

A Multimodal Framework for Customer Service and Incident Management Using Large Language Models

Carlos Díaz-Morales¹, Marcos Checa-Rubio¹, Tomás Bernal-Beltrán², Ronghao Pan², David Barbáchano¹, María del Pilar Salas-Zárate^{3,4}, Mario Andrés Paredes-Valverde^{3,4} and Rafael Valencia-García^{2,*}

¹PANEL SISTEMAS INFORMÁTICOS S.L., C. de Josefa Valcárcel, 9, Cdad. Lineal, 28027 Madrid, Spain

²Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo 30100 Murcia, Spain

³Tecnológico Nacional de México/I.T.S. Teziutlán, Fracción I y II, Teziutlán 73960, Mexico

⁴Medio Melon, Av. Coyoacán 1435, Del Valle, Delegación Benito Juárez, Ciudad de México 03100 Mexico

Abstract

Modern customer service and incident management scenarios require handling heterogeneous data sources, including textual descriptions, images and audio signals. However, deploying general-purpose generative models in regulated enterprise environments remains challenging due to limited domain grounding, lack of output control and susceptibility to hallucinations. In this work, we introduce mCS-LM, a multilingual multimodal system that integrates Large Language Models, Visual Language Models, Audio Language Models and Retrieval-Augmented Generation mechanisms within a unified and modular architecture for customer service and incident management. The system combines multimodal perception with controlled generation to produce structured, domain-consistent outputs from heterogeneous inputs. To enhance reliability, the system incorporates vector databases and retrieval pipelines over corporate knowledge sources, together with schema-driven generation and validation layers that enforce consistency and syntactic correctness. The resulting platform is implemented as a web-based solution with REST API integration, enabling scalable deployment across real-world enterprise settings.

Keywords

Multimodal Generative Models, Retrieval-Augmented Generation, Hallucination Mitigation, Customer Service,

1. Introduction

Customer service and incident management have experienced a significant transformation in recent years, driven by the need to efficiently process large volumes of interactions while preserving quality, personalization and multilingual capabilities [1]. In domains such as insurance, e-commerce, telecommunications and logistics, customer interactions increasingly incorporate heterogeneous sources of information, including textual descriptions, images and audio recordings [2]. Despite these advances, small and medium-sized enterprises (SMEs) often lack the technical and computational resources required to deploy AI solutions capable of effectively handling such multimodal data [3].

Recent developments in Large Language Models (LLMs) [4], Visual Language Models (VLMs) [5] and Audio Language Models (ALMs) [6] have demonstrated strong performance across language understanding, visual reasoning and speech processing tasks. However, their direct application in regulated enterprise environments remains limited [7]. General-purpose models are prone to generating hallucinated content, lack mechanisms to enforce domain-specific constraints, require substantial

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

*Corresponding author.

✉ carlos.diaz@panel.es (C. Díaz-Morales); marcos.checha@panel.es (M. Checa-Rubio); tomas.bernalb@um.es (T. Bernal-Beltrán); ronghao.pan@um.es (R. Pan); david.barbachano@panel.es (D. Barbáchano); maria.sz@teziutlan.tecnm.mx (M. d. P. Salas-Zárate); mario.pv@teziutlan.tecnm.mx (M. A. Paredes-Valverde); valencia@um.es (R. Valencia-García)

🌐 <https://www.panel.es> (C. Díaz-Morales); <https://www.panel.es> (M. Checa-Rubio); <https://www.panel.es> (D. Barbáchano); <https://webs.um.es/valencia> (R. Valencia-García)

🆔 0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0008-7317-7145 (R. Pan); 0000-0003-1818-3434 (M. d. P. Salas-Zárate); 0000-0001-9508-9818 (M. A. Paredes-Valverde); 0000-0003-2457-1791 (R. Valencia-García)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

computational resources and are often not integrated with external corporate knowledge sources [8, 9]. Moreover, existing approaches typically address individual modalities in isolation, rather than providing unified solutions capable of jointly processing text, images and audio [10, 11]. These limitations hinder the development of reliable and controllable multimodal systems for real-world customer service and incident management.

In this work, we present mCS-LM, a multilingual (including Spanish and English) multimodal system that integrates LLMs, VLMs, ALMs and Retrieval-Augmented Generation (RAG) within a modular architecture tailored to customer service and incident management scenarios. The proposed system combines modality-specific perception components with controlled generation strategies, enabling the transformation of heterogeneous inputs into structured, domain-consistent outputs. To improve reliability and factual grounding, the system incorporates multimodal RAG pipelines over corporate knowledge sources, supported by vector databases that dynamically inject relevant contextual information into the generation process. In addition, structured output schemas, validation layers and constrained decoding mechanisms are employed to enforce syntactic correctness, semantic consistency and alignment with domain requirements, providing explicit safeguards against unreliable responses in high-stakes enterprise environments. The project consortium 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 system is deployed as a web-based platform with REST API integration, enabling seamless interaction across communication channels such as web chat, email and instant messaging, while supporting multilingual and multimodal workflows in enterprise environments. As detailed in Section 3, the architecture follows a modular design that separates perception, reasoning, retrieval and generation processes, allowing flexible adaptation to different application domains.

At the system level, multimodal processing modules (see Section 3.1) handle text, image and audio inputs through LLM, VLM and ALM models, whose outputs are projected into a shared semantic representation space. These representations are connected to a multimodal retrieval layer (see Section 3.2) that integrates external knowledge sources, including policy documents, procedural manuals and internal databases. This multimodal RAG component conditions the generation process, reinforcing factual consistency and reducing reliance on parametric knowledge alone.

To address reliability issues in regulated enterprise contexts, the system incorporates a dedicated hallucination control module (see Section 3.3) that constrains the behavior of generative models. Beyond grounding responses through RAG, this module introduces intention-aware routing mechanisms, where user inputs are first classified and then processed using predefined prompts and domain-specific templates, conditioning the response generation process. Combined with validation procedures and controlled decoding strategies, this approach mitigates hallucinations and ensures consistent, structured outputs.

From an interaction perspective, the platform provides two complementary interfaces (Section 3.4) and REST API services within a scalable microservices architecture, facilitating cloud-ready deployment and seamless integration into existing enterprise infrastructures. Regarding the web interface, customers interact with a multimodal conversational system powered by LLM, VLM and ALM components that supports automated assistance and structured incident reporting. In parallel, operational agents access a dedicated dashboard that 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 separation between automated interaction and human supervision allows the system to scale while maintaining control in critical enterprise workflows.

The system is currently in an advanced stage of development and is being evaluated in real operational environments in collaboration with domain experts. This validation process involves representative real-world scenarios, enabling iterative refinement of multimodal understanding, structured generation and control mechanisms. Within this context, the present work focuses on the design and integration of

the proposed architecture, while more comprehensive quantitative evaluation, including benchmarking against baseline approaches, remains part of ongoing work.

2. Background Information

Recent breakthroughs in multimodal Artificial Intelligence, spanning both vision and audio, as well as in Natural Language Processing (NLP) [12, 13], have played a key role in reshaping customer service and incident management solutions. These technological improvements have enabled the rise of large language models, which excel at interpreting language, reasoning within context, and generating coherent text. As a result, they support more natural and interactive user experiences than traditional rule-based systems [14, 15]. At the core of this transformation are transformer-based models, which have become the dominant framework due to their ability to handle multiple languages and adapt to specific domains through approaches like fine-tuning and in-context learning [16, 17]. Despite these advantages, integrating such models into regulated enterprise environments remains complex. Key concerns include the risk of hallucinated outputs, the significant computational resources required, and the ongoing challenge of maintaining consistent and controllable system behavior [1, 18].

At the same time, multimodal approaches have extended the capabilities of language models beyond the text alone. VLMs, for instance, combine visual encoders with language generation components, making it possible to reason over images and produce textual outputs [16, 17]. This enables practical applications such as analyzing photos of damaged equipment, identifying infrastructure issues, or extracting information from technical documents by converting visual inputs into structured descriptions. In a similar vein, ALMs rely on pretrained speech processing systems like Whisper or Wav2Vec to handle spoken data [6]. These models can transcribe and interpret voice interactions, making them especially useful for tasks such as voice-driven incident reporting or customer support [19]. However, despite their growing capabilities, these multimodal systems are generally designed for broad, open-ended use cases. As a result, they often struggle to meet the strict requirements of enterprise environments, where outputs must follow predefined structures and adhere to domain-specific constraints.

RAG has emerged as a key approach to mitigate hallucinations and enhance the reliability of model outputs [9]. By linking generative models with external knowledge sources, typically accessed through vector databases and semantic search. These systems can retrieve relevant information at inference time, leading to more accurate and context-aware responses grounded in up-to-date, domain-specific data. In parallel, a range of efficiency-focused techniques has been developed to make these models more accessible in real-world settings. Approaches such as quantization and distillation reduce computational overhead, while parameter-efficient fine-tuning methods like LoRA [20] and QLoRA [21] allow models to be adapted with significantly fewer resources. These advances are particularly valuable for enabling deployment in SMEs environments, where computational capacity may be limited.

Despite notable advances in text, vision, and speech technologies when developed separately, there is still a lack of unified frameworks that effectively combine LLMs, VLMs, ALMs, and RAG into a single, cohesive system suitable for enterprise use. Current solutions often fall short in key areas, such as enforcing structured outputs, reducing hallucinations, and integrating smoothly with existing customer service operations. This limitation highlights the need for a more comprehensive approach. In particular, it motivates the design of a modular multimodal architecture that can seamlessly bring together perception, knowledge retrieval, controlled text generation, and scalable deployment within one integrated system.

3. System Architecture

Figure 1 illustrates the overall structure of the proposed system. The architecture is composed of five main modules that together enable multimodal interaction, controlled generation and operational management within enterprise environments.

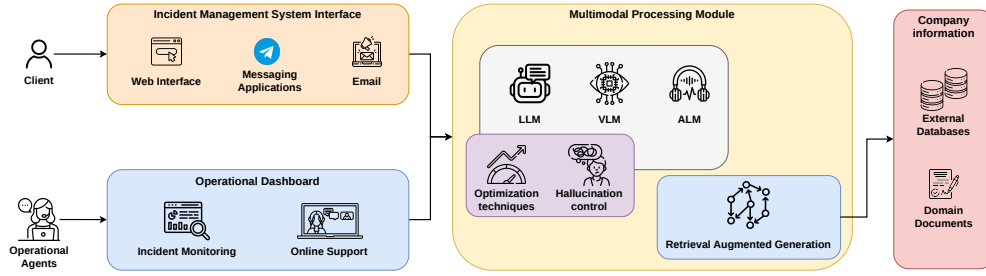


Figure 1: Overview of the modules that conform the mCS-LM system architecture.

The first module corresponds to the multimodal processing layer (see Section 3.1). This component is responsible for ingesting and processing text, image and audio inputs using domain-adapted LLM, VLM and ALM models. It performs perception, feature extraction and semantic alignment, converting heterogeneous inputs into unified representations that can be used for downstream reasoning and generation.

The second module implements the RAG pipeline (see Section 3.2), which links generative models with external corporate knowledge sources. By leveraging vector databases and similarity-based retrieval mechanisms, this module identifies relevant domain-specific information, such as policy documents and procedural guidelines, to enrich the context used during response generation and improve factual grounding.

The third module addresses structured output generation and hallucination control (see Section 3.3). In addition to enforcing domain-specific schemas, validation procedures and constrained decoding strategies, this component incorporates an intention-aware routing mechanism that guides the generation process. By classifying user inputs and mapping them to predefined prompts and templates, the system restricts the output space of the model, reducing the risk of producing inconsistent, out-of-scope or unsupported responses. These mechanisms collectively ensure syntactic validity, semantic coherence and reliable behavior in regulated enterprise settings.

The fourth module corresponds to the incident management system interface (see Section 3.4). This layer provides the main interaction channel for end users, exposing conversational capabilities and REST API services. It enables integration with enterprise communication platforms, including web chat, email and instant messaging systems, while supporting multilingual (including Spanish and English) and multimodal interaction scenarios.

The fifth module is the operational dashboard (see Section 3.5), designed for company agents and supervisors. This interface offers real-time monitoring capabilities through configurable KPIs related to incidents, detected intents, resolution status and escalation processes. It also supports human-in-the-loop intervention, allowing users to review system outputs and handle complex or sensitive cases. 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 following a Software-as-a-Service (SaaS) model, facilitating adoption by SMEs. Core functionalities include multimodal interaction and structured incident reporting, while advanced features, such as extended knowledge base integration, workflow customization and scalable deployment, can be configured according to organizational needs. All components are containerized using Docker, enabling modular deployment, horizontal scalability and flexible orchestration across different infrastructure environments.

The architecture is intentionally designed to be model-agnostic. Rather than depending on a fixed set of LLM, VLM, ALM, embedding or vector database technologies, each component can be replaced or adapted according to the requirements of the target enterprise environment. This design enables the framework to incorporate different model families, retrieval backends and deployment configurations

depending on domain constraints, available infrastructure, data privacy requirements and computational resources. As a result, mCS-LM can evolve with future advances in multimodal AI while preserving the same modular processing, retrieval, control and supervision workflow.

3.1. Multimodal Processing Modules

Within the mCS-LM architecture, multimodal processing acts as the foundational layer that enables the system to interpret and represent information coming from different source. It is designed to handle heterogeneous data (text, image, and audio), by assigning each modality to a specialized model, while ensuring that all outputs can be jointly understood in later stages for reasoning and structured response generation.

To make this architecture feasible in real-world enterprise settings, efficiency is a key consideration. All underlying models, including language, vision, and audio components, are optimized using strategies such as quantization and parameter-efficient tuning methods like QLoRA. These techniques significantly lower computational and hardware requirements without sacrificing model capability, allowing deployment even in resource-constrained environments.

Text inputs are managed through language models tailored to domain like customer support and incident handling. By applying lightweight adaptation techniques, these models can focus on tasks such as interpreting policies, processing claims, or generating incident reports, all while maintaining operational efficiency.

For image data, VLMs are responsible for extracting meaningful information from visual inputs such as damage photos or technical issues. The system supports two distinct processing modes. In one approach, images are first translated into controlled textual descriptions that are later analyzed by the language model, effectively decoupling visual perception from reasoning. In the alternative approach, the vision model directly produces structured outputs, such as formatted reports, based on predefined schemas. This combination provides both flexibility and direct E2E processing capabilities.

Audio data is handled through dedicated speech-based models that convert spoken input into rich semantic representations. These representations are aligned with the same latent space used by the language model, enabling unified processing of transcription, intent recognition, and response generation. This setup supports multilingual speech and integrates naturally into conversational systems.

Finally, all modalities converge within a shared semantic representation space, allowing their outputs to be combined through integration strategies such as late fusion. This unified representation ensures that information from text, images, and audio can be jointly leveraged before retrieval and controlled generation take place. By keeping perception modular and separating it from reasoning, the system achieves a balance between flexibility, scalability, and robustness across different enterprise use cases.

3.2. Retrieval-Augmented Generation Pipeline

The RAG component strengthens the generation process by incorporating external domain-specific knowledge sources into the response pipeline. Rather than relying solely on the parametric knowledge encoded within the LLM, the system augments its outputs with information dynamically retrieved from corporate data sources.

Enterprise documents, such as policy manuals, operational procedures 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 maintain the relevance of the knowledge base in evolving enterprise environments, the system allows incremental updates of the vector store, ensuring that retrieved information remains up-to-date. Newly added or modified documents can be embedded and indexed independently, avoiding the need for

full re-indexing and 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.

In cases where retrieved content may be inconsistent, for example, when outdated documents coexist with updated procedures, the system employs a ranking strategy that integrates semantic similarity, document recency and source reliability. This approach prioritizes the most accurate and up-to-date knowledge, while subsequent generation stages enforce consistency in the final response delivered to the user.

By conditioning the decoding process on retrieved context, the RAG module improves factual grounding and reduces the likelihood of hallucinated or outdated responses. The integration of structured corporate knowledge into the generation workflow ensures alignment with current policies and operational constraints.

Furthermore, the system supports multimodal retrieval settings, where textual queries can be linked to information derived from visual or audio-related metadata when available. This capability extends conventional text-based retrieval approaches and enables consistent grounding across heterogeneous enterprise data sources.

3.3. Structured Output and Hallucination Control

Achieving reliable behavior in regulated enterprise scenarios requires explicit control over the generation process. In addition to retrieval-based grounding, the proposed system incorporates a dedicated module for structured output generation and hallucination mitigation, aimed at constraining model responses and preventing unsupported or out-of-scope content.

The control mechanism is based on a hierarchical intent-aware routing pipeline applied prior to generation. Each user request is initially processed by a global intent classifier that assigns a high-level interaction category, such as conversational intent, domain-specific action, policy-related request, out-of-scope query or unspecified issue. This first-stage classification determines whether the interaction can be directly mapped to a predefined response strategy or requires further specialized analysis.

Inputs assigned to intermediate categories are subsequently passed to specialized classifiers tailored to specific interaction types. For example, conversational requests are handled by dialogue-oriented models, whereas actions related to incident management or policy processing are directed to domain-specific classifiers. This multi-stage strategy decomposes the classification task, reducing its complexity, and enables more precise control over the generation workflow.

The result of the classification process governs both the prompt selection and generation strategy used by the LLM responsible for producing the final response. Instead of generating responses directly from the original input, the system relies on predefined templates and domain-specific prompts associated with each interaction type. This intention-aware routing mechanism explicitly restricts the generation space, reducing the likelihood of hallucinated, ambiguous or semantically inconsistent responses.

For tasks requiring formal outputs, such as incident reports or claim summaries, the system enforces structured schemas during generation. These outputs are further processed by validation layers that check syntactic correctness and schema compliance before being returned to the user or integrated into downstream systems. Overall, the combination of 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 interface represents the primary interaction layer between end users and the mCS-LM system. It allows customers to submit incident reports, provide supporting multimodal evidence and receive automated assistance through conversational interactions.

The interface is designed to handle text, image and audio inputs across various communication channels, including web chat, email and instant messaging services. User requests are automatically

routed to the corresponding multimodal processing components, activating the perception, retrieval and controlled generation pipeline described in previous sections. To ensure coherent interactions, the system maintains conversational context, session information and dialogue history, enabling consistent multi-turn exchanges.

Integration with enterprise environments is achieved through REST API services, allowing the system to be embedded within existing customer platforms, management tools or third-party applications. This API-based architecture facilitates request orchestration, automated ticket generation, structured incident reporting and logging of interactions within corporate systems.

The interface supports multilingual communication at both the input and output levels, enabling users to interact in different languages without requiring manual intervention. By encapsulating the complexity of multimodal processing and retrieval mechanisms, the system provides a unified and user-friendly conversational experience while ensuring compliance with domain-specific constraints and organizational policies.

3.5. Operational Dashboard

The operational dashboard serves as the supervision and management layer of the system, providing tools for agents and administrative staff to oversee system activity. While the conversational interface handles automated multimodal interactions with end users, this component focuses on monitoring, control and human intervention when necessary.

The dashboard provides real-time insights into system activity through configurable KPIs, including information about registered incidents, detected intents, resolution progress and escalation cases. These analytics enable supervisors to evaluate system behavior, detect ambiguous or high risk interactions and analyze workload distribution across incident categories. Detailed interaction-level metadata, such as intent predictions and retrieval outputs, can also be accessed to improve transparency and traceability.

Beyond monitoring capabilities, the dashboard supports human-in-the-loop interaction. Agents are able to inspect AI-generated responses, validate generated structured reports, adjust system decisions when required and manually escalate complex or sensitive cases. This supervision layer ensures that automated decision-making remains aligned with organizational policies and regulatory requirements.

The dashboard is integrated within the same microservices infrastructure as the rest of the system, allowing synchronized access to interaction histories, structured reports and knowledge base updates. By decoupling automated interaction from operational supervision, the architecture enables scalable deployment while preserving control, accountability and expert oversight in critical enterprise scenarios.

4. Conclusions and Further Work

The proposed mCS-LM system brings together the key components required for multimodal customer service and incident management, integrating domain-adapted LLM, VLM and ALM models with a RAG pipeline, a structured output and hallucination control module and dedicated interfaces for both end users and operational agents. By combining perception, retrieval, structured generation and human-in-the-loop supervision within a modular microservices architecture, the system provides a flexible and scalable solution tailored to enterprise environments.

At its current stage, the platform is being evaluated in real operational settings in collaboration with customer service professionals from participating companies. This validation process focuses on assessing multimodal understanding capabilities, the consistency of structured outputs and the effectiveness of the hierarchical intent-based control mechanisms in reducing hallucinations and out-of-scope responses. Within this context, the present work concentrates on the design and integration of the system architecture, while more extensive 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 explore several directions. First, we aim to extend the multimodal RAG component with more advanced cross-modal retrieval techniques, improving the alignment between textual queries

and information derived from visual or audio sources, as discussed in recent work on cross-modal RAG systems [22]. Second, we plan to investigate adaptive intent classification strategies 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 [23]. Third, we will explore feedback-driven optimization 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 [24].

Future validation efforts will also include a quantitative assessment of the impact of model adaptation, comparing the domain-adapted LLM, VLM and ALM components with their corresponding base versions. Particular attention will be given to the hallucination control module by evaluating the hierarchical intent classification pipeline, the correctness of the selected generation prompts, and the reduction of unsupported or out-of-scope responses when compared with direct LLM prompting. These analyses will be complemented with detailed error analysis across modalities and measurements of latency, scalability, and computational cost under realistic deployment conditions.

These research directions are expected to further improve the robustness, scalability and practical applicability of multimodal conversational systems in regulated enterprise contexts. In addition, future efforts will consider the extension of the system to low-resource settings, including regional languages and dialectal variations, taking advantage of its modular and language-agnostic design to support broader linguistic diversity and robustness in real-world multilingual scenarios.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammatical and spelling correction. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

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.

References

- [1] R. Balasubramanian, A. Libarikian, D. McElhaney, Insurance 2030—the impact of ai on the future of insurance, McKinsey & Company (2018) 1–10.
- [2] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, M. Zhou, Superagent: A customer service chatbot for e-commerce websites, in: Proceedings of ACL 2017, system demonstrations, 2017, pp. 97–102.
- [3] A. S. Luccioni, S. Viguier, A.-L. Ligozat, Estimating the carbon footprint of bloom, a 176b parameter language model, Journal of machine learning research 24 (2023) 1–15.
- [4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 1 (2023).
- [5] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, Vision-language pre-training: Basics, recent advances, and future trends, arXiv preprint arXiv:2210.09263 (2022).
- [6] Y. Su, J. Bai, Q. Xu, K. Xu, Y. Dou, Audio-language models for audio-centric tasks: A survey, arXiv preprint arXiv:2501.15177 (2025).
- [7] M. Eling, M. Lehmann, 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 (2018) 359–396.
- [8] E. Stoeckli, C. Dremel, F. Uebernickel, Exploring characteristics and transformational capabilities of insurtech innovations to understand insurance value creation in a digital world, Electronic markets 28 (2018) 287–305.

- [9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in neural information processing systems* 33 (2020) 9459–9474.
- [10] M. Shumanov, L. Johnson, Making conversations with chatbots more personalized, *Computers in human behavior* 117 (2021) 106627.
- [11] A. Xu, Z. Liu, Y. Guo, V. Sinha, R. Akkiraju, A new chatbot for customer service on social media, in: *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 3506–3510.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.
- [14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, *arXiv preprint arXiv:2303.08774* (2023).
- [16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International conference on machine learning*, PmlR, 2021, pp. 8748–8763.
- [17] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: *International conference on machine learning*, PMLR, 2023, pp. 19730–19742.
- [18] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, *ACM Transactions on Information Systems* 43 (2025) 1–55.
- [19] J. Gung, E. Moeng, W. Rose, A. Gupta, Y. Zhang, S. Mansour, Natcs: Eliciting natural customer support dialogues, in: *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 9652–9677.
- [20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., *ICLR* 1 (2022) 3.
- [21] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, *Advances in neural information processing systems* 36 (2023) 10088–10115.
- [22] M. M. Abootorabi, A. Zobeiri, M. Dehghani, M. Mohammadkhani, B. Mohammadi, O. Ghahroodi, M. S. Baghshah, E. Asgari, Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation, *Findings of the Association for Computational Linguistics: ACL 2025* (2025) 16776–16809.
- [23] L. Wang, X. Zhang, H. Su, J. Zhu, A comprehensive survey of continual learning: Theory, method and application, *IEEE transactions on pattern analysis and machine intelligence* 46 (2024) 5362–5383.
- [24] X. Sun, J. A. Bosch, J. De Wit, E. Krahmer, 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*, 2023, pp. 99–101.