



# Automatic Personalized Impression Generation for PET Reports Using Large Language Models

Xin Tie MS, Muheon Shin MD, Ali Pirasteh MD, Ibrahim Nevein MD, Zachary Huemann MS,  
Junjie Hu PhD, Steve Y. Cho MD, Tyler J. Bradshaw PhD

Department of Radiology, University of Wisconsin-Madison, Madison, WI

# Introduction: Radiology Report



Radiological Examination

Reading Physician



**EXAM:** PET CT SKULL BASE TO THIGH.

**HISTORY:** New diagnosis of lymphoma, 3-week history of cough, fevers, night sweats, and 8-pound weight loss.

**TECHNIQUE:** In a rested and fasted state, the patient was injected with F-18 FDG. The patient was imaged after an approximate 1 hour uptake phase delay using a dedicated dual modality PET/CT scanner with concurrent noncontrast CT scan data acquisition used for anatomic correlation and coregistration, and for attenuation correction. Images were processed with and without attenuation correction and with the generation of standardized uptake values (SUVs) as appropriate. The study is reviewed with the CT images, PET images and the fused PET/CT images presented in the standard three projections, along with the MIP 3-D rotational images.

**COMPARISON:** [DATE] neck and chest CT, [DATE] mediastinal mass biopsy

**FINDINGS:**

Skull base/neck: Probable physiologic activity is noted in the nasopharynx. Uptake in the floor of the mouth may be related to artifact from dental hardware. Hypermetabolic nodes are present in both supraclavicular regions, contiguous with superior mediastinal adenopathy.

Chest: There is bilateral hypermetabolic mediastinal lymphadenopathy. For reference, the largest left prevascular mass of confluent adenopathy has a maximum SUV of 13.6. No hypermetabolic hilar adenopathy is identified. No lung lesions are identified. No hypermetabolic axillary nodes.

Abdomen/pelvis/proximal thighs: Clustered hypermetabolic masses are noted in the left upper quadrant at the level of the splenic hilum and adjacent to the stomach and pancreatic tail. Maximum SUV is 12.5. There is presumed physiologic activity along both renal collecting systems and ureters and bladder as well as in the ascending colon. Activity in the right retroperitoneum is at the level of the mid ureter and is favored to be ureteral in origin. Small soft tissue densities are noted on noncontrast CT images in the retroperitoneum, suspicious for adenopathy; however, they are not hypermetabolic.

Musculoskeletal: No focal hypermetabolic lesions are identified in the visualized portions of the skeleton.

**IMPRESSIONS:**

- [1] Intensely hypermetabolic superior mediastinal and bilateral supraclavicular adenopathy.
- [2] Intensely hypermetabolic clustered lymph nodes in the region of the splenic hilum.
- [3] Non-hypermetabolic retroperitoneal adenopathy is suspected. Correlation with a diagnostic CT of the abdomen and pelvis with IV and oral contrast is recommended.

Radiology Report

Healthcare Team



Patient



# Introduction: Clinical Whole-body PET Reports

---

## BACKGROUND INFORMATION

[AGE] years old patient with Gleason 3+4 prostate adenocarcinoma diagnosed in [DATE]. Patient is referred for initial staging.

---

---

## FINDINGS

Physiologic background liver standardized uptake value (...)  
Visualized head/neck: Physiologic uptake of the lacrimal and salivary glands. (...)  
Head / neck lymph nodes: No suspicious head/neck lymph nodes.  
Chest: No lung nodules or abnormal uptake. No pleural or pericardial effusion. (...)  
Hepatobiliary: No abnormal uptake. Photogenic left hepatic cyst. (...)  
Spleen / pancreas / adrenals: No abnormal splenic uptake. (...)  
Kidneys / bladder: No abnormal uptake. (...)  
Bowel / peritoneum: No suspicious bowel uptake or abnormality. (...)  
Pelvic organs: Bilateral foci of uptake bilaterally in the peripheral zone of the prostate, (...)  
Abdominopelvic lymph nodes: No definitive suspicious abdominopelvic lymph nodes. (...)  
Musculoskeletal / soft tissues / skin: No suspicious lesions. (...)

---

Report for a PET/CT skull base to thigh exam

# Introduction: Clinical Whole-body PET Reports

---

## BACKGROUND INFORMATION

[AGE] years old patient with Gleason 3+4 prostate adenocarcinoma diagnosed in [DATE]. Patient is referred for initial staging.

---

---

## FINDINGS

Physiologic background liver standardized uptake value (...)  
Visualized head/neck: Physiologic uptake of the lacrimal and salivary glands. (...)  
Head / neck lymph nodes: No suspicious head/neck lymph nodes.  
Chest: No lung nodules or abnormal uptake. No pleural or pericardial effusion. (...)  
Hepatobiliary: No abnormal uptake. Photogenic left hepatic cyst. (...)  
Spleen / pancreas / adrenals: No abnormal splenic uptake. (...)  
Kidneys / bladder: No abnormal uptake. (...)  
Bowel / peritoneum: No suspicious bowel uptake or abnormality. (...)  
Pelvic organs: Bilateral foci of uptake bilaterally in the peripheral zone of the prostate, (...)  
Abdominopelvic lymph nodes: No definitive suspicious abdominopelvic lymph nodes. (...)  
Musculoskeletal / soft tissues / skin: No suspicious lesions. (...)

---

---

## IMPRESSION

[1] Bilateral focal of uptake in the prostate consistent with known adenocarcinoma.  
[2] No definitive evidence for metastatic disease. Several foci of uptake in the region of the lower lumbar spine, are favored to represent benign uptake, as detailed.  
Follow-up may be helpful.

---

- Referring physicians primarily rely on the **impression** section for clinical decision-making and management<sup>1</sup>

Report for a PET/CT skull base to thigh exam

# Introduction: Clinical Whole-body PET Impressions

## ■ Accuracy and completeness

### FINDINGS

Physiologic background liver standardized uptake value (SUV mean and SUV max) reported for comparison between PET studies: 2.7 and 3.5, previously 2.4 and 3.0.

Visualized head/neck: Dental hardware artifact limiting evaluation of adjacent structures. Postsurgical changes following multiple resections and reconstructions. Intense uptake of the floor the mouth extending posteriorly to the base of the right tongue, right parapharyngeal tissues, and adjacent palatal flap, which correlate with areas enhancement on MR [DATE] and is consistent with biopsy-proven disease recurrence. For example, the SUV max is 7.0 in the surgical bed (PET/CT axial image 25), and, the SUV max is 9.4 at the base of the tongue (PET/CT axial image 32). There is some mass effect on the right internal carotid artery, correlating with chronically occluded vessels seen on recent MR.

Focal uptake associated with left mandibular teeth, approximately at the location of tooth #13, is likely inflammatory/reactive due to recent tooth extraction (PET/CT axial image 30). Compared to PET CT [DATE], interval placement of tracheostomy tube, which is in place with surrounding inflammatory/reactive uptake. Elsewhere, physiologic uptake in the visualized portions of the brain, extraocular muscles, and salivary glands.

Head/neck lymph nodes: No suspicious head/neck lymph nodes.

Lungs: Atelectatic banding/scarring in the right lower lobe. Compared to PET/CT [DATE], there is a new pulmonary nodule in the right lower lobe measuring 2.1 x 1.4 cm with SUV max 4.6 (PET/CT axial image 103). There are other new sub-centimeter pulmonary nodules scattered in the right lung, for example, there is a 9 mm pulmonary nodule in the posterior right lung with mild radiotracer uptake (SUV max 1.4, PET/CT axial image 80). New non radiotracer avid 3 mm pulmonary nodule in the posterior lung base (PET/CT axial image 105). Atelectatic banding/scarring in the posterior left lung base.

Pleura/pericardium: No pleural or pericardial effusion.

Thoracic lymph nodes: Mild uptake in few mediastinal lymph nodes, which are nonspecific but favored to be reactive, for example:

- Para-aortic lymph node, 0.7 x 0.5 cm, SUV max 2.0 (PET/CT axial image 79)

- Subcarinal lymph node, 1.1 x 0.67 m, SUV max 2.9 (PET/CT axial image 84).

Other chest findings: Physiologic myocardial uptake. Inflammatory/reactive uptake in the distal esophagus. Mild gynecomastia.

Hepatobiliary: No abnormal uptake. Spleen: No abnormal uptake. Pancreas: No abnormal uptake. Adrenals: Within normal limits.

Kidneys/bladder: No abnormal uptake. Physiologically excreted tracer activity within the renal collecting system and urinary bladder.

Bowel/peritoneum: Redemonstrated gastrostomy tube is appropriately positioned with surrounding reactive uptake. No suspicious uptake. Areas of bowel uptake are probably inflammatory/physiologic, and benign. Compared to PET/CT [DATE], resolved postsurgical pneumoperitoneum with scarring in the subcutaneous fat in the anterior body wall (PET/CT axial image 185).

Pelvic organs: No abnormal uptake. Vasectomy clips.

Abdominopelvic lymph nodes: No suspicious abdominopelvic lymph nodes.

Musculoskeletal/soft tissues/skin: Mild uptake associated with the right body of the ramus correlating with loss of normal marrow signal on MR [DATE] and suspicious for osseous metastasis. Postsurgical changes of the right proximal humerus.

Other: Scattered aortiliac calcifications.

## Time-consuming and error-prone

### IMPRESSION

[1] FDG PET/CT demonstrates intense uptake of the floor the mouth extending posteriorly to the base of the right tongue, right parapharyngeal tissues, and adjacent palatal flap, which correlates with areas enhancement on MR [DATE] and is consistent with biopsy-proven disease recurrence.

[2] Mild uptake associated with the right body of the ramus correlating with loss of normal marrow signal on MR [DATE] and suspicious for osseous metastasis.

[3] Focal uptake associated with left mandibular teeth, approximately at the location of tooth #13, is likely inflammatory/reactive due to recent tooth extraction. Correlate with physical exam.

[4] Compared to PET/CT [DATE], there are new pulmonary nodules in the lungs, as detailed in findings. Moderate uptake associated with the dominant nodule in the right lower lobe is suspicious for metastasis versus second primary malignancy.

[5] Mild uptake in few mediastinal lymph nodes is nonspecific but favored to be reactive. Attention on follow-up.



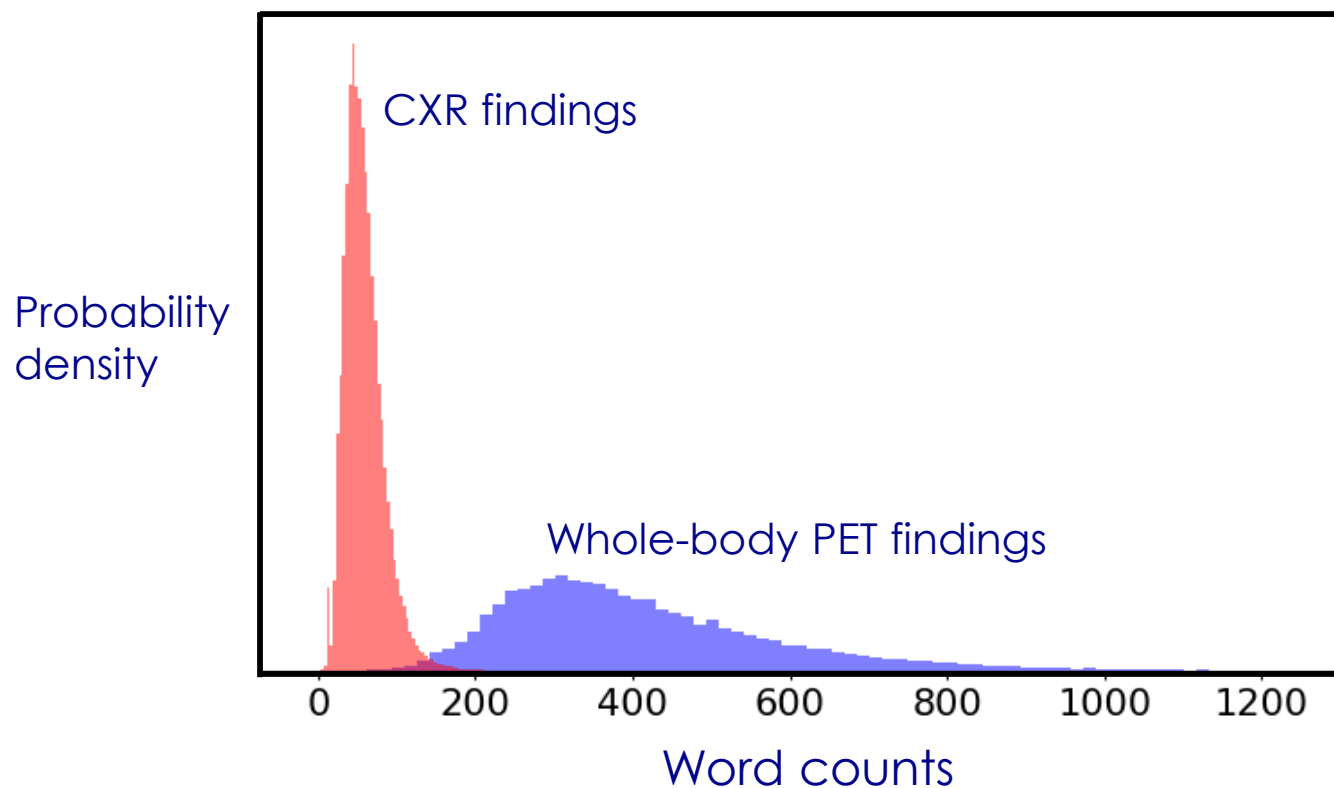
## Distill

# Introduction: Clinical Whole-body PET Impressions

- Accuracy and completeness
- Large language models (LLMs)
  - **Accelerate** the drafting of PET impressions
- **No studies** have focused on impression generation for whole-body PET reports

# Introduction: What makes it different?

- **Length**
  - 250-500 words in the findings section



# Introduction: What makes it different?

## ■ Complexity

- Observations across multiple anatomical regions → **omissions**

---

---

### FINDINGS

Physiologic background liver standardized uptake value (SUV mean and SUV max) (...)

**Visualized head/neck:** Physiologic uptake in the visualized portions of the brain (...)

**Head/neck lymph nodes:** Interval decreased size and FDG activity (...)

-Right level IIb, 1.2 x 0.7 cm transaxially, SUV max 5.5, previously 1.6 x 0.7 cm, SUV max 6.2 (...)

**Lungs / Pleura / pericardium:** No suspicious lung nodules or abnormal uptake. (...)

**Thoracic lymph nodes:** Near complete resolution of supraclavicular/upper mediastinal (...)

**Other chest findings:** Physiologic myocardial uptake. Coronary artery calcifications. (...)

**Hepatobiliary:** No abnormal uptake..

**Spleen / Pancreas / Adrenals:** No abnormal uptake. (...)

**Kidneys / bladder:** No abnormal uptake. Physiologically excreted tracer activity (...)

**Bowel / peritoneum:** No suspicious bowel uptake or abnormality.

**Pelvic organs:** No abnormal uptake..

**Abdominopelvic lymph nodes:** Interval increased FDG uptake in an aortocaval lymph node (...)

**Musculoskeletal / soft tissues / skin:** No suspicious osseous lesions.

---

---



# Introduction: What makes it different?

- **Unique reporting styles** of individual reading physicians
  - Differences in Level of detail, inclusion criterion, phrasing, etc.

---

---

## **IMPRESSION dictated by Physician 1**

[1] FDG PET/CT demonstrates intensely FDG avid lymphadenopathy above and below the diaphragm involving the neck, chest, abdomen, and pelvis, as detailed in the findings, consistent with biopsy-proven mantle cell lymphoma (Deauville score 5).

[2] Marked splenomegaly with diffuse abnormal FDG activity, also suspicious for splenic lymphoma (Deauville 5).

---

---

---

---

## **IMPRESSION dictated by Physician 2**

Neck/thoracic/abdominopelvic nodal as well as splenic lymphoma (Deauville 5).

---

---

# Purpose

- Can LLMs generate PET impressions suitable for clinical use?
  - How to compare different language models?
  - Which language model is the best for this task?
  - Are the LLM impressions useful?

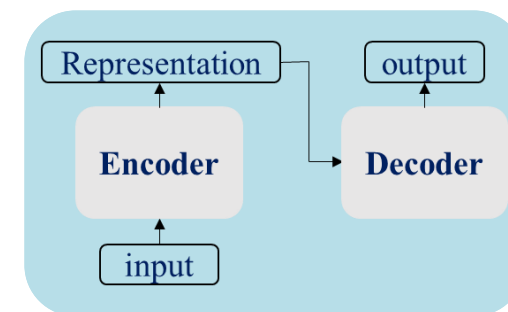
# Dataset

- **37,370** retrospective PET reports from UW healthcare system between Jan 2010 and Jan 2023.
  - **65** different physicians
  - Internal testing: 4000 reports
  - Training: 31,370 reports; Validation: 2000 reports

# Models for PET Report Summarization

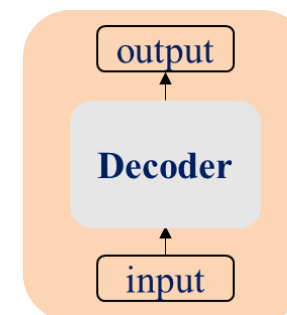
## ■ **Eight** encoder-decoder models

- General-domain pretrained: BART, PEGASUS, T5, Flan-T5
- Medical-knowledge enriched: BioBART, ClinicalT5
- Baseline: Pointer Generator Network (PGN), BERT2BERT



## ■ **Four** decoder-only models

- GPT2, OPT (1.3B), LLaMA (7B), Alpaca (7B)
- LLaMA and Alpaca were fine-tuned with LoRA.



# Personalized Impression Generation

- An extra input token encodes reading physician's identity

**Description:** PET CT SKULL BASE TO THIGH.

**Radiologist:** James

**Findings:**

Physiologic background liver standardized uptake value (SUV mean and SUV max) reported for comparison between PET studies: (...)

Head/neck lymph nodes: No suspicious head/neck lymph nodes.

Lungs: Avid right upper lobe mass (...)

Pleura/pericardium: No pleural or pericardial effusion. (...)

Hepatobiliary: No abnormal uptake. Prior liver transplant. (...)

Pelvic organs: No abnormal uptake.

Abdominopelvic lymph nodes: No suspicious abdominopelvic lymph nodes.

Musculoskeletal/soft tissues/skin: No suspicious osseous lesions. (...)

Other: Aortoiliac atherosclerotic vascular calcifications. (...)

**Indication:**

[AGE] year old [SEX] with new diagnosis of right upper lobe nodule biopsy. (...)

LLM

LLM Impression in James's style

[1] FDG PET/CT demonstrates an intensely FDG avid right upper lobe mass consistent with biopsy-proven primary lung malignancy.

[2] No evidence of other sites of abnormal FDG uptake to suggest locoregional or distant metastatic disease.

# Personalized Impression Generation

- An extra input token encodes reading physician's identity

**Description:** PET CT SKULL BASE TO THIGH.

**Radiologist:** Rachael

**Findings:**

Physiologic background liver standardized uptake value (SUV mean and SUV max) reported for comparison between PET studies: (...)

Head/neck lymph nodes: No suspicious head/neck lymph nodes.

Lungs: Avid right upper lobe mass (...)

Pleura/pericardium: No pleural or pericardial effusion. (...)

Hepatobiliary: No abnormal uptake. Prior liver transplant. (...)

Pelvic organs: No abnormal uptake.

Abdominopelvic lymph nodes: No suspicious abdominopelvic lymph nodes.

Musculoskeletal/soft tissues/skin: No suspicious osseous lesions. (...)

Other: Aortoiliac atherosclerotic vascular calcifications. (...)

**Indication:**

[AGE] year old [SEX] with new diagnosis of right upper lobe nodule biopsy. (...)

LLM Impression in Rachael's style

LLM

Right upper lobe primary lung malignancy without evidence of nodal or distant metastases.

# Purpose

- Can LLMs generate PET impressions suitable for clinical use?
  - **How to compare different language models?**
  - Which language model is the best for this task?
  - Are the LLM impressions useful?

# How to compare different language models?

- Correlated 30 automatic evaluation metrics with physician preferences
  - 200 cases
  - 5-point Likert scale (5 best, 1 worst)
  - Spearman's rank correlation

Category	Corresponding Evaluation Metrics
Lexical overlap-based metrics	ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L, ROUGE-LSUM, BLEU, CHRF, METEOR, CIDEr
Embedding-based metrics	ROUGE-WE-1, ROUGE-WE-2, ROUGE-WE-3, BERTScore, MoverScore
Graph-based metrics	RadGraph
Text generation-based metrics	PRISM, BARTScore, <b>BARTScore+PET *</b> , <b>PEGASUSScore+PET *</b> , <b>T5Score+PET *</b>
Supervised regression-based metrics	S <sup>3</sup> -pyr, S <sup>3</sup> -resp
Question answering-based metrics	UniEval
Reference-free metrics	SummaQA, BLANC, SUPERT, Stats-compression, Stats-coverage, Stats-density, Stats-novel trigram

\* domain adapted



# Results: How to compare different language models?

	Spearman's rank correlation
<b>Inter-reader correlation</b>	0.654
BARTScore+PET	0.568
PEGASUSScore+PET	0.563
T5Score+PET	0.542
UniEval	0.501
BARTScore	0.474
CHRF	0.433
Moverscore	0.420
BLEU	0.412
BERTScore	0.407
ROUGE-1	0.402
ROUGE-2	0.379
ROUGE-L	0.398

# Results: How to compare different language models?

Spearman's rank correlation

<b>Inter-reader correlation</b>	0.654
BARTScore+PET	0.568
PEGASUSScore+PET	0.563
T5Score+PET	0.542
UniEval	0.501
BARTScore	0.474
CHRF	0.433
Moverscore	0.420
BLEU	0.412
BERTScore	0.407
ROUGE-1	0.402
ROUGE-2	0.379
ROUGE-L	0.398

# Results: How to compare different language models?

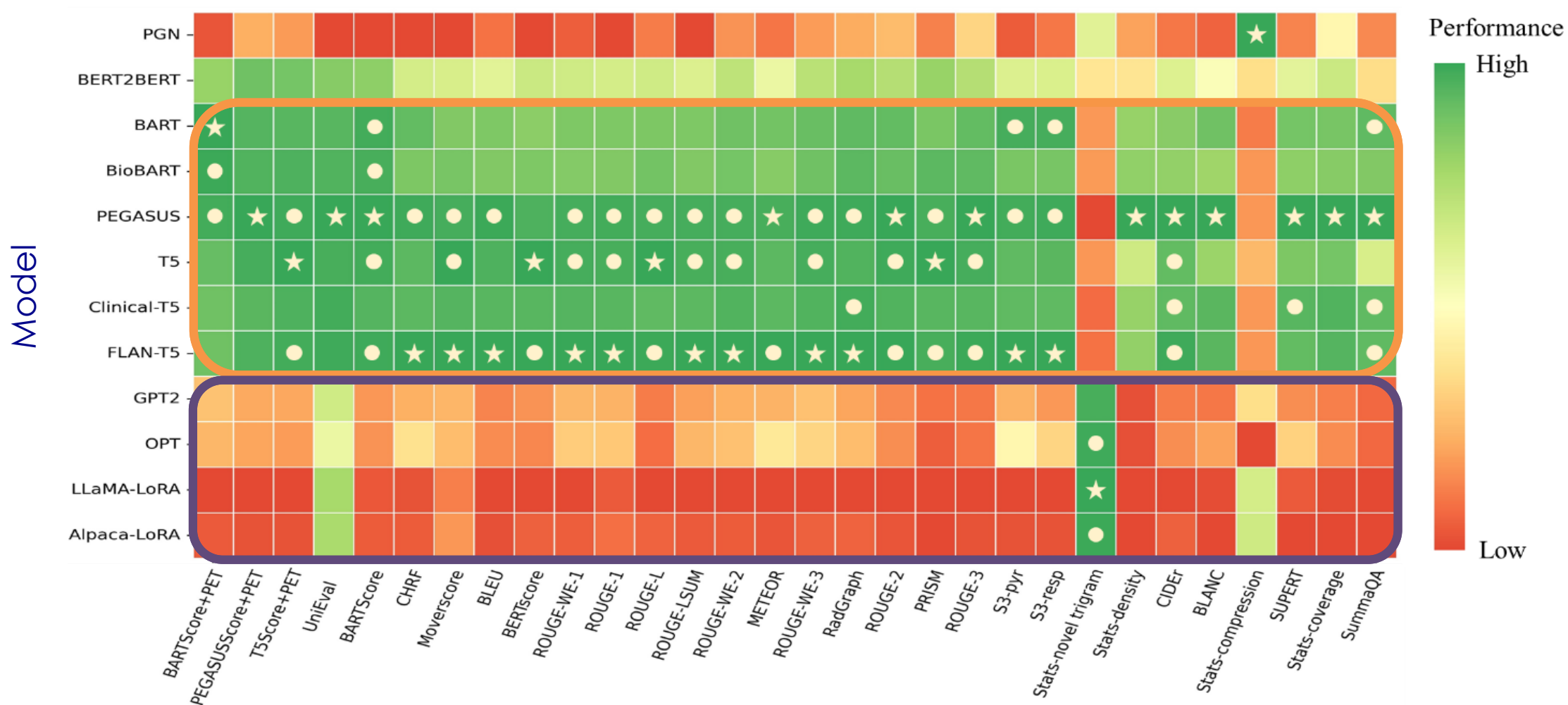
Spearman's rank correlation

<b>Inter-reader correlation</b>	0.654
BARTScore+PET	0.568
PEGASUSScore+PET	0.563
T5Score+PET	0.542
UniEval	0.501
BARTScore	0.474
CHRF	0.433
Moverscore	0.420
BLEU	0.412
BERTScore	0.407
ROUGE-1	0.402
ROUGE-2	0.379
ROUGE-L	0.398

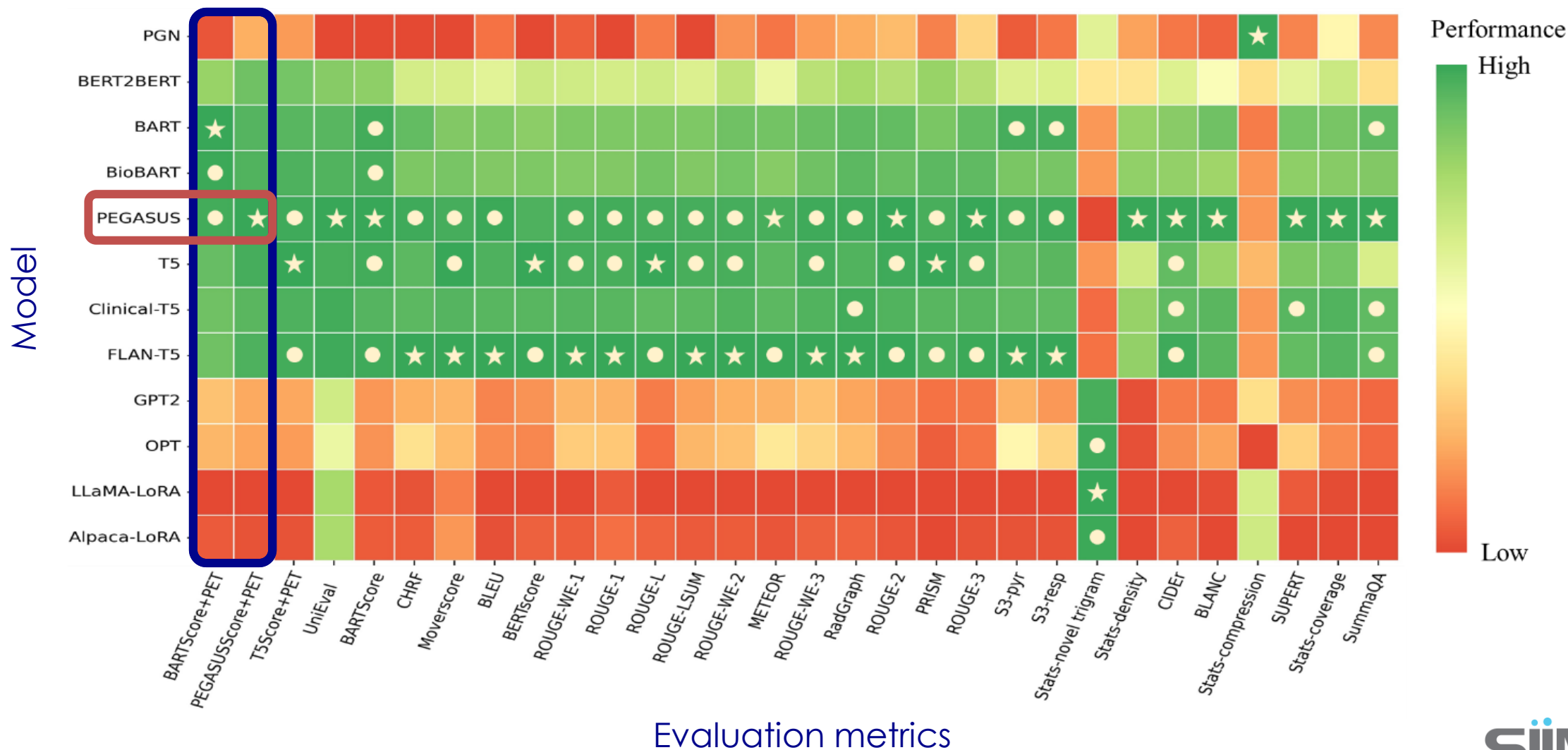
# Purpose

- Can LLMs generate PET impressions suitable for clinical use?
  - How to compare different language models?
  - **Which language model is the best for this task?**
  - Are the LLM impressions useful?

# Results: Which language model is the best?



# Results: Which language model is the best?



# Purpose

- Can LLMs generate PET impressions suitable for clinical use?
  - How to compare different language models?
  - Which language model is the best for this task?
  - **Are the LLM impressions useful?**

# Are the LLM impressions useful?

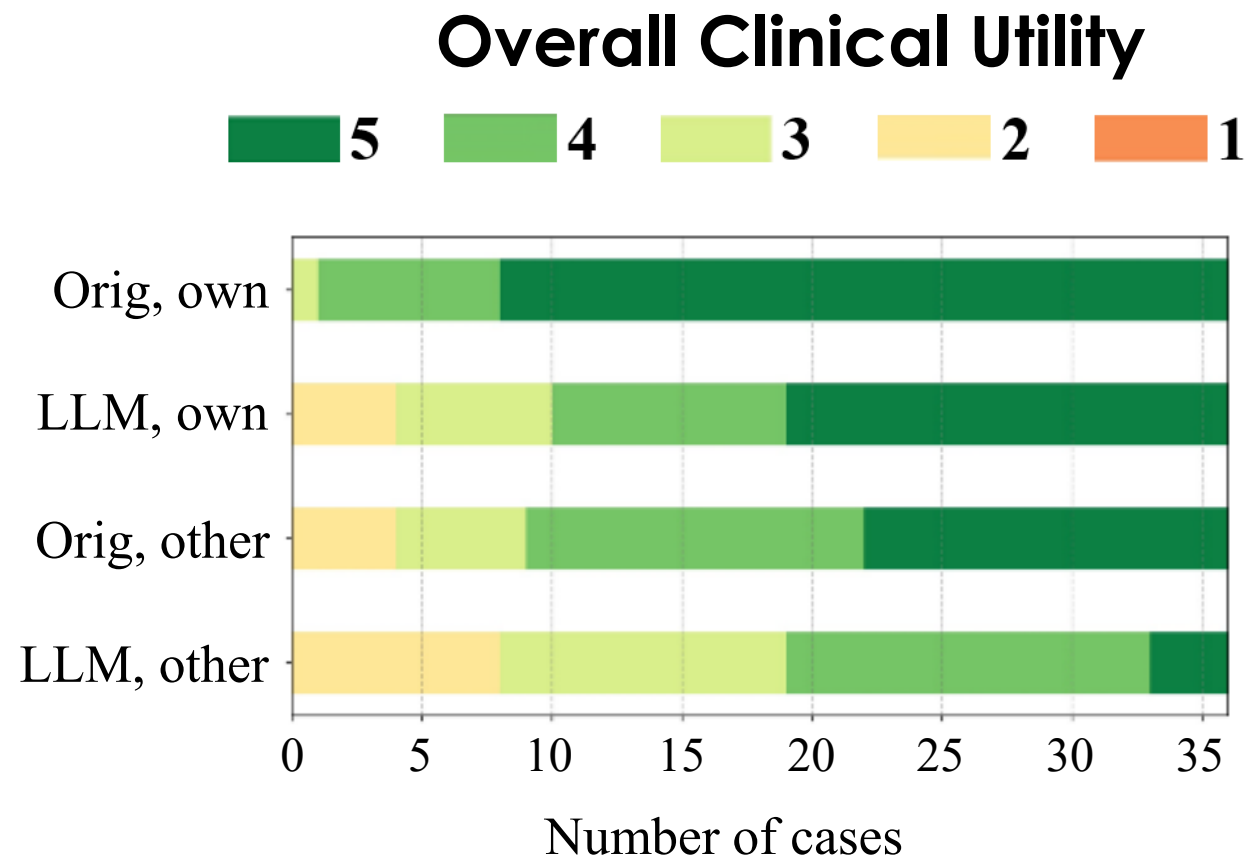
- Three NM physicians reviewed 72 whole-body PET reports
  - **PEGASUS** impressions
  - Each physician independently reviewed 24 cases
  - 12 **their own** reports, 12 **other physicians'** reports
- Scoring system
  - Six quality dimensions (3-point scale, 3 best, 1 worst)
  - Overall utility score (5-point scale, 5 best, 1 worst)



# Are the LLM impressions useful?

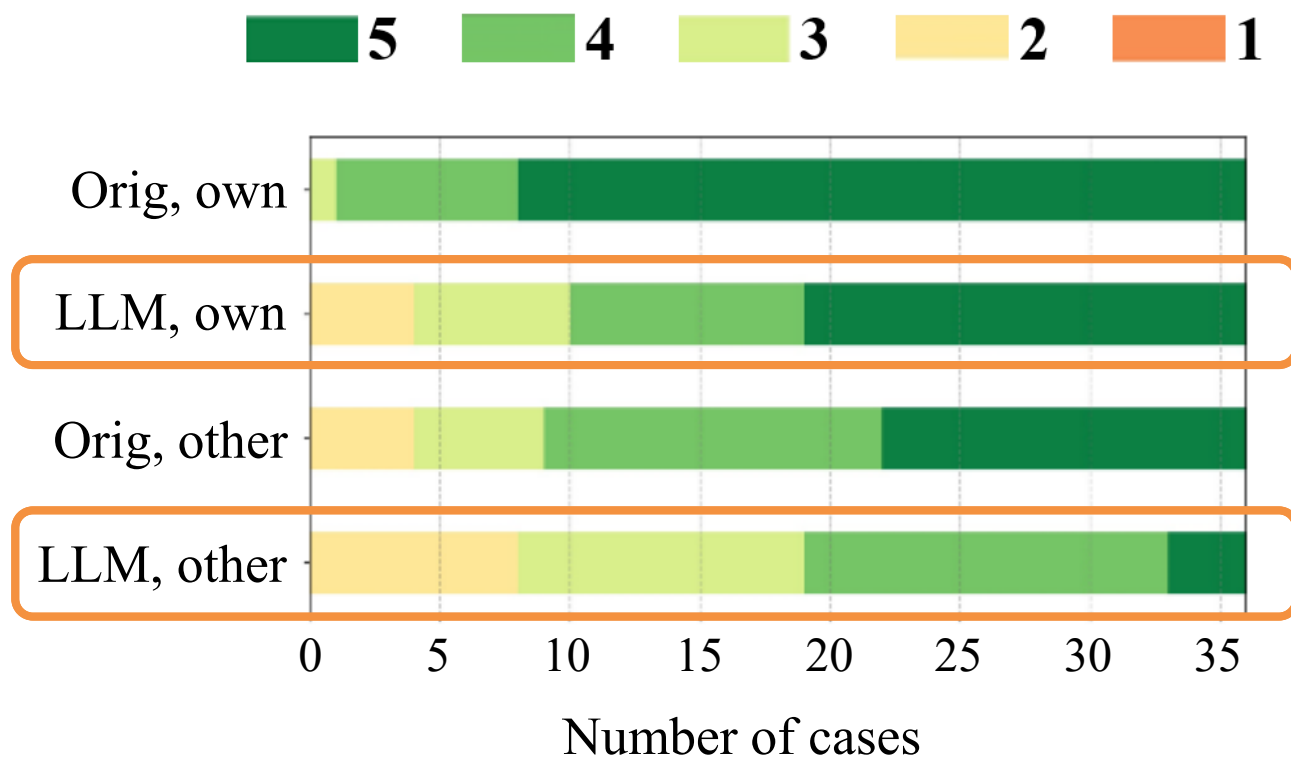
- Three NM physicians reviewed 72 whole-body PET reports
  - **PEGASUS** impressions
  - Each physician independently reviewed 24 cases
  - 12 **their own** reports, 12 **other physicians'** reports
- Scoring system
  - Six quality dimensions (3-point scale, 3 best, 1 worst)
    - Additions, Omissions, Factual correctness, Clarity and organization, Interpretive and technical jargon, Recommendations
  - Overall utility score (5-point scale, 5 best, 1 worst)
    - Scores **3** or above: clinically acceptable

# Results: Are the LLM impressions useful?



# Results: Are the LLM impressions useful?

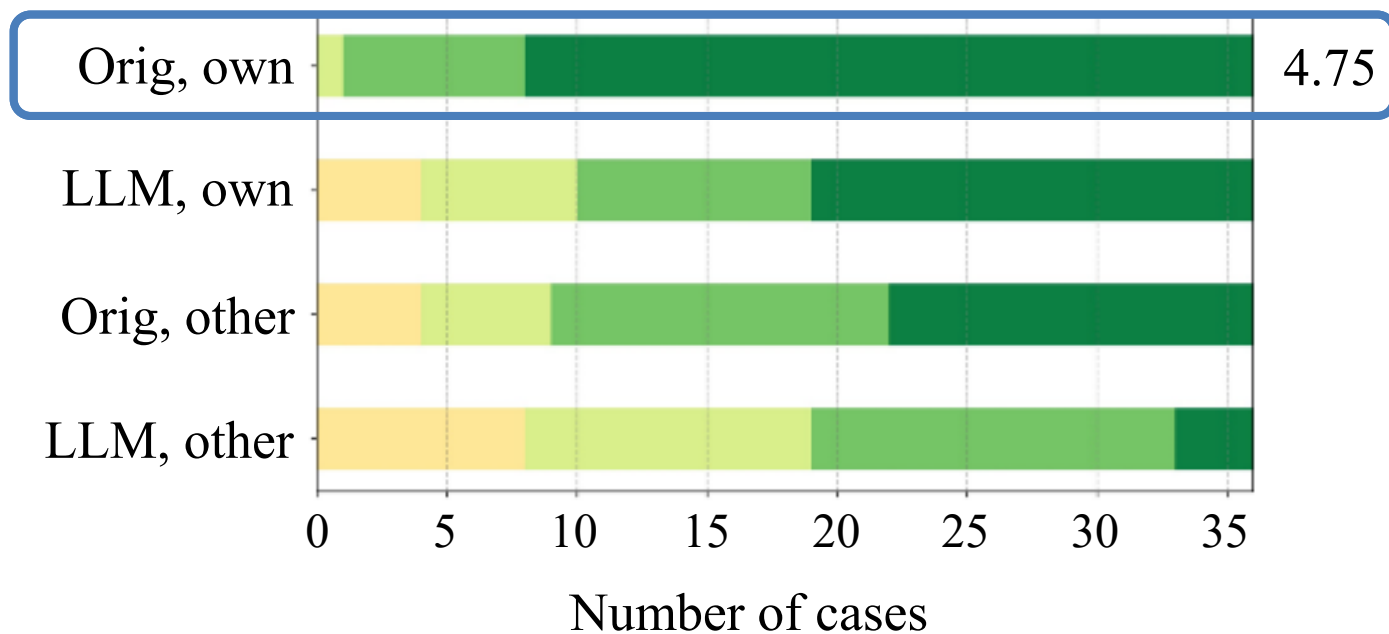
## Overall Clinical Utility



- In total, **83%** of the **PEGASUS-generated impressions** were scored as clinically acceptable (scores 3-5)

# Results: Are the LLM impressions useful?

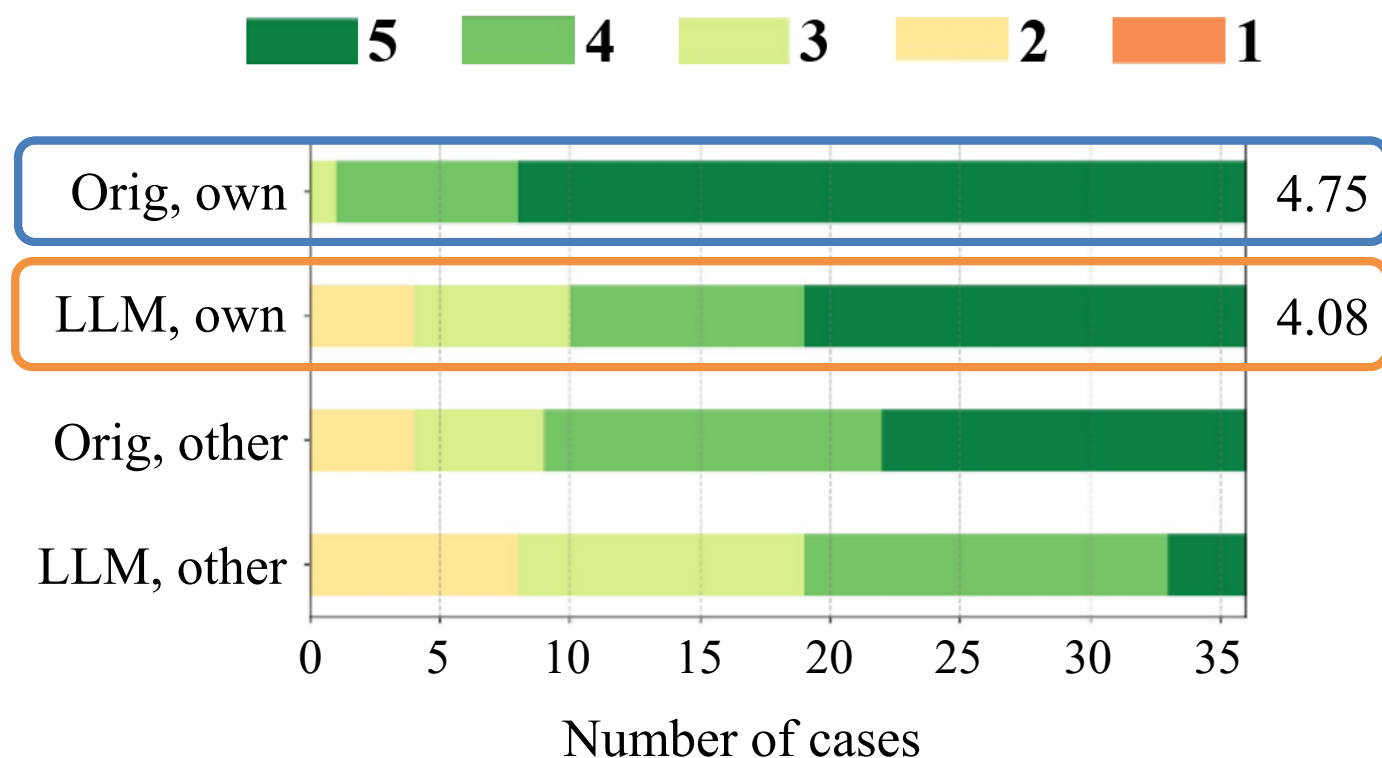
## Overall Clinical Utility



- When the physicians reviewed **their own** reports, the mean utility score of **their own impressions** was 4.75

# Results: Are the LLM impressions useful?

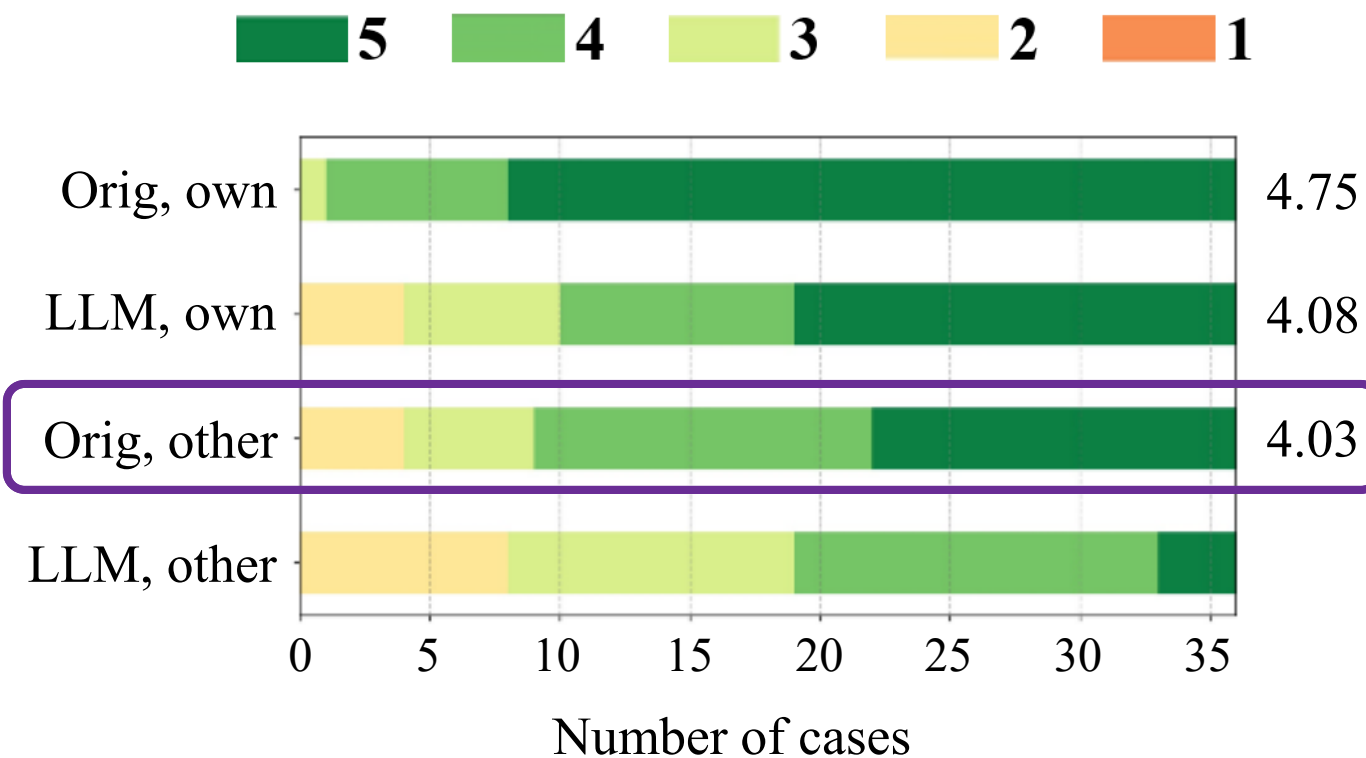
## Overall Clinical Utility



- When the physicians reviewed **their own** reports, the mean utility score of **their own impressions** was 4.75
- The mean utility score of **personalized PEGASUS-generated impressions** was 4.08

# Results: Are the LLM impressions useful?

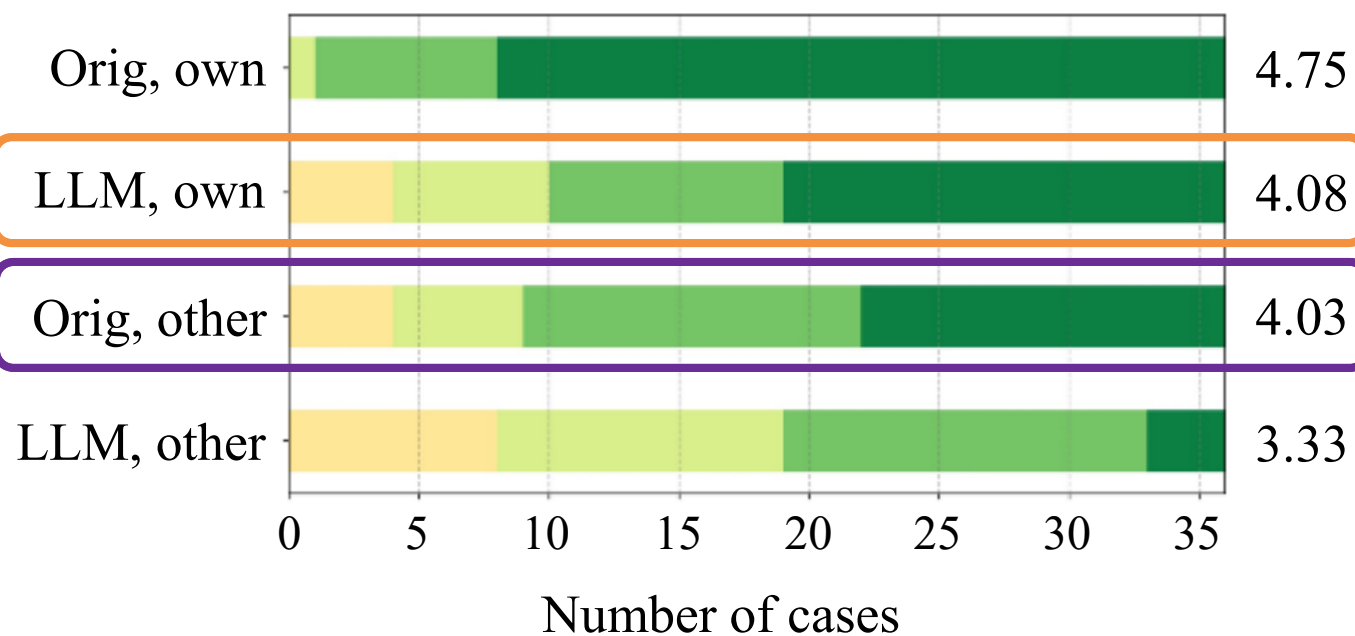
## Overall Clinical Utility



- When the physicians assessed the **clinical impressions** dictated by **other physicians**, the mean utility score was 4.03

# Results: Are the LLM impressions useful?

## Overall Clinical Utility



- **Personalized** LLM impressions are comparable to other physicians' impressions

# Results: Common Mistakes Committed by LLM

Quality dimensions	Clinical impressions	LLM impressions
Additions	2.94	2.89
Omissions	2.94	2.81
Factual correctness	2.97	2.58
Clarity and Organization	2.92	2.92
Interpretive and technical jargon	2.94	2.78
Recommendations	3.00	2.69



# Results: Common Mistakes Committed by LLM

Quality dimensions	Clinical impressions	LLM impressions
Additions	2.94	2.89
Omissions	2.94	2.81
Factual correctness	2.97	2.58*
Clarity and Organization	2.92	2.92
Interpretive and technical jargon	2.94	2.78*
Recommendations	3.00	2.69*

\* denotes the statistically significant difference ( $P < 0.05$ )

# Results: Sample Cases

## Original Clinical Impression

- [1] **FDG PET/CT demonstrates short-term stability of the solitary nodule within the right upper lobe, which has only minimal associated FDG uptake,** not elevated above mediastinal blood pool. Given the lack of significant FDG avidity, and the suggestion of macroscopic fat within the nodule on prior diagnostic chest CT, a benign lesion such as a pulmonary hamartoma is a consideration. Could consider short-term follow-up diagnostic CT chest in 3-6 months to assess for interval change.
- [2] **No evidence of suspected FDG avid local nodal mediastinal/hilar or distant metastatic disease.**
- [3] **Mildly lobulated nodular soft tissue in the subareolar right breast, without significant FDG avidity.** Could correlate with prior mammography, if available. If not, screening mammography is recommended.

## PEGASUS-personalized Impression Overall utility score = 5

- [1] **FDG PET/CT demonstrates a 1.2 x 1.3 cm subpleural solid nodule within the anteroinferior right upper lobe with mild associated FDG uptake,** which is nonspecific by PET criteria for an infectious/inflammatory process versus low-grade primary lung malignancy. Recommend correlation with prior and/or follow-up thoracic CT to assess for interval change in size.
- [2] **No evidence of FDG avid local nodal mediastinal/hilar or distant metastatic disease.**
- [3] **Redemonstration of a mildly lobulated nodular area of soft tissue within the subareolar right breast, unchanged compared to [DATE]. This does not appear to have associated hypermetabolic activity.**

# Results: Sample Cases

## Original Clinical Impression

- [1] **FDG PET/CT demonstrates intense FDG activity associated with the area of nodular soft tissue thickening at the posterior aspect of the left breast, highly suspicious for recurrent/residual malignancy.**
- [2] **Intense FDG avidity associated with the left axillary lymphadenopathy, which is overall unchanged in size and distribution compared to the recent [DATE] CT, also highly suspicious for persistent nodal metastatic disease.**
- [3] No definite evidence of FDG avid osseous metastatic disease. Mild heterogeneous FDG activity, similar to normal bone marrow activity, corresponding to mottled sclerotic and lytic changes in the pelvis, lumbar spine and T8 vertebral body, most likely represents physiologic bone marrow activity at sites of previously treated osseous metastatic disease.

## PEGASUS-personalized Impression Overall utility score = 2

- [1] **FDG PET/CT demonstrates intense FDG activity associated with the enlarging nodular soft tissue at the posterior aspect of the left breast glandular tissue**, which appears to be enlarging on previous CT examinations. **This finding is highly suspicious for recurrent malignancy.**
- [2] **Moderate-intense FDG avid left axillary lymphadenopathy, similar in size and distribution compared to recent chest CT from [DATE], compatible with nodal metastatic disease.**
- [3] Heterogeneous mild to moderate FDG uptake associated with sclerotic and lytic osseous changes in the pelvis, left clavicle, and T8 vertebral body, with no definite correlative CT bone abnormality on our corresponding low-dose noncontrast CT. These findings are nonspecific but favored to represent posttreatment related inflammatory change rather than residual/recurrent disease. Recommend attention to these sites on follow-up imaging.
- [4] No evidence of FDG-avid distant metastatic disease in the chest, abdomen, or pelvis.

# Results: Sample Cases

## Original Clinical Impression

[1] FDG PET/CT demonstrates intense FDG activity associated with the area of nodular soft tissue thickening at the posterior aspect of the left breast, highly suspicious for recurrent/residual malignancy.

[2] Intense FDG avidity associated with the left axillary lymphadenopathy, which is overall unchanged in size and distribution compared to the recent [DATE] CT, also highly suspicious for persistent nodal metastatic disease.

[3] No definite evidence of FDG avid osseous metastatic disease. Mild heterogeneous FDG activity, similar to normal bone marrow activity, corresponding to mottled sclerotic and lytic changes in the pelvis, lumbar spine and T8 vertebral body, **most likely represents physiologic bone marrow activity at sites of previously treated osseous metastatic disease.**

## PEGASUS-personalized Impression Overall utility score = 2

[1] FDG PET/CT demonstrates intense FDG activity associated with the enlarging nodular soft tissue at the posterior aspect of the left breast glandular tissue, which appears to be enlarging on previous CT examinations. This finding is highly suspicious for recurrent malignancy.

[2] Moderate-intense FDG avid left axillary lymphadenopathy, similar in size and distribution compared to recent chest CT from [DATE], compatible with nodal metastatic disease.

[3] Heterogeneous mild to moderate FDG uptake associated with sclerotic and lytic osseous changes in the pelvis, left clavicle, and T8 vertebral body, with no definite correlative CT bone abnormality on our corresponding low-dose noncontrast CT. **These findings are nonspecific but favored to represent posttreatment related inflammatory change rather than residual/recurrent disease.** Recommend attention to these sites on follow-up imaging.

[4] No evidence of FDG-avid distant metastatic disease in the chest, abdomen, or pelvis.

**Factual inaccuracy (minor)**

# Results: Sample Cases

## Original Clinical Impression

[1] FDG PET/CT demonstrates intense FDG activity associated with the area of nodular soft tissue thickening at the posterior aspect of the left breast, highly suspicious for recurrent/residual malignancy.

[2] Intense FDG avidity associated with the left axillary lymphadenopathy, which is overall unchanged in size and distribution compared to the recent [DATE] CT, also highly suspicious for persistent nodal metastatic disease.

[3] No definite evidence of FDG avid osseous metastatic disease. Mild heterogeneous FDG activity, similar to normal bone marrow activity, corresponding to mottled sclerotic and lytic changes in the pelvis, lumbar spine and T8 vertebral body, most likely represents physiologic bone marrow activity at sites of previously treated osseous metastatic disease.

## PEGASUS-personalized Impression Overall utility score = 2

[1] FDG PET/CT demonstrates intense FDG activity associated with the enlarging nodular soft tissue at the posterior aspect of the left breast glandular tissue, which appears to be enlarging on previous CT examinations. This finding is highly suspicious for recurrent malignancy.

[2] Moderate-intense FDG avid left axillary lymphadenopathy, similar in size and distribution compared to recent chest CT from [DATE], compatible with nodal metastatic disease.

[3] Heterogeneous mild to moderate FDG uptake associated with sclerotic and lytic osseous changes in the pelvis, left clavicle, and T8 vertebral body, with no definite correlative CT bone abnormality on our corresponding low-dose noncontrast CT. These findings are nonspecific but favored to represent posttreatment related inflammatory change rather than residual/recurrent disease. Recommend attention to these sites on follow-up imaging.

[4] **No evidence of FDG-avid distant metastatic disease in the chest, abdomen, or pelvis.**

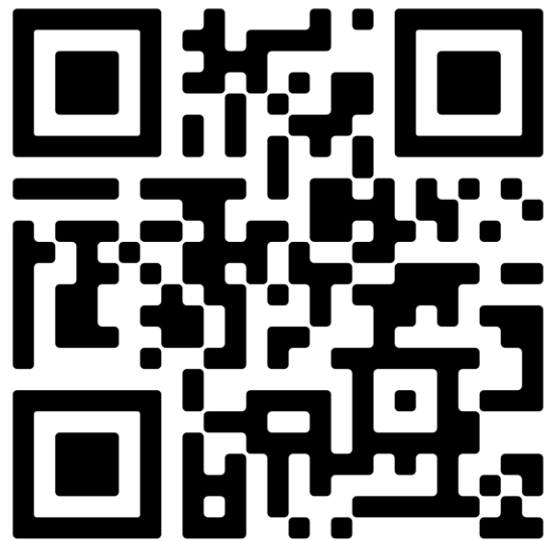
**Factual inaccuracy (major)**

# Summary

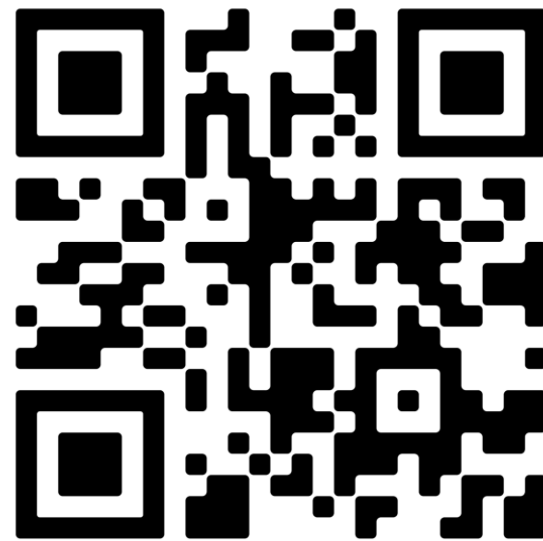
- **Domain-adapted BARTScore and PEGASUScore** exhibited the highest correlations with physician preferences, yet they did not reach the level of inter-reader correlation.
- SOTA encoder-decoder models performed similarly across evaluation metrics. **PEGASUS** emerged as the top-performing model.
- **The large majority** of the LLM impressions were **clinically acceptable**.
- Three physicians considered the overall utility of **personalized PEGASUS-generated impressions** to be comparable to **clinical impressions** dictated by **other physicians**.



Thank you  
xtie@wisc.edu



arXiv



HF model

