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

GPT-Based Automated Classification and Labeling of Surgical Renal Pathology Reports

Satvik Tripathi, University of Pennsylvania; Rithvik Sukumaran; Dana Alkhulaifat, MD; Charles M. Chambers, MCIT; Darco Lalevic, MCIT; Hanna Zafar, MD, MHS; Tessa S. Cook, MD, PhD, CIIP, FSIIM

Introduction/Background

Human annotation of reports to acquire high-quality data for model training can be costly and time-consuming. Leveraging automated labeling with large language models can be a valuable and cost-effective tool to streamline annotation processes. Our aim was to assess GPT-4's performance in labeling renal surgical pathology reports using various prompting-based techniques.

Methods/Intervention

Renal surgical pathology reports from three health systems (n=40) within the same state were labeled by two radiologists with 10 and 14 years of experience as "malignant," "indeterminate," "benign," or "ignore." "Ignore" was used for reports of any pathology not specifically from a renal mass. The reports were distributed equally among the four classes. Prompt engineering for GPT-4 was utilized with zero-, one-, and few-shot learning techniques to classify the reports. The main performance evaluation metric was accuracy.

Results/Outcome

GPT-4 achieved 70%, 77.5%, and 92.5% accuracy with zero-, one- and few-shot learning, respectively. Theincorrect classification was the highest in the "Indeterminate" (n = 4) class for zero-shot prompting and the "Ignore" class for one- and few-shot prompting techniques (n = 5 and 2, respectively). GPT-4 outperformed our existing Deep Learning-based methods.

Conclusion

GPT-4 holds the potential to classify renal surgical pathology reports with significant accuracy, even without extensive training data. The few-shot prompting technique achieved the highest accuracy, demonstrating the model's ability to adapt and learn from minimal examples. This capability could streamline the annotation process, reduce the burden on radiologists, and enable faster data processing. Furthermore, the model's performance in handling varied classes of pathology reports underscores its versatility and potential for broader applications in medical report classification.

Statement of Impact

Automatic labeling of reports can enable prompt identification of important clinical findings, leading to timely intervention and improved treatment outcomes.

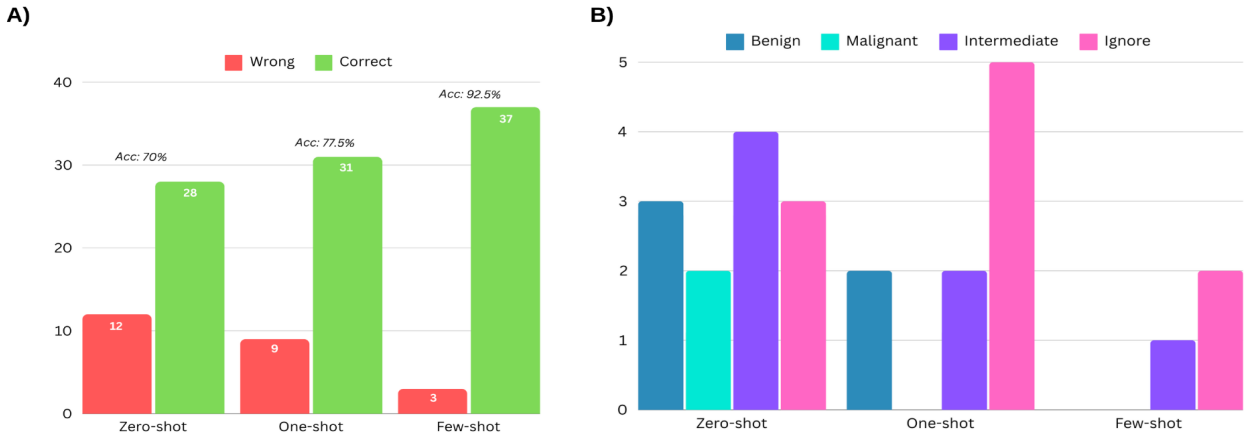

Figure 1. A) Bar chart depicting the distribution of correct and incorrect classifications using zero-, one- and few-shot prompting techniques, along with the accuracy of each technique. B) Bar chart showing the distribution of incorrect classifications for each prompting technique by class.

Keywords

Large Language Models; Automated Labeling; Pathology; Prompt Engineering