医療用大規模日本語視覚言語モデルの構築

安道健一郎 [1,2] 黒瀬優介 [2,1] 菊地智博 [3] 牧元久樹 [3] 小寺聡 [2] 小林和馬 [4,5]
合田和生 [2] 村尾晃平 [5] 吉田浩 [5] 田村孝之 [6] 合田憲人 [6,5] 喜連川優 [6,2] 原田達也 [2,1,5]
[1] 理化学研究所　[2] 東京大学　[3] 自治医科大学　[4] 国立がん研究センター
[5] 国立情報学研究所　[6] 情報・システム研究機構
{ando, kurose, harada}@mi.t.u-tokyo.ac.jp
{tmhk0712, klein2004att, koderasatoshi}@gmail.com
kazumkob@ncc.go.jp　{k-murao, h-yoshida, aida}@nii.ac.jp
tamura.takayuki@rois.ac.jp　{kgoda, kitsure}@tkl.iis.u-tokyo.ac.jp

概要

視覚言語モデル（VLM）は社会に広く普及している一方、医療機関では患者情報を取り扱う関係上、導入があまり進んでいない。そのためオープンな日本語医療 VLM の開発が求められているが、現状は存在していない。本研究では、総パラメータ数 14B のオープンな日本語医療 VLM を初めて構築した。モデルを構築するうえで最も障壁である訓練データの不足を補うため、英語データを加工し約 1,200 万件の日本語医療学習データを作成した。さらに、推論過程を明示的に出力させる CoT 形式データを導入し、推論あり／なしの 2 種のモデルを学習した。評価では、CT 画像および X 線画像に対し、推論ありモデルがオープン VLM の中で最良の性能を示した。

1　はじめに

近年、大規模視覚言語モデル（VLM）の発展により、画像と言語を統合してユーザと対話を行う基盤モデルが急速に普及している。GPT シリーズ [1] や Gemini[2] などを筆頭に活発に VLM が開発されており、国内でも Sarashina[1)] や Asagi[3] など多くのモデルが公開されている。これらの基盤モデルは日常的な対話アシスタントから専門的な産業応用まで、実社会の様々な場面で活用され始めている。

一方、医療分野での活用に目を向けると、国内では医療情報は厳重に保護されているため取り扱いには特段の配慮が求められる。医療機関における生成モデル活用に関しては公的ガイドライン [4] が改定され、クラウド利用を含む環境整備が進みつつあるものの、依然として多くの医療機関では組織外への医療情報の持ち出しは原則的に禁止といった運用を続けている。したがって、医療分野では患者データを外部へ送らずに運用できる、オンプレミス前提のモデルが強く求められる場面が多い。その流れを受けて、医療分野でも高い性能を有するような医療領域特化のオープンな生成モデルが公開され始めており、既に大規模言語モデル（LLM）では多様な医療

図 1: 本研究で構築した日本語医療 VLM の生成例。上の例は CT 画像に対する質問応答。下の例は X 線画像に対する読影レポートを出力しており、推論なし・ありの 2 種類のモデルの結果を図示している。

1) https://huggingface.co/sbintuitions/sarashina2-vision-14b

文書から学習された日本語医療 LLM[2)]が開発されている。しかし、VLM では未だ医療に特化したオープンなモデルは存在していない。

本研究では、新たに医療分野特化のオープンな日本語 VLM を開発した。大規模な医療モデルを開発する際の最大の問題は学習データの不足である。特に日本語の画像テキスト対の公開データは少ない。我々は英語の医療データを活用して日本語の学習データを約 1,200 万件作成することでこの問題に対処した。また、性能向上のための工夫として Chain of thought（CoT）[5] を模した深い推論過程を出力させた。本モデルと訓練に用いたデータの全て、評価に用いたデータの一部は公開予定である。本研究の主な貢献は，以下の通りである：（1）VLM 用の日本語訓練データを作成し、総パラメータ数 14B の日本語医療 VLM を初めて構築した。（2）CT 画像と X 線画像の読影レポート生成で評価し、我々の推論ありモデルはオープンな VLM の中で最も高い性能を示すことが確認された。（3）特別な制約を持たない商用利用可のライセンスを持つモデルやデータを用いて VLM を作成した。

2 データセット構築

日本語の医療分野で公開されている画像テキスト対のデータセット [6, 7, 8] は最大でも数十症例と小規模であり、VLM の学習に使用できる規模のものは存在しない。そこで、本研究では既存の VLM や LLM を活用した複数のステップからなるパイプラインで訓練データを作成する。本研究では後述する Llava[9] 形式のモデルを採用するので 2 ステージの訓練を行う。Stage1 のために画像キャプション対、Stage2 のために VQA のデータが必要である。

2.1 データ収集

データ元として、PubMed Central（PMC）から医療論文の図とキャプションを集めた Open-PMC-18M[10] というデータセットを用いた。このデータセットは PMC から複合図を分離し、医療用 Vision and Language の学習に有効でないグラフなどをフィルタリングすることで既存の PMC 由来のデータセットを高品質にした約 1,800 万の画像テキスト対で構成されている。PMC では個別の論文ごとに著作権のライセンスが規定されている。そのため、

2） https://huggingface.co/SIP-med-LLM/SIP-jmed-llm-3-13b-OP-4k-base

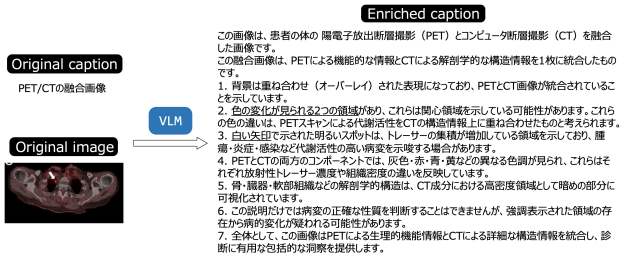

図 2: 再キャプショニングの例。図中のキャプションは英語を日本語に翻訳したもの。

Open-PMC-18M にも商用、非商用ライセンスのデータが混在する。本研究では商用利用可ライセンスのデータのみをフィルタリングして使用した。結果、約 1,200 件のデータが得られ、Stage1 ではこのキャプションデータを訓練データとして用いる。

2.2 再キャプショニング

PMC のキャプションは数フレーズや 1 文などの端的なものが多い（図 2）。よりリッチなキャプションを獲得するため、オリジナルキャプションと画像を VLM に入力して再キャプショニングを行なった。VLM は InternVL3.5 38B[11] を採用した。GPT シリーズなどのクローズドモデルは生成物に対する特殊ライセンスが付与されることが多いので使用しなかった。図 2 の例では、元のキャプションで記述されていなかった矢印や蛍光領域の情報などが増加していることが確認できる。過去の VLM 構築研究 [12, 13, 14, 15] においても類似の処理が行われている例があり、性能の向上が確認されている手法である。

2.3 VQA 生成

前ステップで作成したリッチなキャプションを用いて Visual Question Answering（VQA）の作成を行う。図 3 のように、テキストのみを入出力にして LLM を用いて VQA 生成を行う。これは先行研究 [9, 16, 17, 18] でも多く取り入れられている方法である。PMC のオリジナルキャプションとリッチなキャプションの 2 つを入力し、質問と答えの組を生成するように LLM に指示をする。

本研究では通常の VLM に追加して CoT 形式の VLM も作成する。そのため、答えのみを出力するデータ（**Normal data**）に加えて、ステップバイステップの推論を出力するデータ（**Reasoning data**）を作成する。図 3 に示すように、Reasoning data は 1.

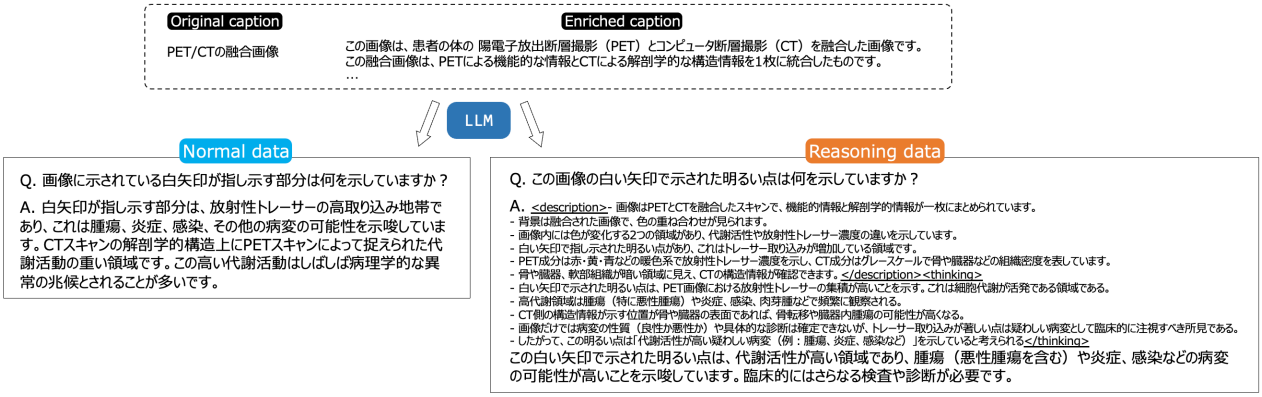

図 3: LLM を用いた VQA 生成の例。入力のキャプションは英語を日本語に翻訳したもの。Reasoning data は CoT 形式を模したデータで、下線部のタグは推論のステップを示す。

画像の詳細な記述、2. 記述に基づいた推論、3. 最終的な答えの 3 ステップに分かれる。データ内では 1、2 のステップをそれぞれ<description>、<thinking>のタグで囲うことで明示的に識別する。この手法はいくつかの研究 [19, 20, 21] でも採用されており、推論ステップの出力を強制することで、性能向上を図る。

LLM は Normal data、Reasoning data 共に GPT-oss 120B[22] の reasoning_effort を high にして使用した。クローズドモデルは生成物に対する特殊ライセンスが付与されることが多いので使用しなかった。

3 モデルの構築

モデルアーキテクチャ 本モデルは LLaVA[9] をベースとしたエンコーダ・デコーダ型の VLM である。LLaVA は画像特徴を抽出するエンコーダ、テキストを生成するデコーダ（LLM）、そして両者を結合するプロジェクター（2 層の MLP）によって構成されている。本モデルでは画像エンコーダに SigLIP[23] を、テキストデコーダに SIP Healthcare generative AI project が提供する日本語医療 LLM である SIP-jmed-llm[3] を用いた。結果、本モデルの総パラメータ数は約 14B となった。

モデルの訓練 モデルの実装には分散学習フレームワークである Megatron-LM[24] を用いた。前述の通り、LLaVA の訓練方法に従って本研究でも 2 段階の学習手法を採用した。Stage1: プロジェクター層のみを学習。Stage2: プロジェクター層とテキストデコーダを学習。Stage1 の訓練には画像キャプション対を、Stage2 の訓練には VQA データを約 1,200 万件用いた。その際、Stage2 の訓練データに **Normal data** を使用して**推論なしモデル**を学習し、**Reasoning data** を使用して**推論ありモデル**を学習した。学習には NVIDIA H200 を 8 枚使用し、stage1 の学習には約 3 日、stage2 の学習には約 6 日を要した。また、データ作成時、再キャプショニングは H200 を 10 ノード（80 枚）用いて 12 日間行った、VQA 作成は H200 を 10 ノード用いて 18 日間行った。

3） SIP-jmed-llm-3-13b-OP-4k-base の一世代前モデル

4 実験設定

4.1 評価データ

評価データとして 2 種類の医療画像データを用いる。それぞれのデータの画像とそれに対する質問応答の例は図 1 に示す。

CT 画像 自治医科大学で収集された消化器癌患者の CT 検査とその読影レポートを元に作成された質問応答データを評価データとして用いる。1 枚の CT 画像に付き 1 組の VQA が医師により付加されており、935 件をテストデータとして用いる。質問、応答の平均文字数はそれぞれ 21 文字、25 文字である。

X 線読影レポート 英語の大規模医療データである MIMIC CXR[25] を日本語に翻訳したデータを評価データとして用いる。MIMIC CXR から 1 枚の X 線画像を元に記述された読影レポートのみをフィルタリングし、GPT-5 を用いて日本語に翻訳した。テストデータとして 300 件、Fine-tuning を行うための訓練データとして 10,050 件を作成した。MIMIC CXR のレポートにはいくつかのセクションが含まれるが、そのうち FINDINGS と IMPRESSION を連結して Answer として採用し、INDICATION、HISTORY、EXAMINATION、TECHNIQUE を患者の

		CT	X 線	
Model	Size	RougeL	RougeL	LAJ
GPT-5	–	–	16.11	3.59
GPT-4.1	–	–	13.95	2.45
Gemma3	27B	15.99	15.50	2.16
Phi4	6B	36.31	16.10	1.77
InternVL	38B	24.88	16.77	2.54
Villa-ip	14B	40.48	16.57	2.06
Sarashina	14B	17.28	10.49	1.58
Asagi	14B	29.13	16.18	1.54
Llava-Med	8B	0.90	0.05	1.01
推論なし	14B	27.36	11.31	1.72
推論あり	14B	44.45	17.26	2.73
推論なし（SFT）	14B	–	30.16	3.14
推論あり（SFT）	14B	–	30.04	3.50

表 1: CT 画像・X 線画像における各モデルの評価結果。LAJ は LLM as a Judge を示す。CT データは外部に持ち出すことができないため、クローズドモデルや LAJ での評価を行っていない。

背景情報として連結してプロンプトに含めた。Answer と背景情報はそれぞれ平均で 241 文字、97 文字である。

4.2 評価指標

CT 画像に対する QA と X 線画像に対する読影レポート生成において評価指標は ROUGE-L[26] F1 を用いる。また、X 線画像タスクに限り、GPT-5 を用いた LLM as a Judge（LAJ）で評価を行う。CT 画像はデータの性質上、API を通して LLM にアクセスすることができなかったので LAJ は行っていない。

4.3 性能比較モデル

実験では本研究で開発した推論なし、推論ありモデルに加えて、X 線読影レポートタスクでは訓練データを用いて Fine-tuning した SFT モデルも使用した。比較対象として、既存の VLM モデルから 9 種類を採用して実験を行った。9 種類のうち、gpt-5-2025-08-07、gpt-4.1-2025-04-14 は汎用クローズド VLM であり、Gemma-3-27b-it、Phi-4-multimodal-instruct、InternVL3_5-38B は英語を中心に学習されたオープン VLM であり、llm-jp-3-vila-14b、sarashina2-vision-14b、Asagi-14B は日本語を中心に学習された汎用オープン VLM である。llava-med-v1.5-mistral-7b は唯一英語の医療データを中心に学習された医療用 VLM である。なお、複数のモデルサイズがある VLM はできる限り 14B 以上のモデルを採用し、存在しない場合は最もサイズの大きいモデルを採用した。各モデルの総パラメータ数は表 1 に示す。

5 評価結果

CT 画像の評価結果を表 1 に示す。本研究で開発した推論ありモデルが最も良い性能を示している。他の全ての英語汎用モデルや日本語汎用モデルを上回っており、日本語医療データを中心に学習することは医療タスクの性能向上に顕著に貢献するということが確認された。さらに、同サイズ帯で最高レベルの性能を有する InternVL の性能を上回ったことから、我々の推論ありモデルは現時点で公開されているオープン VLM の中で最高レベルの性能を持つ日本語医療 VLM と言える。次点で Villa-ip、次に Phi4 の性能が高かった。Llava-Med に関しては、日本語を出力することができずスコアが著しく低い結果となった。推論なしモデルと推論ありモデルを比較すると、推論なしモデルは全モデル中 5 位という中位に留まるが、推論を行うと全モデル中 1 位となり、162%の性能改善を示した。これは、推論過程を出力させることが最終的な出力の質に大きく影響することを表している。

X 線読影レポートの結果に着目すると、ROUGE-L では我々の推論ありモデルが最も性能が良く、次に InternVL が続く。LAJ では GPT-5 が最高性能のモデルとなり、次に我々の推論ありモデルが続く結果となった。しかしながら、オープンモデルの中では我々のモデルが最高性能であった。Fine-tuning を行うと GPT-5 に迫るスコアになったことから、目的タスクが事前に明確な場合はオープンモデルを Fine-tuning することが有力な戦略であることが示唆される。また、MIMIC CXR は広く使用されているデータセットであり、クローズドモデルは既に MIMIC CXR のデータを学習している可能性がある点を留意する必要がある。

6 おわりに

本研究では日本語医療 VLM を始めて構築した。また、推論過程を出力するような VLM を開発し、２種類のデータセットで評価して既存のオープン VLM の性能を上回ることを確認した。本研究で開発したモデルや、学習に用いたデータ、CT を除く評価データは全て公開予定である。

謝辞

本研究は戦略的イノベーション創造プログラム (SIP)「統合型ヘルスケアシステムの構築」JPJ012425，JST ムーンショット型研究開発事業 JPMJMS2011、CREST 課題番号 JP-MJCR2015、JSPS 科研費 JP23K16990、JP23K19971、及び東京大学 Beyond AI 研究推進機構の基礎研究費 (AI 自体の進化) の支援を受けたものです。

参考文献

[1] OpenAI. GPT-5 system card. 2025.

[2] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. **arXiv**, 2025.

[3] 上原 康平, 黒瀬 優介, 安道 健一郎, Jiali Chen, Fan Gao, 金澤 爽太郎, 坂本 拓彌, 竹田 悠哉, Boming Yang, Xinjie Zhao, 村尾 晃平, 吉田 浩, 田村 孝之, 合田 憲人, 喜連川 優, 原田 達也. Asagi: 合成データセットを活用した大規模日本語 VLM. 言語処理学会第 31 回年次大会, 2025.

[4] 厚生労働省. 医療情報システムの安全管理に関するガイドライン 第 6.0 版. 2023.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **NeurIPS**, 2022.

[6] Yuta Nakamura, Shohei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Clinical Comparable Corpus Describing the Same Subjects with Different Expressions. **Stud Health Technol Inform**, 2022.

[7] Shuntaro Yada Shoko Wakamiya Eiji Aramaki Yuta Nakamura, Shouhei Hanaoka. NTCIR-17 MedNLP-SC Radiology Report Subtask Overview: Dataset and Solutions for Automated Lung Cancer Staging. **Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies**, 2023.

[8] Yuta Nakamura, Koji Fujimoto, Jonas Kluckert, Michael Krauthammer, Jun Kanzawa, Akira Katayama, Tomohiro Kikuchi, Ryo Kurokawa, Wataru Gonoi, Peitao Han, et al. Ntcir-18 radnlp 2024 overview: Dataset and solutions for automated lung cancer staging. **Proceedings of the 18th NTCIR Conference**, 2025.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **NeurIPS**, 2023.

[10] Negin Baghbanzadeh, Mohammed Saidul Islam, Sajad Ashkezari, Elham Dolatabadi, and Arash Afkanpour. Open-pmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning. **arXiv**, 2025.

[11] Weiyun Wang, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. **arXiv**, 2025.

[12] Shijie Zhou, Ruiyi Zhang, Yufan Zhou, and Changyou Chen. A high-quality text-rich image instruction tuning dataset via hybrid instruction generation. **COLING**, 2025.

[13] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. **ECCV**, 2024.

[14] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multimodal instruction tuning dataset with extensive diversity. **Science China Information Sciences**, 2024.

[15] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. **arXiv**, 2024.

[16] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. **NeurIPS**, 2023.

[17] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. **arXiv**, 2023.

[18] Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, et al. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. **Findings of ACL**, 2024.

[19] Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. **arXiv**, 2025.

[20] Caption as reward: Enhancing vision-language reasoning through dense visual description. **Submitted to ICLR (under review)**, 2026.

[21] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. **ICCV**, 2025.

[22] OpenAI. gpt-oss-120b & gpt-oss-20b model card. **arXiv**, 2025.

[23] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. **ICCV**, 2023.

[24] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. **arXiv**, 2019.

[25] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. **Scientific data**, 2019.

[26] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. **Text Summarization Branches Out**, 2004.

表 2: モデル学習時の主なハイパーパラメータ

	Stage1	Stage2
バッチサイズ	576	128
訓練イテレーション	21,051	94,728
学習率関連		
— 最大学習率	1e-3	2e-5
— 最小学習率	1e-8	1e-8
— スケジューラー	コサイン	コサイン
— Warmup ratio	0.03	0.03
Weight decay	0.1	0.1
Optimizer	AdamW	AdamW

A　データサイズの詳細

訓練データのサイズを詳細に述べる。Open-PMC-18M には 17,867,999 件のデータが存在する。そこから非商用データを除き、キャプション生成を行うと 12,125,556 件にフィルタリングされる。stage1 はこの 12,125,556 件のキャプションデータで訓練する。さらに 12,125,556 件から VQA を生成し、生成不良なデータをフィルタリングすると 12,125,203 件まで減少する。stage2 は 12,125,203 件で訓練した。

B　学習時のハイパーパラメータ

本研究で用いたモデル学習時の主なハイパーパラメータの詳細を表 2 に示す。

C　本研究で使用したプロンプト

データ作成、LAJ で用いたプロンプトを図 4 に示す。

キャプション生成

Please create a detailed caption for this medical image. Provide the description as a bullet-point list. Use plain text only—do not use Markdown (no asterisks, dashes, numbering, headers, or code blocks). Include every observable structure and feature in the image. Focus strictly on factual visual details and avoid any inference, interpretation, or diagnostic statements.

VQA 生成

下に示す「画像のキャプション」と「画像の詳細なキャプション」に基づいて、質問文と、その質問に対する回答を日本語で書いてください。
- タスクが VQA であることを考慮して、VQA として成立する質問回答を生成してください。
- 画像について質問してください。
- 解答に至るまでの推論過程を示すようにしてください。
- 回答は必ず最初に画像についての情報を一つ一つ描写してから、その後に推論を始め、最後に回答を書いてください。
- 回答中の画像についての描写部分は<description>、推論部分は<thinking>というタグで囲んでください。
- 必ず<description>中の描写に基づいて推論をしてください。それ以外の情報を参照して推論しないでください。参照したい画像情報があれば、全て必ず<description>中に記述してから参照してください。
-「画像のキャプション」や「画像の詳細なキャプション」に書かれている記述を直接参照する形で<description>や<thinking>、VQA を記述しないでください。
-「画像のキャプション」、「画像の詳細なキャプション」という語句を出力中で使わないでください。

LLM as a Judge

Task: Grade the CANDIDATE_REPORT by comparing ONLY with QUESTION + REFERENCE_REPORT (ground truth).
Do NOT use external radiology knowledge. Do NOT invent typical findings. Do NOT be influenced by the QUESTION’s suspected diagnosis; judge what is actually stated in REFERENCE_REPORT.

Core principles
- REFERENCE_REPORT is the only truth. Anything not supported there should NOT be assumed true.
- Evaluate clinical meaning (presence/absence, laterality, location, device presence/position). Ignore style.
- If the CANDIDATE has multiple sections (e.g., Findings/Impression/Reasoning), score the overall clinical message.
- If Findings match GT but Impression contradicts GT, give partial credit (do not auto-score 1

Step 1) Extract “key items” from REFERENCE_REPORT (internally)
A. Key positives (findings that matter clinically or are the point of the question)
B. Key negatives explicitly stated (e.g., “no pneumothorax/effusion/edema/pneumonia”, “no acute cardiopulmonary disease”)
C. Devices/lines and their position statements (PICC/CVC/ETT/NGT/IABP/pacemaker leads etc.)

Step 2) Compare CANDIDATE_REPORT to each key item and label errors
Define error severities (use these exact meanings):

...（省略）

Step 3) Compute score
- Start at 10.
- Subtract penalties: CRITICAL − 4 each, IMPORTANT − 2 each, MINOR − 1 each (MINOR total capped at 2).
- Clamp to integer 1–10.
- Practical anchor calibration:
- 9–10: essentially matches GT; at most minor extras/omissions.
- 6–8: one important error or a couple smaller mismatches, but main clinical message aligns.
- 3–5: major mismatch in impression OR multiple important errors; some correct content remains.
- 1–2: catastrophic (wrong laterality, misses key finding + invents major pathology/device malposition, or multiple critical errors).

図 4: プロンプト一覧