SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Does Size Really Matter? Comparing llama 3 vs 3.1

Suyash Khubchandani, MD, MHA, CARPL.ai, Inc.; Amit Kumar; Vasantha K. Venugopal, MD

Introduction/Background

Prompt engineering in radiological reports involves crafting inputs to guide AI models in generating accurate and useful outputs. This technique helps automate downstream tasks on radiological reports. However, its effectiveness is limited by context length constraints, which can restrict the amount of information the model can process and integrate simultaneously. Commercial LLMs like ChatGPT are bounded by max context window of 16 k tokens where 1.5 tokens are approximately used for one word. For large radiological reports like MRI, it becomes difficult to perform few shots learning when context window is of limited size. In this study we have used llama 3.1 open source model with context size of 128 k tokens against llama 3.

Methods/Intervention

We collected 1000 radiology reports annotated by expert radiologists, classifying each report into findings: acute infarct, intra-axial tumor, intra-axial hemorrhage, extra-axial tumor, and extra-axial hemorrhage. These reports were evaluated using the Llama 3 and Llama 3.1 models under zero-shot, one-shot, and few-shot learning settings. For zero-shot learning, the models received no prior examples. In one-shot learning, each model received one example per finding, and for few-shot learning, multiple examples were provided. The models' performance was compared to determine the impact of different learning strategies on their accuracy in identifying radiological findings.

Results/Outcome

In our study, we observed a significant improvement in accuracy for the llama 3.1 model. For llama 3, its accuracy improved from 66-97% with zero short learning to 95-98 % with few-shot learning, as further prompt engineering was constrained by the limited number of context size. However, by incorporating additional examples into the prompt, LLama 3.1 demonstrated an accuracy of 99.5%.

Conclusion

This enhancement underscores the model's capability to learn effectively from expanded datasets, highlighting the importance of large context windows in achieving superior predictive accuracy for larger reports.

Statement of Impact

Larger context LLMs are better suited for analyzing radiological MRI reports as they can handle more detailed and comprehensive patient data, ensuring accurate diagnosis and interpretation. Their ability to process extensive contextual information enhances downstream tasks, providing more reliable and insightful outcomes compared to smaller context-size LLMs.

Keywords

large language models; natural language processing; radiological reports; classification; context size