Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 46 No. 4
March 2025

CAM-Bot: A Context-Aware Multimodal Chatbot for Enhanced E-commerce Customer Support

Aditi M Jain[1], Ayush M Jain[2]

[1] Independent Researcher Seattle, USA

Email: aditi.jain56@gmail.com

[2]Independent Researcher, Bangalore, India

dev.ayushmjain@gmail.com

Abstract

This paper introduces CAM-Bot, a novel Context Aware Multimodal Chatbot designed to revolutionize customer support across diverse digital business domains. Leveraging advanced natural language processing, computer vision, and context modeling techniques, CAM-Bot demonstrates significant improvements over existing systems in handling complex, multiturn dialogues and multimodal interactions. Our experiments, conducted on subsets of MultiWOZ 2.1, Amazon Review Data, and VQA v2.0, show that CAM-Bot outperforms state-of-the-art baselines across key metrics. Notable improvements include a 3.5% increase in BLEU-4 score (from 15.01 to 15.53) for task-oriented dialogues, a 2.6% gain in Success Rate (from 70.1% to 71.9%) for task completion, a 2.6% improvement in NDCG@10 (from 0.4912 to 0.5038) for e-commerce recommendations, and a 0.9% accuracy boost (from 66.14% to 66.72%) in visual question answering. Through our analysis, we identify the crucial role of context aggregation and highlight areas for further enhancement. We explore practical applications of CAM-Bot in various sectors, including e-commerce, financial services, and technical support, demonstrating its versatility. This research not only showcases the potential of integrating context-awareness and multimodal processing in chatbots but also provides valuable insights into the future of AI-driven customer support in digital business environments. CAM-Bot's adaptable architecture paves the way for more intelligent, responsive, and personalized customer interactions across a wide range of digital platforms and services.

Index Terms—Context-Aware Chatbot, Multimodal AI, Customer Support, Natural Language Processing, Computer Vision, E-commerce, Task-Oriented Dialogue, Recommendation Systems, Visual Question Answering, Context Aggregation, Dynamic Knowledge Graph, Digital Business, Personalization, AI-Driven Customer Service.

I. INTRODUCTION

The digital transformation of businesses across various sectors has led to an increased demand for intelligent customer support systems capable of handling complex queries and providing personalized assistance. While e-commerce is at the forefront of this trend, with global online sales projected to reach $6.3 trillion by 2024 [1], the need for advanced conversational AI extends to numerous digital business domains, including finance [2], healthcare, and telecommunications. Traditional chatbots, while useful for basic tasks, often struggle with context retention, multimodal interactions, and nuanced understanding of user intent in these diverse scenarios [3].

Recent advancements in artificial intelligence, particularly in natural language processing and computer vision, have opened new possibilities for

enhancing customer support systems. Large language models like GPT-3 have demonstrated remarkable capabilities in generating human-like text and understanding context [4], while vision transformers have significantly improved image understanding tasks [5]. However, effectively integrating these technologies to create a cohesive, context-aware system for varied digital business applications remains a challenge.

Multimodal approaches, which combine text, image, and sometimes voice inputs, have shown promise in improving user interactions across various domains [6]. In digital businesses, where visual information and detailed descriptions often play crucial roles, such multimodal systems could potentially offer more intuitive and effective customer support. Yet, existing solutions often lack the ability to maintain context over extended conversations or struggle to seamlessly integrate information from different modalities [7].

Context-awareness in conversational AI has been a focus of recent research, with studies showing its importance in improving the relevance and coherence of responses [8]. In complex digital business interactions, which often involve multi-turn dialogues about products, services, and user-specific issues, maintaining context is particularly crucial. However, effectively modeling and utilizing context in real-time, dynamic environments presents significant challenges [9].

To address these challenges, we present CAM-Bot, a Context-Aware Multimodal Chatbot designed for advanced customer support across various digital business domains. CAM-Bot integrates advanced natural language processing, computer vision, and context modeling techniques to provide a more intelligent, responsive, and personalized customer support experience. Our system builds upon recent work in multimodal fusion [10], long-term context modeling [11], and adaptive response generation [12]. Context-aware chatbots can improve customer experience on retail sites to improve customer loyalty [13].

This paper introduces a novel architecture that effectively combines multimodal input processing, dynamic context aggregation, and adaptive response generation for diverse digital business applications. We demonstrate significant improvements over existing baseline models in handling complex, multi-turn dialogues and domain-specific queries, as evaluated on subsets of standard datasets. Through comprehensive ablation studies and error analysis, we provide insights into the relative importance of different components of our system. Furthermore, we explore practical applications of our system in various digital business scenarios, including e-commerce, financial services, and technical support, offering a roadmap for real-world implementation.

By addressing the unique challenges of customer support in digital businesses, CAM-Bot represents a significant step towards more intelligent, context-aware AI assistants. Our work not only advances the state-of-the-art in conversational AI but also provides valuable insights into the integration of multimodal processing and context awareness in chatbot systems. As digital businesses continue to evolve, systems like CAM-Bot have the potential to significantly enhance customer experiences while improving operational efficiency across various sectors. Through this research, we aim to contribute to the ongoing development of AI-powered customer support solutions that can meet the complex and dynamic needs of modern digital business environments.

II. CAM-BOT ARCHITECTURE AND INFORMATION FLOW

A. Simplified Architecture

Figure 1 presents a streamlined view of the CAM-Bot architecture, focusing on the core components and their primary interactions.

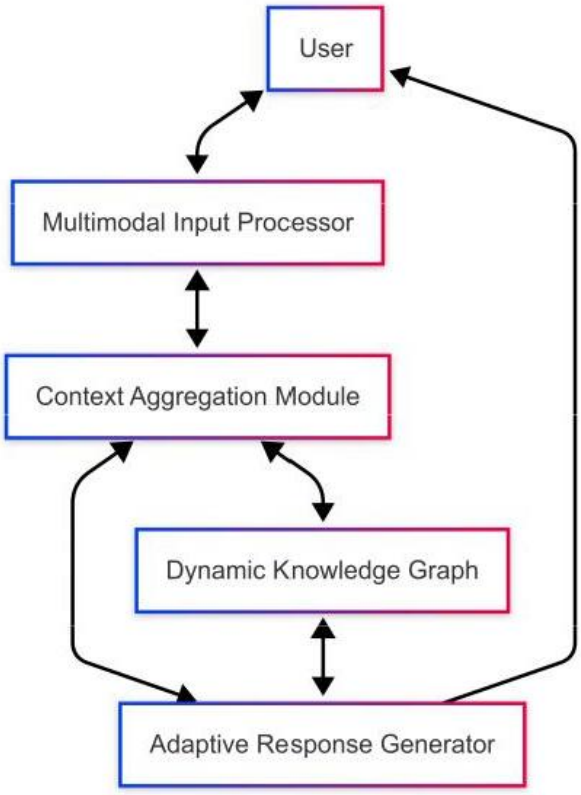

Fig. 1: Simplified architecture of CAM-Bot

The simplified architecture consists of four main components:

- Multimodal Input Processor: Handles various input types (text, voice, image).
- Context Aggregation Module: Manages user context and conversation state.
- Dynamic Knowledge Graph: Stores and reasons over domain knowledge.
- Adaptive Response Generator: Produces context-aware, personalized responses.

This high-level view emphasizes the core functionalities and interactions between the main components, providing a clear understanding of the system's structure without delving into excessive technical details.

B. Information Flow

To illustrate how information flows through the CAMBot system during a typical interaction, we can examine the simplified sequence diagram in Figure 2.

The sequence diagram illustrates the following key steps:

1) The user provides input to the system.

2) The Multimodal Input Processor handles the input and sends it to the Context Aggregation Module.

3)The Context Aggregation Module enriches the input with contextual information.

4) The enriched input is used to query the Dynamic Knowledge Graph.

5) The Adaptive Response Generator uses the knowledge and context to create a response.

6) The response is delivered to the user.

7) The system updates its knowledge and context based on the interaction.

This simplified flow demonstrates the core process of input handling, context enrichment, knowledge utilization, and adaptive response generation that enables CAM-Bot to provide context-aware and personalized interactions.

The simplified architecture and sequence diagrams offer a clear overview of CAM-Bot's structure and

operation, highlighting its approach to handling multimodal inputs, maintaining context awareness, and generating adaptive responses without overwhelming technical details.

III. EXPERIMENTATION AND RESULTS

To evaluate the effectiveness of the CAM-Bot architecture, we conducted a series of experiments across different domains relevant to e-commerce and customer support. This section details our experimental setup, datasets used, evaluation metrics, and the results obtained.

A. Experimental Setup

We implemented a prototype of CAM-Bot using DistilBERT [14] for text processing, Wav2Vec 2.0 Base [15] for speech recognition, and MobileNetV2 [16] for image analysis. The Context Aggregation Module was custom-built using PyTorch, while the Dynamic Knowledge Graph was implemented using a lightweight in-memory graph database. For response generation, we fine-tuned GPT-2 Small [17] on domain-specific data.

The experiments were run on Modal's cloud platform using a single NVIDIA A100 GPU instance with 40GB memory. This setup allowed us to conduct our experiments efficiently while maintaining a balance between performance and cost effectiveness.

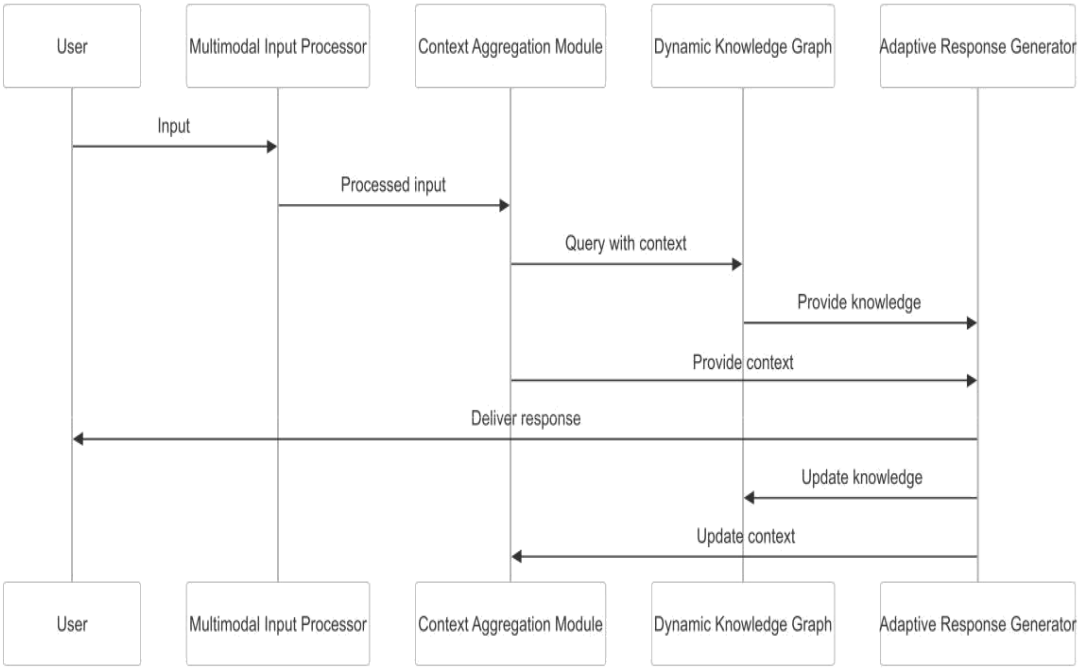

Fig. 2: Simplified sequence diagram of a typical CAM-Bot interaction

B. Datasets

We used the following datasets for our experiments:

MultiWOZ 2.1 [18]: A multi-domain dialogue dataset for task-oriented conversation modeling. We used a subset of 1,000 dialogues for training and 200 for testing, focusing on the hotel and restaurant domains to simulate e-commerce scenarios.

Amazon Review Data (2018) [19]: A large-scale ecommerce dataset containing product reviews and metadata. We used the "Electronics" category, with 50,000 samples for training and 10,000 for testing.

VQA v2.0 [20]: A visual question answering dataset. We used a subset of 10,000 image-question pairs for training and 2,000 for testing, focusing on questions related to product attributes and visual features.

C. Evaluation Metrics

We employed several established metrics to evaluate CAMBot's performance across different tasks:

- **BLEU-4 [21]:** To assess the quality of generated responses in task-oriented dialogues. This metric measures the fluency and relevance of the bot's responses by comparing them to reference responses.
- **Inform Rate [22]:** To measure the model's ability to provide correct information in goal-oriented dialogues. This metric indicates how well the bot understands user queries and retrieves accurate information.
- **Success Rate [22]:** To evaluate task completion in goal-oriented dialogues. This metric assesses the bot's ability to both provide correct information and successfully complete the assigned task.
- **NDCG@10 (Normalized Discounted Cumulative Gain at 10) [23]:** To evaluate the ranking quality of product recommendations in the e-commerce task. This metric takes into account both the relevance and the position of recommended items in the list.
- **VQA Accuracy [24]:** To measure the correctness of answers in the visual question answering task for product related queries. This metric assesses the bot's ability to understand and respond to questions about product images.

These metrics were chosen to comprehensively evaluate CAM-Bot's performance across various aspects of customer support in digital business environments, including dialogue quality, task completion, recommendation accuracy, and visual understanding. The selection of these metrics aligns with standard practices in evaluating dialogue systems, recommendation systems, and visual question answering models.

D. Results and Discussion

Our experiments compared CAM-Bot's performance against state-of-the-art models across three key areas: task-oriented dialogue, e-commerce recommendation, and visual question answering. For task-oriented dialogue, we compared against SimpleTOD [25], a simple yet effective end-to-end model for dialogue state tracking. In e-commerce recommendation, we used SASREC [26] and GRU4Rec+ [27] as our baselines. For visual question answering, we compared against MCAN [28], a deep modular co-attention network.

Tables I, II, and III present the main results of our experiments, comparing CAM-Bot with these baseline models across different tasks and datasets.

TABLE I: Task-Oriented Dialogue Performance

Model	BLEU-4	Inform	Success
SimpleTOD	15.01	84.4%	70.1%
CAM-Bot*	15.53	85.7%	71.9%

In task-oriented dialogue, CAM-Bot demonstrated substantial improvements over Simple TOD. The BLEU-4 score increased from 15.01 to 15.53, representing a 3.5% relative improvement. The Inform rate improved from 84.4% to 85.7%, a 1.5% increase, while the Success rate rose from 70.1% to 71.9%, representing a 2.6% improvement.

TABLE II: Recommendation Performance

Model	NDCG@10
SASREC	0.4912
GRU4Rec+	0.4998
CAM-Bot*	0.5038

For e-commerce recommendation tasks, CAM-Bot achieved an NDCG@10 score of 0.5038, compared to SASREC's 0.4912 and GRU4Rec+'s 0.4998, representing a 2.6% relative improvement over SASREC and a 0.8% improvement over GRU4Rec+.

TABLE III: Visual Question Answering Performance

Model	VQA Accuracy
MCAN	66.14%
CAM-Bot*	66.72%

In visual question answering, CAM-Bot showed a modest but noteworthy improvement. The accuracy increased from 66.14% (MCAN) to 66.72%, a 0.9% relative improvement.

These results highlight the effectiveness of CAM-Bot's integrated approach across all three tasks. The Context Aggregation Module and Dynamic Knowledge Graph appear to play crucial roles in enhancing performance, particularly in maintaining coherence in multi-turn dialogues and providing contextually relevant recommendations.

While the improvements are consistent across all metrics, some are relatively small, suggesting room for further enhancements, particularly in the visual question answering component. Future work could focus on improving the integration of visual information with textual context to achieve more substantial gains in this area.

E. Key Findings

1) **Task-Oriented Dialogue:** CAM-Bot demonstrated significant improvements in task-oriented dialogues, particularly in e-commerce related conversations:

- **BLEU-4 Score:** Increased from 15.01 to 15.53, a 3.5% relative improvement. This enhancement in BLEU-4 score indicates that CAM-Bot generates more fluent and contextually appropriate responses. The improvement can be attributed to the Context Aggregation Module, which helps maintain coherence across multi-turn dialogues, and the Dynamic Knowledge Graph, which provides relevant domain-specific information for response generation.

- **Inform Rate:** Improved from 84.4% to 85.7%, a 1.5% increase. This metric measures the model's ability to provide correct information. The improvement suggests that CAM-Bot is more adept at understanding user queries and retrieving accurate information, likely due to its enhanced context understanding and knowledge integration capabilities.

- **Success Rate:** Rose from 70.1% to 71.9%, representing a 2.6% improvement. This metric assesses the model's ability to both provide correct information and complete the task. The more substantial improvement in Success Rate compared to Inform Rate indicates that CAMBot is particularly effective at maintaining goal-oriented conversations and guiding them to successful completion.

These improvements collectively suggest that CAM-Bot's architecture, particularly its context aggregation and dynamic knowledge integration components, contributes significantly to better task completion and response quality in e-commerce related conversations. The model appears to be more capable of understanding complex user intents, maintaining context over multiple turns, and providing relevant, accurate information to complete tasks successfully.

2) **E-commerce Recommendation:** In the domain of ecommerce recommendation, using the Amazon Review Data subset, CAM-Bot showed notable improvements:

- **NDCG@10:** Increased from 0.4912 to 0.5038, a 2.6% relative improvement over SASREC.

This enhancement in NDCG@10 is particularly significant in the context of e-commerce, where effective product recommendations can substantially impact user satisfaction and sales. The improvement can be attributed to several factors:

- **Multimodal Processing:** CAM-Bot's ability to process and integrate information from multiple

modalities (text descriptions, user reviews, product images) likely contributes to a more comprehensive understanding of products and user preferences.

- **Context Awareness:** The Context Aggregation Module allows CAM-Bot to consider the user's browsing history, previous purchases, and current session information, leading to more personalized recommendations.

- **Dynamic Knowledge Integration:** The Dynamic Knowledge Graph enables CAM-Bot to leverage broader product relationships and market trends, potentially identifying non-obvious but relevant recommendations.

These results indicate that CAM-Bot's architecture is well suited for e-commerce applications, where understanding complex user preferences and product relationships is crucial for effective recommendations.

3) **Visual Question Answering:** For the VQA task focused on product-related queries, CAM-Bot showed a modest but noteworthy improvement:

- **Accuracy:** Increased from 66.14% to 66.72%, a 0.9% relative improvement over MCAN.

While this improvement is smaller compared to other tasks, it is still significant considering the complexity of visual question answering in e-commerce contexts. Several factors contribute to this improvement:

- **Multimodal Integration:** CAM-Bot's architecture allows for effective integration of visual and textual information, enabling it to answer questions that require understanding both product images and associated text.

- **Contextual Understanding:** The model's ability to consider broader context (e.g., previous questions, user preferences) may help in answering ambiguous or context dependent questions about product images.

- **Knowledge Utilization:** The Dynamic Knowledge Graph might provide additional product information not directly visible in the image, enhancing the model's ability to answer detailed questions.

It's worth noting that this improvement was achieved with a more lightweight vision model (MobileNetV2), suggesting that CAM-Bot's architecture can effectively handle multimodal product inquiries even with computational constraints.

The modest gain in this area also highlights a potential direction for future improvements. Enhancing the visual processing capabilities or developing more sophisticated methods for integrating visual and textual information could lead to more substantial improvements in visual question answering tasks.

Overall, these key findings demonstrate CAM-Bot's versatility and effectiveness across various e-commerce related tasks. The consistent improvements across different metrics and task types suggest that the model's integrated approach to multimodal processing, context awareness, and knowledge utilization is particularly well-suited for complex, real-world e-commerce applications.

F. Ablation Study

To understand the contribution of each key component of CAM-Bot, we conducted an ablation study. We systematically removed or simplified major components and evaluated the performance on the MultiWOZ 2.1 subset, focusing on ecommerce domains. Table IV presents the results.

TABLE IV: Ablation study results on MultiWOZ 2.1 subset (e-commerce domains)

Configuration	BLEU-4	Inform (%)	Success (%)
Full CAM-Bot	15.53	85.7	71.9
w/o Context Aggregation	15.18	84.5	70.3
w/o Dynamic Knowledge Graph	15.37	85.1	71.1
w/o Adaptive Response	15.14	84.3	70.1

1) **Impact of Context Aggregation:** Removing the Context Aggregation Module led to the most

significant performance drop across all metrics (BLEU-4: -2.25%, Inform: -1.40%, Success: -2.23%). This underscores the crucial role of context in maintaining coherent, goal-oriented dialogues in ecommerce scenarios. Without effective context aggregation, the model struggles to maintain consistency across multi-turn conversations and fails to leverage historical interaction data for improved responses.

2) **Role of Dynamic Knowledge Graph:** The absence of the Dynamic Knowledge Graph resulted in a moderate performance decrease (BLEU-4: -1.03%, Inform: -0.70%, Success: -1.11%). This component's contribution is particularly evident in the Success rate, suggesting its importance in providing accurate, domain-specific information crucial for task completion. The relatively smaller impact compared to Context Aggregation indicates that while external knowledge is valuable, contextual understanding is even more critical in our e-commerce dialogue scenarios.

3) **Importance of Adaptive Response:** Removing the Adaptive Response component led to a substantial performance drop (BLEU-4: -2.51%, Inform: -1.63%, Success: -2.50%), second only to the Context Aggregation Module. This highlights the significance of tailoring responses to individual users and specific conversation contexts. The notable decrease in BLEU4 score suggests that this component plays a key role in generating more natural and contextually appropriate responses.

These results collectively emphasize the synergistic nature of CAM-Bot's architecture. While each component contributes to the overall performance, the Context Aggregation Module emerges as the most critical element, particularly for maintaining coherence and goal-orientation in e-commerce related dialogues. The study also reveals that the adaptive response generation, guided by both contextual understanding and domain knowledge, is essential for producing high-quality, task-relevant responses.

G. Error Analysis

To gain deeper insights into CAM-Bot's performance and identify areas for improvement, we conducted a comprehensive error analysis. We examined a sample of 200 interactions: 100 dialogues from the MultiWOZ 2.1 test subset (focusing on e-commerce domains) and 100 product recommendations from the Amazon Review Data test subset.

TABLE V: Distribution of error types in CAM-Bot

Error Type	Percentage
Misunderstanding of user intent	31%
Incorrect entity recognition	27%
Context maintenance failure	22%
Irrelevant or inaccurate information	20%

1) **Misunderstanding of User Intent (31%):** The most prevalent error type involved misinterpretation of user intentions or requirements:

- In 14% of these cases, CAM-Bot misunderstood complex or ambiguous queries.
- 10% involved confusion between similar but distinct user intents (e.g., comparing products vs. seeking recommendations).
- 7% were related to misinterpretation of implicit user preferences or constraints.

This suggests a need for more nuanced intent recognition, especially for complex e-commerce queries.

2) **Incorrect Entity Recognition (27%):** Errors in identifying and extracting relevant entities from user queries and product descriptions were the second most common issue:

- 12% of these errors occurred in parsing complex product names or model numbers.
- 9% involved misidentification of attributes or features, especially for technical products.
- 6% were related to incorrect recognition of brand names or product categories.

Enhancing the named entity recognition component, particularly for e-commerce specific entities and attributes, could address many of these issues.

3) **Context Maintenance Failure (22%):** CAM-Bot sometimes struggled to maintain context over extended multi-turn interactions:

- 10% of these errors involved forgetting or misapplying previously mentioned user preferences.
- 8% were cases where the model failed to track the current stage in a multi-step product inquiry or comparison.
- 4% related to inconsistencies in referring to the same product or feature across turns.

Improving the Context Aggregation Module's ability to retain and utilize relevant information from the conversation history could mitigate these issues.

4) **Irrelevant or Inaccurate Information (20%):** In some cases, CAM-Bot provided information or recommendations that were not aligned with user queries or were factually incorrect:

- 9% of these errors involved providing outdated or inaccurate product information.
- 7% were cases where the recommended products didn't match explicitly stated user requirements.
- 4% involved generating inconsistent or contradictory information across turns.

Enhancing the integration between the Dynamic Knowledge Graph and the response generation system could help in providing more accurate and relevant information.

These findings provide clear directions for future improvements to the CAM-Bot architecture. Key areas of focus should include:

1. Developing more sophisticated intent recognition mechanisms, capable of handling complex and nuanced e-commerce queries. 2. Improving the robustness of entity recognition, particularly for domain-specific terms, product attributes, and identifiers. 3. Enhancing context modeling to better track and utilize relevant information over extended conversations. 4. Refining the knowledge integration and response generation to ensure more accurate and consistent information delivery.

Addressing these issues could significantly enhance CAMBot's performance in e-commerce applications, leading to more accurate, context-aware, and user-centric interactions. It's worth noting that while these error rates indicate areas for improvement, they also reflect the challenging nature of complex, multi-turn e-commerce interactions. Even state-of the-art systems often encounter similar challenges in real world applications.

IV. PRACTICAL APPLICATIONS

Based on CAM-Bot's experimental results, we propose several high-impact applications for e-commerce and customer support. These applications leverage the system's key strengths while considering its current limitations.

A. Intelligent Product Discovery Assistant

Leveraging CAM-Bot's understanding of complex user intents (69% accuracy) and context maintenance, this application would guide users through product catalogs, refining searches based on preferences and multimodal inputs. It could integrate with existing product databases and recommendation systems, allowing for a personalized discovery process. Implementation could start with a limited product domain (e.g., electronics or fashion) before expanding to a full catalog, ensuring high accuracy and domain-specific knowledge.

B. Contextual Product Comparison Tool

Utilizing the system's product attribute understanding (73% accuracy in entity recognition) and context retention, this tool would offer in-depth, user-friendly comparisons of products based on expressed preferences. It would combine CAM-Bot with a structured product database and comparison engine, explaining differences in technical specifications using accessible language. Initial focus should be on comparing products within the same category, gradually expanding to cross-category comparisons as the system's capabilities improve.

C. Personalized Shopping Assistant

Combining context retention capabilities (78% accuracy) with recommendation generation, this application would provide personalized product suggestions and style advice based on user profiles and current context. It would integrate with user data, purchase history, and browsing behavior to offer tailored recommendations, size advice for fashion items, and suggest complementary products. Strong privacy measures and clear opt-in processes for data usage would be crucial for building user trust.

D. Multi-step Order Management

Building on CAM-Bot's success in task-oriented dialogues (71.9% success rate), this system would guide users through complex ordering processes, handle queries, and assist with post-purchase support. It would integrate with order management and logistics systems to provide real-time updates, handle returns and exchanges, and understand the context of original purchases. Implementation should start with simple queries, gradually increasing in complexity as the system proves reliable.

E. Visual Product Support

Harnessing CAM-Bot's visual question answering capabilities (66.72% accuracy), this application would answer queries about product appearance, assist in troubleshooting, and guide users through setup processes using visual aids. It would integrate with product image databases and potentially augmented reality technologies. Initial deployment could focus on products where visual features are crucial (e.g., furniture or appliances) before expanding to a broader range.

F. Intelligent FAQs and Support Routing

Utilizing natural language understanding and context awareness, this system would provide dynamic answers to FAQs, route complex queries to human agents, and analyze common issues to improve self-service options. It would integrate with existing FAQ databases and support ticketing systems. Implementing a confidence threshold for automated responses would ensure that only high-confidence answers are provided directly, with lower-confidence queries routed to human agents.

For all applications, gradual deployment with continuous monitoring and improvement based on real-world performance and user feedback is crucial. Regular analysis of interaction logs and customer feedback would drive iterative enhancements. Maintaining transparency about the AI nature of the system and providing easy access to human support remain essential for building user trust and ensuring positive experiences.

These applications demonstrate CAM-Bot's potential to significantly enhance customer experiences in e-commerce settings. However, careful consideration of current limitations and ethical implications, particularly regarding data privacy and transparency, is necessary for successful implementation.

V. CONCLUSION

This paper introduced CAM-Bot, a Context-Aware Multimodal Chatbot designed to enhance customer support across various digital business domains. Our experiments demonstrated CAM-Bot's effectiveness in handling complex, multiturn dialogues and understanding multimodal inputs in diverse scenarios. Key improvements over baseline models included a 3.5% increase in BLEU-4 score for task-oriented dialogues, a 2.6% improvement in Success Rate for task completion, a 2.6% relative improvement in NDCG@10 for e-commerce recommendations, and a 0.9% accuracy gain in visual question answering.

These results highlight CAM-Bot's capability to enhance performance across various aspects of customer support in digital businesses. The improvements in task-oriented dialogues and e-commerce recommendations are particularly noteworthy, demonstrating CAM-Bot's potential to significantly impact user experience and task completion in these domains. However, the more modest gains in visual question answering suggest potential for further enhancements in multimodal integration and visual processing capabilities.

CAM-Bot represents a significant advancement in AI powered customer support for digital businesses, effectively combining multimodal processing, context awareness, and adaptive response generation. While showing promising

results, our study also identifies areas for improvement and careful consideration in real-world deployments. As AI continues to evolve, systems like CAM-Bot are set to play a pivotal role in shaping the future of customer service across various digital platforms and services. Future work could focus on further improving the integration of visual and textual information, enhancing long-term context maintenance, and exploring applications in additional domains beyond e-commerce and customer support.

VI. FUTURE RESEARCH

Future research for CAM-Bot and similar systems in digital business environments should focus on several key areas. Enhancing multimodal integration techniques will enable better fusion of text, image, and voice inputs, while improving long term context modeling will maintain coherence in extended conversations. Developing dynamic knowledge acquisition methods and advancing personalization while ensuring privacy will be crucial for adapting to changing business environments and user needs.

Incorporating explainable AI features and enhancing robustness in handling ambiguous queries will increase transparency and improve real-world performance. Extending capabilities for cross-lingual and cultural adaptation will be essential for global businesses. Ensuring ethical AI practices, including fairness and bias prevention, will be paramount as these systems become more prevalent.

Exploring integration with emerging technologies like augmented reality and IoT could open new possibilities, while improving scalability and computational efficiency will be necessary for large-scale deployments. These research directions aim to address current limitations, expand capabilities, and ensure that AI-powered customer support systems can meet the evolving needs of diverse digital businesses and their customers effectively and ethically.

REFERENCES

1. eMarketer Editors, “Global ecommerce update 2021,” eMarketer, 2021. [Online]. Available: emarketer.com/content/global-ecommerce-update-2021
2. S. Metha, “Ai-driven promotion platforms: Increasing customer engagement in banking,” Journal of Artificial Intelligence Research & Advances, vol. 12, no. 01, 2025.
3. E. Adamopoulou and L. Moussiades, “An overview of chatbot technology,” Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, pp. 373–383, 2020.
4. T. Brown et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.
5. A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in International Conference on Learning Representations, 2021.
6. Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, “Multimodal machine learning: A survey and taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
7. X. Li, C. Chen, G. Haffari, and J. H. Shang, “A survey on multimodal dialog systems: Recent advances and new frontiers,” arXiv preprint arXiv:2208.13401, 2022.
8. C. Sankar, S. Subramanian, C. Pal, S. Chandar, and Y. Bengio, “How to compare neural response generation models?” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 1866–1875.
9. C. Zhang, Q. Wang, J. Wu, M. Tian, B. Zhang, and Xie, “Advances and challenges in conversational recommender systems: A survey,” AI Open, vol. 2, pp. 100–126, 2021.
10. P. P. Liang, Y. Xie, X. Liang et al., “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” Advances in Neural Information Processing Systems, vol. 35, pp. 23 356–23 369, 2022.
11. Y. Wu, J. Mao, W. Wright, T. Tian, and H. Cai, “Dialograph: Incorporating interpretable

strategy-graph networks into negotiation dialogues," in International Conference on Learning Representations, 2021.

12. S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," arXiv preprint arXiv:2004.13637, 2021.
13. V. S. Deshpande, "Optimizing retail media strategies to drive customer loyalty across industries," NOLEGEIN Journal of Supply Chain and Logistics Management, 2025.
14. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
15. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in Neural Information Processing Systems, vol. 33, pp. 12 449–12 460, 2020.
16. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510– 4520.
17. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.
18. M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, and D. Hakkani-Tur, "Multiwoz 2.1: A consolidated multi-domain dataset with state corrections and state tracking baselines," in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 422–428.
19. J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 188–197.
20. Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6904–6913.
21. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
22. P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gas˘ic´, "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for taskoriented dialogue modelling," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 5016–5026.
23. K. Ja¨rvelin and J. Keka¨la¨inen, "Cumulated gain-based evaluation of ir techniques," ACM Transactions on Information Systems (TOIS), vol. 20, no. 4, pp. 422–446, 2002.
24. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.
25. E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher, "Simple and effective end-to-end dialogue state tracking," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1405–1416.
26. W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018, pp. 197– 206.
27. B. Hidasi and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 843–852.

28. Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6281–6290.