Multimodal Conversational Agents: A Survey

Constantinos Romantzis
Department of Computer Science,
University of Nicosia
Nicosia, Cyprus
0009-0003-5156-1756

***Abstract*—The rapid rise of Conversational Agents (CAs) has led to their implementation in almost every area of life and industry. Currently CAs are relying primarily on AI models that process text, also known as Natural Language Processing (NLP), to perform various tasks such as extracting Sentiment from the user's input or recognising Intents and extracting Entities in order to understand the context and provide the appropriate information. All of these tasks are performed to make the interaction between the CA and the user more efficient and human-like. However, this focus on textual inputs not only further disconnects what is supposed to be a human-esque interaction from the true human experience of being able to receive various stimuli to infer information instead of solely relying on text but also leaves significant room for possible improvements by potentially leveraging other modalities to extract useful features. This study aims to review the effect of employing multiple modalities, at the backend level, in order to create more robust Augmented Conversational Agents. Through research we identify the two main approaches to doing that as well as the different instances and combinations of multimodality that have been shown to provide significant improvements.**

***Index Terms*—Multimodality, NLP, NLU, Multimodal Conversational Agents, Chatbots**

I. Introduction

Conversational Agents, or Chatbots as they're widely called, are becoming more and more prevalent in all industries around the globe as they can provide a certain level of intimacy and humanity to the process of information retrieval in the form of Questioning & Answering (Q&A). From their humble beginnings in 1966 with ELIZA [1], the simple back-and-forth bot, to the present day where they're leveraging Large Language Models (LLMs) such as GPT [2] to produce content that, while not always factually reliable, can oftentimes mimic human behaviour to the point of being almost indistinguishable chatbots are becoming increasingly popular in various domains such as customer service, healthcare, education, and entertainment. This is due to their ability to provide personalized, efficient and round-the-clock assistance to users. For instance, in customer service, conversational agents can help customers with their queries and provide them with product information, thus reducing the need for human support. In healthcare, they have been shown to be successful in assisting patients in managing their health conditions, such as diabetes, and reminding them of their medication schedules [3]. In education, they can provide students with quick access to relevant information and assist them in completing their assignments.

As agents rely on complex Natural Language Processing (NLP) models in their backend and since humans do not rely solely on written text but also employ their senses of Sight (images), Hearing (audio formats) and Touch (pressure sensitivity) to infer and reach outcomes it would stand to reason to assume that employing multiple modalities in order to augment the models that the agents rely on would result not only in a more human-esque interactive experience for the user but also in expanded capabilities for the agent that can produce faster and more accurate results.

Despite that, most of the available literature both in the area of Conversational Agents and in the more general sector of NLP modeling appears to be confined solely within the Language Modality and therefore potentially missing out on the rich features that can be extracted and utilized either in direct combination with the Language features or used in parallel when the situation demands it. In this survey we aim to examine the most prevalent and effective techniques of incorporating multimodality into Conversational Agents focusing on both the technical aspects of the implementation but also on the effect they have on the agent's accuracy and ease of use. Those techniques can be further sorted into two categories; the first one being the use of the Conversational Manager as a central connectivity point and the second being the direct augmentation of one model by synthesizing the features extracted from various modalities.

II. Multimodal Augmentation

A. Key Modalities and Multimodality in the context of Computer Science

Modalities are different modes and/or means of communication that can be used to convey information, including verbal, nonverbal, visual, auditory, and haptic cues. In computer science, modalities are used to enhance the interaction between humans and machines, making it more natural, intuitive, and efficient. By combining multiple modalities, multimodal systems can provide a richer and more flexible user experience, as users can choose the modality that best suits their preferences, abilities, and contexts [4].

Some brief examples of modalities and what they can include would be audio in the form of speech and ambient sound, images ranging from static pictures of objects to real-time tracking of gestures, facial expressions and eye movements, and haptic features like the pressure applied while typing [5]. The aforementioned modalities are especially important

when it comes to modality integration in CAs as Speech,which is one of the most common modalities, allows humans to express complex ideas and emotions through language and Ambient Sounds, such as music, voice feedback, or environmental noise, can add context, atmosphere, and emotional resonance to the interaction [6]. In situations where speech is not possible or desirable, Gestures can substitute speech and Facial expressions, such as smiling, frowning, or raising eyebrows, can convey a wide range of emotions and attitudes, and help humans interpret the meaning of verbal and nonverbal cues [7]. Likewise, Eye movements, such as gaze direction, pupil dilation, or blinking rate, can indicate attention, interest, or cognitive load, and provide feedback to the system about the user's mental state and attention to the conversation [8]. Finally, touch, such as tapping, swiping, or squeezing, can enable precise and fine-grained input, and enhance the realism and immersion of virtual environments.

Therefore, Multimodality refers to the integration of multiple modalities into a cohesive and coordinated system that can recognize, generate, and manipulate information from different sources. Multimodal systems can use machine learning algorithms, such as deep neural networks and Transformers, to process and analyze multimodal data, and generate text responses or even multimodal outputs, such as speech synthesis, animation [9], or tactile feedback. Multimodal systems can be used in various domains, such as education, healthcare, entertainment, or transportation, to improve accessibility, engagement, and performance. In addition to Multimodal systems, individual models, primarily LLMs as of recently, can leverage Multimodality directly to incorporate additional features provided by the additional Modalities and thus become Multimodal on their own.

B. Agents Leveraging Multiple Modality-specific Models

As mentioned earlier, Multimodal Conversational Agents can be distinguished into two categories depending on their structure. Here we will examine the structure of using the CA not only as the unit that communicates in the forefront with the user but also as a central connectivity point, a "hub" in other words, where multiple different models each confined to its own modality are brought together and in order for the agent to return an answer that combines their outputs. Logic in the form of Entity and Intent recognition is used to distinguish between the different possible inputs of the user and based on that they're passed along to the correct module to be handled appropriately before returning the necessary information. For instance, a multimodal conversational agent can employ speech recognition models to extract audio features and transcribe spoken language into text, which can be then passed to the NLP models to extract meaning and context from the text, and then tailor the answer appropriately to the visual cues, such as facial expressions and gestures, that the visual recognition models have interpreted [7, 10]. This approach has been documented in a variety of domains and has proven to be successful in adding an additional layer of personalisation and adaptability that allows the conversation to flow in a manner that resembles human-like interactions.

1) Multimodal CAs in Healthcare: Within the domain of Healthcare, and more specifically, in the areas where Prognosis, Treatment and Rehabilitation intersect the concept of CAs being a valuable tool in the Medical Personnel's arsenal is gaining significant traction especially due to their non-invasive nature. In the case of project REA [10], preliminary testing rounds have indicated that a medically inclined CA that incorporates visual and auditory signals to assess the progress and state of the patients rehabilitation while also operating as a round-the-clock, reliable, treatment protocol enforcer by reminding them to take their medication as instructed by their primary physician is worth exploring further. Likewise, similar approaches such as Zenon [7], the CA that uses a combination of NLP and Computer Vision models to perform sentiment analysis while the user is interacting with the agent, have shown during preliminary testing phases that having two different modalities that compliment each other allows for a more robust sentiment analysis as the multimodal facet of it allows each model to pick up on, and combine, subtleties that are unique to each medium. By its very nature, the non-intrusive aspect of this data collection strategy is vital when dealing with patients that may have suffered different degrees of cognitive impairment. Overall, these are very promising studies as they both offer a connection between physicians and their patients without the need for the physicians to actively split their time between repetitive observation thus allowing them to focus on more pressing matters without sacrificing the quality of the intervention. It is also worth noting that, while these CAs are impressive on their own, they are also central components in broader projects as they have been developed to work as part of larger technological and medical frameworks [5, 12] that aim to use the recent technological advancements in hardware capabilities and AI to provide a necessary overhaul to the way patient rehabilitation and treatment is handled in regards to early diagnosis and subsequent treatment.

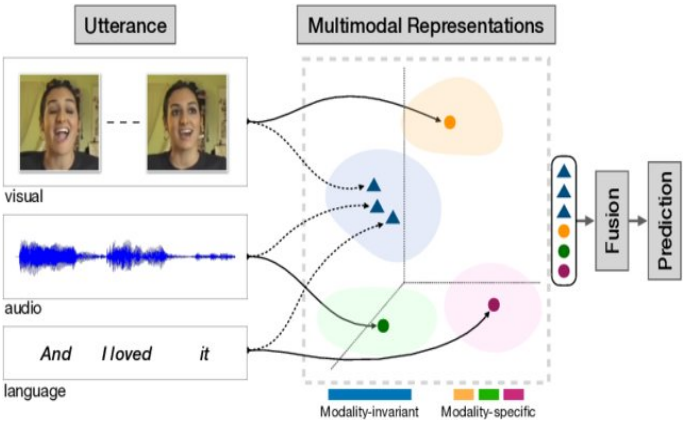

Fig. 1. Sentiment Analysis via Fusion of multiple Modalities [11]. As this is a Multimodal Model (see 2. C) that only handles Sentiment Analysis it can be connected to a CA as an additional model to provide an affective insight.

2) Multimodal CAs in Education: Efforts to Augment the capabilities of CAs are not exclusive to Healthcare and Medicine, and have also been made, in the same vein, in other domains. In recent years, the development of multimodal conversational agents has gained significant attention in the field of e-learning especially due to the COVID-19 Pandemic that forced the education system to shift towards complete e-learning methods during the two main lockdown periods and later on towards hybrid systems as we began to navigate the post-pandemic world. For the scope of this review, two studies pertaining to the area of education and e-learning were selected [9, 13]. The authors argue that traditional e-learning approaches, which are often multimedia-based and passive, do not fully engage learners and limit their potential for active and effective participation [9]. In order for this issue to be addressed, multimodal CAs that facilitate proactive, peer-to-peer interactions among human learners and artificial AI models are proposed [9, 13]. In one of the instances reviewed, it was shown that employing a Multimodal CA that primarily received —voice— as input and proceeded to output audio and animations resulted in a decrease in the required technological literacy needed to interact with it for the school children who were receiving the e-lessons and the teachers who were responsible for the management of the learning material [9]. It also allows the children to learn at their own pace as they each have their own instance of a CA that is trained to understand intents and adapts their level to match the level of the user while also using multiple modalities to simplify and illustrate concepts that would otherwise require a higher educational level when explained in text with technical definitions. The effectiveness of Multimodal CAs in Education was further supported by reception to Gera, the multimodal chatbot of the Geranium educative system [13], which was highly favorable by school teachers who, like the first case [9], praised the system's ease of use as well as its educational potential due to its adaptive and Multimodal engagement with the students. These approaches aim to create a more social and engaging learning experience, fostering motivation and enhancing the overall effectiveness of e-learning while also directly showcasing the value that CAs can provide in the domain of Education.

C. Multimodal Language Models

1) Multimodality with Transformer Networks and Attention: NLP models traditionally process text using deep neural networks such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to train these models and are trained on massive text datasets in order to perform tasks like sentiment analysis, intent recognition, and entity extraction. However, these models have generally struggled with tasks requiring the integration of many modalities, such as image captioning or multimodal question-answering. This is due to the fact that standard models are incapable of capturing the delicate and intricate interactions between different modalities and fail to integrate data from disparate sources. This is why when dealing with CAs using standard NLP models the techniques discussed above are employed which rely on separate models to work in parallel in order for them to handle the different modalities before passing the extracted features back to the agent in a form that can be further processed by the NLP units.

Even though there have been some relatively successful attempts at integrating Features extracted from Images and Audio to standard Language models [14 - 15], due to the relatively recent emergence of, the extremely impressive, Transformer models, which are Deep Neural Networks that process the input all at once by implementing the use of the Attention mechanism [16], CAs have received a significant overhaul in the way they function and such they have become even more prevalent in our daily lives. The use of the Attention mechanism in Transformer networks has also given rise to techniques that aim to enable these language models to locally process multiple modalities and integrate the appropriate features directly into the model without having to rely on the cooperation of other parallel models.

One example of such an application that leverages multimodality would be the, recently introduced, Vision-Language Transformer (VLT) [17]. The VLT is a transformer-based model that can integrate information from both images and text to perform image captioning by being trained on a large dataset of images and captions, and is designed to predict the caption given the input image. The model leverages a transformer-based architecture that can capture the complex relationships between the image and the corresponding caption. The VLT model is trained using a multi-task learning approach, which involves training the model on two tasks simultaneously enabling it to demonstrate state-of-the-art performance on several image captioning benchmarks, including the REFCOCO & REFCOCO+ image segmentation datasets. In practice the model works by producing a specific language feature vector for each of the positions of the image feature based on the interaction between the language information and the corresponding pixel information. The model has also proven to be equally as effective in processing video instead of simply static images.

2) LLMs and Multimodal Chain of Thought: The ability of the Transformer-based Language models to process inputs all at once with the Attention mechanism [16] has allowed researchers to dramatically increase the size of the training corpus. This has resulted in LLMs which are now dominating the industry with new model after model emerging in a short period of time. Contrary to standard NLP models, LLMs are as of now being created with the capability of leveraging multimodality and integrating information from different modalities in mind in order to perform complicated tasks by being trained on huge datasets containing both text and, in some cases, textual representations of other modalities like images or videos in the form of captions. This caption based approach in problem solving, which closely resembles the function of what would be a CA, has proven

to be highly effective but still far from perfect. It works by leveraging another technique that is unique to LLMs called Chain-of-Thought (CoT) Reasoning [18] which forces the model to break the initial prompt down into sequential tasks and process them step by step in order to arrive to a final conclusion that makes use of every piece of information received. This does not only closely resemble how humans perform problem-solving reasoning, meaning that it would increase the human-like element of computer-to-human and human-to-computer interactions but, it is during these CoT steps that image captions are shown to be particularly effective in assisting with problem solving requiring reasoning [19 - 20]. As we saw earlier, though, it is possible to extract features directly from the images themselves instead of relying on captions [17]. This development shifts the focus away from the language-centric approach and allows for the implementation of richer features generated from the actual images thus creating what can only be described as genuinely Multimodal Chain-of-Thought (MCoT) [21] Reasoning. This technique boasts problem solving accuracies that surpass the current, caption based, state-of-the-art by 16.51% (75.17%→91.68% from experimental results) and even surpasses human performance by 3.28% (88.40%→91.68% from experimental results) (MCoT) [21]. The authors of this technique also state that while the direct integration and feature extraction was performed on images due to their perceived importance in terms of Modality ranking, the same technique can be extended to apply to other modalities as well and will investigate the exact performance on audio in future works.

3) Augmented CAs leveraging Multimodal Models : Due to their impressive performance CAs relying solely on Multimodal Models are already available with the most popular example being OpenAI's ChatGPT [22 - 23] which is best described as a conversational model (CM) currently using Generative Pre-trained Transformers (GPT - a type of Transformer Neural Network) [2] to perform generative language tasks. More specifically it uses OpenAI's 4th iteration of GPT (GPT-4) [24 - 25] which leverages techniques similar to the ones studied above meaning it also uses CoT reasoning [18, 21] and Multimodal (Image) features [17, 21] making it the most realistic and human-like LLM currently available. These "Chatbots" can perform a variety of tasks from general chatting to summarizing, paraphrasing and even generating large text paragraphs, providing advice and facts based on their training knowledge (not always reliable due to their "Generative" nature which might produce false results for which the model appears certain - known as "Hallucination"), coding and as of the 4th Iteration [24 - 25] image generation based on other Generative Models proving that once again Multimodality is in the forefront of progress. GPT4 is also the brain behind Microsoft's attempt to introduce Conversational AI into the way internet search results are queried with the introduction of the new Bing AI which acts as a CA that combines GPT4's abilities with the ability to search the internet for additional information.

GPT-based models aside, promising work in CAs using single Multimodal Models exists with emphasis being placed on the CA's ability to receive and provide inputs and outputs that handle more than one modality as exemplified in the case of a CA that successfully leverages the Attention mechanism to make use of a Multimodal Knowledge Base that allows it to receive and serve context-relevant combinations of Images and Text in real time [26].

D. Implementation of Reinforcement Learning from Human Feedback (RLHF)

Even though Reinforcement Learning (RL) is not a Modality in and of itself it can not be omitted from a study dealing with the implementation of Multimodality in order to improve CAs in a manner that makes it more Human-esque. Reinforcement Learning works by introducing an agent that performs tasks based on a numerical Reward & Punishment system where the correct performance of a task results in an increase of points whereas an incorrect performance will result in a decrease [27] essentially mimicking the dopamine-based reward system that the human brain uses to make decision. This is relevant to CAs and especially ones relying on Multimodal Language Models because Reinforcement Learning from Human Feedback (RLHF) is a technique that is used during the finetuning step of these models [24 - 26] in order to teach the agent how to differentiate between all the possible answers it can generate by giving it a framework of which answers are deemed desirable by human users. As the potential features increase exponentially with each modality introduced, having a method of filtering all the undesirable answers that the model can produce is what adds to the agent's ability to remain coherent and safe to use.

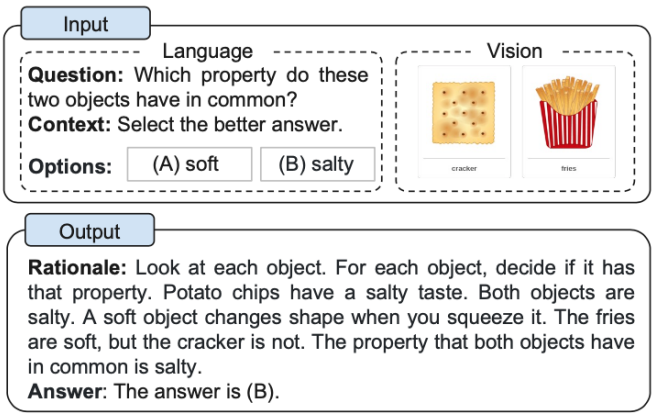

Fig. 2. Multimodal (CoT) LLM leveraging Image Features to assist reasoning [21].

E. Verdict Regarding the feasibility of Multimodally Augmented Conversational Agents

Modality combination has been shown to be achievable in two (2) ways; 1) Separate models each trained to deal with a separate modality perform the feature extractions, infer from those features and then pass their results to the backend of the Conversational Manager which is used as central connectivity point for these models in order for the NLP unit to answer accordingly and 2) Language Models are trained to process Multiple Modalities directly without the need to rely on an ensemble of different models. Due to the demanding nature of such a task the most successful Multimodal Models are primarily Transformer/Attention-based Large Language Models incorporating more advanced techniques such as CoT Reasoning and RLHF in order to properly extract and make the correct use of the additional features.

TABLE I
RESEARCH SYNTHESIS MATRIX

Augmentation Approach	**Modalities**	
	Image	***Audio***
Individual Models	C. Chira et al. Nakano et al. J. Schuir et al. Mavropoulos et al. D. Griol et al.	Chunjong Par J. Schuir et al. Mavropoulos et al. Bubeck et al. *
Multimodal Model	Hazarika et al. A. Anastasopoulos et al. Shao-Yen Tsen et al. Ding et al. A. Zeng et al. ** Z. Zhang et al. Bubeck et al. * L. Liao et al.	Hazarika et al.

*Uses a Multimodal Language Model to process images and utilizes a separate OpenAI model (WhisperAI) to process audio.
**The images are converted to captions before being used.

III. CONCLUSION

Human-like, human-to-computer and computer-to-human interactions are important, not only because of the element of familiarity that they offer, but also because humans use visual and auditory cues to assess their surroundings and to gain insights that would otherwise be lost when the only source of information is restricted to the modalities that require proper and unrestricted articulation and as such CAs that do not support and integrate multiple modalities are forced to rely solely on the user's ability and willingness to express their sentiments accurately and honestly. By combining these different modalities, multimodal conversational agents are able to leverage the variety of information provided by each modality and as such they can perform at a higher, more desirable, level while also creating a more immersive and engaging experience for the user.

REFERENCES

[1] Weizenbaum, Joseph. 'ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine', (1966).
[2] Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 'Improving Language Understanding by Generative Pre-Training', n.d.
[3] Gong et al. 'My Diabetes Coach, a Mobile App–Based Interactive Conversational Agent to Support Type 2 Diabetes Self-Management: Randomized Effectiveness-Implementation Trial'. Journal of Medical Internet Research 22, no. 11 (5 November 2020): e20322. https://doi.org/10.2196/20322.
[4] Furht, Borko, ed. 'Multimodal Interfaces'. In Encyclopedia of Multimedia, 651–52. Boston, MA: Springer US, 2008. https://doi.org/10.1007/978-0-387-78414-4_159.
[5] Mavropoulos et al. 'Smart Integration of Sensors, Computer Vision and Knowledge Representation for Intelligent Monitoring and Verbal Human-Computer Interaction'. Journal of Intelligent Information Systems 57, no. 2 (October 2021): 321–45. https://doi.org/10.1007/s10844-021-00648-7.
[6] Chunjong Park, Chulhong Min, S. Bhattacharya, and F. Kawsar. 'Augmenting Conversational Agents with Ambient Acoustic Contexts', n.d.
[7] C. Chira et al. 'An Affective Multi-Modal Conversational Agent for Non Intrusive Data Collection from Patients with Brain Diseases', n.d.
[8] Nakano, Yukiko I., and Ryo Ishii. 'Estimating User's Engagement from Eye-Gaze Behaviors in Human-Agent Conversations'. In Proceedings of the 15th International Conference on Intelligent User Interfaces, 139–48. Hong Kong China: ACM, 2010. https://doi.org/10.1145/1719970.1719990.
[9] Julian Schuir, Eduard Anton, Marian Eleks, and Frank Teuteberg. 'Tell Me and I Forget, Involve Me and I Learn: Design and Evaluation of a Multimodal Conversational Agent for Supporting Distance Learning', n.d.
[10] Mavropoulos et al. 'A Context-Aware Conversational Agent in the Rehabilitation Domain'. Future Internet 11, no. 11 (1 November 2019): 231. https://doi.org/10.3390/fi11110231.
[11] Hazarika, Devamanyu, Roger Zimmermann, and Soujanya Poria. 'MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis'. arXiv, 19 October 2020. http://arxiv.org/abs/2005.03545.
[12] Maga-Nteve et al. 'A Semantic Technologies Toolkit for Bridging Early Diagnosis and Treatment in Brain Diseases: Report from the Ongoing EU-Funded Research Project ALAMEDA'. In Metadata and Semantic Research, edited by Emmanouel Garoufallou, María-Antonia Ovalle-Perandones, and Andreas Vlachidis, 1537:349–54. Communications in Computer and Information Science. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-98876-0_30.
[13] D. Griol, J. M. Molina, and Araceli Sanchis de Miguel. 'Developing Multimodal Conversational Agents for an Enhanced E-Learning Experience', n.d.
[14] Antonios Anastasopoulos, Shankar Kumar, and H. Liao. 'Neural Language Modeling with Visual Features', n.d.
[15] Shao-Yen Tseng, Shrikanth S. Narayanan, and P. Georgiou. 'Multimodal_Embeddings_From_Language_Models_for_Emotion_Recognition_in_the_Wild' n.d.
[16] Vaswani, Ashish et al. 'Attention Is All You Need'. arXiv, 5 December 2017. http://arxiv.org/abs/1706.03762.
[17] Ding, Henghui, Chang Liu, Suchen Wang, and Xudong Jiang. 'VLT: Vision-Language Transformer and Query Generation for Referring Segmentation'. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, no. 6 (1 June 2023): 7900–7916. https://doi.org/10.1109/TPAMI.2022.3217852.
[18] Wei et al. 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models'. arXiv, 10 January 2023. http://arxiv.org/abs/2201.11903.
[19] Lu, Pan, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 'IconQA: A New Benchmark for Abstract Diagram Understanding and Visual Language Reasoning'. arXiv, 25 July 2022. http://arxiv.org/abs/2110.13214.
[20] Andy Zeng, Adrian S. Wong, Stefan Welker, K. Choromanski, F. Tombari, Aveek Purohit, M. Ryoo, et al. 'SOCRATIC MODELS: COMPOSING ZERO-SHOT MULTIMODAL REASONING WITH LANGUAGE', n.d.
[21] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, G. Karypis, and Alexander J. Smola. 'Multimodal Chain-of-Thought Reasoning in Language Models', n.d.

[22] Deng, Jianyang, and Yijia Lin. 'The Benefits and Challenges of ChatGPT: An Overview'. Frontiers in Computing and Intelligent Systems 2, no. 2 (5 January 2023): 81–83. https://doi.org/10.54097/fcis.v2i2.4465.
[23] Mattas, Puranjay Savar. 'ChatGPT: A Study of AI Language Processing and Its Implications'. International Journal of Research Publication and Reviews 04, no. 02 (2023): 435–40. https://doi.org/10.55248/gengpi.2023.4218.
[24] Bubeck et al. 'Sparks of Artificial General Intelligence: Early Experiments with GPT-4'. arXiv, 13 April 2023. http://arxiv.org/abs/2303.12712.
[25] OpenAI. 'GPT-4 Technical Report'. arXiv, 27 March 2023. http://arxiv.org/abs/2303.08774.
[26] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. 'Knowledge-Aware Multimodal Dialogue Systems', n.d.
[27] Kaelbling, L. P., M. L. Littman, and A. W. Moore. 'Reinforcement Learning: A Survey'. Journal of Artificial Intelligence Research 4 (1 May 1996): 237–85. https://doi.org/10.1613/jair.301.