Evaluating the Pre-Consultation Ability of LLMs using Diagnostic Guideliness

Jean Seo[1], **Gibaeg Kim**[1], **Kihun Shin**[3], **Seungseop Lim**[1], **Hyunkyung Lee**[1], **Wooseok Han**[1], **Jongwon Lee**[4], **Eunho Yang**[1,2]
[1]AITRICS [2]KAIST
[3]Severance Hospital, Yonsei University
[4]College of Medicine, The Catholic University of Korea
{jeanseo}@aitrics.com

Abstract

We introduce **EPAG**, a benchmark dataset and framework designed for **E**valuating the **P**re-consultation **A**bility of LLMs using diagnostic **G**uidelines. LLMs are evaluated directly through HPI-diagnostic guideline comparison and indirectly through disease diagnosis. In our experiments, we observe that small open-source models fine-tuned with a well-curated, task-specific dataset can outperform frontier LLMs in pre-consultation. Additionally, we find that increased amount of HPI (History of Present Illness) does not necessarily lead to improved diagnostic performance. Further experiments reveal that the language of pre-consultation influences the characteristics of the dialogue. By open-sourcing our dataset and evaluation pipeline on https://github.com/seemdog/EPAG, we aim to contribute to the evaluation and further development of LLM applications in real-world clinical settings.

1 Introduction

Large Language Models (LLMs) are increasingly integrated into clinical applications, transforming healthcare industry by automating various tasks (Yang et al., 2023a; Zhou et al., 2024; Thirunavukarasu et al., 2023; Wang et al., 2025). One example is pre-consultation, where LLMs assist history-taking (Wang et al., 2024; Yang et al., 2023b) and decision-making (SAMIEE; Li et al., 2024). However, it is crucial to acknowledge the significant risks involved. Erroneous outputs can result in severe adverse consequences such as mistreatment or incorrect drug prescription, highlighting the necessity of rigorous evaluations (Kim et al., 2025; Ullah et al., 2024).

We propose **EPAG** (**E**valuating the **P**re-consultation **A**bility of LLMs using diagnostic **G**uidelines), a benchmark dataset and evaluation pipeline specifically designed for pre-consultation. Given basic patient information, such as age, sex, and chief complaints, pre-consultation models ask questions to elicit symptoms related to potential diagnoses. EPAG benchmark dataset comprises 520 patient profiles, spanning 26 diseases, 10 ICD-11 chapters, 10 primary specialties, and 22 secondary specialties, along with pre-defined diagnostic guidelines. In EPAG, the pre-consultation dialogue is evaluated through two tasks: (1) HPI-Diagnostic Guideline Comparison, and (2) Disease Diagnosis. In our experiments, eleven LLMs are evaluated across various numbers of dialogue turns.

The main contributions of our work are:

- Developing a systematic framework and constructing a high-quality dataset for evaluating the clinical pre-consultation ability of LLMs.
- Open-sourcing the dataset and pipeline.
- Implementing targeted experiments and sharing the results with in-depth analysis.

2 Related Work

2.1 Medical LLMs in Clinical Applications

Existing clinical chatbot applications include HuatuoGPT (Zhang et al., 2023), ChatDoctor (Li et al., 2023), MedChatZH (Tan et al., 2024), MedAide (Basit et al., 2024), and MILD Bot (Kim et al., 2024). Other medical LLM applications not limited to chatbots are Kumichev et al. (2024); Zhang et al. (2024); Wiest et al. (2024); Ghosh et al. (2024); Waisberg et al. (2024). LLMs have demonstrated diagnostic accuracy comparable to that of physicians in certain contexts (Qian et al., 2021), with existing works primarily focusing on final diagnostic outcomes (McDuff et al., 2023; Singhal et al., 2023; Tu et al., 2024). However, research on patient information collection during LLM pre-consultation remains limited. To address this, we propose a fine-grained framework that evaluates LLM pre-consultation capabilities.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 5: Industry Track, pages 78–94

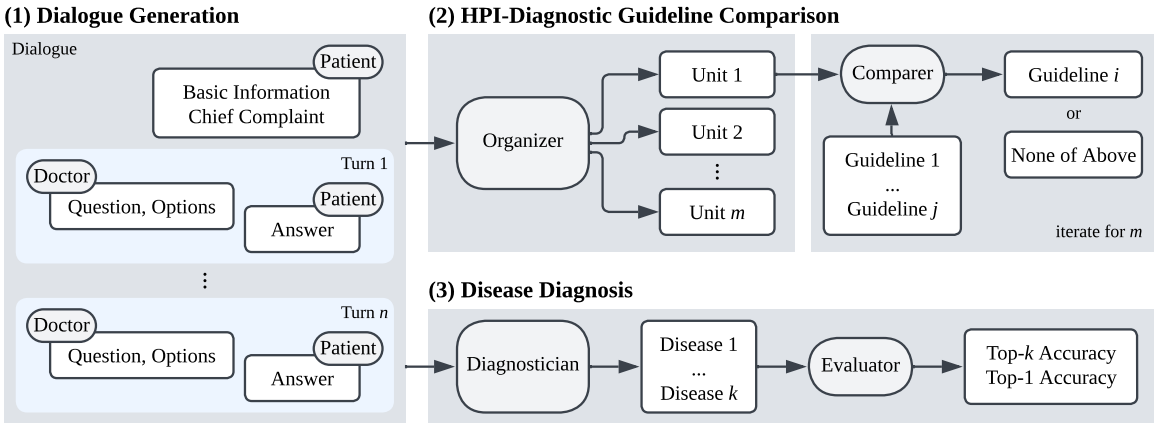

Figure 1: EPAG pipeline. **(1) Dialogue Generation**: The patient-agent acts as a patient given a specific profile, while the doctor-agent conducts a pre-consultation using only the basic information and chief complaint. After n turns, the doctor-agent is assessed through two tasks: **(2) HPI-Diagnostic Guideline Comparison**, where the organizer model extracts HPI units and the comparer model determines which of the diagnostic guidelines is most relevant, and **(3) Disease Diagnosis**, where the dialogue is given to a separate diagnostician-agent for diagnosis.

2.2 Evaluation of Medical LLMs

Multiple-choice QA is widely used for medical evaluation, as demonstrated by Med-HALT (Pal et al., 2023), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and KoreMedMCQA (Kweon et al., 2024). However, it is insufficient for assessing real-world clinical conversational abilities (Bedi et al., 2024; Chen et al., 2024). More sophisticated evaluation frameworks in the clinical domain have been proposed, including MEDIC (Kanithi et al., 2024), LLM-Mini-CEX (Shi et al., 2023), CRAFT-MD (Johri et al., 2025). Other evaluation benchmarks regarding disease diagnosis include works by Hou et al. (2024), Zhu et al. (2025), Bhasuran et al. (2025), Delaunay and Cusido (2024), Sarvari and Al-Fagih (2025), Reese et al. (2025), Gaber et al. (2025). While Winston et al. and Fast et al. propose evaluation pipelines for pre-consultation, their dataset coverage is limited and peripheral.

3 EPAG Benchmark

We assess pre-consultation models designed to collect as much relevant information as possible from the patient, including symptoms, family history, and other factors, referred to as the History of Present Illness (HPI). This section covers the tasks, dataset construction process, and evaluation pipeline of EPAG.

3.1 Evaluation Tasks

As Figure 1 demonstrates, we propose a two-tiered evaluation framework based on the collected HPI.

3.1.1 HPI-Diagnostic Guideline Comparison

For direct evaluation, we focus on how effectively the models capture information necessary for accurate disease identification. The evaluation process involves pre-consultation simulation with a patient-agent exhibiting symptoms of a specific disease and a doctor-agent, which is the subject of evaluation. During this interaction, the doctor-agent asks questions and provides multiple options for the patient-agent to choose from. The HPI collected is then compared against a set of diagnostic guidelines for the specific disease. The diagnostic guidelines represent a collection of essential information for diagnosing a particular disease, curated by human clinicians from trusted sources with further details in Section 3.2.1.

3.1.2 Disease Diagnosis

For indirect evaluation, we assess how well the collected HPI supports accurate diagnoses when provided to a separate diagnostic model. While this is not a direct evaluation of the HPI extracted by LLMs, it is a crucial assessment as one of the eventual goals of LLM pre-consultation is to assist in correct diagnosis and treatment.

3.2 Dataset

Figure 2 shows the dataset construction process.

3.2.1 Diagnostic Guideline

To evaluate whether each dialogue turn elicits meaningful patient information for diagnosis, we construct a gold-label diagnostic guideline dataset.

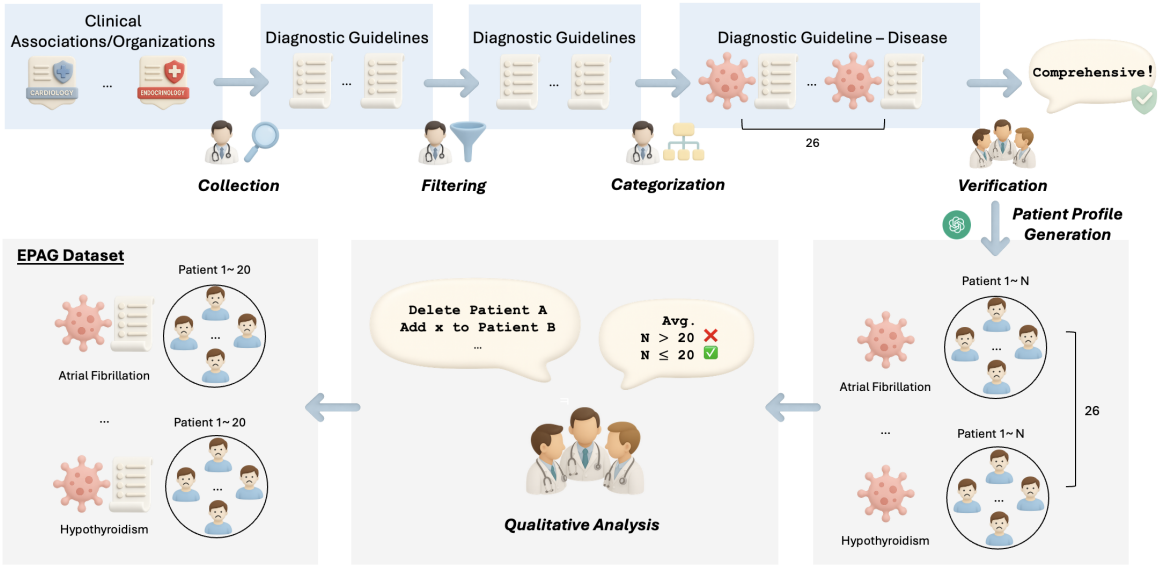

Figure 2: EPAG benchmark dataset construction process. Expert clinicians collect all possible diagnostic guidelines of diseases from credible clinical sources. They then filter diseases based on whether they can be reasonably diagnosed through consultation alone and sufficiently common to ensure unbiased evaluation. Next, clinicians verify that the disease list is comprehensive enough to serve as a generalizable evaluation set. Using the finalized list, synthetic patient profiles are generated and finalized through qualitative analysis by clinicians.

The following steps are implemented by professional clinicians based on credible clinical associations and organizations in Appendix A: (1) collect diagnostic guidelines with explicit references; (2-1) filter diseases that are diagnosable through consultation alone, without reliance on physical exams, X-ray or MRI; (2-2) exclude diseases that are too rare. As exemplified in Appendix B, each diagnostic guideline specifies key symptoms, ancillary symptoms, family history, and other relevant risk factors. Each feature is assigned a weight of either *high* or *medium*.

3.2.2 Disease

As our primary goal is to evaluate language models rather than multi-modal models, we focus on diseases that can be differentiated without reliance on other examination results. Through extensive discussions with clinicians, we identify 26 such diseases spanning 10 primary specialties and 22 secondary specialties. To ensure that the selection of 26 diseases provides sufficient clinical generalizability, clinicians classify them according to the International Classification of Diseases, 11th Revision (ICD-11) [1]. This categorization confirms that the included diseases span a broad range of conditions across 10 ICD-11 chapters, as shown in Table 3, indicating that the dataset covers a clinically diverse and representative scope of diseases that can be reasonably differentiated through history-taking. Each disease is systematically assigned to both primary and secondary specialties following established clinical criteria in Appendix C, reflecting the multidisciplinary nature of real-world patient care.

3.2.3 Patient Profile

We generate diverse patient profiles using OpenAI o3-mini [2]. Expert clinicians then conduct a qualitative review to ensure (i) sufficient diversity across profiles and (ii) adequate clinical detail to support realistic patient–doctor interactions. To minimize bias in the synthetic dataset, we retain 20 profiles per disease, yielding a total of 520 profiles. Each profile contains demographic and clinical information such as age, sex, height, weight, and relevant medical history, representing realistic patient cases. Each patient profile is used to assign a role to the patient-agent, which then interacts with the doctor-agent, simulating realistic scenarios. A sample profile and diversity of patient group can be found in Table 4 and Figure 6 respectively.

3.3 Evaluation Framework

Supposing pre-consultation models that ask questions and provide options to choose from, [Question, Options, Answer] triplets are utilized through-

[1] https://icd.who.int/en

[2] https://openai.com/

Model	HPI-Diagnostic Guideline Comparison Score		Disease Diagnosis Accuracy	
	Not Weighted	Weighted	Top-1	Top-*k*
Human Expert	4.35	7.29	68.24	80.65
LLMs				
GPT-4.1	**4.82**	**8.12**	**74.56**	**83.81**
GPT-4.1-mini	4.46	7.64	69.15	81.36
GPT-4o	4.39	7.59	69.23	81.35
GPT-4o-mini	4.46	7.75	64.62	79.62
Claude-3.7-Sonnet	4.59	**8.12**	69.23	82.31
Claude-3.5-Sonnet	4.62	8.05	72.69	81.35
Claude-3.5-Haiku	4.58	7.84	65.38	80.77
Phi-3.5-mini	3.91	6.88	61.82	78.84
Llama-3.2-3B	3.87	6.8	58.14	72.09
Qwen2.5-7B	3.74	6.51	58.46	76.54
Medgemma-4B 🩺	4.19	7.22	65.93	82.31

Table 1: HPI–diagnostic guideline comparison scores and disease diagnosis accuracies for eleven models, alongside a human expert baseline, over five-turn dialogues. Results exceeding the human baseline are shaded in blue, and those below in red. Stethoscope (🩺) denotes the medically fine-tuned model.

out evaluation.

3.3.1 HPI-Diagnostic Guideline Comparison Score

(1) Response Generation

The doctor-agent is provided with the chief complaint and basic information, including age, sex, height, weight, then generates questions and options. The patient-agent is provided with the full patient profile, and asked to select the appropriate option with the prompt in Table 5. This process is iterated for n times.

(2) Organization

After n turns of pre-consultation, the [Question, Options, Answer] triplets are organized into individual units, each representing a single piece of clinical information, by an organizer model, using the prompt in Table 6. This step is crucial because, in the next phase, we compare each unit against pre-defined diagnostic guidelines to assess whether it matches any. Since a single [Question, Options, Answer] triplet may contain multiple pieces of information, separating them into individual units ensures more accurate comparison. For example:

> **Question:** Are there any other symptoms that occur with chest tightness?
> **Options:** Shortness of breath or difficulty breathing, A feeling of a racing heart, Cold sweats, Dizziness, Vomiting or nausea
> **Answer:** Shortness of breath or difficulty breathing

The number of organized units should be five, not one: (1) Patient has shortness of breath or difficulty breathing, (2) Patient does not have a racing heart, (3) nor cold sweats, (4) nor dizziness, (5) nor vomiting or nausea. In differential diagnosis, the absence of symptoms is as significant as their presence, so the unselected options are treated as separate units. Additionally, to avoid duplicating scores for redundant questions, we deduplicate the information extracted during the organization step. An example is provided in Appendix D.

(3) Comparison

Next, we use a comparer model with the prompt in Table 7 to match each unit with the most relevant diagnostic guideline. If a unit does not match any of the guidelines, the comparer model is instructed to respond with "None of Above." As illustrated in Figure 1, for each of the m units, the comparer performs the comparison process.

(4) Score Calculation

The final score for the pre-consultation dialogue is calculated by awarding 1 point if the unit corresponds to a guideline and 0 point for "None of Above." Since some diagnostic guidelines may be more influential in diagnosing or ruling out certain diseases than others, we also compute a weighted score. Human expert clinicians assign each guideline a significance level of medium or high, as shown in Appendix B. A unit corresponding to a medium-significance guideline earns 1 point, while a high-significance guideline earns 2 points. Both versions of the score are calculated for each patient and averaged across 520 datasets to determine the final score for each doctor-agent.

To verify the reliability of our evaluation pipeline, we conduct a human comparison. For each disease, one is randomly sampled for each disease and evaluated by a human clinician using the same pipeline. After performing an F-test ($p > 0.05$) to ensure equal variances, a T-test confirms that the two sets of scores are statistically similar ($p > 0.05$).

3.3.2 Disease Diagnosis Accuracy

For indirect evaluation of the pre-consultation dialogue, we use an independent diagnostician-agent with the prompt in Table 8. To account for multiple names for the same disease, we consider the prediction correct if the model identifies a parent or child concept of the gold label disease. We employ an evaluator model using the prompt in Table 9 to determine if the predicted disease matches the gold label.

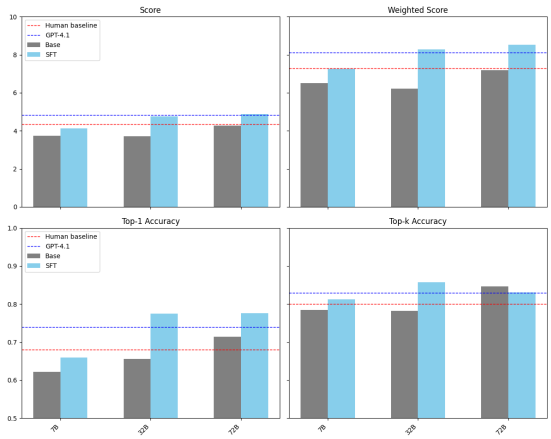

Figure 3: Performance of Qwen-2.5 models (7B, 32B, 72B) before (grey) and after (blue) SFT. Red horizontal line marks human clinician performance, and blue marks GPT-4.1 performance—the strongest model.

4 Experiments

We evaluate eleven models as the doctor-agent, including four from OpenAI [3], three from Anthropic [4], and four open-source LLMs, one of which is medically fine-tuned, and compare their performance to a human baseline. For the human baseline, human clinicians go through the same pre-consultation process as the doctor-agents, while the rest of the pipeline remains unchanged. Figure 8 shows the user interface used by human clinicians. The resulting pre-consultation dialogues are then evaluated using our proposed pipeline. In this experiment, all other components in the pipeline use GPT-4o-mini, with distinct prompts assigned to each role (patient, organizer, comparer, diagnostician, evaluator). To ensure reproducibility, we fix the random seed and set the temperature of each agent to 0. The only variable is the doctor-agent model.

5 Result and Analysis

As shown in Table 1, in five turn dialogues, GPT-4.1 attains the highest performance across all metrics, tying with Claude-3.7-Sonnet on the weighted HPI-diagnostic guideline comparison score. Qwen2.5-7B and Llama-3.2-3B perform worst overall. The human baseline places above all open-source models, but below every proprietary LLM. Contrary to our intuition that medical fine-tuning would elicit decent performance, Medgemma-4B underperforms the human baseline. A plausible explanation is that Medgemma-4B is fine-tuned primarily on existing medical tasks, which may have weakened its instruction-following ability on unseen tasks like pre-consultation. We conduct a series of additional experiments, providing several important takeaways.

Model size does not guarantee performance. Larger or more expensive models are expected to outperform their smaller counterparts across most tasks. This holds true in the GPT-4.1 family, where GPT-4.1 exceeds GPT-4.1-mini on all four metrics. However, GPT-4o-mini outperforms GPT-4o on HPI-diagnostic guideline comparison score. Moreover, Claude-3.5-Sonnet outperforms Claude-3.7-Sonnet, the most expensive model, on the unweighted score and Top-1 accuracy. Although technical reports often emphasize gains from increased scale, our findings suggest that this relationship weakens for clinical pre-consultation.

Task-specific Fine-tuning matters. If model size does not guarantee pre-consultation ability, what does? We hypothesize that once a model's medical knowledge surpasses a certain threshold, its performance depends primarily on how effectively it can leverage that knowledge to generate appropriate questions. This interpretation is supported by the underperformance of Medgemma-4B, despite its presumed advantage in medical knowledge. To test this, we construct a 3k pre-consultation dialogue dataset independent from EPAG—generated by LLMs and rigorously reviewed by clinical experts—and fine-tune Qwen-2.5 models (7B, 32B, 72B) using LoRA (Hu et al., 2021). Figure 3 compares each model's performance before and after supervised fine-tuning. Consistent with our earlier analysis, the baseline models do not exhibit strict monotonic gains with size: while Top-1 accuracy improves as model size increases, the other three metrics rank as 32B < 7B < 72B. After SFT, all models show marked improvements across most metrics, with 32B benefiting the most. Although the base models fall below both the human expert and GPT-4.1, fine-tuned models often exceed the human expert—and notably, 7B and 32B match or even surpass GPT-4.1. Qwen2.5-72B's slight decline in Top-k accuracy after fine-tuning possibly suggests underfitting, likely because our 3k-dialogue dataset is insufficient to fully optimize the largest model but more than adequate for the smallest model, making 32B the optimal size for this dataset. Overall, the peaking

[3] https://openai.com/
[4] https://www.anthropic.com/

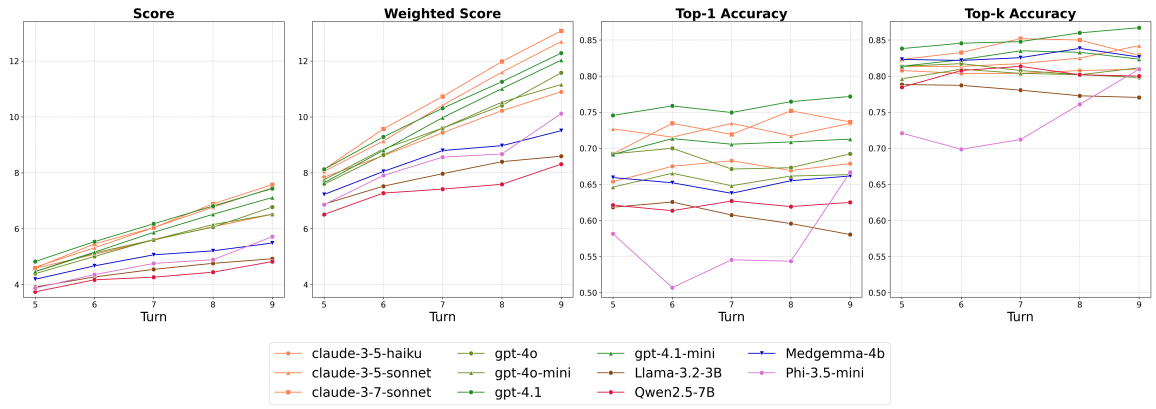

Figure 4: EPAG results across eleven models with number of dialogue turns ranging from five to nine.

performance of fine-tuned Qwen2.5-32B demonstrates that relatively small open-source models, when trained on high-quality, task-specific data, can outperform larger, more expensive models in specialized applications.

Not all HPI directly lead to correct diagnosis. As shown in Figure 4, the amount of HPI increases with the number of dialogue turns, while diagnostic accuracy does not. Appendix E exemplifies why more HPI does not directly correlate with accurate differential diagnosis. If a model fixates on certain keywords that are loosely connected to the correct diagnosis, it may ask numerous guideline-related but clinically less significant questions and even increase the likelihood of misdiagnosis.

Language affects dialogue patterns. With the prior experiments done in Korean, we explore whether the used language makes any difference by comparing English and Korean dialogues with Qwen 2.5 models (7B, 32B, 72B). We hypothesize that English pre-consultations would yield stronger performance as the English training corpus is understood to be much larger than Korean. Surprisingly, Figure 5 shows that Korean dialogues produce higher HPI-diagnostic guideline comparison scores, while English dialogues achieve superior disease diagnosis accuracy. A qualitative review explains this enigma: in English, the model frequently pursues deep, repetitive follow-ups on a single symptom—enhancing diagnostic confidence but generating fewer unique atomic units. By contrast, in Korean sessions it casts a wider net, querying a broader array of symptoms, which boosts HPI scores but dilutes focus and can introduce multiple diagnostic possibilities. This behavior aligns with our earlier finding that *not all HPI directly lead to correct diagnosis.*

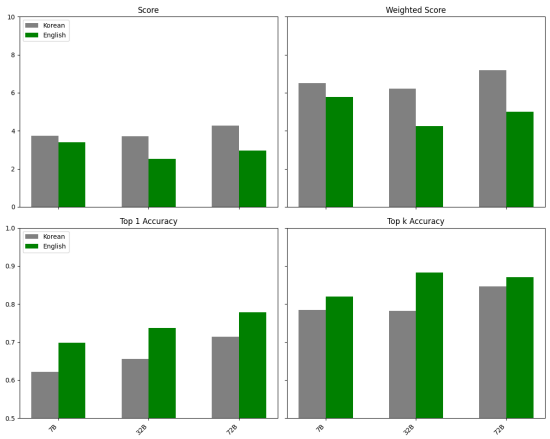

Figure 5: Performance of Qwen-2.5 models (7B, 32B, 72B) on Korean (grey) versus English (green) dialogues.

6 Conclusion

We present **EPAG**, a benchmark dataset and automated pipeline for **E**valuating the **P**re-consultation **A**bility of LLMs using diagnostic **G**uidelines. Experiments show that model size does not guarantee performance, and not all extracted HPI contribute directly to diagnosis, highlighting the need for future research to quantify the impact of each HPI component on specific diagnosis and refine pre-consultation models. Additional studies demonstrate that smaller open-source LLMs can surpass larger proprietary models when fine-tuned with high-quality data, and that the language used during pre-consultation shapes dialogue characteristics.

Limitation and Future Work

The EPAG benchmark dataset includes 26 diseases across 10 ICD-11 chapters but focuses solely on text-based pre-consultation models, excluding diseases that require physical test results, such as X-

rays, or MRIs, which are more common in real-world settings. Therefore, future work should incorporate multi-modal evaluation of pre-consultation models to process inputs beyond text, including medical images.

Ethics Statement

While our proposed evaluation pipeline for assessing the pre-consultation abilities of LLMs demonstrates a high correlation with human evaluation, it has limitations and does not cover all disease categories. As such, the experimental results presented in this paper should not be considered definitive. The selection of a model for any specific clinical application should involve thorough assessment before being deployed in practice.

References

Abdul Basit, Khizar Hussain, Muhammad Abdullah Hanif, and Muhammad Shafique. 2024. Medaide: Leveraging large language models for on-premise medical assistance on edge devices.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. 2024. A systematic review of testing and evaluation of healthcare applications of large language models (llms). *medRxiv*, pages 2024–04.

Balu Bhasuran, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan Han, Zhiyong Lu, and Zhe He. 2025. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *npj Digital Medicine*, 8(1):166.

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. 2024. Evaluating large language models in medical applications: a survey. *arXiv preprint arXiv:2405.07468*.

Julien Delaunay and Jordi Cusido. 2024. Evaluating the performance of large language models in predicting diagnostics for spanish clinical cases in cardiology. *Applied Sciences*, 15(1):61.

Dennis Fast, Lisa C. Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, Alexander Löser, and Keno K. Bressem. Autonomous medical evaluation for guideline adherence of large language models. *npj Digital Medicine*, 7.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. 2025. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Guangya Yu, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: A benchmark for evaluating large language models in multi-step clinical diagnosis.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhou Ran Cai, et al. 2025. An evaluation framework for conversational reasoning in clinical llms during patient interactions. *Nature Medicine*.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Medic: Towards a comprehensive framework for evaluating llms in clinical applications.

Mirae Kim, Kyubum Hwang, Hayoung Oh, Min Ah Kim, Chaerim Park, Yehwi Park, and Chungyeon Lee. 2024. MILD bot: Multidisciplinary childhood cancer survivor question-answering bot. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 665–676, Miami, Florida, US. Association for Computational Linguistics.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. 2025. Medical hallucinations in foundation models and their impact on healthcare.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. 2024. Medsyn: Llm-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer.

Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. 2024. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations.

Brenna Li, Ofek Gross, Noah Crampton, Mamta Kapoor, Saba Tauseef, Mohit Jain, Khai N Truong, and Alex Mariakakis. 2024. Beyond the waiting room: Patient's perspectives on the conversational nuances of pre-consultation chatbots. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards accurate differential diagnosis with large language models.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models.

Han Qian, Bin Dong, Jia-jun Yuan, Fan Yin, Zhao Wang, Hai-ning Wang, Han-song Wang, Dan Tian, Wei-hua Li, Bin Zhang, et al. 2021. Pre-consultation system based on the artificial intelligence has a better diagnostic performance than the physicians in the outpatient department of pediatrics. *Frontiers in Medicine*, 8:695185.

Justin T Reese, Leonardo Chimirri, Yasemin Bridges, Daniel Danis, J Harry Caufield, Michael A Gargano, Carlo Kroll, Andrew Schmeder, Fengchen Liu, Kyran Wissink, et al. 2025. Systematic benchmarking demonstrates large language models have not reached the diagnostic accuracy of traditional rare-disease decision support tools. *medRxiv*, pages 2024–07.

MANA SAMIEE. General practitioners' perspectives on llm chatbots for shared decision-making.

Peter Sarvari and Zaid Al-Fagih. 2025. Rapidly benchmarking large language models for diagnosing comorbid patients: comparative study leveraging the llm-as-a-judge method. *JMIRx Med*, 6:e67661.

Xiaoming Shi, Jie Xu, Jinru Ding, Jiali Pang, Sichen Liu, Shuqing Luo, Xingwei Peng, Lu Lu, Haihong Yang, Mingtao Hu, Tong Ruan, and Shaoting Zhang. 2023. Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Yang Tan, Zhixing Zhang, Mingchen Li, Fei Pan, Hao Duan, Zijie Huang, Hua Deng, Zhuohang Yu, Chen Yang, Guoyang Shen, et al. 2024. Medchatzh: A tuning llm for traditional chinese medicine consultations. *Computers in biology and medicine*, 172:108290.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards conversational diagnostic ai.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, and Andrew G Lee. 2024. Large language model (llm)-driven chatbots for neuro-ophthalmic medical education. *Eye*, 38(4):639–641.

Cai Wang, Qian Chen, Weizi Shao, and Xiaofeng He. 2024. Kemedgpt: Intelligent medical pre-consultation with knowledge-enhanced large language model. In *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pages 386–391.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax?

Isabella C Wiest, Marie-Elisabeth Leßmann, Fabian Wolf, Dyke Ferber, Marko Van Treeck, Jiefu Zhu, Matthias P Ebert, Christoph Benedikt Westphalen, Martin Wermke, and Jakob Nikolas Kather. 2024. Anonymizing medical documents with local, privacy preserving large language models: The llm-anonymizer. *medRxiv*, pages 2024–06.

Caleb Winston, Cleah Winston, Claris Winston, and Chloe Winston. Medical question-generation for pre-consultation with llm in-context learning. In *GenAI for Health: Potential, Trust and Policy Compliance*.

He Yang, Fei Wang, Matthew Greenblatt, Sharon Huang, and Yi Zhang. 2023a. Ai chatbots in clinical laboratory medicine: Foundations and trends. *Clinical chemistry*, 69.

Rui Yang, Ting Tan, Wei Lu, Arun Thirunavukarasu, Daniel Ting, and Nan Liu. 2023b. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024. Llm-based medical assistant personalization with short- and long-term memory coordination.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Principles, applications, and challenges.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. Diagnosisarena: Benchmarking diagnostic reasoning for large language models.

A Source of Diagnostic Guidelines

- Cardiology: American College of Cardiology [5], American Heart Association [6]
- Oncology: National Comprehensive Cancer Network [7]
- Stroke: American Heart Association, American Stroke Association [8]
- Allergy & Immunology: Joint Task Force for Practice Parameters [9], American Academy of Allergy Asthma & Immunology [10], American College of Allergy Asthma and Immunology [11]
- Gastroenterology: American Gastroenterological Association Homepage [12]
- HIV/AIDS: U.S. Preventive Services Task Force [13]
- Pulmonology: Global Initiative for Chronic Obstructive Lung Disease [14], Global Initiative for Asthma [15]
- Nephrology: Improving Global Outcomes [16]
- Diabetes: American Diabetes Association [17]
- General Surgery: American College of Surgeons [18]
- Rheumatology: European League Against Rheumatism [19]
- Endocrinology: American Association of Clinical Endocrinologists [20]

[5] https://www.acc.org/
[6] https://www.heart.org/
[7] https://www.nccn.org/
[8] https://www.stroke.org/en/
[9] https://www.aaaai.org/allergist-resources/statements-practice-parameters/practice-parameters-guidelines
[10] https://www.aaaai.org/
[11] https://acaai.org/
[12] https://gastro.org/
[13] https://www.uspreventiveservicestaskforce.org/uspstf/
[14] https://goldcopd.org/
[15] https://ginasthma.org/
[16] https://kdigo.org/
[17] https://diabetes.org/
[18] https://www.facs.org/
[19] https://www.eular.org/
[20] https://www.aace.com/

B Diagnostic Guideline Example

Weight	Feature
high	Palpable Breast Lump
high	Nipple Discharge, Bloody or Spontaneous
high	Skin Changes: Peau d'orange, Ulceration, Erythema, Thickening
high	New-Onset Nipple Inversion/Retraction
high	Axillary Masses/Lymphadenopathy
medium	Asymmetry in Breast Size/Shape, New Onset
medium	Nipple/Areolar Eczema or Itching
medium	Localized Thickening or Induration
medium	Systemic Symptoms: Weight Loss, Fatigue, Night Sweats, Fever
medium	Pregnancy/Lactation-Related Abnormalities
medium	Post-Surgical or Post-Radiation Breast Changes
high	Family History of Breast Cancer, BRCA Mutation
high	Genetic Predisposition: BRCA1/BRCA2, TP53, PALB2 etc.
high	Prior Biopsy with Atypia or LCIS/ADH
medium	Hormonal Factors: Early Menarche, Late Menopause, HRT Use
high	Prior Chest Radiation Therapy, esp. 10∼30 y/o

Table 2: Diagnostic guidelines for breast cancer.

C Disease Categorization

To enhance the generalization and reliability of our benchmarking system, we adopt the International Classification of Diseases, 11th Revision (ICD-11) as the main categorization of diseases. This approach ensures comprehensive coverage across diverse disease groups. For better alignment with real-world clinical decision-making we assign each disease to a Primary Specialty and, where applicable, one or more Secondary Specialties.

C.1 Primary Specialty Selection Criteria

Each disease is assigned to a Primary Specialty, the leading specialty responsible for the disease's management, based on the following:

1. ICD-11 Disease Classification:
    - Each disease is mapped to its corresponding ICD-11 chapter, which indicates the major body system or disease category it belongs to.
    - The specialty most commonly responsible for managing diseases in each chapter is assigned as the Primary Specialty.
2. International Clinical Guidelines: The Primary Specialty is further validated using well established medical guidelines from globally recognized organizations listed in Appendix A.
3. Standard Medical Practice: The most commonly designated department responsible for managing the disease in hospitals and healthcare settings is selected.

C.2 Secondary Specialty Selection Criteria

Many diseases require collaboration across multiple specialties. A Secondary Specialty, additional specialties that frequently contribute to diagnosis, treatment, or complication management, is assigned in cases where:

1. Multidisciplinary care is essential.
    - Conditions which require involvement from multiple specialties for optimal management.
    - Example: Stroke (8B20)
        - Primary: Neurology (acute treatment and long-term management)
        - Secondary: Cardiology (stroke prevention in atrial fibrillation), Rehabilitation Medicine (post-stroke recovery)
2. Complication management is required.
    - Specialties involved in managing complications related to the primary disease.
    - Example: Diabetes (5A14)
        - Primary: Endocrinology (blood sugar control, metabolic regulation)
        - Secondary: Nephrology (diabetic nephropathy), Cardiology (cardiovascular risk)
3. Surgical vs. Non-Surgical considerations.
    - Conditions where both medical and surgical specialties play a role.
    - Example: Colorectal Cancer (2B91)
        - Primary: Oncology (chemotherapy and cancer management)
        - Secondary: Gastroenterology (diagnosis via colonoscopy), General Surgery (surgical treatment)

By structuring disease classification based on these criteria, we ensure that our benchmark system accurately represents real-world clinical workflows and enhances the applicability of AI-driven medical decision support tools.

D Organized Unit Example

Main Symptom:
I keep coughing and have difficulty breathing.

D.1 [Question, Options, Answer] Triplet

Question: When you cough, do you produce any sputum?
Options: Dry cough with no sputum, White or clear sputum, Yellow or green sputum, Red or brown sputum
Answer: White or clear sputum

Question: When is your difficulty breathing worse?
Options: I have difficulty breathing even when I am at rest, I have difficulty breathing when walking on flat ground, I have difficulty breathing when climbing stairs or going uphill, I only have difficulty breathing when I move quickly or exercise
Answer: I have difficulty breathing when walking on flat ground

Question: How long have you had the coughing and difficulty breathing symptoms?
Options: Less than 2 weeks, 2 weeks to 3 months, 3 months to 6 months, More than 6 months
Answer: More than 6 months

Question: Do you smoke?
Options: I currently smoke, I used to smoke but quit, I have never smoked, I am often exposed to secondhand smoke
Answer: I currently smoke

Question: Do you have any symptoms while sleeping?
Options: I wake up because I can't breathe, I can't sleep due to severe coughing, I need more than one pillow to breathe properly, I snore a lot, I sleep without any special symptoms
Answer: I wake up because I can't breathe

D.2 Organized Units

- White or clear sputum is produced when coughing.
- The difficulty in breathing worsens when walking on flat ground.
- The coughing and difficulty in breathing symptoms have lasted more than 6 months.
- I currently smoke.
- I wake up during sleep because I can't breathe.

E Analysis

The following case involves a patient expected to be diagnosed with *Acute Kidney Injury*. MedGemma-4B is used as the doctor agent model.

Chief Complaint: Decreased urine output and flank pain.
HPI from 5-turn dialogue
There is pain in the right flank. The amount of urine has decreased. Recently had symptoms of a cold. Takes antihypertensive medication regularly. No history of urinary stones.
Diagnosis: *Acute Kidney Injury* (correct)
HPI from 6-turn dialogue
There is pain in the right flank. The amount of urine has decreased. Recently had symptoms of a cold. Takes antihypertensive medication regularly. No history of urinary stones. The flank pain is severe, rated 7 out of 10 in intensity. (Added)
Diagnosis: *Renal Colic due to Urinary Stone* (incorrect)

Although both *Acute Kidney Injury* and *Renal Colic* can present with flank pain, the additional 6th turn provides patient information about the intensity of pain, which may have shifted the model's diagnostic focus away from other relevant symptomatic information. *Renal Colic* typically results from urinary stone, leading to severe pain. In this case, highlighting the severity of flank pain may have caused the model to prioritize pain-centric reasoning, which misled the differential diagnosis toward *Renal Colic*. While the additional information (pain intensity) is clinically relevant and could aid a physician's understanding, it may have inadvertently diverted the model's diagnostic focus.

ICD-11 Chapter	Disease	ICD-11 Code	Primary Specialty	Secondary Specialty
Neoplasms	Breast Cancer	2E65	Oncology	General Surgery
	Prostate Cancer	2C82	Oncology	Urology
	Colorectal Cancer	2B91	Oncology	Gastroenterology, General Surgery
	Lung Cancer	2C25	Oncology	Pulmonology, Thoracic Surgery
	Gastric Cancer	2B72	Oncology	Gastroenterology, General Surgery
Diseases of the Circulatory System	Hypertrophic Cardiomyopathy	BC43.1	Cardiology	Medical Genetics
	Peripheral Artery Disease	BD4Z	Cardiology	Vascular Surgery
	Atrial Fibrillation	BC81.3	Cardiology	Neurology (Stroke Risk), Internal Medicine
	Heart Failure	BD1Z	Cardiology	Endocrinology (Diabetes-related)
Diseases of the Nervous System	Stroke	8B20	Neurology	Cardiology, Rehabilitation Medicine
	Aneurysmal Subarachnoid Haemorrhage	8B01.0	Neurology	Neurosurgery, Emergency Medicine
Diseases of the Immune System	Anaphylaxis	4A84	Allergy & Immunology	Emergency Medicine
	Systemic Sclerosis	4A42	Rheumatology	Pulmonology (Lung fibrosis), Cardiology (Cardiac involvement)
	Systemic Lupus Erythematosus	4A40.0	Rheumatology	Nephrology (Lupus Nephritis), Cardiology (Vascular Complications)
Diseases of the Skin	Atopic Dermatitis	EA80	Allergy & Immunology	Dermatology
Diseases of the Digestive System	Ulcerative Colitis	DD71	Gastroenterology	Rheumatology (Autoimmune-related)
	Nonalcoholic Fatty Liver Disease	DB92.Z	Gastroenterology	Endocrinology (Metabolic Syndrome)
	Irritable Bowel Syndrome with Constipation (IBS-C)	DD91.00	Gastroenterology	Psychiatry (Stress-related IBS)
	Acute Pancreatitis	DC31	Gastroenterology	General Surgery
Certain Infectious or Parasitic Diseases	Human Immunodeficiency Virus (HIV) Infection	1C62	Infectious Diseases	Immunology
Diseases of the Respiratory System	Chronic Obstructive Pulmonary Disease	CA22	Pulmonology	Internal Medicine
	Asthma	CA23	Pulmonology	Allergy & Immunology
	Allergic Rhinitis	CA08.0	Allergy & Immunology	Otorhinolaryngology, Pulmonology
Diseases of the Genitourinary System	Acute Kidney Injury	GB60	Nephrology	Critical Care Medicine
Endocrine, Nutritional or Metabolic Diseases	Diabetes Mellitus	5A14	Endocrinology	Nephrology (Diabetes-related Kidney Disease)
	Hypothyroidism	5A00	Endocrinology	Cardiology (Atrial Fibrillation Risk), Psychiatry (Depression Link)

Table 3: List of 26 diseases consisting EPAG benchmark. Detailed classification of diseases including ICD-11 Chapter, ICD-11 Code, Primary Specialty, and Secondary Specialty are provided.

Patient Profile		
Disease Name	**Breast Cancer**	
Typicality	**Normal**	
Basic Information	**Age**	51
	Sex	Female
	Height	162cm
	Weight	62kg
History of Present Illness	**Location**	Left breast and adjacent axillary region
	Quality	Firm, irregular mass
	Severity	4/10 (Mild pain but significant anxiety)
	Duration	Approximately 3 months
	Timing	Slight variations with menstrual cycle, discovered accidentally during routine examination
	Context	Detected by the patient herself during a routine breast examination
	Modifying Factors	Slight reduction in swelling post-menstruation, no specific alleviating factors
	Associated Signs and Symptoms	Mild nipple discharge, slight fatigue, minimal pain
Additional Information	**Family History**	No family history of breast cancer or similar cancers
	Previous Surgery or Illness	No previous history of breast-related surgery or conditions
	Lifestyle Changes	No recent changes in lifestyle; the patient aims for early detection through screening
	Health Check-ups	Regularly undergoes women's health check-ups
Pain Area	Left chest (pectoral region)	
	Left anterior acromio-clavicular region	
Past Medical History	No history of breast diseases	
	No other chronic illnesses	
Social History	Office worker, full-time	
	Non-smoker, drinks alcohol 1-2 times per week	
	Regular health check-ups and breast self-examination	
Chief Complaint	A firm lump in the left chest, causing anxiety	

Table 4: Sample patient profile with breast cancer.

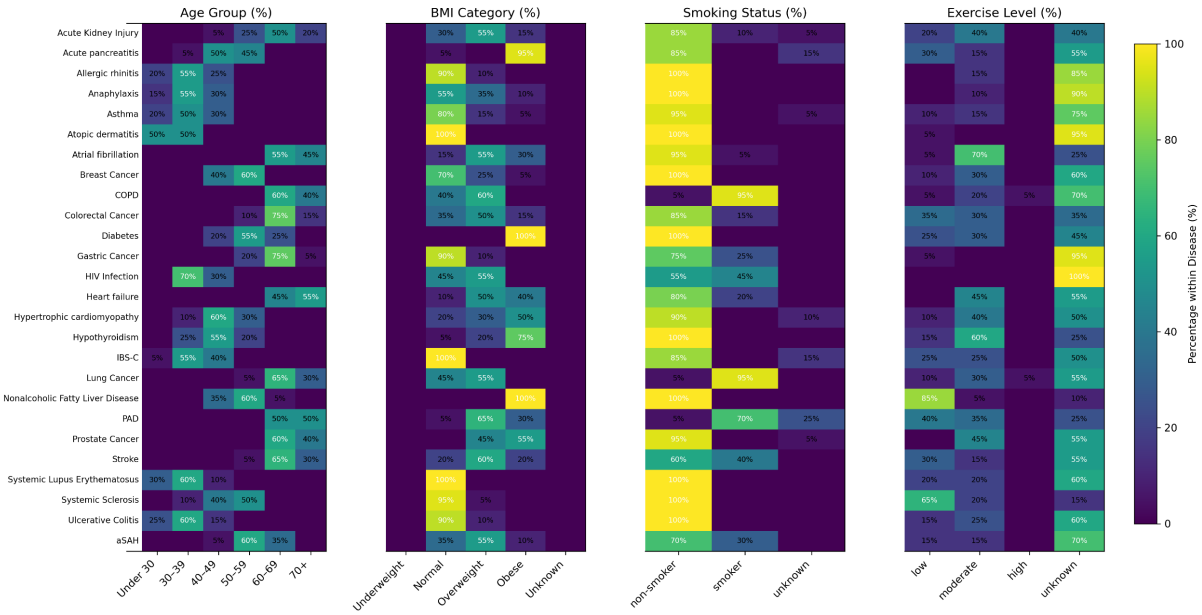

Figure 6: Distribution of age group, BMI category, smoking status, exercise level of patients for each disease.

Original

당신은 아래와 같은 프로필을 가진 환자입니다.
{patient_information}

적절한 진단을 위해 의사의 문진에 충실히 대답해야 합니다. 의사가 제시한 질문지 중 한 가지를 골라 대답하세요.

출력 형식:
Answer:

Translated

You are a patient with the following profile:
{patient_information}

You should faithfully answer the doctor's inquiries for an appropriate diagnosis. Choose one of the questions provided by the doctor and respond.

Output format:
Answer:

Table 5: Patient Agent Prompt

Original

Prompt:
Below is a set of consultation dialogues between a doctor and a patient with disease. The dialogue consists of the patient's chief complaint, a few turns of questions, options, and answer triplets.
The questions and options are given by the doctor, and the answers are provided by the patient.
Your task is to organize the dialogue in a clear, information-based manner using bullet points. Each bullet point should contain only one piece of information.
This structured information is essential for diagnosing the patient's condition, so make sure to extract as much relevant information as possible.

Guidelines for Organizing:
- Do not include the main symptom in the bullet points.
The main symptom is just for reference and should not be summarized in bullet points.
- Focus only on the information that can be inferred from the Question-Options-Answer triplets.
- Each bullet point must present only one piece of information.
- Avoid sentences with multiple clauses.
For example, instead of "The patient has cough and sputum," break it down into two points:
"The patient has a cough" and "The patient has sputum."
- Avoid using demonstrative pronouns (e.g., "this symptom") and pronouns (e.g., "he/she"). Focus on the specific symptoms.
- Organize the information from the patient's perspective, avoiding the doctor as the subject.
- Keep the language neutral and concise, stating only the facts shared by the patient.
- If the question asks about additional symptoms beyond the main symptom and the answer is that no other symptoms are present,
list each symptom option provided in the question and state that the patient does not have each of those symptoms.
For example, instead of "There are no other symptoms," specify each of the symptom option provided:
"There is no family history," "There is no weight loss," "There is no fever."
- Be precise and specific in organizing the information.
For example, if a question asks about "whether the patient has ever had any tests related to lumps or breasts,"
and the answer is "No," do not simply write "The patient has not had any tests."
Instead, write, "The patient has not had any tests related to lumps or breasts."

Example:
{examples}

Input
{input}

Organized Information:
-

Table 6: Organizer Prompt

Original

You are a medical/health expert. Below is a conversation between a disease patient and a doctor.
In this case, evaluate whether [the interview conversation (A)] effectively leads to the [key diagnostic elements (B)], which are pre-defined for specific diseases.
Here, (B) includes not only symptoms but also important elements such as past medical history, family history, and other disease diagnoses.
First, identify if (A) is relevant enough to disease and helpful in drawing out new information to diagnose disease given (H).
If not, output "Irrelevant/Redundant."
If (A) is relevant to disease and helpful in drawing out new information to diagnose disease given (H),
determine whether each item in (B) can be identified through the interview conversation (A).
If two or more (B) items can be identified from (A), output the most relevant (B) item. If no (B) items can be identified through (A), output "None of above."

<Explanation of the provided information>
- **Dialogue History (H)**
This is a prior conversation between the patient and the doctor.
It includes the main symptom the patient reported, the questions the doctor asked to make a diagnosis, the options presented, and the patient's answer.
Sometimes only the main symptom the patient complained about may be provided.
- **Interview Conversation (A)**
This consists of the questions and options the doctor asks the patient for diagnostic purposes.
The patient chooses one option from the given choices to respond.
- **Pre-defined Key Diagnostic Elements List (B)**
Example: Persistent Cough, Hemoptysis (Coughing up Blood), Dyspnea (Shortness of Breath), Chest Pain, Unexplained Weight Loss, Family History of Lung Cancer, Smoking History, etc.

<Important Notes>
1. **Evaluation Criteria**
- Check if the interview conversation (A) is designed to identify (B),
or if it directly helps to determine specific aspects of (B) such as the onset, duration, more exact location and frequency of symptoms.
- If (A) is related to an item in (B) but deviates from the patient disease which is disease, then output "None of above."
- Assess if the questions and options in (A) can effectively elicit relevant information related to (B) from the patient.
2. **Output Criteria** - Provide a brief Reason for
whether (A) can effectively elicit (B)-related information. Do not repeat the questions and options.
- The Reason should be up to two sentences.
- The Final Response should be either [(B) item] or "None of above." or "Irrelevant/Redundant."
- If multiple (B) items can be identified from (A), output only the one most directly related to (A).
If the relevance is judged to be the same, separate the related (B) items using "[OR]" and output them all.
- (H) is for reference only, so the evaluation should focus on whether (A) is related to (B).

Example:
{example}

(H):
{h}
(A):
{a}
(B):
{b}
Reason:

Table 7: Comparer Prompt

Original

You are a medical expert. Given 'patient_info' and 'medical_history', output the suspected disease names in order of highest probability. Output your prediction in English in YAML format.

Instructions:
- Use only specific disease names related to the patient's symptoms.
- Prioritize based on main symptoms, severity, duration, and answers given in the medical history.
- Exclude diseases that don't match the responses or are too generic.
- List the diseases in order of highest probability first.
- Do not provide any extra explanation.

Output format:
Diseases:
- (probable diseases)

Table 8: Diagnostician Agent Prompt

Original

You are a medical expert. Given 'model_predictions' and 'golden_standard', decide if the predictions are correct. Output your reasoning in English in YAML format.

Instructions:
- Accept if the predicted disease is very similar to the actual one.
- Accept synonyms or other expressions for the same disease.
- Accept if the disease names include hierarchical (superior/inferior) relationships.
- Accept medical abbreviations as equivalent to official names.
- Allow regional/cultural expression differences.
- If at least one prediction is correct, consider it acceptable.

Output format:
Reasoning: |
(your reasoning in English)
Result: True/False

Table 9: Evaluator Prompt

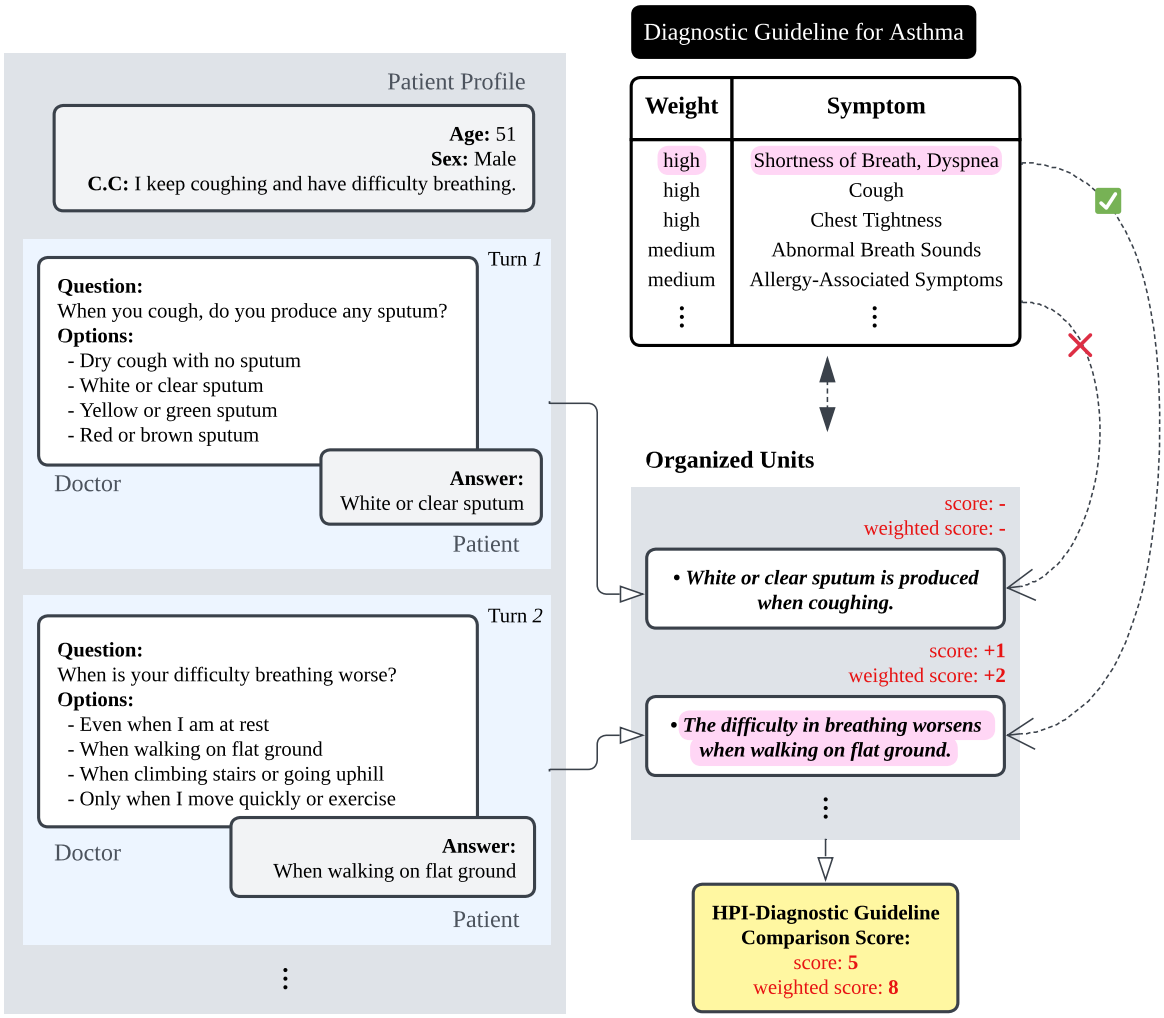

Figure 7: Sample pre-consultation dialogue and HPI-diagnostic guideline comparison process. Given basic patient information, including the chief complaint, the doctor asks questions and the patient selects answers from provided options. The dialogues are organized into atomic units, each of which is compared against a pre-defined diagnostic guideline. Units matching the guideline receive a score; those that do not are not scored.

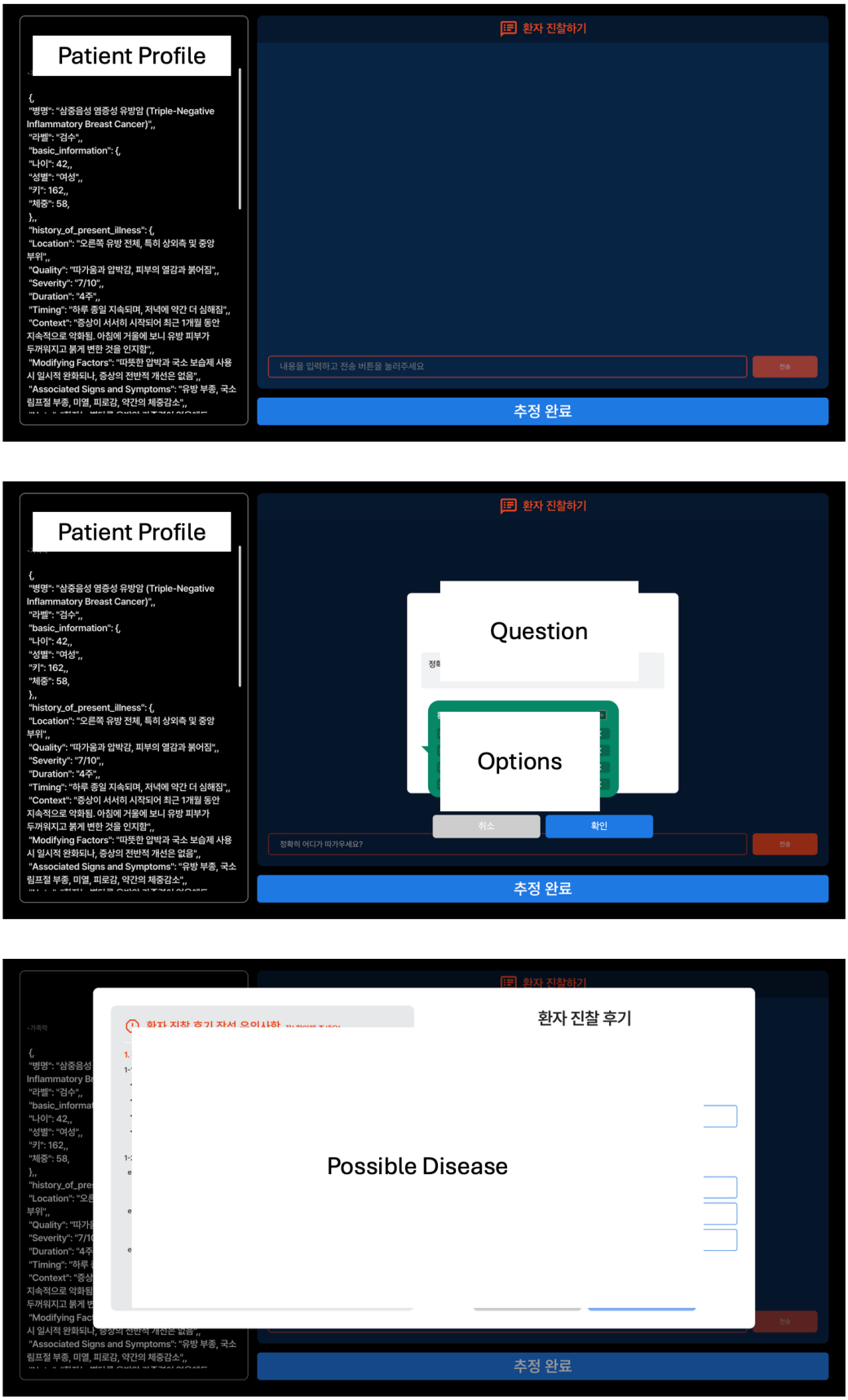

Figure 8: User interface used by human clinicians to simulate pre-consultation dialogues with patient agents. Given the patient profile displayed on the left, clinicians generate questions and response options for the patient agent to select. After each submission, the selected option is shown to the clinician, who then formulates the next question and options. After a series of dialogue turns, clinicians provide a diagnosis of the possible diseases.