BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Enhancing doctor-patient communication using large language models for pathology report interpretation

Xiongwen Yang[1,2], Yi Xiao[3], Di Liu[1,2], Yun Zhang[4], Huiyin Deng[5], Jian Huang[6], Huiyou Shi[7], Dan Liu[8], Maoli Liang[2,9], Xing Jin[1,2], Yongpan Sun[1,2], Jing Yao[1,2], XiaoJiang Zhou[1,2], Wankai Guo[1,2], Yang He[1,2], WeiJuan Tang[1,2] and Chuan Xu[1,2*]

Abstract

Background Large language models (LLMs) are increasingly utilized in healthcare settings. Postoperative pathology reports, which are essential for diagnosing and determining treatment strategies for surgical patients, frequently include complex data that can be challenging for patients to comprehend. This complexity can adversely affect the quality of communication between doctors and patients about their diagnosis and treatment options, potentially impacting patient outcomes such as understanding of their condition, treatment adherence, and overall satisfaction.

Materials and methods This study analyzed text pathology reports from four hospitals between October and December 2023, focusing on malignant tumors. Using GPT-4, we developed templates for interpretive pathology reports (IPRs) to simplify medical terminology for non-professionals. We randomly selected 70 reports to generate these templates and evaluated the remaining 628 reports for consistency and readability. Patient understanding was measured using a custom-designed pathology report understanding level assessment scale, scored by volunteers with no medical background. The study also recorded doctor-patient communication time and patient comprehension levels before and after using IPRs.

Results Among 698 pathology reports analyzed, the interpretation through LLMs significantly improved readability and patient understanding. The average communication time between doctors and patients decreased by over 70%, from 35 to 10 min ($P < 0.001$), with the use of IPRs. The study also found that patients scored higher on understanding levels when provided with AI-generated reports, from 5.23 points to 7.98 points ($P < 0.001$), with the use of IPRs. indicating an effective translation of complex medical information. Consistency between original pathology reports (OPRs) and IPRs was also evaluated, with results showing high levels of consistency across all assessed dimensions, achieving an average score of 4.95 out of 5.

Conclusion This research demonstrates the efficacy of LLMs like GPT-4 in enhancing doctor-patient communication by translating pathology reports into more accessible language. While this study did not directly measure patient outcomes or satisfaction, it provides evidence that improved understanding and reduced communication time may positively influence patient engagement. These findings highlight the potential of AI to bridge gaps between medical professionals and the public in healthcare environments.

*Correspondence:
Chuan Xu
xuchuan89757@163.com
Full list of author information is available at the end of the article

Keywords Large language models, Doctor-patient communication, Surgical oncology scene, Postoperative pathology reports

Introduction

As medical information technology rapidly advances, the application of artificial intelligence (AI) in healthcare is becoming increasingly widespread [1–3]. Notably, Large Language Models (LLMs) have shown potential in the analysis and processing of medical texts [2]. Pathology reports, being critical for diagnosis and treatment decisions, directly impact the quality and efficiency of doctor-patient communication [4]. However, these reports often contain a large amount of professional terminology and complex data, making them difficult for patients to understand. Doctors also face time pressure when explaining these reports. Therefore, enhancing the readability of pathology reports and improving effective communication between doctors and patients has become crucial for improving the quality of medical services. Additionally, insufficient communication between doctors and patients has been identified as a significant factor affecting patient satisfaction and treatment compliance [5]. Studies have shown that good doctor-patient communication can significantly improve patients' understanding and acceptance of treatment plans, thereby affecting treatment outcomes [5, 6].

In recent years, LLMs have made significant progress in understanding and generating natural language, demonstrating their ability to analyze and rewrite medical texts in a manner more understandable to non-professionals [7, 8]. For instance, Steimetz et al. (2024) demonstrated that LLM chatbots can significantly improve the readability of pathology reports while also highlighting some of the limitations such as inaccuracies and hallucinations in the generated reports [9]. This study aims to explore the possibility of using LLMs to enhance the efficiency of doctor-patient communication, particularly by automating the translation of pathology report content into patient-friendly language. This approach aims to reduce cognitive barriers to medical information and promote better patient understanding of their health conditions.

Using routine post-operative pathology reports in oncology, this study designed a universal pathology report interpretation framework through LLMs and developed a corresponding pathology report understanding level assessment scale. This was done to explore the potential and actual effects of LLMs in enhancing doctor-patient communication efficiency.

Therefore, in response to these challenges, this study aims to explore the potential of using LLMs to enhance doctor-patient communication, particularly by simplifying pathology report content into patient-friendly language, and to provide insights on how LLMs can be integrated into clinical practice to improve communication efficiency [10, 11].

By improving the readability of pathology reports, we hope to promote better patient understanding of their health conditions, strengthen trust and communication between doctors and patients, and ultimately enhance the overall quality of medical services and patient satisfaction. Trust in physicians, fostered by effective communication, plays a pivotal role in treatment adherence. Research indicates that patients who trust their healthcare providers are more likely to follow prescribed treatments, which is essential for better health outcomes [12, 13].

Materials and methods

The work has been reported in line with the Standards for Quality Improvement Reporting Excellence (SQUIRE) criteria [14].

Study design

From October to December 2023, text pathology reports of malignant tumors were retrieved from the database of four hospitals. Pathology reports included information on cytology, tissue biopsy examination, and resections. Additionally, all common tumor types were included, except for rare malignant tumors, which were excluded due to limited sample sizes and follow-up data (Fig. 1).

Among the 698 eligible text pathology reports on malignant tumors, 70 reports (5 reports per organ for 14 organs) were randomly selected to develop templates for interpretive reports and corresponding scoring scales. These were used to enable LLMs to reliably generate similar interpretive reports, as well as to produce identical outputs from the remaining 628 reports. Doctors evaluated each report for consistency by comparing the original pathology report (OPR) with the AI-generated simplified report (Interpretive pathology report, IPR). The evaluation focused on whether key diagnostic information, such as tumor type (e.g., carcinoma, lymphoma), tumor stage (e.g., TNM classification), histological features (e.g., cell differentiation), presence of metastasis, and other clinically significant findings (e.g., molecular markers, margins, and lymph node involvement), were accurately represented in the simplified version. Doctors from multiple specialisms, including pathology, oncology, and surgery, participated in this evaluation process. Each

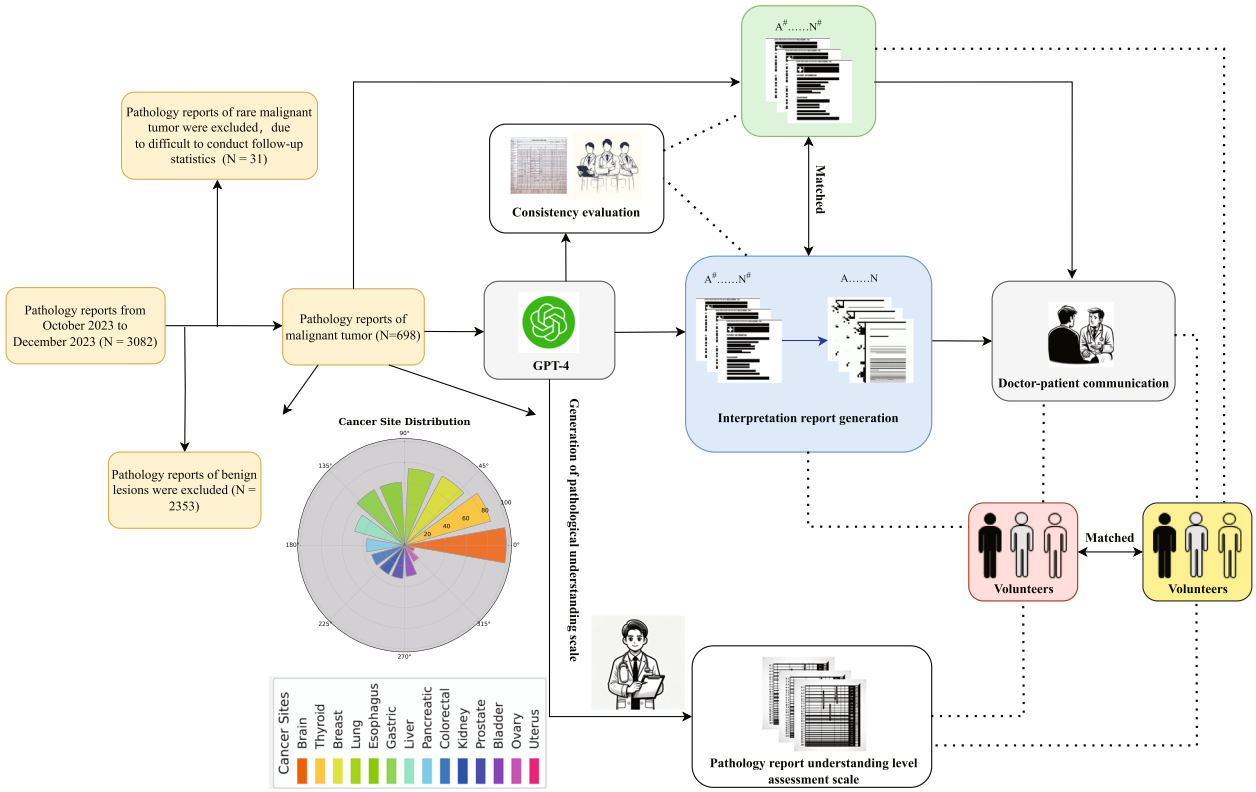

Fig. 1 Study design flow chart. The pathology reports from pathologists (Label ($A^{\#}$……$N^{\#}$)) were fed into the natural language processing (NLP) pipeline to generate new pathology interpretation reports (Label (A……N)). Label ($A^{\#}$……$N^{\#}$) and Label (A……N) were both read and scored by three volunteers, and the results were statistically compared with each other. In addition, the understanding of Label ($A^{\#}$……$N^{\#}$) and Label (A……N) were scored by the volunteers through the pathological score scale. Meanwhile, the doctor-patient communication time after the volunteers read Label ($A^{\#}$……$N^{\#}$) and Label (A……N) was also recorded and statistically analyzed. The pathological score scale was generated by the large language model (LLM), which was modified and organized by pathologist. The dotted lines indicate that both pathologists and/or volunteers participated in the corresponding task of the study and interacted with each other during the process

specialist ensured that the critical diagnostic elements within their field were accurately reflected. No significant differences were observed between specialisms in the consistency of the outcomes, as all specialists prioritized accuracy and clarity in their respective domains. If discrepancies were found, the reports were reviewed and corrected to ensure alignment between the two versions. This process is further illustrated in Fig. 2C.

The baseline health literacy levels of the volunteers were assessed using the Health Literacy Questionnaire (HLQ), ensuring that their understanding of medical terminology was evaluated prior to the study [15]. This assessment helped us control for variations in health literacy among the volunteers. The results of the HLQ assessments are summarized in Table 1. In the study, three volunteers (V_A, V_B, and V_C) with only a high school education and no medical background scored the 698 OPRs using the scoring scales (Fig. 2) and recorded reading time. Then, three other volunteers (V_D, V_E, and V_F) with similar backgrounds scored the IPRs using the scoring scales (Fig. 2) and recorded reading time. Lastly, doctors (with 10–15 years of experience) communicated with volunteers (V_A, V_B, and V_C) based on the OPRs and recorded doctor-patient communication time, and then communicated with volunteers (V_D, V_E, and V_F) based on the IPRs and recorded the time. Figure 1 summarizes the study design.

Scale and template generation

Seventy pathology reports were assigned to an author (X.W.Y) to construct scales and templates (Fig. 2), aimed at evaluating the accuracy and repeatability of IPRs generated by GPT-4 through quantitative metrics.

A pathology report understanding level assessment scale is presented in Fig. 2A. This scale aims to comprehensively assess the understanding level of non-medical background individuals regarding pathology reports. Patient understanding was measured using a custom-designed pathology report understanding level assessment scale, developed based on established health

A

Pathology Report Understanding Level Assessment Scale

Scoring Criteria (Ten-point scale)*:
1. Understanding of Report Structure. 0) Unable to identify the basic structure and various parts of the report (**0 points**). 1) Can identify some parts of the structure (e.g., diagnosis, patient information) but does not fully understand them (**1 point**). 2) Fully understands the report's structure and the content and function of each major section (**2 points**).
2. Terminology Recognition and Understanding. 0) Cannot recognize professional terminology or completely misunderstands the terms (**0 points**). 1) Can recognize some basic medical terms but has limited understanding (**1 point**). 2) Accurately recognizes and fundamentally understands most terms (**2 points**).
3. Interpretation of Results. 1) Unable to interpret the report's results (**0 points**). 2) Can partially interpret results, but misunderstandings exist (**1 point**). 3) Correctly interprets the basic information of the report's results (**2 points**).
4. Extraction of Key Information. 0) Unable to extract key information from the report (**0 points**). 1) Can extract some key information but misses important details (**1 point**). 2) Accurately extracts and understands all key information from the report (**2 points**).
5. Comprehensive Understanding and Application. 0) Unable to comprehensively understand the report content or relate it to health conditions (**0 points**). 1) Has a basic comprehensive understanding but limited ability to relate the report content to health conditions (**1 point**). 2) Not only fully understands the report content but also can effectively relate it to personal or others' health conditions (**2 points**).
Scoring Guide: **Level C (0-4 points)**: Low level of understanding, it is recommended to undertake basic medical knowledge learning to improve understanding of pathology reports. **Level B (5-7 points)**: Basic level of understanding, capable of grasping some key points of the report but still needs to enhance understanding of professional terminology and report structure. **Level A (8-10 points)**: High level of understanding, able to accurately interpret and apply information from the pathology reports.

***** *This scale aims to comprehensively assess the understanding level of non-medical background individuals regarding pathology reports.*

B

Pathology Report Interpretation Template*

1. Report Overview
Report Type: Explain what type of pathology report this is, e.g., a tissue biopsy examination, cytology study, etc. **Case Information:** Briefly summarize the patient's basic information, such as age and gender.
2. Sample Information
Sample Source: Describe how and from where the sample was obtained. **Sample Type:** Describe whether the sample is tissue, cells, fluid, etc.
3. Gross and Microscopic Findings
Findings Description: Use simple language to describe what the pathologist sees under the microscope, such as changes in cells, the state of tissues, etc.
4. Diagnosis Results
Results Explanation: Translate medical terms into easy-to-understand language, explaining the significance of the diagnosis results. If possible, provide comparisons with common diseases or conditions.
5. Recommendations and Explanations
Next Steps: Suggest follow-up medical steps or treatment options based on the diagnosis results. **Health Guidance:** Offer related lifestyle or dietary advice to help understand how to manage or improve the condition.
6. Frequently Asked Questions
Q&A: List some common questions and their answers about the report to help patients and their families better understand the report's content and significance.
Notes: 1)Each section should be adjusted based on the specific contents of the pathology report. 2)Use simple and direct language, avoiding too many medical jargon terms. 3)Where possible, use metaphors or analogies to explain complex medical concepts, making them easier to understand.

***This template is intended as a general framework; specific content needs to be filled in and adjusted according to the actual details of each pathology report. This aims to assist individuals without a medical background in understanding the content and importance of pathology reports.*

C

Pathology Artificial Intelligence Quality Index

1. Accuracy (Whether the information in the GPT-4 report is accurate and consistent with current medical knowledge and the actual content of the pathology report.**)**
Scoring Criteria: 1 point: The report is full of errors and significantly deviates from the actual pathology report content. **2 points:** The report contains multiple errors or misunderstandings. **3 points:** The report is basically accurate, with minor errors. **4 points:** The report is largely accurate, with very few minor errors. **5 points:** The report is completely accurate and entirely consistent with the pathology report content.
2. Interpretation Depth (How GPT-4 interprets the details of the pathology report and whether it can provide in-depth explanations of the pathology results.**)**
Scoring Criteria: 1 point: Almost no interpretation, merely repeats report content. **2 points:** Superficial explanations lacking depth. **3 points:** Provides a certain depth of explanation, but there is room for improvement. **4 points:** Deep and detailed explanations. **5 points:** Provides very insightful and in-depth explanations.
3. Readability (The readability of the report, including the fluency and comprehensibility of the language.**)**
Scoring Criteria: 1 point: The report is difficult to understand, with disorganized language. **2 points:** The report has readability issues, with some paragraphs difficult to understand. **3 points:** The report is generally readable, but there is room for improvement. **4 points:** The report is fluent and easy to understand, with only a few difficult parts. **5 points:** The report is very fluent, with clear and easy-to-understand language.
4. Clinical Relevance (The relevance and usefulness of the report's information to clinical practice.**)**
Scoring Criteria: 1 point: The report information is irrelevant to clinical practice. **2 points:** A few pieces of information are clinically relevant. **3 points:** Part of the report content is helpful to clinical practice. **4 points:** Most of the report content is very useful for clinical practice. **5 points:** The report content is entirely in line with clinical needs and very useful.
5. Overall Evaluation (Considering all the above aspects, the overall satisfaction of the doctor with the GPT-4 generated report.**)**
Scoring Criteria: 1 point: Very dissatisfied. **2 points:** Dissatisfied. **3 points:** Neutral. **4 points:** Satisfied. **5 points:** Very satisfied.

* *Using this scale, doctors can comprehensively evaluate the quality of pathology interpretation reports generated by GPT-4. By summarizing the scores, it's possible to roughly determine GPT-4's level of understanding and interpreting pathology reports, as well as its potential value in clinical applications.*

Fig. 2 **A** Pathology report understanding level assessment scale. **B** Pathology report interpretation template. **C** Pathology Artificial Intelligence Quality Index. The scales and template were designed by large language model (LLM), and the pathologist modified and organized the scale

Table 1 Baseline health literacy levels

Health Literacy Dimension	Average Score
Feeling Understood and Supported by Healthcare Providers	3.92
Having Sufficient Information to Manage My Health	3.83
Actively Managing My Health	3.58
Social Support for Health	3.58
Appraisal of Health Information	3.83
Ability to Actively Engage with Healthcare Providers	3.83
Navigating the Healthcare System	3.42
Ability to Find Good Health Information	3.75
Understanding Health Information Well Enough to Know What to Do	3.92

Scores on the HLQ dimensions range from 1 to 4, with higher scores indicating higher levels of health literacy

literacy principles. The scale drew from the Health Literacy Questionnaire (HLQ) and other key research on health literacy [15–18]. It was designed to assess the clarity, relevance, and ease of understanding of key information in pathology reports, specifically for individuals with no medical background. The scale was refined through pilot testing to ensure its applicability for the study population.

A pathology report interpretation template is depicted in Fig. 2B. This template is intended as a general framework; specific content needs to be filled in and adjusted according to the actual details of each pathology report. This aims to assist individuals without a medical background in understanding the content and importance of pathology reports. The iterative prompt engineering involved multiple steps: First Prompt: "Summarize the pathology report for a layperson." Refinement: "Summarize the pathology report in simple language, explaining the diagnosis, significance, and next steps." Final Prompt: "Translate the pathology report into easy-to-understand language, include diagnosis, clinical significance, treatment options, and follow-up recommendations." The OPRs were generated using the refined templates. Each section of the template was filled with specific details from the pathology reports, ensuring consistency and comprehensibility. Examples of these templates and filled reports are illustrated in Figs. 2B and 3.

A pathology AI quality index is shown in Fig. 2C. This index was developed using GPT-4 and further refined through discussions with pathologists, who finalized the content and scoring criteria. Using this scale, doctors can comprehensively evaluate the quality of pathology interpretation reports generated by GPT-4. By summarizing the scores, it is possible to roughly determine GPT-4's level of understanding and interpreting pathology reports, as well as its potential value in clinical applications. This method was designed to rigorously compare the IPRs generated by GPT-4 against the standards set by the OPRs. The evaluation was conducted across five key dimensions by three pathologists, each with over a decade of professional experience: Accuracy (Dimension A), Interpretative Depth (Dimension B), Readability (Dimension C), Clinical Relevance (Dimension D), and Overall Evaluation (Dimension E). Pathologist X is a general pathologist working in a university hospital with expertise in oncologic pathology; Pathologist Y is a thoracic pathologist with specialization in lung cancer diagnostics, working at a non-university cancer center; and Pathologist Z is a gastrointestinal pathology expert affiliated with a leading academic medical center. All pathologists have extensive experience in analyzing complex pathology reports and contributing to AI-assisted diagnostic models. Their diverse backgrounds ensured a comprehensive evaluation of the pathology reports from different perspectives. This comprehensive review aimed to determine how well the GPT-4-generated reports captured the essence of the OPRs. The results, as adjudicated by the pathologists—referred to as Pathologist X, Pathologist Y, and Pathologist Z.

To evaluate the text complexity of both OPRs and IPRs, we calculated the word count using the word count feature in Microsoft Office 365 (Microsoft Corporation, Redmond, WA, USA). This method provided a quantitative measure of report length, allowing us to compare word counts across different types of malignancies and between OPRs and IPRs.

Patient data anonymization and security

To secure patient data, all identifying information was anonymized before being processed by the LLM/GPT model. The anonymization process ensured that no personal information, such as names, dates of birth, or medical record numbers, was included in the dataset. Additionally, the LLM was used in a secure, isolated environment that complied with data protection regulations, including [specific regulations if applicable, e.g., GDPR

A

Original Pathology Reports

Gross Findings:

Frozen Section: A lobe of lung tissue, measuring 19x10.5x3 cm, partially dissected, bronchus length 0.5 cm, diameter 2 cm. The bronchus was opened to reveal all segments of the lung. The bronchial mucosa was relatively smooth. A gray-white mass, measuring 4x3x2.8 cm, was located 2.5 cm from the bronchial resection margin, close to the pleura. The cut surface of the mass was gray-white and solid, with unclear boundaries from the surrounding tissue. The pleura near the mass was rough. One lymph node was found near the bronchus, measuring 0.5 cm in diameter.

Pathological Diagnosis: Frozen section and paraffin-embedded remaining tissue:

Specimen Name: Lung lobectomy. **Tumor Type:** Peripheral type; solitary. **Tumor Location:** Right lower lobe of the lung.

Tumor Size: 4x3x2.8 cm. **Histological Type:** Invasive adenocarcinoma, non-mucinous type (acinar pattern about 35% (complex glandular structures about 5%), papillary pattern about 5%, solid pattern about 55%, micropapillary pattern about 5%). Histological Grade: Poorly differentiated.

IASLC New Classification: Grade 3. **Pleural Invasion:** PL0. **Airway Spread (STAS):** Not seen. **Vascular Invasion:** Present. **Nerve Invasion:** Present.

Bronchial Margin: Negative. **Distance from Tumor to Bronchial Margin:** 2.5 cm. **Lymph Node Metastasis:** No cancer metastasis in the peribronchial lymph node (0/1).

Immunohistochemistry Results (Slide 3): TTF-1 (+), CK7 (+), NapsinA (+), P63 (few +), CK5/6 (-), P40 (-), Ki67 (about 5% +).

Special Stains (Slide 3): Elastic fibers (+).

Lymph Node Gross Findings: Frozen Section:

Group 2: One gray-brown tissue piece, measuring 1.4x1.2x0.2 cm. Group 4: One gray-brown tissue piece, measuring 1.2x0.7x0.3 cm.

Group 7: One gray-brown tissue piece, measuring 1.6x1.2x0.4 cm. Group 9: One gray-brown tissue piece, measuring 1x0.7x0.3 cm.

Group 10: One gray-brown tissue piece, measuring 1x0.7x0.2 cm. Group 11: Two gray-brown tissue pieces, each measuring 1x0.8x0.2 cm.

Microscopic Findings: Pathological Diagnosis:

Frozen Section: (Group 7) Cancer metastasis in the lymph node (1/1); No cancer metastasis in the other lymph nodes: (Group 2) 0/1, (Group 4) 0/1, (Group 9) 0/1, (Group 10) 0/1, (Group 11) 0/2.

B

Interpretive Pathology Reports

1. Report Overview

Report Type: This is a postoperative pathological report.

Case Information: The patient is a 54-year-old male.

2. Sample Information

Sample Source: The sample was obtained from the right lower lobe of the patient's lung.

Sample Type: The sample is a piece of lung tissue.

3. Gross and Microscopic Findings

Findings Description:

Gross Examination: A portion of lung tissue measuring 19x10.5x3 cm was examined. The bronchus (airway) is 0.5 cm long and 2 cm in diameter. Inside the lung tissue, a gray-white tumor measuring 4x3x2.8 cm was found near the pleura (the lung's outer layer). The tumor has an unclear boundary with the surrounding tissue, and the pleura near the tumor is rough. A lymph node near the bronchus, 0.5 cm in diameter, was also found.

Microscopic Examination: The tumor is identified as an invasive adenocarcinoma, a type of lung cancer. The tumor is classified as a non-mucinous type with various patterns: acinar (35%), papillary (5%), solid (55%), and micropapillary (5%). The tumor is low-grade (less differentiated) and has invaded nerves and blood vessels but not the pleura. The bronchial margin is clear of cancer cells, with a distance of 2.5 cm from the tumor. Lymph nodes near the bronchus do not show cancer spread.

Immunohistochemistry:

Positive markers: TTF-1, CK7, NapsinA, and partially P63.

Negative markers: CK5/6 and P40.

Proliferation marker Ki67 shows about 5% positive cells.

Elastic fibers are present in the sample.

Lymph Node Findings:

Cancer spread is detected in the lymph node from group 7. Other lymph nodes from groups 2, 4, 9, 10, and 11 show no cancer spread.

4. Diagnosis Results

Results Explanation:

The patient has invasive adenocarcinoma in the right lower lobe of the lung.

The tumor is classified as low-grade, indicating it is more aggressive and less differentiated.

Cancer cells have spread to nearby lymph nodes in one group but not in others.

Immunohistochemistry results support the diagnosis of lung adenocarcinoma.

The clear surgical margin suggests the tumor has been removed with a good safety margin.

5. Recommendations and Explanations

Next Steps:

Follow-up appointments with an oncologist for further evaluation and treatment planning.

Potential treatments include surgery, chemotherapy, radiation therapy, or targeted therapy, depending on the cancer stage and overall health of the patient.

Health Guidance:

Maintain a healthy diet and lifestyle to support overall well-being.

Avoid smoking and exposure to lung irritants to improve lung health.

Regular follow-ups and imaging studies as recommended by the healthcare provider.

6. Frequently Asked Questions

Q: What is adenocarcinoma?

A: Adenocarcinoma is a type of cancer that originates in glandular tissues, commonly found in the lungs.

Q: What does low-grade tumor mean?

A: A low-grade tumor is less differentiated, meaning the cancer cells look less like normal cells and tend to grow and spread more aggressively.

Q: What is the significance of lymph node involvement?

A: Cancer spread to lymph nodes often indicates a more advanced stage, requiring comprehensive treatment.

Q: What are the next steps after this diagnosis?

A: The patient will need to consult with an oncologist to discuss further treatment options, which may include surgery, chemotherapy, or radiation therapy.

Fig. 3 Application of interpretive pathology report (IPR). **A** Original pathology report (OPR). **B** Corresponding IPR

or HIPAA]. These measures ensured that no sensitive patient data was exposed or accessible outside the study, safeguarding patient confidentiality while allowing for accurate AI-generated pathology report analysis.

Statistical analyses

The data are presented as either mean ± standard deviation, minimum and maximum values. We evaluated the data from groups through Shapiro–Wilk test for normality test. If the data followed normal distribution, t test was used for statistical analysis; otherwise, Mann–Whitney U test was used for statistical analysis. The relationships between continuous variables were determined using Spearman's correlation analysis. A $P<0.05$ was deemed to indicate statistical significance. All statistical calculations were carried out using R software, version 4.3.2 (Lucent Technologies, Murray Hill, NJ, USA).

Results

Characteristics of sample

Between October and December 2023, a total of 3,082 patients were screened at four institutions, as illustrated in Fig. 1. Of these, 2,353 patients were excluded due to pathologically confirmed benign tumors. Additionally, 31 patients with rare malignant tumors were excluded due to challenges associated with follow-up data collection, which primarily included the geographical dispersion of patients, variability in hospital record-keeping practices, and inconsistent communication channels across institutions. Consequently, the study included 698 patients for further analysis. The majority of the study cohort were female, as detailed in Table 2. The participants' ages ranged widely from 24 to 82 years, with an average age of 55.27 years. A significant proportion, approximately 85.67%, were below the age of 65.

Table 2 Basic characteristics of patients

Cancer Sites	Patients	Age (years)[a]	Sex (M, F)
All sites	698	55.27 ± 12.66 (24, 82)	290 (41.55%), 408 (58.45%)
Brain	32	58.16 ± 11.25 (34, 79)	13 (40.62%), 19 (59.38%)
Thyroid	76	44.53 ± 11.75 (24, 74)	32 (42.11%), 44 (57.89%)
Breast	86	50.98 ± 11.32 (25, 80)	0 (0.00%), 86 (100.00%)
Lung	98	58.04 ± 11.64 (32, 82)	49 (50.00%), 49 (50.00%)
Esophagus	10	63.10 ± 7.28 (50, 71)	7 (70.00%), 3 (30.00%)
Gastric	30	55.30 ± 12.42 (25, 80)	18 (60.00%), 12 (40.00%)
Liver	32	61.53 ± 12.47 (35, 82)	24 (75.00%), 8 (25.00%)
Pancreatic	18	56.39 ± 11.63 (37, 76)	15 (83.33%), 3 (16.67%)
Colorectal	74	61.03 ± 12.77 (27, 82)	31 (41.89%), 43 (58.11%)
Kidney	61	62.08 ± 10.02 (34, 82)	31 (50.82%), 30 (49.18%)
Prostate	37	72.89 ± 3.70 (67, 82)	37 (100.00%), 0 (0.00%)
Bladder	50	70.06 ± 5.75 (58, 81)	33 (66.00%), 17 (34.00%)
Ovary	61	52.11 ± 6.27 (39, 68)	0 (0.00%), 61(100.00%)
Uterus	33	53.45 ± 3.24 (47, 59)	0 (0.00%), 33 (100.00%)

[a] Data are means ± SDs, with ranges in parentheses

Text data extractions

As shown in Table 3, the average word count of OPRs was 549.98. Notably, brain malignancies had the lowest average word count for their OPRs, at 406.78, whereas ovarian malignancies had the highest, at 961.21. The analysis also revealed an average of 19.73 medical terms per OPR across all studied categories of malignant tumors. Prostate malignancies had the fewest average medical terms, at 14.46, while ovarian malignancies had the most, averaging 30.43 medical terms.

We observed that the average word count for OPRs across all types of malignant tumors was 549.98, while the average word count for IPRs was significantly higher at 787.44. Liver malignancies had the lowest average word count for OPRs (441.41) and IPRs (775.25). In contrast, ovarian malignancies had the highest average word count for OPRs (961.21), while esophagus malignancies had the highest average word count for IPRs (833.80). This suggests that although there is significant variation in the word count of OPRs among different malignancies ($P<0.001$), the variation in IPR word counts is less pronounced ($P=0.088$, Figs. 4 and 5).

Moreover, the word count for the OPRs of ovarian malignant tumors was higher than that for the IPRs ($P<0.001$), whereas the word counts for the OPRs of other cancer types were lower than those for the IPRs ($P<0.001$).

Consistency evaluation of expression content

To assess the fidelity and quality of IPRs relative to OPRs, we utilized a consistency evaluation scale developed with GPT-4, as shown in Fig. 2C. The results, as adjudicated by the pathologists—referred to as Pathologist X, Pathologist Y, and Pathologist Z—showed no significant statistical differences in their assessments across the dimensions. Remarkably, all dimensions consistently scored 4 or higher, with Readability (Dimension C) notably achieving a unanimous score of 5, as detailed in Table 4.

Pathology report reading time

Two groups of volunteers separately read OPRs (V_A, V_B, and V_C) and IPRs (V_A, V_B, and V_C), with their reading times recorded (Table 5, Fig. 4 and 6). The average reading time for OPRs across all types of malignant tumors was 401.76 s. Notably, brain malignancies had the shortest average reading time at 305.47 s, whereas ovarian malignancies had the longest at 700.64 s, indicating

Table 3 Characteristics of pathology reports

Cancer Sites	Pathology reports	OPRs (Word count)*	OPRs (medical terms)*	IPRs (Word count)*	*P* value**
All sites	698	549.98 ± 154.72	19.73 ± 5.22	787.44 ± 53.51	< 0.001
		(304, 1154)	(10, 34)	(657, 875)	
Brain	32	406.78 ± 28.00	16.81 ± 2.89	786.47 ± 51.31	< 0.001
		(306, 454)	(10, 22)	(701, 874)	
Thyroid	76	434.45 ± 52.39	16.84 ± 3.00	789.34 ± 56.78	< 0.001
		(304, 564)	(10, 24)	(695, 875)	
Breast	86	485.23 ± 56.85	19.43 ± 3.22	785.42 ± 56.48	< 0.001
		(398, 686)	(11, 25)	(697, 874)	
Lung	98	552.62 ± 29.48	20.96 ± 3.61	785.42 ± 56.48	< 0.001
		(500, 598)	(12, 32)	(697, 874)	
Esophagus	10	448.90 ± 33.09	14.70 ± 2.63	833.80 ± 41.58	< 0.001
		(400, 498)	(12, 20)	(764, 875)	
Gastric	30	497.67 ± 34.91	15.30 ± 2.77	780.03 ± 47.01	< 0.001
		(443, 597)	(12, 21)	(701, 860)	
Liver	32	441.41 ± 70.15	14.75 ± 2.69	775.25 ± 51.74	< 0.001
		(306, 570)	(10, 20)	(698, 853)	
Pancreatic	18	461.89 ± 39.06	16.00 ± 4.64	788.61 ± 54.91	< 0.001
		(403, 517)	(10, 24)	(700, 875)	
Colorectal	74	513.89 ± 56.49	16.47 ± 2.88	784.73 ± 51.07	< 0.001
		(366, 654)	(12, 24)	(696, 874)	
Kidney	61	500.87 ± 30.93	21.95 ± 2.96	790.98 ± 50.48	< 0.001
		(428, 553)	(16, 28)	(702, 874)	
Prostate	37	453.11 ± 60.71	14.46 ± 2.26	808.08 ± 49.10	< 0.001
		(343, 568)	(10, 19)	(702, 873)	
Bladder	50	679.28 ± 47.49	23.48 ± 1.97	776.00 ± 55.78	< 0.001
		(602, 785)	(20, 28)	(697, 873)	
Ovary	61	961.21 ± 67.47	30.43 ± 1.88	781.70 ± 54.90	< 0.001
		(802, 1154)	(28, 34)	(657, 874)	
Uterus	33	671.48 ± 56.57	22.52 ± 3.02	797.94 ± 46.99	< 0.001
		(519, 777)	(18, 29)	(711, 871)	

Intentionally minimized to ensure the reports are accessible to a non-medical audience. The goal of the IPRs is to enhance understanding for patients and laypersons, which is why medical terms were avoided in the report generation process

OPRs Original pathology reports, *IPRs* Interpretive pathology reports

* Data are means ± SDs, with ranges in parentheses

** The OPRs and IPRs of different cancer sites were analyzed statistically

statistically significant differences in reading times for OPRs across tumor types ($P < 0.001$). In contrast, the average reading time for IPRs was 430.67 s, with the shortest for liver malignancies at 418.88 s, and the longest for esophagus tumors at 452.10 s. No significant differences were observed in the reading times for IPRs across the tumor types ($P = 0.413$).

A comparison of the reading times between OPRs and IPRs for all types of malignant tumors revealed that OPRs were generally read faster than IPRs, with a statistically significant difference ($P < 0.001$). However, for bladder, ovarian, and uterus malignancies, the reading times were longer for OPRs compared to IPRs, with these differences also being statistically significant ($P < 0.001$ for each).

Understanding level assessment

The evaluation further involved a multidimensional scoring of OPRs and IPRs using the Pathology Report Understanding Level Assessment Scale, as shown in Table 5, Figs. 2A and 6. Across all types of malignant tumors, the average score for OPRs was 5.23. In comparison, the average score for IPRs was significantly higher, at 7.98. This disparity in scoring between OPRs and IPRs across all tumor types was statistically significant ($P < 0.001$).

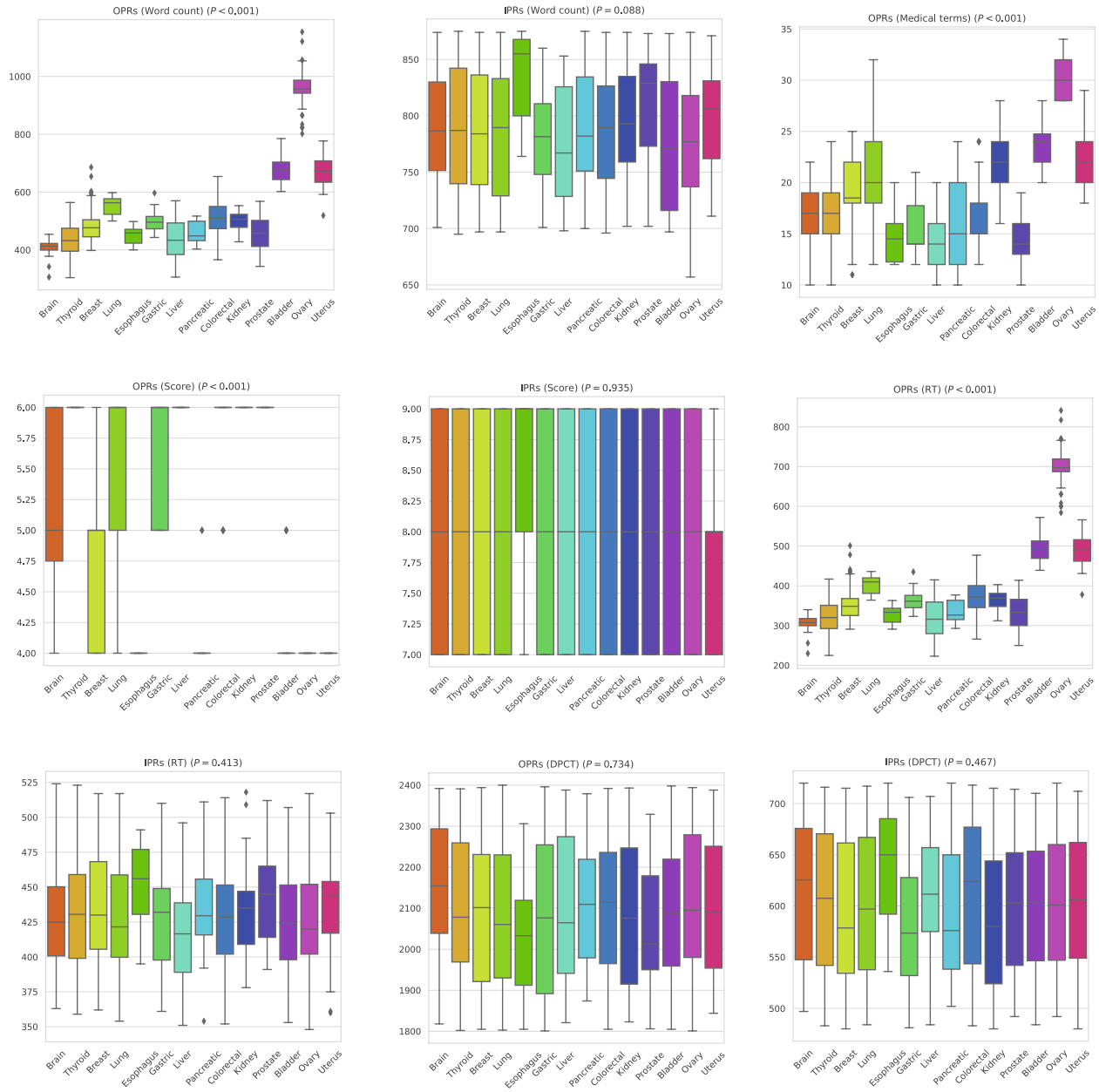

Fig. 4 Comparative analysis of original pathology reports (OPRs) and interpretive pathology reports (IPRs) metrics across cancer sites. RT: Reading time. DPCT: Doctor-patient communication time

Doctor-patient communication

After volunteers (A, B and C) finished reading the OPRs, the doctor engaged in simulated doctor-patient communication with the volunteers to explain the patient's condition and recorded the communication time (Table 5, Figs. 4 and 6D). Across all types of malignant tumors, the average communication time was 2091.25 s. Specifically, brain malignancies exhibited the longest average communication time at 2154.41 s, while prostate malignancies had the shortest at 2062.03 s. Statistical analysis revealed no significant differences in communication times across the different tumor types ($P=0.734$). Additionally, after volunteers (D, E and F) finished reading the IPRs, the doctor conducted simulated doctor-patient communication based on the report content, explained the patient's condition, and recorded the communication time. Across all types of malignant tumors, the average communication time was 599.15 s. The longest average

Fig. 5 Original pathology reports (OPRs) vs. interpretive pathology reports (IPRs) comparison (word count, score, reading time and doctor-patient communication time) by cancer site. RT: Reading time. DPCT: Doctor-patient communication time

communication time occurred with esophagus malignancies, at 638.30 s, while the shortest was for gastric malignancies, at 581.80 s. Statistical analysis indicated no significant differences in communication times among the various types of malignant tumors ($P=0.467$). Further analysis showed that, regardless of the tumor type, the communication time after reading the OPRs was significantly longer than that after reading the IPRs, a difference that was statistically significant ($P<0.001$).

Correlation of OPRs and IPRs metrics

We analyzed the correlation between various metrics of OPRs and IPRs, as illustrated in Fig. 6. This heatmap provides a clear and intuitive display of the correlations among nine key metrics within OPRs and IPRs. It reveals a strong correlation between word count, medical terms, score, and reading time for OPRs. The figure serves as a visually intuitive tool to identify both the strength and the direction of relationships between these metrics.

Discussion

Our research on the application of GPT-4-generated IPRs in enhancing doctor-patient communication supports the expanding role of AI within healthcare, offering valuable insights that are particularly relevant to surgical settings. The principal outcomes of our study substantiate the integration of AI to augment patient comprehension and communication efficacy. Comparatively

Table 4 Evaluation of consistency between original radiology reports and interpretive radiology reports

Cancer Site	Dimension A (Accuracy)	Dimension B (Interpretation Depth)	Dimension C (Readability)	Dimension D (Clinical Relevance)	Dimension E (Overall Evaluation)
All sites	4.95	4.95	5	4.92	4.84
Brain	5	4.97	5	4.91	4.91
Thyroid	4.93	4.96	5	4.83	4.83
Breast	4.94	4.94	5	4.91	4.8
Lung	4.95	4.94	5	4.93	4.83
Esophagus	5	5	5	4.9	4.9
Gastric	4.97	4.97	5	4.9	4.83
Liver	5	4.94	5	4.97	4.91
Pancreatic	4.89	4.89	5	4.89	4.67
Colorectal	4.96	4.97	5	4.95	4.88
Kidney	4.93	4.95	5	4.97	4.85
Prostate	4.95	4.95	5	4.95	4.84
Bladder	4.96	4.92	5	4.96	4.86
Ovary	4.93	4.95	5	4.93	4.82
Uterus	4.94	4.97	5	4.88	4.79

reviewing recent scholarly work situates our study within the modern scientific discourse, emphasizing the novel contributions and prospective advancements our findings introduce to the field [2, 8].

Across all types of malignant tumors, the use of IPRs resulted in significantly higher patient understanding scores compared to traditional OPRs, with an average improvement from 5.23 to 7.98 on the Pathology Report Understanding Level Assessment Scale. Furthermore, the study found a substantial reduction in doctor-patient communication time when using IPRs, decreasing from an average of 2091.25 s to 599.15 s, underscoring the potential time-saving benefits of AI-assisted reports. These findings suggest that AI-generated reports can enhance doctor-patient communication while also improving overall healthcare efficiency.

In addition to improving communication time and comprehension, the consistency evaluation conducted by pathologists highlighted that the IPRs generated by GPT-4 were highly accurate, scoring consistently across dimensions such as Accuracy, Interpretative Depth, and Readability. This consistency in evaluation across different tumor types supports the robustness of the AI-generated reports, indicating their potential for widespread clinical application. The strong correlation observed between OPR and IPR metrics further emphasizes the effectiveness of the AI model in maintaining clinical relevance while simplifying report content for patient understanding. This enhanced understanding is critical as it directly influences patient engagement and empowerment. Patients who grasp their medical conditions and the logic behind their treatment options are more inclined to adhere to recommended treatments and engage in proactive health management. This link between comprehension and compliance is well-documented in healthcare literature, with our data providing robust evidence of AI's pivotal role in fostering this understanding [19–22].

Moreover, recent studies have increasingly acknowledged AI's capability to enhance the accessibility and comprehensibility of medical documentation. For instance, Amin et al. employed three prominent large language models—ChatGPT, Google Bard, and Microsoft Bing—to simplify radiology reports [23]. Subsequently, they solicited assessments from pertinent clinical practitioners concerning the accuracy of each model's output. Nevertheless, the research did not address the comprehensibility of these simplified radiology reports for individuals lacking a medical background. Consequently, the applicability of large language models in making radiological information accessible to a broader, non-specialist audience remains unverified [23]. Truhn et al. utilized GPT-4 to generate structured pathology reports, demonstrating that structured reports generated by large language models are consistent with those produced by pathologists [24]. This indicates that LLMs could potentially be employed routinely to extract ground truth data for machine learning from unstructured pathology reports in the future. However, this study focused only on evaluations by professionals and lacks an assessment of the usability of AI-generated reports in broader scenarios. Similarly, Steimetz et al. examined methods for

Table 5 Volunteers' evaluation of the original pathology reports and the interpretive pathology reports generated based on GPT-4

Cancer Sites	V (A, B, C)[a] OPRs (RT)[b]	V (D, E, F)[a] IPRs (RT)	*P*	V (A, B, C) OPRs (Score)	V (D, E, F) IPRs (Score)	*P*	V (A, B, C) OPRs (DPCT)	V (D, E, F) IPRs (DPCT)	*P*
All sites	401.76 ± 112.06 (223, 841)	430.67 ± 37.81 (348, 524)	< 0.001	5.23 ± 0.88 (4, 6)	7.98 ± 0.82 (7, 9)	< 0.001	2091.25 ± 170.90 (1801, 2400)	599.15 ± 69.31 (480, 720)	< 0.001
Brain	305.47 ± 21.55 (230, 340)	428.06 ± 37.64 (363, 524)	< 0.001	5.03 ± 0.74 (4, 6)	8.00 ± 0.76 (7, 9)	< 0.001	2154.41 ± 148.01 (1818, 2392)	612.66 ± 69.27 (497, 720)	< 0.001
Thyroid	321.33 ± 38.80 (225, 417)	432.11 ± 39.31 (359, 523)	< 0.001	6.00 ± 0.00 (6, 6)	7.87 ± 0.82 (7, 9)	< 0.001	2104.17 ± 165.55 (1802, 2391)	601.67 ± 69.50 (483, 716)	< 0.001
Breast	354.56 ± 41.56 (291, 501)	435.30 ± 40.24 (362, 517)	< 0.001	4.90 ± 0.72 (4, 6)	7.95 ± 0.85 (7, 9)	< 0.001	2083.97 ± 183.14 (1805, 2394)	590.48 ± 70.42 (480, 715)	< 0.001
Lung	402.52 ± 21.47 (364, 436)	428.65 ± 40.78 (354, 517)	< 0.001	5.41 ± 0.69 (4, 6)	7.98 ± 0.81 (7, 9)	< 0.001	2074.30 ± 169.04 (1803, 2400)	597.28 ± 71.82 (484, 717)	< 0.001
Esophagus	326.90 ± 24.18 (291, 363)	452.10 ± 32.22 (395, 491)	< 0.001	4.00 ± 0.00 (4, 4)	8.40 ± 0.84 (7, 9)	< 0.001	2030.00 ± 163.08 (1805, 2306)	638.30 ± 65.62 (536, 720)	< 0.001
Gastric	362.53 ± 25.41 (323, 435)	428.00 ± 37.60 (361, 510)	< 0.001	5.70 ± 0.47 (5, 6)	8.03 ± 0.85 (7, 9)	< 0.001	2079.67 ± 198.41 (1801, 2396)	581.80 ± 67.32 (481, 706)	< 0.001
Liver	321.41 ± 51.14 (223, 415)	418.88 ± 37.06 (351, 496)	< 0.001	6.00 ± 0.00 (6, 6)	8.03 ± 0.86 (7, 9)	< 0.001	2088.75 ± 181.83 (1821, 2388)	612.81 ± 59.89 (484, 707)	< 0.001
Pancreatic	336.56 ± 28.48 (293, 377)	433.72 ± 37.05 (354, 511)	< 0.001	4.17 ± 0.38 (4, 5)	7.89 ± 0.90 (7, 9)	< 0.001	2103.56 ± 148.86 (1874, 2379)	593.33 ± 73.37 (502, 720)	< 0.001
Colorectal	374.41 ± 41.24 (266, 477)	429.35 ± 38.25 (352, 514)	< 0.001	5.89 ± 0.31 (5, 6)	8.00 ± 0.81 (7, 9)	< 0.001	2095.24 ± 161.22 (1805, 2392)	607.85 ± 73.10 (483, 718)	< 0.001
Kidney	364.85 ± 22.54 (312, 403)	431.59 ± 30.75 (378, 518)	< 0.001	6.00 ± 0.00 (6, 6)	8.08 ± 0.80 (7, 9)	< 0.001	2088.69 ± 181.35 (1823, 2393)	585.89 ± 67.33 (480, 715)	< 0.001
Prostate	330.03 ± 44.33 (250, 414)	440.27 ± 32.33 (391, 512)	< 0.001	6.00 ± 0.00 (6, 6)	7.89 ± 0.88 (7, 9)	< 0.001	2062.03 ± 147.32 (1806, 2329)	598.19 ± 69.89 (492, 714)	< 0.001
Bladder	494.98 ± 34.63 (439, 572)	424.60 ± 37.97 (353, 507)	< 0.001	4.16 ± 0.37 (4, 5)	8.02 ± 0.77 (7, 9)	< 0.001	2085.76 ± 175.08 (1805, 2398)	602.08 ± 68.73 (484, 710)	< 0.001
Ovary	700.64 ± 49.20 (584, 841)	426.98 ± 38.05 (348, 517)	< 0.001	4.00 ± 0.00 (4, 4)	7.95 ± 0.85 (7, 9)	< 0.001	2112.07 ± 177.29 (1801, 2394)	602.00 ± 65.65 (492, 720)	< 0.001
Uterus	489.27 ± 41.26 (378, 566)	436.06 ± 36.86 (360, 503)	< 0.001	4.00 ± 0.00 (4, 4)	7.94 ± 0.75 (7, 9)	< 0.001	2092.79 ± 167.61 (1844, 2388)	598.55 ± 69.28 (480, 712)	< 0.001

OPRs Original pathology reports, *IPRs* Interpretive pathology reports, *RT* Reading time, *DPCT* doctor-patient communication time

[a] Volunteers A, B, and C were high school educated people with non-medical backgrounds, aged 50, 50, and 52 years old, and their genders were male, female, and female, respectively. In addition, the matched volunteers D, E and F are also high school educated people with non-medical background, their ages are 50, 51 and 51 years old respectively, and their genders are male, female and female respectively

[b] Data are means ± SDs, with ranges in parentheses

simplifying medical documents to improve patient comprehension, finding that enhancing readability directly impacts patient engagement and satisfaction [9]. In addition, Singhal et al. showed that LLMs effectively encode clinical knowledge, reinforcing their potential in improving healthcare communication [8]. Harrer further discusses the ethical considerations and complexities of integrating large language models into medical systems, emphasizing the importance of thoroughly evaluating their real-world applications to ensure both patient safety and accuracy [11].

Building on previous research, our study simulated interactions between doctors and patients regarding the interpretation of postoperative pathology reports in surgical settings [9, 23, 24]. It demonstrated the universal applicability of explanations generated by large language models across different demographic groups. This research goes beyond simply translating and simplifying professional reports; it highlights the importance of such models as bridges between professional and non-professional domains, thereby expanding the use of large language models in real-world healthcare settings.

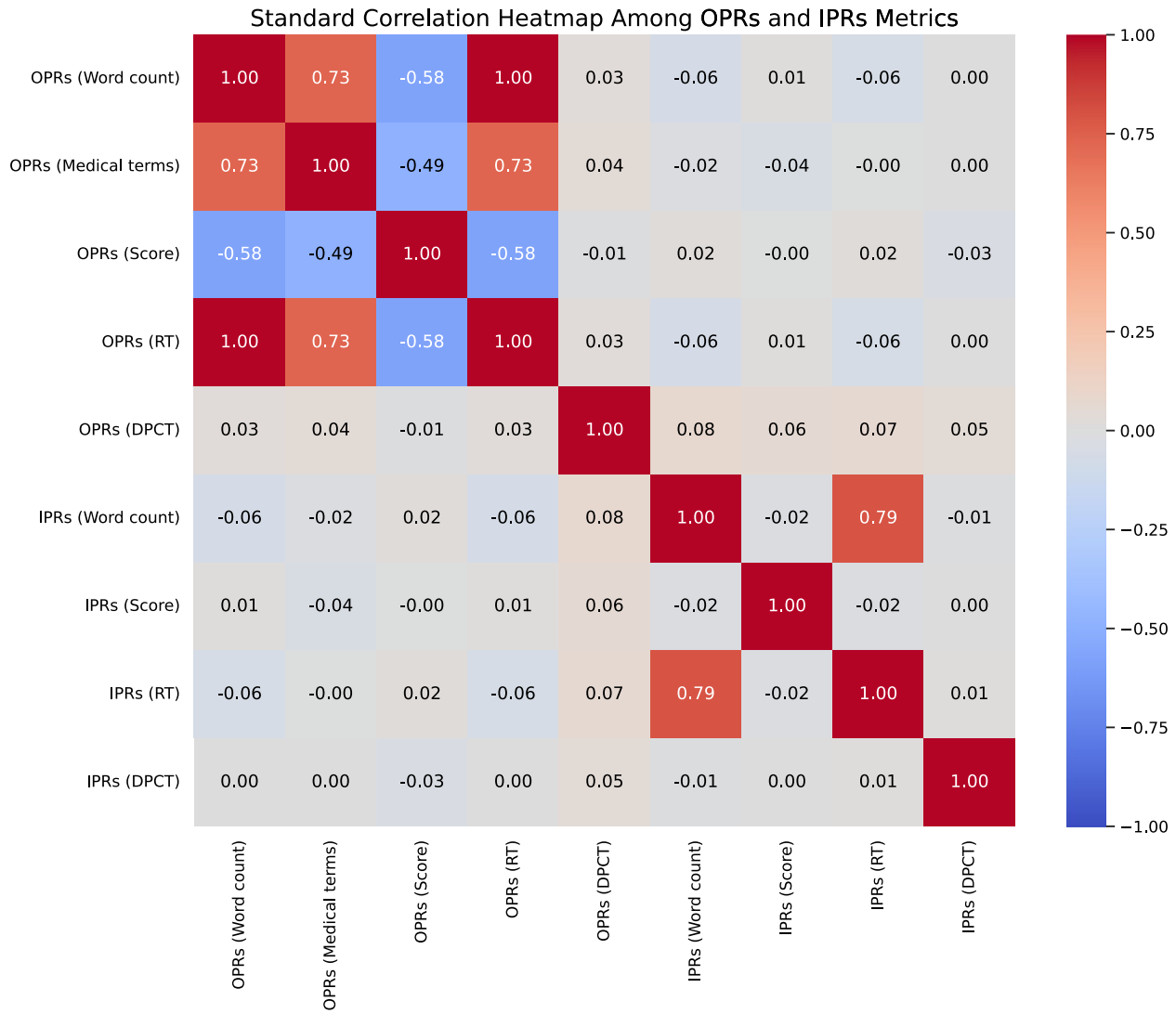

Fig. 6 Correlation heatmap of original pathology reports (OPRs) and interpretive pathology reports (IPRs). RT: Reading time. DPCT: Doctor-patient communication time

Another significant observation from our study was the reduction in communication time between doctors and patients. The average duration for doctors to explain pathological reports decreased dramatically from approximately 35 min with OPRs to about 10 min with IPRs, marking a reduction of over 70% in communication time. This efficiency gain is especially critical in surgical settings where time is scarce, and the cognitive load on patients is substantial due to the stress and complexity of their medical situations. By minimizing the time needed to convey essential information, doctors can dedicate more time to addressing patient concerns, answering questions, and providing personalized care. Additionally, this efficiency may lead to increased patient throughput, essential in high-demand environments like surgical units. The scarcity of medical resources globally further underscores the importance of these findings, suggesting that large language models can significantly alleviate the strain on healthcare resources.

Additionally, our study demonstrates that the IPRs generated by GPT-4 show a high degree of consistency with the OPRs, as evaluated across key dimensions such as accuracy, interpretative depth, and readability. These findings underscore the robustness of the evaluative framework in verifying that the IPRs accurately represent the key insights of the OPRs. This framework not only ensures that the generated reports are consistent with the original medical data, but also plays a crucial role in

maintaining the integrity and reliability of the pathology interpretation process. By systematically comparing multiple dimensions, the framework provides a comprehensive assessment that helps to identify potential discrepancies and ensures the clinical relevance of the reports. This rigorous approach allows for the use of AI-generated reports with greater confidence in real-world medical settings, ultimately contributing to more efficient doctor-patient communication and improved healthcare outcomes. With proper training and model adjustments, LLMs like GPT-4 can achieve high levels of accuracy and reliability in interpreting and simplifying complex surgical pathology reports, vital for patient recovery and comprehension post-surgery.

The implications of these findings for clinical practice are profound. Integrating AI-generated IPRs into healthcare systems can be achieved through several practical steps. First, hospitals and clinics can implement AI models like GPT-4 to automatically generate simplified, patient-friendly pathology reports alongside traditional reports. These AI-generated reports can be shared with patients via patient portals or during face-to-face consultations. Additionally, training healthcare providers to utilize AI-generated reports as communication tools during consultations can further enhance patient understanding. By offering easy-to-understand summaries, patients are more likely to engage with their care plans, leading to greater satisfaction and better adherence to treatment, ultimately contributing to improved health outcomes. Additionally, reducing the time spent on routine explanations can alleviate workload pressures on healthcare professionals, potentially enhancing job satisfaction and reducing burnout.

However, it is important to note that this study was conducted in a Chinese-speaking region, and all pathology reports, whether original or interpretive, were written in Chinese. The language and cultural background may influence the generalizability of our findings. During the template generation and evaluation process, we carefully considered the use of Traditional Chinese Medicine (TCM) terminology and the specific structure of Chinese pathology reports. Therefore, in real-world applications, it is crucial to take cultural and linguistic contexts into account when applying the conclusions of this study.

While our study utilized volunteers to simulate patient interactions, we acknowledge the potential differences between volunteers and real patients. Real patients in clinical settings often experience a range of emotions, such as anxiety, fear, and distress, which can influence their behavior, decision-making, and communication efficiency. Studies have shown that patients under emotional distress may struggle with comprehension and retention of medical information, potentially impacting their ability to engage in effective communication with healthcare providers [25]. In contrast, volunteers in our study, who were aware of the non-threatening nature of the environment, did not experience these emotional stressors. As such, future research should aim to include real patients to better capture the complexity of clinical interactions and the impact of emotional states on communication outcomes.

Despite the promising results, our study acknowledges several key limitations that warrant careful consideration. These limitations highlight areas for cautious interpretation of the results and suggest potential avenues for future research to address these gaps. First, our study's heavy reliance on the capabilities of GPT-4, a specific version of Large Language Models developed by OpenAI, raises questions about the generalizability of our findings. While GPT-4 is renowned for its sophisticated natural language processing capabilities, it represents only one example of such technologies. Different LLMs may exhibit varying effectiveness based on their training data and underlying algorithms. Future research could explore the performance of other LLMs in similar tasks to verify if the observed benefits are replicable across different AI platforms. Second, the demographic and geographic diversity of our patient sample was confined to specific hospitals within a limited region, which may restrict the applicability of our results to other settings where patient populations differ significantly in terms of language, culture, and healthcare practices. Additionally, the sample size, while sufficient for statistical analysis, may not fully capture the variability and complexity of patient experiences across broader populations. Expanding the sample size and including a more diverse patient group in future studies could provide insights into how different populations interact with and benefit from AI-generated reports. Third, the primarily quantitative nature of our study provides a robust statistical foundation for evaluating the effectiveness of AI in improving patient understanding and communication efficiency. However, this approach may overlook the nuanced human aspects of doctor-patient interactions that are better captured through qualitative methods. Future studies might incorporate qualitative research techniques, such as in-depth interviews or focus groups, to gather more comprehensive insights into how patients and healthcare providers perceive and value the AI-generated interpretive reports. Fourth, one limitation of this study is the exclusion of hallucinations, a commonly reported error in LLM/GPT models, from the evaluation. Hallucinations refer to instances where the model generates information that is factually incorrect or fabricated, which could potentially affect the interpretation of AI-generated pathology reports. However, in this study, our primary focus was

on evaluating the accuracy, consistency, and readability of the pathology reports, specifically in relation to diagnostic content. As such, hallucinations were not included in the scope of this assessment. Future research should aim to investigate the occurrence of hallucinations in medical text generation and their potential implications for clinical practice, especially when using AI models in high-stakes decision-making environments. Fifth, we acknowledge the small number of volunteers and the potential impact on baseline characteristics. Different groups were chosen to avoid bias introduced by familiarity with the report format. However, controlling for baseline characteristics is crucial. The health literacy levels of the volunteers were assessed and considered in the analysis. Therefore, these limitations underscore the need for cautious interpretation of our study results and highlight the importance of addressing these areas in future research. By expanding the scope, diversity, and depth of research into the use of AI in healthcare, we can better understand the capabilities and limitations of these technologies and work towards maximizing their benefits while minimizing potential drawbacks.

Conclusion

In conclusion, our study demonstrates the potential benefits of using large language models (LLMs) like GPT-4 in the healthcare setting, particularly in processing and interpreting pathology reports. While the findings highlight the efficiency and accuracy of GPT-4 in generating interpretive pathology reports, we do not claim that patient outcomes or patient satisfaction were directly improved based on this study alone. Instead, this research illustrates the promise of AI tools in enhancing healthcare communication and streamlining clinical workflows, offering insights into the evolving role of AI in healthcare delivery. Future studies will be required to further investigate the impact of LLMs on patient satisfaction and clinical outcomes in diverse and real-world settings.

Acknowledgements
We acknowledge parts of this article were generated with GPT-4 (powered by OpenAI's language model; https://chat.openai.com/), but the output was confirmed by the authors. Thanks to the colleagues in the department of pathology for their help in this paper, your excellent work has made our research more efficient.

Authors' contributions
Xiongwen Yang and Yi Xiao wrote the main manuscript text. Di Liu and Huiyou Shi validated and conducted formal analysis. Huiyin Deng, Jian Huang, and Yun Zhang curated the data. Dan Liu, Maoli Liang, Jing Yao, XiaoJiang Zhou, Wankai Guo, and Yang He contributed to conceptualization and project administration. Xing Jin, Yongpan Sun, WeiJuan Tang, and Chuan Xu provided methodology and conducted review and editing. Chuan Xu also supervised the project, handled visualization, and secured funding.

Funding
Supported by Talent Fund of Guizhou Provincial People's Hospital.

Data availability
The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Declarations

Ethics approval and consent to participate
All procedures involving collection of tissue were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This retrospective compliance study was approved by the Ethics Review Committee of Guizhou Provincial People's Hospital (Ethics Number: 2024004), the Third Affiliated Hospital of Sun Yat-sen University (Ethics Number: B2023074), the Third Xiangya Hospital, Central South University (Ethics Number: 2024011), and Jiangxi Cancer Hospital (Ethics Number: JC2024006). Written informed consent was obtained from individual or guardian participants.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Thoracic Surgery, Guizhou Provincial People's Hospital, No. 83, Zhongshan East Road, Guiyang, Guizhou 550000, China. [2]NHC Key Laboratory of Pulmonary Immunological Diseases, Guizhou Provincial People's Hospital, Guiyang, Guizhou 550000, China. [3]Department of Cardio-Thoracic Surgery, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, Guangdong, China. [4]Department of Pathology, Guizhou Provincial People's Hospital, Guiyang, Guizhou, China. [5]Department of Anesthesiology, the Third Xiangya Hospital of Central South University, Changsha, Hunan, China. [6]Department of Thoracic Surgery, Jiangxi Cancer Hospital, Nanchang, Jiangxi, China. [7]Department of Radiology, Guizhou Provincial People's Hospital, Guiyang, Guizhou, China. [8]Department of Medical Records and Statistics, Guizhou Provincial People's Hospital, Guiyang, Guizhou, China. [9]Department of Respiratory Medicine, Guizhou Provincial People's Hospital, Guiyang, Guizhou, China.

Received: 10 June 2024 Accepted: 23 December 2024
Published online: 23 January 2025

References

1. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, et al. A large language model for electronic health records. NPJ Digital Med. 2022;5(1):194.
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40.
3. Yang X, Chu XP, Huang S, Xiao Y, Li D, Su X, Qi YF, Qiu ZB, Wang Y, Tang WF, et al. A novel image deep learning-based sub-centimeter pulmonary nodule management algorithm to expedite resection of the malignant and avoid over-diagnosis of the benign. Eur Radiol. 2024;34(3):2048–61.
4. Mossanen M, True LD, Wright JL, Vakar-Lopez F, Lavallee D, Gore JL. Surgical pathology and the patient: a systematic review evaluating the primary audience of pathology reports. Hum Pathol. 2014;45(11):2192–201.
5. Dunsch F, Evans DK, Macis M, Wang Q. Bias in patient satisfaction surveys: a threat to measuring healthcare quality. BMJ Glob Health. 2018;3(2):e000694.
6. Farley H, Enguidanos ER, Coletti CM, Honigman L, Mazzeo A, Pinson TB, Reed K, Wiler JL. Patient Satisfaction Surveys and Quality of Care: An Information Paper. Ann Emerg Med. 2014;64(4):351–7.
7. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023;330(9):866–9.
8. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
9. Steimetz E, Minkowitz J, Gabutan EC, Ngichabe J, Attia H, Hershkop M, Ozay F, Hanna MG, Gupta R. Use of Artificial Intelligence

Chatbots in Interpretation of Pathology Reports. JAMA Netw Open. 2024;7(5):e2412767.
10. Winograd A. Loose-lipped large language models spill your secrets: The privacy implications of large language models. Harvard J Law Technol. 2023;36(2):615.
11. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. 2023;90: 104512.
12. Birkhäuer J, Gaab J, Kossowsky J, Hasler S, Krummenacher P, Werner C, Gerger H. Trust in the health care professional and health outcome: A meta-analysis. PLoS ONE. 2017;12(2):e0170988.
13. Haskard Zolnierek KB, DiMatteo MR. Physician Communication and Patient Adherence to Treatment: A Meta-Analysis. Med Care. 2009;47(8):826.
14. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Safety. 2016;25(12):986–92.
15. Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health. 2013;13(1):658.
16. Dewalt DA, Berkman ND, Sheridan S, Lohr KN, Pignone MP. Literacy and health outcomes: a systematic review of the literature. J Gen Intern Med. 2004;19(12):1228–39.
17. Paasche-Orlow MK, Wolf MS. The causal pathways linking health literacy to health outcomes. Am J Health Behav. 2007;31(Suppl 1):S19-26.
18. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med. 2011;155(2):97–107.
19. Kravitz RL, Hays RD, Sherbourne CD, DiMatteo MR, Rogers WH, Ordway L, Greenfield S. Recall of recommendations and adherence to advice among patients with chronic medical conditions. Arch Intern Med. 1993;153(16):1869–78.
20. McDonald HP, Garg AX, Haynes RB. Interventions to enhance patient adherence to medication prescriptions: scientific review. JAMA. 2002;288(22):2868–79.
21. Schillinger D, Piette J, Grumbach K, Wang F, Wilson C, Daher C, Leong-Grotz K, Castro C, Bindman AB. Closing the loop: physician communication with diabetic patients who have low health literacy. Arch Intern Med. 2003;163(1):83–90.
22. Hibbard JH, Greene J. What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. Health Aff (Millwood). 2013;32(2):207–14.
23. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology. 2023;309(2):e232561.
24. Truhn D, Loeffler CM, Müller-Franzes G, Nebelung S, Hewitt KJ, Brandner S, Bressem KK, Foersch S, Kather JN. Extracting structured information from unstructured histopathology reports using generative pre-trained transformer 4 (GPT-4). J Pathol. 2024;262(3):310–9.
25. Oben P. Understanding the Patient Experience: A Conceptual Framework. J Patient Exp. 2020;7(6):906–10.

Publisher's Note