

Supplementary Materials

A. Inter-rater consistency analysis on qualitative metrics

As shown in **Table S1**, in this study, we used Gwet's AC1 for inter-rater consistency analysis. Cohen's κ and ICC are known to be sensitive to marginal distribution imbalances and rater-specific scoring tendencies in multiclass or ordinal data. Such conditions can lead to systematic underestimation of κ values, even when the actual inter-rater agreement is high, which is a phenomenon commonly referred to as the " κ paradox." Gwet's AC1 mitigates this issue by adopting a more robust estimation of chance agreement, thereby reducing the influence of prevalence and rater bias. As a result, AC1 provides a more stable and reliable measure of inter-rater consistency in settings with skewed category distributions or pronounced rater tendencies.

Notably, we dichotomized the five-point Likert scale at a cutoff of 4 (scores 4–5 as high and 1–3 as low) when calculating Gwet's AC1. The reasons are multi-fold. First, the primary aim of model evaluation is to determine whether a report has clinical utility; the distinction between scores 4 and 5 generally reflects stylistic or rhetorical preferences rather than substantive disagreement, whereas the difference between scores 1–3 and ≥ 4 represents a meaningful shift in usability. Dichotomization, therefore, aligns more closely with the underlying decision threshold. Second, although AC1 is more robust than Cohen's κ , it can still be influenced by imbalanced marginal distributions and rater tendencies toward adjacent categories; in settings where ratings cluster in the mid-to-high range, minor discrepancies between 4 and 5 or between 2 and 3 can artificially depress agreement. Applying a predefined clinical threshold and collapsing the scale reduces the weight of such non-essential discrepancies, improving statistical stability and interpretability. In our results, AC1 increased substantially after dichotomization, indicating that most rater differences occurred within the same score band rather than across the clinical threshold, thereby more accurately reflecting consensus on clinical utility.

Table S1 | Gwet's AC1 for inter-rater consistency analysis on the 900 Chinese cases

	Comprehensiveness	Explainability	Conciseness	Unbiasedness
DeepSeek-R1 (Rea.)	0.899	0.937	0.468	0.912
DeepSeek-R1 (Con.)	0.691	0.824	0.969	0.750
DeepSeek-V3_0324	0.802	0.861	0.952	0.822
GPT-4.5	0.877	0.830	0.911	0.797

Note: 0–0.20: slight; 0.2–0.4: fair; 0.4–0.6: moderate; 0.6–0.8: substantial; 0.8–1.0: almost perfect agreement.

1 **B. DeepSeek-R1 output examples**

2 **Table S2 | Example of a lung case**

Item	Content
Raw Chinese findings (from a radiologist)	双肺支气管血管束增多，右肺上叶体积减小，右肺上叶尖后段见一不规则肿块影，其内可见血管穿行，较大截面大小约 53mm×24mm，CT 值约为 84HU，增强呈明显强化，周围散在多发结节状、条索状密度增高影。右肺上叶部分段支气管及右主支气管部分闭塞。余肺散在结节，长径均小于 5mm。各叶、段支气管通畅。纵隔 4 区见肿大淋巴结，短径约 10mm。两侧胸腔未见积液。
Raw Chinese impression (from a radiologist)	1、右肺上叶不规则肿块影，考虑肺 Ca 并阻塞性炎症，请结合临床及病理。2、纵隔 4R 区淋巴结肿大。
Raw Chinese DeepSeek-R1 (Rea.)	嗯，我现在需要根据用户提供的 CT 影像结果来做出诊断结论。首先，我要仔细看看这些影像发现。 双肺支气管血管束增多，这可能意味着有炎症或者慢性病变，比如慢性支气管炎，但也可能是其他情况。右肺上叶体积减小，这可能提示有肺不张或者纤维化，导致肺叶收缩。接下来，右肺上叶尖后段有一个不规则的肿块，尺寸较大，约 53mm×24mm，CT 值 84HU，增强扫描明显强化。肿块内部有血管穿行，这可能说明肿块血供丰富，可能是恶性的，因为恶性肿瘤通常会有更多血管生成。不过，有些良性肿瘤或者感染性病变也可能有丰富的血供，需要进一步考虑。 周围有散在的多发结节和条索状密度增高影，这可能与肿瘤周围的炎症反应、淋巴扩散或转移有关。右肺上叶部分段支气管和右主支气管部分闭塞，这可能是因为肿瘤压迫或直接侵犯支气管，导致气道阻塞，这也常见于中央型肺癌。余肺的小结节，长径小于 5mm，可能是转移灶或者炎性结节，但需要结合其他表现来看。 纵隔 4 区有肿大淋巴结，短径 10mm，这可能提示淋巴结转移，但正常淋巴结的大小标准可能因位置而异，一般短径超过 10mm 通常考虑异常，这里刚好是 10mm，可能处于临界值，但结合其他表现，可能需要考虑转移可能性。两侧胸腔没有积液，这排除了胸水的可能，但恶性病变有时也可能没有积液。 综合分析，主要考虑右肺上叶的恶性肿瘤，尤其是肺癌可能性大。肿块形态不规则，增强明显强化，内部血管穿行，伴随支气管闭塞和纵隔淋巴结肿大，这些都支持恶性肿瘤的诊断。周围的多发结节和条索影可能代表肿瘤周围的浸润或转移。需要排除的包括结核、真菌感染或其他肉芽肿性疾病，但结合增强 CT 的表现，恶性肿瘤更可能。 因此，诊断结论应该是右肺上叶的恶性肿瘤，可能为肺癌，并伴有纵隔淋巴结转移。但用户要求仅给出诊断结论，不需要其他建议。
Raw Chinese DeepSeek-R1 (Con.)	右肺上叶尖后段周围型肺癌（伴纵隔淋巴结转移可能）。

Translated English findings (from a radiologist)	There is an increase in bronchovascular markings in both lungs. The right upper lobe shows volume reduction, and an irregular mass is observed in the apical-posterior segment of the right upper lobe, with vessels traversing through it. The largest cross-section measures approximately 53 mm × 24 mm, with a CT attenuation value of about 84 HU. The lesion exhibits marked enhancement after contrast administration, and multiple surrounding nodular and cord-like high-density shadows are scattered nearby. Partial occlusion of some segmental bronchi in the right upper lobe and part of the right main bronchus is noted. Scattered nodules are present in the remaining lung fields, all with long diameters less than 5 mm. The bronchi of each lobe and segment are otherwise patent. An enlarged lymph node is seen in the mediastinal region (station 4), with a short-axis diameter of approximately 10 mm. No pleural effusion is detected on either side.
Translated English impression (from a radiologist)	1. Irregular mass in the right upper lobe of the lung, suggestive of lung carcinoma with associated obstructive inflammation. Correlation with clinical findings and pathology is recommended. 2. Enlarged lymph node in the right lower paratracheal (4R) mediastinal region.
Translated English DeepSeek-R1 (Rea.)	Well, I now need to make a diagnostic conclusion based on the CT imaging results provided by the patient. First, I need to carefully examine all the imaging findings. There is an increase in bronchovascular bundles in both lungs, which may indicate inflammation or chronic lesions such as chronic bronchitis, but other possibilities cannot be excluded. The right upper lobe shows volume reduction, which may suggest atelectasis or fibrosis leading to contraction of the lobe. Next, there is an irregular mass in the apical-posterior segment of the right upper lobe, relatively large in size, about 53 mm × 24 mm, with a CT value of 84 HU and showing marked enhancement on contrast scan. Blood vessels can be seen traversing within the mass, which may indicate a rich vascular supply and a possible malignancy, as malignant tumors often show increased angiogenesis. However, some benign tumors or infectious lesions may also have an abundant blood supply, so this should be further considered. Multiple scattered nodules and cord-like high-density shadows are seen around the lesion, which may be related to an inflammatory reaction, lymphatic spread, or metastasis around the tumor. Partial obstruction of some segmental bronchi in the right upper lobe and of the right main bronchus is present, which may be caused by tumor compression or direct invasion of the bronchus, leading to airway obstruction—this is also commonly seen in central-type lung cancer. The remaining lung fields show small nodules with a long diameter of less than 5 mm, which may represent metastatic foci or inflammatory nodules, but should be evaluated in conjunction with other findings.

	An enlarged lymph node is noted in mediastinal station 4, with a short-axis diameter of 10 mm. This may suggest lymph node metastasis, though the normal size standard of lymph nodes varies by location. In general, nodes with a short-axis diameter exceeding 10 mm are considered abnormal. In this case, it is right at the threshold, but combined with other findings, metastasis should be considered. No pleural effusion is seen on either side, which rules out the possibility of pleural fluid, though malignant lesions sometimes present without effusion. Comprehensive analysis suggests that the main consideration is a malignant tumor in the right upper lobe, most likely lung cancer. The irregular shape of the mass, marked enhancement, intralesional vascularity, associated bronchial obstruction, and mediastinal lymph node enlargement all support the diagnosis of malignancy. The multiple surrounding nodules and cord-like shadows may represent peritumoral infiltration or metastases. Differential diagnoses include tuberculosis, fungal infection, or other granulomatous diseases, but based on the enhanced CT findings, malignancy is more likely. Therefore, the diagnostic conclusion should be a malignant tumor of the right upper lobe, likely lung cancer, with possible mediastinal lymph node metastasis. However, as the user requested, only the diagnostic conclusion needs to be provided, without additional recommendations, so the expression should remain concise.
Translated English DeepSeek-R1 (Con.)	Peripheral lung cancer in the apical-posterior segment of the right upper lobe (with possible mediastinal lymph node metastasis).

- 1 Note: Translation was performed using GPT-5 (accessed from its official website). The prompt for
- 2 translation: *Please translate the following text into standard and professional English + {Raw*
- 3 *Chinese text}*.

1 **Table S3 | Radiologist's scores of the above lung case**

Metric	Score
DeepSeek-R1 (Con.)	
Missed primary diagnoses (MPD)	0
Missed secondary diagnoses (MSD)	1 (Omitted the tumor-related obstructive inflammation noted in the original report)
Primary misdiagnoses (PMisD)	0
Secondary misdiagnoses (SMisD)	0
Comprehensiveness	2, 2, 2
Explainability	2, 2, 2
Conciseness	4, 5, 5
Unbiasedness	2, 2, 2
DeepSeek-R1 (Rea.)	
Missed primary diagnoses (MPD)	0
Missed secondary diagnoses (MSD)	0
Primary misdiagnoses (PMisD)	0
Secondary misdiagnoses (SMisD)	0
Comprehensiveness	4, 5, 5
Explainability	5, 5, 4
Conciseness	3, 3, 3
Unbiasedness	4, 5, 5

2 Note: For the qualitative metrics, the individual ratings of all three evaluators are illustrated.

3

4 In this example, DeepSeek-R1 (Rea.) preserves the logical reasoning underlying the diagnostic
5 impression, delineating evidential strength and uncertainty regarding the nature of the pulmonary
6 mass and mediastinal lymph nodes. This transparency supports the formulation of verifiable cT and
7 cN assessments and guides appropriate next-step evaluations, such as EBUS-TBNA or PET-CT,
8 before final staging. In contrast, DeepSeek-R1 (Con.) is very concise and omits the tumor-related
9 obstructive inflammation noted in the original report (an MSD). Such omission may bias staging
10 accuracy and therapeutic decision-making. In addition, DeepSeek-R1 (Con.) provides an overly
11 deterministic statement of the peripheral lung cancer without sufficient evidence. Therefore, from the
12 standpoint of ensuring accurate staging and evidence-based management, DeepSeek-R1 (Rea.)
13 offers superior practical utility by aligning interpretive reasoning with clinical decision-making
14 processes. DeepSeek-R1 (Rea.) thus holds greater clinical value for accurate tumor staging and
15 management planning.

1 Table S4 | Example of a breast case

Item	Content
Raw Chinese findings (from a radiologist)	左乳头后方可见肿块影，大小约 28×12mm，形态不规则，边缘不规则，可见毛刺，其内见多个点状、环形钙化影，周边组织受牵拉，乳头明显内陷，周围皮肤增厚。右乳外象限可见肿块影，大小约 23×19mm，形态不规则，边缘分叶，右乳见散在点状钙化影。双侧腋部见多个增大淋巴结，大者短径约 11mm（左侧）双侧胸大肌未见异常。右乳皮肤、皮下、乳头、乳晕未见异常。
Raw Chinese impression (from a radiologist)	左乳头后方肿块，累及左乳头及皮肤，左侧腋部淋巴结增大，BI-RADS 5 类，建议活检。 右乳外象限 9 点肿块，BI-RADS 4b 类，建议活检。 右腋部淋巴结可见，注意随访。 双乳散在纤维腺体型：ACR b
Raw Chinese DeepSeek-R1 (Rea.)	好的，我现在需要根据用户提供的钼靶影像结果来给出诊断结论。 首先，我要仔细阅读并理解这些影像描述的每一个细节。 左乳头后方的肿块影大小是 28×12mm，形态和边缘都不规则，还有毛刺。内部有多个点状和环形钙化，周围组织被牵拉，乳头明显内陷，皮肤增厚。这些特征看起来挺典型的恶性征象，比如毛刺和钙化形态可能提示恶性肿瘤，尤其是浸润性导管癌。乳头内陷和皮肤增厚可能意味着肿瘤牵拉周围组织或侵犯皮肤。 右乳外象限的肿块 23×19mm，形态不规则，边缘分叶，伴有散在点状钙化。虽然分叶状边缘可能见于良性肿瘤，但形态不规则和钙化也需警惕恶性可能。不过右乳的皮肤、皮下、乳头乳晕都正常，可能恶性程度较左乳低，但也不能排除。 双侧腋窝淋巴结增大， 左侧短径 11mm。正常淋巴结短径通常小于 10mm，所以增大可能提示转移。 不过需要考虑是否有其他原因导致淋巴结肿大，比如感染，但结合乳房内的恶性征象，转移的可能性更大。 综合分析，双侧乳房都有不规则肿块伴钙化，腋下淋巴结肿大，尤其左乳的特征非常符合乳腺癌的表现。右乳的肿块也需要考虑恶性可能，但需要更多信息确认。综合来看，最可能的诊断是双侧乳腺癌，左侧可能更严重，伴有腋窝淋巴结转移。不过需要排除其他可能性，但根据影像描述，恶性可能性很高。
Raw Chinese DeepSeek-R1 (Con.)	乳腺癌（BI-RADS 5 类）
Translated English findings (from a radiologist)	A mass is visible behind the left nipple, measuring approximately 28 × 12 mm. The lesion has an irregular shape and margin, with spiculated projections. Multiple punctate and ring-like calcifications are observed within the mass. Surrounding tissue shows traction, the nipple is markedly retracted, and the adjacent skin is thickened. In the outer quadrant of the right breast, a mass measuring approximately 23 × 19 mm is seen, also irregular in shape with lobulated margins. Scattered punctate calcifications are present in the right breast.

	Multiple enlarged lymph nodes are visible in both axillary regions, the largest (on the left) having a short-axis diameter of about 11 mm. No abnormalities are detected in either pectoralis major muscle. The skin, subcutaneous tissue, nipple, and areola of the right breast appear normal.
Translated English impression (from a radiologist)	A mass is present behind the left nipple, involving the left nipple and overlying skin. Enlarged lymph nodes are observed in the left axilla. Assessment: BI-RADS Category 5 — highly suggestive of malignancy; biopsy is recommended. A mass is seen at the 9 o'clock position in the outer quadrant of the right breast. Assessment: BI-RADS Category 4b — moderate suspicion for malignancy; biopsy is recommended. Visible lymph nodes are noted in the right axilla — follow-up is advised. Both breasts show scattered fibroglandular tissue: ACR Category b.
Translated English DeepSeek-R1 (Rea.)	Okay. Based on the mammography findings provided by the patient, I now need to give a diagnostic conclusion. First, I will carefully read and understand every detail described in the imaging report. The mass behind the left nipple measures 28×12 mm, with an irregular shape and margins, and spiculated projections. Inside the mass, there are multiple punctate and ring-like calcifications, with retraction of surrounding tissue, obvious nipple inversion, and skin thickening. These features are quite typical signs of malignancy — for example, the spiculations and calcification patterns may suggest a malignant tumor, especially invasive ductal carcinoma. Nipple inversion and skin thickening may indicate that the tumor is pulling the surrounding tissue or invading the skin. The mass in the outer quadrant of the right breast measures 23×19 mm, also with an irregular shape and lobulated margins, accompanied by scattered punctate calcifications. Although lobulated margins can be seen in benign tumors, the irregular shape and calcifications warrant suspicion of malignancy. However, since the skin, subcutaneous tissue, nipple, and areola of the right breast are all normal, the degree of malignancy may be lower than that of the left breast, though it cannot be ruled out. There is bilateral axillary lymph node enlargement, with the left side showing a short-axis diameter of 11 mm. Normally, the short axis of lymph nodes is less than 10 mm, so the enlargement may indicate metastasis. However, other causes, such as infection, should also be considered. Still, combined with the malignant features in the breasts, metastasis is more likely. In summary, both breasts show irregular masses with calcifications and enlarged axillary lymph nodes. The features of the left breast are particularly consistent with breast cancer. The right breast mass also requires consideration of malignancy, but needs further confirmation. Overall, the most likely diagnosis is bilateral breast cancer, with the left

	side being more severe and accompanied by axillary lymph node metastasis. Although other possibilities should be excluded, based on the imaging description, the likelihood of malignancy is very high.
Translated English DeepSeek-R1 (Con.)	Breast cancer (BI-RADS Category 5)

Note: Translation was performed using GPT-5 (accessed from its official website). The prompt for translation: *Please translate the following text into standard and professional English + {Raw Chinese text}*.

Table S5 | Radiologist's scores of the above breast case

Metric	Score
DeepSeek-R1 (Con.)	
Missed primary diagnoses (MPD)	0
Missed secondary diagnoses (MSD)	1 (Omitted involvement of the left nipple and overlying skin and enlargement of lymph nodes in the left axilla)
Primary misdiagnoses (PMisD)	0
Secondary misdiagnoses (SMisD)	0
Comprehensiveness	2, 1, 2
Explainability	2, 2, 2
Conciseness	5, 5, 5
Unbiasedness	3, 2, 2
DeepSeek-R1 (Rea.)	
Missed primary diagnoses (MPD)	0
Missed secondary diagnoses (MSD)	0
Primary misdiagnoses (PMisD)	0
Secondary misdiagnoses (SMisD)	0
Comprehensiveness	5, 4, 5
Explainability	5, 5, 4
Conciseness	2, 3, 2
Unbiasedness	5, 4, 5

Note: For the qualitative metrics, the individual ratings of all three evaluators are illustrated.

In this breast case, the output from DeepSeek-R1 (Con.) remained concise but exhibited notable limitations. It omitted key secondary findings (an MSD), including involvement of the left nipple and overlying skin, as well as lymph node enlargement in the left axilla, and its conclusions were overly deterministic yet insufficiently specific. In contrast, DeepSeek-R1 (Rea.) organized all relevant details and explicitly noted that “the short axis of the lymph nodes is less than 10 mm, so the enlargement may indicate metastasis,” providing a more interpretable and clinically credible explanation.

1 C. Human-in-the-loop evaluation



2

3 **Fig S1 | The custom graphical interface used for human-in-the-loop evaluation.** This interface

4 includes three main units: Raw findings (upper left), Model output (lower left), and the evaluation

5 unit (right). For each case, the raw findings and one random model output from DeepSeek-R1 (Rea.)

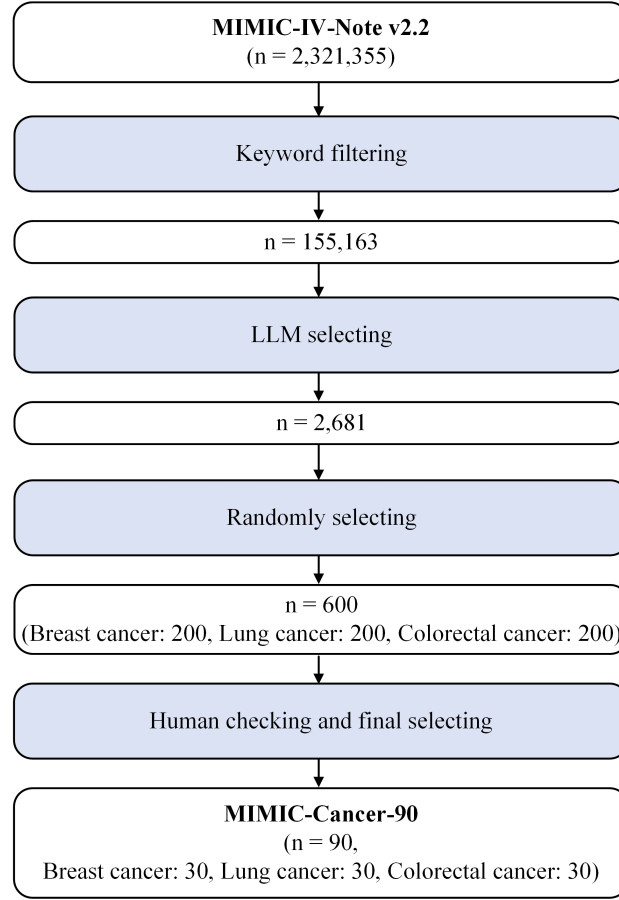
6 or DeepSeek-R1 (Con.) were illustrated to the readers. The readers were told to click the start button

7 (“*Start evaluating current case*”) and then evaluate one case. After finishing one case evaluation,

8 they were requested to click the finish button (“*Finish current case evaluation*”) to start the next

9 case evaluation. The evaluation time between the two clicks was recorded.

1 D. MIMIC-Cancer-90 dataset



2
3 **Fig S2 | Construction procedures of MIMIC-Cancer-90.**

4
5 As illustrated in **Fig S2**, starting with the *radiology report* ($n = 2,321,355$) provided by the
6 MIMIC-IV-Note v2.2 dataset, four main steps were conducted to acquire the MIMIC-Cancer-90
7 dataset:

8 (i) Step 1, keyword filtering. A keyword-based text mining approach was implemented to
9 identify clinical notes containing potential references to three common cancer types: breast cancer,
10 lung cancer, and colorectal cancer (**Table S6**). For each cancer type, a list of regular-expression
11 patterns was defined to capture relevant terminology, including disease names, histological
12 descriptions, imaging findings, and diagnostic markers. Through this step, 155,163 cases were left.

13 (ii) Step 2, LLM selection. A multi-step natural language processing (NLP) pipeline was
14 designed to process the radiology reports acquired in Step 1. The procedures comprised three major
15 stages:

16 Step 2.1 Extraction of findings and impression sections.

17 Each full-text radiology report was submitted to GPT-4.1 (Version: 2025-04-14) via calling its
18 official API. GPT-4.1 was instructed, using a custom prompt, to parse the report and separate it into
19 two standardized sections: Findings and Impression (**Table S7**). If either section was absent, the
20 output explicitly indicated “None”, and such reports were excluded. The extraction was performed
21 in a structured JSON format to ensure machine-readability.

Step 2.2 Selection of eligible reports.

The extracted content was subsequently evaluated by the same GPT model to determine its suitability for the study evaluation (**Table S8**). A report was considered eligible only if the Findings section was complete and unambiguous; the Impression section contained a suspected or confirmed malignant neoplasm; the Impression did not reference patient history, prior diagnoses, or other external cancer-related information; and the Impression could be logically inferred from the Findings alone. The model returned a binary decision (“YES” or “NO”), and only reports with a “YES” response were retained.

Step 2.3 Cancer type classification.

Eligible reports were classified into breast cancer, lung cancer, or colorectal cancer categories based on keyword pattern matching within the report text. Reports containing the keyword “breast” were labeled as Breast Cancer; those containing “lung” were labeled as Lung Cancer; and those containing any of “colon”, “rectal”, or “colorectal” were labeled as Colorectal Cancer. Reports that did not match any of these predefined lexical patterns were assigned the label Unknown.

With the above procedures, 2,681 cases were selected.

(iii) Step 3, random selection. Randomly select 600 cases from the above 2,681 ones, with 200 cases for breast cancer, 200 for lung cancer, and 200 for colorectal cancer.

(iv) Step 4, human checking and final selection. One senior radiologist checked the 600 cases from Step 3 again to filter out those ineligible cases that GPT-4.1 failed to rule out. Finally, randomly select 90 cases (30 for each of the three cancer types) to constitute the MIMIC-Cancer-90 dataset.

Table S6 | Keyword-based text mining

Cancer type	Keyword / Regular expression patterns
Breast Cancer	breast cancer; breast carcinoma; ductal carcinoma; invasive ductal carcinoma; lobular carcinoma; invasive lobular carcinoma; breast mass; spiculated mass; BI[\s-]?RADS\s*[45]; microcalcifications
Lung Cancer	lung cancer; pulmonary carcinoma; bronchogenic carcinoma; non[\s-]?small cell lung cancer; NSCLC; SCLC; lung mass; lung lesion; spiculated nodule; malignant lung nodule; mediastinal lymphadenopathy
Colorectal Cancer	colorectal cancer; colon cancer; rectal cancer; adenocarcinoma of (the)?(colon rectum); colonic mass; rectal mass; apple[\s-]?core lesion; annular constricting lesion; bowel wall thickening

Note: The patterns include literal keywords and regular expressions. Case-insensitive matching was applied during text mining.

1 **Table S7 | API calling details of GPT-4.1 for finding and impression extraction**

Item	Content
Calling parameter	Temperature = 0.2, top_p = 0.95
Prompt	Raw in Chinese: 你是放射学专家。请从以下医学报告中提取两个标准结构段落： 1. Findings（检查所见） 2. Impression（影像印象） 如果某一部分缺失，请明确说明为“无”。 请严格以 JSON 格式返回，格式如下： <pre>{ "findings": "...", "impression": "..." }</pre> 放射报告全文如下： ----- {full_text} ----- Translated in English: You are a radiology expert. Please extract two standardized structured sections from the following medical report: 1. Findings 2. Impression If any section is missing, clearly state it as “None”. Return your response strictly in JSON format, as shown below: <pre>{ "findings": "...", "impression": "..." }</pre> Full radiology report: ----- {full_text} -----

2 Note: The raw prompts were input into GPT-4.1 in Chinese.

1 **Table S8 | API calling details of GPT-4.1 for eligible report selection**

Item	Content
Calling parameter	Temperature = 0.2, top_p = 0.95
Prompt	Raw in Chinese: 你是一个放射学 AI 训练专家，正在评估以下放射报告内容是否适合用于大语言模型进行“根据 Findings 预测 Impression”评估任务。 以下是放射报告的内容： Findings: {findings} Impression: {impression} 判断标准如下，4 项要求必须同时严格满足： 1. Findings 是完整、清晰的； 2. Impression 必须怀疑或诊断癌症（恶性肿瘤）； 3. Impression 不能提及病史、临床诊断、前期检查或其他外部已知癌症信息； 4. Impression 应该是只能根据 Findings 推理出来的结论。 请只回答 YES 或 NO。 Translated in English: You are a radiology AI training expert evaluating whether the following radiology report is suitable for use in a large language model task designed to predict the Impression based on the Findings. The report content is as follows: Findings: {findings} Impression: {impression} Evaluation criteria — all four of the following conditions must be strictly met: 1. The Findings section is complete and clear. 2. The Impression must suspect or diagnose cancer (malignancy). 3. The Impression must not mention medical history, clinical diagnosis, prior exams, or other externally known cancer information. 4. The Impression should be a conclusion that can be inferred solely from the Findings. Please answer only with YES or NO.

2 Note: The raw prompts were input into GPT-4.1 in Chinese.

E. Prompt and API calling parameters of DeepSeek-R1 and comparator models

Table S9 | Prompt and API calling parameters for the Chinese reports

Item	Content
API calling parameter	Temperature = 0, top_p = 0.95
Prompt	Raw in Chinese: 根据以下{乳腺/肺/结直肠}区域的{CT/乳腺 X 线/磁共振}影像所见生成诊断印象：{影像所见具体内容}。只生成诊断印象，忽略处理建议。{Findings} Translated in English: Please generate a diagnostic impression based on the following {CT/MG/MRI} imaging findings of the {breast/lung/colorectal} region: {Image findings}. Only generate the impression and ignore the management recommendations.

Note: The raw prompts were input into models (DeepSeek-R1, DeepSeek-V3_0324, and GPT-4.5) in Chinese. The low *Temperature* was chosen to generate high-confidence responses with limited randomness, which is particularly important for tasks requiring factual consistency.

Table S10 | Prompt and API calling parameters for the English reports

Item	Content
API calling parameter	Temperature = 0, top_p = 0.95
Prompt	Raw in English: Please generate a diagnostic impression based on the following {CT/MG/MRI} imaging findings of the {breast/lung/colorectal} region: {Image findings}. Only generate the impression and ignore the management recommendations.

Note: The raw prompts were input into models (DeepSeek-R1, DeepSeek-V3_0324, and GPT-4.5) in English. The low *Temperature* was chosen to generate high-confidence responses with limited randomness, which is particularly important for tasks requiring factual consistency.