Pragmatic Radiology Report Generation: Exploring Reproducibility and Extensions

Ethan Rasmussen

Department of Computer Science
University of Illinois at Urbana-Champaign
Email: ethanmr3@illinois.edu

Abstract

This paper explores the reproduction and extension of *Pragmatic Radiology Report Generation* (1).

Video Presentation: https://mediaspace.illinois.edu/media/t/1_26xbhxx1

Code Repository: https://github.com/ethanrasmussen/llm_radiology

Introduction & Reproducibility

Radiology report generation—the task of automatically producing a coherent, clinically useful narrative given imaging studies—has seen significant progress with transformer based models. However, standard image to text formulations often overlook the *pragmatic* context in which radiologists operate, leading to omitted negative findings or hallucinations of uninferable details. *Pragmatic Radiology Report Generation* (1) introduces a **pragmatic perspective**, showing that the *indication* (the clinical reason for imaging) drives the inclusion of negative observations and that cleaning ground truth reports of uninferable information is critical for reliable evaluation. By jointly conditioning on both the image and the indication—and by developing novel evaluation metrics that disentangle positive correctness from negative mention accuracy—their LLaMA based model achieves substantial gains in both traditional and pragmatics inspired metrics, compared to previous methods.

Scope of Reproducibility:

- **Dataset preprocessing:** I replicated the original authors' cleaning framework using details from their paper, applying the same pragmatic report cleaning rules on the MIMIC-CXR dataset.
- **Model fine tuning:** I initially replicated their Llama2 7B fine tuning, then extended to Llama3. Due to computational limitations, I had to modify the training & fine-tuning code to utilize quantized low-rank adaptation (qLoRA).
- **Baselines and extension:** I compared against retrieval based baselines and two external radiology LLMs (*ClinicalGPT* (3) and *Radiology-Llama2* (4)), and further evaluated different prompting styles.

I further extend the initial study by:

1. **Applying the same pragmatic fine tuning to Llama3**, using a subset of the original MIMIC-CXR dataset (15,000 studies) due to storage and compute limits on a single A100 GPU in Google Colab.
2. **Benchmarking ClinicalGPT and Radiology Llama2** under identical prompts and cleaning.
3. **Exploring advanced prompting**, such as chain-of-thought and structured (2) prompts, alongside base instruction prompting.

Methodology

Environment:

- **Language & Runtime:** Python 3.10 (Google Colab standard).
- **Hardware:** Single NVIDIA A100 (40GB VRAM) GPU for reproduction, as opposed to the original experiment's 4 x A100 80GB GPUs.
- **Key Dependencies:**
  - **Primary Libraries:** PyTorch, Transformers, Numpy
  - **For qLoRA:** PEFT, bitsandbytes
  - **Labeling & Data Cleaning:** CheXbert, deepspeed
  - **Fine-Tuning:** OpenAI, Huggingface, Accelerate, Tokenizers
  - **Others:** Google, Pandas, Pillow, F1CheXbert

Data Acquisition & Preprocessing:

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Component	Description	Format / Values	Count / Example
Patients	De-identified subjects	Patient ID (e.g. `p1234567`)	65,379 unique patients
Studies	Individual imaging exams per patient	Study UID (e.g. `s5678901`)	227,835 total studies
Images	Chest X-ray images (frontal & lateral)	DICOM (`JPEG available in MIMIC-CXR-JPG`)	377,110 images
View Positions	Acquisition perspective	Categorical: PA, AP, Lateral	e.g. PA
Reports	Free-text radiology interpretation	Plain text with sections	100–500 words
Indications	Clinical reason for the exam	Free-text	e.g. "Rule out pneumonia"

Table 1: Components of MIMIC-CXR dataset.

- **Dataset from Original Paper:** *MIMIC-CXR*(10): 377,110 chest X ray images with de-identified radiology reports. Each report yields one of four *CheXbert*(9) labels (positive, negative, uncertain, missing) over fourteen thoracic conditions (14 numerical labels per report).

- **Dataset in Reproduction:** In my reproduction, I utilized MIMIC-CXR reports paired with JPG images from the MIMIC-CXR-JPG release (chosen for storage efficiency and compatibility), matching the original authors' use of DICOMs, but adapted for JPG inputs.

- **Data Notes & Processing:**

  - **Dataset Components:** The components of the MIMIC-CXR dataset are detailed above in ’Table 1’.
  - **Subset Selection:** Due to storage constraints, I sampled 15,000 studies, rather than the full 227,827 studies used by the original paper.
  - **Cleaning Rules:** I followed the original authors' report cleaning rules (e.g., removing prior comparisons, communications, recommendations, view mentions, and procedural notes) via Flan T5 rule composition prompting.
  - **Tokenization & Formatting:** Reports and prompts structured with "### Instruction", "### Input", and "### Response", truncated to 512 tokens.
  - **Impression Pruning (Extension):** As part of an extension, I applied *impression pruning* as a data cleaning step after the initial inference. (*Note: this is a post-process step, not pre-process*) After inference, I applied a simple parser to extract only the *Impression* section from each generated report. This was not used by Nguyen et al.(1), and I collected results with & without impression pruning.

- **Dataset Download Instructions:**

  1. **Get access**
     - Create a PhysioNet account and complete the required training (CITI ”Data or Specimins Only Research” course).
     - Find the MIMIC-CXR project page, formally accept the data use agreement (DUA), and wait for your requested approval.
  2. **Download files**
     - Acquire `mimic-cxr-2.0.0-studies.csv` from the MIMIC-CXR project page.
     - Use `wget` or the PhysioNet API client to download the image files and plain-text reports from the project page. To download JPEG files rather than the default DICOM files, download from the MIMIC-CXR-JPG project page.
  3. **Organize & process**
     - Unpack/unzip the downloaded files, and re-organize them to your desired directory structure. The directory structure used in this reproduction is detailed within the codebase (link in abstract section).
     - If necessary for your use case, resize or crop images, normalize pixel intensities, and augment images.
     - Tokenize and clean the plain-text reports. This may not be needed in all use cases, but is critical for the experiment outlined in this paper.

Model:

- **Pragmatic LLaMA (Original):** Fine tuned LLaMA 7B on clean impressions, proposed by Nguyen et al.(1). I replicated this model in my reproduction, conditioning on indication & predicted positive conditions using quantized low rank adaptation (qLoRA). While qLoRA wasn't used in the original paper, it was necessary for my reproduction due to computational resource constraints.

  - **Model Architecture:** LLaMA 7B Transformer; with vision component separate (ResNet 50 classifier) for positive condition prediction. The vision component's output labels feed into the LLaMA transformer's input prompt, alongside the report indication. The model will then output an impression (i.e. radiology report).
  - **Original Code:** The original training pipeline for Pragmatic LLaMA is made publicly available by the University of Chicago's Human + AI Lab (*CHAI*) at: https://github.com/ChicagoHAI/llm_radiology
  - **Visual:** See ’Figure 1’ on the following page for more detail.

- **Finetuned Pragmatic LLaMA-3 (Extension):** The original paper utilized LLaMA-2-7B, finetuning it on MIMIC-CXR via Alpaca prompts. As an extension, I performed a similar finetuning processing, using LLaMA-3.1-8B-Instruct as a base model with quantized Low Rank Adaptation (qLoRA).

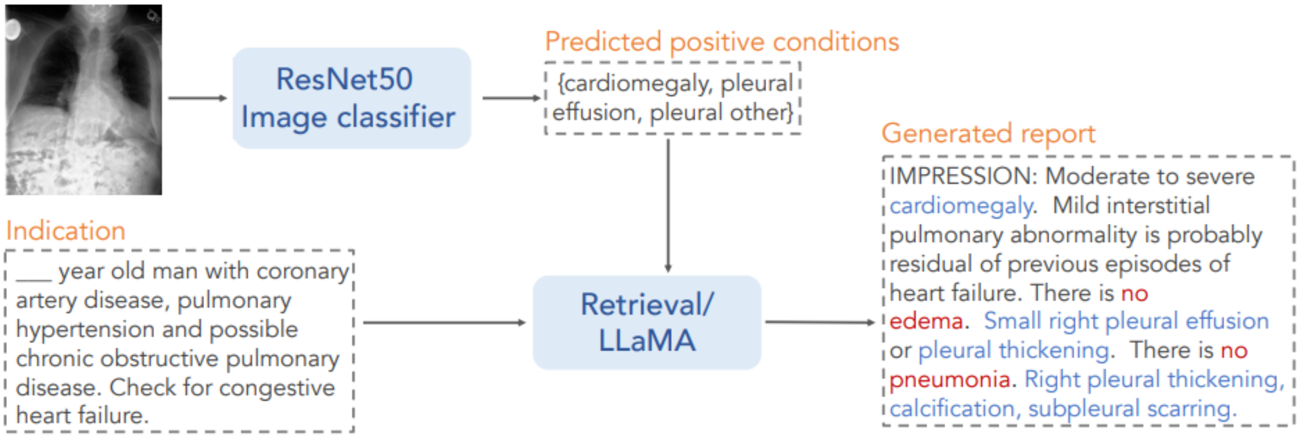
ResNet50
Image classifier
Predicted positive conditions
{cardiomegaly, pleural effusion, pleural other}
Indication
___ year old man with coronary artery disease, pulmonary hypertension and possible chronic obstructive pulmonary disease. Check for congestive heart failure.
Retrieval/
LLaMA
Generated report
IMPRESSION: Moderate to severe cardiomegaly. Mild interstitial pulmonary abnormality is probably residual of previous episodes of heart failure. There is no edema. Small right pleural effusion or pleural thickening. There is no pneumonia. Right pleural thickening, calcification, subpleural scarring.

Figure 1: Model overview from original paper. Positive findings in blue, and negative mentions in red.

Training:

- **Hyperparameters Used in Reproduction:**
  - **Batch size** = 1
  - **Gradient accumulation steps** = 4
  - **Learning rate** = $1 * 10^{-4}$
- **Computational Requirements:**
  - **GPU unit** = Single NVIDIA A100 (40GB VRAM)
  - **Epochs** = 5
  - **Epoch runtime** = approx. 1 hour
  - **Approximate GPU hours** = approx. 5 hours
- **Loss Function:** Standard cross-entropy over tokens; qLoRA parameters updated while keeping base weights frozen.
  - **Formula:** $L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{V} y_{i,j}\log(p_{i,j})$

Evaluation:

I employed both **standard** and **pragmatics-inspired** metrics (evaluated on the *Impression* only):

1. **Positive F1 / F1-5:**
   - **Explanation:** Positive F1 measure the correctness of positive condition mentions. F1-5 specifically measures for the five most frequent positive labels.
   - **Formula (F1):** $F1^{+} = \frac{2\,P^{+}\,R^{+}}{P^{+}+R^{+}}$
   - **Formula (F1-5):** $F1_5^{+} = \frac{2\,P_5^{+}\,R_5^{+}}{P_5^{+}+R_5^{+}}$
2. **Negative F1 / F1-5:**
   - **Explanation:** Negative F1 measures the accuracy of negative mentions, with F1-5 evaluating specifically for the five most frequent negated labels.
   - **Formula (F1):** $F1^{-} = \frac{2\,P^{-}\,R^{-}}{P^{-}+R^{-}}$
   - **Formula (F1-5):** $F1_5^{-} = \frac{2\,P_5^{-}\,R_5^{-}}{P_5^{-}+R_5^{-}}$
3. **BERTScore:**
   - **Explanation:** *BERTScore* (7) is used to evaluate the quality of text generation tasks, like machine translation or summarization, by measuring the semantic similarity between a generated text and a reference text.
   - **Formula:** $\mathrm{BERTScore} = F_B = \frac{2\,P_B\,R_B}{P_B+R_B}$
4. **BLEU-2:**
   - **Explanation:** *BLEU score* (8) that only considers the precision of 2-grams (two-word sequences) in the generated translation when compared to a reference translation.
   - **Formula:** $\mathrm{BLEU}_2 = \mathrm{BP}\ \exp\Big(12\big(\ln p_1 + \ln p_2\big)\Big)$
5. **Hallucination Heuristic:**
   - **Explanation:** Heuristic measuring the percentage of reports containing any uninferable information.
   - **Formula:** $HallucinationRate = 100 \times \frac{1}{N}\sum_{i=1}^{N} h_i$

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Original/Paper (full dataset)	0.360	0.137	0.307	0.417	0.05	0.127	0.158
Reproduction (subset)	0.409	0.168	0.171	0.204	0.119	0.249	0.44

Table 2: Pragmatic LLaMA Evaluation Metrics: Original vs. Reproduction on Subset

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Pragmatic LLaMA-2 (reproduced)	0.409	0.168	0.171	0.204	0.119	0.249	0.44
Pragmatic LLaMA-3 (new extension)	0.376	0.191	0.087	0.053	0.099	0.057	0.06
Radiology-Llama2 (external)	0.21	0.047	0.139	0.208	0.055	0.127	0.6
ClinicalGPT (external)	0.114	0.008	0.201	0.318	0.0	0.0	0.97

Table 3: Evaluation Metrics Across Models (on subset of MIMIC-CXR)

Results

Reproduction of the Original Pragmatic LLaMA:

I first assess how closely my reproduction on a subset of MIMIC-CXR matches the original paper's metrics. Table 2 above compares the key metrics for both the paper's full-scale model and my subset-based run, all evaluated on the *Impression* sections:

- **Language overlap (BERTScore & BLEU-2):** My subset reproduction actually yields *higher* raw scores (BLEU-2 +0.031, BERTScore +0.049). However, this is likely due to the subset, rather than a model improvement. This likely reflects the reduced diversity of a smaller test set, where fewer rare phrasings and edge-case findings dampen penalty from mismatched wording.
- **Positive condition accuracy (PosF1/PosF1-5):** Here, a *noticeable* drop (PosF1 from 0.307 to 0.171) can be observed. With fewer positive examples in the subset, the model sees less variety in condition mentions, which reduces recall on rarer positives.
- **Negative mention accuracy (NegF1/NegF1-5):** Conversely, negative-mention scores more than double (NegF1 from 0.050 to 0.119). The smaller subset happened to contain proportionally more negative-heavy cases, which improved the model's ability to learn to omit implausible positives.
- **Hallucination heuristic:** At 44.0%, hallucinations nearly triple versus the original 15.8%. This jump emphasizes the necessity of large, well-balanced test sets to reliably assess over-generation of uninferable details.

Overall, while my subset reproduction captures the **directional trends** (improved language metrics but varied clinical-accuracy trade-offs), the **absolute values differ**, underlining how dataset scale and composition materially affect both overlap and clinical metrics.

Extensions to LLaMA 3 & External Models:

Next, I apply the same fine-tuning pipeline to LLaMA-3 and benchmark two radiology-specialized LLMs on the identical 15,000-study split. Table 3 presents those results:

- **LLaMA-3 extension:** Compared to LLaMA-2, LLaMA-3 achieves the *highest BLEU-2* (0.191) and **drastically** *reduces hallucinations* to 6.0% (better than both the reproduced and original LLaMA-2 version). However, its positive-mention recall plummets (PosF1 = 0.087), suggesting that LLaMA-3's broader pretraining leads it to under-mention findings when only lightly fine-tuned on a small subset.
- **Comparison with Radiology-Llama2 & ClinicalGPT:** I compared two recently published radiology-focused LLM's, *ClinicalGPT* (3) and *Radiology-Llama2* (4), against Pragmatic LLaMA. Both domain-specialized models underperform on language metrics and exhibit very high hallucination (60% and 97%, respectively). This indicates that **specialization alone**—without pragmatic conditioning on indication and careful report cleaning—does not guarantee clinically accurate summaries. However, it also indicates a mismatch with the expected output format (i.e. output includes additional unexpected text, but is not necessarily ”wrong”). I explore this further with my introduction of ”*impression pruning*.”

Extension/Ablation – Prompt Styles & ”Impression Pruning”:

To explore the effects of prompting and post-processing, I compare three prompt types (Base, Chain-of-Thought, Structured) with and without “impression pruning” in Table 4. Impression pruning is a simple post-inference processing step I added, which first uses Flan-T5-XL to remove unnecessary information (for example, if the LLM outputs both a ”Indication” and ”Impression” section, Flan-T5-XL is instructed to output only the impression). Then, a simple text parser is applied which removes any remaining section title(s). These experiments showed the following:

- **Impression pruning** consistently **boosts negative mention scores** across all prompting styles with Pragmatic LLaMA-3, Radiology-Llama2, and ClinicalGPT, con-

Prompt/Prune Style	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Base	0.409	0.168	0.171	0.204	0.119	0.249	0.44
Base + Pruning	0.335	0.078	0.14	0.197	0.092	0.196	0.2
Chain-of-Thought	0.406	0.162	0.136	0.107	0.126	0.243	0.43
Chain-of-Thought + Pruning	0.352	0.082	0.126	0.107	0.09	0.203	0.24
Structured	0.173	0.056	0.135	0.123	0.083	0.216	0.33
Structured + Pruning	0.331	0.078	0.115	0.096	0.08	0.207	0.12

Table 4: Reproduced Pragmatic LLaMA: Results Across Prompting Styles, with & without Impression Pruning

firming that extracting only the *impression* improves clinical metrics. However, this doesn't appear to be true for Pragmatic LLaMA-2.

- **Impression pruning** has significant **positive impact on hallucination rates** as well, implying that this is an effective way to prevent models from presenting uninferable information.
- **Structured prompting**, particularly when *combined with impression pruning*, **reduces hallucination rates** across nearly all models.

Additional Tables & Evaluation Metrics:

To see the evaluation metrics compared across all five models (original/paper, reproduced, LLaMA-3, Radiology-Llama2, & ClinicalGPT), please find the tables beyond the bibliography.

Discussion

My reproducibility study confirms that the **pragmatic perspective**—conditioning on both *image* and *indication*, alongside targeted cleaning of ground truth—significantly enhances report generation, reducing hallucinations and boosting clinical-efficacy metrics.

Reproducibility: Dataset processing and model fine tuning steps were straightforward to replicate using the provided codebase, with several modifications to run in Google Colab and with less compute.

Ease: Fine-tuning the model was relatively easy given the authors' original approach. Minimal adaptation was required to implement quantized low-rank adaptation (qLoRA), which was possible by leveraging Huggingface's Trainer API and the provided report cleaning framework.

Difficulties:

- Managing the MIMIC-CXR subset and ensuring consistency in train/test splits demanded careful bookkeeping due to distribution shifts in negative mentions.
- Reproduction following the authors' original training hyperparameters was simply not possible, due to my reduced hardware access.
- Some steps required manual movement of files within my directory, and the provided codebase had to be altered in several locations due to hard-coded file paths and broken relative paths.

Recommendations to future authors:

1. Provide Docker images and/or more detailed environment setup instructions, to allow for easier reproduction.
2. Explore the differences between the performance of natively multi-modal models against text-only models (such as Pragmatic Llama, which only takes image input indirectly), evaluating accuracy, clinical efficacy, and hallucination rates between the two model classes.
3. Explore newer reasoning models, which may produce higher quality results, but may also require additional data post-processing and pruning.

Conclusion

My reproduction, as well as the original work by Nguyen et al.(1), demonstrates that radiology report generation benefits profoundly from a pragmatic re-formulation: by explicitly incorporating indication in inputs and rigorously cleaning report data, LLaMA based models achieve state of the art performance with greatly reduced hallucinations. Extensions to newer LLaMA variants and comparisons with other domain specific models highlight both the robustness of the pragmatic approach and the continuing need for precise evaluation practices. Such reproducible pipelines pave the way for reliable, clinically trustworthy AI assistants in radiology and beyond.

Author Contributions

This paper reproduction project was completed in its entirety by **Ethan Rasmussen** (*ethanmr3@illinois.edu*), from the University of Illinois at Urbana-Champaign. This includes this full paper, as well as code created in the reproduction of *Pragmatic Radiology Report Generation* (1).

Additional/Glossary Tables

Full evaluation metrics for all models, spanning all three prompting styles, with and without impression pruning, are available on the page beyond the bibliography. Please note that the results for **”Pragmatic LLaMA (original/paper)”** are intended as a reference point. Unlike the metrics from the other four models, they weren’t calculated against the MIMIC-CXR subset. Instead, they were taken verbatim from the paper *Pragmatic Radiology Report Generation*.(1)

References

[1] Nguyen, D., Chen, C., He, H., & Tan, C. (2023). Pragmatic Radiology Report Generation. *Proceedings of Machine Learning Research (ML4H)*, 225, 1–16. 2

[2] Sonoda, Y. et al. (2024). Structured clinical reasoning prompt enhances LLM’s diagnostic capabilities in diagnosis please quiz cases. *Japanese Journal of Radiology*, 43:586–592 3

[3] Wang, G. et al. (2023). ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data. *arXiv*, 2306.09968v1 4

[4] Liu, Z. et al. (2023). Radiology Llama2: Best in Class Large Language Model for Radiology. *arXiv*, 2309.06419v1 5

[5] Wang, L., Chen, X., Deng, X. et al. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*. 7, 41 6

[6] Jain, S. et al. (2021). RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *arXiv*, 2106.14463 7

[7] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *arXiv*, 1904.09675 8

[8] Papineni, K., Roukos, S., Ward, T. Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the Association for Computational Linguistics (ACL)*, 40:311–318 9

[9] Smit, A. et al. (2020). CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *arXiv*, 2004.09167 10

[10] Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Nature: Scientific Data*. 6, 317

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Pragmatic LLaMA (original/paper)	0.36	0.137	0.307	0.417	0.05	0.127	0.158
Pragmatic LLaMA (reproduced/subset)	0.409	0.168	0.171	0.204	0.119	0.249	0.44
Pragmatic LLaMA-3 (extension)	0.376	0.191	0.087	0.053	0.099	0.057	0.06
Radiology-Llama2 (external)	0.21	0.047	0.139	0.208	0.055	0.127	0.6
ClinicalGPT (external)	0.114	0.008	0.201	0.318	0.0	0.0	0.97

Table 5: All Model Results: Base Prompting, no Pruning

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Pragmatic LLaMA (original/paper)	0.36	0.137	0.307	0.417	0.05	0.127	0.158
Pragmatic LLaMA (reproduced/subset)	0.335	0.078	0.14	0.197	0.092	0.196	0.20
Pragmatic LLaMA-3 (extension)	0.312	0.02	0.074	0.031	0.153	0.175	0.04
Radiology-Llama2 (external)	0.234	0.058	0.13	0.156	0.062	0.095	0.40
ClinicalGPT (external)	0.114	0.02	0.184	0.265	0.019	0.021	0.38

Table 6: All Model Results: Base Prompting, with Impression Pruning

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Pragmatic LLaMA (original/paper)	0.36	0.137	0.307	0.417	0.05	0.127	0.158
Pragmatic LLaMA (reproduced/subset)	0.406	0.162	0.136	0.107	0.126	0.243	0.43
Pragmatic LLaMA-3 (extension)	0.291	0.087	0.195	0.23	0.06	0.156	0.37
Radiology-Llama2 (external)	0.225	0.047	0.162	0.254	0.054	0.095	0.44
ClinicalGPT (external)	0.09	0.007	0.137	0.172	0.0	0.0	0.26

Table 7: All Model Results: Chain-of-Thought Prompting, no Pruning

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Pragmatic LLaMA (original/paper)	0.36	0.137	0.307	0.417	0.05	0.127	0.158
Pragmatic LLaMA (reproduced/subset)	0.352	0.082	0.126	0.107	0.09	0.203	0.24
Pragmatic LLaMA-3 (extension)	0.304	0.08	0.196	0.24	0.083	0.20	0.28
Radiology-Llama2 (external)	0.228	0.049	0.128	0.199	0.047	0.122	0.41
ClinicalGPT (external)	0.098	0.008	0.14	0.175	0.0	0.0	0.25

Table 8: All Model Results: Chain-of-Thought Prompting, with Impression Pruning

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Pragmatic LLaMA (original/paper)	0.36	0.137	0.307	0.417	0.05	0.127	0.158
Pragmatic LLaMA (reproduced/subset)	0.173	0.056	0.135	0.123	0.083	0.216	0.33
Pragmatic LLaMA-3 (extension)	0.132	0.046	0.086	0.077	0.007	0.018	0.04
Radiology-Llama2 (external)	-0.009	0.005	0.059	0.054	0.0	0.0	0.12
ClinicalGPT (external)	0.001	0.004	0.052	0.015	0.01	0.025	0.03

Table 9: All Model Results: Structured Prompting, no Pruning

Model	BERTScore	BLEU-2	Pos.F1	Pos.F1-5	Neg.F1	Neg.F1-5	Hallucination
Pragmatic LLaMA (original/paper)	0.36	0.137	0.307	0.417	0.05	0.127	0.158
Pragmatic LLaMA (reproduced/subset)	0.331	0.078	0.115	0.096	0.08	0.207	0.12
Pragmatic LLaMA-3 (extension)	0.315	0.111	0.076	0.053	0.014	0.036	0.05
Radiology-Llama2 (external)	0.20	0.031	0.071	0.055	0.05	0.095	0.05
ClinicalGPT (external)	0.232	0.036	0.035	0.0	0.016	0.042	0.05

Table 10: All Model Results: Structured Prompting, with Impression Pruning