Literature Review

The research by Ghosh et al. (2024) aims to solve the issue of medical question summarization for patient cases with priority given to Hindi-English code-mixed clinical inquiries. The research introduces MMCQS as an initiative to develop the Multimodal Medical Code-mixed Question Summarization Dataset (MMCQSD) which unifies visual elements with Hindi-English code-mixed medical queries[1]. Through integration patients gain an expanded medical representation that shows their condition from multiple dimensions. The authors have developed MedSumm as a framework that takes advantage of Vision Language Models (VLMs) and Large Language Models (LLMs) for this work. The healthcare decision-making benefits from their approach while understanding patient queries reaches deeper levels which creates new possibilities for personalized medical treatments[1].

Sharma et al. (2024) created the THAR dataset which contains 11,549 comments from YouTube written in Hinglish code-mixed language dedicated to hate speech identifica of Islam, Hinduism and Christianity [2]. The THAR dataset uses its resources to solve challenges in detecting hate speech within minimal language resources. The researchers applied MuRIL which represents a transformer-based and deep learning model to achieve binary classification scores of 0.78 macro and weighted F1 and multi-class classification scores of 0.65 and 0.72. The research requires additional hate speech detection models and datasets targeting low-resource languages to move forward in the field of hate-speech detection[2].

The authors of Ghosh et al (2024) developed CLIPSyntel as a system that uses Contrastive Language Image Pretraining and Large Language Models to advance multimodal medical query summarization. To better understand patient queries they combined MMQS dataset elements through the coupling of medical image elements with their corresponding textual questions[3]. The system framework contains four units including medical disorder detection followed by relevant context generation followed by context filteration and visually-tinged summary development. The healthcare decision-making process gets improved and patient needs gain personalized understanding when CLIPSyntel adds visual information from images to create medically informed summaries[3].

The research team of Schlegel et al. (2023) developed PULSAR as the model which transforms physician-patient dialogues into medical reports for ImageClef 2023 MediQA-Sum. This study relied on MTS-Dialog and ACI-BENCH datasets as it focused on turning patient interactions into standardized medical records. Pre-training the model on medical corpora specifically contributed to better understanding and processing of specialized medical terminology and context. The training set received

added value through synthesized dialogues which large LLMs produced to simulate genuine patient interactions[4]. The model received subsequent data-fitted training from the MTS-Dialog and ACI-BENCH natural data sources to make it specialized for summary generation. The research discovered multiple shortcomings such as inadequate evaluation of pre-training methods for particular medical domains alongside insufficient methods for creating synthetic medical data including infrequent medical situations. The research study highlighted the need for robust out-of-domain evaluation for determining generalization potential while suggesting upcoming work targeting the fusion of text together with imaging data as well as structured information to enhance the summary production process[4].

Large language models served as the foundation for Tie et al. (2024) to perform their research on automatic PET report impression generation. An analysis of 37,370 PET reports from 2010 to 2022 at their institution enabled authors to fine-tune twelve freely available LLMs for generating report impressions from combined patient information and results data. The identification token used for each reading physician served to instruct the models about specific stylistic choices during report generation[5]. The studies demonstrated that physician-preferred scores came from BARTScore and PEGASUS Score in their domain-specific versions so researchers selected PEGASUS for expert evaluation. A study found that physicians evaluated PEGASUS-generated reports written in their individual speech patterns to be professionally appropriate in 89% of cases with a mean rating of 4.08 out of 5. This research shows how properly trained LLMs enable quick PET report generation through customized text production thus addressing medical imaging needs for efficient summarization capability[5].

The paper by Dalal et al. (2024) introduces MedicoVerse as a new summary method for pharma industry regulatory reports that extend beyond typical lengths. A two-part method uses hierarchical clustering together with natural language processing models which operate at the state-of-the-art level for effective processing of extensive text documents. The primary operational component of the technique makes use of SapBERT model to generate word embeddings before hierarchical agglomerative clustering takes place. The clustering groups data into organized structures while improving summary effectiveness[6]. The bart-larg-cnn-samsun model generates each individual summary from the clusters before combining them to generate the complete document summary. Through this research project teams gain better awareness of their regulatory environment by simplifying the processing of lengthy regulation documents which enables them to make informed decisions[6].

Scientists at Singh et al. (2024) demonstrate the difficulty of detecting misogyny in multimedia online materials that use Hindi-English language mixing in memes. Their work introduces MIMIC which contains hand-annotated memes found within 5,054 mixed Hindi-English language samples. Two classification tasks have been established for the MIMIC dataset which includes binary classification for misogynistic content detection and multi-label classification for misogyny classification types including objectification prejudice and humiliation[7]. The authors highlight the important lack of resources available to analyze multimodal material in code-mixed languages which

belong to low-resource communities. The authors show that multimodal methods outperform text-based and image-based approaches in detecting misogynistic content as they test three different computational models[7]

For sequence-to-sequence models, Bidirectional and Auto-Regressive Transformers (BART) became a popular denoising autoencoder which addresses natural language processing tasks including generation and translation and understanding according to Lewis et al. in 2019[8]. BART applies random text corruptions to input through token masking and sentence shuffling alongside deletion before it trains models to recover original text. The method employs a Transformer-based architecture with a bidirectional encoder and left-to-right decoder to perform general operations of BERT and GPT models. In addition to attaining state-of-the-art results on abstractive conversation, question answering and summarization, and machine translation, BART demonstrates strong skills across a variety of tasks[8].

Khanna et al. (2023) conducted the AI-MIRACLE study to determine if ChatGPT 4.0 (a big language model) could effectively perform multilingual radiology report translation while making them easier to understand among patients. The analysis investigated the five language groups that Americans most commonly use while studying reader comprehension levels for the simplified and translated reports[9]. Effectiveness testing of ChatGPT 4.0 translation services included simplification of chosen radiology reports into Tagalog, Vietnamese, Spanish, Arabic and Mandarian languages while an initial verification stage used Hindi as the test language. The study evaluated doctor performance by conducting surveys with medical professionals who assessed both the accuracy and the speech flow of translation outputs. The research obtained data from 24 participants who evaluated mixed results showing that the model achieved success throughout multiple languages. A new research demonstrates how ChatGPT 4.0 breaks language barriers to enhance patient comprehension of medical jargon in healthcare practice. These results demonstrate the need for AI model performance improvement through broader medical situation and language trainingreports[9].

Keshav et al. (2023) establish a multimodal approach to classify code-mixed speech primarily directed at Hindi-English language combinations. The research compares two methods: a 3-shot Few Shot Learning (FSL) technique and a Fully Connected Neural Network (FCNN) approach using a self-built dataset of movie and political evaluations. The FCNN achieved a 61.53% classification accuracy yet the FSL exceeded it significantly with 99.83% accuracy according to [10]. Given the wide divergence in performance levels between these algorithms this proves the superiority of few-shot learning techniques used in code-mixed speech sentiment assessment. Research contributes to filling a gap by directly implementing pre-trained speech models for sentiment analysis which supports the suitability of multimodal methods in this sector[10].

A detailed evaluation by Niu et al. (2024) can be found in their paper[11]. Authors detail the transformation of Large Language Models (LLMs) into Multimodal Large Language Models (MLLMs) in addition to their expanding application areas in medical practice. The authors evaluate MLLM applications across clinical decision support

systems along with medical imaging techniques as well as patient empowerment solutions and research use cases[11]. The review pays special attention to Multi-modal Large Language Models because they excel at combining various forms of data types such as text, images and audio to generate comprehensive patient health insights. The research demands more attention to data constraints technology barriers and ethical issues because the authors identify them as limiting factors for MLLM implementation in healthcare[11].

In order to identify abusive messages and troll from code-mixed social media material that includes both Hindi and English content, Shekhar et al. (2021) created a self-learning HLSTM model. The research adopts a transliteration framework to manage code-mixed data through script conversion from Hindi to Roman script for maintaining consistency[12]. The analysis points out a failure in existing hate-speech detection models for Hinglish languages since these systems cannot process language mixing effectively. This HLSTM model addresses the problem through word and character embedding approaches that help capture contextual information achieving a hateful detection rate of 97.49%. This evaluation demonstrates how specialized models should handle the complex behavioral patterns within code-mixed text processing[12].

In their approach to detect hate speech within Hindi-English code-switched text Sharma et al. (2021) present the MoH (Map Only Hindi) pipeline. The system first identifies languages before converting Romanized Hindi text into Devanagari script through transliteration and then applies fine-tuning on BERT and MuRIL multilingual language models[13]. Prior models struggle with code-mixed language processing for hate speech because they lack ability to process language mixing behavior which leads to an existence gap in this domain. According to the MoH methodology the authors achieved a 6% better performance compared to baseline and outperformed previous transliteration libraries by 15%. The value of domain-specific preprocessing combined with language model adjustments enables better processing of code-switched text according to[13].

A collection of fundamental language models operating with parameters ranging from 7 billion to 65 billion was released by the research group Touvron et al. (2023) as LLaMA (Large Language Model Meta AI[14]). These models have received training through open datasets containing content from CommonCrawl web pages as well as GitHub repos alongside Wikipedia across multiple languages and Project Gutenberg books and scientific articles from both ArXiv and Stack Exchange Q and A. Research showed that there exists a difference between what open-source language models offer researchers and the performance capabilities obtainable through proprietary large-scale tools[14].

The research group of Tunstall et al. developed Zephyr in 2023 as an alignment technique between smaller language models and user intent through the use of distilled direct preference optimization (dDPO). Zephyr-7B became a 7-billion-parameter model through AI-synthesized feedback refinement that provides results similar to large models during chat benchmark performance tests[15]. Zephyr-7B delivers superior performance than its larger counterpart Llama2-Chat-70B which has a parameter size of 70 billion for intent alignment tests. The training process completes in several hours through operations on 16 A100 GPUs. The results prove that well-trained small

models achieve excellent performance levels for natural language understanding and generation operations[15].

Wang et al. (2023) developed a multi-modal retrieval system that generates effective summaries from chest X-ray reports to enhance factual accuracy in produced summaries[16]. The researchers apply the MIMIC-CXR and MIMIC-III datasets to develop their approach through the combination of image similarity generation for additional text features with few-shot learning enabled by chain-of-thought prompting and ensemble techniques that boost performance. This research discovers an important knowledge gap that exists within the improvement of multimodal data fusion methods to achieve more accurate radiology report summaries. The technique attained a second-place position among 11 participating teams in the RadSum23 shared task while being the most advanced method in F1RadGraph scoring[16].

Kumar and Sachdeva (2022) developed a deep neural detection system that analyses cyberbullying from textual social content as well as visual contents and info-graphics. The CapsNet-ConvNet architecture merges CapsNet capsules for text processing with ConvNet capsules for visual processing. The extraction of text from visual information for info-graphics happens through the utilization of the Google Lens system [17]. Each modality prediction goes through a perceptron-based decision-level late fusion to determine whether the content shows bullying behaviour or not. A mixed-modal dataset consisting of 10,000 posts and comments obtained from YouTube, Instagram, and Twitter sites was used to assess the model which demonstrated a remarkable AUC–ROC of 0.98. The study addresses the research gap by highlighting efficient multimodal detection methods for cyberbullying which need processing diverse data to enhance accuracy [17].

In 2021 Yadav et al. presented a reinforcement learning method for abstractive query summarization through two novel question-aware semantic reward mechanisms for detecting question-type and question-focus detection[18]. The model receives these incentives which guide its production of semantic summaries that retain essential medical entities and question intent. The method achieved superior performance in benchmark tests which established its ability to create diverse accurate question summaries beyond existing models according to research[18].

Zhao and his coauthors developed SkinGPT-4 which serves as an realtime diagnostic tool for Dermatology through a sophisticated visual language model. The system used a fine-tuned version of MiniGPT-4 to learn complete skin disease image data totaling 52,929 pictures from both public and private sources. The system enables users to add skin images through its multimodal structure so the system assesses different skin condition features directly [19]. SkinGPT-4 delivers precise medical assessments together with interactive therapy choices after processing visual data through natural language capabilities which addresses the shortage of dermatologists and visual skin condition interpretation difficulties. Local deployment functionality of the system ensures user privacy while providing trustworthy diagnosis options for skin sicknesses to users[19].

Zhang et al. (2019) developed BERTScore which uses BERT contextual embeddings to create a machine algorithm for evaluating text generation tasks by measuring candidate and reference sentence parallels. BERScore represents an upgrade from

BLEU and ROUGE because instead of counting exact token overlaps it uses cosine similarities between contextualized token embeddings to determine semantic value matches [20]. The method provides more exact measures of semantic text quality since it goes beyond basic token match analysis. The authors demonstrated BERTScore by comparing it against 363 machine translation and image captioning models using human evaluation methods which produced better model selection performance compared to existing metrics. The assessment strength of BERTScore stands out through its approach to difficult examples such as adversarial paraphrases[20].

Touvron et al. (2023) introduced LLaMA 2 which serves as an improved version of initial LLaMA models through releasing base and fine-tuned chat models at three different size variants of 13 billion, 7 billion, and 70 billion. These models underwent training using 40% additional publicly available data compared to their previous versions in order to enhance their operational capabilities[21]. Openness is a critical characteristic of LLaMA 2 because it enables various business applications but prevents military or espionage-related deployment. The LLaMA 2 launch has successfully unified closed and open AI systems thus making greater AI innovations and research access possible[21].

Tiwari et al. (2022) proposed "Dr. Can See" and it represents a novel multimodal disease diagnosis virtual assistant which duplicates human doctor diagnostic procedures by processing symptoms through visual and text-based media. Through its Context-aware Symptom Image Identification module the system adopts discourse context techniques and visual data for effective symptom recognition[22]. Reinforcement learning methods enable the assistant to perform interactive user navigation through symptom reporting sequences and disease diagnosis procedures. The research team achieves a major accomplishment by developing an English multi-modal conversational medical dialogue dataset with marked symptom specifications and visual data points to resolve the database shortfalls in this domain. The authors made their codebase publicly available to let researchers and developers advance both research and development of this work[22].

The researchers at Mrini et al. (2021) developed a combination of question summarization techniques with Recognizing Question Entailment (RQE) to enhance understanding of consumer health inquiries. Their approach which involved joint training with augmented data showed important improvements on four biomedical datasets by reaching an 8% elevation in accuracy together with a 2.5% augmentation in ROUGE-1 scores[23]. The research team established during human evaluations that jointly trained models produced both trusted and helpful summaries. The authors have released their code together with two question summarization datasets that originated from a large medical conversation corpus and apply to this investigation[23].

Kumar et al. (2023) proposed a new abstractive summarization technique with special focus on comment sensitivity. They use the content of the article along with the comments provided by the users and produce summaries that comprise the opinion not only of the content but also the readers[24]. It resolves the problem of capturing varied opinions and sentiment presented by the users in the comments and produces

summaries that are richer and more representative. They tested the model on appropriate data sets and showed its ability to produce summaries that are comparable to the users' points of view[24].

Gupta et al. (2022) presented two datasets, MedVidCL and MedVidQA, to enable medical video understanding and question-answering research. MedVidCL contains 6,617 'medical non-instructional', 'medical instructional', , and 'non-medical' category videos to enable the creation of systems to differentiate them[25]. MedVidQA consists of 3,010 disease-related questions with timestamped visual answers extracted from real-world medical videos to enable the Medical Visual Answer Localization (MVAL) task. These datasets try to enable cross-modal understanding between medical terms and videos, which will enable the creation of apps that are useful to the general population and healthcare professionals alike[25].

Das and Gambäck (2014) presented an early work on sentence-level language identification of code-mixed Indian social media through Hindi-English code-switching. Their research created a system dependent on a series of features including lexical, context, and orthographic cues for identifying the language of each word in a sentence[26]. This piece of work addresses the issues created by the informal unstructured and nature of social media messages, where the users often switch languages within a sentence or utterance. The results of the research have been the foundation for subsequent work on natural language processing tasks for code-mixed data, particularly in multilingual societies where this type of mixing of languages is prevalent[26].

By constraining the amount of trainable parameters, Hu and colleagues in 2021 presented Low-Rank Adaptation (LoRA), an adaptation strategy that effectively fine-tunes large language models[27]. By adding trainable rank decomposition matrices to each layer of the Transformer structure and freezing pre-trained model weights, LoRA considerably lowers the number of trainable parameters for downstream tasks. Due to its exponentially lower computational and storage costs, it is now widely used in the machine learning field, especially for jobs demanding the fine-tuning of diffusion models[27].

Chin-Yew Lin proposed ROUGE (Recall-Oriented Understudy for Gisting Evaluation) in 2004, a collection of metrics to measure automatic summarization and machine translation systems in similarity with machine-generated and human-generated reference summaries or translations[28]. ROUGE-L measures longest common subsequence, ROUGE-N measures n-gram overlap, and ROUGE-S measures skip-bigram co-occurrence statistics. ROUGE scores increase as the similarity with reference summaries increases, and ROUGE is a common method in natural language processing to estimate text quality generated[28].

Liu et al. came up with Medical-VLBERT in 2021, a model that utilizes visual and textual information to automatically generate medical reports from COVID-19 Computed Tomography scans. The model uses a distinct learning approach where it pre-trains on medical texts and applies transfer learning to produce professional medical sentences from visual observations[29]. As part of an attempt to address the data scarcity challenge that comes with COVID-19 data, researchers trained the model on the vast Chinese CX-CHR dataset as well as subsequently fine-tuning it on a

dedicated COVID-19 CT dataset. This resulted in the model scoring both report generation performance and state-of-the-art terminology prediction[29].

In 2023, Zhu et al. proposed MiniGPT-4, a model to improve vision-language understanding by combining powerful large language models (LLMs) with vision. MiniGPT-4 is a combination of a pretrained vision encoder and a pretrained LLM through a shared projection layer, which aligns vision features and language representations well[30]. The simple yet effective structure allows MiniGPT-4 to carry out a variety of vision-language tasks, including creating rich image interpretation and visual question answering, and its performance competes with larger models like GPT-4. The structural design of the model allows efficient training using minimum computational resources and supports wider areas of application for multimodal AI research[30].

Zhang et al. in 2023 conducted an extensive survey of vision-language model (VLM) use to vision tasks, referring to the use of natural language supervision for visual learning. The study takes into account various VLM architectures, including CLIP, which aligns text and visual representations via contrastive learning to enable models to carry out tasks like zero-shot image classification and cross-modal retrieval[31]. The study also takes into account the role of large-scale pretraining towards different datasets, enabling the generalization ability of VLMs towards multiple vision tasks to be facilitated. The researchers also touch on challenges and trends in the future, hinting at the need for more effective training processes and more sophisticated multimodal data processing[31].

A Large Language Model with 7 Billion parameters in it called Mistral 7B was proposed by Jiang et al. in 2023 with the goal of enhancing efficiency and performance. Mistral 7B outperforms Llama 1 34B on reasoning, math, and coding activities, and outperforms Llama 2 13B on a variety of benchmarks[32]. The model handles sequences of unlimited length with low inference costs by using sliding window attention (SWA) and grouped-query attention (GQA) to speed up inference. Furthermore, Mistral 7B. Instruct, an instruction-tuned version, has been created and is superior than Llama 2 13B in terms of its ability to speak in both automatic and human tests. The Apache 2.0[32] license applies to both models[32].

References

[1] Ghosh, A., *et al.*: Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In: European Conference on Information Retrieval. Springer, ??? (2024)

[2] Sharma, D., Singh, A., Singh, V.K.: Thar-targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. ACM Transactions on Asian and Low-Resource Language Information Processing (2024)

[3] Ghosh, A., *et al.*: Clipsyntel: Clip and llm synergy for multimodal question summarization in healthcare. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38 (2024)

[4] Schlegel, V., et al.: Pulsar at mediqa-sum 2023: Large language models augmented by synthetic dialogue convert patient dialogues to medical records (2023)

[5] Tie, X., *et al.*: Personalized impression generation for pet reports using large language models. Journal of Imaging Informatics in Medicine **37**(2), 471–488 (2024)

[6] Dalal, A., Ranjan, S., Bopaiah, Y., *et al.*: Text summarization for pharmaceutical sciences using hierarchical clustering with a weighted evaluation methodology. Scientific Reports **14**, 20149 (2024)

[7] Singh, A., Sharma, D., Singh, V.K.: Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. ACM Transactions on Asian and Low-Resource Language Information Processing (2024). Just Accepted (April 2024)

[8] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)

[9] Khanna, P., *et al.*: Artificial intelligence in multilingual interpretation and radiology assessment for clinical language evaluation (ai-miracle). Journal of Personalized Medicine **14**(9), 923 (2024)

[10] Keshav, S., Lal, G.J., Premjith, B.: Multimodal approach for code-mixed speech sentiment classification. In: Advances in Signal Processing, Embedded Systems and IoT: Proceedings of Seventh ICMEET-2022, pp. 553–563. Springer, ??? (2023)

[11] Niu, Q., et al.: From text to multimodality: Exploring the evolution and impact of large language models in medical practice (2024)

[12] Shekhar, S., *et al.*: Hatred and trolling detection transliteration framework using hierarchical lstm in code-mixed social media text. Complex Intelligent Systems **9**(3), 2813–2826 (2023)

[13] Sharma, A., Kabra, A., Jain, M.: Ceasing hate withMoH: Hate Speech Detection in Hindi-English Code-Switched Language (2021)

[14] Touvron, H., Lavril, T., Izacard, G., et al.: Llama: Open and efficient foundation language models (2023)

[15] Tunstall, L., Beeching, E., Lambert, N., et al.: Zephyr: Direct Distillation of LM Alignment. arXiv preprint (2023)

[16] Wang, T., Zhao, X., Rios, A.: Utsa-nlp at radsum23: Multi-modal retrieval-based chest x-ray report summarization. In: The 22nd Workshop on Biomedical Natural

Language Processing and BioNLP Shared Tasks (2023)

[17] Kumar, A., Sachdeva, N.: Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. Multimedia Systems **28**(6), 2043–2052 (2022)

[18] Yadav, S., Gupta, D., Abacha, A.B., Demner-Fushman, D.: Reinforcement Learning for Abstractive Question Summarization with Question-Aware Semantic Rewards (2021)

[19] Zhou, J., He, X., Sun, L., et al.: SkinGPT-4: An Interactive Dermatology Diagnostic System with Visual Large Language Model (2023)

[20] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT (2019)

[21] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., et al.: LLaMA 2: Open Foundation and Fine-Tuned Chat Models (2023)

[22] Tiwari, M., Manthena, M., Saha, S., Bhattacharyya, P., Dhar, M., Tiwari, S.: Dr. can see: Towards a multi-modal disease diagnosis virtual assistant. In: Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM), pp. 1935–1944 (2022)

[23] Mrini, K., Dernoncourt, F., Chang, W., Farcas, E., Nakashole, N.: Joint summarization-entailment optimization for consumer health question understanding. In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pp. 58–65 (2021)

[24] Kumar, R., Chakraborty, R., Tiwari, A., Saha, S., Saini, N.: Diving into a sea of opinions: Multi-modal abstractive summarization with comment sensitivity. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM), pp. 1117–1126 (2023)

[25] Gupta, D., Attal, K., Demner-Fushman, D.: A Dataset for Medical Instructional Video Classification and Question Answering. arXiv preprint (2022)

[26] Das, A., Gambäck, B.: Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text (2014)

[27] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank Adaptation of Large Language Models (2021)

[28] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[29] Liu, G., Liao, Y., Wang, F., Zhang, B., Zhang, L., Liang, X., Wan, X., *et al.*:

Medical-vl-bert: Medical visual language bert for covid-19 ct report generation with alternate learning. IEEE Transactions on Neural Networks and Learning Systems **32**(9), 3786–3797 (2021)

[30] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models (2023)

[31] Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-Language Models for Vision Tasks: A Survey (2023)

[32] Jiang, A.Q., Sablayrolles, A., Mensch, A., et al.: Mistral 7B (2023)