*Digitalization and Management Innovation IV*
*A. Lau and A.J. Tallón-Ballesteros (Eds.)*

*doi:10.3233/FAIA250737*

# Comparative Analysis of Chatbot Systems

Hengsheng XU[a], Linkun WAN[a], Yunyin LI[a], Jiaxi LIU[a] and Adela S.M. LAU[b,c,1]
[a] *Guangzhou College of Commerce, China*
[b] *Data Science Lab, Department of Statistics and Actuarial Science, School of Computing and Data Science, The University of Hong Kong, Hong Kong*
[c]*Dr Adela Education Limited*
ORCiD ID: Hengsheng Xu https://orcid.org/0009-0008-9351-7530
Linkun Wan https://orcid.org/0009-0002-5389-5226
Yunyin Li https://orcid.org/0009-0006-9090-2324
Jiaxi Liu https://orcid.org/0009-0004-7715-1453
Adela S.M. Lau https://orcid.org/0000-0001-5918-8309

**Abstract.** Existing research on chatbot evaluation suffers from inconsistent assessment standards, fragmented criteria, and insufficient coverage of critical dimensions like legal compliance and ethical alignment, which hinders reliable benchmarking of chatbots' performance. Our study proposes a comprehensive framework for such evaluation and systematically compares five chatbot systems: Tidio (Rule-Based), GPT-4o (AI-Powered), Claude 3.5 Sonnet (LLM), Watson Assistant (Enterprise), and Qwen2.5-Max (Multilingual) in terms of their accuracy, safety, legal compliance, generalizability of performance, and ethical alignment. We conclude that while chatbots enhance efficiency in healthcare (97.34% patient education completeness) and e-commerce (30%–40% cost reduction), critical limitations persist. Recommendations include: (1) retrieval-augmented generation (RAG) for hallucination reduction, (2) ethical governance frameworks (e.g., AILuminate), and (3) domain-specialized tuning. Cross-sector collaboration and standardized evaluations are essential for responsible deployment of AI.

**Keywords.** Chatbot Large Language Models, AI-Powered Conversational Agents, GPT-4o

## 1. Introduction

The rapid integration of chatbots and large language models (LLMs) into business and healthcare operations is transforming service delivery, driven by advances in natural language processing (NLP) and generative AI. Current applications range from customer support and patient education to legal analysis and supply chain optimization, with the global chatbot market projected to grow at a CAGR of 23.94%, reaching $61.97 billion by 2035 [1]. However, significant challenges persist, including inconsistent evaluation standards for factual accuracy, safety vulnerabilities, legal compliance gaps, and ethical risks like bias and hallucinations. This study addresses these gaps by proposing a comprehensive, multi-dimensional evaluation framework. We systematically compare

[1] Corresponding Author: Adela S.M. Lau, Department of Statistics and Actuarial Science, Run Run Shaw Building, The University of Hong Kong, Pokfulam Road, Hong Kong. Email: adelalau@hku.hk
Dr Adela Education Limited is a HKU startup and spinoff.

five representative chatbot systems: Tidio (Rule-Based), ChatGPT (AI-Powered), Claude 3.5 Sonnet (LLM), Watson Assistant (Enterprise), and Qwen2.5-Max (Multilingual) in terms of five validated criteria: factual accuracy, safety, legal compliance, generalizability of performance, and ethical alignment. Our goal is to establish objective benchmarks to guide the responsible deployment of conversational AI technologies.

## 2. Literature Review

### *2.1. Types and capabilities of chatbot and large language models on the market*

Rule-based chatbots (e.g., Lego's Sophia, Tidio) rely on pre-programmed scripts for specific queries, something which is effective in structured tasks but not in complex dialogues. AI-powered agents (e.g., Siri, ChatGPT) use ML/NLP for dynamic interactions, adapting to unstructured conversations and improving industrial productivity and fraud detection [2]. LLMs (e.g., GPT-4, Gemini) excel in text generation and multilingual tasks, with Gemini outperforming peers in readability ($p < 0.05$), accuracy (80%), and quality (80%) [3]. Model inspection frameworks are crucial to mitigate shortcut learning.

### *2.2. Review of methods (criteria and parameters) for comparing large language models*

Factual accuracy: Benchmarks (e.g., FELM) [1] annotate domains (knowledge, science, math) using F1/accuracy metrics. Safety: SafetyBench (11,435 MCQs) and S-Eval [4] (220k+ prompts) test hate speech/violence robustness. Legal Compliance: Tools (e.g., LinksAI) [5] assess citation accuracy/ hallucination minimization, requiring interpretability. Generalization: Cross-domain tests measure adaptation speed/efficiency. Ethical Alignment: Audits evaluate bias/fairness via quantitative indicators [6].

### *2.3. Large language models and their comparison*

GPT-4o excels in math/coding but has high jailbreak vulnerability [4]. Claude 3.5 Sonnet performs well in data extraction yet faces security risks, while adhering to GDPR/HIPAA for legal compliance. Qwen2.5-Max dominates Chinese tasks, especially legal document analysis, aligning with local frameworks. All models exhibit safety concerns, though Claude demonstrates the best regulatory alignment.

### *2.4. Current Applications of Chatbots and Potential Applications*

NLP/generative AI advances drive chatbot integration for efficiency/scalability. Key uses: Customer service (e.g., HERA automates onboarding/tracking); Healthcare (GPT-4 excels in patient education (97.34% vs 93.61%) [7][8] but lags in complex decisions (96.88% vs 97.99%), necessitating oversight. Market projected at $61.97B by 2035 (CAGR 23.94%) [9]. Emerging: Supply chain (inventory/threat prediction) and cybersecurity. Challenges: Bias, privacy, adaptability limits. The future requires empathetic AI/self-learning balanced with expertise.

### *2.5. Summary*

Strengths: LLMs excel at natural language tasks like text generation and translation, scale effectively with data, and power human-like chatbots for customer service and education. Limitations: Prone to hallucinations, biased output, and high computational costs; struggle with reasoning, long-term context retention, and domain-specific accuracy. Future Potential: Key applications in enterprise (automated reports, legal analysis), healthcare (diagnostic support), education (adaptive tutoring), and research (literature synthesis).

## 3. Research method

### *3.1. Data Source*

This study evaluates five chatbot systems: Rule-Based (Tidio), AI-Powered (ChatGPT, GPT-4), Large Language Model (Claude 3.5 Sonnet), Enterprise Virtual Agent (Watson Assistant), and Multilingual Bot (Qwen2.5-Max, Chinese). Selection was based on popularity, public benchmarks, and industrial relevance, with testing conducted via publicly available APIs or demo interfaces.

Datasets were utilized to evaluate specific dimensions: Factual accuracy was assessed using FELM and a self-constructed medical corpus, focusing on F1 score and hallucination rate. Safety was appraised through the high-risk subset of SafetyBench, emphasizing harmful responses and jailbreak vulnerabilities. Legal compliance was examined with 30 GDPR and 20 PIPL articles, with a focus on clause coverage and misquotations. Generalizability was tested on C-Eval and MATH-500, examining disparities in the accuracy on one domain from the accuracy on other domains. Ethical alignment was evaluated through bias scenarios, with responses manually coded.

### *3.2. Measurement Instrument*

- Factual accuracy: Measured using the FELM benchmark, assessing accuracy on five domains: world knowledge, science, mathematics, medicine, and social sciences. Metrics include F1 score and balanced accuracy.
- Safety: Assessed through SafetyBench, covering 11,435 prompts across seven risk categories, and S-Eval, which measures attack resistance and content policy compliance.
- Legal Compliance: Benchmarks such as LinksAI and region-specific frameworks evaluate the adherence to GDPR, HIPAA, and Chinese legal standards, focusing on the accuracy of the citations, legal clarity, and hallucination rate.
- Performance Generalizability: Cross-domain evaluations measure adaptation to new tasks, using metrics like adaptation speed and cross-task success rate.
- Ethical Alignment: Evaluated through ethical risk prompts, assessing fairness, transparency, and bias, with both qualitative and quantitative analysis of responses.

### 3.3. Procedure

Each chatbot was tested with a standardized set of prompts from various industries (healthcare, legal, e-commerce, public services, and education). Responses were assessed using a scoring rubric based on five criteria. Quantitative metrics (response time, accuracy) were recorded automatically, while qualitative attributes (ethical tone, compliance clarity) were annotated by two researchers (Kappa ≥ 0.8). Statistical analysis used ANOVA and Tukey's post-hoc test ($p < 0.05$). This framework ensures objective, reproducible assessments for the analysis in Chapters 4 and 5. Table 1 qualitatively compares five chatbot systems in terms of five critical dimensions: technical capability, safety, legal compliance, adaptability, and ethical risks.

**Table 1.** Functional Characteristics Matrix. Provides a qualitative comparison of chatbot systems using five critical operational dimensions, highlighting their strengths and limitations in real-world deployment contexts.

| Model | Functionality | Security | Legal Compliance | Adaptability | Ethics |
|---|---|---|---|---|---|
| **GPT-4o** | Strong performance | High risk | Medium level | Strength in interdisciplinary | High risks |
| **Claude 3.5** | Medium | Highly secure | Compliant with GDPR and HIPAA | Strength in structured tasks | Medium |
| **Qwen2.5-Max** | Chinese priority | Medium | Adaptation to Chinese Law | Limited adaptability | cultural bias |
| **Tidio** | Low | Highly secure | Low | limited adaptability | No risks |
| **Watson Assistant** | Medium | Highly secure | High, compliance with IBM framework | Medium | No risks |

## 4. Results

### 4.1. Types of Chatbots and their Capabilities

The primary categories of chatbots that dominate the market include Rule-Based Chatbots (e.g., Lego's Sophia, Tidio), which operate via predefined scripts, excelling in structured workflows (e.g., FAQs, order tracking) but failing in unstructured dialogues.

Another consists of AI-Powered Conversational Agents (e.g., ChatGPT, Claude): These leverage NLP/ML for dynamic interactions, learning from user inputs to handle ambiguity. Applications include fraud detection (financial sector) and multilingual customer support.

Large Language Models (LLMs) (e.g., GPT-4o, Gemini) are advanced generative models with contextual text synthesis and cross-domain knowledge. Gemini outperforms peers in readability (80%) and response quality.

### 4.2. Evaluations

- Factual accuracy: Benchmarks like FELM assess accuracy in science, math, and world knowledge using F1 scores [1].

- Safety: Tools like SafetyBench evaluate responses in terms of harmful content, with Claude 3.5 receiving "Very Good" safety ratings, while GPT-4o shows high risk and vulnerabilities [4].
- Legal Compliance: Frameworks assess adherence to GDPR (EU), HIPAA (US), or Chinese regulations. Qwen2.5-Max excels in Chinese legal tasks [10], while other models show varying degrees of compliance [11].

### *4.3. LLM Comparisons*

- GPT-4o: Strong in technical performance but prone to hallucinations (18.7%) and shows medium GDPR compliance [3][11].
- Claude 3.5: Superior in safety and security, with strong EU GDPR and US HIPAA compliance [4][5].
- Qwen2.5-Max: Dominates Chinese language tasks and is suited for Chinese legal frameworks but shows limited adaptability [10][11].
- Tidio: Low performance in complex data processing, excels in safety and security, with no hallucinations [11].
- Watson Assistant: Specializes in customization for entrepreneurs, with high safety and security but requires customization [11].

### *4.4. Applications*

- Hallucination Rate: Tidio performs the best (0%), while GPT - 4o has the highest rate (18.7%)[4]. This shows that Tidio's content has a high level of credibility.
- Latency: Tidio is the fastest (120 ms), followed by Watson (680 ms), and GPT - 4o is the slowest (850 ms). Tidio is suitable for scenarios with high real-time requirements.
- Security (ASR): There are no successful jailbreak cases in Tidio, while Claude 3.5 and Qwen2.5 have relatively higher rates (15.6% and 33.8% respectively). Tidio is more secure and reliable.
- Compliance with Chinese legal framework: Qwen2.5 scores the highest (4.9/5) and is suitable for applications with strict Chinese legal compliance requirements [10].
- Interdisciplinary: Qwen2.5 shows the best performance (16.3%) and has great potential in cross-domain applications.

Table2 is a comprehensive comparison, providing key performance indicators of several multimodal large-scale models, aiming to provide a clear reference for understanding their characteristics and applicability.

**Table 2**. Quantitative Performance Metrics for Several Multi-Modal Large-Scale Models.

| Metric | GPT-4o | Claude 3.5 | Qwen2.5-Max | Tidio | Watson Assistant |
|---|---|---|---|---|---|
| **Hallucination Rate (%)** | 18.7 | 12.3 | 15.2 | 0.0 | 8.9 |
| **Latency (ms)** | 850 | 920 | 1100 | 120 | 680 |
| **ASR (%)** | 42.1 | 15.6 | 33.8 | 0.0 | 5.3 |
| **Compliance with Chinese legal framework** | 3.8/5 | 4.2/5 | 4.9/5 | 2.1/5 | 3.5/5 |
| **Generalizability** | 6.2 | 9.8 | 16.3 | - | 12.4 |

### *4.5. Strengths, Limitations and Future*

- Strengths: Scalable automation (e.g., multilingual support), contextual adaptability.
- Limitations: Hallucinations (Qwen: 20% error rate), bias risks, high computing costs.
- Future: Integration with IoT, safety standardization (e.g., AILuminate), and domain-specific tuning (e.g., DeepSeek-R1 for math).
- Safety Expansion: SafetyBench reveals critical LLM vulnerabilities, rendering urgent the need for industry-wide protocols.
- Data Sources: Ethical curation (e.g., ChineseSafe benchmarks) mitigates bias.

Research Priorities: Mitigate hallucinations via retrieval-augmented generation (RAG); enhance cross-lingual compliance.

## 5. Discussion

Chatbots such as HERA in Indonesia provide significant efficiency gains by automating customer onboarding, order tracking, and appointment scheduling [12], reducing reliance on human agents while maintaining conversational quality [11]. In healthcare, AI-powered chatbots expedite patient education, alleviating staff workload [7][8]. However, different models have different levels of effectiveness. For instance, GPT-4o exhibits strong generalizability, performing well on different knowledge domains (6.2%) but suffers from a high hallucination rate (18.7%) and latency (850 ms) [1], posing risks in high-stakes applications. In contrast, Claude 3.5 shows lower hallucination rates (12.3%) and robust compliance with GDPR and HIPAA [4], making it safer for sensitive data processing.

Challenges persist, particularly in complex decision-making and data privacy. While Watson Assistant excels in compliance (IBM framework) and latency (680 ms), its adaptation requires customization [11]. Qwen2.5-Max, tailored for Chinese legal frameworks (4.9/5), suffers from cultural bias [6], while Tidio, though secure and fast (120 ms latency), lacks complex data processing capabilities. Ethical concerns, such as bias in NLP models [5][6] and job displacement, remain critical. For example, GPT-4o's high ASR (42.1%) indicates potential risks of misuse [1], and Claude 3.5's lack of transparency raises accountability issues [5].

Future advances must prioritize contextual learning, reduce hallucination rates [1], and ensure ethical AI deployment [4][6] to maximize chatbots' potential across industries.

## 6. Conclusion

This study rigorously compares five chatbots, Tidio, GPT-4o, Claude 3.5, Watson Assistant, and Qwen2.5-Max, in terms of factual accuracy, safety, compliance, and ethics. GPT-4o excels technically (92% on HumanEval [3], 96.4% on MATH-500 [3]) but shows high hallucination rates (15-20%) [1] and jailbreak vulnerabilities. Claude 3.5 leads in regulatory compliance (GDPR/HIPAA) [5], [7], while Qwen2.5-Max dominates on Chinese tasks (92.2% on C-Eval) [10] but struggles cross-domain (83% on C-

SimpleQA). SafetyBench reveals universal risks [4], with all models showing bias and factual inaccuracies in specialized fields (FELM) [1].

In practice, GPT-4o achieves 97.34% completeness in healthcare education [7], while reducing e-commerce costs by 30%-40% [12]. Three proposals have been advanced for addressing these problems: (1) RAG for hallucination reduction, (2) ethical frameworks like AILuminate, and (3) domain-specific models (e.g., DeepSeek-R1). Standardized evaluations and cross-sector collaboration are crucial for responsible deployment.

## References

[1] Chen, S, Zhao, Y, Zhang, J, Chern, I, Gao,S, Liu, P, He, J. FELM: Benchmarking factuality evaluation of large language models. Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023 Dec;44502-4452, doi:10.48550/arXiv.2310.00741.

[2] Exploring agentic AI use cases and industry applications in 2025. https://techkors.com/agentic-ai-use-cases/

[3] Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal. 2024 Mar;6:100048, doi:10.1016/j.nlp.2023.100048.

[4] Zhang, Z, Lei, L, Wu, L, Sun, R, Huang, Y, Long, C, Liu, X, Lei, X, Tang, J, Huang, M. SafetyBench: Evaluating the safety of large language models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024 Jun; 15537–15553, doi:10.18653/v1/2024.acl-long.830.

[5] Doshi-Velez F, Kim F. Towards a rigorous science of interpretable machine learning, 2017, doi:10.48550/arXiv.1702.08608.

[6] Hajikhani A, Cole C. A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. Quantitative Science Studies. 2024;5(3):736–756, doi:10.1162/qss_a_00310.

[7] Bai X, Wang S, Zhao Y, Feng M, Ma W, Liu X. Application of AI Chatbot in Responding to Asynchronous Text-Based Messages From Patients With Cancer: Comparative Study. Journal of Medical Internet Research, 2025;27:e67462. doi: 10.2196/67462

[8] Büker M, Mercan G. Readability, accuracy, appropriateness, and quality of AI chatbot responses as a patient information source on root canal retreatment: A comparative assessment. International Journal of Medical Informatics. 2025;201:e105948, doi: 10.1016/j.ijmedinf.2025.105948.

[9] Markets RA. Chatbot market industry forecast report 2025-2035, with profiles of Acuvate, Aivo, Botsify, eGain, Haptik, Helpshift, Inbenta, LiveChat, ManyChat, Next IT, SmartBots, Yellow Messenger and more," GlobeNewswire, [Online]. Available: https://www.globenewswire.com/news-release/2025/5/15/3081937/28124/en/Chatbot-Market-Industry-Forecast-Report-2025-2035-with-Profiles-of-Acuvate-Aivo-Botsify-eGain-Haptik-Helpshift-Inbenta-LiveChat-ManyChat-Next-IT-SmartBots-Yellow-Messenger-and-more.html.

[10] Park, A, Lee, SB. Examining AI and Systemic Factors for Improved Chatbot Sustainability. Journal of Computer Information Systems. 2024;64(6):728-742, doi:10.1080/08874417.2023.2251416.

[11] IBM. Chatbot. 2025 Mar 12, https://www.ibm.com/cn-zh/think/topics/chatbots.

[12] HERA. How HERA, AI chatbot Indonesia is transforming customer support with real task execution. https://heracx.ai/article/how-hera-ai-chatbot-indonesia-is-transforming-customer-support-with-real-task-execution.