common questions like “how to change your delivery address” was around 85%, but for long-tail questions such as “can overseas addresses be changed?”, the matching accuracy dropped to as low as 30%, and the cost of expanding rules was enormous [2]. In the statistical stage of customer service robot testing in e-commerce situations, the accuracy of intents rose to 75% yet there was still some difficulty in comprehensive meaning comprehension [3].

3.2. Breakthroughs in the era of deep learning and large models

In 2016, Google carried out the experimental study about emotional dialogue modeling by DailyDialog dataset, which provided significant empirical support with respect to the benefit of neural networks in emotional interaction. This data set was the first to be annotated with emotions for a multi-round dialogue at scale when this was first done; it included more than 10,000 sets of every day dialogues, which covers the seven basic emotions and labels both the level of emotion in a sentence and the label of the entire dialogue, forming a reference for evaluating models. The experiment used an emotional dialogue model based on LSTM: The encoder uses LSTM to grasp the emotional flow of the dialogue history, and the decoder has emotional embedding vectors to make the answers have consistent emotions. The results show that the emotional consistency rate of the generated response is 78%, which is 23% higher than that of the statistical model. These datasets have very much demonstrated the advantage of neural networks’ benefits to emotional interactions [10]. However, the emotion consistency of this model in long conversations >5 rounds is sharply reduced to 58%, which shows that early neural networks have obvious problems in context modeling - the "long-range dependence forget" problem of LSTM made early emotional information diluted little by little as the number of rounds of conversation increased, making it difficult to keep emotion consistent.

ChatGPT has been applied to many fields and shown great practical value. In the school learning environment, students use it a lot when revising their papers to help them check for logical gaps, grammar mistakes, and even adapt the expression style based on different subjects' rules, like changing colloquial expressions to academic ones if needed. According to the data from an educational platform, it is found that the use of this teaching aid makes student revision of papers more efficient by about 40%, and the repetition rate drops by around 15-20%. In the medical field, ChatGPT can be smart with medical records: by using technologies such as text named information recognition and relationship extraction, ChatGPT can extract important information like patients' age, symptoms, and previous medications from unnecessary handwritten or audio medical records, and then display the information in a structured way. The time required to compile a single medical file has dropped from 15 minutes to three minutes, while the accuracy of the extracted information is as high as 92%, meaning a lot fewer human errors.

DOI: 10.54254/2755-2721/2025.LD27277

As for the customer service scenario, the application of ChatGPT directly improved the solving efficiency. Traditional customer service based on fixed FAQ matching, the handling of complex issues was less than 50%. But ChatGPT could understand deeply of the needs of users (such as how to issue it again after canceling the order) and generate its own solutions. From the data we can see that after integrating ChatGPT, a certain e-commerce platform improved the problem-solving rate to +40% [8], the proportion of the first time solving user's problems increased from 58% to 98%, and the average number increased from 80 to 150 pieces per day.

In 2024, it showed off again by doing really good jobs in professional areas, further showing how great big models are at having really deep conversations. the USMLE, its pass rate is 85%; the USMLE includes three parts: basic medicine, clinical knowledge, and clinical skills, making it the exam with the hardest questions. The passing rate of human candidates is around 73%. GPT-4 not only can correctly memorize the diagnostic criteria of rare diseases, but also can gradually understand the reason through the details in the case description (such as “fever 3 days, history of mosquito bites”). And can also provide a variety of different options, approaching the dialogue exchange of a professional doctor. It has fully proven its practical value in professional conversation [9].

4. Current challenges and future prospects

4.1. Existing technical challenges

Model is obviously poor at the context understanding. Extremely long conversations often lose some key information from the previous context, and when identifying implicit intentions (e.g., irony, metaphor), its accuracy rate is less than 60% [11].

Large language models also face problems related to facts and truthfulness and have the tendency to produce false information. Even GPT-4 makes mistakes with about 15-20% of questions on factual Q&A tests [12].

These limits are from a bunch of earlier studies. Early work looked at the context processing that language models do. designing multi-round conversation datasets for example MultiWOZ the long conversation subset of DailyDialog researchers systematically evaluate the model's information retention capability under different lengths of conversations, when the number of conversation rounds exceeds 10 the model's recall rates of key information will sharply decline. This gave me a starting point to study loss of information in very long talks. As for the implicit intent recognition, the researcher created special datasets including complex semantics like irony or metaphor (for example, Irony Detection Dataset, Metaphor Corpus), and after comparing the human annotations with the model’s predictions, set up the problem’s difficulty level and also pointed out how poorly the model dealt with non-literal meaning.

In the study of factual accuracy, we mainly need to carry out evaluation standards and quantitative analysis of the behaviors of models. Researchers have made the evaluation process of factual correctness more systematic by creating huge sets of factual question and answers. This comprises of TruthfulQA, FEVER, which covers different domains like historical facts, knowledge about science, and common-sense judgments. early models like GPT-3 have error rates in these benchmark tests generally higher than 30%, so later studies mainly aimed at how to reduce the production of false content. The 15%-20% error rate of GPT-4 can be directly seen from the comparisons of those prerequisite tasks, which has become an important basis for people to measure the factual trustworthiness of the current models.

DOI: 10.54254/2755-2721/2025.LD27277

4.2. Future direction of development

Multimodal fusion interaction intends to fuse different modalities like sight and sound into a whole, so as to realize the understanding and generation of 'language + picture + sound'. Like if someone uploads a photo of a certain dish and asks for the steps to cook it, this model could mix what it sees in photos with some recipe facts to tell how to cook that food better [13].

Knowledge enhancement and fact checking rely on the retrieval-augmented generation (RAG), which connects Wikipedia and its own knowledge store. Generate a response, the model will check the correctness of the facts before it, so as to avoid generating incorrect content [14].

Personalized personal memory mechanism focuses on developing a dynamic way to update user's information, it records all the preferences, history of interactions and the required scene in order to make "one-to-one" personalized conversations possible. For example, when interacting with a high school student user, the model gives a recommendation for scientific popularization content appropriate to a high school student's educational level [15].

5. Conclusion

The development process of chatbots represents the development process of natural language. From the rule-based chat templates, to general conversations with large language models. Each breakthrough is accompanied by a better understanding of the meaning and essence of language. The rule stage set the stage for interaction, the statistical stage introduced data-driven thinking, the deep learning stage achieved breakthroughs in semantic modeling, and the large language model stage opened the possibility for general intelligence.

Chatbots now still struggle when it comes to understanding context, being authentic and personal. But multi-modal fusion methods, knowledge improvement, individual memory and other technological routes give the idea of direction. With the deep integration of natural language and artificial intelligence technologies, Chatbots are to play a larger role in areas like education, healthcare, and the provision of life service offerings; they have the potential to become an essential part of how people connect with the digital world.

References

[1] Weizenbaum, J. (2025). The Technical Architecture and Philosophical Reflections of ELIZA: A Study on the Reproduction of the 1966 Dialogue Script. Journal of Artificial Intelligence Research, 82, 456-489.
[2] Wallace, R. (2019). AIML 2.0 Standard and Dialogue System Design. IEEE Transactions on Affective Computing, 10(3), 456-470.
[3] Huang, M., et al. (2020). Intent Recognition Based on Hidden Markov Model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). ACL 2020 Oral Presentation, Vancouver, Canada
[4] Komeili, M., et al. (2023). The Phenomenon of Semantic Drift in Long Conversations. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). ACL 2023 Long Paper, Toronto, Canada
[5] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. NeurIPS. Advances in Neural Information Processing Systems (NeurIPS). NeurIPS 2014 Oral Presentation, Montréal, Canada
[6] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. ICLR. International Conference on Learning Representations (ICLR). ICLR 2015 Oral Presentation, San Diego, USA
[7] Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS). NeurIPS 2017 Spotlight Presentation, Long Beach, USA
[8] Evans, O., et al. (2023). Truthful AI: Addressing Hallucination. Nature Machine Intelligence, 5(7), 689-701.
[9] OpenAI. (2023). GPT-4 Technical Report

DOI: 10.54254/2755-2721/2025.LD27277

[10] Li, S., et al. (2017). DailyDialog: A Multi-round Dialogue Dataset. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). ACL 2017 Short Paper, Vancouver, Canada
[11] Komeili, M., et al. (2022). Contextual Understanding in Long Conversations. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). ACL 2022 Long Paper, Dublin, Ireland
[12] Radford, A., et al. (2021). Learning Transferable Visual Models. Advances in Neural Information Processing Systems (NeurIPS). NeurIPS 2021 Oral Presentation, Virtual Conference
[13] Lewis, P., et al. (2023). Retrieval-Augmented Generation for LLM. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). ACL 2023 Oral Presentation, Toronto, Canada
[14] Yan Canping. (2024). Practice of Multi-modal Fusion Based on Large Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). CVPR 2024 Oral Presentation, Seattle, USA
[15] Zhang, X. F., Beauchamp, N., & Wang, L. (2025). PRIME: Large Language Model Personalization with Cognitive Memory and Thought Processes. International Conference on Learning Representations (ICLR). ICLR 2025 Oral Presentation, Paris, France