mCS-LM: Multimodal Customer Service and Incident Management Systems Based on Large Language Models

Carlos Díaz-Morales[1], Marcos Checa-Rubio[1], Tomás Bernal-Beltrán[2], Ronghao Pan[2], David Barbáchano[1], María del Pilar Salas-Zárate[3,4], Mario Andrés Paredes-Valverde[3,4], Rafael Valencia-García[2]

[1] PANEL SISTEMAS INFORMÁTICOS S.L., C. de Josefa Valcárcel, 9, Cdad. Lineal, 28027 Madrid, Spain
[2] Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo 30100 Murcia, Spain
[3] Tecnológico Nacional de México/I.T.S. Teziutlán, Fracción I y II, Teziutlán 73960, Mexico
[4] Medio Melon, Av. Coyoacán 1435, Del Valle, Delegación Benito Juárez, Ciudad de México 03100 Mexico
carlos.diaz@panel.es, marcos.checa@panel.es, tomas.bernalb@um.es, ronghao.pan@um.es, david.barbachano@panel.es, maria.sz@teziutlan.tecnm.mx, mario.pv@teziutlan.tecnm.mx, valencia@um.es

Abstract

Customer service and incident management increasingly rely on multimodal evidence, combining text, images and audio. However, general-purpose models lack domain grounding, structured output control and reliability guarantees required in regulated enterprise environments, often leading to hallucinated responses and limiting their practical deployment. This paper presents mCS-LM, a multilingual multimodal framework that integrates Large Language Models (LLMs), Visual Language Models (VLMs), Audio Language Models (ALMs) and Retrieval-Augmented Generation (RAG) within a modular and traceable architecture tailored to customer service and incident management. The system introduces complementary processing flows: (i) perception modules for visual and audio understanding aligned with LLM-based reasoning, and (ii) structured report generation from multimodal evidence through supervised fine-tuning using QLoRA and efficient adaptation techniques. To mitigate hallucinations and improve factual reliability, the framework incorporates vector databases and multimodal RAG pipelines that retrieve domain-specific knowledge from external corporate sources. Formal structural schemas and validation mechanisms enforce output consistency and syntactic correctness. The platform is deployed as a web-based system with REST API integration, enabling scalable multimodal interaction across channels such as instant messaging, email and web chat. Experimental results demonstrate that multimodal generative models can be specialized for structured, domain-constrained enterprise tasks while maintaining computational viability and robustness.

Keywords: Multimodal Conversational Systems, Customer Service Automation, Large Language Models, Visual Language Models, Audio Language Models, Retrieval-Augmented Generation, Hallucination Mitigation

1. Introduction

Customer service and incident management have undergone significant digital transformation in recent years, driven by the need to handle large volumes of interactions efficiently while maintaining quality, personalization and multilingual support (Balasubramanian et al., 2018). In sectors such as insurance, e-commerce, telecommunications and logistics, customer interactions increasingly involve multimodal evidence, including textual descriptions, images documenting damages and audio recordings (Cui et al., 2017). Small and medium-sized enterprises (SMEs), however, often lack the technical and computational resources required to deploy advanced AI-driven solutions capable of processing and reasoning over such heterogeneous data (Luccioni et al., 2023).

Recent advances in Large Language Models (LLMs) (Zhao et al., 2023), Visual Language Models (VLMs) (Gan et al., 2022) and Audio Language Models (ALMs) (Su et al., 2025) have demonstrated strong capabilities in conversational reasoning, visual understanding and speech processing. Nevertheless, the direct adoption of these models in regulated enterprise environments remains challenging (Eling and Lehmann, 2018). General-purpose models are prone to hallucinations, struggle to produce domain-constrained structured outputs, require costly computational infrastructure and often lack integration with external corporate knowledge sources (Stoeckli et al., 2018; Lewis et al., 2020). Furthermore, existing solutions typically address text, image or audio processing in isolation, rather than providing a unified multimodal framework (Shumanov and Johnson, 2021; Xu et al., 2017). This gap limits the development of reliable and controllable multimodal assistants suitable for real-world customer service and incident management workflows.

This paper presents the mCS-LM project, a multilingual (including Spanish and English) multimodal framework designed to integrate LLMs, VLMs, ALMs and Retrieval-Augmented Generation (RAG) within a modular and scalable architecture tailored to customer service and incident management. The system combines perception modules for text, image and audio processing with

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 38–45

structured generation mechanisms and multimodal RAG pipelines supported by vector databases to systematically mitigate hallucinations and enhance factual grounding. In addition, formal structural schemas, output validation layers and controlled generation constraints enforce syntactic correctness, domain alignment and semantic consistency, providing explicit safeguards against unreliable responses in high-stakes enterprise environments. The project consortiun is formed by Panel Sistemas S.L. (Spain), TECNOMOD research group at the University of Murcia (Spain), Medio Melón S.A. (Mexico) and Instituto Tecnológico Teziutlán (Mexico), as part of an International Technological Cooperation Project with unilateral certification and monitoring (reference UNI-20240017).

The resulting platform is deployed as a web-based system with REST API integration, enabling seamless interaction across channels such as instant messaging, email and web chat, while supporting multilingual and multimodal communication in enterprise environments. As detailed in Section 3, the architecture is organized into modular components that separate perception, reasoning, retrieval and generation processes, allowing flexible adaptation to different customer service domains such as insurance, e-commerce, telecommunications and logistics.

The overall architecture integrates specialized multimodal processing modules (see Section 3.1) responsible for handling text, image and audio inputs through LLM, VLM and ALM components, whose outputs are aligned within a unified semantic representation space. These modules are connected to a multimodal RAG layer (see Section 3.2), which integrates external corporate knowledge sources, such as policy documents, procedural manuals and internal databases, stored in vector databases for efficient semantic search over domain-specific corporate knowledge sources. This retrieval layer conditions generation and reinforces factual grounding.

To ensure reliability in regulated enterprise contexts, the system combines multiple complementary mechanisms for hallucination mitigation (see Section3.3). Beyond grounding responses through RAG, a dedicated hallucination control module introduces intention-aware routing and controlled generation strategies. This module performs prior intent classification and interaction-type identification, conditioning the response generation process and explicitly constraining the model's output space according to domain-specific schemas and predefined prompts. Together with validation layers and controlled decoding mechanisms, these safeguards mitigate hallucinations and prevent syntactic or semantic inconsistencies.

The web interface (Section 3.4) provides user-facing interaction and REST API services within a scalable microservices architecture, facilitating cloud-ready deployment and seamless integration into existing enterprise infrastructures. From the end-user perspective, customers interact directly with the multimodal conversational layer powered by LLM, VLM and ALM components, receiving automated responses and structured incident reports. In parallel, company agents access a dedicated dashboard focused on operational management. This interface provides real-time analytics and configurable Key Performance Indicators (KPIs) related to registered incidents, detected intents, resolution status and escalation cases. It also enables agents to monitor system activity, review AI-generated outputs and intervene in complex or high-risk interactions requiring human supervision. This dual-layer design separates automated conversational processing from operational oversight, ensuring both scalability and controlled human involvement.

The platform is currently in its final stages of development and is undergoing validation within real operational environments in collaboration with customer service professionals from the participating companies. These domain experts evaluate the system using representative real-world scenarios, including both common and complex incident cases, allowing iterative refinement of multimodal understanding, structured report generation and hallucination control mechanisms.

This practitioner-driven validation process ensures that the framework aligns with practical workflow requirements, domain-specific constraints and the most critical edge cases encountered in day-to-day operations. Within this scope, the present work emphasizes the design and integration of the proposed architecture, while more comprehensive quantitative evaluation, including benchmarking against baseline models, is part of ongoing work as the system continues to evolve within real operational settings.

2. Background Information

The development of advanced customer service and incident management systems is driven by recent progress in Natural Language Processing (NLP) (Vaswani et al., 2017; Devlin et al., 2019) and multimodal Artificial Intelligence, including vision (He et al., 2016) and audio (Yang et al., 2025). LLMs have demonstrated remarkable capabilities in natural language understanding, contextual reasoning and text generation, enabling more flexible and conversational interfaces compared to traditional rule-based systems (Brown et al., 2020; Achiam et al., 2023). Transformer-based architectures have become the dominant paradigm, supporting multilingual interaction and domain adap-

tation through fine-tuning and in-context learning strategies (Radford et al., 2021; Li et al., 2023). However, their deployment in regulated enterprise environments introduces challenges related to hallucinations, computational cost and controllability (Balasubramanian et al., 2018; Huang et al., 2025).

In parallel, VLMs extend language reasoning to visual inputs, integrating image encoders with language decoders to enable multimodal understanding and generation (Radford et al., 2021). These models allow the interpretation of images such as damaged products, infrastructure incidents or technical documentation, transforming visual evidence into structured textual representations (Li et al., 2023). Similarly, ALMs leverage pretrained speech encoders such as Whisper or Wav2Vec to transcribe and interpret spoken interactions (Su et al., 2025), supporting voice-based incident reporting and customer support scenarios (Gung et al., 2023). Despite these advances, multimodal models are typically optimized for open-ended tasks rather than domain-constrained structured generation required in enterprise workflows.

To address factual inconsistencies and hallucinations, RAG architectures have gained prominence (Lewis et al., 2020). By combining generative models with external knowledge retrieval mechanisms supported by vector databases and semantic search, RAG systems improve grounding and ensure access to up-to-date domain-specific information. Additionally, model optimization techniques such as quantization, distillation and parameter-efficient fine-tuning methods like LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) have been proposed to reduce computational requirements while maintaining performance, facilitating deployment in SMEs environments.

Although significant progress has been made in text, vision and speech processing independently, comprehensive frameworks that integrate LLMs, VLMs, ALMs and RAG within a unified, controllable and enterprise-ready architecture remain limited. Existing solutions often lack explicit mechanisms for structured output enforcement, hallucination mitigation and seamless integration into operational customer service infrastructures. This gap motivates the development of a modular multimodal framework capable of combining perception, retrieval, controlled generation and deployment scalability within a single system.

3. System Architecture

Figure 1 shows the overall architecture of the system. The platform is organized into five main modules. The first module corresponds to the multimodal processing layer (see Section 3.1), which is responsible for handling text, image and audio inputs through domain adjusted LLM, VLM and ALM models. This module performs perception, encoding and semantic alignment of heterogeneous data sources, transforming multimodal inputs into unified representations suitable for downstream reasoning.

The second module focuses on the RAG pipeline (see Section 3.2), which connects the generative models with external corporate knowledge bases. Through vector databases and semantic search mechanisms, this layer retrieves relevant domain-specific information, such as policy documents and procedural manuals, to ground responses and enhance factual consistency.

The third module addresses structured output and hallucination control (see Section 3.3). In addition to enforcing domain-specific schemas, controlled decoding strategies and validation mechanisms, this module incorporates an intention-aware classification and routing layer that conditions the response generation process before decoding takes place. By identifying the interaction type and constraining the generation space through predefined prompts and domain-aligned templates, the system explicitly restricts the model's behavior, reducing the likelihood of producing unsupported, out-of-scope or fabricated content. These combined safeguards ensure syntactic correctness, semantic consistency and reliable responses in regulated enterprise contexts.

The fourth module corresponds to the incident management system interface (see Section 3.4), which provides the user-facing conversational layer and REST API services. This component enables seamless integration with enterprise communication channels such as instant messaging, email and web chat, supporting multilingual (including Spanish and English) and multimodal interaction workflows.

The fifth module corresponds to the operational dashboard (see Section 3.5), designed for company agents and supervisors. This interface provides real-time analytics and configurable KPIs related to registered incidents, detected intents, resolution status and escalation cases. It enables monitoring of system activity, review of AI-generated outputs and human intervention in complex or high-risk interactions. By separating conversational automation from operational oversight, the architecture ensures scalability while maintaining controlled human involvement in enterprise workflows. The following subsections describe these modules in more detail.

The platform is designed to be offered under a Software-as-a-Service (SaaS) model, allowing flexible adoption by SMEs. Basic functionality supports multimodal interaction and structured incident reporting, while advanced features, such as extended knowledge base integration, customiz-

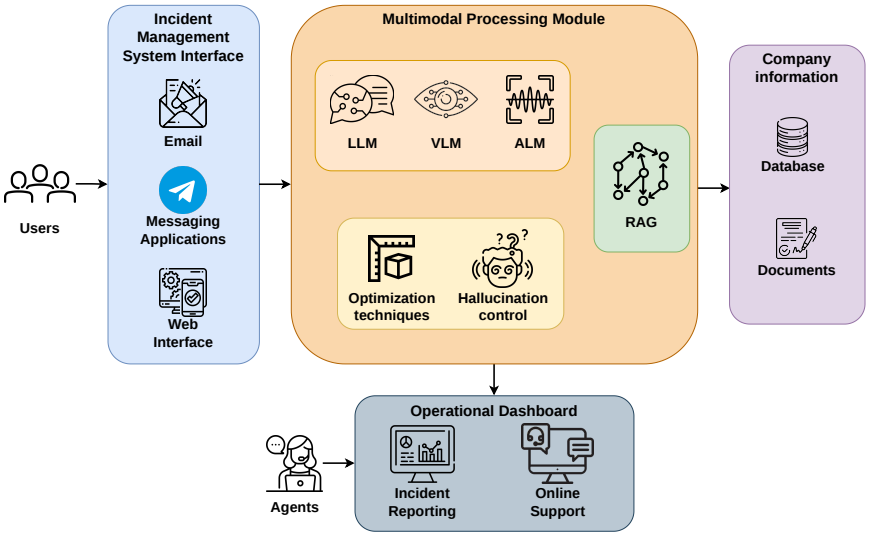

Figure 1: Overview of the modules that conform the mCS-LM system architecture.

able workflows and enterprise-level scalability, can be configured according to organizational requirements. All services and software components are containerized using Docker, enabling modular deployment, horizontal scalability and flexible orchestration depending on performance constraints and infrastructure availability.

3.1. Multimodal Processing Modules

The multimodal processing layer constitutes the perception and semantic encoding backbone of the mCS-LM architecture. It is responsible for handling heterogeneous inputs (text, images, and audio) through specialized domain-adapted models, enabling unified downstream reasoning and structured response generation.

To ensure computational viability in enterprise environments, all multimodal components, including LLMs, VLMs and ALMs, are optimized through techniques such as quantization and parameter-efficient fine-tuning using QLoRA. These strategies reduce hardware requirements while preserving performance, facilitating deployment in SMEs infrastructures.

Textual interactions are processed through LLMs fine-tuned for customer service and incident management tasks. Domain adaptation is achieved through parameter-efficient fine-tuning strategies, including QLoRA, allowing the system to specialize in policy interpretation, claim management and incident reporting scenarios while maintaining computational efficiency.

Visual inputs are handled by VLMs to transform visual evidence, such as photographs of damages, defective products or infrastructure incidents, into structured textual representations. The VLM component operates under two complementary processing strategies. In the descriptive flow, visual content is first converted into a controlled textual description that is subsequently processed by the LLM reasoning module, separating perception from linguistic inference. In the generative flow, a fine-tuned VLM produces domain-specific structured outputs (e.g., JSON incident reports) from raw images. This dual design enables both modular reasoning over image-derived descriptions and end-to-end structured generation aligned with predefined reporting schemas.

Audio interactions are processed through ALMs to extract contextual embeddings from spoken inputs. These embeddings are projected into a shared semantic space aligned with the LLM decoder, allowing the system to perform transcription, intent understanding and response generation from voice-based interactions. This design supports multilingual speech inputs and enables seamless integration of audio within multimodal conversational workflows.

To ensure coherent multimodal reasoning, the outputs of the LLM, VLM and ALM components are aligned within a unified semantic representation space. This alignment facilitates late-fusion integration strategies, enabling the system to combine textual, visual and acoustic information before retrieval and controlled generation stages. By separating perception from reasoning and maintaining modular specialization of each modality, the architecture achieves both flexibility and scalability across diverse enterprise domains.

3.2. Retrieval-Augmented Generation Pipeline

The RAG pipeline enhances the generative capabilities of the multimodal models by grounding responses in external domain-specific knowledge sources. Instead of relying exclusively on the parametric knowledge encoded within the LLM, the system dynamically retrieves relevant information from corporate repositories before response generation.

Domain documents, including policy manuals, procedural guidelines and internal databases, are preprocessed, segmented and embedded into vector representations using sentence-level embedding models. These embeddings are stored in vector databases that enable efficient semantic search through similarity-based retrieval mechanisms. When a user query is received, its representation is used to retrieve the most relevant contextual fragments, which are then incorporated into the generation prompt.

To ensure that retrieved information remains up-to-date in dynamic enterprise environments, the system supports incremental updates of the vector database. New or modified documents can be embedded and indexed without requiring full re-indexing, enabling efficient maintenance of the knowledge base. Additionally, document-level metadata, such as timestamps and versioning information, is incorporated into the retrieval process to prioritize more recent and relevant content.

When potentially conflicting information is retrieved (e.g., legacy documents versus updated protocols), the system applies a ranking strategy that combines semantic similarity, document recency and source reliability. This prioritization ensures that the most current and authoritative information is used to condition generation, while the structured output module enforces consistency in the final response delivered to the user.

This retrieval layer conditions the decoding process, reinforcing factual consistency and reducing the likelihood of hallucinated or outdated responses. By integrating structured corporate knowledge into the generation workflow, the system ensures that responses remain aligned with current policies, operational procedures and domain constraints.

The architecture also supports multimodal retrieval scenarios, allowing textual queries to be associated with information derived from images or audio metadata when available. This design extends traditional text-only RAG approaches and enables unified grounding across heterogeneous enterprise data sources.

3.3. Structured Output and Hallucination Control

Ensuring reliability in regulated enterprise environments requires explicit mechanisms to control the behavior of generative models. Beyond retrieval-based grounding, the mCS-LM framework incorporates a dedicated structured output and hallucination control module designed to constrain generation and reduce unsupported or out-of-scope responses.

The control pipeline introduces a hierarchical intention-aware classification process prior to response generation. Each incoming user message is first processed by a global intent classifier that performs a high-level categorization of the interaction (e.g., conversational intent, domain-specific action, policy-related claim, out-of-scope request or unspecified issue). This initial classification determines whether the interaction can be directly mapped to a predefined response strategy or requires further specialized analysis.

Messages assigned to intermediate categories are routed to a second-stage classifier tailored to the corresponding interaction type. For instance, conversational intents are processed by a dialogue-oriented classifier, while domain-specific actions related to incident management or policy handling are redirected to dedicated domain classifiers. This two-step classification strategy reduces task complexity, isolates interaction types and provides finer control over subsequent generation stages.

The outcome of this hierarchical classification process conditions the prompt selection and generation strategy used by the LLM responsible for producing the final response. Instead of generating outputs directly from the raw input, the model operates under predefined domain-aligned templates and constrained decoding configurations associated with the detected interaction type. This intention-aware routing mechanism explicitly restricts the model's output space, reducing the likelihood of hallucinated, ambiguous or semantically inconsistent responses.

In addition, structured generation schemas are enforced for tasks requiring formal outputs, such as incident reports or claim summaries. Validation layers verify syntactic correctness and schema compliance before responses are delivered to the user or forwarded to downstream systems. Together, hierarchical intent classification, controlled prompt selection, constrained decoding and schema validation form a multi-layer safeguard strategy that enhances reliability, interpretability and domain alignment in enterprise workflows.

3.4. Incident Management System Interface

The incident management system interface constitutes the user-facing interaction layer of the mCS-LM framework. It enables customers to report incidents, submit supporting evidence and receive automated assistance through multimodal conversational workflows.

The interface supports text, image and audio inputs across multiple communication channels, including web chat, email and instant messaging platforms. Incoming interactions are routed to the appropriate multimodal processing modules, triggering the perception, retrieval and controlled generation pipeline described in previous sections. The system manages conversational context, session state and interaction history to ensure coherent multi-turn exchanges.

Through REST API services, the interface can be integrated into existing enterprise infrastructures, customer portals or third-party applications. This API-driven design allows flexible orchestration of requests, enabling automated ticket creation, structured incident reporting and interaction logging within corporate management systems.

Multilingual support is provided at both the understanding and generation levels, allowing customers to interact in different languages without requiring manual intervention. By abstracting the complexity of the underlying multimodal and retrieval components, the interface delivers a seamless conversational experience while maintaining alignment with domain constraints and enterprise policies.

3.5. Operational Dashboard

The operational dashboard provides the management and supervision layer of the mCS-LM framework, designed for company agents and administrative personnel. While the incident management system interface enables automated multimodal interaction with end users, the dashboard focuses on operational oversight, monitoring and human intervention when required.

The dashboard offers real-time visibility into registered incidents, detected intents, resolution status and escalation cases through configurable KPIs. These analytics allow supervisors to monitor system performance, identify high-risk or ambiguous interactions and assess workload distribution across incident categories. Interaction-level metadata, including intent classifications and retrieval traces, can be inspected to enhance transparency and traceability.

In addition to monitoring capabilities, the dashboard enables human-in-the-loop intervention. Agents can review AI-generated outputs, validate structured reports, modify responses when necessary and manually escalate complex or sensitive cases. This supervision layer ensures that automated decision-making remains aligned with organizational policies and regulatory requirements.

The operational dashboard is integrated within the same microservices architecture as the conversational interface, allowing synchronized access to interaction logs, structured reports and knowledge base updates. By separating conversational automation from operational management, the system achieves scalability without sacrificing control, accountability or human expertise in critical enterprise workflows.

4. Conclusions and Further Work

The current version of mCS-LM integrates the core modules required for multimodal incident management, including domain-adapted LLM, VLM and ALM components, a RAG pipeline, a structured output and hallucination control module and dedicated interfaces for both end users and operational agents. The architecture combines perception, retrieval, controlled generation and human-in-the-loop supervision within a modular and scalable microservices framework suitable for enterprise environments.

The platform is currently undergoing validation in real operational settings with customer service professionals from participating companies. This validation phase focuses on assessing multimodal understanding accuracy, structured report consistency and the effectiveness of hierarchical intent classification in reducing hallucinations and out-of-scope responses. In this context, the present work emphasizes the design and integration of the proposed architecture, while more comprehensive quantitative evaluation, including benchmarking against baseline models, is part of ongoing work as the system continues to evolve within real operational settings.

Future work will focus on several directions. First, we plan to enhance the multimodal RAG pipeline by incorporating more advanced cross-modal retrieval strategies, enabling tighter alignment between textual queries and visual or audio-derived knowledge representations, as highlighted in recent studies on cross-modal RAG systems (Abootorabi et al., 2025). Second, we will explore adaptive intent classification mechanisms based on continual learning, allowing the hallucination control module to evolve as new interaction patterns emerge, in line with emerging research on continual learning for generative AI frameworks (Wang et al., 2024). Third, we aim to investigate automated feedback loops between the operational dashboard and the generation modules, leveraging agent corrections to refine prompts and improve structured output quality over

time, a strategy that aligns with recent proposals for integrating human feedback into generative systems (Sun et al., 2023).

These developments will strengthen the robustness, scalability, and practical applicability of multimodal conversational systems in regulated enterprise contexts. Additionally, future work will explore the extension of the proposed framework to low-resource settings, including regional languages and dialectal variations, leveraging its modular and language-agnostic design to support broader linguistic diversity and robustness in real-world multilingual scenarios.

Acknowledgements

This work is being funded by CDTI and the European Regional Development Fund (ERDF EU-/FEDER UE)-a way of making Europe, through project mCS-LM IDI-20250122.

5. Ethical Considerations and Limitations

The data used for system validation and adaptation consist of enterprise-provided incident records, policy documents and internally generated test cases. All materials are handled within controlled corporate environments in accordance with applicable data protection regulations. No personal or sensitive information is publicly released as part of this research.

The system is designed to operate within enterprise infrastructures, ensuring that multimodal inputs (including text, images and audio) are processed under organizational data governance policies. When required, anonymization and access control mechanisms are applied to prevent unauthorized exposure of customer information.

6. Bibliographical References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ramnath Balasubramanian, Ari Libarikian, and Doug McElhaney. 2018. Insurance 2030—the impact of ai on the future of insurance. *McKinsey & Company*, pages 1–10.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, system demonstrations*, pages 97–102.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Martin Eling and Martin Lehmann. 2018. The impact of digitalization on the insurance value chain and the insurability of risks. *The Geneva papers on risk and insurance-issues and practice*, 43(3):359–396.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*.

James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. 2023. Natcs: Eliciting natural customer support dialogues. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9652–9677.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of machine learning research*, 24(253):1–15.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Michael Shumanov and Lester Johnson. 2021. Making conversations with chatbots more personalized. *Computers in human behavior*, 117:106627.

Emanuel Stoeckli, Christian Dremel, and Falk Uebernickel. 2018. Exploring characteristics and transformational capabilities of insurtech innovations to understand insurance value creation in a digital world. *Electronic markets*, 28(3):287–305.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*.

Xin Sun, Jos A Bosch, Jan De Wit, and Emiel Krahmer. 2023. Human-in-the-loop interaction for continuously improving generative model in conversational agent for behavioral intervention. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 99–101.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10181.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).