# Towards Interpretable Radiology Report Generation via Concept Bottlenecks using a Multi-Agentic RAG

Hasan Md Tusfiqur Alam[1][⋆][0000−0003−1479−7690], Devansh Srivastav[1][0000−0002−8858−7402], Md Abdul Kadir[1,2][0000−0002−8420−2536], and Daniel Sonntag[1,2][0000−0002−8857−8709]

[1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[2] University of Oldenburg, Oldenburg, Germany
{hasan.alam,devansh.srivastav,abdul.kadir,daniel.sonntag}@dfki.de

**Abstract.** Deep learning has advanced medical image classification, but interpretability challenges hinder its clinical adoption. This study enhances interpretability in Chest X-ray (CXR) classification by using concept bottleneck models (CBMs) and a multi-agent Retrieval-Augmented Generation (RAG) system for report generation. By modeling relationships between visual features and clinical concepts, we create interpretable concept vectors that guide a multi-agent RAG system to generate radiology reports, enhancing clinical relevance, explainability, and transparency. Evaluation of the generated reports using an LLM-as-a-judge confirmed the interpretability and clinical utility of our model's outputs. On the COVID-QU dataset, our model achieved 81% classification accuracy and demonstrated robust report generation performance, with five key metrics ranging between 84% and 90%. This interpretable multi-agent framework bridges the gap between high-performance AI and the explainability required for reliable AI-driven CXR analysis in clinical settings. Our code will be released at https://github.com/tifat58/IRR-with-CBM-RAG

**Keywords:** Interpretable Radiology Report Generation · Concept Bottleneck Models · Multi-Agent RAG · Explainable AI · LLMs · VLMs

## 1 Introduction and Related Work

Deep learning has significantly improved diagnostic accuracy in medical imaging, especially in chest X-ray (CXR) analysis for conditions such as pneumonia, lung cancer, and tuberculosis [31,16]. However, these models often function as "*black boxes*", limiting interpretability and clinician trust in critical fields like radiology [12]. Traditional CXR classification methods lack transparency, providing little insight into model predictions. Although automated radiology report generation could streamline diagnostics and improve reporting consistency [18],

---
⋆ Corresponding author. Email: hasan.alam@dfki.de

it remains challenging to interpret and validate, which hinders its adoption in clinical settings that demand explainability [15].

To address these challenges, we combine concept bottleneck models (CBMs) [19] with a multi-agent Retrieval-Augmented Generation (RAG) [20,21] system to enhance interpretability in CXR classification and report generation (Fig. 1). While RAG systems have shown promise for radiology report generation by improving factual accuracy and reducing irrelevant content [27,3,33,36,22], they face limitations: retrieval-only methods can introduce noise or redundancy, and generative models risk producing clinically inconsistent outputs. Our approach addresses these issues by modeling relationships between visual features and clinical concepts, creating interpretable concept vectors that support accurate classification and clear, clinically relevant reporting.

CBMs enable interpretability by introducing concept layers where each neuron corresponds to a human-interpretable clinical concept [19,7]. Although traditional CBMs require extensive human annotation, which limits scalability, recent adaptations leverage vision-language models like CLIP [26] to align visual features with concepts [25,39]. While these models typically incorporate complex components like residual connections and large concept sets, recent work by [37] suggests that simpler designs may be more effective for medical data, achieving robust representation without extensive modifications. By bridging the gap between AI models and the interpretability necessary for effective medical decision-making, our approach aims to build an interpretable CXR interpretation system that empowers radiologists with greater insight into the AI's decision-making process, supporting more informed and trustworthy diagnoses.

## 2    Methodology

Our approach leverages concept bottleneck models (CBMs) [19], medical image embeddings and multi-agent RAG [20] to ensure robustness and interpretability in medical image classification and report generation as shown in Fig. 2, in two stages. First, disease classification with associated concept contributions (sec. 2.1). Second, robust report generation using relevant clinical documents and descriptors from stage 1 (sec. 2.2).

### 2.1    Interpretable Classification using Concept Bottleneck

Following the [25], we adopt CBMs with automatic concept discovery for CXR classification. We use GPT-4 [1] to query a set of $N = 20$ medical descriptors (concepts) for each disease category following prompt questionnaires described in [37]. These concepts are aggregated into a concept set $C = \{c_1, c_2, \cdots, c_N\}$ for each disease category. To extract image embeddings, we utilize the ChexAgent model [9], a multimodal vision language model (VLM) tuned explicitly for CXR interpretation. ChexAgent outputs image embeddings $V \in \mathbb{R}^{H \times D}$, where $H$ is the image height, and $D$ is the embedding dimension. Large language models have demonstrated effectiveness in encoding clinical knowledge [30]. For each

Fig. 1: Input (*blue*) to output (*green*) pipeline: Given a CXR as input, the Concept Bottleneck Model predicts clinical attributes (concepts) and their contributions in an intermediate step, followed by predicting the disease class. The multi-agent RAG system then generates a comprehensive report, incorporating clinical interpretations and insights drawn from relevant clinical documents

concept $c_i$, a text embedding $t_i \in \mathbb{R}^D$ is generated using the Mistral Embed Model [17]. This model is chosen for its efficiency and accuracy in embedding textual data into a high-dimensional vector space, suitable for subsequent similarity computations.

Given an image embedding $V$ and a set of concept embeddings $t_i$, we computed a similarity matrix $M_{i,j} \in \mathbb{R}^{H \times N}$ using cosine similarity, where $i$ indexes over the image embedding pixels, and $j$ indexes over the concept embeddings. To reduce the dimensionality and focus on the most salient features, max pooling is applied to each similarity matrix $M_{i,j}$. This results in a singular value $s_i = \max(M_i)$ for each concept, forming a concept vector $e = (s_1, s_2, \cdots, s_N)$. After that, the concept vector $e$ is normalized to a scale between 0 and 1 to maintain interpretability. A fully connected layer $W_F \in \mathbb{R}^{M_c \times N}$ is then applied to classify the images into $M_c$ categories. The classification logits are computed as $z_i = W_F \cdot e_i$, where $z_i$ is the logit vector for the $i$-th image. Log-softmax is applied to the logits to obtain the log probabilities. The model is trained using categorical cross-entropy loss [4]. During inference steps, the model then provides the predicted class and associated concepts and their contributions, which are explicitly used in the report generation step.

Fig. 2: Proposed architecture for the interpretable report generation. (*Top*) For a CXR image, disease class, and concept contribution scores are predicted using a CBM model with automatic concept discovery. (*Bottom*) Based on these contributions, a multi-agent RAG system generates reports using relevant clinical documents. The chain-of-thought reasoning ensures that detected features contribute to accurate classification and report generation, with the final output evaluated for robustness and clinical relevance by LLM as judge [40] evaluation.

## 2.2   Robust Explanation-based Radiology Report Generation

We collected clinical documentation for each disease category from the National Institutes of Health (NIH). To facilitate efficient retrieval and analysis, embeddings for each document $d_i$ were generated using the OpenAI embedding model [5], yielding vectors $\mathbf{v}_i \in \mathbb{R}^n$ such that $\mathbf{v}_i = f_{\text{embed}}(d_i)$, where $f_{\text{embed}}$ is the embedding function. These embeddings were stored in Qdrant vector database [3] $\mathcal{Q} = \{\mathbf{v}_i \mid i = 1, 2, \dots, N\}$, with $N$ as the number of documents. The multi-agent framework for Retrieval-Augmented Generation (RAG) includes specialized agents for each disease category, each implemented as a Reasoning and Acting (ReAct) agent [38]. For a classified disease category $M_c$, the corresponding ReAct agent $A_C$ is activated to retrieve relevant document embeddings $D_C$ from $\mathcal{Q}$, based on similarity to an input query $q$: $D_C = \{\mathbf{e}_j \mid \text{sim}(\mathbf{v}_j, q) \geq \tau\}$, where $\text{sim}(\cdot)$ is a similarity function and $\tau$ a relevance threshold.

Alongside the ReAct agents, the framework incorporates a Radiologist Agent $A_R$ and a Medical Writer Agent $A_W$. The Radiologist Agent $A_R$ uses the activated ReAct agent $A_C$ as a tool to retrieve relevant clinical information from $D_C$ and, based on identified concepts $c_k$ from earlier stages, calculates an influence score $s_k = \text{influence}(c_k, D_C)$ for each concept, where $\text{influence}(\cdot)$ assesses relevance and diagnostic significance. These scores are assembled into a summary vector $\mathbf{s} = (s_1, s_2, \dots, s_m)$, with $m$ as the number of relevant concepts. The Med-

---

[3] https://qdrant.tech/

ical Writer Agent $A_W$ receives this vector $\mathbf{s}$ and composes a report by applying a generation function $f_{\text{gen}}$, resulting in report $= f_{\text{gen}}(\mathbf{s})$. The final output $y$ for a query $q$ in category $C$ can be represented as $y = A_W(A_R(A_C(q, \mathcal{Q})))$, where each agent sequentially processes and enhances the information to generate a detailed radiology report. This framework was implemented using CrewAI [4] and facilitated by LlamaIndex [5], ensuring efficient retrieval and high-quality report generation.

## 3   Results and Discussion

We evaluate the performance of both the interpretable classification using CBMs and the report generation module on the COVID-QU Dataset [10], compiled by Qatar University, comprising 33,920 CXR images across three classes: COVID-19 (11,956 images), Pneumonia (11,263 images), and Normal (10,701 images).

We compare the classification results with two types of baselines: visual encoders: CLIP [26], Bio-VIL [2], and two other methods that are trained on CBMs: Label-free CBM[25] and Robust CBM [37]. As shown in Table 1, Our classification model attains an accuracy of 81% on the Covid-QU dataset. To further evaluate the interpretability and robustness of the concept contribution, we explored concept intervention techniques to correct the predictive output of the model [29,32]. In Fig. 3b, we observe that correcting 3-4 concepts for misclassified test samples based on decreasing order in contribution scores yields a significant increase in performance. In Fig. 3a, the model performance on the test set is evaluated while removing the concept contributions in descending (Max Contribution), ascending (Min Contribution), and Randomly. The sharp decline in performance in the case of Max Contribution validates that the models indeed learn to predict from the concept contributions. The x-axis for Fig. 3a is in log-scale for better interpretation.

To evaluate the effectiveness of our report generation process using the multiagentic approach, we generate medical reports for all three classes: Pneumonia, COVID-19, and Normal. We compare the report generated with the same generated using GPT-4 and single-agent RAG, where one single agent is responsible for retrieving and generating reports from relevant documents. Each report is

---

Table 1: Classification Performance Comparision.

| Model | Covid-QU | Intepretability |
|---|---|---|
| CLIP [26] | 0.47 | No |
| Bio-VIL [37] | 0.78 | No |
| Label-free CBM [25] | 0.72 | Yes |
| Robust CBM [37] | 0.78 | Yes |
| Ours | **0.81** | Yes |

Table 2: Clustering Evaluation for Report Generation Approaches

| Metric | GPT4 | Single Agent | Multi-Agent |
|---|---|---|---|
| Silhouette | 0.37 | 0.41 | 0.27 |
| Davies-Bouldin | 1.11 | 0.96 | 1.44 |
| Calinski-Harabasz | 69.94 | 93.99 | 44.78 |
| Dunn | 0.54 | 0.73 | 0.36 |

(a) Model performance while removing concept features using different strategies.

(b) Model performance on intervening and correcting concepts for misclassified cases.

Fig. 3: Evaluation of the Robustness of Concept set of the classification model.

transformed into high-dimensional embeddings using the Mistral Embed Model, capturing the latent features. We apply t-distributed Stochastic Neighbor Embedding (t-SNE) [24] to reduce the dimensionality for visualization and analysis. To quantify how well the reports are clustered for different diseases, we compute Silhouette Score [28], Davies-Bouldin Index, [11], Calinski-Harabasz Index [6], and Dunn Index [14]. The clustering evaluation metrics in Table 2 offer insights into the quality of clustering for the different approaches. While the Single Agent method achieves the highest Silhouette Score (0.41), lowest Davies-Bouldin Index (0.96), and highest Calinski-Harabasz Index (93.99), indicating the tightest and most distinct clusters, it sacrifices clinical interpretability. In contrast, the Multi-Agent approach has slightly lower metrics, such as a Silhouette Score of 0.27 and a Davies-Bouldin Index of 1.44, yet this method reflects the clinical reality more accurately. Specifically, the COVID-19 and Pneumonia clusters show some proximity in the Multi-Agent approach, as shown in the t-SNE plot in Fig. 4, which is medically justified due to the biological overlap between these conditions. Additionally, the Normal cluster remains well-separated, confirming the model's ability to differentiate fundamentally different health states. Thus, despite lower numerical scores, the Multi-Agent approach better captures the nuances required for effective medical report generation.

To further evaluate the generated reports, we employed LLM as a Judge [40] approach, where five different LLMs — Llama 3.1 [13], Mistral [17], Gemma2



(a) Multi Agent RAG            (b) Single Agent RAG            (c) GPT 4

Fig. 4: t-SNE visualization of the embeddings of generated reports

Table 3: Evaluation of Report Generation Approaches using LLM as Judge

| Model | Semantic Similarity | | | Accuracy | | | Correctness | | | Clinical Usefulness | | | Consistency | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GPT4 | Single Agent | Multi Agent | GPT4 | Single Agent | Multi Agent | GPT4 | Single Agent | Multi Agent | GPT4 | Single Agent | Multi Agent | GPT4 | Single Agent | Multi Agent |
| Llama 3.1 8B | 0.84 | 0.82 | 0.84 | 0.91 | 0.87 | 0.91 | 0.92 | 0.88 | 0.91 | 0.89 | 0.85 | 0.84 | 0.93 | 0.85 | 0.89 |
| Mistral 7B | 0.79 | 0.88 | 0.89 | 0.84 | 0.88 | 0.94 | 0.85 | 0.85 | 0.95 | 0.88 | 0.92 | 0.96 | 0.86 | 0.88 | 0.96 |
| Gemma 2 9B | 0.77 | 0.79 | 0.85 | 0.80 | 0.80 | 0.82 | 0.81 | 0.83 | 0.87 | 0.69 | 0.67 | 0.78 | 0.76 | 0.77 | 0.83 |
| LLaVA 9B | 0.78 | 0.80 | 0.80 | 0.83 | 0.87 | 0.89 | 0.86 | 0.86 | 0.91 | 0.78 | 0.82 | 0.86 | 0.80 | 0.83 | 0.89 |
| GPT 3.5 Turbo | 0.79 | 0.75 | 0.82 | 0.84 | 0.78 | 0.86 | 0.86 | 0.79 | 0.88 | 0.81 | 0.75 | 0.86 | 0.84 | 0.76 | 0.88 |
| Average | 0.79 | 0.80 | 0.84 | 0.84 | 0.84 | 0.88 | 0.86 | 0.84 | 0.90 | 0.81 | 0.80 | 0.86 | 0.84 | 0.81 | 0.89 |

[34], LLaVA [23], and GPT-3.5 Turbo [5], were used to assess the reports for Semantic Similarity, Accuracy, Correctness, Clinical Usefulness, and Consistency against a ground-truth reference generated by Dragonfly-Med [8], a multimodal biomedical visual-language model by Together AI, fine-tuned from Llama 3 that achieved state-of-the-art performance on several benchmarks. Table 3 shows that using a multi-agent approach in report generation significantly enhances performance across all metrics compared to single-agent methods. For example, models like Mistral 7B exhibit notable improvements when employing the multi-agent approach, with Correctness increasing from 0.85 (Single Agent) to 0.95 (Multi-Agent) and Clinical Usefulness from 0.92 to 0.96. For further qualitative analysis, we use the Mixture of Agents (MoA) [35] approach. For our evaluation, Llama3.1 [13] and Mistral [17] act as proposer agents, while Medllama2 serves as the aggregator agent. The qualitative feedback from this process is used to perform a binary classification using GPT4-o, determining whether the generated report is clinically valid. As shown in the table below, the MoA method achieves scores of 0.81 with GPT-4, 0.82 with the Single Agent approach, and 0.85 with the Multi-Agent approach. The incremental improvement in scores—from 0.81 to 0.85—indicates that leveraging multiple agents enhances the quality and clinical validity of the generated reports, which aligns with the findings of [21].

## 4   Conclusion and Future Work

This paper introduces an interpretable framework that integrates CBMs with a multi-agent RAG system for CXR classification and report generation. By capturing relationships between visual features and clinical concepts, our model provides competitive performance and clinically relevant explanations. The multi-agent RAG system enhances report quality and relevance, validated through evaluations by LLMs. This work bridges high-performing AI with the interpretability essential for clinical use, promising reliable AI-driven chest X-ray analysis. Future efforts will be to validate this approach to other imaging modalities and further refine the multi-agent system for adaptability and robustness.

## Acknowledgments

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
3. Bernardi, M.L., Cimitile, M.: Report generation from x-ray imaging by retrieval-augmented generation and improved image-text matching. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
4. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
5. Brown, T.B.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
6. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods $3$(1), 1–27 (1974)
7. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., Dvijotham, K.: Interactive concept bottleneck models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 5948–5955 (2023)
8. Chen, K., Thapa, R., Chalamala, R., Athiwaratkun, B., Song, S.L., Zou, J.: Dragonfly: Multi-resolution zoom supercharges large visual-language model. arXiv preprint arXiv:2406.00977 (2024)
9. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208 (2024)
10. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? Ieee Access $8$, 132665–132676 (2020)
11. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence (2), 224–227 (1979)
12. De Falco, I., De Pietro, G., Sannino, G.: Classification of covid-19 chest x-ray images by means of an interpretable evolutionary rule-based approach. Neural Computing and Applications $35$(22), 16061–16071 (2023)
13. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
14. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics $4$(1), 95–104 (1974)
15. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery $9$(4), e1312 (2019)
16. Hwang, E.J., Park, S., Jin, K.N., Im Kim, J., Choi, S.Y., Lee, J.H., Goo, J.M., Aum, J., Yim, J.J., Cohen, J.G., et al.: Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA network open $2$(3), e191095–e191095 (2019)
17. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)

18. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
19. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
20. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
21. Li, J., Zhang, Q., Yu, Y., Fu, Q., Ye, D.: More agents is all you need. arXiv preprint arXiv:2402.05120 (2024)
22. Liang, S., Sánchez, P., Sonntag, D.: Optimizing relation extraction in medical texts through active learning: A comparative analysis of trade-offs. In: Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024). pp. 23–34 (2024)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
24. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
25. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. arXiv preprint arXiv:2304.06129 (2023)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
27. Ranjit, M., Ganapathy, G., Manuel, R., Ganu, T.: Retrieval augmented chest x-ray report generation using openai gpt models. In: Machine Learning for Healthcare Conference. pp. 650–666. PMLR (2023)
28. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
29. Shin, S., Jo, Y., Ahn, S., Lee, N.: A closer look at the intervention procedure of concept bottleneck models. In: International Conference on Machine Learning. pp. 31504–31520. PMLR (2023)
30. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
31. Srivastav, D., Bajpai, A., Srivastava, P.: Improved classification for pneumonia detection using transfer learning with gan based synthetic image augmentation. In: 2021 11th international conference on cloud computing, data science & engineering (confluence). pp. 433–437. IEEE (2021)
32. Steinmann, D., Stammer, W., Friedrich, F., Kersting, K.: Learning to intervene on concept bottlenecks. arXiv preprint arXiv:2308.13453 (2023)
33. Sun, L., Zhao, J., Han, M., Xiong, C.: Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. arXiv preprint arXiv:2407.15268 (2024)
34. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al.: Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118 (2024)
35. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., Zou, J.: Mixture-of-agents enhances large language model capabilities. arXiv preprint arXiv:2406.04692 (2024)

36. Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., Yao, H.: Mmed-rag: Versatile multimodal rag system for medical vision language models. arXiv preprint arXiv:2410.13085 (2024)
37. Yan, A., Wang, Y., Zhong, Y., He, Z., Karypis, P., Wang, Z., Dong, C., Gentili, A., Hsu, C.N., Shang, J., et al.: Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 (2023)
38. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)
39. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
40. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023)