Appendix A Supplementary Figures

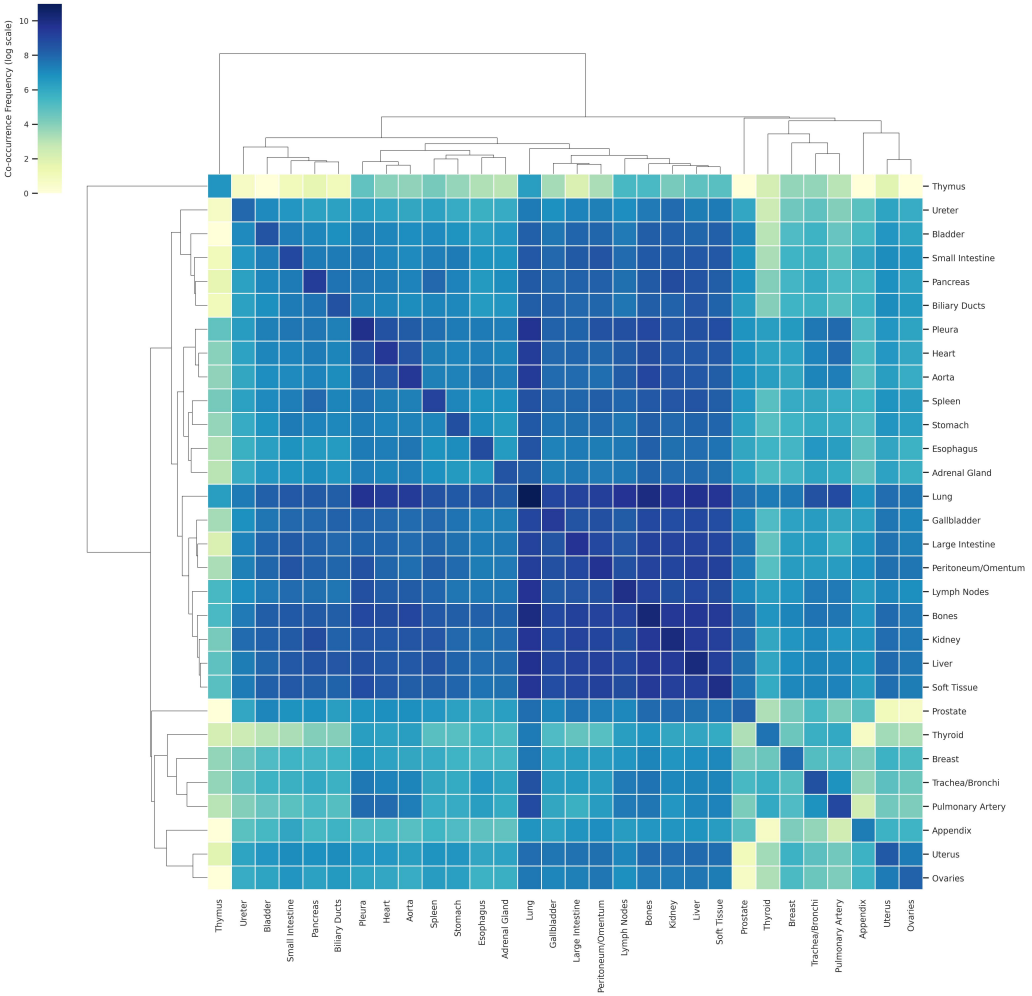


Fig. A1: **Organ-level abnormality co-occurrence matrix in the CTRgDB dataset.** The matrix shows multi-organ abnormality co-occurrence across 30 major organs in CTRgDB. In the abdomen, abnormalities of the liver, kidneys, peritoneum and lymph nodes frequently co-occur, consistent with common metastatic spread patterns. In the thorax, lung abnormalities co-occur broadly with abnormalities in multiple other organs. Together, these patterns show that CTRgDB captures complex multi-organ abnormality involvement.

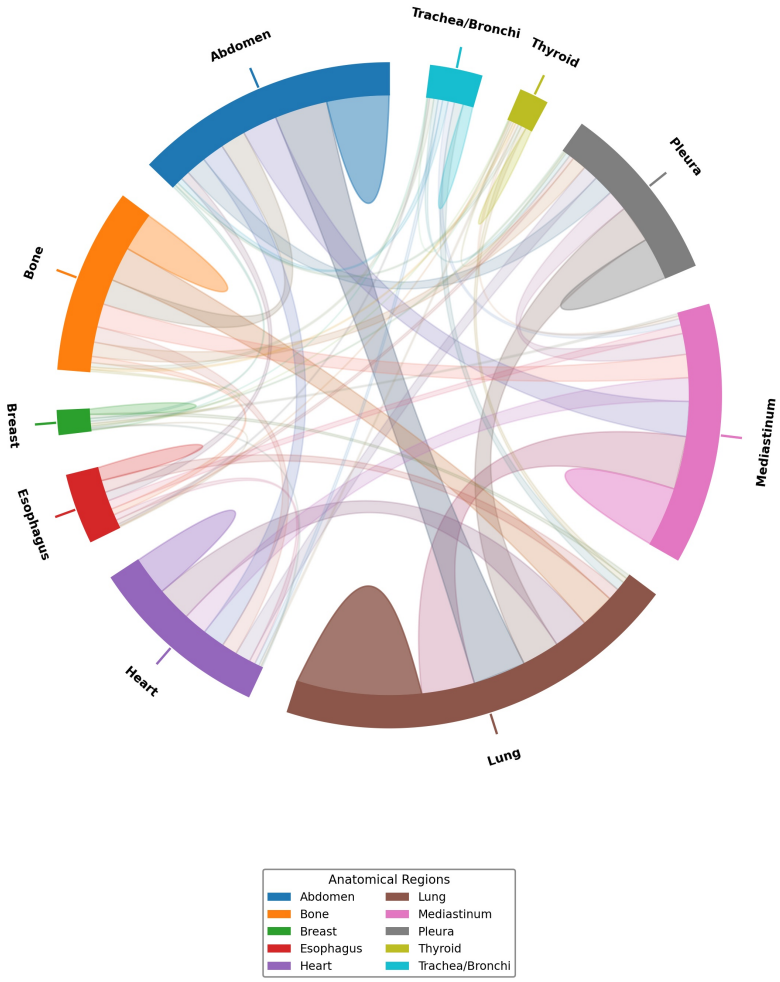


Fig. A2: **Region-level abnormality co-occurrence chord diagram for chest CTs in CTRgDB.** The diagram shows the frequency of multi-region abnormality involvement across 10 major thoracic regions in CTRgDB. Abnormalities in the mediastinum frequently co-occur with those in the lungs, consistent with the common spread of neoplastic and inflammatory processes to the mediastinal region. CTRgDB captures complex multi-region abnormality involvement in chest CT examinations.

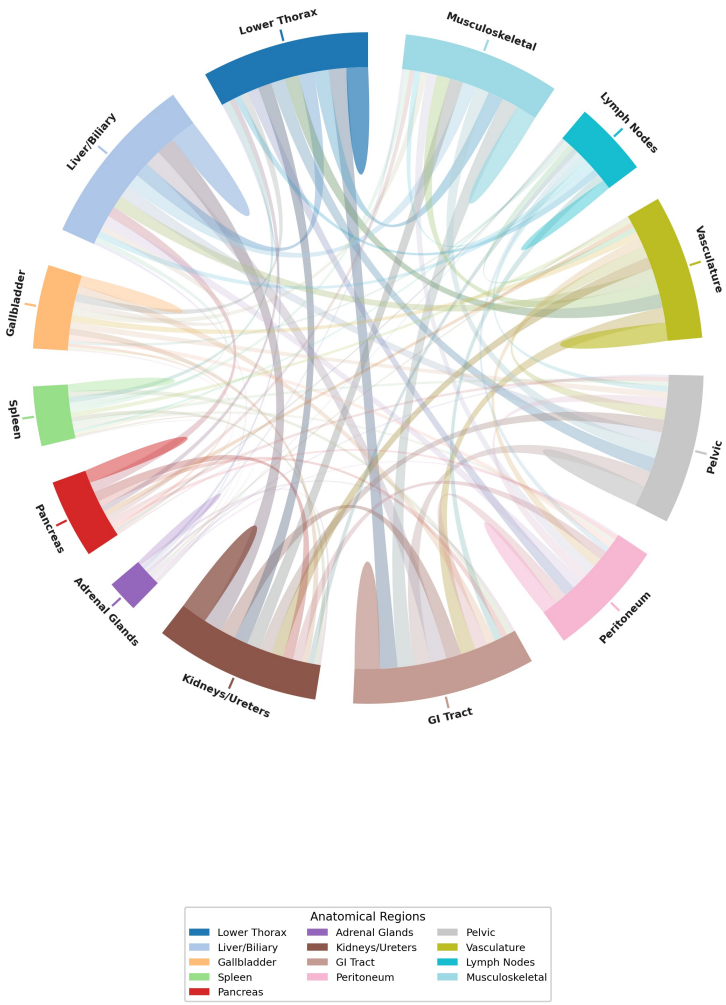


Fig. A3: **Region-level abnormality co-occurrence chord diagram for abdominal CTs in CTRgDB.** The diagram shows the frequency of multi-region abnormality involvement across 13 major abdominal regions in CTRgDB. CTRgDB captures complex multi-region abnormality involvement in abdominal CT examinations.

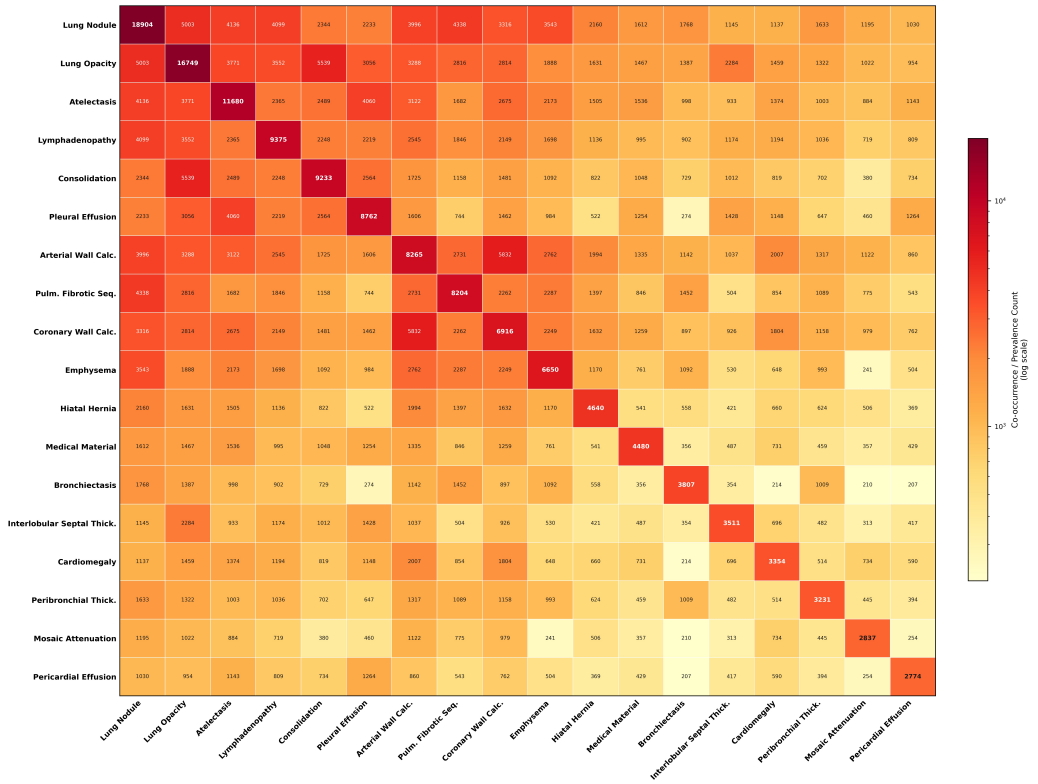


Fig. A4: **Disease-level co-occurrence matrix for chest CTs in CTRgDB.** The matrix shows co-occurrence patterns among 18 major diseases defined in CT-Rate, highlighting frequent multi-disease involvement in individual chest CT examinations. These results demonstrate that CTRgDB captures complex disease-level abnormalities in chest CTs.

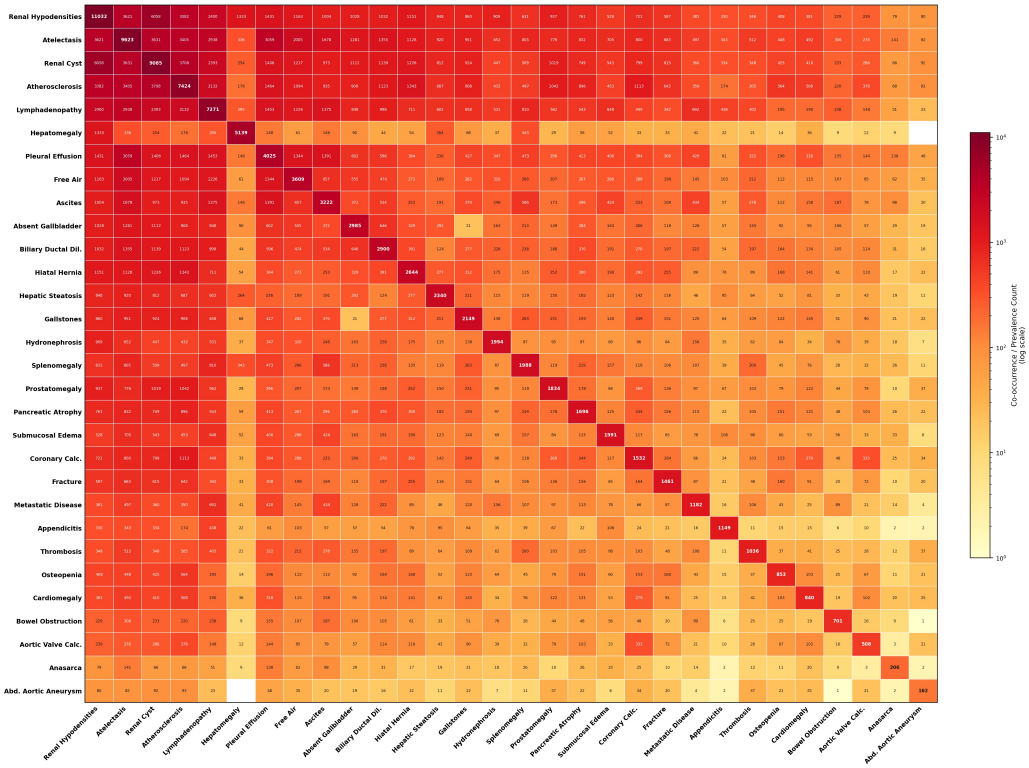


Fig. A5: **Disease-level co-occurrence matrix for abdominal CTs in CTRgDB.** The matrix illustrates co-occurrence patterns among 30 major abdominal diseases defined with reference to Merlin, revealing frequent coexisting disease entities within individual abdominal CT examinations. This distribution supports the ability of CTRgDB to represent complex disease-level involvement in abdominal CTs.

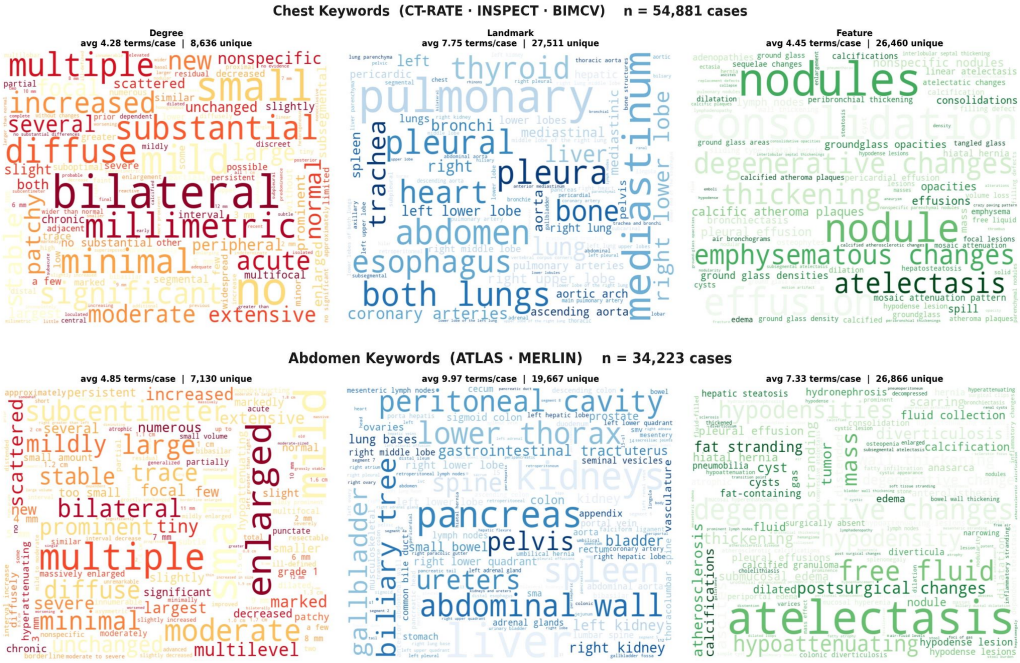
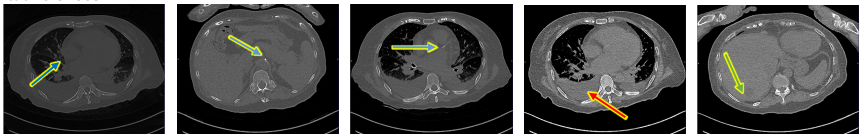


Fig. A6: **Word clouds in CTRgDB reports.** The word clouds highlight the rich diagnostic vocabulary used in chest and abdominal CT reports in CTRgDB. Reports frequently describe abnormalities in terms of degree, landmark and feature, enabling precise and clinically informative characterization. These patterns indicate that CTRgDB provides detailed diagnostic descriptions at the finding level.

Representative Slices



Reference report

The thoracic cage is symmetric. Multiple small pulmonary nodules are seen in both lungs, the largest located in the anterior basal segment of the right lower lobe, measuring approximately 8 × 4 mm, with well-defined margins. Focally increased lucency is noted in the dorsal segment of the left lower lobe. Scattered linear fibrotic opacities are present in both lungs. No enlarged mediastinal or hilar lymph nodes are identified. The trachea, bronchi, and their major branches are patent. Cardiac size and contour are normal. Calcified plaques are seen in the aorta and coronary arteries. Within the scanned range, a nodular hyperdense lesion is noted in the gallbladder, measuring approximately 21 × 17 mm.

Generated reports

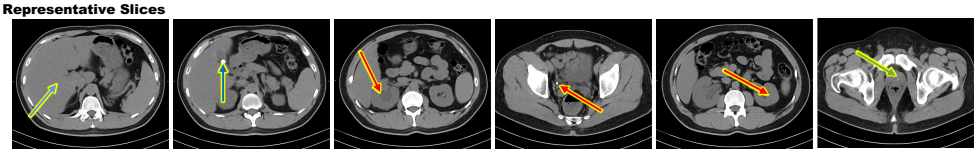
AI-written abdomen: in the upper abdominal organs included in the study area; a stone of approximately 15 mm in diameter was observed in the gallbladder lumen. bone: normal. breast: normal. esophagus: normal. heart: calcific atherosclerotic plaques are present in lad and rca coronary artery traces. lung: when examined in the lung parenchyma window; mild emphysematous changes were observed in both lungs. there are millimetric nonspecific nodules in both lungs. mild linear atelectasis was observed in the right lung lower lobe superior segment, left lung lingular segment and bilateral lower lobes. mediastinum: normal. pleura: normal. thyroid: normal. trachea and bronchie: normal.
Readability: 9 There are no spelling or grammatical errors, the descriptions are in the correct order, the logic is clear and fluent.
Completeness: 9 Basically covers all four lesions.
Accuracy: 8 Lacks detailed description of the size and location of the pulmonary nodules.

Human-written The thoracic cage is symmetric. Diffuse increased lucency is seen in both lungs, consistent with mild emphysematous change. Small pulmonary nodules are present bilaterally, the largest located in the right lower lobe, measuring approximately 5 mm in diameter. A linear fibrotic opacity is seen in the left lower lobe. No enlarged mediastinal or hilar lymph nodes are identified. The trachea, bronchi, and their major branches are patent. Cardiac size and contour are normal. No obvious pleural abnormality is seen. Within the scanned range, the liver and spleen show no abnormality. A nodular hyperdense lesion is noted within the gallbladder. No other significant abnormality is identified.
Readability: 9 There are no spelling or grammatical errors, the descriptions are in the correct order, the logic is clear and fluent.
Completeness: 8 Lacks diagnosis of arterial calcification.
Accuracy: 7 Lacks description of gallstone size; deviation in the size of the pulmonary nodule.

AI-assisted human-written The thoracic cage is symmetric. Increased lucency is seen in both lungs, consistent with mild emphysematous change. Small pulmonary nodules are present in both lungs, the largest located in the right lower lobe, measuring approximately 7 mm in diameter. Scattered linear fibrotic opacities are seen in both lungs. No enlarged mediastinal or hilar lymph nodes are identified. The trachea, bronchi, and their major branches are patent. Calcifications are noted in the walls of the aorta and coronary arteries. No obvious pleural abnormality is seen. Within the scanned range, a hyperdense focus is noted in the gallbladder. No other significant abnormality is identified.
Readability: 9 There are no spelling or grammatical errors, the descriptions are in the correct order, the logic is clear and fluent.
Completeness: 9 Basically covers all four lesions.
Accuracy: 8 Lacks description of gallstone size.

Fig. A7: **Case study of human-AI collaboration in chest CT report generation.** Astra identified all abnormalities but provided insufficient characterization of a pulmonary nodule. The radiologist missed arteriosclerosis. Human–AI collaboration accurately detected all abnormalities with precise characterization.

Representative Slices



Reference report

The right kidney is enlarged in volume. The left kidney is normal in size and contour. Multiple nodular hyperdense foci are seen in the left kidney, the largest measuring approximately 5 mm in diameter, consistent with renal calculi. No obvious abnormal attenuation is identified in the right renal parenchyma. A nodular hyperdense focus measuring approximately 6 × 5 mm is seen in the right proximal ureter, with a CT attenuation value of approximately 791 HU, consistent with a ureteral calculus. Dilatation of the right ureter proximal to the stone, as well as the right renal pelvis and calyces, is noted, consistent with obstructive hydroureteronephrosis. Scattered inflammatory stranding is seen in the right perinephric and periureteral fat. The left perinephric space is clear. No obvious dilatation of the left renal pelvis, calyces, or ureter is seen, and no hyperdense calculus is identified in the left ureter. The adrenal glands are normal in size and morphology bilaterally, with homogeneous attenuation. The urinary bladder is adequately distended, with no obvious abnormal intravesical density and no bladder wall thickening. A small nodular hyperdense focus is seen in the prostate, likely representing calcification. No obvious abnormality is identified in the seminal vesicles bilaterally. No enlarged lymph nodes are seen in the retroperitoneum or pelvis. No ascites or pelvic free fluid is identified. Diffuse decreased attenuation of the liver is noted, consistent with hepatic steatosis. A nodular hyperdense focus is seen in the gallbladder, measuring approximately 11 mm in diameter, likely representing a gallstone.

Generated reports

AI-written lower thorax: normal. liver and biliary tree: mild diffuse hypoattenuation of the liver parenchyma, consistent with steatosis. gallbladder: cholelithiasis is noted within the gallbladder neck. spleen: normal. pancreas: normal. adrenal glands: normal. kidneys and ureters: severe hydronephrosis of the right kidney with mild hydroureter. multiple obstructing stones are seen in the midportion of the ureter measuring up to 0.5 cm in diameter. mild uroepithelial enhancement of the left kidney. multiple bilateral nonobstructing renal stones are seen within both kidneys, measuring up to 1 mm in size on the right and 3 mm in size on the left. gastrointestinal tract: normal. peritoneum: normal. pelvic: normal. vasculature: patent lymph nodes: normal. musculoskeletal: normal.
Readability: 9 There are no spelling or grammatical errors, the descriptions are in the correct order, the logic is clear and fluent.
Completeness: 8 Lacks diagnosis of calcification within the prostate.
Accuracy: 8 Lacks detailed description of the gallbladder stone.

Human-written The liver is normal in size and contour. The gallbladder is not distended, with no obvious wall thickening. A nodular hyperdense focus is seen within the gallbladder lumen, measuring approximately 1.5 cm in diameter, consistent with a gallstone. No dilatation of the intrahepatic or extrahepatic bile ducts is identified. The spleen is normal in size and morphology, with homogeneous parenchymal attenuation. The pancreas is normal in size and morphology, with no focal abnormal enhancement identified in the pancreatic parenchyma. The peripancreatic fat planes are preserved. The left kidney is normal in size, contour, and attenuation. A punctate hyperdense focus is seen in the left renal pelvis, consistent with a small calculus. A punctate hyperdense focus measuring approximately 0.7 cm in diameter is seen in the distal abdominal segment of the right ureter, consistent with a ureteral calculus. Dilatation of the right ureter proximal to the stone and of the right renal pelvis is noted, consistent with obstructive hydronephrosis. The right kidney is enlarged in volume, and thickening of the right perirenal fascia is present. The left perinephric space is clear. No left ureteral dilatation is seen. No obvious space-occupying lesion is identified in the gastrointestinal tract. The urinary bladder is well distended, with no wall thickening or abnormal intraluminal density. The prostate is not enlarged, with punctate calcification noted within it. The seminal vesicles are normal in size, morphology, and attenuation bilaterally, with no abnormal enhancement. The bilateral vesicoseminal angle structures are preserved. No enlarged lymph nodes are identified in the abdominal cavity, retroperitoneum, or pelvis. No ascites or pelvic free fluid is seen. No definite osseous destruction is identified within the scanned field.
Readability: 8 The language lacks conciseness.
Completeness: 8 Lacks diagnosis of fatty liver.
Accuracy: 8 Insufficient detail in description of renal abnormality, slightly redundant.

AI-assisted human-written The liver is normal in size and contour. Diffuse decreased attenuation of the hepatic parenchyma is noted, lower than that of the spleen on the same level, consistent with hepatic steatosis. The gallbladder is not distended, and no obvious wall thickening is seen. A punctate hyperdense focus is noted at the gallbladder neck, consistent with a small gallstone. No dilatation of the intrahepatic or extrahepatic bile ducts is identified. The spleen is normal in size and morphology, with homogeneous parenchymal attenuation and no focal abnormal enhancement. The pancreas is normal in size and morphology, with no focal abnormal enhancement identified in the pancreatic parenchyma. The peripancreatic fat planes are preserved. Multiple punctate hyperdense foci are seen in the left renal pelvis, consistent with calculi. A nodular hyperdense focus is seen in the distal right ureter, consistent with a ureteral calculus. Dilatation of the right ureter proximal to the stone and of the right renal pelvis is noted. Mild thickening of the right perirenal fascia is present. The left kidney is normal in size, contour, and attenuation, and the left perinephric space is clear. No dilatation of the left ureter is seen. No obvious space-occupying lesion is identified in the gastrointestinal tract. The urinary bladder is well distended, with no wall thickening or abnormal intraluminal density. The prostate is not enlarged, with punctate calcifications noted within it. The seminal vesicles are normal in size, morphology, and attenuation bilaterally, with no abnormal enhancement. The bilateral vesicoseminal angle structures are preserved. No enlarged lymph nodes are identified in the abdominal cavity, retroperitoneum, or pelvis. No ascites or pelvic free fluid is seen. No definite osseous destruction is identified within the scanned field.
Readability: 8 The language lacks conciseness.
Completeness: 10 Accurately diagnoses all major and minor lesions.
Accuracy: 9 Descriptions of all abnormalities are basically accurate and complete.

Fig. A8: **Case study of human-AI collaboration in abdomen CT report generation.** Astra missed prostatic calcification, while the radiologist missed hepatic steatosis. Human–AI collaboration accurately identified all abnormalities with precise characterization.

Appendix B Supplementary Tables

Dataset	Model	micro Precision [95% CI]	micro Recall [95% CI]	micro F1 [95% CI]
CT-Rate	Astra-Base-Ori	0.5618 [0.5431, 0.5803]	0.3694 [0.3538, 0.3838]	0.4458 [0.4302, 0.4596]
	Astra-Base	0.5613 [0.5445, 0.5780]	0.4301* [0.4174, 0.4428]	0.4870* [0.4748, 0.4994]
	Astra	0.5789** [0.5632, 0.5948]	0.5242** [0.5103, 0.5381]	0.5502** [0.5380, 0.5628]
BIMCV	Astra-Base-Ori	0.3112 [0.2927, 0.3301]	0.2739 [0.2568, 0.2917]	0.2914 [0.2750, 0.3078]
	Astra-Base	0.3398* [0.3186, 0.3601]	0.2578 [0.2402, 0.2752]	0.2932 [0.2759, 0.3103]
	Astra	0.3572** [0.3373, 0.3774]	0.3405** [0.3222, 0.3578]	0.3486** [0.3305, 0.3652]
Inspect	Astra-Base-Ori	0.4306 [0.4161, 0.4449]	0.4524 [0.4385, 0.4667]	0.4412 [0.4288, 0.4537]
	Astra-Base	0.4700* [0.4542, 0.4851]	0.4321 [0.4181, 0.4465]	0.4503 [0.4372, 0.4628]
	Astra	0.4720 [0.4576, 0.4867]	0.4759** [0.4625, 0.4892]	0.4739** [0.4616, 0.4865]
Merlin	Astra-Base-Ori	0.4727 [0.4650, 0.4808]	0.4313 [0.4237, 0.4392]	0.4510 [0.4443, 0.4584]
	Astra-Base	0.5131* [0.5044, 0.5218]	0.4220 [0.4146, 0.4299]	0.4631* [0.4560, 0.4701]
	Astra	0.5300** [0.5215, 0.5387]	0.4619** [0.4539, 0.4699]	0.4936** [0.4863, 0.5008]

Table B1: **Ablation study: RadBERT classification performance (micro Precision, Recall, F1) on in-distribution datasets** (95% confidence intervals estimated via 2,000 bootstrap iterations). *Astra*: RL tuned model; *Astra-Base*: sft model with preprocessed reports; *Astra-Base-Ori*: sft model with original reports. * represents a significant improvement between Astra-Base and Astra-Base-Ori with $P < 0.05$, **represents a significant improvement between Astra and Astra-Base with $P < 0.05$.

Dataset	Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
CT-Rate	Astra-Base-Ori	0.2776 [0.2720, 0.2830]	0.2466 [0.2371, 0.2565]	0.3316 [0.3204, 0.3423]	0.2531 [0.2428, 0.2635]	0.2653 [0.2503, 0.2795]	0.2742 [0.2655, 0.2827]
	Astra-Base	0.3343* [0.3275, 0.3413]	0.3079* [0.2971, 0.3176]	0.4252* [0.4145, 0.4365]	0.3244* [0.3124, 0.3358]	0.3755* [0.3602, 0.3900]	0.3582* [0.3489, 0.3673]
	Astra	0.3510** [0.3436, 0.3580]	0.3844** [0.3739, 0.3946]	0.5140** [0.5033, 0.5252]	0.4159** [0.4034, 0.4273]	0.4457** [0.4306, 0.4601]	0.4400** [0.4300, 0.4493]
BIMCV	Astra-Base-Ori	0.2226 [0.2179, 0.2274]	0.2962 [0.2856, 0.3072]	0.2619 [0.2501, 0.2734]	0.1911 [0.1809, 0.2016]	0.2414 [0.2270, 0.2562]	0.2477 [0.2387, 0.2565]
	Astra-Base	0.2262 [0.2204, 0.2323]	0.2612 [0.2488, 0.2734]	0.2575 [0.2443, 0.2698]	0.1833 [0.1719, 0.1945]	0.2590 [0.2425, 0.2757]	0.2402 [0.2299, 0.2503]
	Astra	0.2624** [0.2567, 0.2684]	0.4133** [0.4023, 0.4240]	0.3783** [0.3663, 0.3897]	0.2755** [0.2649, 0.2863]	0.3448** [0.3294, 0.3595]	0.3530** [0.3441, 0.3620]
Inspect	Astra-Base-Ori	0.2904 [0.2855, 0.2953]	0.2815 [0.2733, 0.2895]	0.3598 [0.3512, 0.3684]	0.3236 [0.3145, 0.3327]	0.3243 [0.3151, 0.3332]	0.3223 [0.3158, 0.3288]
	Astra-Base	0.3047* [0.2994, 0.3098]	0.2962* [0.2880, 0.3045]	0.3847* [0.3758, 0.3935]	0.3432* [0.3332, 0.3531]	0.3502* [0.3400, 0.3598]	0.3436* [0.3365, 0.3505]
	Astra	0.3255** [0.3202, 0.3310]	0.3564** [0.3481, 0.3647]	0.4417** [0.4327, 0.4500]	0.3780** [0.3682, 0.3880]	0.3958** [0.3867, 0.4055]	0.3930** [0.3854, 0.4000]
Merlin	Astra-Base-Ori	0.3178 [0.3152, 0.3205]	0.3000 [0.2951, 0.3048]	0.3187 [0.3139, 0.3233]	0.3115 [0.3072, 0.3156]	0.3902 [0.3838, 0.3964]	0.3301 [0.3261, 0.3337]
	Astra-Base	0.3442* [0.3415, 0.3468]	0.3085 [0.3034, 0.3136]	0.3402* [0.3356, 0.3448]	0.3292* [0.3247, 0.3334]	0.4262* [0.4200, 0.4325]	0.3510* [0.3472, 0.3548]
	Astra	0.3544** [0.3520, 0.3569]	0.3692** [0.3644, 0.3742]	0.4045** [0.4000, 0.4090]	0.3706** [0.3664, 0.3747]	0.4639** [0.4578, 0.4699]	0.4021** [0.3983, 0.4058]

Table B2: **Ablation study: fine-grained caption metrics (Rate-Score, FORTE) on in-distribution datasets** (95% confidence intervals estimated via 2,000 bootstrap iterations). F. means FORTE metric. *Astra*: RL tuned model; *Astra-Base*: sft model with preprocessed reports; *Astra-Base-Ori*: sft model with original reports. * represents a significant improvement between Astra-Base and Astra-Base-Ori with $P < 0.05$, **represents a significant improvement between Astra and Astra-Base with $P < 0.05$.

Dataset	Model	micro P. [95% CI]	micro R. [95% CI]	micro F1 [95% CI]
CTRG-Chest	Astra-Base-Ori	0.5556 [0.4342, 0.6812]	0.1228 [0.0877, 0.1617]	0.2011 [0.1479, 0.2571]
	Astra-Base	0.6901* [0.5873, 0.7888]	0.1744* [0.1329, 0.2188]	0.2784* [0.2196, 0.3386]
	Astra	0.6618 [0.5781, 0.7379]	0.3203** [0.2684, 0.3742]	0.4317** [0.3718, 0.4907]
AMOS-MM	Astra-Base-Ori	0.3662 [0.2914, 0.4476]	0.1368 [0.1043, 0.1731]	0.1992 [0.1565, 0.2461]
	Astra-Base	0.4151* [0.3542, 0.4762]	0.2895* [0.2400, 0.3412]	0.3411* [0.2885, 0.3921]
	Astra	0.4051** [0.3590, 0.4534]	0.4211** [0.3658, 0.4744]	0.4129** [0.3694, 0.4569]
Inhouse-Chest-1	Astra-Base-Ori	0.4162 [0.3882, 0.4442]	0.3288 [0.3007, 0.3578]	0.3674 [0.3422, 0.3913]
	Astra-Base	0.5158* [0.4818, 0.5498]	0.3408* [0.3155, 0.3675]	0.4104* [0.3851, 0.4365]
	Astra	0.5675** [0.5361, 0.6002]	0.4355** [0.4094, 0.4615]	0.4928** [0.4679, 0.5176]
Inhouse-Abdomen-1	Astra-Base-Ori	0.1789 [0.1538, 0.2050]	0.2092 [0.1757, 0.2439]	0.1928 [0.1653, 0.2198]
	Astra-Base	0.1988* [0.1765, 0.2202]	0.3825* [0.3374, 0.4261]	0.2616* [0.2338, 0.2879]
	Astra	0.2453** [0.2263, 0.2654]	0.5239** [0.4840, 0.5626]	0.3342** [0.3109, 0.3575]
Inhouse-Abdomen-2	Astra-Base-Ori	0.3516 [0.3044, 0.4022]	0.3147 [0.2657, 0.3643]	0.3321 [0.2853, 0.3789]
	Astra-Base	0.2818 [0.2528, 0.3102]	0.5268* [0.4760, 0.5743]	0.3672* [0.3321, 0.4000]
	Astra	0.3659** [0.3437, 0.3885]	0.7599** [0.7204, 0.7995]	0.4939** [0.4706, 0.5167]
Inhouse-Abdomen-3	Astra-Base-Ori	0.3521 [0.3191, 0.3884]	0.3055 [0.2752, 0.3358]	0.3271 [0.2989, 0.3571]
	Astra-Base	0.3054 [0.2802, 0.3318]	0.3979* [0.3617, 0.4335]	0.3456* [0.3180, 0.3729]
	Astra	0.3811** [0.3576, 0.4041]	0.6046** [0.5736, 0.6377]	0.4675** [0.4443, 0.4898]

Table B3: **Ablation study: RadBERT classification performance (micro Precision, Recall, F1) on external validation datasets** (95% confidence intervals estimated via 2,000 bootstrap iterations). P. means precision. R. means recall. *Astra*: RL tuned model; *Astra-Base*: SFT model with preprocessed reports; *Astra-Base-Ori*: SFT model with original reports. * represents a significant improvement between Astra-Base and Astra-Base-Ori with $P < 0.05$; ** represents a significant improvement between Astra and Astra-Base with $P < 0.05$.

Dataset	Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
CTRG-Chest	Astra-Base-Ori	0.1991 [0.1922, 0.2065]	0.1273 [0.1082, 0.1473]	0.2160 [0.1931, 0.2394]	0.0769 [0.0577, 0.0965]	0.0685 [0.0438, 0.0936]	0.1222 [0.1061, 0.1389]
	Astra-Base	0.1872 [0.1791, 0.1956]	0.1169 [0.0933, 0.1407]	0.1614 [0.1320, 0.1906]	0.0608 [0.0414, 0.0803]	0.0939* [0.0586, 0.1300]	0.1083 [0.0879, 0.1294]
	Astra	0.2434** [0.2348, 0.2525]	0.2550** [0.2324, 0.2783]	0.3918** [0.3690, 0.4131]	0.1408** [0.1195, 0.1634]	0.1346** [0.1063, 0.1649]	0.2306** [0.2151, 0.2467]
AMOS-MM	Astra-Base-Ori	0.2025 [0.1964, 0.2087]	0.1220 [0.1030, 0.1406]	0.1507 [0.1331, 0.1684]	0.1196 [0.1030, 0.1375]	0.0516 [0.0326, 0.0729]	0.1110 [0.0979, 0.1238]
	Astra-Base	0.2389* [0.2335, 0.2443]	0.2480* [0.2266, 0.2704]	0.2602* [0.2410, 0.2786]	0.1710* [0.1547, 0.1883]	0.0913* [0.0730, 0.1107]	0.1926* [0.1798, 0.2060]
	Astra	0.2507** [0.2449, 0.2561]	0.2880** [0.2684, 0.3070]	0.3084** [0.2904, 0.3258]	0.1973** [0.1807, 0.2138]	0.1036** [0.0826, 0.1258]	0.2243** [0.2117, 0.2370]
Inhouse-Chest-1	Astra-Base-Ori	0.2484 [0.2410, 0.2564]	0.2143 [0.1992, 0.2295]	0.3545 [0.3338, 0.3752]	0.1938 [0.1786, 0.2089]	0.1888 [0.1611, 0.2162]	0.2379 [0.2233, 0.2521]
	Astra-Base	0.2533* [0.2459, 0.2609]	0.2029 [0.1869, 0.2191]	0.3168 [0.2951, 0.3373]	0.2136* [0.1957, 0.2308]	0.2704* [0.2365, 0.3039]	0.2509* [0.2359, 0.2663]
	Astra	0.2874** [0.2792, 0.2957]	0.2745** [0.2605, 0.2890]	0.4805** [0.4647, 0.4957]	0.2641 [0.2489, 0.2788]	0.3227** [0.2888, 0.3562]	0.3355** [0.3231, 0.3473]
Inhouse-Abd.-1	Astra-Base-Ori	0.1929 [0.1871, 0.1991]	0.1588 [0.1408, 0.1762]	0.1796 [0.1620, 0.1990]	0.1160 [0.0993, 0.1325]	0.0396 [0.0252, 0.0557]	0.1235 [0.1121, 0.1353]
	Astra-Base	0.2343* [0.2290, 0.2401]	0.2420* [0.2240, 0.2602]	0.2609* [0.2427, 0.2799]	0.1639* [0.1488, 0.1798]	0.0534* [0.0382, 0.0697]	0.1800* [0.1696, 0.1906]
	Astra	0.2509** [0.2452, 0.2564]	0.3029** [0.2883, 0.3185]	0.3500** [0.3345, 0.3664]	0.2169** [0.2020, 0.2308]	0.0631** [0.0471, 0.0807]	0.2332** [0.2243, 0.2422]
Inhouse-Abd.-2	Astra-Base-Ori	0.2437 [0.2339, 0.2540]	0.1563 [0.1329, 0.1780]	0.2876 [0.2634, 0.3114]	0.2172 [0.1911, 0.2418]	0.1894 [0.1480, 0.2309]	0.2126 [0.1909, 0.2341]
	Astra-Base	0.2765* [0.2700, 0.2832]	0.3309* [0.3099, 0.3513]	0.3890* [0.3706, 0.4085]	0.2979* [0.2789, 0.3176]	0.2850* [0.2582, 0.3136]	0.3257* [0.3126, 0.3389]
	Astra	0.3157** [0.3097, 0.3216]	0.4701** [0.4538, 0.4869]	0.5270** [0.5125, 0.5415]	0.4234** [0.4078, 0.4403]	0.3871** [0.3562, 0.4175]	0.4519** [0.4426, 0.4613]
Inhouse-Abd.-3	Astra-Base-Ori	0.2364 [0.2298, 0.2428]	0.1755 [0.1559, 0.1949]	0.2263 [0.2093, 0.2431]	0.1939 [0.1766, 0.2099]	0.1883 [0.1585, 0.2176]	0.1960 [0.1814, 0.2101]
	Astra-Base	0.2843* [0.2769, 0.2915]	0.2801* [0.2647, 0.2956]	0.3107* [0.2946, 0.3250]	0.2305* [0.2157, 0.2446]	0.2133* [0.1861, 0.2381]	0.2587* [0.2468, 0.2698]
	Astra	0.3143** [0.3077, 0.3216]	0.4053** [0.3927, 0.4182]	0.4512** [0.4376, 0.4650]	0.3215** [0.3069, 0.3353]	0.2640** [0.2352, 0.2916]	0.3605** [0.3506, 0.3704]

Table B4: **Ablation study: fine-grained caption metrics (Rate-Score, FORTE) on external validation datasets** (95% confidence intervals estimated via 2,000 bootstrap iterations). F. means FORTE metric. *Astra*: RL tuned model; *Astra-Base*: sft model with preprocessed reports; *Astra-Base-Ori*: sft model with original reports. * represents a significant improvement between Astra-Base and Astra-Base-Ori with $P < 0.05$, **represents a significant improvement between Astra and Astra-Base with $P < 0.05$.

Metric	Reward Function					
	Base	**NLG**	**RadBERT**	**Rate-S.**	**FORTE-O.**	**FORTE**
BLEU-1	0.4058 [0.3936, 0.4179]	0.4993 [0.4894, 0.5089]	0.4559 [0.4462, 0.4646]	0.3314 [0.3231, 0.3392]	0.4036 [0.3913, 0.4154]	0.4923 [0.4838, 0.5002]
BLEU-2	0.3140 [0.3041, 0.3239]	0.3949 [0.3868, 0.4025]	0.3450 [0.3373, 0.3522]	0.2505 [0.2444, 0.2564]	0.3177 [0.3076, 0.3276]	0.3793 [0.3722, 0.3860]
BLEU-3	0.2491 [0.2406, 0.2575]	0.3178 [0.3106, 0.3245]	0.2691 [0.2626, 0.2755]	0.1944 [0.1896, 0.1990]	0.2558 [0.2472, 0.2642]	0.2965 [0.2900, 0.3028]
BLEU-4	0.2027 [0.1953, 0.2103]	0.2604 [0.2537, 0.2668]	0.2150 [0.2090, 0.2208]	0.1542 [0.1501, 0.1580]	0.2102 [0.2026, 0.2178]	0.2373 [0.2311, 0.2434]
METEOR	0.2161 [0.2121, 0.2204]	0.2494 [0.2459, 0.2529]	0.2200 [0.2162, 0.2237]	0.2372 [0.2344, 0.2401]	0.2193 [0.2156, 0.2231]	0.2373 [0.2338, 0.2407]
ROUGE-L	0.4156 [0.4077, 0.4232]	0.4499 [0.4423, 0.4574]	0.3922 [0.3852, 0.3991]	0.3550 [0.3479, 0.3624]	0.4210 [0.4128, 0.4295]	0.4165 [0.4096, 0.4234]
RadBERT-P.	0.5658 [0.5498, 0.5835]	0.5504 [0.5351, 0.5660]	0.5565 [0.5410, 0.5718]	0.5301 [0.5155, 0.5452]	0.5913 [0.5752, 0.6088]	0.5570 [0.5423, 0.5725]
RadBERT-R.	0.4635 [0.4502, 0.4781]	0.5125 [0.4991, 0.5275]	0.5377 [0.5237, 0.5524]	0.5453 [0.5310, 0.5597]	0.4308 [0.4184, 0.4436]	0.5397 [0.5252, 0.5544]
RadBERT-F1	0.5096 [0.4971, 0.5222]	0.5308 [0.5186, 0.5428]	0.5470 [0.5353, 0.5591]	0.5376 [0.5253, 0.5501]	0.4985 [0.4865, 0.5107]	0.5482 [0.5360, 0.5611]
Rate-Score	0.3891 [0.3825, 0.3960]	0.4069 [0.4000, 0.4138]	0.3868 [0.3800, 0.3934]	0.4333 [0.4260, 0.4404]	0.4052 [0.3980, 0.4122]	0.4121 [0.4052, 0.4190]
F.-Degree	0.3898 [0.3816, 0.3984]	0.4245 [0.4165, 0.4324]	0.3671 [0.3591, 0.3759]	0.3988 [0.3907, 0.4065]	0.4228 [0.4150, 0.4311]	0.4549 [0.4467, 0.4631]
F.-Landmark	0.4958 [0.4868, 0.5048]	0.5434 [0.5344, 0.5521]	0.4810 [0.4713, 0.4909]	0.5456 [0.5364, 0.5547]	0.5226 [0.5137, 0.5315]	0.5597 [0.5510, 0.5685]
F.-Feature	0.3403 [0.3286, 0.3521]	0.4042 [0.3929, 0.4153]	0.3583 [0.3464, 0.3700]	0.4062 [0.3950, 0.4164]	0.3678 [0.3564, 0.3792]	0.4154 [0.4037, 0.4262]
F.-Impression	0.3833 [0.3687, 0.3986]	0.4243 [0.4108, 0.4390]	0.4013 [0.3871, 0.4156]	0.4211 [0.4075, 0.4343]	0.4170 [0.4021, 0.4314]	0.4400 [0.4258, 0.4544]
F.-Overall	0.4023 [0.3941, 0.4109]	0.4491 [0.4407, 0.4574]	0.4020 [0.3931, 0.4108]	0.4429 [0.4346, 0.4506]	0.4326 [0.4243, 0.4406]	0.4675 [0.4589, 0.4754]

Table B5: **Ablation study on GRPO reward function** (n = 1,564 cases on CT-Rate; 95% confidence intervals estimated via 2,000 bootstrap iterations). *Base*: SFT baseline without RL; *NLG*: RL with NLG reward only; *RadBERT*: RL with RadBERT reward only; *Rate-Score*: RL with Rate-Score reward only; *FORTE-Ori*: RL with original FORTE reward; *FORTE*: RL with updated FORTE reward (Astra). F. means FORTE metric.

Structured Abdominal CT Report Extraction Prompt:
System Message: You are an expert radiology report processor specializing in extracting structured information from abdomen CT reports.
User Message: Your task is to analyze a given abdomen CT report, apply a set of specific rules, and output the results as a single, valid JSON object.

Target Anatomical Regions:
You must extract information for the following 13 specific anatomical regions ONLY:
1. lower thorax: includes lower chest and lung bases
2. liver and biliary tree: includes liver and biliary tree (bile ducts), but NOT gallbladder
3. gallbladder: gallbladder only (separate from liver)
4. spleen: spleen only
5. pancreas: pancreas only
6. adrenal glands: adrenal glands only
7. kidneys and ureters: includes kidneys and ureters
8. gastrointestinal tract: includes stomach, intestines, appendix; be careful not to miss the appendix
9. peritoneum: includes peritoneal space, peritoneal cavity, and abdominal wall
10. pelvic: includes pelvic organs, bladder, prostate and seminal vesicles, uterus and ovaries
11. vasculature: vasculature system
12. lymph nodes: lymph nodes
13. musculoskeletal: includes bones and muscles

Rule 1: Remove Non-Diagnostic Text
Before any other processing, you must identify and completely remove any text that represents communication between clinicians, summary codes, or procedural notes. This includes but is not limited to phrases like "FINDINGS GIVEN TO Dr. ...", "SUMMARY 4:...", "END OF IMPRESSION:", and "AT THE CONCLUSION OF THE EXAMINATION...".

Rule 2: Process Each Target Region
For each of the 13 target regions listed above, find the corresponding section in the report and apply these sub-rules:
A. **Delete Negative Findings:** Remove all sentences, clauses, or phrases that describe normal, negative, or absent findings. Examples include "The liver is normal in size and shape", "with coordinated proportion of liver lobes", "unremarkable appearance", and "without evidence of...".
B. **Rewrite Comparative Statements:** If a finding is compared to a prior scan (for example, "stable", "unchanged", or "increased/decreased in size"), rewrite the sentence to describe only the current state of the finding. For example, "The 2 cm nodule is stable" becomes "There is a 2 cm nodule." If a finding has resolved, it is now normal and should be removed according to Rule 2A.
C. **Preserve Positive Findings:** For general non-comparative positive findings, the original statements should be retained in their entirety **without any modifications**.

Rule 3: Generate Structured JSON Output
The final output MUST be a single, valid JSON object and nothing else. Do not include any introductory text, explanations, or markdown code block markers.
The JSON object must have keys corresponding to ALL 13 predefined anatomical regions.
For each region:
1. If there are any positive findings remaining after processing, the value should be a string containing those findings.
2. If a region had no positive findings (i.e., it was normal, all findings were negative, or the section was not mentioned in the report), the value MUST be the string "normal".

Report: {report}
Processed Report (JSON): ...

Table B6: Prompt used for structured information extraction from abdominal CT reports.

Structured Chest CT Report Extraction Prompt:
System Message: You are an expert radiology report processor specializing in extracting structured information from chest CT reports.
User Message: Your task is to analyze a given chest CT report, apply a set of specific rules, and output the results as a single, valid JSON object.

Target Anatomical Regions:
You must extract information for the following 10 specific anatomical regions ONLY:
1. abdomen: include all findings related to liver, gallbladder, pancreas, spleen, kidneys, adrenals, gastrointestinal tract, abdominal vessels, and abdominal lymph nodes, etc.
2. bone: bone
3. breast: breast
4. esophagus: esophagus
5. heart: heart
6. lung: lung
7. mediastinum: mediastinum area
8. pleura: pleura
9. thyroid: thyroid
10. trachea and bronchi: trachea and bronchi

Rule 1: Remove Non-Diagnostic Text
Before any other processing, you must identify and completely remove any text that represents communication between clinicians, summary codes, or procedural notes. This includes but is not limited to phrases like "FINDINGS GIVEN TO Dr. ...", "SUMMARY 4:...", "END OF IMPRESSION:", and "AT THE CONCLUSION OF THE EXAMINATION...", etc.

Rule 2: Process Each Primary Target Region
For each of the 10 primary target regions listed above, find the corresponding section in the report and apply these sub-rules:
A. **Delete Negative Findings:** Remove all sentences, clauses, or phrases that describe normal, negative, or absent findings. Examples include "The lungs are clear", "No pleural effusion is seen", "unremarkable appearance", and "without evidence of...".
B. **Rewrite Comparative Statements:** If a finding is compared to a prior scan (for example, "stable", "unchanged", or "increased/decreased in size"), rewrite the sentence to describe only the current state of the finding. For example, "The 2 cm nodule is stable" becomes "There is a 2 cm nodule." If a finding has resolved, it is now normal and should be removed according to Rule 2A.
C. **Preserve Positive Findings:** For general non-comparative positive findings, the original statements should be retained in their entirety **without any modifications**.

Rule 5: Generate Structured JSON Output
The final output MUST be a single, valid JSON object and nothing else. Do not include any introductory text, explanations, or markdown code block markers.
The JSON object must have keys corresponding to ALL 10 primary anatomical regions. If Rule 3 is triggered, an 11th key, "brain", must also be included.
For each region:
1. If there are any positive findings remaining after processing, the value should be a string containing those findings.
2. If a region had no positive findings (i.e., it was normal, all findings were negative, or the section was not mentioned in the report), the value MUST be the string "normal".
Report: `{report}`
Processed Report (JSON): ...

Table B7: Prompt used for structured information extraction from chest CT reports.

Abdominal CT Example Case:
Input Report: "Abdomen: The liver is normal in size and shape, with coordinated proportion of liver lobes. Multiple mixed high- and low-density lesions are found in the liver, with CT values of approximately 17–186 HU. The largest lesion is in the posterior right lobe of the liver measuring approximately 69mm x 88mm. On the enhancement scan, no significant enhancement is seen in the lesions, with a low-density ring seen surrounding the lesions. Patchy slightly low-density foci are seen in the left lobe of the liver(S4), measuring approximately 14.3mm x 5.8mm, with delayed isodense enhancement. The intrahepatic duct system and common bile duct are not dilated and normal in course. The portal structure is clear. The position and morphology of gallbladder are normal. The spleen is enlarged, approximately 8 rib units, with no abnormal density. A round low-density lesion with no enhanced is found in the left kidney, with a diameter of approximately 4.3mm. Pelvis: The local lumen of the rectum and sigmoid colon is dilated, with increased wall thickness, measuring approximately 29mm x 55mm. On the enhancement scan, the lesions show lower uneven enhancement than normal intestinal walls. Multiple small lymph nodes are seen around the intestine. The myometrium enhancement is uneven."

Expected Output:


```
{
  "lower thorax": "normal",
  "liver and biliary tree": "Multiple mixed high- and low-density lesions are found in the liver, with CT values
of approximately 17--186 HU. The largest lesion is in the posterior right lobe of the liver measuring approximately
69mm x 88mm. On the enhancement scan, no significant enhancement is seen in the lesions, with a low-density ring
seen surrounding the lesions. Patchy slightly low-density foci are seen in the left lobe of the liver(S4), measuring
approximately 14.3mm x 5.8mm, with delayed isodense enhancement.",
  "gallbladder": "normal",
  "spleen": "The spleen is enlarged, approximately 8 rib units.",
  "pancreas": "normal",
  "adrenal glands": "normal",
  "kidneys and ureters": "A round low-density lesion with no enhanced is found in the left kidney, with a diameter
of approximately 4.3mm.",
  "gastrointestinal tract": "The local lumen of the rectum and sigmoid colon is dilated, with increased wall
thickness, measuring approximately 29mm x 55mm. On the enhancement scan, the lesions show lower uneven enhancement
than normal intestinal walls.",
  "peritoneum": "normal",
  "pelvic": "The myometrium enhancement is uneven.",
  "vasculature": "normal",
  "lymph nodes": "Multiple small lymph nodes are seen around the intestine.",
  "musculoskeletal": "normal"
}
```


Table B8: Illustrative example of structured information extraction from an abdominal CT report.

Chest CT Example Case:
Input Report: "LUNGS: Mild centrilobular emphysema. No focal consolidation. Several small nodules are stable. HEART AND MEDIASTINUM: The cardiomediastinal silhouette is normal in size. No pericardial effusion. No mediastinal lymphadenopathy. PLEURA: No pleural effusion or pneumothorax. BONES: Degenerative changes of the thoracic spine are present. FINDINGS GIVEN TO Dr. ..., AT THE CONCLUSION OF THE EXAMINATION, BY THE ONCALL RADIOLOGY RESIDENT. END OF IMPRESSION: SUMMARY 2: No acute abnormality."

Expected Output:


```
{
  "abdomen": "normal",
  "bone": "Degenerative changes of the thoracic spine are present.",
  "breast": "normal",
  "esophagus": "normal",
  "heart": "normal",
  "lung": "Mild centrilobular emphysema. There are several small nodules.",
  "mediastinum": "normal",
  "pleura": "normal",
  "thyroid": "normal",
  "trachea and bronchi": "normal"
}
```


Table B9: Illustrative example of structured information extraction from a chest CT report.

Prompt Template for Medical Term Extraction (FORTE Pre-processing):

You are an expert medical terminologist. Your task is to extract specific medical terms from a given CT report and categorize them into four distinct dimensions.

The four dimensions are:
1. Degree: Terms describing size, intensity, severity, or quantity (e.g., "mild", "diffuse", "multiple", "enlarged").
2. Landmark: Anatomical locations, organs, or spatial descriptors (e.g., "liver", "pancreas", "right kidney", "pericholecystic").
3. Feature: Visual characteristics of diseases or primary abnormal findings (e.g., "fat stranding", "thickening", "lesion", "stone", "ascites").
4. Impression: Final diagnostic conclusions or disease names (e.g., "cholecystitis", "pancreatitis", "cirrhosis", "renal stone").

Instructions:
• Analyze the provided CT report.
• Extract all relevant terms for each of the four categories.
• For each category, list the unique terms found in the report.
• If no terms are found for a category, provide an empty list `[]`.
• Your output **MUST** be a single JSON object with four keys: `"degree"`, `"landmark"`, `"feature"`, `"impression"`. The value for each key should be a list of strings. Do not add any other text or explanations before or after the JSON object.

Example Input Report:
```
''IMPRESSION: Findings consistent with right upper lobe pneumonia. FINDINGS: There is a focal area of consolidation in the right upper lobe. Mild pleural effusion is noted. The mediastinum is unremarkable. No evidence of pulmonary embolism.''
```

Example Output JSON:
```
{
  "degree": ["acute", "diffuse", "mild", "small", "unremarkable", "enlarged"],
  "landmark": ["gallbladder", "pericholecystic", "pancreas", "liver"],
  "feature": ["thickening", "fat stranding", "stone"],
  "impression": ["cholecystitis", "hepatic steatosis"]
}
```

Now, process the following CT report:

CT Report:
{report_text}

Output JSON:

Table B10: Prompt used to extract structured medical terms from raw CT reports to update FORTE keywords. Given a CT report, the DeepSeek is instructed to identify and categorize terms into four dimensions: *Degree* (severity/quantity descriptors), *Landmark* (anatomical locations), *Feature* (visual abnormality characteristics), and *Impression* (diagnostic conclusions). The extracted terms were then manually reviewed and merged with synonyms to build the final FORTE terminology lexicon.

Model	BLEU-1 [95% CI]	BLEU-2 [95% CI]	BLEU-3 [95% CI]	BLEU-4 [95% CI]	ROUGE-L [95% CI]	METEOR [95% CI]
Astra	0.4866* [0.4769, 0.4958]	0.3803* [0.3723, 0.3878]	0.3052* [0.2978, 0.3119]	0.2501* [0.2433, 0.2566]	0.4409* [0.4330, 0.4490]	0.2397* [0.2359, 0.2434]
Gemini-3	0.1503 [0.1394, 0.1606]	0.1171 [0.1083, 0.1257]	0.0982 [0.0907, 0.1056]	0.0850 [0.0785, 0.0918]	0.3665 [0.3568, 0.3769]	0.1455 [0.1414, 0.1496]
Qwen3-VL-8B	0.2132 [0.2005, 0.2265]	0.1602 [0.1505, 0.1702]	0.1328 [0.1245, 0.1413]	0.1143 [0.1070, 0.1216]	0.3293 [0.3204, 0.3384]	0.1407 [0.1368, 0.1447]
HuluMed-32B	0.3001 [0.2868, 0.3137]	0.2324 [0.2219, 0.2432]	0.1886 [0.1798, 0.1977]	0.1575 [0.1498, 0.1653]	0.3902 [0.3815, 0.3988]	0.1826 [0.1785, 0.1866]
HuluMed-14B	0.3025 [0.2885, 0.3166]	0.2322 [0.2216, 0.2432]	0.1877 [0.1786, 0.1967]	0.1562 [0.1482, 0.1641]	0.3801 [0.3714, 0.3886]	0.1794 [0.1753, 0.1836]
HuluMed-7B	0.2970 [0.2828, 0.3107]	0.2269 [0.2160, 0.2373]	0.1832 [0.1741, 0.1920]	0.1527 [0.1449, 0.1604]	0.3688 [0.3603, 0.3777]	0.1756 [0.1714, 0.1797]
M3D	0.1841 [0.1717, 0.1967]	0.1404 [0.1306, 0.1505]	0.1166 [0.1083, 0.1252]	0.1000 [0.0926, 0.1076]	0.3219 [0.3145, 0.3293]	0.1397 [0.1356, 0.1436]
RadFM	0.1065 [0.0955, 0.1177]	0.0869 [0.0778, 0.0963]	0.0758 [0.0679, 0.0840]	0.0674 [0.0603, 0.0747]	0.3376 [0.3289, 0.3463]	0.1318 [0.1274, 0.1361]
Lingshu-32B	0.2873 [0.2766, 0.2979]	0.2068 [0.1984, 0.2152]	0.1645 [0.1572, 0.1718]	0.1371 [0.1308, 0.1436]	0.3020 [0.2961, 0.3076]	0.1467 [0.1433, 0.1502]
Lingshu-7B	0.0611 [0.0540, 0.0684]	0.0510 [0.0450, 0.0572]	0.0452 [0.0398, 0.0508]	0.0406 [0.0357, 0.0457]	0.3539 [0.3434, 0.3640]	0.1302 [0.1259, 0.1346]
MedGemma-27B	0.1174 [0.1058, 0.1294]	0.0952 [0.0858, 0.1049]	0.0826 [0.0745, 0.0910]	0.0733 [0.0661, 0.0809]	0.3540 [0.3431, 0.3650]	0.1331 [0.1290, 0.1371]
MedGemma-4B	0.0257 [0.0221, 0.0295]	0.0228 [0.0195, 0.0262]	0.0209 [0.0179, 0.0241]	0.0192 [0.0165, 0.0222]	0.3730 [0.3614, 0.3848]	0.1303 [0.1257, 0.1350]
R2GenGPT	0.3330 [0.3197, 0.3459]	0.2620 [0.2513, 0.2725]	0.2133 [0.2041, 0.2221]	0.1779 [0.1698, 0.1856]	0.4198 [0.4120, 0.4277]	0.1980 [0.1937, 0.2021]
LLaVA	0.2012 [0.1891, 0.2129]	0.1641 [0.1541, 0.1740]	0.1387 [0.1300, 0.1474]	0.1198 [0.1122, 0.1275]	0.4172 [0.4081, 0.4263]	0.1754 [0.1709, 0.1796]

Table B11: **Natural language generation metrics on the CT-Rate dataset** ($n = 1564$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.5789* [0.5632, 0.5948]	0.5242* [0.5103, 0.5381]	0.5502* [0.5380, 0.5628]
Gemini-3	0.4134 [0.3917, 0.4363]	0.1828 [0.1714, 0.1944]	0.2535 [0.2397, 0.2677]
Qwen3-VL-8B	0.3561 [0.3244, 0.3867]	0.0695 [0.0611, 0.0779]	0.1163 [0.1030, 0.1292]
HuluMed-32B	0.4687 [0.4494, 0.4868]	0.3210 [0.3058, 0.3360]	0.3810 [0.3660, 0.3952]
HuluMed-14B	0.4373 [0.4182, 0.4558]	0.3212 [0.3065, 0.3366]	0.3704 [0.3561, 0.3850]
HuluMed-7B	0.4682 [0.4484, 0.4882]	0.2515 [0.2368, 0.2660]	0.3272 [0.3113, 0.3426]
M3D	0.2446 [0.2262, 0.2637]	0.0997 [0.0915, 0.1082]	0.1417 [0.1311, 0.1528]
RadFM	0.2850 [0.2600, 0.3125]	0.0644 [0.0577, 0.0716]	0.1050 [0.0949, 0.1162]
Lingshu-32B	0.2102 [0.1857, 0.2347]	0.0573 [0.0505, 0.0644]	0.0901 [0.0796, 0.1008]
Lingshu-7B	0.2462 [0.2167, 0.2757]	0.0449 [0.0389, 0.0508]	0.0759 [0.0661, 0.0854]
MedGemma-27B	0.2907 [0.2048, 0.3810]	0.0050 [0.0031, 0.0071]	0.0098 [0.0061, 0.0140]
MedGemma-4B	0.3643 [0.2857, 0.4468]	0.0101 [0.0076, 0.0129]	0.0197 [0.0148, 0.0251]
R2GenGPT	0.5118 [0.4941, 0.5300]	0.3580 [0.3448, 0.3714]	0.4213 [0.4079, 0.4348]
LLaVA	0.5406 [0.5220, 0.5613]	0.2751 [0.2630, 0.2876]	0.3647 [0.3510, 0.3782]

Table B12: **RadBERT classification performance (micro Precision, Recall, F1) on the CT-Rate dataset** ($n = 1{,}564$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * indicates the best performance across all models for each metric.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.3510* [0.3436, 0.3580]	0.3844* [0.3739, 0.3946]	0.5140* [0.5033, 0.5252]	0.4159* [0.4034, 0.4273]	0.4457* [0.4306, 0.4601]	0.4400* [0.4300, 0.4493]
Gemini-3	0.2225 [0.2190, 0.2261]	0.1817 [0.1727, 0.1910]	0.2243 [0.2151, 0.2340]	0.1409 [0.1325, 0.1494]	0.1459 [0.1340, 0.1579]	0.1732 [0.1659, 0.1810]
Qwen3-VL-8B	0.2058 [0.2022, 0.2095]	0.0963 [0.0884, 0.1045]	0.1831 [0.1743, 0.1918]	0.0648 [0.0578, 0.0723]	0.0636 [0.0544, 0.0741]	0.1020 [0.0955, 0.1090]
HuluMed-32B	0.2528 [0.2479, 0.2576]	0.2274 [0.2177, 0.2366]	0.3059 [0.2948, 0.3168]	0.2278 [0.2174, 0.2380]	0.2388 [0.2252, 0.2528]	0.2500 [0.2414, 0.2585]
HuluMed-14B	0.2534 [0.2486, 0.2584]	0.2276 [0.2177, 0.2369]	0.2875 [0.2769, 0.2983]	0.2157 [0.2055, 0.2257]	0.2402 [0.2265, 0.2541]	0.2427 [0.2340, 0.2514]
HuluMed-7B	0.2361 [0.2315, 0.2410]	0.1926 [0.1827, 0.2026]	0.2708 [0.2590, 0.2830]	0.1889 [0.1781, 0.1997]	0.1923 [0.1776, 0.2062]	0.2111 [0.2018, 0.2202]
M3D	0.2038 [0.2009, 0.2066]	0.1352 [0.1272, 0.1435]	0.1989 [0.1900, 0.2078]	0.0740 [0.0677, 0.0801]	0.0677 [0.0590, 0.0762]	0.1189 [0.1131, 0.1248]
RadFM	0.1881 [0.1856, 0.1905]	0.0688 [0.0619, 0.0754]	0.0891 [0.0799, 0.0978]	0.0556 [0.0489, 0.0621]	0.0537 [0.0450, 0.0620]	0.0668 [0.0608, 0.0725]
Lingshu-32B	0.2232 [0.2201, 0.2264]	0.1158 [0.1083, 0.1238]	0.2071 [0.1999, 0.2141]	0.0654 [0.0591, 0.0718]	0.0730 [0.0631, 0.0827]	0.1153 [0.1100, 0.1204]
Lingshu-7B	0.1690 [0.1666, 0.1715]	0.0635 [0.0572, 0.0701]	0.0601 [0.0542, 0.0665]	0.0426 [0.0374, 0.0479]	0.0464 [0.0387, 0.0542]	0.0532 [0.0487, 0.0579]
MedGemma-27B	0.1622 [0.1599, 0.1646]	0.0327 [0.0279, 0.0378]	0.0546 [0.0487, 0.0604]	0.0147 [0.0113, 0.0185]	0.0069 [0.0039, 0.0104]	0.0272 [0.0241, 0.0305]
MedGemma-4B	0.1513 [0.1490, 0.1535]	0.0191 [0.0143, 0.0240]	0.0336 [0.0278, 0.0398]	0.0132 [0.0096, 0.0170]	0.0119 [0.0073, 0.0170]	0.0194 [0.0157, 0.0234]
R2GenGPT	0.3084 [0.3021, 0.3151]	0.2798 [0.2699, 0.2897]	0.3829 [0.3724, 0.3941]	0.2886 [0.2770, 0.3000]	0.3167 [0.3023, 0.3317]	0.3170 [0.3081, 0.3259]
LLaVA	0.2911 [0.2847, 0.2972]	0.2658 [0.2556, 0.2758]	0.3025 [0.2908, 0.3132]	0.2503 [0.2381, 0.2625]	0.2988 [0.2827, 0.3143]	0.2793 [0.2696, 0.2891]

Table B13: **Fine-grained caption metrics (Rate-Score, FORTE) on the CT-Rate dataset** ($n = 1564$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * indicates the best performance across all models for each metric. F. means FORTE metric.

Model	BLEU-1 [95% CI]	BLEU-2 [95% CI]	BLEU-3 [95% CI]	BLEU-4 [95% CI]	ROUGE-L [95% CI]	METEOR [95% CI]
Astra	0.3898* [0.3786, 0.4016]	0.2844* [0.2757, 0.2936]	0.2185* [0.2111, 0.2264]	0.1759* [0.1691, 0.1828]	0.3928 [0.3845, 0.4003]	0.2065* [0.2027, 0.2103]
Gemini-3	0.1673 [0.1562, 0.1790]	0.1313 [0.1221, 0.1406]	0.1085 [0.1007, 0.1166]	0.0929 [0.0860, 0.1001]	0.3478 [0.3382, 0.3571]	0.1526 [0.1484, 0.1568]
Qwen3-VL-8B	0.2719 [0.2596, 0.2844]	0.1992 [0.1894, 0.2089]	0.1591 [0.1507, 0.1674]	0.1339 [0.1266, 0.1414]	0.3100 [0.3017, 0.3177]	0.1504 [0.1466, 0.1541]
HuluMed-32B	0.2971 [0.2911, 0.3034]	0.1816 [0.1779, 0.1854]	0.1196 [0.1169, 0.1225]	0.0859 [0.0835, 0.0885]	0.2410 [0.2375, 0.2445]	0.1708 [0.1683, 0.1734]
HuluMed-14B	0.3453 [0.3390, 0.3494]	0.2297 [0.2253, 0.2328]	0.1628 [0.1590, 0.1656]	0.1241 [0.1204, 0.1269]	0.2555 [0.2509, 0.2600]	0.1593 [0.1561, 0.1623]
HuluMed-7B	0.2876 [0.2734, 0.3013]	0.2111 [0.2004, 0.2215]	0.1660 [0.1571, 0.1746]	0.1370 [0.1292, 0.1445]	0.3084 [0.3013, 0.3154]	0.1587 [0.1546, 0.1625]
M3D	0.2537 [0.2404, 0.2673]	0.1869 [0.1767, 0.1973]	0.1479 [0.1391, 0.1565]	0.1229 [0.1152, 0.1304]	0.3026 [0.2956, 0.3097]	0.1514 [0.1475, 0.1553]
RadFM	0.1306 [0.1194, 0.1424]	0.1059 [0.0966, 0.1154]	0.0894 [0.0813, 0.0977]	0.0776 [0.0704, 0.0851]	0.3311 [0.3214, 0.3405]	0.1434 [0.1388, 0.1478]
Lingshu-32B	0.2749 [0.2627, 0.2869]	0.2004 [0.1905, 0.2098]	0.1595 [0.1509, 0.1677]	0.1336 [0.1259, 0.1410]	0.3010 [0.2941, 0.3076]	0.1502 [0.1463, 0.1540]
Lingshu-7B	0.1236 [0.1124, 0.1352]	0.0999 [0.0906, 0.1094]	0.0841 [0.0761, 0.0925]	0.0729 [0.0657, 0.0802]	0.3377 [0.3280, 0.3470]	0.1438 [0.1394, 0.1481]
MedGemma-27B	0.1764 [0.1648, 0.1888]	0.1368 [0.1275, 0.1465]	0.1120 [0.1040, 0.1200]	0.0948 [0.0878, 0.1018]	0.3300 [0.3203, 0.3394]	0.1484 [0.1442, 0.1525]
MedGemma-4B	0.1688 [0.1572, 0.1799]	0.1328 [0.1233, 0.1421]	0.1090 [0.1009, 0.1171]	0.0923 [0.0852, 0.0994]	0.3350 [0.3260, 0.3439]	0.1517 [0.1474, 0.1558]
R2GenGPT	0.3525 [0.3408, 0.3638]	0.2590 [0.2498, 0.2679]	0.2007 [0.1929, 0.2081]	0.1624 [0.1554, 0.1690]	0.3698 [0.3618, 0.3776]	0.1857 [0.1816, 0.1897]
LLaVA	0.2607 [0.2475, 0.2743]	0.2012 [0.1907, 0.2119]	0.1630 [0.1541, 0.1721]	0.1371 [0.1291, 0.1451]	0.3899 [0.3807, 0.3991]	0.1792 [0.1750, 0.1834]

Table B14: **Natural language generation metrics on the Merlin dataset** ($n = 1{,}000$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.5263 [0.5069, 0.5463]	0.4714* [0.4536, 0.4895]	0.4973* [0.4810, 0.5137]
Gemini-3	0.2873 [0.2571, 0.3163]	0.0845 [0.0740, 0.0947]	0.1305 [0.1158, 0.1450]
Qwen3-VL-8B	0.0945 [0.0789, 0.1100]	0.0363 [0.0301, 0.0425]	0.0525 [0.0436, 0.0612]
HuluMed-32B	0.2900 [0.2641, 0.3170]	0.1276 [0.1160, 0.1396]	0.1772 [0.1620, 0.1929]
HuluMed-14B	0.2778 [0.2556, 0.3006]	0.1542 [0.1415, 0.1675]	0.1983 [0.1828, 0.2144]
HuluMed-7B	0.2320 [0.2076, 0.2580]	0.0846 [0.0747, 0.0949]	0.1240 [0.1102, 0.1384]
M3D	0.2026 [0.1741, 0.2295]	0.0537 [0.0455, 0.0615]	0.0849 [0.0723, 0.0966]
RadFM	0.2134 [0.1707, 0.2574]	0.0202 [0.0156, 0.0248]	0.0369 [0.0287, 0.0451]
Lingshu-32B	0.1406 [0.1183, 0.1652]	0.0378 [0.0317, 0.0444]	0.0595 [0.0500, 0.0698]
Lingshu-7B	0.1813 [0.1496, 0.2131]	0.0375 [0.0307, 0.0446]	0.0622 [0.0511, 0.0738]
MedGemma-27B	0.1748 [0.1146, 0.2381]	0.0072 [0.0045, 0.0101]	0.0139 [0.0088, 0.0194]
MedGemma-4B	0.2375 [0.1978, 0.2777]	0.0329 [0.0271, 0.0393]	0.0578 [0.0478, 0.0686]
R2GenGPT	0.4685 [0.4503, 0.4856]	0.4007 [0.3842, 0.4174]	0.4319 [0.4169, 0.4467]
LLaVA	0.5165 [0.4970, 0.5376]	0.3566 [0.3404, 0.3731]	0.4219 [0.4062, 0.4382]

Table B15: **RadBERT classification performance (micro Precision, Recall, F1) on the Merlin dataset** (n = 1,000 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.3564* [0.3509, 0.3618]	0.3765* [0.3648, 0.3872]	0.4107* [0.4001, 0.4209]	0.3750* [0.3654, 0.3844]	0.4684* [0.4549, 0.4809]	0.4076* [0.3992, 0.4157]
Gemini-3	0.2154 [0.2117, 0.2193]	0.0975 [0.0881, 0.1069]	0.1299 [0.1219, 0.1386]	0.1170 [0.1094, 0.1250]	0.1743 [0.1617, 0.1870]	0.1296 [0.1222, 0.1373]
Qwen3-VL-8B	0.2007 [0.1981, 0.2035]	0.1001 [0.0917, 0.1085]	0.0954 [0.0890, 0.1017]	0.0735 [0.0671, 0.0799]	0.0543 [0.0470, 0.0618]	0.0808 [0.0757, 0.0859]
HuluMed-32B	0.2115 [0.2088, 0.2143]	0.1265 [0.1195, 0.1336]	0.1968 [0.1889, 0.2048]	0.1445 [0.1378, 0.1511]	0.1203 [0.1110, 0.1298]	0.1442 [0.1391, 0.1496]
HuluMed-14B	0.2171 [0.2143, 0.2199]	0.1389 [0.1313, 0.1459]	0.1416 [0.1337, 0.1496]	0.1382 [0.1309, 0.1456]	0.1240 [0.1135, 0.1354]	0.1357 [0.1293, 0.1419]
HuluMed-7B	0.1963 [0.1934, 0.1992]	0.0997 [0.0916, 0.1079]	0.1153 [0.1072, 0.1236]	0.0717 [0.0652, 0.0788]	0.0744 [0.0648, 0.0837]	0.0903 [0.0843, 0.0962]
M3D	0.1891 [0.1865, 0.1919]	0.0912 [0.0827, 0.0996]	0.1064 [0.0990, 0.1140]	0.0635 [0.0575, 0.0697]	0.0476 [0.0396, 0.0555]	0.0772 [0.0716, 0.0827]
RadFM	0.1687 [0.1668, 0.1708]	0.0431 [0.0368, 0.0497]	0.0442 [0.0385, 0.0503]	0.0225 [0.0185, 0.0265]	0.0153 [0.0104, 0.0205]	0.0313 [0.0273, 0.0353]
Lingshu-32B	0.1955 [0.1930, 0.1980]	0.0442 [0.0384, 0.0498]	0.0880 [0.0811, 0.0949]	0.0612 [0.0547, 0.0678]	0.0468 [0.0392, 0.0549]	0.0600 [0.0554, 0.0648]
Lingshu-7B	0.1767 [0.1743, 0.1790]	0.0371 [0.0312, 0.0435]	0.0635 [0.0570, 0.0704]	0.0382 [0.0329, 0.0436]	0.0413 [0.0335, 0.0489]	0.0450 [0.0403, 0.0497]
MedGemma-27B	0.1719 [0.1699, 0.1741]	0.0195 [0.0154, 0.0239]	0.0570 [0.0505, 0.0639]	0.0322 [0.0275, 0.0368]	0.0218 [0.0167, 0.0276]	0.0326 [0.0291, 0.0364]
MedGemma-4B	0.1775 [0.1755, 0.1797]	0.0392 [0.0335, 0.0450]	0.0644 [0.0577, 0.0711]	0.0398 [0.0345, 0.0453]	0.0248 [0.0190, 0.0307]	0.0421 [0.0378, 0.0465]
R2GenGPT	0.3229 [0.3176, 0.3285]	0.2965 [0.2861, 0.3072]	0.3140 [0.3043, 0.3237]	0.2882 [0.2793, 0.2976]	0.3851 [0.3713, 0.3985]	0.3210 [0.3130, 0.3292]
LLaVA	0.3204 [0.3144, 0.3267]	0.2640 [0.2530, 0.2757]	0.3009 [0.2901, 0.3110]	0.2749 [0.2652, 0.2846]	0.3928 [0.3769, 0.4078]	0.3082 [0.2990, 0.3172]

Table B16: **Fine-grained caption metrics (Rate-Score, FORTE) on the Merlin dataset** (n = 1,000 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	BLEU-1 [95% CI]	BLEU-2 [95% CI]	BLEU-3 [95% CI]	BLEU-4 [95% CI]	ROUGE-L [95% CI]	METEOR [95% CI]
Astra	0.4622* [0.4465, 0.4770]	0.3912* [0.3769, 0.4047]	0.3479* [0.3342, 0.3609]	0.3140* [0.3010, 0.3265]	0.5394 [0.5276, 0.5512]	0.2685* [0.2620, 0.2753]
Gemini-3	0.3954 [0.3860, 0.4052]	0.3092 [0.3008, 0.3187]	0.2616 [0.2537, 0.2706]	0.2262 [0.2186, 0.2348]	0.4176 [0.4071, 0.4286]	0.2216 [0.2167, 0.2265]
Qwen3-VL-8B	0.3549 [0.3433, 0.3676]	0.2845 [0.2741, 0.2957]	0.2478 [0.2382, 0.2585]	0.2212 [0.2120, 0.2313]	0.4168 [0.4048, 0.4294]	0.2216 [0.2163, 0.2270]
HuluMed-32B	0.2623 [0.2553, 0.2699]	0.1993 [0.1932, 0.2059]	0.1651 [0.1596, 0.1712]	0.1412 [0.1359, 0.1469]	0.3456 [0.3368, 0.3551]	0.2093 [0.2055, 0.2135]
HuluMed-14B	0.3750 [0.3639, 0.3865]	0.2973 [0.2878, 0.3076]	0.2547 [0.2460, 0.2641]	0.2228 [0.2146, 0.2316]	0.4050 [0.3954, 0.4142]	0.2229 [0.2180, 0.2282]
HuluMed-7B	0.2800 [0.2694, 0.2919]	0.2154 [0.2067, 0.2251]	0.1806 [0.1731, 0.1890]	0.1554 [0.1486, 0.1631]	0.3467 [0.3385, 0.3544]	0.2103 [0.2060, 0.2145]
M3D	0.4314 [0.4196, 0.4371]	0.3531 [0.3425, 0.3595]	0.3092 [0.2987, 0.3159]	0.2752 [0.2653, 0.2819]	0.4237 [0.4143, 0.4335]	0.2248 [0.2193, 0.2308]
RadFM	0.4017 [0.3865, 0.4162]	0.3451 [0.3313, 0.3580]	0.3108 [0.2980, 0.3232]	0.2821 [0.2701, 0.2940]	0.4546 [0.4446, 0.4645]	0.2291 [0.2229, 0.2354]
Lingshu-32B	0.3036 [0.2974, 0.3099]	0.2373 [0.2329, 0.2419]	0.2020 [0.1982, 0.2060]	0.1770 [0.1735, 0.1806]	0.3451 [0.3393, 0.3508]	0.2149 [0.2108, 0.2191]
Lingshu-7B	0.3408 [0.3227, 0.3593]	0.2990 [0.2826, 0.3155]	0.2724 [0.2571, 0.2876]	0.2495 [0.2352, 0.2641]	0.5024 [0.4903, 0.5146]	0.2343 [0.2273, 0.2411]
MedGemma-27B	0.3974 [0.3855, 0.4096]	0.3309 [0.3194, 0.3425]	0.2931 [0.2820, 0.3045]	0.2644 [0.2535, 0.2754]	0.4372 [0.4239, 0.4501]	0.2258 [0.2197, 0.2316]
MedGemma-4B	0.2954 [0.2765, 0.3151]	0.2681 [0.2508, 0.2862]	0.2498 [0.2334, 0.2666]	0.2335 [0.2180, 0.2493]	0.5359 [0.5216, 0.5502]	0.2411 [0.2334, 0.2488]
R2GenGPT	0.4488 [0.4319, 0.4643]	0.3824 [0.3672, 0.3967]	0.3408 [0.3264, 0.3541]	0.3080 [0.2943, 0.3206]	0.5305 [0.5186, 0.5426]	0.2587 [0.2515, 0.2660]
LLaVA	0.3590 [0.3403, 0.3767]	0.3189 [0.3018, 0.3353]	0.2924 [0.2765, 0.3080]	0.2703 [0.2553, 0.2853]	0.5626 [0.5491, 0.5762]	0.2610 [0.2534, 0.2688]

Table B17: **Natural language generation metrics on the Inspect dataset** (n = 1,000 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.4844 [0.4587, 0.5109]	0.4808 [0.4565, 0.5055]	0.4825* [0.4598, 0.5046]
Gemini-3	0.2693 [0.2482, 0.2895]	0.4153* [0.3888, 0.4422]	0.3267 [0.3050, 0.3471]
Qwen3-VL-8B	0.2690 [0.2438, 0.2957]	0.2284 [0.2057, 0.2524]	0.2471 [0.2246, 0.2703]
HuluMed-32B	0.1487 [0.1370, 0.1601]	0.4004 [0.3742, 0.4271]	0.2169 [0.2018, 0.2317]
HuluMed-14B	0.2180 [0.1970, 0.2406]	0.2323 [0.2095, 0.2556]	0.2249 [0.2049, 0.2456]
HuluMed-7B	0.1678 [0.1494, 0.1870]	0.2107 [0.1883, 0.2337]	0.1868 [0.1678, 0.2057]
M3D	0.1883 [0.1652, 0.2109]	0.1458 [0.1277, 0.1645]	0.1644 [0.1449, 0.1843]
RadFM	0.1814 [0.1541, 0.2098]	0.1040 [0.0879, 0.1212]	0.1322 [0.1124, 0.1520]
Lingshu-32B	0.2367 [0.2123, 0.2617]	0.1909 [0.1715, 0.2119]	0.2114 [0.1909, 0.2326]
Lingshu-7B	0.1681 [0.1436, 0.1945]	0.1080 [0.0921, 0.1259]	0.1315 [0.1128, 0.1519]
MedGemma-27B	0.2198 [0.1608, 0.2805]	0.0270 [0.0185, 0.0366]	0.0481 [0.0334, 0.0647]
MedGemma-4B	0.1939 [0.1346, 0.2549]	0.0216 [0.0146, 0.0293]	0.0389 [0.0264, 0.0522]
R2GenGPT	0.4414 [0.4123, 0.4699]	0.3943 [0.3697, 0.4198]	0.4165 [0.3925, 0.4403]
LLaVA	0.4926 [0.4601, 0.5244]	0.3376 [0.3129, 0.3629]	0.4006 [0.3743, 0.4255]

Table B18: **RadBERT classification performance (micro Precision, Recall, F1) on the Inspect dataset** (n = 1,000 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.3305* [0.3202, 0.3408]	0.3601* [0.3450, 0.3747]	0.4449* [0.4296, 0.4595]	0.3793* [0.3620, 0.3969]	0.4028* [0.3852, 0.4190]	0.3968* [0.3836, 0.4091]
Gemini-3	0.2335 [0.2286, 0.2386]	0.1817 [0.1707, 0.1932]	0.2519 [0.2402, 0.2633]	0.2285 [0.2134, 0.2427]	0.2074 [0.1936, 0.2222]	0.2174 [0.2078, 0.2275]
Qwen3-VL-8B	0.2140 [0.2083, 0.2196]	0.1452 [0.1334, 0.1570]	0.2025 [0.1893, 0.2148]	0.1667 [0.1517, 0.1809]	0.1767 [0.1616, 0.1908]	0.1728 [0.1617, 0.1826]
HuluMed-32B	0.2189 [0.2144, 0.2233]	0.1819 [0.1699, 0.1934]	0.2457 [0.2337, 0.2570]	0.1711 [0.1580, 0.1849]	0.2135 [0.1972, 0.2291]	0.2024 [0.1920, 0.2122]
HuluMed-14B	0.2082 [0.2028, 0.2136]	0.1546 [0.1422, 0.1672]	0.2144 [0.2015, 0.2281]	0.1490 [0.1354, 0.1625]	0.1612 [0.1464, 0.1757]	0.1698 [0.1584, 0.1805]
HuluMed-7B	0.2018 [0.1969, 0.2066]	0.1364 [0.1256, 0.1474]	0.1843 [0.1718, 0.1965]	0.1167 [0.1055, 0.1288]	0.1284 [0.1154, 0.1425]	0.1415 [0.1316, 0.1514]
M3D	0.1946 [0.1896, 0.1995]	0.1262 [0.1144, 0.1380]	0.1914 [0.1786, 0.2043]	0.1008 [0.0896, 0.1129]	0.1139 [0.1021, 0.1262]	0.1331 [0.1240, 0.1421]
RadFM	0.1920 [0.1875, 0.1963]	0.1187 [0.1052, 0.1323]	0.1439 [0.1296, 0.1584]	0.0910 [0.0775, 0.1045]	0.1040 [0.0906, 0.1179]	0.1144 [0.1035, 0.1258]
Lingshu-32B	0.2234 [0.2190, 0.2279]	0.1203 [0.1093, 0.1319]	0.2262 [0.2152, 0.2367]	0.1413 [0.1304, 0.1522]	0.1629 [0.1503, 0.1756]	0.1627 [0.1540, 0.1714]
Lingshu-7B	0.1751 [0.1703, 0.1801]	0.0886 [0.0787, 0.0982]	0.1038 [0.0914, 0.1164]	0.0699 [0.0589, 0.0812]	0.0626 [0.0522, 0.0738]	0.0812 [0.0724, 0.0900]
MedGemma-27B	0.1636 [0.1593, 0.1677]	0.0386 [0.0302, 0.0473]	0.0859 [0.0747, 0.0967]	0.0473 [0.0377, 0.0570]	0.0491 [0.0389, 0.0593]	0.0552 [0.0468, 0.0631]
MedGemma-4B	0.1443 [0.1403, 0.1481]	0.0311 [0.0231, 0.0397]	0.0580 [0.0461, 0.0699]	0.0340 [0.0258, 0.0437]	0.0317 [0.0231, 0.0407]	0.0387 [0.0312, 0.0464]
R2GenGPT	0.2907 [0.2805, 0.3030]	0.2830 [0.2695, 0.2972]	0.3719 [0.3560, 0.3870]	0.3139 [0.2976, 0.3310]	0.3130 [0.2975, 0.3294]	0.3204 [0.3088, 0.3326]
LLaVA	0.2584 [0.2487, 0.2673]	0.2476 [0.2320, 0.2630]	0.3278 [0.3088, 0.3462]	0.2800 [0.2608, 0.2989]	0.2716 [0.2542, 0.2887]	0.2818 [0.2670, 0.2960]

Table B19: **Fine-grained caption metrics (Rate-Score, FORTE) on the Inspect dataset** (n = 1,000 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	BLEU-1 [95% CI]	BLEU-2 [95% CI]	BLEU-3 [95% CI]	BLEU-4 [95% CI]	ROUGE-L [95% CI]	METEOR [95% CI]
Astra	0.4020* [0.3914, 0.4117]	0.3032* [0.2941, 0.3116]	0.2536* [0.2456, 0.2612]	0.2206* [0.2132, 0.2278]	0.3988 [0.3915, 0.4058]	0.1968* [0.1922, 0.2013]
Gemini-3	0.2749 [0.2619, 0.2879]	0.2196 [0.2089, 0.2303]	0.1909 [0.1813, 0.2004]	0.1699 [0.1613, 0.1786]	0.3903 [0.3811, 0.3995]	0.1722 [0.1679, 0.1767]
Qwen3-VL-8B	0.3272 [0.3163, 0.3376]	0.2545 [0.2455, 0.2633]	0.2198 [0.2115, 0.2279]	0.1957 [0.1881, 0.2032]	0.3655 [0.3570, 0.3747]	0.1695 [0.1654, 0.1737]
HuluMed-32B	0.3460 [0.3365, 0.3552]	0.2613 [0.2536, 0.2687]	0.2195 [0.2127, 0.2263]	0.1907 [0.1845, 0.1971]	0.3370 [0.3306, 0.3433]	0.1700 [0.1664, 0.1737]
HuluMed-14B	0.3466 [0.3371, 0.3558]	0.2615 [0.2539, 0.2690]	0.2197 [0.2128, 0.2265]	0.1908 [0.1846, 0.1972]	0.3374 [0.3307, 0.3435]	0.1702 [0.1665, 0.1739]
HuluMed-7B	0.3470 [0.3377, 0.3560]	0.2618 [0.2542, 0.2693]	0.2199 [0.2131, 0.2267]	0.1911 [0.1848, 0.1974]	0.3369 [0.3304, 0.3432]	0.1703 [0.1667, 0.1740]
M3D	0.2957 [0.2831, 0.3079]	0.2346 [0.2241, 0.2448]	0.2038 [0.1944, 0.2131]	0.1816 [0.1731, 0.1901]	0.3590 [0.3519, 0.3662]	0.1690 [0.1648, 0.1734]
RadFM	0.2496 [0.2360, 0.2635]	0.2063 [0.1946, 0.2179]	0.1832 [0.1727, 0.1937]	0.1653 [0.1556, 0.1749]	0.3838 [0.3754, 0.3918]	0.1698 [0.1653, 0.1744]
Lingshu-32B	0.3435 [0.3347, 0.3522]	0.2622 [0.2546, 0.2698]	0.2228 [0.2158, 0.2299]	0.1960 [0.1894, 0.2025]	0.3340 [0.3274, 0.3405]	0.1687 [0.1650, 0.1724]
Lingshu-7B	0.1578 [0.1445, 0.1709]	0.1372 [0.1255, 0.1485]	0.1252 [0.1146, 0.1355]	0.1154 [0.1056, 0.1249]	0.4188 [0.4088, 0.4286]	0.1702 [0.1655, 0.1753]
MedGemma-27B	0.2247 [0.2083, 0.2403]	0.1909 [0.1775, 0.2035]	0.1721 [0.1601, 0.1836]	0.1577 [0.1467, 0.1682]	0.4037 [0.3932, 0.4139]	0.1696 [0.1648, 0.1744]
MedGemma-4B	0.1141 [0.1035, 0.1253]	0.1028 [0.0932, 0.1131]	0.0956 [0.0865, 0.1053]	0.0894 [0.0808, 0.0985]	0.4350 [0.4243, 0.4454]	0.1732 [0.1681, 0.1785]
R2GenGPT	0.2116 [0.1960, 0.2280]	0.1790 [0.1656, 0.1926]	0.1613 [0.1493, 0.1736]	0.1478 [0.1368, 0.1593]	0.4433 [0.4327, 0.4530]	0.1835 [0.1783, 0.1889]
LLaVA	0.3224 [0.3080, 0.3357]	0.2567 [0.2452, 0.2677]	0.2220 [0.2118, 0.2317]	0.1974 [0.1883, 0.2062]	0.4146 [0.4057, 0.4239]	0.1804 [0.1752, 0.1855]

Table B20: **Natural language generation metrics on the BIMCV dataset** (n = 1,505 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.3572* [0.3373, 0.3774]	0.3405* [0.3222, 0.3578]	0.3486* [0.3305, 0.3652]
Gemini-3	0.2884 [0.2688, 0.3098]	0.2419 [0.2251, 0.2595]	0.2631 [0.2461, 0.2816]
Qwen3-VL-8B	0.2449 [0.2207, 0.2698]	0.1217 [0.1075, 0.1355]	0.1626 [0.1450, 0.1792]
HuluMed-32B	0.1800 [0.1614, 0.1994]	0.1241 [0.1103, 0.1381]	0.1469 [0.1319, 0.1616]
HuluMed-14B	0.1790 [0.1606, 0.1980]	0.1233 [0.1097, 0.1370]	0.1460 [0.1311, 0.1607]
HuluMed-7B	0.1809 [0.1629, 0.1999]	0.1244 [0.1105, 0.1382]	0.1475 [0.1323, 0.1624]
M3D	0.1650 [0.1460, 0.1836]	0.0905 [0.0797, 0.1015]	0.1169 [0.1037, 0.1298]
RadFM	0.1830 [0.1607, 0.2057]	0.0827 [0.0717, 0.0944]	0.1139 [0.0994, 0.1287]
Lingshu-32B	0.1691 [0.1499, 0.1902]	0.0858 [0.0750, 0.0978]	0.1138 [0.1001, 0.1287]
Lingshu-7B	0.1777 [0.1491, 0.2067]	0.0447 [0.0370, 0.0530]	0.0714 [0.0594, 0.0839]
MedGemma-27B	0.1792 [0.1293, 0.2335]	0.0140 [0.0096, 0.0194]	0.0260 [0.0178, 0.0358]
MedGemma-4B	0.2429 [0.1704, 0.3158]	0.0126 [0.0083, 0.0175]	0.0239 [0.0159, 0.0329]
R2GenGPT	0.2942 [0.2717, 0.3167]	0.1795 [0.1641, 0.1960]	0.2229 [0.2054, 0.2406]
LLaVA	0.2694 [0.2493, 0.2885]	0.2013 [0.1853, 0.2165]	0.2304 [0.2137, 0.2462]

Table B21: **RadBERT classification performance (micro Precision, Recall, F1) on the BIMCV dataset** (n = 1,505 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.2624* [0.2567, 0.2684]	0.4133* [0.4023, 0.4240]	0.3783* [0.3663, 0.3897]	0.2755* [0.2649, 0.2863]	0.3448* [0.3294, 0.3595]	0.3530* [0.3441, 0.3620]
Gemini-3	0.2076 [0.2035, 0.2115]	0.2286 [0.2184, 0.2398]	0.2066 [0.1957, 0.2167]	0.1586 [0.1482, 0.1686]	0.1596 [0.1471, 0.1720]	0.1884 [0.1802, 0.1963]
Qwen3-VL-8B	0.1954 [0.1918, 0.1990]	0.1683 [0.1586, 0.1786]	0.1669 [0.1572, 0.1770]	0.1041 [0.0957, 0.1127]	0.1043 [0.0940, 0.1141]	0.1359 [0.1287, 0.1429]
HuluMed-32B	0.1904 [0.1870, 0.1939]	0.1728 [0.1632, 0.1833]	0.1747 [0.1643, 0.1849]	0.0836 [0.0767, 0.0906]	0.0814 [0.0712, 0.0910]	0.1281 [0.1214, 0.1350]
HuluMed-14B	0.1904 [0.1870, 0.1939]	0.1730 [0.1630, 0.1836]	0.1753 [0.1647, 0.1855]	0.0832 [0.0762, 0.0905]	0.0819 [0.0717, 0.0913]	0.1284 [0.1215, 0.1353]
HuluMed-7B	0.1907 [0.1873, 0.1941]	0.1702 [0.1603, 0.1805]	0.1736 [0.1631, 0.1841]	0.0832 [0.0762, 0.0902]	0.0817 [0.0715, 0.0911]	0.1272 [0.1204, 0.1339]
M3D	0.1803 [0.1772, 0.1834]	0.1755 [0.1651, 0.1860]	0.1552 [0.1458, 0.1647]	0.0665 [0.0598, 0.0730]	0.0547 [0.0463, 0.0637]	0.1130 [0.1066, 0.1189]
RadFM	0.1769 [0.1739, 0.1798]	0.1280 [0.1187, 0.1378]	0.1153 [0.1060, 0.1253]	0.0729 [0.0652, 0.0806]	0.0468 [0.0385, 0.0548]	0.0907 [0.0846, 0.0971]
Lingshu-32B	0.2010 [0.1975, 0.2046]	0.1571 [0.1481, 0.1665]	0.1709 [0.1620, 0.1797]	0.0739 [0.0672, 0.0804]	0.0780 [0.0690, 0.0877]	0.1200 [0.1140, 0.1262]
Lingshu-7B	0.1598 [0.1566, 0.1630]	0.0927 [0.0840, 0.1017]	0.0766 [0.0679, 0.0855]	0.0394 [0.0340, 0.0453]	0.0448 [0.0361, 0.0534]	0.0634 [0.0577, 0.0693]
MedGemma-27B	0.1479 [0.1453, 0.1506]	0.0571 [0.0492, 0.0650]	0.0500 [0.0428, 0.0572]	0.0239 [0.0190, 0.0293]	0.0277 [0.0211, 0.0347]	0.0397 [0.0345, 0.0449]
MedGemma-4B	0.1400 [0.1374, 0.1428]	0.0417 [0.0346, 0.0495]	0.0508 [0.0427, 0.0596]	0.0175 [0.0130, 0.0226]	0.0165 [0.0113, 0.0224]	0.0317 [0.0267, 0.0368]
R2GenGPT	0.1823 [0.1782, 0.1862]	0.1947 [0.1813, 0.2086]	0.1752 [0.1635, 0.1870]	0.1284 [0.1170, 0.1400]	0.1856 [0.1694, 0.2029]	0.1710 [0.1601, 0.1821]
LLaVA	0.2189 [0.2143, 0.2235]	0.2636 [0.2525, 0.2749]	0.2513 [0.2396, 0.2632]	0.1430 [0.1333, 0.1530]	0.1970 [0.1817, 0.2114]	0.2137 [0.2052, 0.2225]

Table B22: **Fine-grained caption metrics (Rate-Score, FORTE) on the BIMCV dataset** (n = 1,505 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.6618 [0.5781, 0.7379]	0.3203* [0.2684, 0.3742]	0.4317* [0.3718, 0.4907]
Gemini-3	0.5529 [0.4456, 0.6566]	0.1673 [0.1237, 0.2098]	0.2568 [0.1954, 0.3137]
Qwen3-VL-8B	0.4000 [0.2667, 0.5349]	0.0702 [0.0418, 0.1003]	0.1194 [0.0734, 0.1667]
HuluMed-32B	0.5676 [0.4890, 0.6457]	0.2989 [0.2474, 0.3522]	0.3916 [0.3333, 0.4485]
HuluMed-14B	0.5060 [0.4024, 0.6087]	0.1495 [0.1079, 0.1937]	0.2308 [0.1717, 0.2881]
HuluMed-7B	0.4615 [0.3414, 0.5902]	0.1068 [0.0722, 0.1444]	0.1734 [0.1212, 0.2274]
M3D	0.4121 [0.3403, 0.4850]	0.2669 [0.2162, 0.3185]	0.3240 [0.2685, 0.3782]
RadFM	0.4483 [0.2727, 0.6471]	0.0463 [0.0238, 0.0727]	0.0839 [0.0444, 0.1286]
Lingshu-32B	0.5062 [0.3944, 0.6104]	0.1439 [0.1018, 0.1851]	0.2240 [0.1636, 0.2798]
Lingshu-7B	0.6154 [0.5195, 0.7093]	0.1993 [0.1536, 0.2455]	0.3011 [0.2402, 0.3602]
MedGemma-27B	0.6250 [0.4000, 0.8336]	0.0356 [0.0146, 0.0590]	0.0673 [0.0282, 0.1100]
MedGemma-4B	0.8000 [0.5000, 1.0000]	0.0285 [0.0107, 0.0491]	0.0550 [0.0211, 0.0930]

Table B23: **RadBERT classification performance (micro Precision, Recall, F1) on the external CTRG-Chest dataset** ($n = 324$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.2434 [0.2348, 0.2525]	0.2550* [0.2324, 0.2783]	0.3918* [0.3690, 0.4131]	0.1408 [0.1195, 0.1634]	0.1346 [0.1063, 0.1649]	0.2306* [0.2151, 0.2467]
Gemini-3	0.2386 [0.2284, 0.2489]	0.2326 [0.2096, 0.2542]	0.2693 [0.2461, 0.2897]	0.1072 [0.0844, 0.1302]	0.0962 [0.0679, 0.1256]	0.1763 [0.1582, 0.1938]
Qwen3-VL-8B	0.2149 [0.2055, 0.2243]	0.1483 [0.1258, 0.1710]	0.2199 [0.1944, 0.2443]	0.0792 [0.0582, 0.1008]	0.0588 [0.0344, 0.0861]	0.1265 [0.1080, 0.1443]
HuluMed-32B	0.2187 [0.2102, 0.2279]	0.2056 [0.1819, 0.2295]	0.2660 [0.2376, 0.2929]	0.1466 [0.1245, 0.1705]	0.1771 [0.1418, 0.2129]	0.1988 [0.1801, 0.2172]
HuluMed-14B	0.2148 [0.2051, 0.2250]	0.1608 [0.1373, 0.1873]	0.2148 [0.1897, 0.2393]	0.1033 [0.0763, 0.1317]	0.1216 [0.0851, 0.1600]	0.1501 [0.1276, 0.1742]
HuluMed-7B	0.2054 [0.1977, 0.2138]	0.1620 [0.1352, 0.1874]	0.2085 [0.1850, 0.2327]	0.0663 [0.0496, 0.0836]	0.0767 [0.0475, 0.1070]	0.1284 [0.1102, 0.1467]
M3D	0.2200 [0.2116, 0.2286]	0.1925 [0.1701, 0.2147]	0.2466 [0.2230, 0.2704]	0.1019 [0.0814, 0.1229]	0.1183 [0.0860, 0.1539]	0.1648 [0.1467, 0.1833]
RadFM	0.1997 [0.1947, 0.2047]	0.0166 [0.0086, 0.0254]	0.0525 [0.0368, 0.0682]	0.0307 [0.0163, 0.0459]	0.0220 [0.0051, 0.0442]	0.0304 [0.0198, 0.0420]
Lingshu-32B	0.2380 [0.2295, 0.2473]	0.1656 [0.1474, 0.1843]	0.2873 [0.2662, 0.3083]	0.0969 [0.0777, 0.1157]	0.1178 [0.0855, 0.1510]	0.1669 [0.1509, 0.1827]
Lingshu-7B	0.1977 [0.1889, 0.2068]	0.1300 [0.1063, 0.1554]	0.1566 [0.1292, 0.1842]	0.1085 [0.0862, 0.1326]	0.1996 [0.1521, 0.2495]	0.1486 [0.1239, 0.1742]
MedGemma-27B	0.1554 [0.1495, 0.1619]	0.0408 [0.0216, 0.0628]	0.0562 [0.0335, 0.0829]	0.0322 [0.0156, 0.0521]	0.0400 [0.0155, 0.0686]	0.0423 [0.0247, 0.0629]
MedGemma-4B	0.1586 [0.1533, 0.1640]	0.0331 [0.0181, 0.0506]	0.0624 [0.0422, 0.0854]	0.0244 [0.0118, 0.0380]	0.0119 [0.0000, 0.0297]	0.0329 [0.0219, 0.0462]

Table B24: **Fine-grained caption metrics (Rate-Score, FORTE) on the external CTRG-Chest dataset** ($n = 324$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.4051 [0.3590, 0.4534]	0.4211* [0.3658, 0.4744]	0.4129* [0.3694, 0.4569]
Gemini-3	0.3387 [0.2203, 0.4584]	0.0553 [0.0332, 0.0788]	0.0950 [0.0582, 0.1339]
Qwen3-VL-8B	0.1389 [0.0861, 0.1953]	0.0526 [0.0312, 0.0755]	0.0763 [0.0463, 0.1081]
HuluMed-32B	0.2949 [0.2335, 0.3580]	0.1684 [0.1305, 0.2055]	0.2144 [0.1696, 0.2573]
HuluMed-14B	0.2075 [0.1594, 0.2651]	0.1158 [0.0879, 0.1470]	0.1486 [0.1146, 0.1868]
HuluMed-7B	0.3392 [0.2697, 0.4214]	0.1526 [0.1191, 0.1893]	0.2105 [0.1667, 0.2575]
M3D	0.2230 [0.1600, 0.2883]	0.0868 [0.0598, 0.1156]	0.1250 [0.0880, 0.1621]
RadFM	0.1724 [0.0435, 0.3333]	0.0132 [0.0027, 0.0261]	0.0244 [0.0050, 0.0481]
Lingshu-32B	0.1985 [0.1357, 0.2696]	0.0684 [0.0452, 0.0944]	0.1018 [0.0672, 0.1387]
Lingshu-7B	0.4815 [0.3000, 0.6667]	0.0342 [0.0183, 0.0526]	0.0639 [0.0347, 0.0964]
MedGemma-27B	0.4286 [0.1538, 0.6923]	0.0158 [0.0050, 0.0294]	0.0305 [0.0097, 0.0557]
MedGemma-4B	0.4070 [0.3068, 0.5165]	0.0921 [0.0643, 0.1250]	0.1502 [0.1071, 0.1982]

Table B25: **RadBERT classification performance (micro Precision, Recall, F1) on the external AMOS-MM dataset** (n = 400 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.2507* [0.2449, 0.2561]	0.2880* [0.2684, 0.3070]	0.3084* [0.2904, 0.3258]	0.1973* [0.1807, 0.2138]	0.1036* [0.0826, 0.1258]	0.2243* [0.2117, 0.2370]
Gemini-3	0.1847 [0.1799, 0.1897]	0.0752 [0.0578, 0.0930]	0.0977 [0.0816, 0.1134]	0.0680 [0.0536, 0.0819]	0.0429 [0.0246, 0.0632]	0.0710 [0.0584, 0.0833]
Qwen3-VL-8B	0.1872 [0.1826, 0.1920]	0.0845 [0.0683, 0.1012]	0.0895 [0.0770, 0.1017]	0.0565 [0.0453, 0.0680]	0.0240 [0.0128, 0.0373]	0.0636 [0.0543, 0.0728]
HuluMed-32B	0.1984 [0.1932, 0.2040]	0.1662 [0.1451, 0.1886]	0.1619 [0.1420, 0.1820]	0.1154 [0.0993, 0.1322]	0.0253 [0.0082, 0.0466]	0.1172 [0.1024, 0.1320]
HuluMed-14B	0.1894 [0.1843, 0.1947]	0.1521 [0.1312, 0.1720]	0.1284 [0.1118, 0.1459]	0.0903 [0.0742, 0.1073]	0.0282 [0.0108, 0.0473]	0.0997 [0.0866, 0.1133]
HuluMed-7B	0.1987 [0.1932, 0.2043]	0.1518 [0.1291, 0.1739]	0.1520 [0.1329, 0.1724]	0.0949 [0.0797, 0.1113]	0.0365 [0.0200, 0.0556]	0.1088 [0.0952, 0.1236]
M3D	0.1877 [0.1835, 0.1919]	0.1021 [0.0840, 0.1209]	0.1047 [0.0897, 0.1194]	0.0726 [0.0602, 0.0860]	0.0238 [0.0080, 0.0426]	0.0758 [0.0650, 0.0871]
RadFM	0.1745 [0.1697, 0.1796]	0.0763 [0.0604, 0.0928]	0.0741 [0.0590, 0.0900]	0.0524 [0.0406, 0.0642]	0.0132 [0.0000, 0.0306]	0.0540 [0.0440, 0.0647]
Lingshu-32B	0.1846 [0.1804, 0.1887]	0.0672 [0.0541, 0.0818]	0.0928 [0.0777, 0.1085]	0.0621 [0.0510, 0.0738]	0.0245 [0.0103, 0.0396]	0.0617 [0.0520, 0.0715]
Lingshu-7B	0.1614 [0.1579, 0.1650]	0.0293 [0.0170, 0.0424]	0.0337 [0.0234, 0.0448]	0.0131 [0.0072, 0.0197]	0.0135 [0.0000, 0.0301]	0.0224 [0.0146, 0.0317]
MedGemma-27B	0.1764 [0.1722, 0.1804]	0.0329 [0.0223, 0.0438]	0.0779 [0.0636, 0.0914]	0.0436 [0.0329, 0.0542]	0.0164 [0.0028, 0.0320]	0.0427 [0.0338, 0.0518]
MedGemma-4B	0.1800 [0.1762, 0.1837]	0.0325 [0.0223, 0.0430]	0.0860 [0.0699, 0.1034]	0.0654 [0.0522, 0.0798]	0.0249 [0.0081, 0.0451]	0.0522 [0.0415, 0.0632]

Table B26: **Fine-grained caption metrics (Rate-Score, FORTE) on the external AMOS-MM dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.5675 [0.5361, 0.6002]	0.4355* [0.4094, 0.4615]	0.4928* [0.4679, 0.5176]
Gemini-3	0.2158 [0.1834, 0.2508]	0.1076 [0.0908, 0.1267]	0.1436 [0.1219, 0.1678]
Qwen3-VL-8B	0.3049 [0.2510, 0.3622]	0.0601 [0.0451, 0.0762]	0.1005 [0.0771, 0.1244]
HuluMed-32B	0.3467 [0.3153, 0.3778]	0.2638 [0.2358, 0.2911]	0.2996 [0.2728, 0.3265]
HuluMed-14B	0.3298 [0.3024, 0.3567]	0.2999 [0.2710, 0.3288]	0.3142 [0.2877, 0.3394]
HuluMed-7B	0.3488 [0.3089, 0.3867]	0.1508 [0.1274, 0.1740]	0.2105 [0.1818, 0.2378]
M3D	0.3481 [0.3054, 0.3907]	0.1323 [0.1124, 0.1537]	0.1917 [0.1664, 0.2187]
RadFM	0.6170 [0.5538, 0.6869]	0.0930 [0.0782, 0.1094]	0.1617 [0.1380, 0.1876]
Lingshu-32B	0.2735 [0.2136, 0.3391]	0.0489 [0.0366, 0.0618]	0.0830 [0.0629, 0.1040]
Lingshu-7B	0.2743 [0.2120, 0.3394]	0.0497 [0.0371, 0.0628]	0.0842 [0.0632, 0.1056]
MedGemma-27B	0.3529 [0.1429, 0.5385]	0.0048 [0.0008, 0.0094]	0.0095 [0.0016, 0.0184]
MedGemma-4B	0.3571 [0.1818, 0.5263]	0.0080 [0.0034, 0.0130]	0.0157 [0.0066, 0.0253]

Table B27: **RadBERT classification performance (micro Precision, Recall, F1) on the external Inhouse-Chest-1 dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.2874* [0.2792, 0.2957]	0.2745* [0.2605, 0.2890]	0.4805* [0.4647, 0.4957]	0.2641* [0.2489, 0.2788]	0.3227* [0.2888, 0.3562]	0.3355* [0.3231, 0.3473]
Gemini-3	0.2275 [0.2219, 0.2334]	0.2086 [0.1937, 0.2237]	0.2678 [0.2522, 0.2839]	0.1006 [0.0866, 0.1154]	0.0925 [0.0690, 0.1167]	0.1674 [0.1555, 0.1796]
Qwen3-VL-8B	0.1985 [0.1921, 0.2059]	0.1060 [0.0903, 0.1212]	0.1597 [0.1404, 0.1796]	0.0647 [0.0511, 0.0796]	0.0588 [0.0371, 0.0829]	0.0973 [0.0833, 0.1117]
HuluMed-32B	0.2313 [0.2238, 0.2389]	0.1891 [0.1737, 0.2047]	0.2863 [0.2623, 0.3095]	0.1656 [0.1498, 0.1810]	0.1717 [0.1437, 0.1970]	0.2032 [0.1885, 0.2166]
HuluMed-14B	0.2440 [0.2367, 0.2514]	0.2086 [0.1950, 0.2233]	0.3118 [0.2895, 0.3331]	0.1845 [0.1695, 0.1995]	0.1782 [0.1520, 0.2039]	0.2208 [0.2072, 0.2343]
HuluMed-7B	0.2156 [0.2089, 0.2227]	0.1496 [0.1314, 0.1680]	0.2366 [0.2158, 0.2560]	0.1265 [0.1096, 0.1432]	0.1324 [0.1056, 0.1588]	0.1613 [0.1440, 0.1768]
M3D	0.2037 [0.1985, 0.2093]	0.1456 [0.1299, 0.1611]	0.2111 [0.1927, 0.2298]	0.0987 [0.0832, 0.1146]	0.0986 [0.0723, 0.1258]	0.1385 [0.1248, 0.1527]
RadFM	0.1953 [0.1904, 0.2006]	0.0826 [0.0700, 0.0962]	0.1806 [0.1620, 0.1988]	0.1004 [0.0858, 0.1151]	0.1572 [0.1215, 0.1934]	0.1302 [0.1165, 0.1441]
Lingshu-32B	0.1741 [0.1697, 0.1787]	0.0798 [0.0690, 0.0914]	0.0653 [0.0524, 0.0785]	0.0238 [0.0163, 0.0314]	0.0326 [0.0165, 0.0499]	0.0504 [0.0419, 0.0595]
Lingshu-7B	0.1735 [0.1693, 0.1782]	0.0786 [0.0679, 0.0896]	0.0650 [0.0522, 0.0779]	0.0238 [0.0163, 0.0314]	0.0325 [0.0165, 0.0497]	0.0500 [0.0415, 0.0589]
MedGemma-27B	0.1541 [0.1500, 0.1586]	0.0125 [0.0067, 0.0193]	0.0217 [0.0134, 0.0313]	0.0083 [0.0026, 0.0149]	0.0109 [0.0000, 0.0286]	0.0134 [0.0071, 0.0213]
MedGemma-4B	0.1672 [0.1624, 0.1725]	0.0380 [0.0297, 0.0470]	0.0584 [0.0454, 0.0715]	0.0220 [0.0143, 0.0303]	0.0182 [0.0037, 0.0344]	0.0341 [0.0261, 0.0422]

Table B28: **Fine-grained caption metrics (Rate-Score, FORTE) on the external Inhouse-Chest-1 dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.2453 [0.2263, 0.2654]	0.5239* [0.4840, 0.5626]	0.3342* [0.3109, 0.3575]
Gemini-3	0.0522 [0.0281, 0.0808]	0.0279 [0.0143, 0.0439]	0.0364 [0.0190, 0.0560]
Qwen3-VL-8B	0.0095 [0.0019, 0.0183]	0.0100 [0.0020, 0.0192]	0.0097 [0.0020, 0.0189]
HuluMed-32B	0.2165 [0.1816, 0.2511]	0.1833 [0.1508, 0.2175]	0.1985 [0.1654, 0.2304]
HuluMed-14B	0.0870 [0.0629, 0.1134]	0.0837 [0.0609, 0.1078]	0.0853 [0.0620, 0.1097]
HuluMed-7B	0.1119 [0.0842, 0.1415]	0.0956 [0.0720, 0.1201]	0.1031 [0.0779, 0.1286]
M3D	0.0914 [0.0640, 0.1190]	0.0737 [0.0509, 0.0967]	0.0816 [0.0570, 0.1061]
RadFM	0.0503 [0.0227, 0.0798]	0.0179 [0.0079, 0.0306]	0.0264 [0.0119, 0.0435]
Lingshu-32B	0.0603 [0.0382, 0.0842]	0.0478 [0.0294, 0.0672]	0.0533 [0.0333, 0.0747]
Lingshu-7B	0.0830 [0.0506, 0.1216]	0.0398 [0.0235, 0.0589]	0.0538 [0.0321, 0.0790]
MedGemma-27B	0.2105 [0.0556, 0.4000]	0.0080 [0.0019, 0.0168]	0.0154 [0.0037, 0.0323]
MedGemma-4B	0.4762 [0.4174, 0.5397]	0.1992 [0.1651, 0.2343]	0.2809 [0.2397, 0.3234]

Table B29: **RadBERT classification performance (micro Precision, Recall, F1) on the external Inhouse-Abdomen-1 dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.2509* [0.2452, 0.2564]	0.3029* [0.2883, 0.3185]	0.3500* [0.3345, 0.3664]	0.2169* [0.2020, 0.2308]	0.0631 [0.0471, 0.0807]	0.2332* [0.2243, 0.2422]
Gemini-3	0.1659 [0.1619, 0.1697]	0.0391 [0.0285, 0.0508]	0.0558 [0.0418, 0.0703]	0.0322 [0.0222, 0.0420]	0.0141 [0.0027, 0.0296]	0.0353 [0.0269, 0.0440]
Qwen3-VL-8B	0.1699 [0.1659, 0.1739]	0.0405 [0.0301, 0.0523]	0.0631 [0.0532, 0.0739]	0.0285 [0.0220, 0.0357]	0.0000 [0.0000, 0.0000]	0.0331 [0.0278, 0.0390]
HuluMed-32B	0.2035 [0.1973, 0.2098]	0.1684 [0.1492, 0.1885]	0.1722 [0.1530, 0.1918]	0.1396 [0.1209, 0.1583]	0.0598 [0.0361, 0.0850]	0.1350 [0.1202, 0.1504]
HuluMed-14B	0.1626 [0.1586, 0.1669]	0.0502 [0.0369, 0.0640]	0.0546 [0.0430, 0.0670]	0.0414 [0.0298, 0.0531]	0.0134 [0.0000, 0.0312]	0.0399 [0.0308, 0.0497]
HuluMed-7B	0.1853 [0.1806, 0.1900]	0.1032 [0.0862, 0.1212]	0.1330 [0.1164, 0.1495]	0.0780 [0.0641, 0.0920]	0.0365 [0.0174, 0.0588]	0.0877 [0.0757, 0.1002]
M3D	0.1840 [0.1790, 0.1888]	0.0944 [0.0776, 0.1125]	0.1145 [0.0979, 0.1305]	0.0615 [0.0497, 0.0740]	0.0260 [0.0113, 0.0430]	0.0741 [0.0633, 0.0855]
RadFM	0.1513 [0.1477, 0.1549]	0.0133 [0.0061, 0.0217]	0.0268 [0.0178, 0.0357]	0.0266 [0.0169, 0.0370]	0.0155 [0.0000, 0.0332]	0.0205 [0.0123, 0.0299]
Lingshu-32B	0.1841 [0.1799, 0.1887]	0.0594 [0.0464, 0.0728]	0.0921 [0.0783, 0.1066]	0.0656 [0.0545, 0.0776]	0.0118 [0.0024, 0.0229]	0.0572 [0.0494, 0.0654]
Lingshu-7B	0.1605 [0.1572, 0.1639]	0.0247 [0.0159, 0.0344]	0.0474 [0.0348, 0.0600]	0.0174 [0.0103, 0.0257]	0.0000 [0.0000, 0.0000]	0.0224 [0.0169, 0.0287]
MedGemma-27B	0.1871 [0.1818, 0.1924]	0.0315 [0.0200, 0.0441]	0.1184 [0.1021, 0.1348]	0.0373 [0.0263, 0.0489]	0.0252 [0.0000, 0.0580]	0.0531 [0.0423, 0.0647]
MedGemma-4B	0.1876 [0.1835, 0.1918]	0.0463 [0.0351, 0.0583]	0.1456 [0.1256, 0.1646]	0.0901 [0.0755, 0.1050]	0.0218 [0.0071, 0.0391]	0.0760 [0.0653, 0.0865]

Table B30: **Fine-grained caption metrics (Rate-Score, FORTE) on the external Inhouse-Abdomen-1 dataset** (n = 400 cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.3659* [0.3437, 0.3885]	0.7599* [0.7204, 0.7995]	0.4939* [0.4706, 0.5167]
Gemini-3	0.1491 [0.1029, 0.1994]	0.1119 [0.0769, 0.1502]	0.1278 [0.0884, 0.1690]
Qwen3-VL-8B	0.0569 [0.0348, 0.0797]	0.0723 [0.0438, 0.1002]	0.0637 [0.0385, 0.0886]
HuluMed-32B	0.1890 [0.1527, 0.2267]	0.1515 [0.1189, 0.1860]	0.1682 [0.1342, 0.2020]
HuluMed-14B	0.1650 [0.1316, 0.1978]	0.1538 [0.1190, 0.1879]	0.1592 [0.1258, 0.1899]
HuluMed-7B	0.1178 [0.0859, 0.1503]	0.1049 [0.0762, 0.1353]	0.1110 [0.0813, 0.1410]
M3D	0.1278 [0.0974, 0.1624]	0.1212 [0.0925, 0.1517]	0.1244 [0.0951, 0.1562]
RadFM	0.0508 [0.0213, 0.0860]	0.0210 [0.0091, 0.0341]	0.0297 [0.0128, 0.0486]
Lingshu-32B	0.1656 [0.1278, 0.2046]	0.1235 [0.0939, 0.1555]	0.1415 [0.1093, 0.1739]
Lingshu-7B	0.1176 [0.0714, 0.1654]	0.0420 [0.0242, 0.0603]	0.0619 [0.0362, 0.0887]
MedGemma-27B	0.1768 [0.1355, 0.2128]	0.0207 [0.0136, 0.0299]	0.0371 [0.0247, 0.0524]
MedGemma-4B	0.2429 [0.1481, 0.3455]	0.0396 [0.0221, 0.0590]	0.0681 [0.0385, 0.0989]

Table B31: **RadBERT classification performance (micro Precision, Recall, F1) on the external Inhouse-Abdomen-2 dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.3157* [0.3097, 0.3216]	0.4701* [0.4538, 0.4869]	0.5270* [0.5125, 0.5415]	0.4234* [0.4078, 0.4403]	0.3871* [0.3562, 0.4175]	0.4519* [0.4426, 0.4613]
Gemini-3	0.1790 [0.1732, 0.1848]	0.0947 [0.0743, 0.1147]	0.1156 [0.0969, 0.1334]	0.0788 [0.0617, 0.0961]	0.0988 [0.0663, 0.1284]	0.0970 [0.0794, 0.1136]
Qwen3-VL-8B	0.1771 [0.1712, 0.1830]	0.0599 [0.0462, 0.0743]	0.1056 [0.0889, 0.1224]	0.0630 [0.0502, 0.0758]	0.0356 [0.0218, 0.0499]	0.0660 [0.0551, 0.0772]
HuluMed-32B	0.1967 [0.1897, 0.2045]	0.1482 [0.1246, 0.1727]	0.1810 [0.1583, 0.2033]	0.1435 [0.1193, 0.1689]	0.1042 [0.0745, 0.1367]	0.1442 [0.1239, 0.1649]
HuluMed-14B	0.1989 [0.1916, 0.2063]	0.1585 [0.1364, 0.1805]	0.1925 [0.1700, 0.2146]	0.1489 [0.1277, 0.1705]	0.1017 [0.0727, 0.1320]	0.1504 [0.1326, 0.1685]
HuluMed-7B	0.1989 [0.1925, 0.2054]	0.1323 [0.1128, 0.1515]	0.1856 [0.1646, 0.2058]	0.1181 [0.0998, 0.1376]	0.0907 [0.0645, 0.1167]	0.1317 [0.1164, 0.1466]
M3D	0.1802 [0.1747, 0.1857]	0.1002 [0.0816, 0.1201]	0.1132 [0.0953, 0.1320]	0.0714 [0.0567, 0.0866]	0.0449 [0.0263, 0.0660]	0.0824 [0.0698, 0.0959]
RadFM	0.1642 [0.1584, 0.1707]	0.0735 [0.0550, 0.0931]	0.0994 [0.0800, 0.1193]	0.0520 [0.0368, 0.0679]	0.0239 [0.0073, 0.0426]	0.0622 [0.0493, 0.0753]
Lingshu-32B	0.1843 [0.1793, 0.1898]	0.0525 [0.0393, 0.0670]	0.1466 [0.1258, 0.1670]	0.0744 [0.0594, 0.0898]	0.0493 [0.0311, 0.0681]	0.0807 [0.0689, 0.0929]
Lingshu-7B	0.1666 [0.1609, 0.1724]	0.0505 [0.0344, 0.0678]	0.0921 [0.0751, 0.1095]	0.0447 [0.0310, 0.0583]	0.0255 [0.0074, 0.0484]	0.0532 [0.0409, 0.0668]
MedGemma-27B	0.1554 [0.1515, 0.1597]	0.0084 [0.0019, 0.0166]	0.0371 [0.0250, 0.0505]	0.0202 [0.0091, 0.0323]	0.0087 [0.0000, 0.0213]	0.0186 [0.0109, 0.0272]
MedGemma-4B	0.1731 [0.1676, 0.1792]	0.0349 [0.0227, 0.0477]	0.1040 [0.0816, 0.1270]	0.0744 [0.0560, 0.0929]	0.0351 [0.0154, 0.0584]	0.0621 [0.0478, 0.0771]

Table B32: **Fine-grained caption metrics (Rate-Score, FORTE) on the external Inhouse-Abdomen-2 dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.

Model	**micro Precision** [95% CI]	**micro Recall** [95% CI]	**micro F1** [95% CI]
Astra	0.3811 [0.3576, 0.4041]	0.6046* [0.5736, 0.6377]	0.4675* [0.4443, 0.4898]
Gemini-3	0.1544 [0.1223, 0.1851]	0.1207 [0.0964, 0.1449]	0.1354 [0.1087, 0.1617]
Qwen3-VL-8B	0.0989 [0.0742, 0.1232]	0.0706 [0.0524, 0.0890]	0.0824 [0.0615, 0.1025]
HuluMed-32B	0.1906 [0.1567, 0.2256]	0.1297 [0.1057, 0.1555]	0.1543 [0.1269, 0.1821]
HuluMed-14B	0.1316 [0.1018, 0.1622]	0.0988 [0.0766, 0.1216]	0.1129 [0.0884, 0.1376]
HuluMed-7B	0.1670 [0.1333, 0.2008]	0.1078 [0.0861, 0.1283]	0.1310 [0.1058, 0.1555]
M3D	0.1497 [0.1140, 0.1847]	0.0719 [0.0535, 0.0911]	0.0971 [0.0733, 0.1216]
RadFM	0.0704 [0.0372, 0.1075]	0.0180 [0.0092, 0.0277]	0.0286 [0.0147, 0.0438]
Lingshu-32B	0.1241 [0.0923, 0.1562]	0.0629 [0.0459, 0.0809]	0.0835 [0.0616, 0.1054]
Lingshu-7B	0.1034 [0.0642, 0.1449]	0.0270 [0.0165, 0.0379]	0.0428 [0.0263, 0.0599]
MedGemma-27B	0.2222 [0.0909, 0.3500]	0.0103 [0.0038, 0.0177]	0.0196 [0.0074, 0.0336]
MedGemma-4B	0.5135 [0.3448, 0.6744]	0.0244 [0.0138, 0.0365]	0.0466 [0.0266, 0.0686]

Table B33: **RadBERT classification performance (micro Precision, Recall, F1) on the external Inhouse-Abdomen-3 dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement.

Model	Rate-Score [95% CI]	F.-Degree [95% CI]	F.-Landmark [95% CI]	F.-Feature [95% CI]	F.-Impression [95% CI]	F.-Overall [95% CI]
Astra	0.3143* [0.3077, 0.3216]	0.4053* [0.3927, 0.4182]	0.4512* [0.4376, 0.4650]	0.3215* [0.3069, 0.3353]	0.2640* [0.2352, 0.2916]	0.3605* [0.3506, 0.3704]
Gemini-3	0.2034 [0.1976, 0.2092]	0.1050 [0.0911, 0.1201]	0.1289 [0.1151, 0.1438]	0.0919 [0.0798, 0.1043]	0.1012 [0.0782, 0.1231]	0.1068 [0.0942, 0.1191]
Qwen3-VL-8B	0.1938 [0.1891, 0.1985]	0.0829 [0.0694, 0.0966]	0.0929 [0.0806, 0.1053]	0.0677 [0.0579, 0.0771]	0.0277 [0.0162, 0.0393]	0.0678 [0.0591, 0.0757]
HuluMed-32B	0.2058 [0.1998, 0.2122]	0.1536 [0.1324, 0.1754]	0.1657 [0.1457, 0.1849]	0.1140 [0.0978, 0.1295]	0.0871 [0.0608, 0.1139]	0.1301 [0.1132, 0.1460]
HuluMed-14B	0.1925 [0.1872, 0.1981]	0.1002 [0.0844, 0.1172]	0.1215 [0.1039, 0.1391]	0.0687 [0.0555, 0.0819]	0.0628 [0.0416, 0.0853]	0.0883 [0.0748, 0.1021]
HuluMed-7B	0.2028 [0.1972, 0.2079]	0.1212 [0.1048, 0.1378]	0.1513 [0.1355, 0.1665]	0.0891 [0.0745, 0.1028]	0.0605 [0.0398, 0.0798]	0.1055 [0.0928, 0.1172]
M3D	0.1860 [0.1818, 0.1903]	0.0941 [0.0802, 0.1087]	0.1053 [0.0925, 0.1183]	0.0691 [0.0568, 0.0815]	0.0453 [0.0274, 0.0656]	0.0785 [0.0682, 0.0891]
RadFM	0.1690 [0.1650, 0.1733]	0.0517 [0.0397, 0.0641]	0.0555 [0.0443, 0.0669]	0.0369 [0.0273, 0.0480]	0.0171 [0.0052, 0.0304]	0.0403 [0.0323, 0.0489]
Lingshu-32B	0.1929 [0.1886, 0.1978]	0.0648 [0.0535, 0.0760]	0.1190 [0.1044, 0.1341]	0.0646 [0.0536, 0.0754]	0.0381 [0.0227, 0.0545]	0.0716 [0.0629, 0.0805]
Lingshu-7B	0.1734 [0.1696, 0.1774]	0.0383 [0.0283, 0.0496]	0.0591 [0.0487, 0.0707]	0.0234 [0.0164, 0.0308]	0.0149 [0.0048, 0.0280]	0.0339 [0.0275, 0.0414]
MedGemma-27B	0.1730 [0.1692, 0.1773]	0.0313 [0.0221, 0.0415]	0.0615 [0.0498, 0.0739]	0.0320 [0.0231, 0.0417]	0.0205 [0.0078, 0.0357]	0.0363 [0.0287, 0.0447]
MedGemma-4B	0.1689 [0.1654, 0.1722]	0.0210 [0.0126, 0.0308]	0.0429 [0.0312, 0.0559]	0.0157 [0.0093, 0.0235]	0.0181 [0.0059, 0.0339]	0.0244 [0.0169, 0.0329]

Table B34: **Fine-grained caption metrics (Rate-Score, FORTE) on the external Inhouse-Abdomen-3 dataset** ($n = 400$ cases; 95% confidence intervals estimated via 2,000 bootstrap iterations). * represents a significant improvement between Astra and the second-best baseline with $P < 0.05$, otherwise shows a non-significant improvement. F. means FORTE metric.