Supplementary Figures

2025-11-07

Table of contents

Supplementary Figure 1

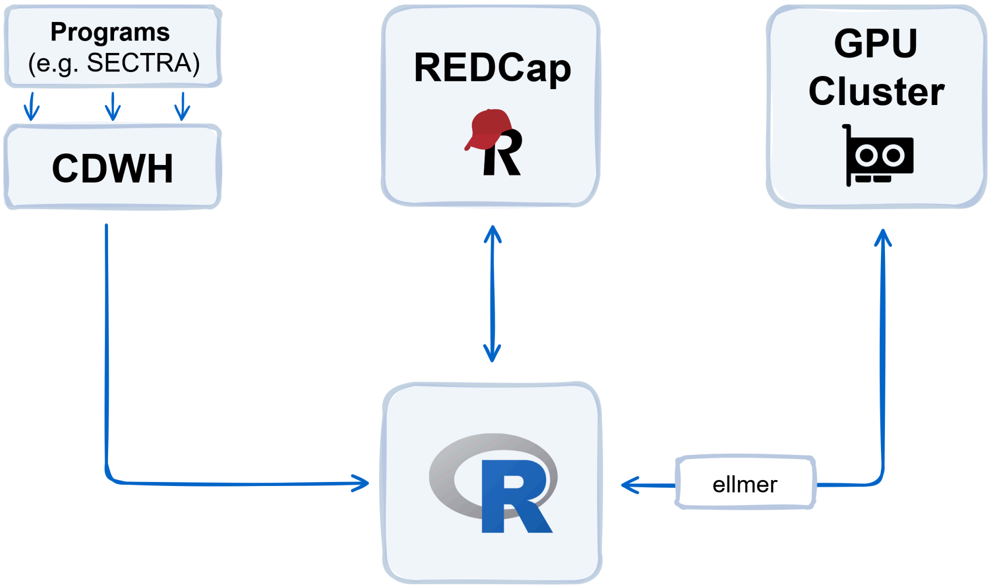

Figure 1: This figure shows the flow of data in our study. The unstructured radiology reports and metadata are pulled from the Clinical Datawarehouse (CDWH) using a database interface in R and uploaded to the study database hosted on a local REDCap instance. The imaging reports are processed by Large Language Models (LLMs) on a local GPU cluster. For communication between R and the LLMs we use the ellmer package. Abbreviations: CDWH, Clinical Datawarehouse; GPU, Graphics Processing Unit.

Supplementary Figure 2

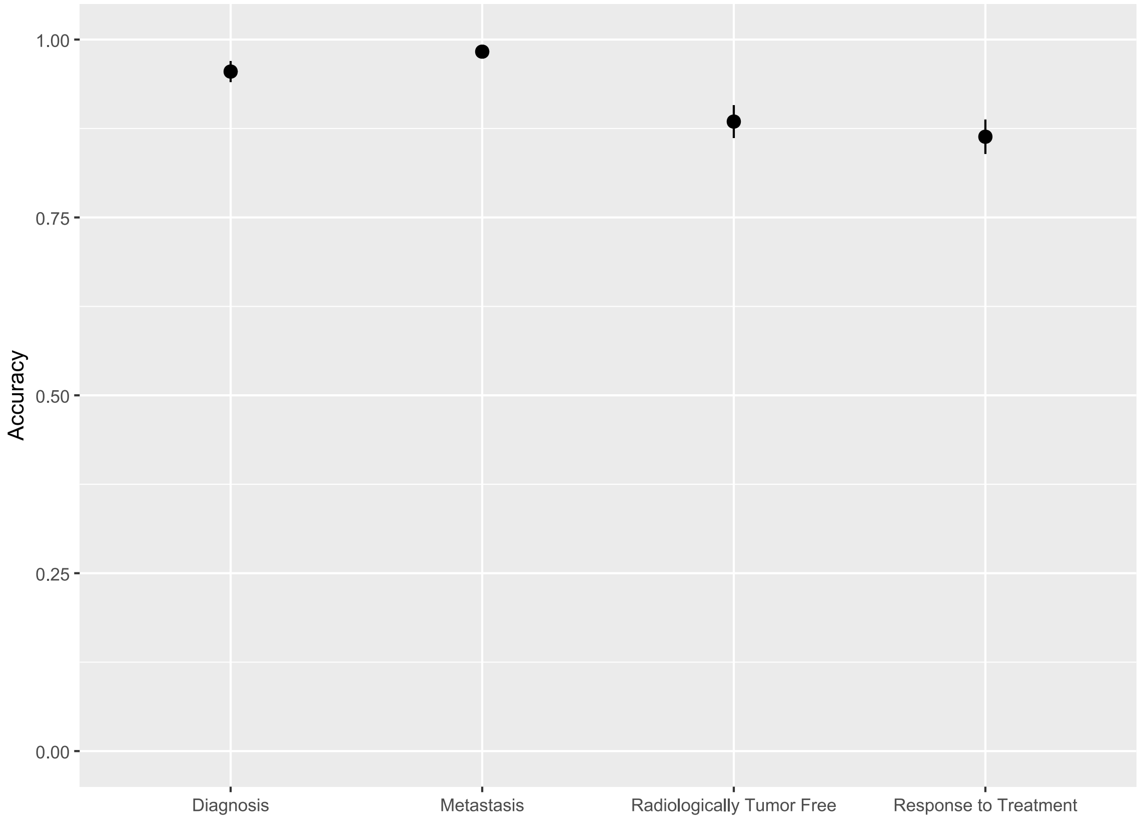

Figure 2: The plot shows the accuracy with 95% confidence intervals of human extractors compared to the ground truth, stratified by task.

Supplementary Figure 3

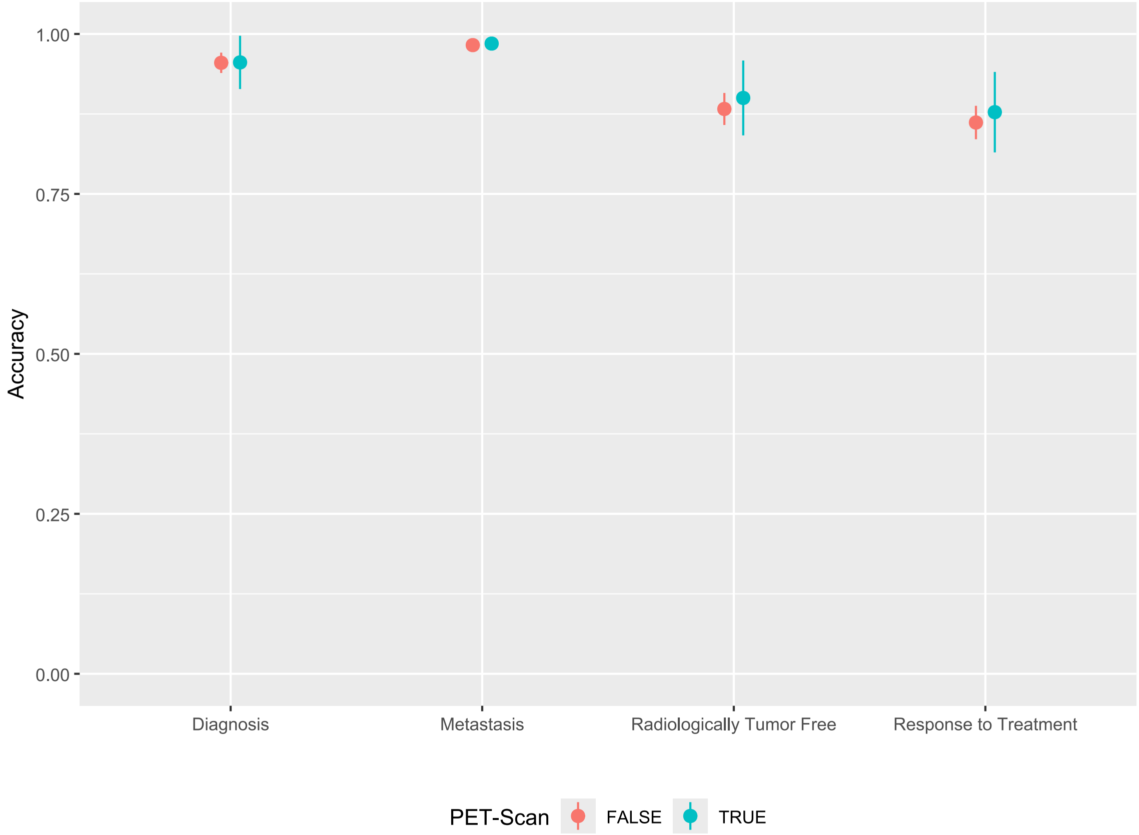

Figure 3: The plot shows the accuracy with 95% confidence intervals of human extractors compared to the ground truth, stratified by task and PET-CT vs non-PET-CT scans.

Supplementary Figure 4

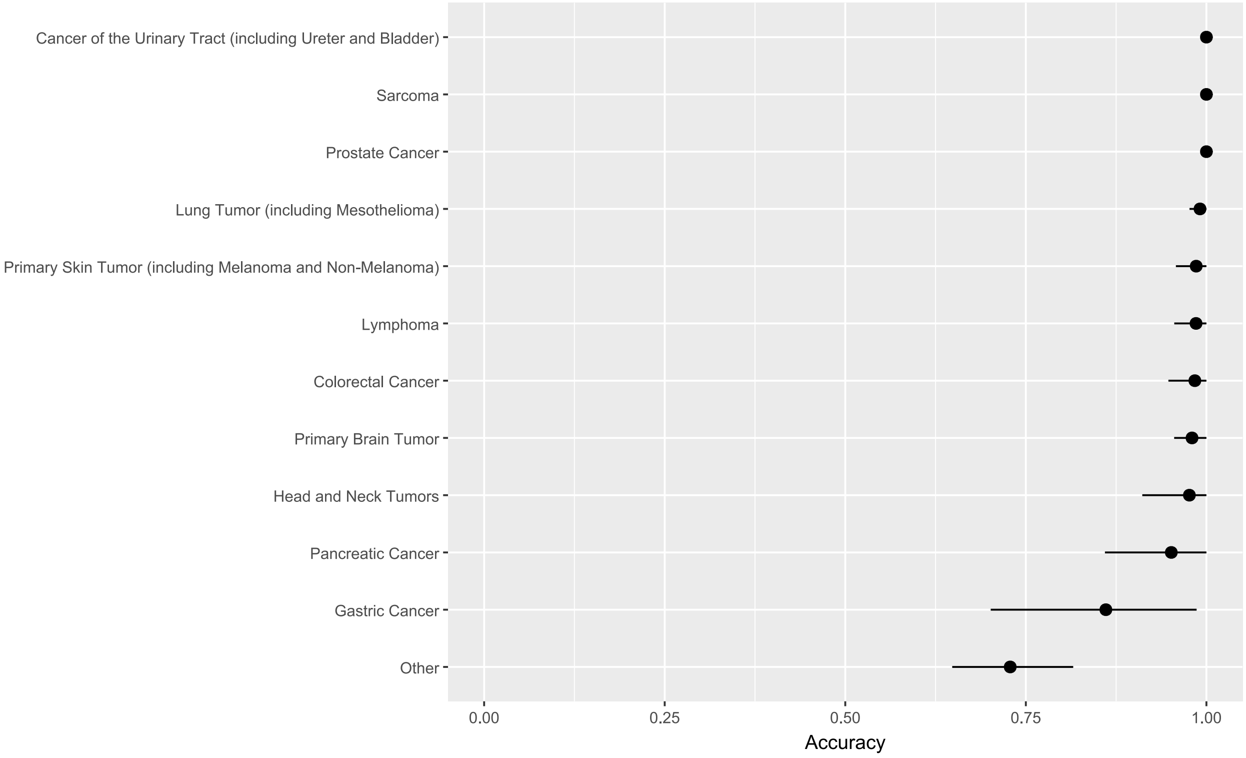

Figure 4: This plot shows the accuracy with 95% confidence intervals of human extractors compared to the ground truth, stratified by extracted diagnosis. Diagnoses with less than 20 extractions were grouped into 'Other'. Confidence intervals were calculated using cluster bootstraping resampling (1000 samples).

Supplementary Figure 5

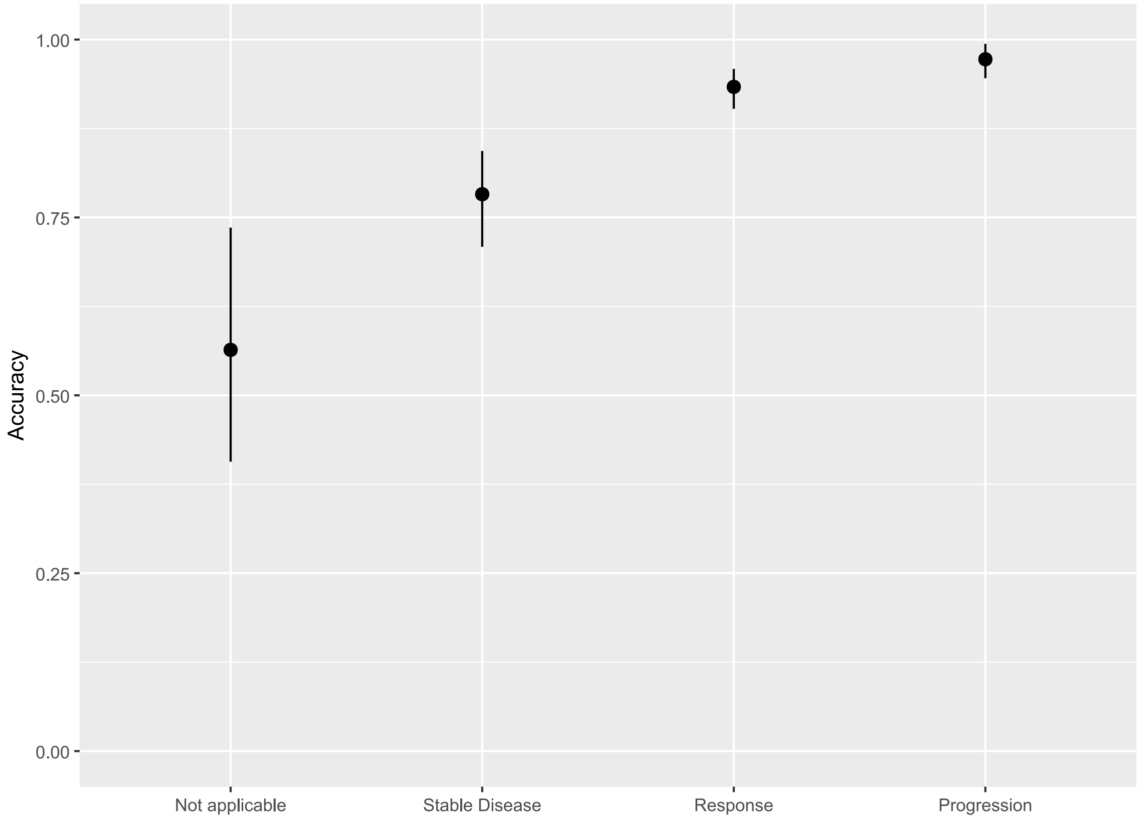

Figure 5: This plot shows the accuracy with 95% confidence intervals of human extractors compared to the ground truth, stratified by extracted response for Response to Treatment (non-PET-CT scans). Confidence intervals were calculated using cluster bootstraping resampling (1000 samples).

Supplementary Figure 6

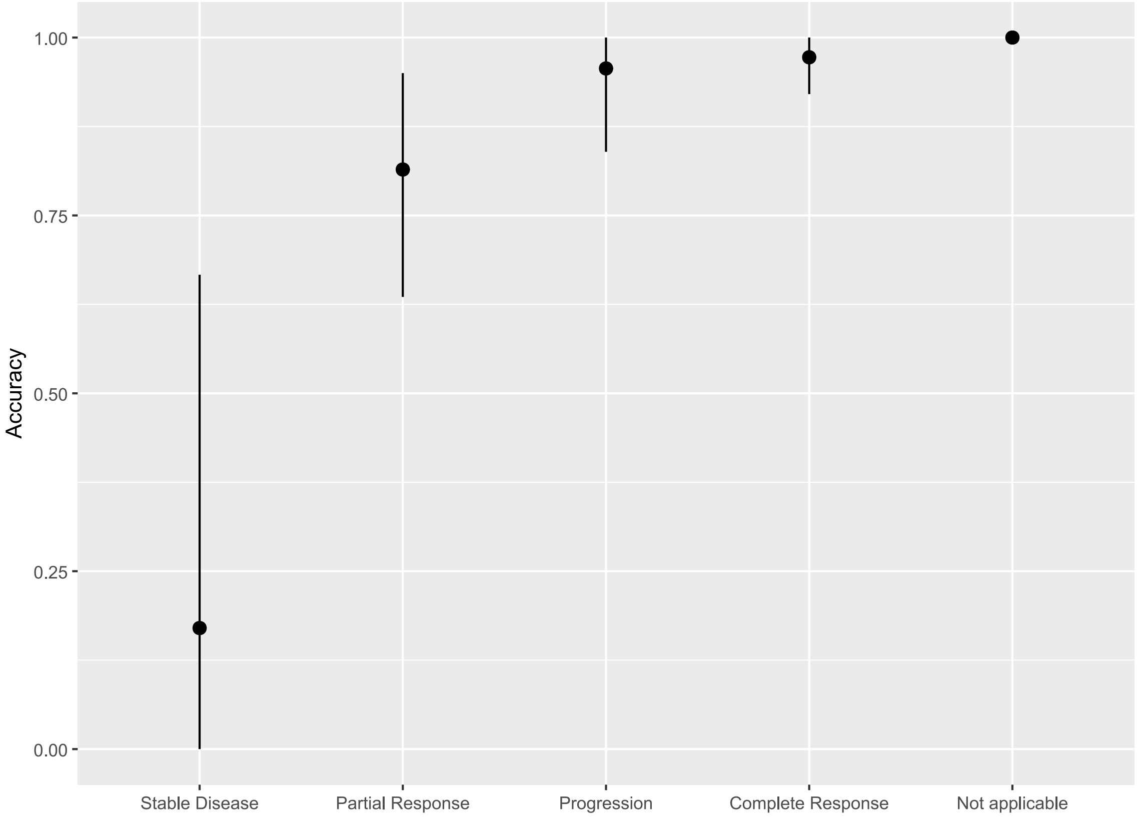

Figure 6: This plot shows the accuracy with 95% confidence intervals of human extractors compared to the ground truth, stratified by extracted response for Response to Treatment (PET-CT scans). Confidence intervals were calculated using cluster bootstraping resampling (1000 samples).

Supplementary Table 1

	Accuracy (95% CI)	Difference (95% CI)	P-value NI
Diagnosis			
human	95.2% (93.4–97.0%)		—
gpt-oss:120b	94.0% (91.3–96.7%)	−1.2 (-3.7 to 1.4)	0.002
qwen3:32b	69.0% (63.8–74.2%)	−26.2 (-31.4 to −20.9)	1.000
qwq:32b	54.0% (48.4–59.6%)	−41.2 (-46.8 to −35.5)	1.000
mistral-small:24b	90.0% (86.6–93.4%)	−5.2 (-8.4 to −1.9)	0.540
llama3.3:70b	92.0% (88.9–95.1%)	−3.2 (-5.9 to −0.5)	0.091
Radiologically Tumor Free			
human	87.6% (84.9–90.4%)		—
gpt-oss:120b	80.6% (76.1–85.1%)	−7.0 (-11.2 to −2.9)	0.832
qwen3:32b	75.6% (70.7–80.5%)	−12.0 (-16.4 to −7.7)	0.999
qwq:32b	79.9% (75.4–84.5%)	−7.7 (-11.9 to −3.5)	0.897
mistral-small:24b	70.1% (64.9–75.3%)	−17.5 (-22.3 to −12.7)	1.000
llama3.3:70b	71.9% (66.8–77.0%)	−15.7 (-20.4 to −11.0)	1.000
Overall			
human	96.1% (95.6–96.7%)		—
gpt-oss:120b	94.2% (93.4–95.1%)	−1.9 (-2.7 to −1.1)	<0.001
qwen3:32b	90.7% (89.7–91.6%)	−5.5 (-6.4 to −4.5)	0.830
qwq:32b	90.2% (89.3–91.2%)	−5.9 (-6.9 to −5.0)	0.972
mistral-small:24b	90.6% (89.6–91.7%)	−5.5 (-6.6 to −4.5)	0.845
llama3.3:70b	91.2% (90.1–92.3%)	−5.0 (-6.0 to −3.9)	0.466

Table 1: Accuracy with 95% confidence intervals for diagnosis, radiological absence of tumor, and across all tasks, by human extractors and LLMs, along with comparisons to human performance.

Supplementary Figure 7

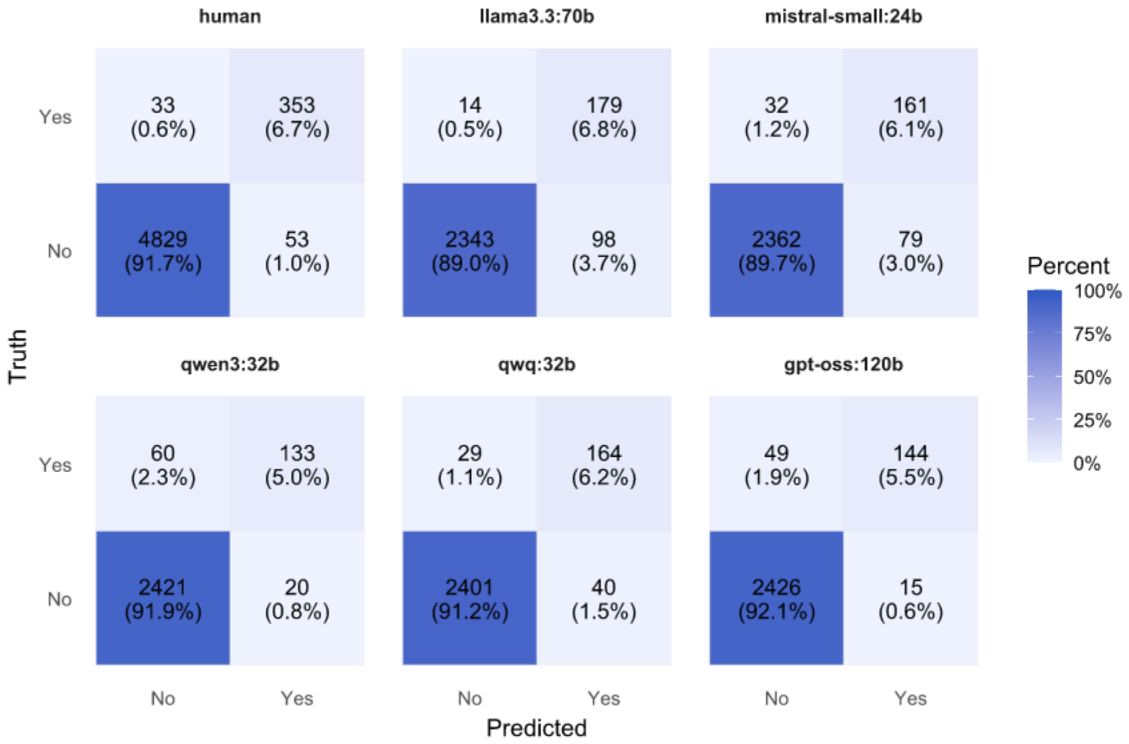

Figure 7: Confusion matrices for humans and LLMs for metastasis classification. As humans extracted in duplicate, the confusion matrix for humans is based on both extractions.

Supplementary Figure 8

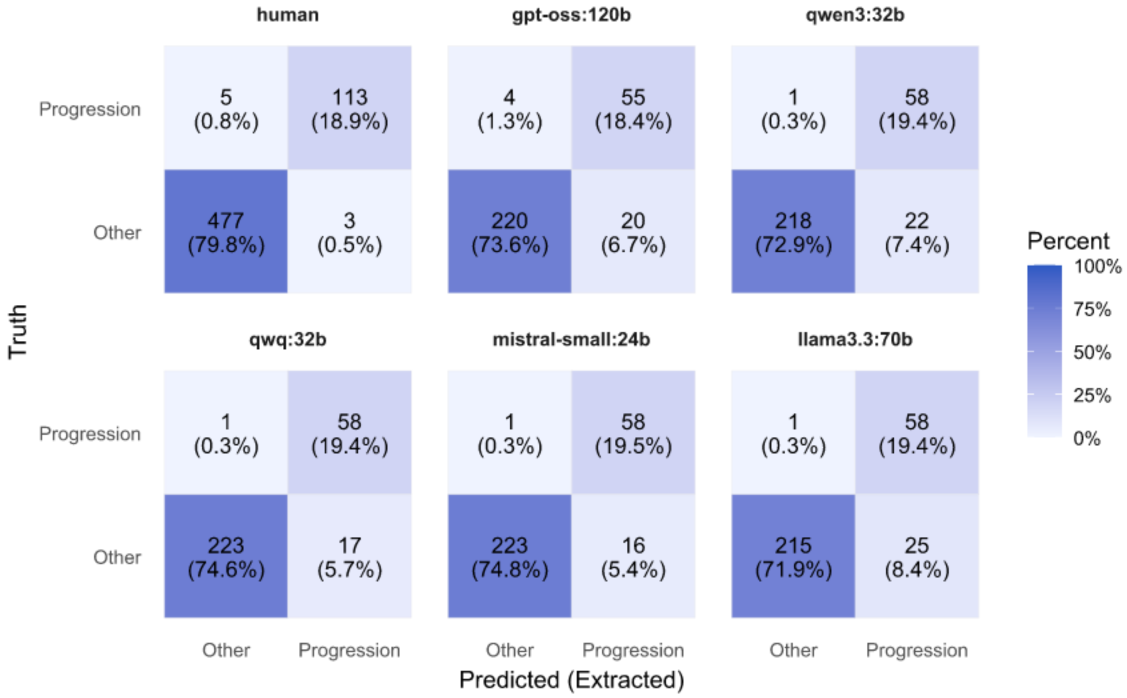

Figure 8: Confusion matrices for humans and LLMs for reponse to treatment classification, when response to treatment is dichotomized into progression vs other. As humans extracted in duplicate, the confusion matrix for humans is based on both extractions.

Supplementary Table 2

	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
human	91.5% (88.7–94.2%)	98.9% (98.6–99.2%)	86.9% (83.1–90.8%)	99.3% (99.1–99.6%)
gpt-oss:120b	74.6% (67.3–81.9%)	99.4% (99.1–99.7%)	90.6% (86.0–95.1%)	98.0% (97.4–98.7%)
qwen3:32b	68.9% (61.5–76.3%)	99.2% (98.8–99.6%)	86.9% (81.1–92.7%)	97.6% (97.0–98.2%)
qwq:32b	85.0% (79.4–90.5%)	98.4% (97.8–98.9%)	80.4% (74.9–85.8%)	98.8% (98.3–99.3%)
mistral-small:24b	83.4% (78.1–88.8%)	96.8% (96.1–97.5%)	67.1% (60.4–73.8%)	98.7% (98.2–99.1%)
llama3.3:70b	92.7% (88.9–96.6%)	96.0% (95.1–96.8%)	64.6% (58.3–70.9%)	99.4% (99.1–99.7%)

Table 2: Sensitivity, Specificity, PPV, and NPV for Metastasis Detection by Human Extractors and LLMs

Supplementary Table 3

Group	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
human	95.9% (95.9–95.9%)	83.7% (83.7–83.7%)	97.4% (97.4–97.4%)	99.0% (99.0–99.0%)
gpt-oss:120b	93.8% (93.8–93.8%)	75.4% (75.4–75.4%)	74.1% (74.1–74.1%)	98.3% (98.3–98.3%)
qwen3:32b	98.5% (98.5–98.5%)	66.3% (66.3–66.3%)	73.1% (73.1–73.1%)	99.6% (99.6–99.6%)
qwq:32b	98.5% (98.5–98.5%)	68.5% (68.5–68.5%)	77.2% (77.2–77.2%)	99.6% (99.6–99.6%)
mistral-small:24b	98.4% (98.4–98.4%)	58.5% (58.5–58.5%)	78.2% (78.2–78.2%)	99.5% (99.5–99.5%)
llama3.3:70b	98.5% (98.5–98.5%)	62.8% (62.8–62.8%)	70.6% (70.6–70.6%)	99.6% (99.6–99.6%)

Table 3: Sensitivity, Specificity, PPV, and NPV for response to treatment, when dichotomized into progression vs other, by human and LLMs.

Supplementary Table 4

	Mean Time per Report (minutes)
human	2.0 (1.0–4.0)
llama3.3:70b	0.6
mistral-small:24b	0.2
qwen3:32b	1.1
qwq:32b	0.8
gpt-oss:120b	0.8