Scientific Journal of Impact Factor (SJIF): 7.3
E-ISSN (O): 2348-4470
P-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 13,Issue 4,April 2026

Chatbots Using Natural Language Processing for Customer Support: Architecture, Techniques, Applications, and Future Directions

Nazma A. Inamdar[1], Ganesh Anil Ghewari[2], Shaikh Nihalahamad Aslam[2]

[1]*Department of Information Technology, Government Polytechnic, Nanded*
[2]*Department of Information Technology, Walchand College of Engineering, Sangli*

Abstract — The development of artificial intelligence and natural language processing (NLP) technologies has realized a shift in the paradigm of handling customer services by companies. The paper will present a detailed description of NLP-based chatbot applications to assist customers, including the architectural paradigms, basic NLP components, deep learning algorithms, industry-related applications, performance metrics, and ethical standards. We address rule-based, retrieval-based, and generative models and specifically the transformer-based models, such as BERT, DistilBERT, GPT- 3, and DialoGPT. We discuss such problems as context retention, multilingual support, unbalanced classes and more advanced approaches, such as retrieval-augmented generation (RAG), sentiment analysis, and voice-based communication. It is a compilation of published benchmark findings, case studies in e-commerce, banking, healthcare and hospitality and analysis of key performance indicators. The paper concludes by identifying the open research challenges and future directions, such as multimodal interaction, explainable AI, and ethical AI development.

Index Terms—chatbots, natural language processing, customer support, dialogue management, BERT, transformer models, large language models, retrieval-augmented generation, conversational AI, sentiment analysis.

Keywords- chatbots, natural language processing, customer support, dialogue management, BERT, transformer models, large language models, retrieval-augmented generation, conversational AI, sentiment analysis.

I. INTRODUCTION

The rapidly growing exponential demand of online services and online commerce has placed broader requirements on scalable, responsive and intelligent customer support systems than ever before. The traditional contact centers that are operated by human beings are increasingly being strained by the rising cost of doing business, customer needs that require the centers to be open 24/7, and by the product ecosystem complexities [1]. Artificial intelligence (AI) chatbots enabled by advancements in the natural language processing (NLP) field have proven to be a highly attractive way out of such problems, enabling companies to automate a decent portion of customer interactions without losing service quality. The subdiscipline of AI known as NLP, which deals with the computational knowledge and generation of human language, offers the basis upon which chatbots are able to read user intent, extract meaningful objects, generate coherent responses, and alter conversational behavior based on the context [2]. Chatbot technology has changed radically since the pioneering ELIZA system written by Weizenbaum in 1966 [3]--no longer a pattern-matching rule engine, but a neural architecture, with open-domain conversation. The advent of the Transformer architecture by Vaswani et al. [4] and the later implementation of large-scale pre-trained language models including BERT [5] and GPT-3 [6] have specifically spurred forward, with the models performing almost as well as humans on various language understanding tasks.

However, in spite of these developments, there are still a lot of challenges. Ambiguous entries, multi-turn context tracking, low-resource languages, and factual accuracy- weaknesses are also still problems with chatbots, which can ruin customer confidence without immediate action [7]. Real-world deployments are further complicated by regulatory issues (data privacy, algorithmic bias, and regulatory compliance) and ethical implications of data privacy, algorithmic bias, and regulatory compliance [8].

The current paper provides a survey of the state of the art in terms of NLP-based customer support chatbots based on peer-reviewed literature, publicly released industry reports, and benchmark datasets. We contribute: (i) a taxonomy of chatbot architectures and NLP methods; (ii) a review of domain-specific applications in four industries; (iii) evaluation frameworks and performance measures; (iv) advanced methods such as RAG and emotional intelligence; and (v) open research problems and future directions.

The rest of the paper will be structured in the following way. In section II, the basic principles of NLP are discussed. Section III describes chatbot architectures. The NLP models and data processing methods are described in section IV. Application surveys are in Section V. Section VI explains the advantages of chatbots that are powered by AI. Section VII examines difficulties and constraints. Section VIII looks into state-of-the-art techniques. Section IX deals with human-AI co-operation. The performance evaluation is presented in section X. Section XI reviews industry implementations. Section XII describes the future directions and Section XIII is a conclusion.

II. FUNDAMENTALS OF NATURAL LANGUAGE PROCESSING IN CHATBOTS

NLP is a collection of computational problems which collectively allow machines to analyze, understand, and produce human language. Some of the NLP elements are especially relevant in the case of customer support chatbots.

A. Intent Recognition
The objective of recognizing the communicative goal of the user in a natural language utterance is known as intent recognition. When querying about the status of my order, an example query should be classified in intent ORDER_STATUS. State-of-the-art methods encode this as a text classification task and use deep neural networks, typically trained on large corpora, to provide high accuracy despite scanty in-domain training data [9]. Liu et al. [10] showed that BERT models (with fine-tuning) significantly outperform classical SVM and CNN baselines at intent classification benchmarks (with F1 score greater than 0.97 when using the SNIPS dataset).

B. Entity Extraction
Slot filling and named entity recognition (NER) are algorithms that recognize and classify particular bits of information in user utterances, including product names, order numbers, dates, and locations. These are the objects that fill the slots of structured dialogue state that is accessed by downstream systems to satisfy requests [11]. Transformer-based sequence labeling models are able to reach state-of-the-art performance on NER benchmarks like CoNLL-2003 [12].

C. Sentiment Analysis
Sentiment analysis identifies the polarity of emotion, whether positive, negative or neutral, and in more fine-grained versions, the emotion (e.g., anger, satisfaction, frustration) conveyed in a user message [13]. Real-time sentiment detection can be used in customer support settings to allow chatbots to adjust their style of response, give priority to escalations to human operators, and indicate interactions that need to be reviewed by the manager. The Stanford Sentiment Treebank was presented by Socher et al. [14] and showed fine-grained sentiment classification to be effective with recursive neural networks, with later studies based on BERT-based models showing even higher accuracy on this task [15].

D. Tokenization and Text Preprocessing.
The most common preprocessing pipelines involve tokenizing (dividing text into token), normalization (lowercasing, punctuation marks), and stop word removal, lemmatization or stemming, and out-of-vocabulary processing [16]. Recent subword tokenization algorithms have been developed to overcome the vocabulary coverage issue, most prominently, Byte Pair Encoding (BPE) [17] and WordPiece [5], which encode rare and novel words as a sequence of subword units, a feature especially useful in customer support contexts with domain-specific vocabulary and product names.

III. ARCHITECTURE AND DESIGN OF CUSTOMER SUPPORT CHATBOTS

A. Key Architectural Components

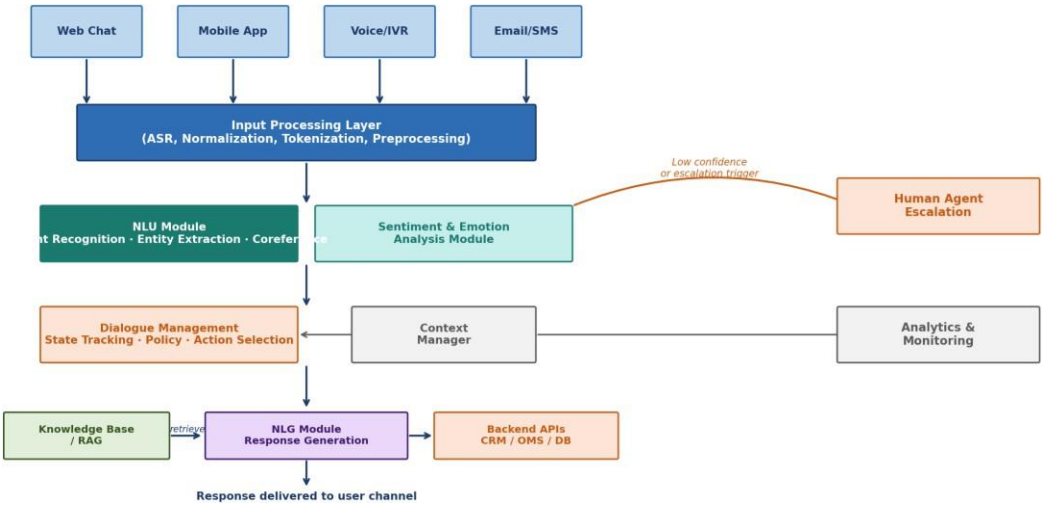

Fig. 1. End-to-End Architecture of an NLP-Based Customer Support Chatbot

Module	Function	Key Technologies
Natural Language Understanding (NLU)	Intent recognition, entity extraction, coreference resolution	BERT, DistilBERT, BiLSTM-CRF
Dialogue Management (DM)	State tracking, policy learning, action selection	LSTM, Transformer, Reinforcement Learning
Natural Language Generation (NLG)	Response template filling, generative response synthesis	GPT-3, T5, DialoGPT
Backend Integration	API calls, database queries, knowledge base lookup	REST APIs, SQL, Vector DBs
Context Manager	Multi-turn context retention, session management	Attention mechanisms, Memory Networks

TABLE I. Core Architectural Modules of NLP-Based Customer Support Chatbots

B. Language Models and Embeddings.
The representational core of the modern NLP systems is represented by word embeddings, dense vectors that store semantic relationships among words. The concept of distributional semantics in low-dimensional vectors was introduced by Word2Vec [19] and GloVe [20]. Embeddings obtained with ELMo [21] and transformer-based systems, including contextual embeddings, produce token representations based on the surrounding context, which significantly enhances task performance on disambiguation and understanding.

C. Dialogue Management Systems.
The multi-turn conversational flow is controlled by dialogue management, which determines what the system should do based on the current state of the dialogue. There are two main paradigms: (1) rule-based finite state machines, which adhere to pre-written conversation graphs and are very predictable but tend to break when users are non-cooperative; and
(2) statistical/neural dialogue managers, which are learned and can generalize to unseen conversation paths [22]. Deep reinforcement learning (RL) has been used to optimize dialogue policy, and the dialogue is modeled as a Markov Decision Process and policies that maximize rewards are learned by interacting with user simulators [23].

D. Chatbot Frameworks and Technologies.
A number of open source and commercial chatbot frameworks are available. AIML (Artificial Intelligence Markup Language) [24] and other rule-based systems are simple and interpretable, but have limited scalability. Retrieval-based models such as FAISS [25] and dense retrieval models represent a corpus of candidate responses, and choose the most relevant response at inference time, which provides high factual grounding. Models that are generative and have encoder-decoder or decoder-only architecture (e.g., DialoGPT [26], Blenderbot [27]) can generate new, contextually sensitive answers but are prone to hallucination. The hybrid systems integrate both retrieval and generation and take advantage of the factual dependability of the former and fluency of the latter [28].

IV. NLP MODELS AND TECHNIQUES FOR CUSTOMER SUPPORT

A. Deep Learning Methods.
Some of the earliest deep architectures to be used on NLP were Convolutional Neural Networks (CNNs), which performed well on sentence classification tasks by learning local n-gram features [29]. Recurrent Neural Networks (RNNs), and especially the Long Short-Term Memory (LSTM) networks [30], were more appropriate to the study of sequential language modeling, in that they were able to capture long-range dependencies. This was changed by the introduction of the attention mechanism [31] and later the Transformer architecture [4] which allowed sequences to be processed in parallel and significantly improved the performance of nearly all NLP benchmarks.

B. Transformer Models
Devlin et al. [5] introduced BERT (Bidirectional Encoder Representations from Transformers), a deep bidirectional transformer pre-trained on large text collections via masked language modeling and next sentence prediction tasks, which transformed NLP. Downstream task fine-tuning BERT only needs relatively small volumes of labeled data, and is thus highly feasible in domain-specific customer support applications. DistilBERT [32] can reduce size by 40 percent and has only a slight performance loss, which is appropriate in a deployment with latency constraints.
A cross-lingual pre-trained model specifically, XLM-RoBERTa [33], when deployed in multinational customer support scenarios, is especially relevant as it demands similar performance across a wide range of linguistic markets. Pre-training on customer support corpora with domain-adaptive has also been demonstrated to enhance performance on domain-specific tasks [34].

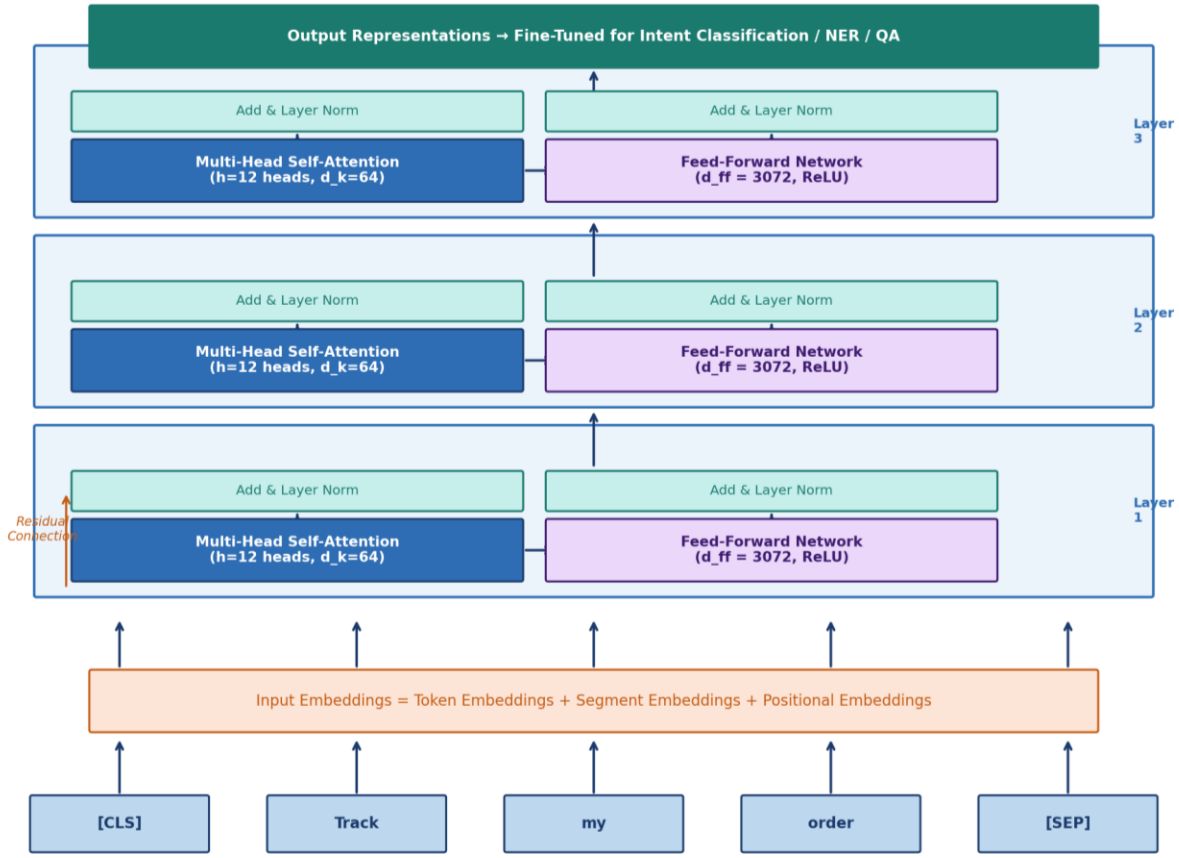

Fig. 2. Transformer Encoder Architecture (BERT) with Multi-Head Self-Attention and Feed-Forward Layers

C. Large Language Models.

GPT family of models (GPT [35], GPT-2 [36], GPT-3 [6]) uses autoregressive language modeling, i.e. the ability to predict the next token based on all previous tokens, to generate open-ended text. GPT-3, which has 175 billion parameters, has a strong few-shot learning performance, and it can perform well on a variety of NLP tasks with just a few in-context examples [6]. Nonetheless, GPT-3 is prone to hallucination the creation of plausible and factually inaccurate text, an essential shortcoming in a customer support scenario where factual accuracy is essential [37]. Reinforcement Learning models with Human Feedback (RLHF) like InstructGPT [38] have demonstrated increased compliance with human intent and reduced occurrence of harmful output generation.

D. Data Processing and Feature Extraction

Lemmatization simplifies the word forms to their canonical lemma (e.g., running to run), decreasing the size of the vocabulary and enhancing generalization [16]. Semantic similarity can be efficiently searched in the repositories of responses with semantic embeddings that are computed by sentence-level models (e.g., Sentence-BERT [39]). The TF-IDF, mutual information, and chi-squared tests are not outdated feature selection techniques that can be useful both in classical machine learning pipelines and as and interpretability tool [40].

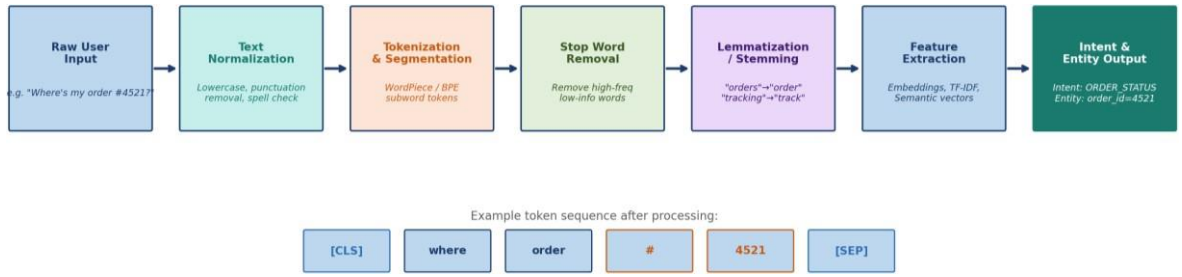

Fig. 3. NLP Text Processing Pipeline: from Raw User Input to Intent and Entity Extraction

V. APPLICATIONS AND USE CASES

A. Retail and E-Commerce.
The chatbots are implemented on e-commerce platforms throughout the customer experience, including product discovery and after-sale services. Natural language product search: In which a query is provided by the user and is mapped to a semantic embedding and compared against a product embedding index helps to provide more intuitive product discovery compared to a query-based search [42]. Order management systems with chatbots are capable of answering questions related to order status, approximate delivery time, and the process of returns in real time and decrease the number of support tickets, which need to be processed by human agents [43]. Examples of this grand paradigm on a large scale are Amazon Alexa and Shopify conversational commerce [44].

B. Banking and Financial Services.
Banking chatbots have to handle many repetitive questions regarding account balances, transaction history, and branch locations. Advanced executions embrace directed processes on loan applications and investment onboarding [45]. Notably, the effective antifraud platform needs to be integrated with the financial chatbots: the suspicious transaction behavior detected by the machine learning algorithms can lead to the automatic warnings found on the chatbot interface [46]. This industry has compliance standards that mandate chatbots to maintain a history of communication and provide explicit disclosure of the automated nature of the service [8].

C. Healthcare
Patient engagement The interaction between patients and healthcare chatbots can be enhanced by triaging symptoms, scheduling appointments, medication reminders, and providing general health information [47]. Babylon Health and Ada Health symptom checker are examples of commercial implementations. Chatbots that interact with patients should comply with strict regulations on data protection (HIPAA in the United States, GDPR in the European Union) and take specific care to ensure that they do not go too far and offer information and advice instead of a diagnosis [48]. Research has shown that properly designed healthcare chatbots can decrease unwarranted visits to the emergency department and enhance medication compliance [49].

D. Food Delivery and Hospitality
The food delivery business uses chatbots to place orders, deliveries in real-time, and address complaints. Mapping APIs allow the chatbot to deliver precise estimated time of arrival (ETA) notifications and actively inform about delays [50]. Chatbots in hospitality are used to handle reservation systems, address frequently asked questions about amenities, and check-in and check-out processes thereby reducing the workload of the front desk and enhancing the guest experience [51].

VI. ADVANTAGES OF AI-POWERED CHATBOTS

Operational and strategic benefits of AI-based customer support chatbots are highly covered in the academic literature and in industry reports. These advantages can be divided into three dimensions that we classify.

A. Operational Efficiency
Chatbots are very accessible 24/7 without the shift differentials and fatigue effects that restrains the performance of human agents [52]. Automation can solve high-volume, low-complexity queries, such as FAQ answers and order status queries, reducing Average Handle Time (AHT) and enabling human agents to deal with high-value interactions that are complex [53]. According to a State of Service Report published by Salesforce (2022), 83% of customer service decision- makers stated that chatbots are creating cost savings [54].

B. Improvement of Customer Experience.
Another distinction that needs to be mentioned when it comes to AI-powered chatbots is the customization of the latter: depending on the history of the customers, their purchasing patterns, as well as the preferences they might be expected to have, chatbots will be able to offer them personal recommendations, actively present the information that they might be interested in the most, and change the style of communication to the particular user [55]. Multi-channel implementation Multi-channel implementation (web, mobile, SMS, social media) is a strategy to guarantee quality services on touchpoints [56]. The existence of response mechanisms that maintain a history of conversations across a sequence of interactions is context-conscious, and assists in delivery of a coherent, relationship-like customer experience.

C. Business Performance Metrics.
Evaluations of the use of chatbots quantitatively consistently indicate an improvement in major customer experience indicators. In a study by IBM [57], AI chatbots have the potential to save customer service costs by as much as 30%. It has been demonstrated that chatbot-aided communication can enhance the rates of First-Contact Resolution (FCR) or the percentage of queries that do not require an escalation or follow-up, as it allows quicker access to information and more

uniform responses [58]. The score of Customer Satisfaction (CSAT) increases when chatbots are properly designed and can smoothly transfer to human agents when necessary [59].

VII. CHALLENGES AND LIMITATIONS

A. Technical Challenges

The problem of dealing with uncertain inputs is inherent: natural language is polysemous by nature and users often convey the same meaning in very different ways when they use the same words to convey different meanings [7]. Sarcasm, irony and implicit negation are especially challenging to the NLP models that are inclined to rely on surface patterns [60]. Multi-part queries which involve multiple intents are complex and need more elaborate dialogue management strategies to break down and deal with each element of the query.

B. Data and Model Issues.

Supervised learning methods need large amounts of labeled training data, which is costly to gather and label in domain-specific customer care areas [61]. As observed in Section IV-D, imbalance in classes is a widespread issue. Model hallucination: the creation of sure but factually incorrect outputs: is a major threat when used in a customer-facing application; retrieval-augmented generation (Section VIII-D) is a promising protection measure [62]. The process of language support of low-resource languages, especially those with small pre-training corpora, is still an open problem [33].

C. Practice Implementation Problems.

Interoperability with existing enterprise systems: CRM systems, ERP systems, custom databases are often subject to custom API development and data reconciliation work [63]. Multi-turn conversation management needs a strong context tracking across sessions, with topic switching and anaphora resolving. Model compression, quantization, or hardware acceleration may be required due to the real-time processing demands (typically less than 200 ms latency to provide a good user experience). Mobile device constraints also limit the complexity of on-device NLP models: limited memory and compute.

D. Ethical and Security Concerns.

The interaction of chatbots with customers will inevitably involve the collection and processing of personal data, which will increase the liability of customers regarding the privacy regulations like the General Data Protection Regulation (GDPR) [65] and the California Consumer Privacy Act (CCPA) [66]. Systematic errors in model behaviour that disfavour certain demographic groups are called algorithmic bias and have been reported in NLP systems and should be proactively monitored and avoided [67]. Bad actors can also seek adversarial attacks (designed inputs that aim to evoke the wrong, unhealthy or secretive response) and this needs strong input verification and result-filtering [68]. Transparency policies provide that users are made aware when they are dealing with an automated system as opposed to a human agent [8].

VIII. TECHNIQUES ADVANCED AND INNOVATIVE

A. Sentiment Analysis and Emotional Intelligence.

State-of-the-art customer support chatbots use multi-class emotion recognition models to identify not only positive/negative polarity but also particular feelings like frustration, urgency and satisfaction [69]. Sentiment scoring in real-time also allows automatically adjusting the response: when a chatbot realizes that the user is getting frustrated, it can automatically activate an empathetic response template and focus on escalation to a human agent. Rashlin et al. [70] have shown that in open-domain situations, transformer models trained on empathetically annotated dialogue datasets can produce more empathetic responses.

B. Personalization and Context Awareness

User preference learning: User behavior, communication style and service history model development allows the interactions to become more personal as time goes on [71]. history kept conversation Inter-session conversation history maintenance is keeping of the context of past interaction effectively. The response is relevant in the rapidly changing business environment by dynamically creating content, because of real-time indicators such as existing promotions, out of stock, or service outages [72].

C. Retrieval-Augmented Generation (RAG)

RAG, which was postulated by Lewis et al. [73], is a generative language model that incorporates a retrieval component that retrieves in a knowledge base the relevant documents and then generates a response. Such a composite approach has a major impact in reducing hallucination since the replies are anchored on recovered factual content with no loss of the fluency of neural production. Knowledge base in customer support apps may consist of product documentation, frequently asked questions (FAQ), policy manuals and a list of solved tickets in the past. Traditional sparse retrieval

(BM25) has been mostly replaced by dense retrieval models using bi-encoder structures [74] to retrieve semantic documents in RAG pipelines.

D. Voice-Enabled Chatbots

Voice chatbots combine text-to-speech (TTS) and automatic speech recognition (ASR) systems to enable natural speech interactions with users who may use them as voice interfaces or need hands-free interactions [75]. Deep learning ASR systems such as DeepSpeech [76] and Whisper [77] have shown human parity performance on common benchmarks. The future of chatbots is multi-modal chatbots, which combine a text, voice and visual input modes, to enable more rich modalities of interaction in applications such as visual product troubleshooting [78].

IX. HYBRID SOLUTIONS: HUMAN AND AI COLLABORATION.

However, not all communication with the customer care can be all-automatic and must be. Critically emotional cases (e.g. bereavement, severe financial hardship), cases of a legal nature that are demanding of innovative problem-solving solutions, such as these, should not be handled by algorithms [79]. Hybrid systems provide successful interaction by applying smart escalation policies: once chatbot confidence is low enough, when some trigger events happen (e.g. negative sentiment exceeded a limit after four consecutive turns), or when it is explicitly requested by the user to make a human agent [80].

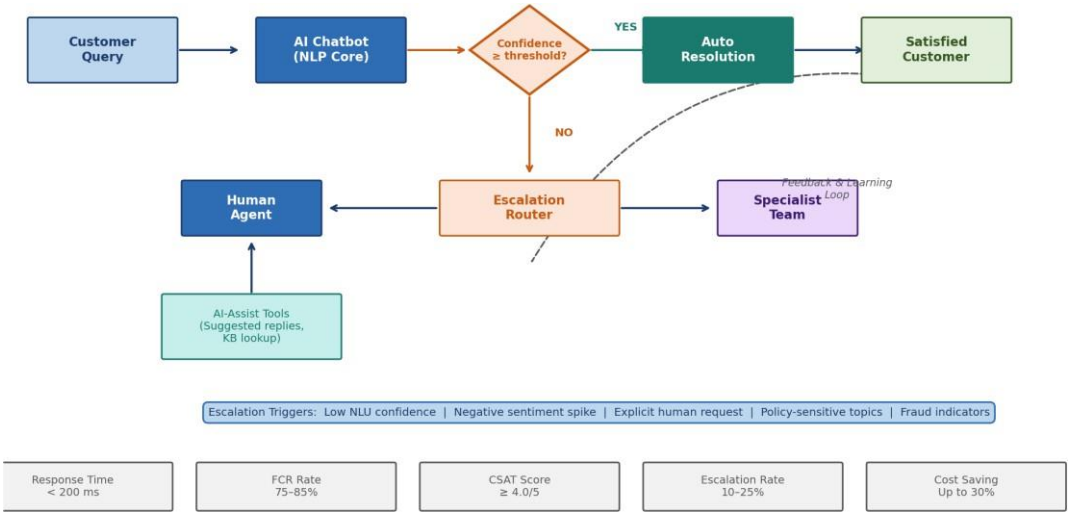

Fig. 4. Hybrid AI–Human Collaboration Framework Showing Escalation Logic and Performance Metrics

Hybrid systems ought to divide labor in a way that would maximize the difference in the relative strengths of both: chatbots are good at responding instantly, consistent enforcement of policy, and processing high-volume engagements concurrently, whereas human agents are good at empathy, complex reasoning, and exceptions handling [81]. Staffing efficiency can be greatly enhanced by workforce management systems which dynamically assign human agent capacity based on predicted chatbot escalations. They also support AI-assisted agents, including real-time proposed responses, automated knowledge retrieval, sentiment alerting, and others, to enhance their functionality [82].

X. PERFORMANCE EVALUATION METRICS

Strict assessment is needed to assess the work of chatbots and direct the process of improvement of the system. The literature and industry evaluation frameworks utilize a mix of automated measures and human judgments.

Metric	Definition	Typical Benchmark
Intent Classification Accuracy	% of utterances assigned correct intent	> 90% (SNIPS, ATIS)
Entity F1 Score	Harmonic mean of precision and recall over entity spans	> 85% (CoNLL-2003)
BLEU Score	N-gram overlap between generated and reference responses	0.30-0.45 (DialoGPT)
Response Latency	Time from query submission to response delivery (ms)	< 200 ms (production SLA)

Metric	Definition	Typical Benchmark
First-Contact Resolution (FCR)	% of queries resolved without escalation or follow-up	70-85% (industry median)
Customer Satisfaction (CSAT)	Post-interaction survey score (1-5 scale)	> 4.0 / 5.0
Escalation Rate	% of interactions transferred to human agents	10-25% (target range)
Task Completion Rate	% of user-initiated tasks successfully completed by chatbot	> 80%

TABLE II. Key Performance Evaluation Metrics for Customer Support Chatbots

It is significant to mention that automated measures like BLEU and ROUGE, although convenient, do not correlate perfectly with the human judgments of quality of responses in conversation contexts [83]. The best known approach to the comprehensive evaluation of chatbots is human assessment protocols that evaluate fluency, coherence, correctness, and user experience [84].

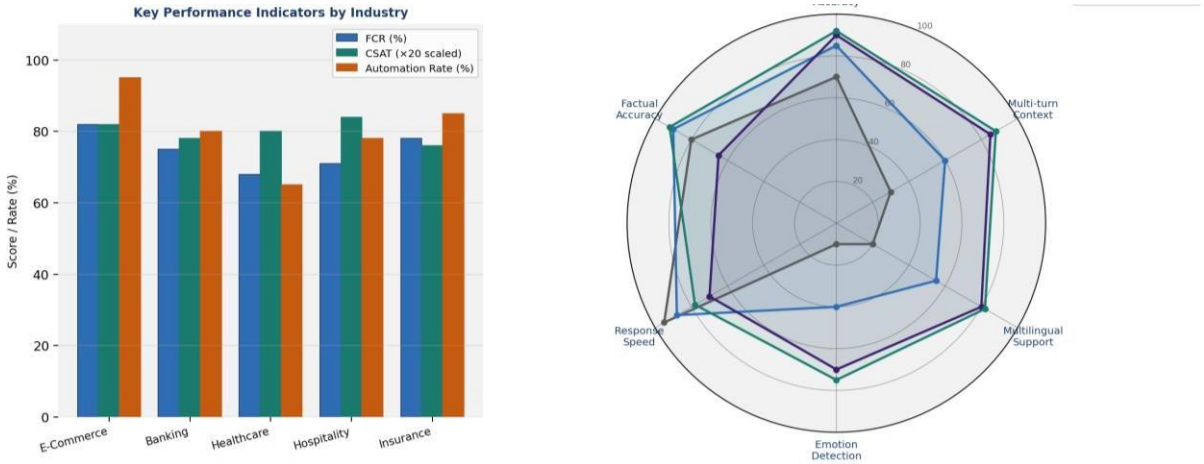

Fig. 5. Industry Application Metrics: KPI Comparison by Sector and Capability Radar Chart

XI. INDUSTRY-SPECIFIC IMPLEMENTATIONS

A. E-Commerce Platforms

Alexa by Amazon and the Alexa Customer Service Skills is an advanced implementation of voice-based conversational AI at scale, with millions of daily customer interactions in the form of product inquiries, order management, and smart home control [44]. The AliMe chatbot implemented on the Taobao and Tmall marketplaces uses a hybrid architecture of retrieval-generation and processes more than 95 percent of pre-sale customer questions on its own during the busiest times of the year, including the Singles Day [85]. The conversational AI that Shopify implements through its AI assistant, which is embedded on merchant dashboards and customer storefronts, illustrates how conversational AI can be embedded into the infrastructure of e-commerce platforms [86].

B. Banking Services

By 2023, one such virtual assistant, Erica by Bank of America, which was introduced in 2018, had assisted more than 32 million customers and assisted more than 1.5 billion interactions [87]. Erica uses NLP in a variety of actions such as balance inquiries, transaction search, bill pay reminders, and personalized financial insights. The conversational AI in retail banking is maturing similarly in HSBC chatbot Amy and DBS Bank digibank chatbot. Notably, such deployments support strong human escalation channels in the context of sophisticated financial guidance and interactions that are sensitive to compliance [88].

C. Healthcare Systems

The AI-driven symptom checker and triage service of Babylon Health is being deployed in various health systems around the world, one of them being the National Health Service (NHS) of the UK. A clinical validation study by Razzaki et al. [89] compared Babylon to the general practitioners in terms of their diagnostic accuracy and revealed similar performance on a standardized clinical vignette test. The Digital Health Center of Excellence of the FDA has started to develop regulatory standards of AI-enabled clinical decision support solutions [90].

D. Insurance Sector

The AI Jim chatbot, a part of Lemonade, automates the insurance claims system whereby policyholders can file and in a simple case, claim and settle claims fully within minutes through a chat with a conversational interface, a feat that used to take days of manual processing [91]. The chatbot uses machine learning-based fraud detection algorithms to filter suspicious claims to be reviewed by humans, both meeting efficiency and risk management goals [92].

XII. FUTURE DIRECTIONS AND RESEARCH OPPORTUNITIES

A. Emerging Technologies

Generative AI is progressing at a very fast rate. The GPT-4 [93] models and the following ones show significantly better reasoning and instruction following, as well as factual accuracy than the predecessors, and now the ceiling of what can be accomplished in automated customer support is much higher. The ability to use tools and call functions- the ability to enable language models to communicate with external APIs and databases is increasing the range of tasks that can be automated.

B. Explainable AI and Transparency.

Both customers and organizations gain advantages in knowing the rationale behind a chatbot arriving at a specific conclusion or suggestion. The fact that dialogue models can be developed to be inherently interpretable instead of being post-hoc explained as black-box models is also a major research frontier [96].

C. Research Gaps and Opportunities

There are various research gaps that should be looked at. The retention of context in long multi-session dialogues is still quite difficult; the attention processes of current models decay in the presence of very long contexts, and effective memory architectures to support persistent user modeling are also under research [97]. The ability to deal with truly intricate, multi-step customer service processes, which involve planning, reasoning, and orchestration of multiple back- end systems, necessitates additional innovations in autonomous agent models [98]. The improved multilingual and cross- lingual assistance, especially of morphologically rich and low-resource languages, demands further investment in multilingual pre-training and cross-lingual transfer methods [33]. The creation of AI, including bias auditing systems, fair training objectives, and participatory design processes including marginalized groups should be considered part of the development lifecycle early on [99].

XIII. CONCLUSION

The current paper has provided a thorough review of NLP-based chatbot systems in customer support with a summary of the basic principles, architecture paradigms, deep learning models, application in the industry, performance indicators, and future research opportunities. The materials discussed have demonstrated that carefully developed NLP chatbots can deliver a lot of value in terms of operations, money, and customer experience in any industry. Nonetheless, to realize these benefits, significant focus should be paid to the selection of models, quality of data, rigor of evaluation, integration of systems, and ethical considerations.

The fast development of large language models, retrieval-augmented generation, and multimodal AI is pushing the limits of what automated customer support can accomplish. At the same time, the complexity and delicacy of many customer communications demonstrate the continuation of the effective human-AI cooperation models. Future research should emphasize more on context retention, complex query processing, multi-lingual support, explainability and ethical development of AI to realize the full potential of conversational AI in customer support without losing customer trust and regulatory compliance.

REFERENCES

[1] D. Acemoglu and P. Restrepo, "Robots and Jobs: Evidence from US Labor Markets," Journal of Political Economy, vol. 128, no. 6, pp. 2188–2244, 2020.

[2] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. (draft). Prentice-Hall, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[3] J. Weizenbaum, "ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine," Communications of the ACM, vol. 9, no. 1, pp. 36–45, Jan. 1966.

[4] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[6] T. B. Brown et al., "Language Models Are Few-Shot Learners," in Advances in Neural Information Processing Systems, vol. 33, 2020.

[7] M. Chakrabarti, S. Chakraborty, and S. Mukherjee, "Challenges in Building Intelligent Open-Domain Dialog Systems," ACM Transactions on Information Systems, vol. 39, no. 4, pp. 1–28, 2021.

[8] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the European Union, Apr. 2016.

[9] Q. Chen et al., "BERT for Joint Intent Classification and Slot Filling," arXiv:1902.10909, 2019.

[10] B. Liu and I. Lane, "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling," in Proc. Interspeech, 2016, pp. 685–689.

[11] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55–75, 2018.

[12] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in Proc. CoNLL-2003, 2003.

[13] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 2nd ed. Cambridge University Press, 2020.

[14] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in Proc. EMNLP, 2013, pp. 1631–1642.

[15] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in China National Conference on Chinese Computational Linguistics, 2019, pp. 194–206.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in Proc. ACL System Demonstrations, 2014, pp. 55–60.

[17] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in Proc. ACL, 2016, pp. 1715–1725.

[18] S. Young et al., "POMDP-Based Statistical Spoken Dialogue Systems: A Review," Proceedings of the IEEE, vol. 101, no. 5, pp. 1160–1179, 2013.

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in Advances in Neural Information Processing Systems, vol. 26, 2013.

[20] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Proc. EMNLP, 2014, pp. 1532–1543.

[21] M. E. Peters et al., "Deep Contextualized Word Representations," in Proc. NAACL-HLT, 2018, pp. 2227–2237.

[22] J. Williams, A. Raux, and M. Henderson, "The Dialog State Tracking Challenge Series: A Review," Dialogue & Discourse, vol. 7, no. 3, pp. 4–33, 2016.

[23] P. Dhingra et al., "End-to-End Reinforcement Learning of Dialogue Agents for Information Access," in Proc. ACL, 2017, pp. 484–495.

[24] R. S. Wallace, "The Anatomy of ALICE," in Parsing the Turing Test, R. Epstein, G. Roberts, and G. Beber, Eds. Springer, 2009, pp. 181–210.

[25] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021.

[26] Y. Zhang et al., "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation," in Proc. ACL System Demonstrations, 2020, pp. 270–278.

[27] S. Roller et al., "Recipes for Building an Open-Domain Chatbot," in Proc. EACL, 2021, pp. 300–325.

[28] K. Shuster et al., "Retrieval Augmentation Reduces Hallucination in Conversation," in Findings of EMNLP, 2021.

[29] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP, 2014, pp. 1746–1751.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in Proc. ICLR, 2015.

[32] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT," arXiv:1910.01108, 2019.

[33] A. Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," in Proc. ACL, 2020, pp. 8440–8451.

[34] D. Gururangan et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," in Proc. ACL, 2020.

[35] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI Blog, 2018.

[36] A. Radford et al., "Language Models Are Unsupervised Multitask Learners," OpenAI Blog, 2019.

[37] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[38] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," in Advances in Neural Information Processing Systems, vol. 35, 2022.

[39] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. EMNLP, 2019.

[40] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[42] K. Balog and T. Kenter, "Personal Knowledge Graphs: A Research Agenda," in Proc. ICTIR, 2019.

[43] S. Xu, P. Sun, and Z. Zhao, "Design and Implementation of an Intelligent Customer Service System for E-Commerce," IEEE Access, vol. 8, pp. 140012–140024, 2020.

[44] Amazon Alexa, "Alexa Developer Documentation," Amazon Web Services, 2023. [Online]. Available: https://developer.amazon.com/en-US/alexa

[45] F. Alt, A. B. Sahami Shirazi, A. Schmidt, U. Kramer, and Z. Akturan, "Comparing Formal and Informal Address Forms for Speech-Based User Interfaces," in Proc. ACM CHI, 2010.

[46] V. Van Vlasselaer et al., "APATE: A Novel Approach for Automated Credit Card Transaction Fraud Detection Using Network-Based Extensions," Decision Support Systems, vol. 75, pp. 38–48, 2015.

[47] L. Laranjo et al., "Conversational Agents in Healthcare: A Systematic Review," Journal of the American Medical Informatics Association, vol. 25, no. 9, pp. 1248–1258, 2018.

[48] U.S. Department of Health and Human Services, "Guidance on HIPAA and Health Apps," Office for Civil Rights, 2023.

[49] K. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," The Canadian Journal of Psychiatry, vol. 64, no. 7, pp. 456–464, 2019.

[50] Uber Technologies, "Uber Eats Help and Support," 2023. [Online]. Available: https://help.uber.com/ubereats

[51] B. Prentice, R. Nguyen, and E. Nusair, "Chatbots in the Hotel Industry: Key Applications and Use Cases," Cornell Hospitality Quarterly, vol. 61, no. 4, pp. 438–450, 2020.

[52] N. Sabharwal, Cognitive Virtual Assistants Using Google Dialogflow. Apress, 2020.

[53] McKinsey & Company, "The Next Frontier of Customer Engagement: AI-Enabled Customer Service," McKinsey Digital, 2023.

[54] Salesforce Research, "State of Service, 5th Edition," Salesforce, 2022. [Online]. Available: https://www.salesforce.com/resources/research-reports/state-of-service/

[55] Y. Li et al., "Personalized Dialogue Generation with Diversified Traits," arXiv:1901.09672, 2019.

[56] Gartner, "Gartner Magic Quadrant for Enterprise Conversational AI Platforms," Gartner Research, 2023.

[57] IBM Institute for Business Value, "The Value of Training AI on Your Data," IBM, 2022.

[58] NICE, "2022 Digital-First Customer Experience Report," NICE Systems, 2022.

[59] H. Chung and S. Ko, "Effects of Chatbot Service Quality on Customer Satisfaction," Journal of Retailing and Consumer Services, vol. 71, 2023.

[60] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," Information Fusion, vol. 37, pp. 98–125, 2017.

[61] A. M. Turing, "Computing Machinery and Intelligence," Mind, vol. 59, no. 236, pp. 433–460, 1950.

[62] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in Proc. ACL, 2022.

[63] R. T. Fielding, "Architectural Styles and the Design of Network-Based Software Architectures," Ph.D. dissertation, Univ. of California, Irvine, 2000.

[64] G. Hinton et al., "Distilling the Knowledge in a Neural Network," arXiv:1503.02531, 2015.

[65] Information Commissioner's Office, "Guide to the UK GDPR," ICO, 2023. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/

[66] California Department of Justice, "California Consumer Privacy Act (CCPA)," 2018. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[67] E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in Proc. ACM FAccT, 2021, pp. 610–623.

[68] N. Carlini et al., "Extracting Training Data from Large Language Models," in Proc. USENIX Security Symposium, 2021.

[69] S. Poria, E. Cambria, and A. Gelbukh, "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis," in Proc. EMNLP, 2015.

[70] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset," in Proc. ACL, 2019.

[71] S. Zhang et al., "Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?," in Proc. ACL, 2018.

[72] E. Dinan et al., "Wizard of Wikipedia: Knowledge-Powered Conversational Agents," in Proc. ICLR, 2019.

[73] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020.

[74] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020.

[75] A. Hannun et al., "Deep Speech: Scaling Up End-to-End Speech Recognition," arXiv:1412.5567, 2014.

[76] C. Lüscher et al., "RWTH ASR Systems for LibriSpeech: Hybrid vs Attention," in Proc. Interspeech, 2019.

[77] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. ICML, 2023.

[78] J.-B. Alayrac et al., "Flamingo: A Visual Language Model for Few-Shot Learning," in Advances in Neural Information Processing Systems, vol. 35, 2022.

[79] A. Følstad and M. Skjuve, "Chatbots for Customer Service: User Experience and Motivation," in Proc. ACM Conversational User Interfaces, 2019.

[80] A. B. Chakraborty et al., "Towards Facilitating Empathic Conversations in Online Mental Health Support," in Proc. The Web Conference, 2020.

[81] M. Amershi et al., "Software Engineering for Machine Learning: A Case Study," in Proc. ICSE-SEIP, 2019.

[82] S. Shum, X. He, and D. Li, "From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots," Frontiers of Information Technology & Electronic Engineering, vol. 19, no. 1, pp. 10–26, 2018.

[83] C. Liu et al., "How NOT to Evaluate Your Dialogue System," in Proc. EMNLP, 2016.

[84] J. Deriu et al., "Survey on Evaluation Methods for Dialogue Systems," Artificial Intelligence Review, vol. 54, no. 1, pp. 755–810, 2021.

[85] J. Li et al., "AliMe Assist: An Intelligent Assistant for Creating an Innovative E-Commerce Experience," in Proc. ACM CIKM, 2017, pp. 2495–2498.

[86] Shopify Inc., "Shopify Magic and Sidekick," Shopify, 2023. [Online]. Available: https://www.shopify.com/magic

[87] Bank of America, "Bank of America Reports Record Full Year 2023 Net Income of $26.5 Billion," Investor Relations, Jan. 2024.

[88] DBS Bank, "DBS digibank AI," DBS, 2023. [Online]. Available: https://www.dbs.com/digibank/

[89] S. Razzaki et al., "A Comparative Study of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis," arXiv:1806.10698, 2018.

[90] U.S. Food and Drug Administration, "Artificial Intelligence and Machine Learning in Software as a Medical Device," FDA, 2023. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

[91] Lemonade, "Lemonade's AI Claims Bot Sets World Record," Lemonade Blog, 2017. [Online]. Available: https://www.lemonade.com/blog/lemonade-sets-new-world-record/

[92] T. Bhowmik, N. Liu, and J. Bian, "AI-Enabled Fraud Detection in Insurance," IEEE Transactions on Emerging Topics in Computing, vol. 10, no. 2, pp. 1124–1136, 2022.

[93] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.

[94] Google DeepMind, "Gemini: A Family of Highly Capable Multimodal Models," arXiv:2312.11805, 2023.

[95] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.

[96] Z. C. Lipton, "The Mythos of Model Interpretability," Queue, vol. 16, no. 3, pp. 31–57, 2018.

[97] A. Dong et al., "Survey of Long-Context Language Models," arXiv:2402.01364, 2024.

[98] Y. Shen et al., "HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face," in Advances in Neural Information Processing Systems, vol. 36, 2023.

[99] A. Selbst et al., "Fairness and Abstraction in Sociotechnical Systems," in Proc. ACM FAccT, 2019, pp. 59–68.

