

Large Language Models in Summarizing Radiology Report Impressions for Lung Cancer in Chinese: Evaluation Study

Danqing Hu, Shanyuan Zhang, Qing Liu, Xiaofeng Zhu, Bing Liu

Submitted to: Journal of Medical Internet Research
on: August 19, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 29

 Figures 30

 Figure 1..... 31

 Figure 2..... 32

 Figure 3..... 33

 Figure 4..... 34

 Figure 5..... 35

 Multimedia Appendixes 36

 Multimedia Appendix 1..... 37

Large Language Models in Summarizing Radiology Report Impressions for Lung Cancer in Chinese: Evaluation Study

Danqing Hu^{1, 2*} PhD; Shanyuan Zhang^{3*} MD; Qing Liu^{4*} MD; Xiaofeng Zhu² MSc; Bing Liu³ MD

¹Jiangsu Key Laboratory of Intelligent Medical Image Computing School of Future Technology Nanjing University of Information Science and Technology Nanjing CN

²Zhejiang Lab Hangzhou CN

³Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education) Department of Thoracic Surgery II Peking University Cancer Hospital and Institute Beijing CN

⁴Department of Radiology Peking University Cancer Hospital and Institute Beijing CN

*these authors contributed equally

Corresponding Author:

Bing Liu MD

Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education)

Department of Thoracic Surgery II

Peking University Cancer Hospital and Institute

52 Fucheng Road

Haidian District

Beijing

CN

Abstract

Background: Large language models (LLMs), such as ChatGPT, have demonstrated impressive capabilities in various natural language processing tasks, particularly in text generation. However, their effectiveness in summarizing radiology report impressions remains uncertain.

Objective: This study aims to evaluate the capability of nine LLMs, i.e., Tongyi Qianwen, ERNIE Bot, ChatGPT, Bard, Claude, Baichuan, ChatGLM, HuatuoGPT, and ChatGLM-Med, in summarizing radiology report impressions.

Methods: We collected three types of radiology reports, i.e., CT, PET-CT, and Ultrasound reports, from Peking University Cancer Hospital and Institute. Using these reports, we created zero-shot, one-shot, and three-shot prompts with or without complete example reports as inputs to generate impressions. We employed both automatic quantitative evaluation metrics and five human evaluation metrics (completeness, correctness, conciseness, verisimilitude, and replaceability) to assess the generated impressions. Two thoracic surgeons (ZSY and LB) and one radiologist (LQ) compared the generated impressions with reference impressions, scoring them according to the five human evaluation metrics.

Results: In the automatic quantitative evaluation, ERNIE Bot, Tongyi Qianwen, and ChatGPT demonstrated the best overall performance in generating impressions for CT, PET-CT, and Ultrasound reports, respectively. In the human semantic evaluation, ERNIE Bot outperformed the other LLMs in terms of conciseness, verisimilitude, and replaceability on CT impression generation, while its completeness and correctness scores were comparable to those of other LLMs. Tongyi Qianwen excelled in PET-CT impression generation, with the highest scores for correctness, conciseness, verisimilitude, and replaceability. Claude achieved the best conciseness, verisimilitude, and replaceability scores on US impression generation, and its completeness and correctness scores are close to the best results obtained by other LLMs. The generated impressions were generally complete and correct but lacked conciseness and verisimilitude. Although one-shot and few-shot prompts improved conciseness and verisimilitude, clinicians noted a significant gap between the generated impressions and those written by radiologists.

Conclusions: Current large language models can produce radiology impressions with high completeness and correctness but fall short in conciseness and verisimilitude, indicating they cannot yet fully replace impressions written by radiologists.

(JMIR Preprints 19/08/2024:65547)

DOI: <https://doi.org/10.2196/preprints.65547>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [http://www.jmir.org/](#)

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Large Language Models in Summarizing Radiology Report Impressions for Lung Cancer in Chinese: Evaluation Study

Abstract

Background: Large language models (LLMs), such as ChatGPT, have demonstrated impressive capabilities in various natural language processing tasks, particularly in text generation. However, their effectiveness in summarizing radiology report impressions remains uncertain.

Objective: This study aims to evaluate the capability of nine LLMs, i.e., Tongyi Qianwen, ERNIE Bot, ChatGPT, Bard, Claude, Baichuan, ChatGLM, HuatuoGPT, and ChatGLM-Med, in summarizing Chinese radiology report impressions for lung cancer.

Methods: We collected 100 Chinese CT, PET-CT, and Ultrasound reports each from Peking University Cancer Hospital and Institute. All these reports were from patients with suspected or confirmed lung cancer. Using these reports, we created zero-shot, one-shot, and three-shot prompts with or without complete example reports as inputs to generate impressions. We employed both automatic quantitative evaluation metrics and five human evaluation metrics (completeness, correctness, conciseness, verisimilitude, and replaceability) to assess the generated impressions. Two thoracic surgeons (ZSY and LB) and one radiologist (LQ) compared the generated impressions with reference impressions, scoring them according to the five human evaluation metrics.

Results: In the automatic quantitative evaluation, ERNIE Bot, Tongyi Qianwen, and Claude demonstrated the best overall performance in generating impressions for CT, PET-CT, and Ultrasound reports, respectively. In the human semantic evaluation, ERNIE Bot outperformed the other LLMs in terms of conciseness, verisimilitude, and replaceability on CT impression generation, while its completeness and correctness scores were comparable to those of other LLMs. Tongyi Qianwen excelled in PET-CT impression generation, with the highest scores for correctness, conciseness, verisimilitude, and replaceability. Claude achieved the best conciseness, verisimilitude, and replaceability scores on US impression generation, and its completeness and correctness scores are close to the best results obtained by other LLMs. The generated impressions were generally complete and correct but lacked conciseness and verisimilitude. Although one-shot and few-shot prompts improved conciseness and verisimilitude, clinicians noted a significant gap between the generated impressions and those written by radiologists.

Conclusions: Current large language models can produce radiology impressions with high completeness and correctness but fall short in conciseness and verisimilitude, indicating they cannot yet fully replace impressions written by radiologists.

Keywords: Large language model; impression summarization; radiology report.

Introduction

Clinical documentation plays an indispensable role in healthcare practice, serving as a primary medium for conveying patient information in clinical settings. Clinicians routinely spend a substantial amount of time summarizing vast amounts of textual information, such as compiling diagnostic reports, writing progress notes, or synthesizing a patient's treatment history [1, 2]. With the increasing adoption of electronic medical record systems, the burden of clinical documentation expanded rapidly, contributing to clinician stress and burnout [3]. A recent study indicates that clinicians spend up to two hours on documentation for every hour of patient interaction [4]. This issue may be more prominent in countries with large populations and limited healthcare resources.

As one of the most important clinical documents, radiology reports record essential information from

patient imaging data such as CT scans, PET scans, MRIs, X-rays, and ultrasound examinations. These reports typically consist of two main sections, namely findings and impressions. The findings section presents the radiologist's observations from the images, while the impressions section provides a concise summary of the observed abnormalities and corresponding diagnoses or suspicions with tendencies. Radiology reports, especially the impressions, are crucial for patient diagnosis, disease progression assessment, and treatment planning. Impression summarization refers to using models to simulate physicians to condense lengthy and detailed findings into concise and informative impressions [5-7], which is a key application of text summarization in the medical field [8]. This technique may greatly relieve the workload of radiologists and reduce their possible errors and omissions, thereby improving the accuracy of clinical evaluations [9-11].

Recently, large language models (LLMs) such as ChatGPT [12] and GPT-4 [13] have captured worldwide attention due to their astonishing text-generation capabilities. Through pre-training on vast amounts of data, LLMs demonstrate remarkable performance on unseen downstream tasks using zero-shot, one-shot, or few-shot prompts without parameter updates [14]. By reinforcement fine-tuning with human feedback (RLHF) [15], the LLMs are further guaranteed to produce harmless and unbiased content that aligns with human expectations. The great success of prompt-based LLMs has led to a paradigm shift in NLP research [16-21], thereby bringing new opportunities for radiology report impression summarization.

To enable LLMs to effectively perform specific tasks like impression summarization, it is essential to craft instructions that LLMs can accurately interpret and execute, a process known as prompt engineering. Few-shot prompting is one of the most effective prompt engineering techniques, which supplies the LLMs with some examples to refine their responses. When applied to impression summarization, few-shot prompting can help the LLMs capture the critical information radiologists prioritize and the writing style they use, which is crucial for generating more realistic impressions. Although some studies have applied prompt-based LLMs to this task [2, 22, 23], they only focus on limited types of reports, typically the X-ray reports, and lack detailed clinical expert evaluation of the generated results [22, 23] or only evaluate the LLMs on English reports in a zero-shot manner [2].

In this study, we conduct a systematic study to explore the capability of prompt-based LLMs in summarizing the impressions of various types of Chinese radiology reports using zero-shot and few-shot prompts. By leveraging automatic quantitative and clinical expert evaluations, we aim to clarify the current status of LLMs in Chinese radiology report impression summarization and the gap between the current achievements and requirements for application in clinical practice.

Methods

Study Overview

To evaluate the LLMs for impression summarization, we first collect three types of Chinese radiology reports, i.e., PET-CT, CT, and ultrasound reports from Peking University Cancer Hospital and Institute. Using the collected reports, we evaluate the zero-shot, one-shot, and three-shot performance of impression summarization of five commercially available LLMs, including Tongyi Qianwen, ERNIE Bot, ChatGPT, Bard, and Claude, and four open source LLMs, including Baichuan, ChatGLM, HuatuoGPT, and ChatGLM-Med. Since the LLMs' output not only contains the generated impression but also contains some content unrelated to the impression, such as some explanations about how they generate the impressions or disclaimers indicating the generated impressions are for reference only. We manually extract the impression-related content from the outputs for the automatic quantitative and human semantic evaluations. The overall pipeline is shown

in Figure 1.

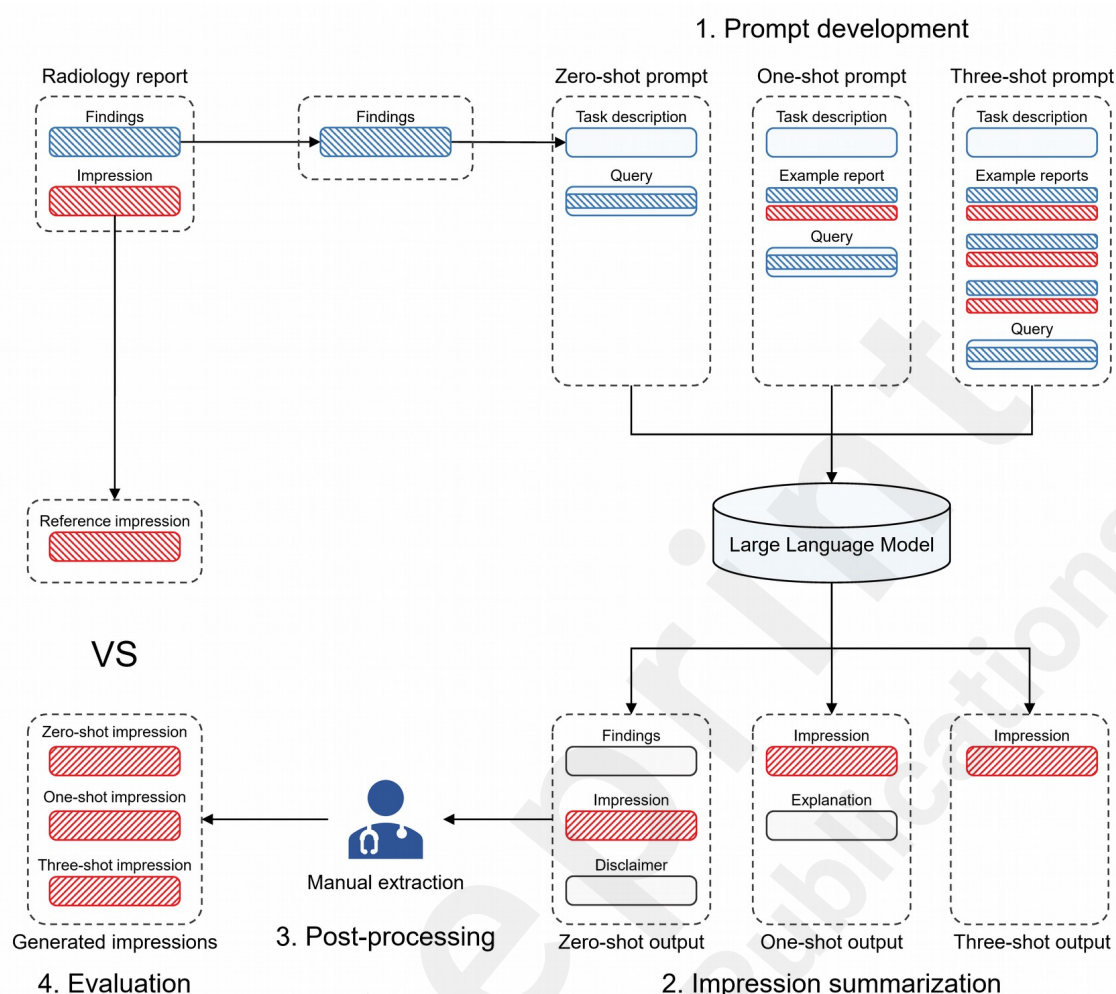


Figure 1.

Study overview: 1. Prompt development: we first developed the zero-shot, one-shot, and three-shot prompts for the collected radiology reports using the prompt templates. 2. Impression summarization: we input the prompts into LLMs and collected the outputs with summarized impressions. 3. Post-processing: we manually extracted the impression-related contents from the LLMs' outputs. 4. Evaluation: we conducted automatic and human evaluations of the generated impressions.

Materials

We collected three types of radiology reports, i.e., whole body PET-CT, chest CT, and neck and abdomen ultrasound (US) reports from Peking University Cancer Hospital and Institute. The relevant patients are all outpatients and inpatients of the Department of Thoracic Surgery II with suspected or confirmed lung cancer. After removing the incomplete reports, we finally obtained 867 PET-CT reports, 819 CT reports, and 1487 ultrasound reports. We randomly selected 100 reports from each type of report for automatic quantitative and human evaluations. We manually reviewed the selected reports to make sure that no patient identification information was recorded in these reports to protect the patient privacy.

Large Language Models

To conduct a comprehensive evaluation, we select five commercially available and four open source

LLMs with different architectures and parameter sizes. The introduction of the selected LLMs are listed below. Table 1 shows the summary of those LLMs.

Tongyi Qianwen. Tongyi Qianwen is an LLM chat product developed by Alibaba Cloud. The latest Tongyi Qianwen 2.0 extends the Qwen model [24] to a few hundred billion parameters, achieving a substantial upgrade from its predecessor in understanding complex instructions, reasoning, memorizing, and preventing hallucinations. We used Tongyi Qianwen v2.1.1 [25] to generate the impressions.

ERNIE Bot. ERNIE Bot (Wenxin Yiyan) is an LLM chat product developed by Baidu based on their ERNIE (Enhanced Representation through Knowledge Integration)[26] and PLATO (Pre-trained Dialogue Generation Model)[27] models. Based on the supervised fine-tuning, RLHF, and knowledge, search, dialogue enhancements, the ERNIE Bot achieves a more precise understanding of Chinese language and its practical applications. We used ERNIE Bot v2.5.2 [28] to generate the impressions.

ChatGPT. ChatGPT is the most impactful LLM developed by OpenAI, raising the trend of prompt-based LLMs worldwide. ChatGPT is an advanced version of instructionGPT [15], which first fine-tunes GPT-3 [14] using human-written demonstrations of the desired output to prompts and then further fine-tuning the model through the RLHF strategy to align language models with user intent. We accessed the ChatGPT via website interface [29] to obtain the generated impressions before January 11, 2024.

Bard. Bard is an LLM chat product powered by PaLM 2 (Pathways Language Model 2) [30] developed by Google AI. PaLM 2 is a transformer-based model trained using a mixture of objectives and multilingual datasets, achieving better performances on natural language generation, code generation, translation, and reasoning than its predecessor, PaLM [31]. We accessed the Bard via website interface [32] to obtain the generated impressions before January 12, 2024.

Claude. Claude is a series of LLM chat products developed by Anthropic. Claude models are transformers pre-trained to predict the next word in large amounts of text. They are fine-tuned with Constitutional AI [33] to make them harmless and helpful without relying on extensive human feedback. In this study, we accessed the latest Claude 3.5 Sonnet via the website interface [34] before October 12, 2024 to produce the impressions.

Baichuan. Baichuan-13B is an open-source LLM developed by Baichuan Intelligence. The baichuan-13B model has 130 billion parameters trained on 1.4 trillion tokens. It supports both Chinese and English and achieves competitive performance in standard Chinese and English benchmarks among models of its size. We used the Baichuan-13B-Chat [35] to generate the impressions.

ChatGLM. ChatGLM3-6B is the latest open-source model in the ChatGLM [36] series developed by Tsinghua University. The ChatGLM3-6B has a more powerful base model trained on a more diverse dataset, sufficient training steps, and a more reasonable training strategy, showing strong performance on language understanding, reasoning, coding, etc. We used the ChatGLM3-6b [37] to generated the impressions.

HuatuoGPT. HuatuoGPT [38] is an open-source LLM developed by the Shenzhen Research Institute of Big Data. HuatuoGPT-7B first uses the Baichuan-7B as the backbone model and then uses the distilled data from ChatGPT and real-world data from doctors to supervised fine-tuning and reinforcement learning with mixed feedback to achieve state-of-the-art results in performing medical

consultation. We used the HuatuoGPT-7B [39] to generate the impressions.

ChatGLM-Med. ChatGLM-Med is an open-source LLM developed by the Harbin Institution of Technology. The ChatGLM-Med employs the ChatGLM-6B as the base model and fine-tunes on a Chinese medical instruction dataset developed by a medical knowledge graph and GPT-3.5 to improve better question-answering results in the medical field. We used the ChatGLM-Med [40] to generate the impressions.

Table 1. The detailed information of the LLMs used in this study.

Model	Developer	Type	License	Language support	Maximum input character limits	Version/Last access date
Tongyi Qianwen	Alibaba Cloud	General	Commercially available	Chinese, English	10000 characters ^a	Tongyi Qianwen v2.1.1/ January 11, 2024
ERNIE Bot	Baidu	General	Commercially available	Chinese, English	2000 characters ^a	ERNIE Bot v2.5.2/ January 11, 2024
ChatGPT	OpenAI	General	Commercially available	Chinese, English	8192 tokens ^a	GPT-3.5/ January 11, 2024
Bard	Google	General	Commercially available	Chinese, English	32000 characters ^a	Bard/January 12, 2024
Claude	Anthropic	General	Commercially available	Chinese, English	200000 characters	Claude 3.5 Sonnet/October 12, 2024
Baichuan	Baichuan Intelligence	General	Open-source	Chinese, English	4096 tokens	Baichuan-13B-Chat/ January 19, 2024
ChatGLM	Tsinghua University	General	Open-source	Chinese, English	8192 tokens	ChatGLM3-6B/ January 19, 2024
HuatuoGPT	Shenzhen Research Institute of Big Data	Medical	Open-source	Chinese, English	4096 tokens	HuatuoGPT-7B/ January 19, 2024
ChatGLM-Med	Harbin Institution of Technology	Medical	Open-source	Chinese, English	2048 tokens	ChatGLM-Med/ January 19, 2024

^aThe maximum length of text that can be entered on the web interface.

Impression Summarization using LLMs

To explore the capability of LLMs to summarize the impression in a zero-shot or few-shot manner, we first design the zero-, one-, and three-shot prompts, as shown in Figure 2. The language of the prompts is also listed in the Multimedia Appendix I. The zero-shot prompt is composed of two main components, i.e., task description and query. In the task description, we first define the role of the LLMs as a radiologist and then specify the task of generating impressions based on the findings from CT, PET-CT, or ultrasound radiology reports. Additionally, we include an instruction emphasizing the need for concise and clear impressions. In the query section, we format the findings of the reports using a “Findings:<xxx>” template and provide an “Impressions:<>” template with an empty placeholder “<>”, where the LLMs are expected to generate and insert the generated impressions. For the few-shot prompts, we insert one or three example reports between the task description and query sections as the one-shot and three-shot prompts, respectively. The example reports are formatted using the same “Findings:<xxx>” and “Impressions:<xxx>” templates to ensure that LLMs can learn and adhere to the desired response format. These prompts allow us to investigate the impact of the number of examples on the quality of the impression generation. Note that the example reports are randomly selected from the collected reports and different from the report in the query. Since the maximum input text lengths supported by LLMs are different, to fairly evaluate and compare the performance of LLMs, we do not conduct experiments when some prompts exceed the maximum input text length of LLMs (one-shot PET-CT prompt for ERNIE Bot, ChatGLM_Med, three-shot PET-CT prompt for ERINE Bot, Baichuan, HuatuoGPT, and ChatGLM_Med).

Using the developed prompts, we collect the outputs of the five commercially available LLMs from their corresponding websites manually. And we deploy the four open-source LLMs with the default hyperparameters on our server to obtain their outputs. Note that, besides the summarized impression, the LLMs usually generate some other content such as the findings, the future examination advice, the explanation of the response, etc. To accurately evaluate the generated impressions, we conduct a post-processing procedure to remove the unrelated content from the outputs to keep the impressions only for further quantitative and human evaluations.

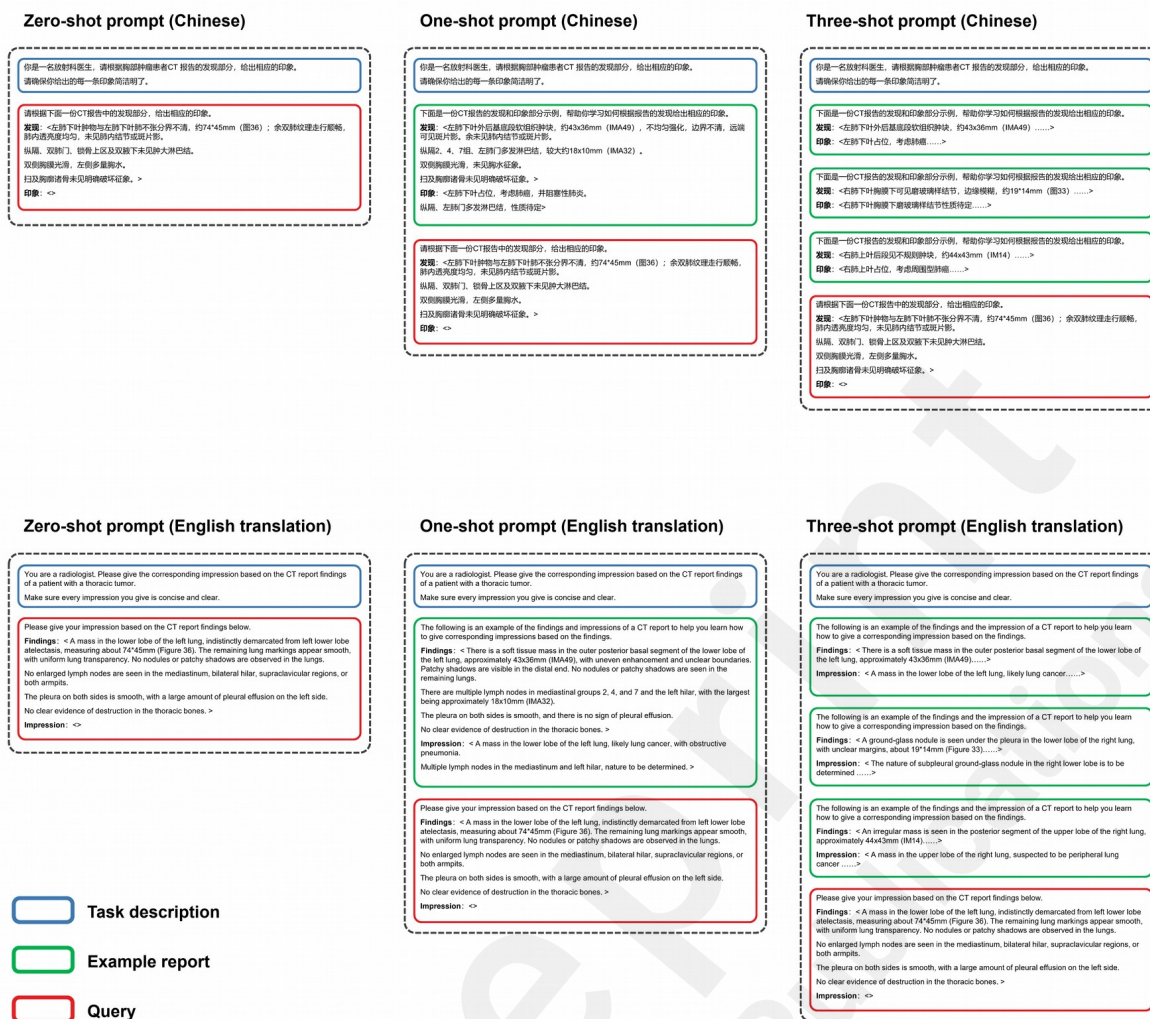


Figure 2. Zero-shot, one-shot, and three-shot prompt examples in Chinese and English. The prompts translated into English are provided for the reader's benefit, and these translations were not used in the study.

Automatic Quantitative Evaluation Metrics

In this study, we select 3 metrics widely used in text generation research to evaluate the generated impression against the reference impression. The values of these metrics range from 0 to 1, where a higher value indicates a better result. A brief introduction of the metrics is listed below:

BLEU. BiLingual Evaluation Understudy (BLEU) [41] score measures the number of position-independent matches of the n-grams of the candidate with the n-grams of the reference, focusing on the precision of the n-grams.

ROUGE-L. Recall-Oriented Understudy for Gisting Evaluation using Longest Common Subsequence (ROUGE-L) [42] measures the longest common subsequence (LCS) between the candidate and reference to calculate the LCS-based F-measure.

METEOR. Metric for Evaluation of Translation with Explicit Ordering (METEOR) [43] measures the harmonic mean of precision and recall calculated based on the mapping between unigrams with

the least number of crosses by exact, stemming, and synonym matching.

Human Semantic Evaluation Metrics

Although the automatic quantitative evaluation metrics above have shown some correlations with human judgments, they are not sufficient to evaluate the difference between the generated and reference impressions in semantics. Therefore, we define 5 human evaluation metrics, i.e., 1) Correctness, 2) Completeness, 3) Conciseness, 4) Verisimilitude, and 5) Replaceability, in this study to evaluate the semantics of the generated impressions. The definitions are listed below. We recruit three clinical experts (ZSY, LB, and LQ) to annotate the generated impression. We use a 5-point Likert scale for each evaluation metric. A higher value indicates a better result. Note that during human semantic evaluation, the clinical experts can only view the generated impressions, reference impressions, and the corresponding findings. For each evaluated impression, they were blinded to both the identity of the LLM that generated the impression and the type of prompt used (zero-shot, one-shot, or three-shot prompt).

Completeness. Completeness measures how completely the information in the generated impression covers the information in the reference impression. The five answer statements for the 5-point Likert scale of Completeness are 1 for "Very incomplete", 2 for "Relatively incomplete", 3 for "Neutral", 4 for "Relatively completeness", and 5 for "Very correct".

Correctness. Correctness measures how correct the information in the generated impression is compared to the information in the reference impression. The five answer statements for the 5-point Likert scale of Correctness are 1 for "Very incorrect", 2 for "Relatively incorrect", 3 for "Neutral", 4 for "Relatively correct", and 5 for "Very correct".

Conciseness. Conciseness measures how much redundant information is in the generated impression. The five answer statements for the 5-point Likert scale of Conciseness are 1 for "Very redundant", 2 for "Relatively redundant", 3 for "Neutral", 4 for "Relatively concise", and 5 for "Very concise".

Verisimilitude. Verisimilitude measures how similar the generated impression is to the reference impression in readability, grammar, and writing style. The five answer statements for the 5-point Likert scale of Verisimilitude are 1 for "Very fake", 2 for "Relatively fake", 3 for "Neutral", 4 for "Relatively verisimilar", and 5 for "Very verisimilar".

Replaceability. Replaceability measures whether the generated impression can replace the reference impression. The five answer statements for the 5-point Likert scale of Replaceability are 1 for "Very irreplaceable", 2 for "Relatively irreplaceable", 3 for "Neutral", 4 for "Relatively replaceable", and 5 for "Very replaceable".

Ethical Considerations

Ethical approval for this study was granted by the Ethics Committee of Peking University Cancer Hospital (2022KT128) before this study. Informed consent was waived due to the retrospective design of this study. Data were stored securely with access restricted to the research team and we removed all identifying information from the collected radiology reports before analysis. No personally identifiable information was included in the study or supplementary materials.

Statistical analysis

In this study, we used the Mann-Whitney U test to compare the human semantic evaluation results between zero-shot, one-shot, and three-shot prompt strategies. A $p < .05$ was considered statistically significant. The statistical analyses were conducted using the Scipy 1.7.3 Python package.

Results

Automatic Quantitative Evaluation Results

Tables 2, 3, and 4 show the BLEU, ROUGE-L, and METEOR values of the nine LLMs. We notice that ERNIE Bot obtains the overall best results for CT impression summarization, Tongyi Qianwen achieves the best performance for PET-CT impression summarization, and Claude shows the best performance for ultrasound impression summarization. Note that the best LLMs are all commercially available models. Moreover, all the best results are obtained based on the one-shot or few-shot prompts, indicating LLMs can learn from the example reports in the prompt to generate better impressions, but more is not necessarily better. Figure 3 illustrates the experimental results more intuitively.

Table 2. Automatic quantitative evaluation results of the generated CT impressions.

Prompt type	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE -L	METEO R
Zero-shot	Tongyi Qianwen	0.028	0.022	0.017	0.014	0.218	0.155
	ERNIE Bot	0.116	0.094	0.079	0.067	0.306	0.247
	ChatGPT	0.084	0.065	0.051	0.041	0.254	0.202
	Bard	0.085	0.067	0.055	0.044	0.275	0.214
	Claude	0.191	0.149	0.118	0.094	0.295	0.253
	Baichuan	0.061	0.047	0.038	0.031	0.233	0.172
	ChatGLM	0.026	0.020	0.016	0.012	0.203	0.155
	HuatuoGPT	0.113	0.084	0.066	0.053	0.259	0.230
	ChatGLM-Med	0.171	0.115	0.082	0.062	0.162	0.166
One-shot	Tongyi Qianwen	0.201	0.163	0.135	0.113	0.337	0.277
	ERNIE Bot	0.483	0.400	0.339	0.289	0.495	0.498
	ChatGPT	0.218	0.174	0.142	0.116	0.335	0.288
	Bard	0.293	0.235	0.195	0.162	0.397	0.352
	Claude	0.328	0.260	0.211	0.171	0.400	0.358
	Baichuan	0.118	0.089	0.070	0.055	0.300	0.268
	ChatGLM	0.365	0.280	0.219	0.171	0.334	0.331
	HuatuoGPT	0.191	0.149	0.122	0.101	0.317	0.293
	ChatGLM-Med	0.192	0.133	0.098	0.075	0.170	0.169
Three-shot	Tongyi Qianwen	0.253	0.207	0.172	0.145	0.366	0.317
	ERNIE Bot	0.440	0.367	0.311	0.264	0.483	0.467
	ChatGPT	0.300	0.244	0.203	0.170	0.386	0.354
	Bard	0.362	0.297	0.249	0.209	0.441	0.414

	Claude	0.365	0.296	0.246	0.203	0.445	0.403
	Baichuan	0.282	0.225	0.185	0.153	0.373	0.347
	ChatGLM	0.218	0.169	0.135	0.108	0.320	0.290
	HuatuoGPT	0.154	0.121	0.099	0.082	0.311	0.282
	ChatGLM-Med	0.159	0.102	0.068	0.049	0.154	0.152

Table 3. Automatic quantitative evaluation results of the generated PET-CT impressions.

Prompt type	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE -L	METEOR
Zero-shot	Tongyi Qianwen	0.452	0.347	0.274	0.221	0.323	0.390
	ERNIE Bot	0.341	0.267	0.218	0.181	0.311	0.337
	ChatGPT	0.220	0.166	0.130	0.105	0.250	0.256
	Bard	0.399	0.301	0.239	0.194	0.306	0.343
	Claude	0.415	0.320	0.249	0.197	0.323	0.378
	Baichuan	0.129	0.098	0.078	0.063	0.234	0.233
	ChatGLM	0.129	0.097	0.077	0.063	0.224	0.223
	HuatuoGPT	0.256	0.191	0.153	0.126	0.233	0.258
	ChatGLM-Med	0.098	0.072	0.057	0.047	0.139	0.227
One-shot	Tongyi Qianwen	0.469	0.365	0.293	0.239	0.348	0.438
	ERNIE Bot ^a	-	-	-	-	-	-
	ChatGPT	0.263	0.199	0.155	0.124	0.266	0.290
	Bard	0.348	0.268	0.217	0.179	0.299	0.333
	Claude	0.390	0.311	0.251	0.204	0.345	0.387
	Baichuan	0.091	0.070	0.056	0.047	0.225	0.218
	ChatGLM	0.233	0.175	0.137	0.111	0.246	0.264
	HuatuoGPT	0.240	0.179	0.142	0.115	0.231	0.257
	ChatGLM-Med ^a	-	-	-	-	-	-
Three-shot	Tongyi Qianwen	0.434	0.337	0.271	0.221	0.361	0.463
	ERNIE Bot ^a	-	-	-	-	-	-
	ChatGPT	0.363	0.277	0.218	0.175	0.291	0.334
	Bard	0.446	0.350	0.285	0.236	0.323	0.369
	Claude	0.415	0.342	0.284	0.239	0.379	0.427
	Baichuan ^a	-	-	-	-	-	-
	ChatGLM	0.223	0.169	0.135	0.111	0.224	0.243
	HuatuoGPT ^a	-	-	-	-	-	-
	ChatGLM-Med ^a	-	-	-	-	-	-

^aThe length of some prompts exceed the maximum input text length of the LLM.

Table 4. Automatic quantitative evaluation results of the generated Ultrasound impressions.

Prompt type	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE -L	METEO R
Zero-shot	Tongyi Qianwen	0.000	0.000	0.000	0.000	0.142	0.101
	ERNIE Bot	0.027	0.024	0.021	0.019	0.277	0.209
	ChatGPT	0.006	0.005	0.004	0.004	0.193	0.140
	Bard	0.016	0.014	0.012	0.011	0.249	0.184
	Claude	0.106	0.085	0.066	0.050	0.315	0.267
	Baichuan	0.008	0.007	0.006	0.005	0.237	0.172
	ChatGLM	0.003	0.002	0.002	0.002	0.182	0.132
	HuatuoGPT	0.010	0.008	0.007	0.006	0.182	0.156
	ChatGLM-Med	0.218	0.175	0.147	0.127	0.191	0.185
One-shot	Tongyi Qianwen	0.153	0.127	0.103	0.082	0.342	0.322
	ERNIE Bot	0.176	0.153	0.133	0.118	0.379	0.346
	ChatGPT	0.136	0.114	0.095	0.080	0.306	0.284
	Bard	0.130	0.107	0.090	0.076	0.325	0.289
	Claude	0.409	0.345	0.287	0.236	0.418	0.419
	Baichuan	0.053	0.045	0.039	0.035	0.284	0.233
	ChatGLM	0.092	0.074	0.060	0.050	0.267	0.260
	HuatuoGPT	0.016	0.013	0.011	0.009	0.197	0.174
	ChatGLM-Med	0.205	0.168	0.146	0.131	0.173	0.162
Three-shot	Tongyi Qianwen	0.180	0.157	0.137	0.120	0.441	0.404
	ERNIE Bot	0.213	0.195	0.179	0.166	0.498	0.454
	ChatGPT	0.246	0.224	0.204	0.186	0.490	0.459
	Bard	0.179	0.155	0.136	0.122	0.414	0.368
	Claude	0.420	0.384	0.349	0.318	0.561	0.520
	Baichuan	0.104	0.086	0.073	0.062	0.276	0.230
	ChatGLM	0.052	0.042	0.034	0.028	0.220	0.193
	HuatuoGPT	0.010	0.008	0.007	0.006	0.180	0.149
	ChatGLM-Med	0.225	0.186	0.162	0.145	0.187	0.188

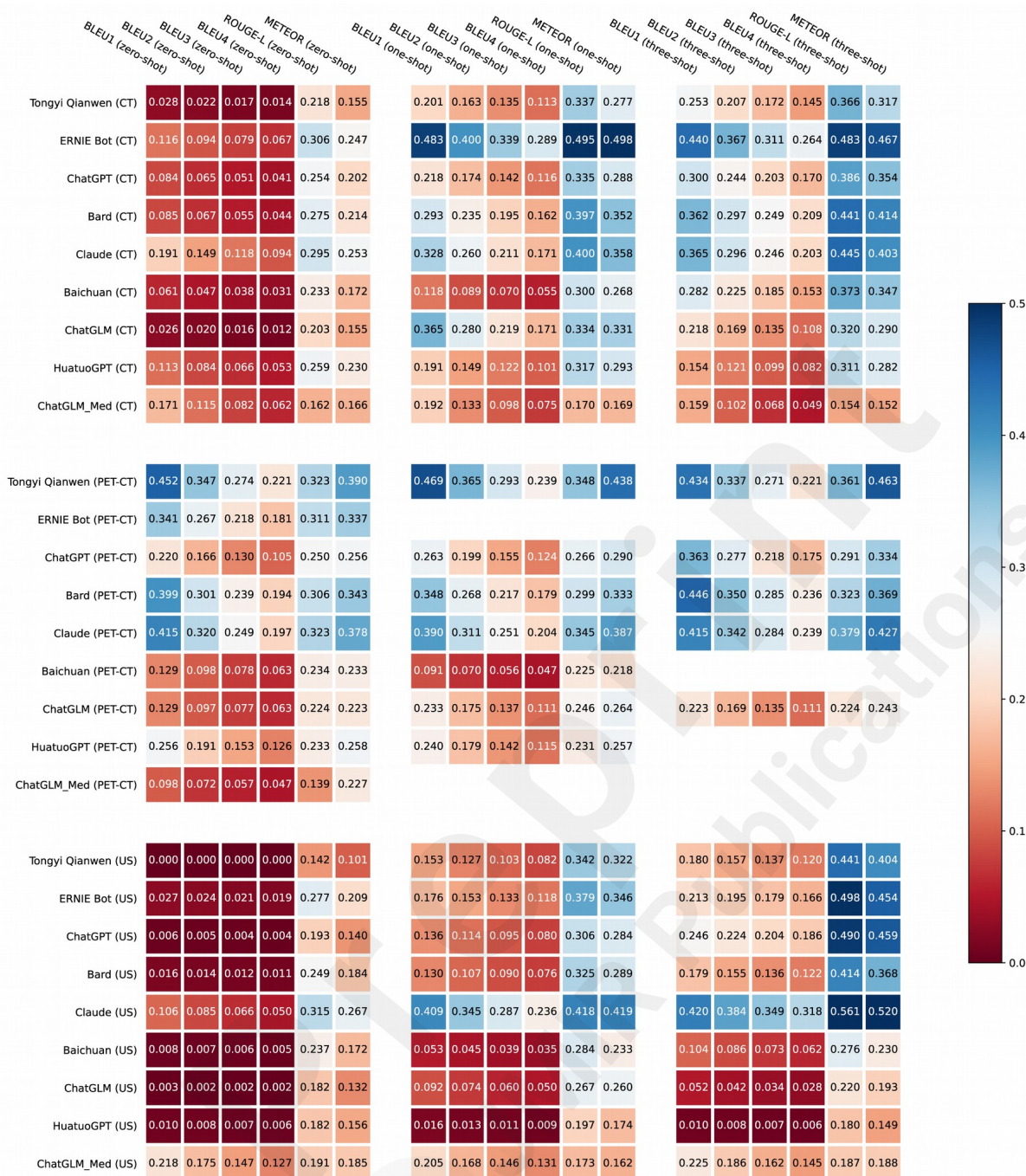


Figure 3. Heatmap visualization of automatic quantitative assessment for the generated CT, PET-CT, and US impressions.

Human Semantic Evaluation Results

Quality Evaluation

As the LLMs may produce undesired context, we first manually review the quality of the generated impressions. After review, we summarize five types of errors in generated impressions, i.e., refuse-to-answer, truncated-output, repeated-output, no-output, and English-output errors.

All five commercially available LLMs produce high-quality impressions with no truncated-output, repeated-output, or English-output errors. Only the Bard model refuses to provide answers for 4 PET-CT impression summarization prompts in one-shot and three-shot manners, respectively.

Different from the commercially available LLMs, the quality of generated impressions varies a lot among the four open-source LLMs. Figure 4 shows the errors of the open-source LLMs. The baichuan model achieves high-quality results, where only one output has the no-output error. The ChatGLM model also achieves good results when using zero-shot prompts, with only one truncated-output error. However, the ChatGLM model obtains many no-output errors when using few-shot prompts. Most of the no-output errors are due to the direct copy of the query section in the prompt but no generated impression. The two medical LLMs, HuatuoGPT and ChatGLM-Med, suffer serious errors in summarizing impressions. HuatuoGPT obtains truncated output and repeated-output errors in over 40% of PET-CT impression summarization tasks. Although the percentage of errors in the CT and ultrasound impression summarization decreases, 13.67% and 18.67% of summarized impressions still contain repeated-output errors, respectively. Note that HuatuoGPT is more prone to obtain repeated output errors when using few-shot prompts. The ChatGLM-Med obtains truncated output, repeated-output, and no-output errors in over 40% of generated PET-CT impressions. And 13.33% of generated CT impressions and 22.67% of generated ultrasound impressions have truncated output, repeated output, and no output errors.

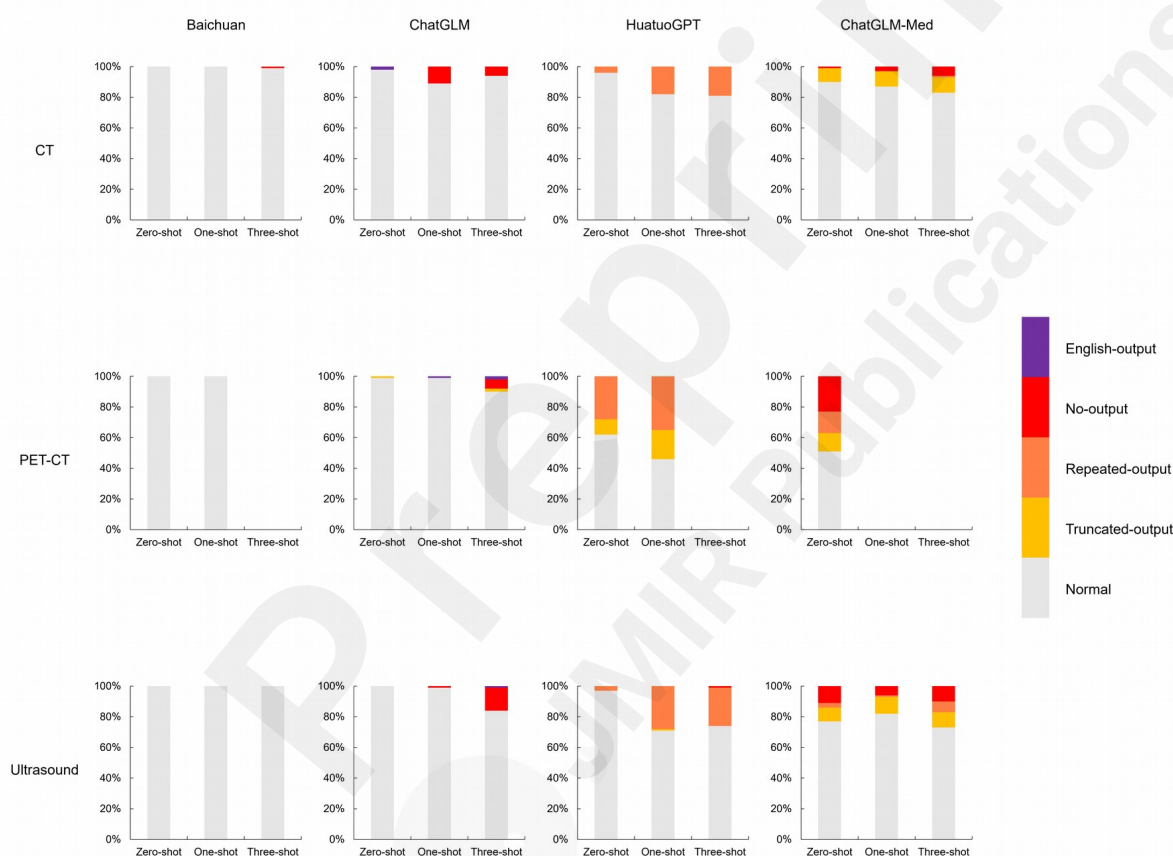


Figure 4. Quality assessment of the open-source LLMs for CT, PET-CT, and US impression generation.

Semantic Evaluation

Based on the automatic quantitative and manual quality evaluation, we note that the five commercially available LLMs achieve better impression summarization than the four open-source LLMs with higher BLEU, ROUGE-L, and METEOR values and better generation qualities. Therefore, we further evaluate the semantics of the generated impressions of the five commercially available LLMs. We define 5 human evaluation metrics: 1) Completeness, 2) Correctness, 3) Conciseness, 4) Verisimilitude, 5) Replaceability. The human evaluation results are shown in Tables 5, 6, and 7. Figure 5 illustrates the human evaluation results in a more intuitive way.

Table 5. Averaged human semantic evaluation results of the generated CT impressions.

Metric	Prompt type	Tongyi Qianwen	ERNIE Bot	ChatGP T	Bard	Claude
Completeness	Zero-shot	4.80	4.73	4.68	4.74	4.55
	One-shot	4.73	4.59	4.70	4.65	4.58
	Three-shot	4.77	4.70	4.70	4.59	4.58
Correctness	Zero-shot	4.25	4.07	3.82	3.84	3.84
	One-shot	4.26	4.24	4.13	4.05	3.96
	Three-shot	4.33	4.25	4.12	4.05	3.98
Conciseness	Zero-shot	1.41	2.46	2.22	2.16	2.49
	One-shot	2.96	4.49	3.02	3.72	3.20
	Three-shot	3.19	4.26	3.55	4.05	3.34
Verisimilitude	Zero-shot	2.45	2.87	2.47	2.50	2.86
	One-shot	3.39	4.11	3.21	3.46	3.40
	Three-shot	3.49	4.02	3.57	3.74	3.56
Replaceability	Zero-shot	2.47	2.69	2.41	2.46	2.89
	One-shot	3.13	3.54	2.94	2.99	3.35
	Three-shot	3.22	3.40	3.11	3.16	3.49

Table 6. Averaged human semantic evaluation results of the generated PET-CT impressions.

Metric	Prompt type	Tongyi Qianwen	ERNIE Bot	ChatGP T	Bard	Claude
Completeness	Zero-shot	3.87	3.94	3.98	3.74	4.19
	One-shot	3.53	-	3.92	3.84	4.30
	Three-shot	3.52	-	3.86	3.78	4.31
Correctness	Zero-shot	3.73	3.57	3.24	3.50	3.43
	One-shot	3.59	-	3.55	3.45	3.43
	Three-shot	3.55	-	3.54	3.47	3.45
Conciseness	Zero-shot	3.58	2.24	1.88	3.25	2.68
	One-shot	3.90	-	2.14	2.91	2.06
	Three-shot	4.13	-	2.84	3.14	2.10
Verisimilitude	Zero-shot	3.32	2.50	1.91	2.61	2.80
	One-shot	3.38	-	2.45	2.70	2.52
	Three-shot	3.39	-	2.89	2.86	2.44
Replaceability	Zero-shot	2.99	2.25	1.88	2.46	2.72
	One-shot	2.88	-	2.35	2.42	2.52
	Three-shot	2.77	-	2.66	2.53	2.49

Table 7. Averaged human semantic evaluation results of the generated Ultrasound impressions.

Metric	Prompt type	Tongyi Qianwen	ERNIE Bot	ChatGP T	Bard	Claude
Completeness	Zero-shot	4.49	4.44	4.33	4.28	4.36
	One-shot	4.44	4.34	4.14	4.12	4.20
	Three-shot	4.51	4.46	4.37	4.14	4.49
Correctness	Zero-shot	4.00	3.82	3.81	3.35	3.58

	One-shot	4.02	4.05	3.85	3.58	3.69
	Three-shot	4.04	4.16	4.08	3.87	4.00
Conciseness	Zero-shot	1.38	2.43	2.17	2.42	2.08
	One-shot	3.35	3.49	3.50	3.61	4.21
	Three-shot	3.65	3.76	3.86	3.70	3.76
Verisimilitude	Zero-shot	2.09	2.89	2.70	2.63	2.57
	One-shot	3.39	3.43	3.29	3.26	3.91
	Three-shot	3.61	3.74	3.71	3.47	4.01
Replaceability	Zero-shot	2.27	2.78	2.70	2.37	2.50
	One-shot	3.16	3.26	3.14	2.86	3.36
	Three-shot	3.42	3.61	3.52	3.13	3.71

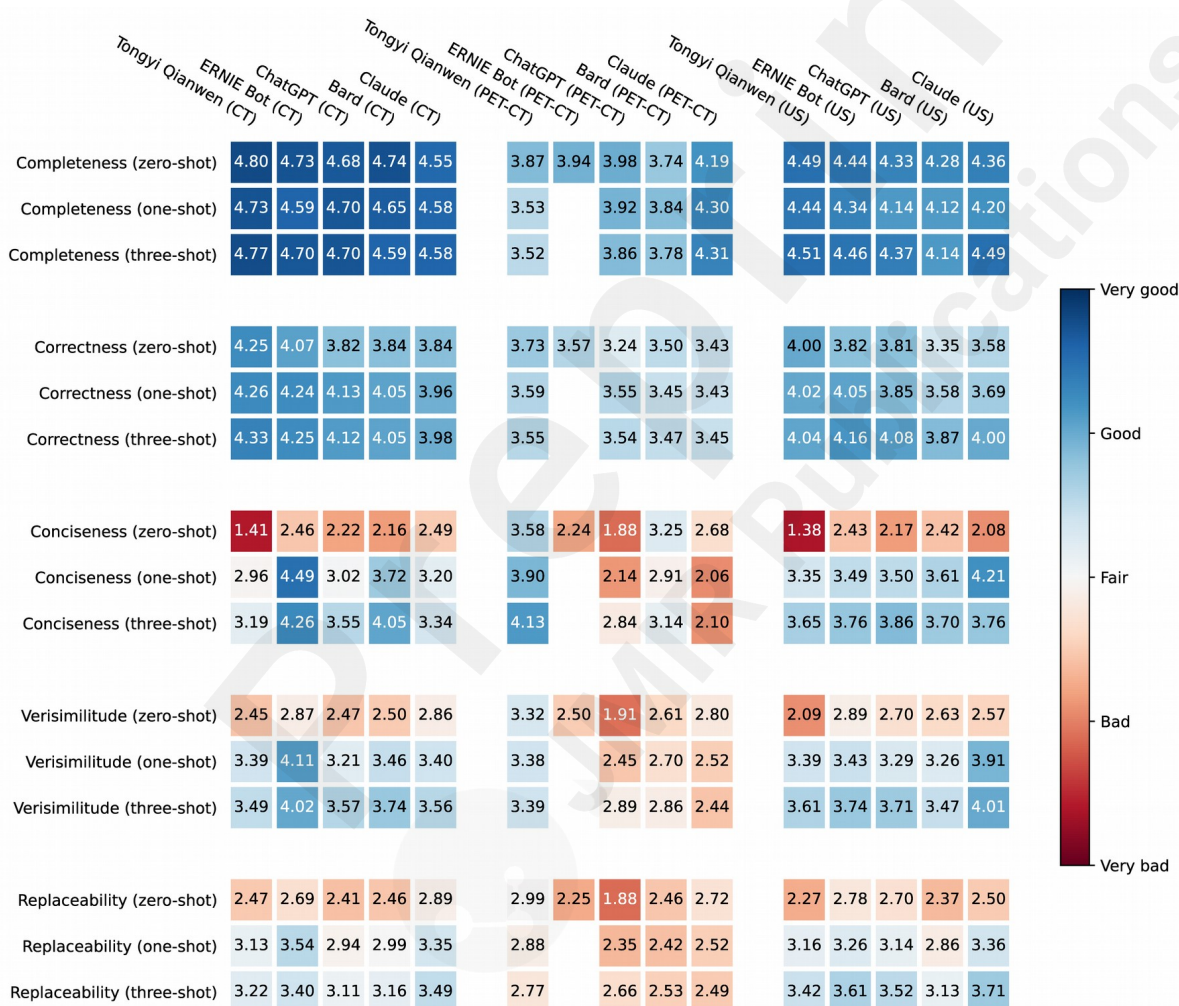


Figure 5. Heatmap visualization of Averaged human semantic assessment for the generated CT, PET-CT, and US impressions.

In the term of completeness, the generated CT and US impressions are generally better than the generated PET-CT impressions. Clinical experts rate all of the generated CT and US impressions by different LLMs as between "Relatively complete" and "Very complete" (the best scores are 4.80 for CT impressions and 4.51 for US impressions), while the PET-CT impressions generated by all but Claude are only close to "Relatively complete". By comparing different prompt types, we note that some impressions using few-shot prompts achieve lower completeness score than the impressions

using zero-shot prompts, but the decrease is limited.

In the term of correctness, the generated CT and US impressions also achieve good results (the best scores are 4.33 for CT impressions and 4.16 for US impressions), which are between "Relatively correct" and "Very correct". The generated PET-CT impressions obtain the 3.73 for the correctness, not reaching the "Relatively correct" level. We also note that, when using few-shot prompts, the generated CT and US impressions get higher correctness scores, but lower correctness scores for the generated PET-CT impressions compared with using zero-shot prompts.

In the term of conciseness, the generated CT, PET-CT, and US impressions obtain good results. Clinicians rate the generated CT, PET-CT, and US impressions as between "Relatively concise" and "Very concise" (the best scores are 4.49 for CT impressions, 4.13 for PET-CT impressions, and 4.21 for US impressions). When using few-shot prompts, the conciseness scores of the generated impressions achieve significant improvements compared with the generated impressions using zero-shot prompts.

In the term of verisimilitude, the generated CT and US impressions score more than 4 point (the best scores are 4.11 for CT impressions and 4.01 for US impressions), while the generated PET-CT impressions score between "Neutral" and "Relatively verisimilar" (the best score is 3.39 for PET-CT impressions). Note that using few-shot prompts can also improve the verisimilitude of the generated impressions significantly.

To comprehensively evaluate the semantics of the generated impressions, clinical experts rate the replaceability of these impressions. We find that the impressions generated by LLMs are not yet at the level that can replace manually written impressions. The generated CT and US impressions only achieved the replaceability scores of 3.54 and 3.71, which are between "Neutral" and "Relatively replaceable", while the generated PET-CT impressions have an even lower replaceability score of 2.99, which is only close to "Neutral".

When comparing the performances of different LLMs, we note that Tongyi Qianwen achieves the best results on the PET-CT impression generation task, with the best results in 4 of the 5 human evaluation metrics, i.e., correctness, conciseness, verisimilitude, and replaceability. ERNIE Bot outperforms the other LLMs on the CT impression generation task with the highest scores in conciseness, verisimilitude, and replaceability, and comparable scores in completeness and correctness. For US impression generation task, Claude achieves better results than other LLMs in conciseness, verisimilitude, and replaceability and comparable results in completeness and conciseness.

To analyze the evaluation variances between the clinical experts, we also list the evaluation results of each clinical expert in the Multimedia Appendix I Tables A1-A9, and Figure A1-A3. Based on the results, we note that there are differences in the scores of different clinical experts. Clinician I's scores are relatively low. He/She thinks that none of the three types of generated impressions can reach the level of replacing manually written impressions (the best score is 2.97 for US impressions). Clinician II's scores are in the middle. He/She thinks that the generated CT and US impressions are close to replacing manually written impressions (the best scores are 3.99 for CT impressions and 3.82 for US impressions), but the generated PET-CT impressions are just neutral in the replaceability (the best score is 3.02 for PET-CT impressions). Clinician III's scores are relatively high than the others. He/She thinks the generated PET-CT and CT impressions are close to replacing manually written impressions (the best scores are 4.08 for CT impressions and 3.98 for PET-CT impressions), and the generated US impressions can basically replace the manually written impressions (the best

score is 4.56 for US impressions).

Although the absolute values of the scores are different between clinical experts, the changing trends of impression scores under different prompt types are similar. Using few-shot prompts can improve most of the conciseness, verisimilitude, and replaceability scores significantly, but may lead to lower completeness and correctness scores. We also illustrate the significant test results in Multimedia Appendix I Figures A4-A12.

Discussion

Principal Results

In this study, we aim to explore the capability of the LLMs in summarizing radiology report impressions. Automatic quantitative and human semantic evaluations are conducted to measure the gap between the generated and reference impressions.

Commercially available LLMs vs open source LLMs

To have a comprehensive evaluation of the state-of-the-art LLMs, in this study, we select five commercially available LLMs, i.e., ChatGPT, Bard, ERNIE Bot, Tongyi Qianwen, and Claude and four open source LLMs, i.e., Baichuan, ChatGLM, HuatuoGPT, and ChatGLM-Med. According to the automatic quantitative evaluation, we can note that the commercially available LLMs outperform the open-source LLMs. Besides, the open-source LLMs exhibit more output errors in the generated impressions, such as the refuse-to-answer, truncated-output, repeated-output, no-output, and English-output errors. These errors are almost absent in the outputs of the commercially available LLMs. When using few-shot prompts, the commercially available LLMs can benefit more than the open-source LLMs, thus achieving higher improvements in the automatic quantitative evaluation metrics. The differences between the performance of commercially available and open-source LLMs may be due to the commercially available LLMs usually have more parameters, use more training data to train, employ more advanced closed-source algorithms to optimize, and are developed as web applications with better engineering implementations. The gap between the commercially available and open-source LLMs indicates that more computing resources and specialized engineering groups are critical for better LLMs, which has become the main obstacle for most research groups.

No best model for all impression summarization tasks

Based on the evaluation results, no single LLM can achieve the best results in all impression summarization tasks. Tongyi Qianwen, ERNIE Bot, and Claude achieve the best overall performance in the PET-CT, CT, and US impression summarization tasks, respectively. Although the experimental results indicate the evaluated LLMs are very competitive with each other and no one can outperform others in all impression summarization tasks significantly, we note that the LLMs optimized specifically for Chinese like Tongyi Qianwen and ERNIE Bot achieve better results in CT and PET-CT impression summarization than LLMs such as ChatGPT, Bard, and Claude. This finding suggests the necessity to build LLMs for specific languages, which can achieve better performance on language-specific tasks. Besides, we also observed that the performance of the LLMs varies across different report types, particularly in terms of BLEU score. In Figure 3, the BLEU scores for US impressions are significantly lower than those for CT and PET-CT impressions. The primary reason for this discrepancy is that the reference US impressions are much shorter than the reference CT and PET-CT impressions, resulting in fewer character matches between the generated and reference impressions. When using the zero-shot prompt, the generated US impressions are much longer than the reference impressions, leading to extremely low BLEU scores. Although the use of few-shot

prompts makes the generated US impressions more concise, the BLEU scores remain lower than those for CT and PET-CT due to the fewer matched characters.

Effect of the few-shot prompt

In this study, we also explore the effect of the few-shot prompt on impression summarization. Based on the experimental results, we note that the few-shot prompt can significantly improve the performance of LLMs on all automatic quantitative evaluation metrics and some human evaluation metrics, including conciseness, verisimilitude, and replaceability. For correctness and completeness, using few-shot prompts may lead to some performance degradation, but usually not significant. When further comparing the performance of LLMs using one-shot and three-shot prompts, we find that the more examples provided, the better the impressions generated are not achieved. For example, Tongyi Qianwen achieves the best BLEU values when using one-shot prompts and the best ROUGE-L and METEOR values when using three-shot prompts for PET-CT impression summarization. ERINE Bot outperforms the other LLMs in automatic metrics for CT impression summarization when using one-shot prompts but Claude achieves the best automatic metrics for US impression summarization when using three-shot prompts. Although there is an overall trend that using few-shot prompts will improve the performance of LLMs in generating impressions, it seems unclear how many examples a prompt should include to be most effective.

Clinical application

Note that to evaluate the semantics of the generated impressions, we first extract the impressions from the generated text manually and then conduct the human evaluation. Therefore, the current experimental results may be higher than those obtained by evaluating the original outputs. We list the automatic quantitative results in Appendix I Tables A10-A12 and Figure A13. We also show the difference in results between using the extracted impressions and original outputs in Figure A14. We can note that all results obtained by evaluating extracted impressions are higher than those on original outputs. However, among all LLMs, Tongyi Qianwen, ERNIE Bot, and ChatGPT show small differences between these results, indicating they can follow the instructions well to generate the text we desire. Although the Bard and Claude achieves comparable performance based on the extracted impressions, its original outputs contain much more impression-unrelated content, reducing its usability in summarizing impressions in real clinical practice.

According to the evaluation of clinical experts, the impressions generated by the LLMs can not directly substitute for those written by radiologists. However, using LLMs to summarize clinical text like radiology findings is still valuable. First, it can help clinicians improve the efficiency of writing clinical documents. In clinical practice, writing clinical documents like radiology reports, admission records, progress notes, and discharge summaries has become a heavy burden for radiologists and clinicians. To alleviate this problem, we can use the LLMs to summarize the related structured or unstructured electronic health records as a preliminary clinical note, and then the clinicians conduct the final review. Moreover, LLMs also have the potential to combine the abnormal findings from multiple types of reports to provide comprehensive evaluations like cancer staging, which can help clinicians assess the patient's status better. For cancer patients who usually undergo a long diagnosis and treatment process, we can also employ the LLMs to summarize the whole diagnosis and treatment timeline, which is very important and valuable for the development of the next treatment plan. Second, the impressions generated by LLMs demonstrate a high level of completeness and correctness, with evaluation scores ranging from 4.0 (good) to 5.0(very good)). As a result, clinicians may find it useful to reference the diagnoses or suspicions of abnormal findings suggested by LLMs

before making their own judgments. However, whether LLMs can really help clinicians with different experiences to make more accurate diagnoses remains unclear, requiring further investigation in future studies. Thirdly, we can use LLMs to facilitate the research. Based on the summarization ability, LLMs can effectively extract key information from clinical documents to identify eligible patients for specific studies.

Comparisons to Prior Work

In this study, we investigated the performance of widely used commercial and open-source LLMs in summarizing Chinese radiology report impressions for lung cancer. Based on our analysis, we found a gap between the impressions generated by LLMs and those written by radiologists. Previous research has also explored the application of LLMs in clinical text summarization. For instance, Liu et al. [22] assessed 29 different LLMs for generating impressions from radiological reports, utilizing open-source datasets such as the MIMIC-CXR and OpenI datasets. However, their study was limited to English X-ray reports and relied solely on automatic quantitative evaluation. Sun et al. [23] evaluated GPT-4 for radiological impression generation, introducing four human semantic metrics, i.e., coherence, comprehensiveness, factual consistency, and harmfulness, to assess the generated impressions. They found that radiologists outperformed GPT-4 in radiology report generation. Despite incorporating human semantic evaluation, their analysis was only confined to 50 English chest radiograph reports. Veen et al. [2] conducted a comprehensive evaluation of clinical text summarization, encompassing various radiology reports, progress notes, doctor-patient conversations, and patient questions. They also fine-tuned LLMs for better performance and they concluded that LLMs can outperform medical experts in multiple clinical text summarization tasks. However, their evaluation was restricted to zero-shot prompts and exclusively involved English clinical texts.

Compared with prior works, our study is the first to evaluate LLMs in summarizing three types of Chinese radiology report impressions, i.e., CT, PET-CT, and US, and to explore the impact of different prompting strategies, i.e., zero-shot, one-shot, and few-shot, on generation performance. Significantly, we conducted a large-scale human semantic evaluation of the generated impressions, providing more robust and convincing results. This study better demonstrates the capabilities of LLMs in non-English clinical text summarization, offering valuable insights into their applicability in diverse linguistic and clinical contexts.

Limitations

To comprehensively evaluate the LLMs' impression summarizing ability, we select three types of radiology reports, i.e., PET-CT, CT, and Ultrasound reports. We should note that all reports are from lung cancer patients treated in a single medical center, which indicates the patient population is homogeneous and the writing style of the reports is relatively uniform. So, the results in this study may differ from the average performance of LLMs in summarizing the impressions of reports from patients with different diseases or medical centers. In the future, we will try to collect more radiology reports from different patients and medical centers to evaluate the LLMs to obtain more robust results.

The most important contribution of this study is that we invite three clinical experts to manually evaluate the impressions generated by the LLMs from the point of view of semantics so that we can find out the gap between the reports generated by LLMs and those written by radiologists. However, manual evaluation is time-consuming and tedious, which is the biggest obstacle for large-scale

evaluation. Therefore, in this study, we only evaluated 100 generated impressions for each report type. Moreover, the number of clinicians involved in this study was also limited. As we know, the subjectivities of clinicians may vary based on their expertise and background. For example, an oncologist, a radiologist, and a pulmonologist may have different preferences for the radiology impressions of a patient with lung cancer. Furthermore, for complex diseases like lung cancer that typically require multidisciplinary treatment, surgical oncologists, radiation oncologists, and medical oncologists may prioritize different information. In this study, we only recruited three clinical experts, i.e., two thoracic surgeons and one radiologist, and just reported their individual evaluation results in the supplement but did not delve into the differences in their preferences and subjectivities. In the future, we will try to recruit more clinicians with different expertise from different departments and conduct subgroup analyses to explore their different views on the impressions generated by LLMs.

Currently, LLMs are updated very quickly. Since human evaluation is very time-consuming, we can not perform real-time human evaluation of the latest LLMs. In the future, we will try to evaluate the latest LLMs and compare them with their previous versions to find out the changes in the performance of impression summarization of radiology reports. Moreover, many LLMs have demonstrated competitive performance on various medical tasks, such as medical question answering, information extraction, and medical literature summarization. However, to the best of our knowledge, this study is the first to investigate the capability of LLMs in summarizing Chinese radiology impressions. Before this study, only a few studies explored this topic, and those were focused on English reports. Although the languages and types of reports are different, some of our findings align with previous studies, such as comparable ROUGE-L scores. But, there are also some differences, such as different expert opinions on the generated impressions. Due to the constraints of human evaluation, we can only choose a limited number of LLMs to evaluate in this study, excluding several notable LLMs like Med-PaLM and LLaVA-Med. In future work, we plan to expand our evaluation to include more models to draw a more comprehensive picture of LLMs for radiology impression summarization. Furthermore, fine-tuning LLMs with localized data or designing more specialized prompts, such as using the chain-of-thought strategy or requiring output in JSON format, may improve the quality of generated impressions and reduce irrelevant content. We will try to explore the effectiveness of these techniques in the future.

Conclusions

In this study, we explore the capability of LLMs in summarizing radiology report impressions using automatic quantitative and human evaluations. The experimental results indicate that there is a gap between the impressions generated by LLMs and written by radiologists. LLMs achieve great performance in completeness and correctness, but are not good in conciseness and verisimilitude. Although the few-shot prompt can improve the LLMs' performance in conciseness and verisimilitude, clinicians still believe that the LLMs can not replace the radiologist in summarizing impressions, especially for PET-CT reports with long findings.

Acknowledgements

This work was supported by the Beijing Natural Science Foundation (L222020), the National Natural Science Foundation of China (No.82402447), the National Key R&D Program of China (No.2022YFC2406804), the Capital's funds for health improvement and research (No.2024-1-1023), and the National Ten-thousand Talent Program.

Conflicts of Interest

None declared.

Abbreviations

LLMs: large language models

CT: computed tomography

PET: positron emission tomography

NLP: natural language processing

Multimedia Appendix 1

References

1. Golob JFJ, Como JJ, Claridge JA. The painful truth: The documentation burden of a trauma surgeon. *Journal of Trauma and Acute Care Surgery*. 2016;80(5):742-7. PMID: 01586154-201605000-00009. doi: 10.1097/ta.0000000000000986.
2. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*. 2024 2024/02/27. doi: 10.1038/s41591-024-02855-5.
3. Gesner E, Gazarian P, Dykes P. The Burden and Burnout in Documenting Patient Care: An Integrative Literature Review. *Stud Health Technol Inform*. 2019 Aug 21;264:1194-8. PMID: 31438114. doi: 10.3233/shti190415.
4. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*. 2016 2016/12/06;165(11):753-60. doi: 10.7326/M16-0961.
5. Zhang Y, Ding DY, Qian T, Manning CD, Langlotz CP, editors. *Learning to Summarize Radiology Findings*. 2018 October; Brussels, Belgium: Association for Computational Linguistics.
6. Jiang Z, Cai X, Yang L, Gao D, Zhao W, Han J, et al. Learning to Summarize Chinese Radiology Findings With a Pre-Trained Encoder. *IEEE Transactions on Biomedical Engineering*. 2023;70(12):3277-87. doi: 10.1109/TBME.2023.3280987.
7. Cai X, Liu S, Han J, Yang L, Liu Z, Liu T. ChestXRyBERT: A Pretrained Language Model for Chest Radiology Report Summarization. *IEEE Transactions on Multimedia*. 2023;25:845-55. doi: 10.1109/TMM.2021.3132724.
8. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*. 2024;25(1). doi: 10.1093/bib/bbad493.
9. Gershanik EF, Lacson R, Khorasani R. Critical finding capture in the impression section of radiology reports. *AMIA Annu Symp Proc*. 2011;2011:465-9. PMID: 22195100.
10. Bruls RJM, Kwee RM. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into Imaging*. 2020 2020/11/23;11(1):121. doi: 10.1186/s13244-020-00925-z.
11. Kasalak Ö, Alnahwi H, Toxopeus R, Pennings JP, Yakar D, Kwee TC. Work overload and diagnostic errors in radiology. *European Journal of Radiology*. 2023 2023/10/01/;167:111032. doi: <https://doi.org/10.1016/j.ejrad.2023.111032>.
12. OpenAI. Introducing ChatGPT. 2022 [cited 2024 March]; Available from: <https://openai.com/blog/chatgpt>.

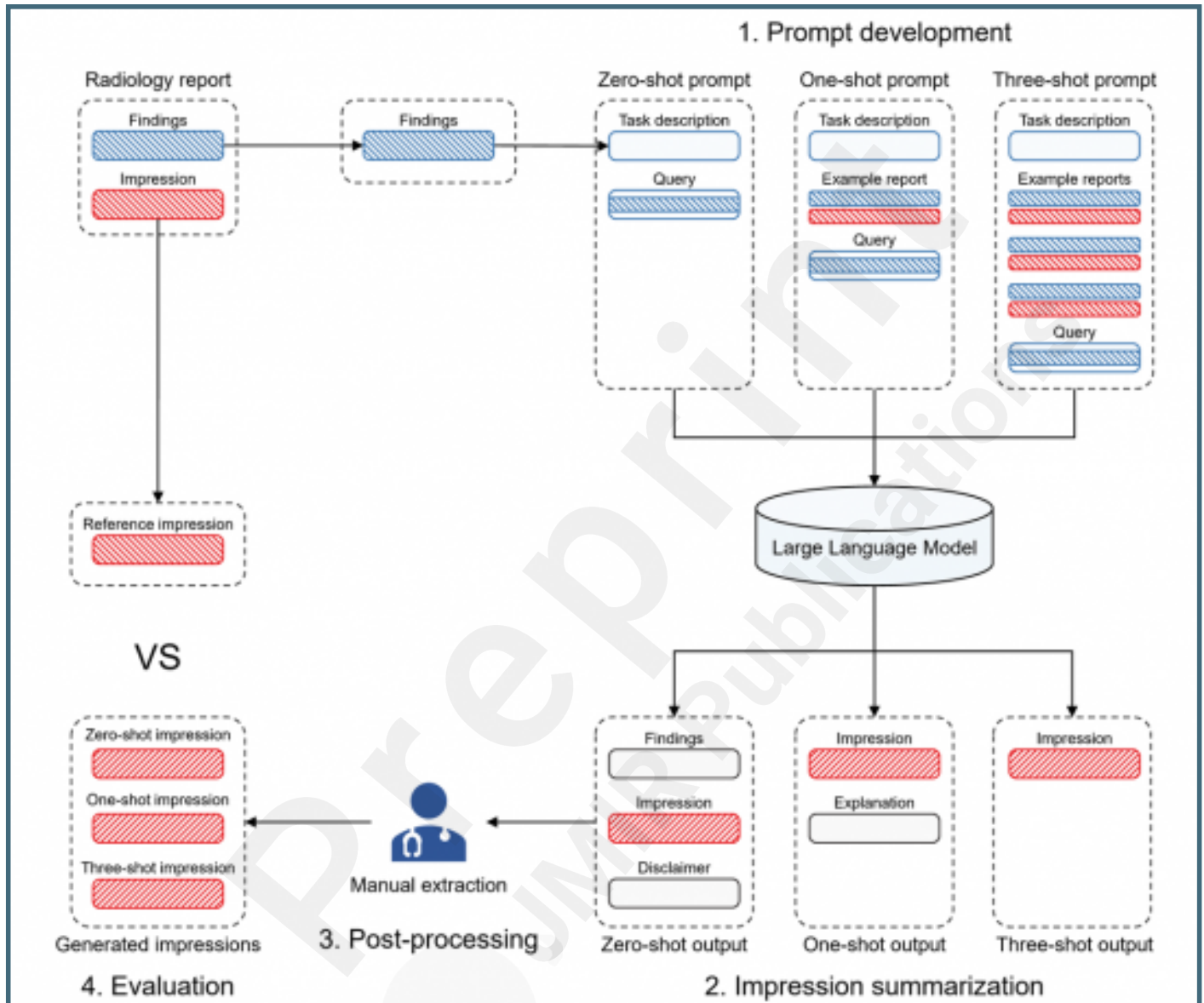
13. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.
14. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al., editors. Language models are few-shot learners. *Advances in neural information processing systems*; 2020; 2020.
15. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al., editors. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*; 2022; 2022.
16. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*. 2023 2023/08/24;6(1):158. doi: 10.1038/s41746-023-00896-7.
17. Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. *International Journal of Medical Informatics*. 2024;183:105321. doi: <https://doi.org/10.1016/j.ijmedinf.2023.105321>.
18. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*. 2024. doi: 10.1093/jamia/ocad259.
19. Chen Q, Du J, Hu Y, Keloth VK, Peng X, Raja K, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv preprint arXiv:230516326. 2023.
20. Doshi R, Amin KS, Khosla P, Bajaj S, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. *Radiology*. 2024;310(3):e231593. PMID: 38530171. doi: 10.1148/radiol.231593.
21. Keloth VK, Hu Y, Xie Q, Peng X, Wang Y, Zheng A, et al. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*. 2024. doi: 10.1093/bioinformatics/btae163.
22. Liu Z, Zhong T, Li Y, Zhang Y, Pan Y, Zhao Z, et al. Evaluating large language models for radiology natural language processing. arXiv preprint arXiv:230713693. 2023.
23. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, et al. Evaluating GPT-4 on Impressions Generation in Radiology Reports. *Radiology*. 2023;307(5):e231259. PMID: 37367439. doi: 10.1148/radiol.231259.
24. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen technical report. arXiv preprint arXiv:230916609. 2023.
25. Cloud A. Tongyi Qianwen. 2024 [cited 2024 January]; Available from: <https://tongyi.aliyun.com/qianwen/>.
26. Sun Y, Wang S, Li Y, Feng S, Chen X, Zhang H, et al. Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:190409223. 2019.
27. Bao S, He H, Wang F, Wu H, Wang H, editors. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. *The 58th Annual Meeting of the Association for Computational Linguistics*; 2020 July; Online: Association for Computational Linguistics; 2020.
28. Baidu. ERNIE Bot. 2024 [cited 2024 January]; Available from: <https://yiyao.baidu.com/>.
29. OpenAI. ChatGPT. 2024 [cited 2024 January]; Available from: <https://chatgpt.com/>.
30. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. Palm 2 technical report. arXiv preprint arXiv:230510403. 2023.
31. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*. 2023;24(240):1-113.
32. AI G. Bard. 2024 [cited 2024 January]; Available from: <https://gemini.google.com/app>.
33. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:221208073. 2022.

34. Anthropic. Claude. 2024 [cited 2024 October]; Available from: <https://claude.ai/new>.
35. Technology BI. Baichuan-13B: A 13B large language model developed by Baichuan Intelligent Technology. GitHub2023 [cited 2024 March]; Available from: <https://github.com/baichuan-inc/Baichuan-13B>.
36. Du Z, Qian Y, Liu X, Ding M, Qiu J, Yang Z, et al., editors. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. 2022 May; Dublin, Ireland: Association for Computational Linguistics.
37. GLM T. ChatGLM3-6B. 2024 [cited 2024 March]; Available from: <https://huggingface.co/THUDM/chatglm3-6b>.
38. Zhang H, Chen J, Jiang F, Yu F, Chen Z, Chen G, et al., editors. HuatuoGPT, Towards Taming Language Model to Be a Doctor. 2023 December; Singapore: Association for Computational Linguistics.
39. FreedomIntelligence. HuatuoGPT. 2023 [cited 2024 March]; Available from: <https://github.com/FreedomIntelligence/HuatuoGPT>.
40. Wang H, Liu C, Zhao S, Qin B, Liu T. Med-ChatGLM. GitHub; 2023 [cited 2024 March]; Available from: <https://github.com/SCIR-HI/Med-ChatGLM>.
41. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. {USA}: Association for Computational Linguistics; 2002. p. 311–8 , numpages = 8.
42. Lin C-Y, editor. ROUGE: A Package for Automatic Evaluation of Summaries. 2004 July; Barcelona, Spain: Association for Computational Linguistics.
43. Banerjee S, Lavie A, editors. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. 2005 June; Ann Arbor, Michigan: Association for Computational Linguistics.

Supplementary Files

Figures

Study overview: 1. Prompt development: we first developed the zero-shot, one-shot, and three-shot prompts for the collected radiology reports using the prompt templates. 2. Impression summarization: we input the prompts into LLMs and collected the outputs with summarized impressions. 3. Post-processing: we manually extracted the impression-related contents from the LLMs' outputs. 4. Evaluation: we conducted automatic and human evaluations of the generated impressions.



Zero-shot, one-shot, and three-shot prompt examples in Chinese and English. The prompts translated into English are provided for the reader’s benefit, and these translations were not used in the study.

Zero-shot prompt (Chinese)

Task description

You are a radiologist. Please give the corresponding impression based on the CT report findings of a patient with a thoracic tumor.

Example report

Findings: The boundary between the mass and adjacent structures is clear. The mass is located in the lower lobe of the left lung, approximately 45x35mm (SMA), with uneven enhancement. The bronchovascular bundle is enlarged. No enlarged lymph nodes were found in the mediastinum, bilateral hilar, and subcarinal areas.

Impression: A mass in the lower lobe of the left lung, considered to be lung cancer.

Query

Based on the CT report findings, please give the corresponding impression.

One-shot prompt (Chinese)

Task description

You are a radiologist. Please give the corresponding impression based on the CT report findings of a patient with a thoracic tumor.

Example report

Findings: The boundary between the mass and adjacent structures is clear. The mass is located in the lower lobe of the left lung, approximately 45x35mm (SMA), with uneven enhancement. The bronchovascular bundle is enlarged. No enlarged lymph nodes were found in the mediastinum, bilateral hilar, and subcarinal areas.

Impression: A mass in the lower lobe of the left lung, considered to be lung cancer.

Query

Based on the CT report findings, please give the corresponding impression.

Three-shot prompt (Chinese)

Task description

You are a radiologist. Please give the corresponding impression based on the CT report findings of a patient with a thoracic tumor.

Example report

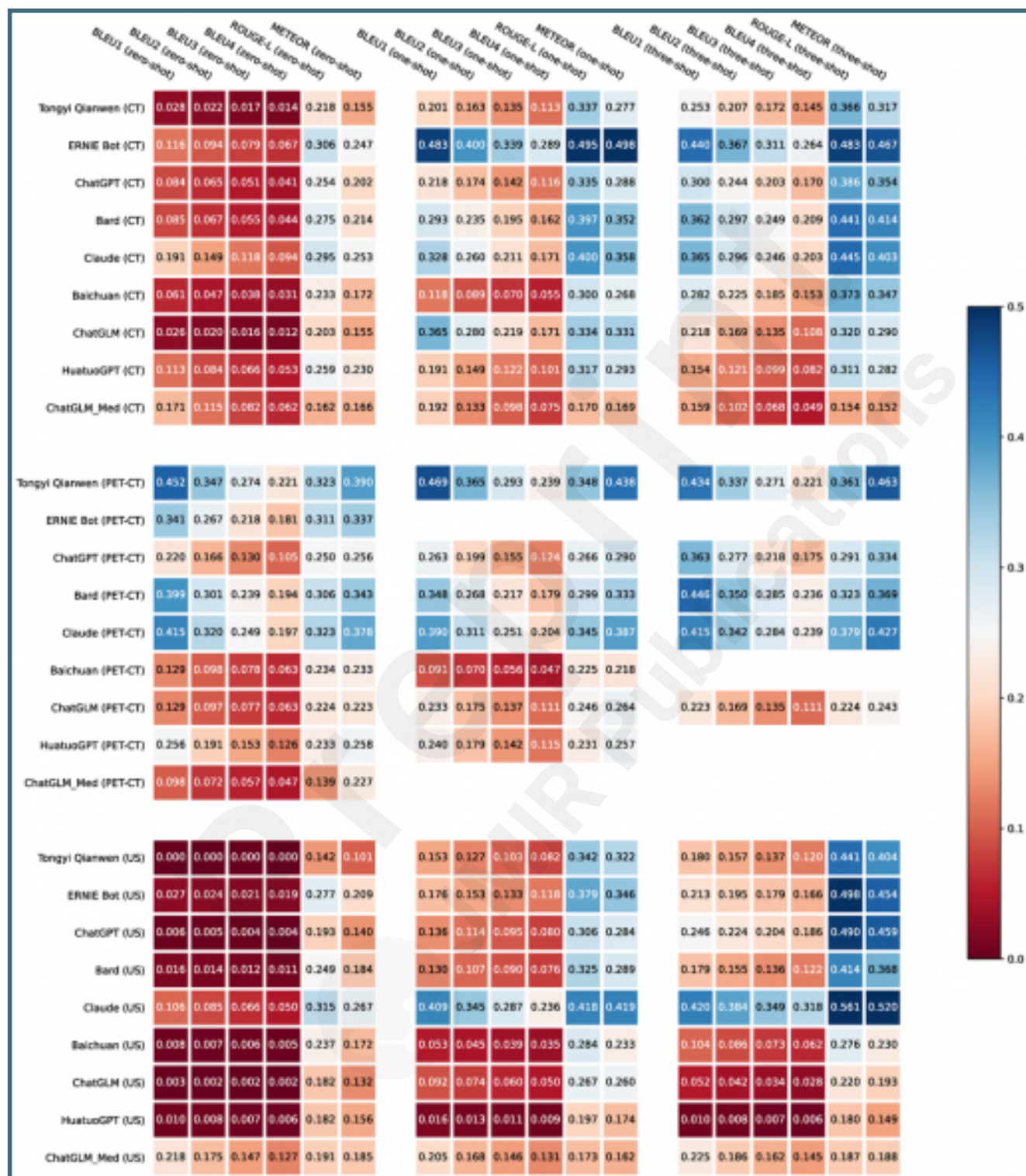
Findings: The boundary between the mass and adjacent structures is clear. The mass is located in the lower lobe of the left lung, approximately 45x35mm (SMA), with uneven enhancement. The bronchovascular bundle is enlarged. No enlarged lymph nodes were found in the mediastinum, bilateral hilar, and subcarinal areas.

Impression: A mass in the lower lobe of the left lung, considered to be lung cancer.

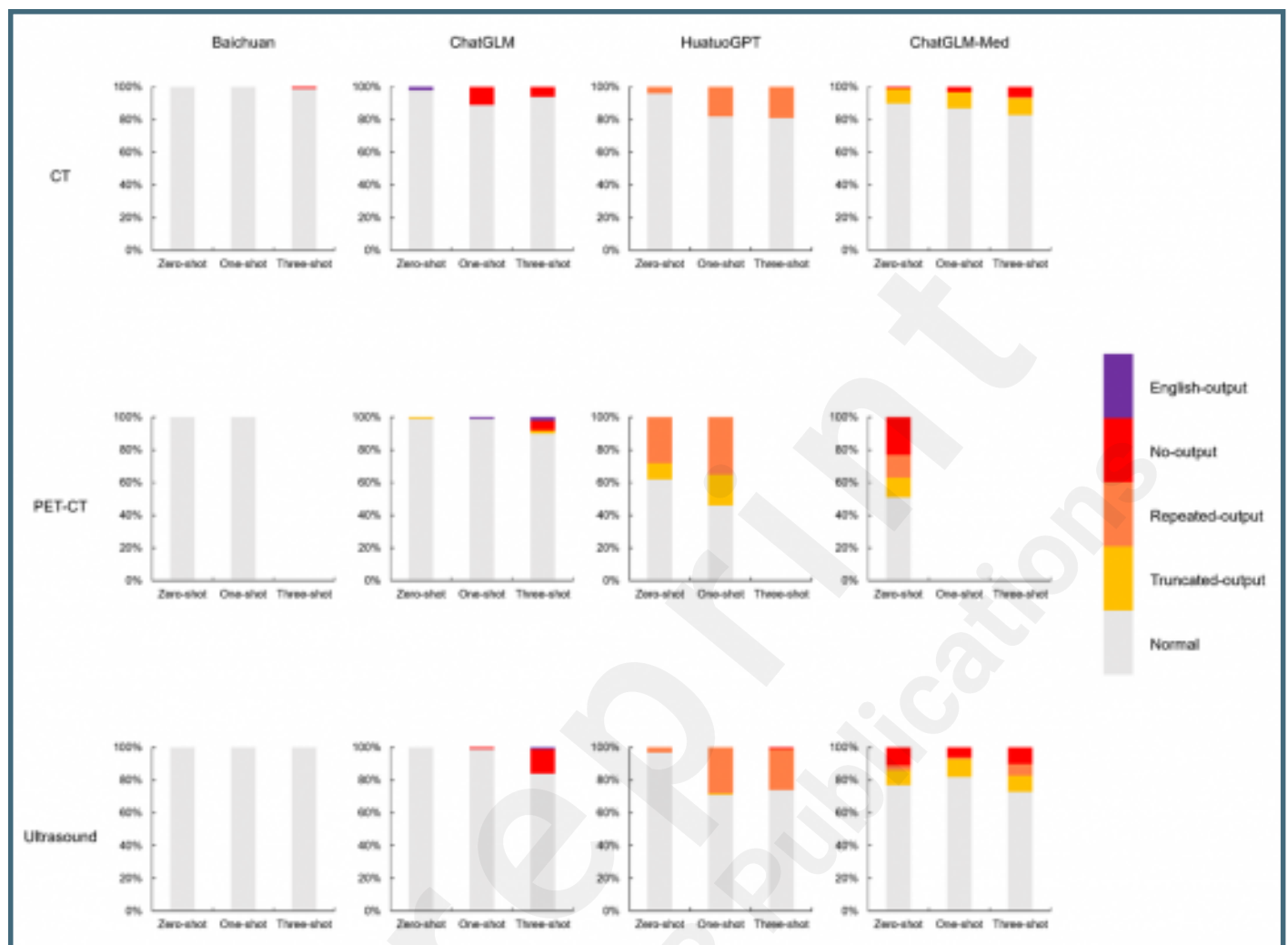
Query

Based on the CT report findings, please give the corresponding impression.

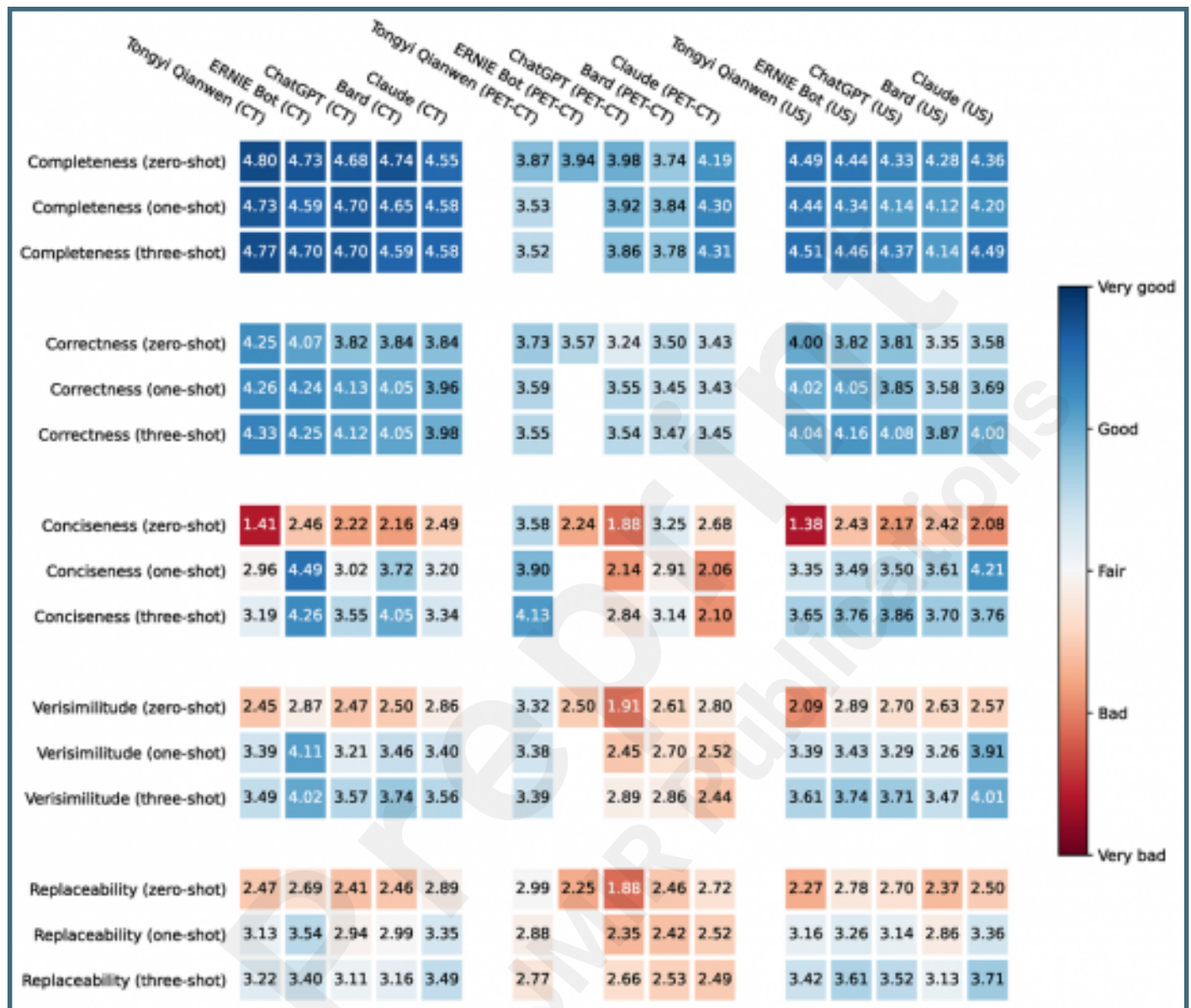
Heatmap visualization of automatic quantitative assessment for the generated CT, PET-CT, and US impressions.



Quality assessment of the open-source LLMs for CT, PET-CT, and US impression generation.



Heatmap visualization of Averaged human semantic assessment for the generated CT, PET-CT, and US impressions.



Multimedia Appendixes

Supplement.

URL: <http://asset.jmir.pub/assets/4da65c1962027548e442a7b518eb3a1b.docx>

