Comparative Analysis of Pre-trained Language Models for Patient Visit Recommendations

Pei-Ying Yang, Shin-En Peng, Shih-Chuan Chang, Yung-Chun Chang*
Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan.
{m946112007, m946112003, m946112004, changyc}@tmu.edu.tw

Abstract

As healthcare specialization advances, patients increasingly struggle to select the appropriate medical departments due to the intersectionality of their symptoms, which complicates the diagnosis and treatment process. To address this issue, the development of artificial intelligence has propelled digital triage systems to the forefront, becoming crucial tools in guiding patients effectively through this complex landscape. This study conducts a comprehensive comparative analysis of pre-trained language models (PLMs), including Bidirectional Encoder Representations from Transformers (BERT), RoBERTa, BlueBERT, Llama, and the Taide series tailored for Mandarin Chinese, on patient data from Taiwan's online medical consultation platform, Taiwan e-Clinic. The focus is on evaluating the efficacy of these models in recommending patient visits, specifically for Mandarin Chinese-speaking patients, to identify the most effective framework for clinical application. Our findings indicate significant differences in the models' abilities to recommend appropriate departments, which has important implications for enhancing digital healthcare services, especially in post-pandemic scenarios. The results demonstrate that PLMs have the potential to understand patient complaints and improve healthcare accessibility, enabling quicker medical recommendations.

1 Introduction

The increasing specialization in medical departments, while beneficial for targeted treatment, has made it more challenging for patients to select the appropriate department for their needs. This difficulty arises not only from the diversity of medical conditions but also from the overlapping nature of symptoms, making it hard for non-professionals to determine the right department. Additionally, patients often do not actively choose healthcare providers, partly because they feel the choice is not crucial, or their options are limited, and the available information is insufficient or unsuitable for decision-making (Chambers et al., 2019). The development of artificial intelligence (AI) has significantly changed the patient's pathway to seeking medical consultation. Digital triage systems have been used for many years. Previous studies have shown that digital triage systems are safe for patients and that algorithm-based triage methods are often more effective at avoiding risks than traditional healthcare professionals, thus gaining wide patient satisfaction (A. Victoor et al., 2012). With the proliferation of Online Symptom Checkers, more patients are using these applications for initial symptom diagnosis and finding appropriate care paths. For example, in Australia, data shows that among 36 AI-based Symptom Checkers, about 52% of cases are correctly ranked in the top three possible diagnoses, demonstrating that AI-based Symptom Checkers have a higher correct diagnosis rate than other types (M. G. Hill et al., 2020). Therefore, these applications have been widely adopted in the UK and Australia, helping users understand potential causes of symptoms and directing them to appropriate care facilities (Painter et al., 2022).

In Taiwan, despite the high density of healthcare facilities, many individuals remain uncertain about which department to visit during their initial consultation. The internet has become an integral part of daily life for Taiwanese citizens. According to 2023 statistics from the Taiwan Network Information Center, approximately 40% of the population uses the internet throughout the day, while only about 15% do not use the internet at all (TWNIC 2023). Consequently, when faced with

the challenge of selecting an appropriate medical department, many patients turn to online resources for assistance. 台灣e院 (Taiwan e-Clinic)[1], as a prominent online medical consultation service, offers a platform where patients can seek advice from professional doctors via the internet. This practice is increasingly common in modern healthcare services, yet it also presents challenges regarding the accuracy of information and users' understanding. Additionally, the time patients spend waiting for responses or searching through extensive articles to find answers to similar health questions can be considerable. If Artificial Intelligence (AI) methods could automatically provide accurate department recommendations based on patient inquiries, it would significantly enhance the efficiency and effectiveness of medical consultations, thus improving the overall healthcare delivery system.

In recent years, NLP technology has seen expansive application within the healthcare sector, substantially improving models' capacity to interpret unstructured healthcare data (Niu et al., 2024; Reeves et al., 2021; S. Datta et al., 2019). The introduction of PLMs marks a significant advancement in NLP capabilities, further enhancing their effectiveness. Empirical studies have demonstrated that utilizing PLMs to support healthcare professionals in clinical classification tasks can lead to exceptionally high levels of accuracy (Williams et al., 2024; Liu et al., 2024). This integration of advanced NLP tools not only optimizes clinical workflows but also contributes to more precise and efficient patient care outcomes.

In late 2019, the World Health Organization (WHO) issued an alert regarding an emerging virus characterized by cough and fever, later identified as SARS-CoV-2 in 2020 (Lu et al., 2019). This virus rapidly escalated into a global pandemic (Wang et al., 2020), fundamentally altering the landscape of medical consultations and triage processes until the WHO lifted its pandemic status at the end of 2022. Such unprecedented circumstances have led to fluctuating patterns in the utilization of online medical consultations before, during, and after the pandemic, presenting a unique opportunity for academic exploration (Bartczak et al., 2022).

In light of this, our study seeks to harness the capabilities of NLP technologies, particularly focusing on the utility of PLMs in enhancing patient visit recommendations. Despite the availability of NLP models supporting multiple languages, there is a notable scarcity of models tailored for Mandarin Chinese in the healthcare context.

Our research is poised to fill this gap by focusing on two primary objectives: (1). Assess and compare the effectiveness of various NLP techniques in delivering medical consultation advice, specifically utilizing Mandarin Chinese data and queries from the Taiwan e-Clinic platform. (2). Explore the shifts in online consultation patterns across three critical periods: before, during, and after the SARS-CoV-2 pandemic. By addressing these aims, our study not only highlights the potential of PLMs to revolutionize patient triage and consultation processes but also captures the dynamic changes in healthcare interactions in the face of a global health crisis.

This research promises to unveil insightful trends and contribute pioneering solutions to the field, aiming to captivate both readers and reviewers with its innovative approach and potential impact on future healthcare delivery.

2 Relative Work

Recent advancements in the use of Large Language Models (LLMs) as medical aids have demonstrated significant potential in improving healthcare outcomes. Panagoulias et al., (2024) showed that the capabilities of the multimodal LLM, GPT-4-Vision-Preview, which excelled in interpreting pathology-related questions and images, achieving an accuracy rate of 84%. Concurrently, Paslı et al., (2024) reported that ChatGPT performed on par with clinical triage teams in emergency department settings, showcasing its robustness in critical healthcare tasks. Further, Wang et al., (2024) developed the DRG-Llama model using Llama-7B, trained with MIMIC-IV data, which not only provided Diagnosis-Related Group (DRG) classifications with a top-1 accuracy of 54.6% but also achieved a top-3 accuracy of 86.5%, thereby surpassing earlier models like ClinicalBERT and CAML. These studies collectively underscore the growing trend of deploying LLMs as effective medical aids, approaching, and in some cases, matching human performance levels in healthcare applications.

[1] https://sp1.hso.mohw.gov.tw/doctor/

Additionally, our research addresses the inherent challenges associated with multilabel text classification, particularly the problem of data imbalance which is prevalent in such tasks. In response to these challenges, innovative methods have been proposed to enhance classification accuracy. For instance, De Angeli et al., (2021) introduced a Class-Specialized Ensemble approach utilizing TextCNN for the classification of tumor histology and subsites, which yielded a micro F_1-score of 0.79 on external datasets, outperforming traditional CNN and ensemble methods. Similarly, in 2020, Cai, Song, Liu, and Zhang developed the HBLA method leveraging BERT.

This approach integrates label semantics with fine-grained text information to achieve superior F1-scores compared to conventional CNN and Seq2Seq Attention models (L. Cai et al., 2020). These methodologies not only advance the field of multilabel classification but also contribute to the broader application of NLP technologies in handling complex medical datasets.

3 Materials and Method

3.1 Dataset

Due to the significant impact and changes caused by SARS-CoV-2 on community healthcare service and clinical medical departments and care in Taiwan from 2019 to 2023, we decided to focus our research on how the pandemic influenced the habits of the general public in seeking online consultations. Therefore, we collected data from the Taiwan e-Clinic, spanning five years from January 1, 2019, to December 31, 2023, to observe these data. This data will help us analyze and understand the changes in online consultation habits before, during, and after the pandemic.

Each record in our dataset includes several key pieces of information: page number, title, questioner's gender, questioner's age group, the question posed by the questioner, responding doctor's information, their reply, and the associated medical department. As per the Taiwan e-Clinic website, the medical department recommended by the responding doctor aligns with their specialty, which we utilized as the label for our dataset. After filtering out records with missing titles, genders, or ages, we successfully compiled a dataset comprising 55,742 records. Additionally, after observing the overall data, we found that the top ten most frequently consulted departments accounted for 77% of the total dataset. This indicates that most inquiries are concentrated on key departments. Therefore, to better focus and identify the relationship between public consultations and specific departments, we decided to filter the overall data to include only records related to these

Dept. # (%)	O&G 12628 (29.5)	URO 7537 (17.6)	OPH 3944 (9.2)	GI 3451 (8.1)	SURG 2941 (6.9)	MED 2864 (6.7)	ORTHO 2472 (5.8)	CV 2416 (5.6)	PSY 2317 (5.4)	DENT 2269 (5.3)	Overall 42839 (100)
Gender											
Female	10130 (80.2)	801 (10.6)	1911 (48.5)	1566 (45.4)	1395 (47.4)	1310 (45.7)	1235 (50.0)	1008 (41.7)	1169 (50.5)	1336 (58.9)	21861 (51.03)
Male	2498 (19.8)	6736 (89.4)	2033 (51.5)	1885 (54.6)	1546 (52.6)	1554 (54.3)	1237 (50.0)	1408 (58.3)	1148 (49.5)	933 (41.1)	20978 (48.97)
Age Group											
Minor	3605 (28.6)	1736 (23.0)	649 (16.5)	540 (15.7)	346 (11.8)	325 (11.4)	416 (16.8)	218 (9.0)	455 (19.6)	268 (11.8)	8558 (19.98)
Young Adult	8435 (66.8)	5149 (68.3)	2595 (65.8)	2310 (66.9)	2064 (70.2)	2050 (71.6)	1483 (60.0)	1467 (60.7)	1506 (65.0)	1639 (72.2)	28698 (66.99)
Mid-age Adult	584 (4.6)	634 (8.4)	686 (17.4)	589 (17.1)	525 (17.9)	480 (16.8)	551 (22.3)	704 (29.1)	355 (15.3)	359 (15.8)	5467 (12.76)
Elderly	4 (0.03)	18 (0.2)	14 (0.4)	12 (0.4)	6 (0.2)	9 (0.3)	22 (0.9)	27 (1.1)	1 (0.04)	3 (0.1)	116 (0.3)

Table 1: Dataset Descriptive Statistics.

top ten departments. This resulted in a dataset of 42,849 records, which we will use for our study.

The medical departments included in our study are as follows: Obstetrics and Gynecology (O&G), Urology (URO), Ophthalmology (OPH), Gastroenterology (GI), Surgery (SURG), Internal Medicine (MED), Orthopedics (ORTHO), Cardiology (CV), Psychiatry (PSY), and Dentistry (DENT). Statistical details for each department are meticulously presented in Table 1. These departments also showed significant gender differences among questioners due to their direct relation to physiological structures. The gender distribution in other departments was roughly equal. In terms of age distribution, the largest proportion of questioners in each department was in the Young Adult category (20-39 years old). According to the Taiwan Network Information Center's 2023 survey of internet users in Taiwan, the internet usage rate among those under 39 is

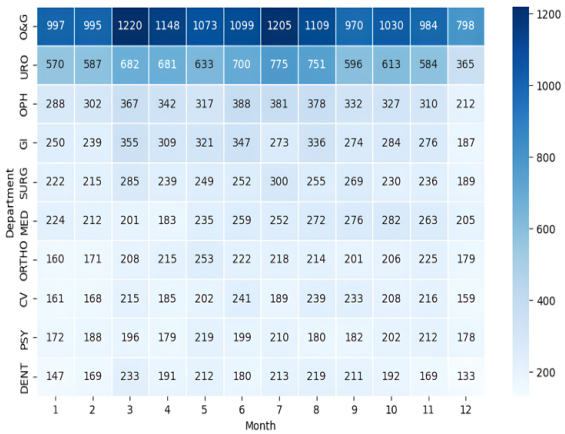


Figure 2: Number of questions asked by each department per month

98.40%, with a slight decline in usage as age increases, which may explain why this age group dominates the questioners in all departments.

Figure 1 shows the number of responses from specialist doctors in each department over different months across five years. Due to Taiwan's subtropical location, the summer months (July and August) are exceptionally hot, leading to a significant increase in infection rates, including urinary tract infections (UTIs) and infections related to the female reproductive system. For example, a 2022 global study indicated that UTIs incidence increases during adolescence and peaks at around 35 years of age for both men and women (Yang et al., 2024). Additionally, vaginal infections are a primary reason for women seeking medical treatment (Shroff et al., 2024).

By analyzing Figure 1 and Table 1, we can observe that the data trends in our study are consistent with global trends. This also explains why consultations for Obstetrics and Gynecology (O&G) and Urology (URO) peak in July and August. However, the difference compared to other months is not significantly pronounced. These findings highlight the seasonal impact on infection rates and underscore the importance of targeted healthcare resources during peak months to manage the increased demand for medical consultations related to these conditions. Consultations for other departments were evenly distributed across the months. However, considering that the primary objective of this study is to analyze the content of online queries, and the number of inquiries per month across different departments is evenly

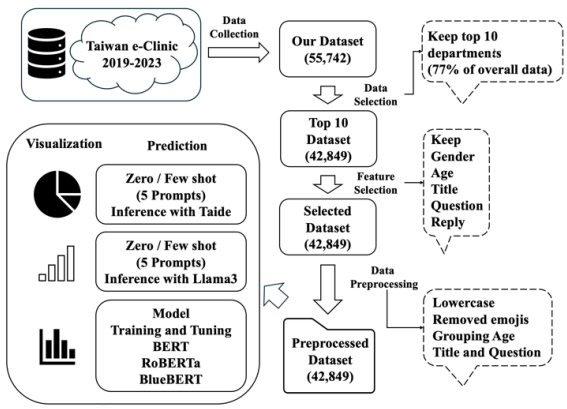


Figure 2: Data Processing and Model Training Workflow for Taiwan e-Clinic

distributed, we did not include time as an input variable. This represents a limitation of our study.

3.2 Pre-trained language models for Patient Visit Recommendations

As shown in Figure 2, our research methodology involves several critical stages. Initially, we extracted data from Taiwan e-Clinic covering the years 2019-2023. Subsequently, we eliminated records containing incomplete data and refined the dataset to encompass only the top ten departments by consultation volume. The retained data included essential elements such as the inquirer's demographics, query titles, substantive questions,

physicians' responses, and the corresponding medical specialties. Next, we cleaned the text data to prepare it for analysis. We then performed text analysis and department classification using different PLMs. This comprehensive approach ensures that we leverage the strengths of multiple models to gain insightful results from the text data. This study analyzes five PLMs and is divided into two parts to comprehensively evaluate their performance. The first part utilizes encoder-only bidirectional models from the Transformer architecture, specifically BERT (Vaswani et al., 2017) and its variant RoBERTa (Liu et al., 2019), as well as the BlueBERT (Peng et al., 2019), which is pre-trained on medical texts to align with our objectives. The second part extends the Transformer architecture to open-source LLM series. We selected two models: Llama3-8B(Llama3) (Touvron et al., 2023), the latest version released in 2024 by Meta[2], and Taide-7B[3] (Taide), a Taiwanese model based on the Llama2 architecture pre-trained on Mandarin Chinese texts, tailored for Mandarin Chinese-speaking regions.

For the text preprocessing, we converted all English text to lowercase and manually removed emojis to reduce model misinterpretation. Since our study aims to provide appropriate medical department recommendations based on patients' online complaints, we retained all original text from both the questioner and the doctor's reply, despite typographical errors and misspellings, to better reflect the real-world usage of online language. To better integrate age-related variables into our model, we converted age group data into categorical text and incorporated this into the textual description of each query. Based on the age categories provided by the Taiwan e-Clinic website, which are segmented into 10-year intervals ranging from 0 to 109 years, we categorized the age data as follows: ages 0-19 years are labeled as 'Minor', 20-39 years as 'Young Adult', 40-69 years as 'Midlife', and 70 years and above as 'Elder' (data distribution is illustrated in Table 1). We consider age information to be a crucial textual descriptor in our dataset. Consequently, BERT-based models utilize the '[SEP]' separator to clearly demarcate this demographic data from the patient's question, facilitating a more structured input that enhances the model's capacity to discern and interpret the underlying meanings of sentences more effectively.

On the other hand, LLMs directly process the integrated text descriptions that include age category information, allowing for a more holistic interpretation of the text without the need for explicit separators. This approach leverages the advanced capabilities of LLMs to understand and analyze complex and nuanced data representations inherently embedded in natural language queries.

To further refine the training of the LLMs and enhance their performance, we implemented two advanced prompt tuning methods: Zero-shot Learning (ZSL) and Few-shot learning (FSL). In the ZSL approach (Li Fei-Fei et al., 2006), the LLM is exposed to a single instance of a patient's chief complaint and is tasked with suggesting the appropriate medical department based solely on this input. This method tests the model's ability to generalize from minimal data. Conversely, FSL (Archit et al., 2022) involves the use of a set of five unique patient complaints and their corresponding departmental recommendations, randomly selected from our dataset. These examples serve as a pattern for the LLM, guiding it to infer and generate department suggestions by recognizing and learning from the provided examples. This strategy not only helps in enhancing the model's predictive accuracy but also aids in understanding the contextual nuances of different medical scenarios.

4 Experiments

4.1 Experimental Settings

In this study, we explored the efficacy of five PLMs in providing Patient Visit Recommendations, assessing their performance and applicability in clinical settings. For the models BERT, RoBERTa, and BlueBERT, the parameters we used were as follows: epoch: 10, batch size: 16, warmup steps: 200, weight decay: 0.1, learning rate: 3.5e-5.

To ensure robust model evaluation, we employed 5-fold cross-validation. We split the dataset into five parts, and in each iteration, one part is selected as the validation set, while the remaining four parts are used as the training set. This process is repeated five times until each part has been used as the validation set. Finally, the results are averaged to obtain the model performance score. For the LLMs, we obtained authorization from Meta and Taide, using their code published on Hugging Face as the basis for

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[3] https://huggingface.co/taide/TAIDE-LX-7B-Chat

PLMs	Model Performance (%)
	Accuracy / Precision / Recall / F_1-score
BERT	0.82 / 0.79 / 0.78 / 0.78
RoBERTa	**0.89 / 0.88 / 0.88 / 0.88**
BlueBERT	0.70 / 0.65 / 0.64 / 0.70
Llama3-ZSL	0.82 / 0.27 / 0.41 / 0.42
Llama3-FSL	0.81 / 0.43 / 0.56 / 0.51
Taide-ZSL	0.41 / 0.59 / 0.41 / 0.43
Taide-FSL	0.45 / 0.70 / 0.45 / 048

Table 2: Performance of compared models.

training. Additionally, we included our custom parameters: temperature: 0.1, top_k: 50, top_p: 0.95, to control the output of the generative models. We also used various metrics to evaluate model performance, including Accuracy, Precision, Recall, and F_1-score. Furthermore, we utilized a macro-averaging technique to aggregate scores across various categories, thereby obtaining an overall performance assessment of the models

4.2 Result and Discussion

As shown in Table 2, we evaluated the performance of different NLP models and LLMs. We also compared the effects of different prompt tuning methods on the performance of LLMs. It can be observed that the RoBERTa model achieved the best performance across all metrics, with an accuracy of 0.89. For imbalanced data, the F_1-score also showed excellent performance, indicating RoBERTa's superior ability to understand the complaints of Mandarin Chinese patients and accurately recommend the appropriate medical departments. BERT also demonstrated strong performance, with an accuracy of 0.82 and a F_1-score of 0.78, making it the second-best model among all tested. This shows that BERT can understand the semantics of the text and providing correct classifications. However, BlueBERT, which was pre-trained on a specialized medical corpus, showed moderate performance with an accuracy and F_1-score of only 0.70.

However, two LLMs produced less than satisfactory results. Both Llama3, trained in multiple languages, and Taide, designed specifically for Mandarin Chinese, demonstrated the limitations of generative language models in precise classification tasks. In the ZSL setting, Llama3 achieved an accuracy score of 0.81, but the other three metrics were only between 0.27 and 0.42. With the FSL setting, where sample text was added, the accuracy score remained unchanged, and the other three scores improved only slightly to between 0.43 and 0.51. The ZSL Taide model achieved a slightly better Precision score of 0.61 compared to Llama3 but lagged in the other three metrics. Similarly, in the FSL Taide model, except for achieving a Precision score of 0.70, the other three metrics only showed slight improvements.

By comparing two models with two different prompt tuning methods and their performance metrics, it is evident that FSL significantly enhances the model's text comprehension and classification performance, especially for LLMs with many class labels and some underrepresented categories. However, when comparing two different LLMs under the same prompt tuning method, it is clear that Llama3 achieves substantially higher Accuracy scores of 0.81 and 0.82 compared to Taide, and also surpasses Taide in terms of F_1-score. Despite this, Llama3's Precision is only 0.27 and 0.43, with Recall at 0.41 and 0.56 (compared to Taide's 0.41 and 0.45). This indicates that while Llama3 excels in overall prediction accuracy, it faces challenges in correctly predicting the specific department for a patient, whereas Taide demonstrates greater precision in predicting the correct department for patients.

These results indicate that while Llama3 and Taide can achieve reasonable accuracy in some situations, their ability to identify categories (recall) and avoid false positives (precision) is limited, especially when there is insufficient demonstration or only limited examples available. This suggests that generative language models may require more extensive training and fine-tuning, particularly with more targeted and high-quality data, to enhance their performance in specific classification tasks.

In this study, we found that encoder-only bidirectional models such as BERT and its variant RoBERTa achieved the best prediction scores compared to LLMs. RoBERTa, in particular, exhibited superior performance due to its extensive pre-training with more data and time, as well as the use of Dynamic Masking technology, which enhances semantic representation and generalization to new data (Liu et al., 2019). Despite BlueBERT focus on medical-related

corpora (Peng et al., 2019), its advantage in understanding medical language was diminished because patients do not typically use highly specialized medical terms when describing symptoms online. This allowed RoBERTa, with its better generalization capability, to outperform BlueBERT in performance scores. We attribute this performance gap to two main reasons. First, LLMs excel in generating coherent and contextually appropriate text, while the bidirectional attention mechanisms in models like RoBERTa enable them to capture complex relationships between words and sentences, leading to better text classification. Consequently, LLMs may struggle to match the performance of models like RoBERTa in tasks requiring precise classification and understanding. Additionally, our experimental results revealed that LLMs still suffer from hallucination problems (Liu et al., 2024; Gabrijela et al., 2024), such as generating non-existent or incorrect department names and providing excessive responses, which significantly reduces their classification performance.

We noticed that Taide, a model specifically trained for Mandarin Chinese, did not outperform Llama3. Although Taide builds upon the Llama2 architecture, it falls short in comparison to Llama3, particularly in terms of training parameters and duration of training. Llama3 is specifically designed for multi-language semantic understanding and generation, aiming for a broader linguistic scope. Consequently, despite the predominance of Mandarin Chinese in our input text, Llama3's more extensive training parameters enabled it to outperform Taide. This outcome suggests that LLMs tailored for a single language may still require a substantial increase in training parameters to achieve a competitive edge in text classification tasks for that specific language.

Additionally, both LLM models demonstrate a decline in reasoning ability and accuracy in generating correct text when training data is limited. This highlights the limitations of handling classification tasks without sufficient training data. LLMs rely on large amounts of training data to learn complex language expressions. When training data is relatively insufficient, it greatly impacts the diversity of the data, leading to the model's inability to make accurate classifications when faced with different types of inputs. Although providing example texts slightly improved the model's performance, it still could not overcome the lack of generalization ability. According to experiments by other researchers, regardless of the language model (Ding et al., 2023; Radiya-Dixit et al., 2020; Zhang et al., 2023), incorporating fine-tuning mechanisms and using techniques such as Low-Rank Adaptation (LoRA) can minimize training loss to the greatest extent and further improve the classification ability of language models. This will also become the focus of our future research.

Despite the current limitations, we believe that generative LLMs hold considerable potential in providing effective responses. Recent research initiatives have begun to explore the feasibility of training LLMs as clinical diagnostic assistants. Looking ahead, with further refinements and enhancements, these models could be more effectively integrated into clinical healthcare settings. Such advancements would enable the provision of more professional and reliable online consultation platforms for patients, significantly enhancing the quality of digital healthcare services.

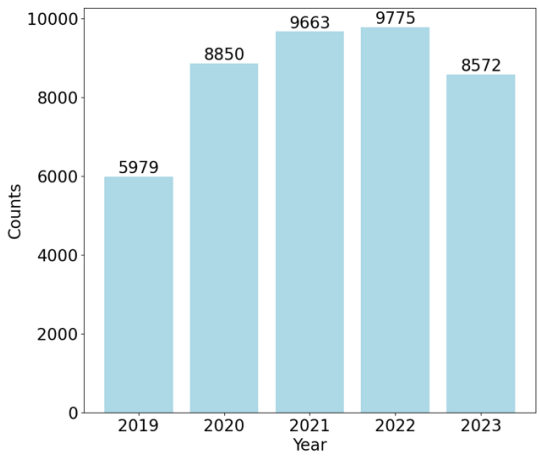


Figure 3: Number of questions over the years.

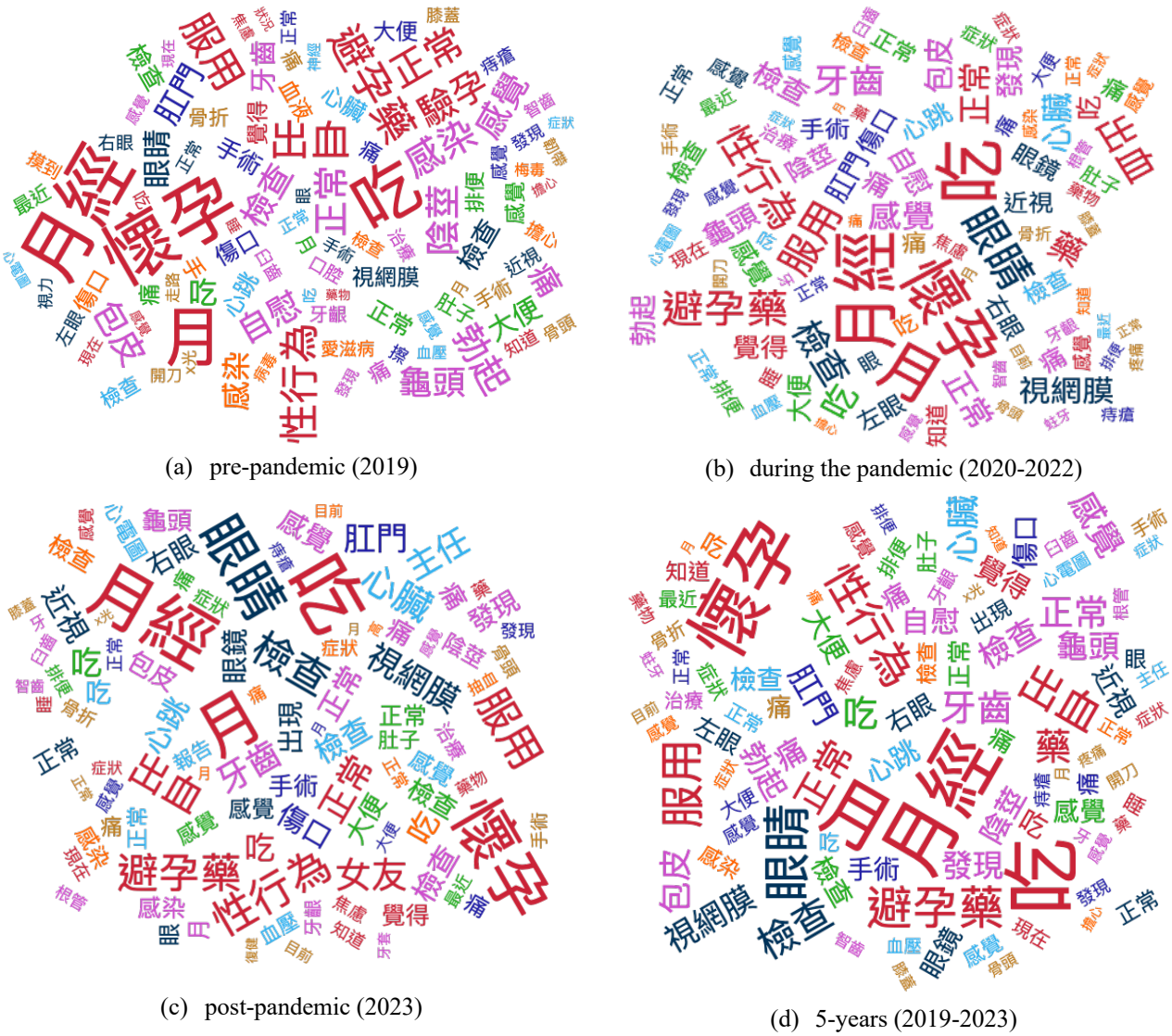
4.3 Keyword Trend Analysis

To further understand the differences in key terms used by questioners across different years, we used Python's wordcloud package to generate word clouds. During the creation of word clouds and frequency analysis for each medical department, we employed the MONPA Chinese segmentator[4] (Hsieh et al., 2017), which is specifically designed

[4] https://github.com/monpa-team/monpa

for Mandarin Chinese, to perform part-of-speech tagging and word segmentation. To facilitate the analysis of the impact of the pandemic on user queries, we divided the years into three segments: 2019 (pre-pandemic), 2020-2022 (during the pandemic), and 2023 (post-pandemic). We extracted words with parts of speech classified as verbs and nouns to better understand the symptoms emphasized by patients. Subsequently, we generated three word clouds for the three time periods, extracting the top 10 frequently used words for each of the 10 tags in each period. Each word cloud contained a total of 100 words, with 10 different colors representing the high-frequency words corresponding to each of the 10 labels. The number of inquiries over the years is shown in Figure 3, which provides a quantitative overview of the data we analyzed. This figure illustrates the volume of patient inquiries, allowing us to correlate the frequency of specific symptoms and concerns with the different phases of the pandemic. By examining both the word clouds and the inquiry volumes, we gained a comprehensive understanding of the changing landscape of patient concerns and the impact of the pandemic on online medical consultations.

According to Figure 3, the number of online consultations peaked in the mid to late pandemic period (2021-2022), highlighting the impact of the pandemic on patient inquiry and consultation methods. Government-enforced strict isolation measures and the promotion of telemedicine led patients to seek online consultations as their first option when experiencing physical discomfort or needing medical advice. This not only reduced the risk of infection from in-person visits but also alleviated the burden on healthcare facilities and conserved medical resources. However, as the



(a) pre-pandemic (2019)

(b) during the pandemic (2020-2022)

(c) post-pandemic (2023)

(d) 5-years (2019-2023)

Figure 4: Word Cloud Representation of Different Covid-19 Periods Online Medical Consultations

pandemic subsided and isolation measures were lifted in 2023, the number of online inquiries dropped to levels even lower than pre-pandemic (2019). We speculate that some patients may doubt the effectiveness of online consultations, particularly for issues requiring physical examinations. During the pandemic, online consultations were often a necessity rather than a preference, leading some individuals to revert to traditional in-person consultations post-pandemic.

Additionally, Figure 4 shows the trends and changes in patient concerns and common symptoms across different pandemic periods. We have placed the corresponding Mandarin Chinese and English translations in A1 of the Appendices. Overall, the keywords associated with each medical department clearly reflect the likely symptoms related to those departments. For instance, "O&G" frequently includes words like "month period" and "懷孕 (pregnancy)," while "Ophthalmology" features terms like "眼睛(eyes)" and "近視(myopia)." Across all departments, keywords often relate to patients' "feelings," the timing or duration of symptoms or pain, and post-treatment home care questions such as diet and medication. However, comparing keywords across different periods revealed minimal changes, possibly due to consistent user habits. From Table 1, patients who use online consultations often seek advice on sensitive issues, which did not change significantly during the pandemic, resulting in stable keyword patterns. However, we believe that these LLMs still have great potential in providing responses. Since the outbreak of the pandemic, the demand for online consultation systems has grown significantly, and other research has also begun to explore training LLMs as clinical diagnosis assistants (C. Wu et al., 2024). Tools based on LLM models, such as automated medical record keeping, personalized medicine, and health monitoring and alert systems (Sambare et al., 2024; Vicente et al., 2020; Yuan et al., 2024), can be adjusted in various ways to be more practically applied in clinical settings, offering patients more professional online consultation platforms. Therefore, our future research should focus more on effectively addressing the diversity and quantification of training data while developing more adaptable models to handle specific application scenarios, especially in the Mandarin Chinese healthcare domain. Through these efforts, we can maximize the potential of LLMs and promote their practical application across various industries, particularly in the healthcare field where Mandarin Chinese is used.

Finally, we used Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) to present the sentences and their attention weights learned by the RoBERTa model. In Figure 5, we randomly selected one of the posts as an example. We also placed the corresponding Mandarin Chinese and English translations in Section A2 of the Appendices. It can be observed that the model is capable of effectively identifying sentences or words related to O&G from the text and assigning appropriate weights. The blue color represents keywords related to O&G, while the teal color represents keywords unrelated to O&G. For instance, the model correctly identified key phrases related to O&G such as "懷孕的機會(chance of pregnancy)", "女友的排卵期 (girlfriend's ovulation period)", " (used a condom before sex)", and "性行為後的 20 分鐘內服用 Anlitin

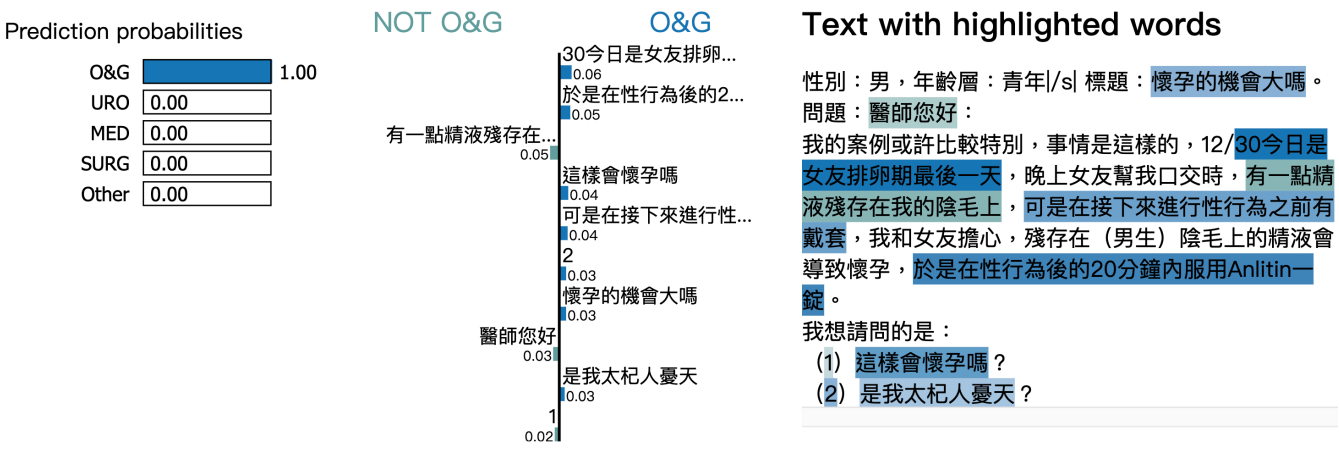


Figure 5: Presented the keywords and their weights using Local Interpretable Model-agnostic Explanations (LIME) of RoBERTa

(took Anlitin within 20 minutes of sex)”. Anlitin is an oral emergency contraceptive. On the other hand, the model also accurately identified irrelevant key sentences, such as “精液殘存在陰毛上 (semen remains in pubic hair)”. The questions containing the keyword “精液 (semen)” are mostly found in Uro; however, our model was able to classify them into categories unrelated to O&G. This demonstrates the model’s ability to provide correct medical department recommendations based on specific contexts or keywords.

5 Conclusion

This study has demonstrated that RoBERTa outperforms other language models in classifying medical departments from patient complaints. This superiority is attributed primarily to its extensive pre-training and dynamic masking, which collectively enhance its semantic understanding and generalization capabilities. Despite BlueBERT's specialization in medical terminology, its performance is compromised in the context of the more colloquial language prevalent in online consultations, rendering RoBERTa more effective. Furthermore, models like RoBERTa significantly surpass LLMs in classification tasks. LLMs, while adept at generating coherent text, tend to falter in precision-based classification, with hallucination issues further undermining their performance. Addressing these hallucination problems and enhancing the training of models specifically for Mandarin Chinese will be pivotal in our future research.

Additionally, our keyword trend analysis revealed that online medical consultations peaked during the mid to late stages of the pandemic (2021-2022), driven by isolation measures and the promotion of telemedicine. However, the number of consultations witnessed a decline post-pandemic (2023), falling below pre-pandemic levels, likely due to patient skepticism and discomfort with virtual interactions. This trend underscores a significant research opportunity to further enhance and optimize online consultation platforms, aiming to restore patient confidence and improve the overall efficacy of digital healthcare services.

Acknowledgments

This study was supported by the National Science and Technology Council of Taiwan under grants NSTC 113-2627-M-006-005-, NSTC 113-2221-E-038-019-MY3, and National Health Research Institutes under grants NHRI-13A1-PHCO-1823244.

References

Archit Parnami, & Lee, M. 2022. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. *ArXiv (Cornell University).* https://doi.org/10.48550/arxiv.2203.04291.

A. Victoor, P. Rademakers, J. Delnoij, and D. de Jong. 2012. Determinants of Patient Choice of Healthcare Providers: A Scoping Review. *BMC Health Services Research*, 12(1). https://doi.org/10.1186/1472-6963-12-272.

Bartczak, K. T., Milkowska-Dymanowska, J., Piotrowski, W. J., & Bialas, A. J. 2022. The utility of telemedicine in managing patients after COVID-19. *Scientific Reports*, *12*(1), 21392. https://doi.org/10.1038/s41598-022-25348-2.

Chambers, D., Cantrell, A. J., Johnson, M., Preston, L., Baxter, S. K., Booth, A., & Turner, J. 2019. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open*, 9(8), e027743. https://doi.org/10.1136/bmjopen-2018-027743.

C. Wu, Z. Lin, W. Fang, and Y. Huang. 2024. A Medical Diagnostic Assistant Based on LLM. *Communications in computer and information science,* 135–147. https://doi.org/10.1007/978-981-97-1717-0_12.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence.* https://doi.org/10.1038/s42256-023-00626-4.

Gabrijela Perković, Antun Drobnjak, & Ivica Botički. 2024. *Hallucinations in LLMs: Understanding and Addressing Challenges*. https://doi.org/10.1109/mipro60963.2024.10569238.

Hsieh, Y.-L., Chang, Y.-C., Huang, Y.-J., Yeh, S.-H., Chen, C.-H., & Hsu, W.-L. 2017. MONPA: Multi-objective Named-entity and Part-of-speech Annotator for Chinese using Recurrent Neural Network. *ACL Anthology*, 80–85. https://aclanthology.org/I17-2014/.

K. De Angeli, S. Gao, I. Danciu, E. B. Durbin, X. C. Wu, A. Stroup, J. Doherty, S. Schwartz, C. Wiggins, M. Damesyn, L. Coyle, L. Penberthy, G. D. Tourassi, and H. J. Yoon. 2022. Class imbalance in out-of-

distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types. *Journal of Biomedical Informatics,* 125, 103957. https://doi: 10.1016/j.jbi.2021.103957.

L. Cai, Y. Song, T. Liu, and K. Zhang. 2020. A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification. *IEEE Access,* 8, 152183-152192. https://doi: 10.1109/ACCESS.2020.3017382.

Li Fei-Fei, Fergus, R., & Perona, P. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611. https://doi.org/10.1109/tpami.2006.79.

Liu, D., Han, Y., Wang, X., Tan, X., Liu, D., Qian, G., Li, K., Pu, D., & Yin, R. 2024, April 27. *Evaluating the Application of ChatGPT in Outpatient Triage Guidance: A Comparative Study*. ArXiv.org. https://doi.org/10.48550/arXiv.2405.00728.

Liu, F., Liu, Y., Shi, L., Huang, H., Wang, R., Yang, Z., & Zhang, L. 2024, April 1. *Exploring and Evaluating Hallucinations in LLM-Powered Code Generation*. ArXiv.org. https://doi.org/10.48550/arXiv.2404.00971.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. 2019, July 26. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. ArXiv.org. https://arxiv.org/abs/1907.11692.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., & Chen, J. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224). https://doi.org/10.1016/s0140-6736(20)30251-8.

M. G. Hill, J. S. McCabe, and A. B. Bonner. 2020. The Quality of Diagnosis and Triage Advice Provided by Free Online Symptom Checkers and Apps in Australia. *Medical Journal of Australia*, 212(11). https://doi.org/10.5694/mja2.50600.

Niu, H., Omitaomu, O. A., Langston, M. A., Olama, M., Ozmen, O., Klasky, H. B., Laurio, A., Ward, M., and Nebeker, J. 2024. EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records. *Journal of Biomedical Informatics,* 150, 104605. https://doi.org/10.1016/j.jbi.2024.104605.

Painter, A., Hayhoe, B., Riboli-Sasco, E., & El-Osta, A. 2022. Online Symptom Checkers: Recommendations for a vignette-based clinical evaluation standard (Preprint). *Journal of Medical Internet Research*. https://doi.org/10.2196/37408.

Panagoulias, Dimitrios P, Virvou, M., & Tsihrintzis, G. A. 2024. Evaluating LLM -- Generated Multimodal Diagnosis from Medical Images and Symptom Analysis. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.01730.

Paslı, S., Şahin, A. S., Beşer, M. F., Topçuoğlu, H., Yadigaroğlu, M., & İmamoğlu, M. 2024. Assessing the precision of artificial intelligence in ED triage decisions: Insights from a study with ChatGPT. *The American journal of emergency medicine,* 78, 170–175. https://doi.org/10.1016/j.ajem.2024.01.037.

Peng, Y., Yan, S., & Lu, Z. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *ArXiv:1906.05474 [Cs]*. https://arxiv.org/abs/1906.05474.

Radiya-Dixit, E., & Wang, X. 2020, June 3. *How fine can fine-tuning be? Learning efficient language models*. Proceedings.mlr.press; PMLR. https://proceedings.mlr.press/v108/radiya-dixit20a.html.

Reeves, R. M., Christensen, L., Brown, J. R., Conway, M., Levis, M., Gobbel, G. T., Shah, R. U., Goodrich, C., Ricket, I., Minter, F., Bohm, A., Bray, B. E., Matheny, M. E., and Chapman, W. 2021. Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. *Journal of Biomedical Informatics*, 120, 103851. https://doi.org/10.1016/j.jbi.2021.103851.

Ribeiro, M. T., Singh, S., & Guestrin, C. 2016, February 16. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. ArXiv.org. https://arxiv.org/abs/1602.04938.

Sambare, D.G., Bhute, H.A., Banait, D.S., Bobhate, G.Y., Amir, D.A., & Bhattacharya, S. 2024. Autonomous Healthcare Systems: Deep Learning-Based IoT Solutions for Continuous Monitoring and Adaptive Treatment. *Journal of Electrical Systems.* 20(1). https://doi.org/10.52783/jes.780.

S. Datta, E. V. Bernstam, and K. Roberts. 2019. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *Journal of Biomedical Informatics,* 100, 103301. https://doi.org/10.1016/j.jbi.2019.103301.

Shroff, S. 2023. Infectious Vaginitis, Cervicitis, and Pelvic Inflammatory Disease. *Medical Clinics*, *107*(2), 299–315. https://doi.org/10.1016/j.mcna.2022.10.009.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv:2302.13971 [Cs]*. https://arxiv.org/abs/2302.13971.

TWNIC. (2023). *2023 台灣網路報告*. Twnic.tw. https://report.twnic.tw/2023/.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. 2017, June 12. *Attention Is All You Need*. ArXiv.org. https://arxiv.org/abs/1706.03762.

Vicente, A.M., Ballensiefen, W. & Jönsson, JI. 2020. How personalised medicine will transform healthcare by 2030: the ICPerMed vision. *J Transl Med,* 18, 180. https://doi.org/10.1186/s12967-020-02316-w.

Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. 2020. A novel coronavirus outbreak of global health concern. *The Lancet*, *395*(10223), 470–473. https://doi.org/10.1016/s0140-6736(20)30185-9.

Wang, H., Gao, C., Dantona, C., Hull, B., & Sun, J. 2024. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *Npj Digital Medicine,* 7(1), 1–9. https://doi.org/10.1038/s41746-023-00989-3.

Williams, C. Y. K., Zack, T., Miao, B. Y., Sushil, M., Wang, M., Kornblith, A. E., & Butte, A. J. 2024. Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department. *JAMA Network Open*, 7(5), e248895. https://doi.org/10.1001/jamanetworkopen.2024.8895.

Yang, X., Chen, H., Zheng, Y., Qu, S., Wang, H., & Yi, F. 2022. Disease burden and long-term trends of urinary tract infections: A worldwide report. *Frontiers in Public Health*, *10*(888205). https://doi.org/10.3389/fpubh.2022.888205.

Yuan, D., Rastogi, E., Naik, G., Rajagopal, S. P., Goyal, S., Zhao, F., Chintagunta, B., & Ward, J. 2024, June 1. *A Continued Pretrained LLM Approach for Automatic Medical Note Generation* (K. Duh, H. Gomez, & S. Bethard, Eds.). ACLWeb; Association for Computational Linguistics. https://aclanthology.org/2024.naacl-short.47/.

Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., & Qiao, Y. 2023. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *ArXiv.org*. https://doi.org/10.48550/arXiv.2303.16199.

A Appendices

A.1 Mandarin-English Keyword Comparison

In Figure 4, we display the relevant Mandarin Chinese keywords for each medical department during different phases of the pandemic. After excluding the keywords that appeared repeatedly over the five years, we provide the Mandarin Chinese keywords along with their corresponding English translations, as shown in Table 3. It is noteworthy that in Mandarin Chinese, synonymous words can be expressed using different characters or phrases. Additionally, the same word may represent different parts of speech with the same meaning, such as "感覺" and "覺得," both meaning "feel," while "感覺" can also be used as the noun "feeling."

A.2 LIME of RoBERTa Tanslation

In Figure 5, we present the Attention weight map learned from the RoBERTa model training, visualized using LIME. We have also provided the Chinese content along with the corresponding English translation, as shown in Table 4.

Mandarin	English	Mandarin	English	Mandarin	English	Mandarin	English
骨頭	Bone	龜頭	Glans	自慰	Masturbate	開刀	Operate
發現	Find	勃起	Erection	右眼	Right eye	排便	Defecation
心跳	Pulse	月	Month	智齒	Wisdom tooth	眼睛	eyes
出現	Appear	視網膜	Retina	擔心	Worry	痔瘡	Hemorrhoids
檢查	Examine	覺得	Feel	出血	Bleeding	睡	Sleep
牙	Tooth	陰莖	Penis	心電圖	Electrocardiogram	蛀牙	Cavity
包皮	Foreskin	最近	Recently	白齒	Molar tooth	月經	Menstruation
骨折	Fracture	症狀	Symptom	焦慮	Anxiety	現在	Now
感覺	Feeling	藥物	Medicine	手術	Operation	避孕藥	Birth-control pills
痛	Pain	治療	Treat	根管	Root Canal	主任	Director
目前	Currently	牙齒	Tooth	大便	Stool	懷孕	Pregnant
眼鏡	Glasses	牙齦	Gum	感染	Infect	疼痛	Pain
眼	eye	心臟	Heart	膝蓋	Knee	服用	Take
左眼	Left eye	藥	Medicine	近視	Myopia	傷口	Wound
x 光	X-ray	性行為	Sex	血壓	Blood Pressure	正常	Normal
肚子	Stomach	吃	Eat	知道	Know	肛門	Anus

Table 3: Keyword Comparison Table

Mandarin	**Translation**
性別：男，年齡層：青年 標題：懷孕的機會大嗎。 問題：醫師您好： 我的案例或許比較特別，事情是這樣的，12/30 今天是女友排卵期最後一天，晚上女友幫我口交時，有一點精液殘存在我的陰毛上，可是在接下來進行性行為之前有戴套，我和女友擔心，殘存在（男生）陰毛上的精液會導致懷孕。於是在性行為後的 20 分鐘內服用 Anlitin一錠。 我想請問的是： (1) 這樣會懷孕嗎？ (2) 是我太杞人憂天？	Gender: Male Age group: Youth Title: Is the chance of pregnancy high? Question: Hello doctor, My case might be a bit unusual. Here's what happened: Today, 12/30, is the last day of my girlfriend's ovulation period. In the evening, my girlfriend performed oral sex on me, and some semen was left on my pubic hair. However, before we proceeded with intercourse, I used protection. My girlfriend and I are worried that the semen left on my (male) pubic hair could lead to pregnancy. So, we took an Anlitin pill within 20 minutes after intercourse. I would like to ask: (1) Is there a chance of pregnancy? (2) Am I overthinking this?

Table 4: LIME of RoBERTa Translation