

# ChatGPT の過剰回答に対する自己フィードバック機構を組み込んだ医療面接試験向け仮想模擬患者

進藤尚希<sup>1</sup> 宇都雅輝<sup>1</sup>

<sup>1</sup> 電気通信大学大学院 情報理工学研究所  
{shindo\_naoki,uto}@ai.lab.uec.ac.jp

## 概要

近年、医療系学生が受験する客観的臨床能力試験における医療面接を標準化する手法の一つとして、人工知能技術に基づく仮想模擬患者が多数提案されている。しかし既存手法の多くは、ルールベースのアプローチを採用しており、回答ルールの構築などに大きな負担を要する。そこで本研究では、ChatGPT を用いて人手によるルール構築を必要としない仮想模擬患者の実現を目指す。ただし、ChatGPT は受検者の質問に対して過剰に回答する場合があるため、提案手法ではそのような過剰回答を抑制する機構を新たに導入し、試験目的に則した回答を出力させることを目指す。

## 1 はじめに

医療系学生が受験する共用試験の一つである客観的臨床能力試験 (Objective Structured Clinical Examination ; OSCE) における医療面接課題では、特定の診療シナリオに沿って回答するように訓練された標準模擬患者 (Standardized Patient ; SP) に対して受検者が質問を行うプロセスを評価することで、臨床技能と態度を測定する [1, 2]。一方でこのような評価では、SP や評価者が評価結果に影響を与えるバイアス要因となりうる [3]。評価者に由来するバイアスの影響については、テスト理論の一つである項目反応理論の拡張モデル [4, 5, 6, 7] を用いるなどの対応策が検討されている。一方で、SP に由来するバイアスを取り除く方法論としては、十分に訓練された SP を採用する方法が一般的である。しかし、SP の十分な訓練には膨大なコストがかかる。

このような問題を解決するために、人工知能技術に基づく仮想模擬患者 (Virtual Standardized Patient ; VSP) を用いて、SP の標準化を行うアプローチが注目されている。VSP を実現する既存手法は、人手に

より事前定義された回答のルール集合から、質問の内容に対応する回答文を選択することで対話を進行するルールベースを用いた手法が広く採用されてきた。一般に、質問と回答文とのマッチングの方法は、人手で設計された厳格なパターンマッチングによる手法 [8, 9, 10, 11] と、深層学習を用いた手法 [12, 13, 14, 15, 16] に分類できる。深層学習を用いた手法では、質問の意図を推論することにより、パターンマッチングによる手法よりも柔軟な回答の選択を実現したが、モデルの訓練のための質問意図がラベル付けされた質問-回答の対のデータセットを必要とする。またいずれの手法も、回答を人手で設計したルール集合から選択する仕組みであるため、ルール作成のコストが大きいという問題を有する。

これらの問題を解決するために、本研究では、ChatGPT に診療シナリオと回答の仕方の指示を与えることで、SP として回答を行う VSP の実現を目指す。ChatGPT に基づく VSP は、人手でのルール設計やラベル付き訓練データセットなどを必要とせずに、与えたシナリオに存在しない内容に関する質問に対しても、文脈を考慮した自然な質問応答を行える利点を有する。一方で、受検者からの質問に直接的に関係しないシナリオまで参照し、過剰な回答を返してしまうことがある。過剰な回答を返した際の例を付録 A.1 に示す。このような回答は、質問を通して模擬患者から適切な情報を引き出せたかどうかを測定する OSCE の試験目的において不適切な振舞いである。そこで本研究では、二つの事前学習された BERT (Bidirectional Encoder Representations from Transformers)[17] を用いて、受検者の質問に対する ChatGPT の過剰な回答を検出し、それらを ChatGPT にフィードバックして適切な回答を再出力させる手法を提案する。具体的には、一つ目の BERT は受検者の質問文と個々のシナリオ文との関連度の予測を行い、二つ目の BERT は受検者の質問文に対する

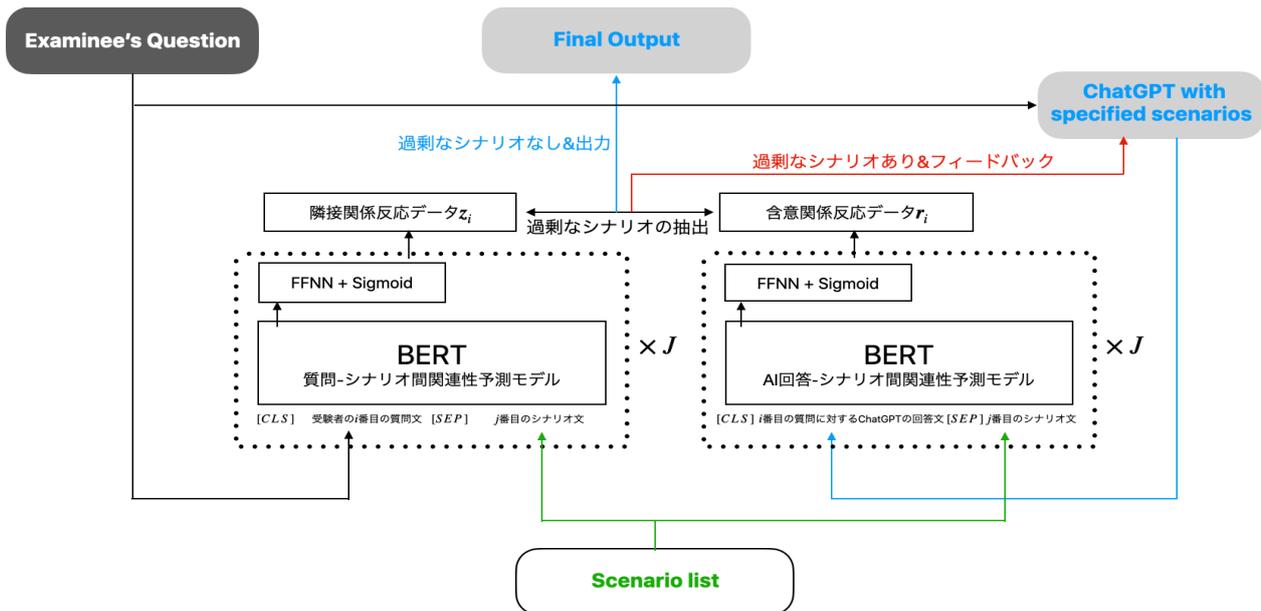


図 1: 提案手法の概略図

ChatGPT の回答が個々のシナリオ文を含意しているかの予測を行う。そして、この二つの BERT の予測結果を比較することで、ChatGPT の回答が過剰なシナリオを含んでいるかを判定し、その判定結果をもとに過剰回答を抑制した回答を再出力させるフィードバックを ChatGPT に与える。本研究では、提案手法を 4 つの症状に関するシナリオに適用する実験を通して、提案手法の有効性を評価する。

## 2 提案手法

上記の通り、提案手法では、受検者の質問に対する ChatGPT の過剰回答の検出を、二つの BERT を組み合わせることで実現する。一つ目の BERT は、OSCE 形式で実施された医療面接の対話データセット [18] でファインチューニングを行なったモデルであり、受検者の質問文と個々のシナリオ文との対応関係を予測する。二つ目のモデルは、一般言語理解評価ベンチマーク (General Language Understanding Evaluation ; GLUE)[19] タスクの一つである含意判定タスクでファインチューニングを行なったモデルであり、受検者の質問に対する ChatGPT の回答文が、個々のシナリオ文を含意しているかを予測する。これら二つの BERT の予測結果を比較し、質問に直接関係しないシナリオへの言及を含む回答を検出して、ChatGPT にフィードバックを行い、適切な回答を再出力させる。提案手法の概略図を図 1 に示す。以降では、提案手法の個々の手続きについて詳述する。

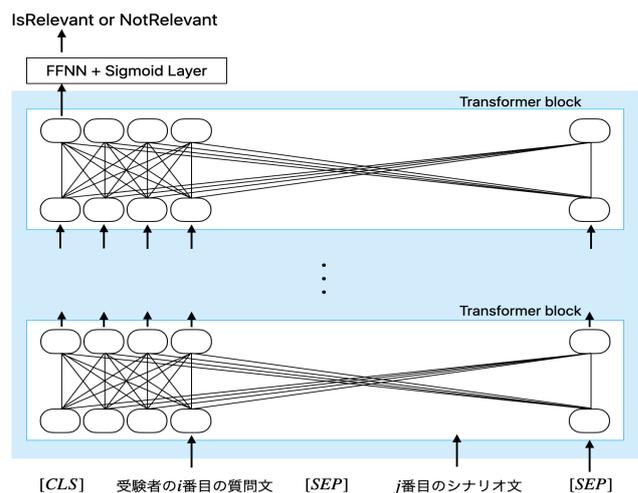


図 2: 質問-シナリオ間関連性予測モデルの概略図

### 2.1 質問とシナリオの対応関係を予測する BERT モデル

質問とシナリオの対応関係を予測する BERT モデルでは、個々のシナリオが、受検者の質問と関係しているかを判定する。具体的には、BERT の事前学習タスクの一つである Next Sentence Prediction (NSP)[20] と同様に、質問とシナリオの二文間の関係が隣接文として成り立つかどうかの 2 値分類を全てのシナリオに対して行うモデルを構築する。以降では、この BERT モデルを「質問-シナリオ間関連性予測モデル」と呼ぶ。

モデルの概念図を図 2 に示す。受検者の  $i$  番目の質問の単語系列と、 $j$  番目のシナリオの単語系列

を, [CLS] と [SEP] という特殊トークンを挿入してモデルに入力する. この入力の [CLS] トークンに対応する BERT の分散表現ベクトル  $\mathbf{x}_{ij} \in \mathbb{R}^{768}$  に対して, 次式で与えられる Linear Layer with Sigmoid Activation を適用する.

$$v_{ij} = \sigma(\mathbf{W}\mathbf{x}_{ij} + b) \quad (1)$$

ここで,  $\sigma$  はシグモイド関数を,  $\mathbf{W}$  と  $b$  はそれぞれ重みパラメータとバイアスパラメータを表す. 式 (1) より得られた  $v_{ij}$  は, 0 から 1 の間をとる連続値である.  $v_{ij}$  に対して, 閾値を 0.5 に設定し,  $i$  番目の質問に対する  $j$  番目のシナリオが隣節文として成り立つか否かの 2 値分類を次式で計算する.

$$z_{ij} = \begin{cases} 1 & v_{ij} \geq 0.5 \\ 0 & \text{else} \end{cases} \quad (2)$$

ここで,  $z_{ij}$  は,  $i$  番目の質問に対して  $j$  番目のシナリオが隣接関係にある場合に 1 を取り, 隣接関係なければ 0 を取る変数である. このプロセスを全てのシナリオに対して繰り返し, それぞれのシナリオが受検者の質問に関連しているかどうかを表す次式の隣接関係反応データ  $\mathbf{z}_i$  を得る.

$$\mathbf{z}_i = \{z_{ij} \mid j \in 1, \dots, J\} \quad (3)$$

## 2.2 ChatGPT の回答とシナリオの対応関係を予測する BERT モデル

ChatGPT の回答とシナリオの対応関係を予測する BERT モデルでは, ChatGPT の回答がどのシナリオを参照しているかを判定する. 具体的には, GLUE の含意判定タスクでファインチューニングされた, 図 2 と同様の構造を持つ BERT を用いる. 以降では, この BERT を「AI 回答-シナリオ間関連性予測モデル」とする.

ここで, AI 回答-シナリオ間関連性予測モデルへの入力が例えば,

“[CLS] シナリオ A, B, C に言及した ChatGPT の回答文 [SEP] シナリオ A の文章”

であるとすると, モデルは, シナリオ A の重複をもって含意と判定するはずである. このプロセスを全てのシナリオに適用することで,  $i$  番目の質問に対する ChatGPT の回答の含意関係反応データ  $\mathbf{r}_i$  を次のように得る.

$$\mathbf{r}_i = \{r_{ij} \mid j \in 1, \dots, J\} \quad (4)$$

ここで,  $r_{ij}$  は,  $i$  番目の質問に対する ChatGPT の回答が  $j$  番目のシナリオを含意しているとき 1, そうでないとき 0 を取る変数である.

## 2.3 過剰なシナリオの検出とそれらのフィードバック手法

2.1 節と, 2.2 節で説明した二つの反応データ  $\mathbf{z}_i, \mathbf{r}_i$  を用いて, 質問には直接関係ないが, ChatGPT の回答には含まれている過剰なシナリオを検出する. ここで, これらの反応データの要素は, 0 と 1 の 2 値データを取るため,  $j$  番目の要素同士を比較すると, 考えられる組み合わせは  $(z_{ij}, r_{ij}) \in \{(0,0), (0,1), (1,0), (1,1)\}$  の 4 通りである. これらのうち  $(z_{ij}, r_{ij}) = (0,1)$  のケースは「受検者の質問に関連しないシナリオであるにも関わらず, ChatGPT の回答にはそのシナリオへの言及が認められる」ことを意味する. したがって, このケースを検出することで, ChatGPT の過剰回答を検出できると期待される. したがって, この比較を全てのシナリオ  $j \in \{1, \dots, J\}$  について行うことで,  $i$  番目の質問に対する ChatGPT の回答に含まれる過剰なシナリオを検出する.

過剰なシナリオが検出された場合, それらを削減して適切な回答を再出力するように, 再帰的に ChatGPT にフィードバックを与える. フィードバックのためのプロンプトへの入力は以下の通りである.

“Please refrain from mentioning scenario of { 過剰なシナリオ } in this doctor’s question.”

ChatGPT はフィードバックを考慮した回答を再出力する. フィードバック後の回答に対し, 上述の二つの反応データの比較を再度行い, フィードバック後の回答が過剰なシナリオを含んでいるか確認を行う. これらの手順を, 過剰なシナリオが削減されるまで再帰的に繰り返す. ただし, 本研究では, 5 回目のフィードバックでも過剰なシナリオを削減しきれなかった場合は, 5 回目の出力を最終的な出力として回答する.

## 2.4 質問-シナリオ間関連性予測モデルのファインチューニング

提案手法では, 2.1 節の質問-シナリオ間関連性予測モデルを, カナダ・ウェスタン大学研究チームが開発した医療面接データセット [18] を用いて, NSP タスクでファインチューニングを行った. このデータセットは, 272 セットのシナリオに関する, OSCE 形式でシミュレートされた医療面接が収録されている. ただし, 本データセットには, モデルの訓練精度に影響を与える可能性のある文脈を推測できない短い回答が多く存在するため, それらを排除, また

表 1: 4 つのシナリオに対する提案手法の性能評価

評価指標	モデル	シナリオ 1	シナリオ 2	シナリオ 3	シナリオ 4	平均
Recall	w/ FT	<b>0.70</b>	<b>0.95</b>	<b>0.46</b>	<b>0.70</b>	<b>0.70</b>
	w/o FT	0.40	0.45	0.39	0.17	0.35
Specificity	w/ FT	0.97	0.95	0.98	0.98	0.97
	w/o FT	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>0.99</b>

† w/ FT はファインチューニングあり, w/o FT はファインチューニングなしを表す

表 2: フィードバック後の過剰回答の修正率

対話の基礎統計量 (合計)	シナリオ 1	シナリオ 2	シナリオ 3	シナリオ 4	平均
過剰な回答の回数	20	12	7	16	13.75
モデルが過剰と予測した回数	16	14	4	13	11.75
過剰なシナリオを削減できた回数	16	9	1	12	9.50
過剰なシナリオを削減できた割合	0.80	0.75	0.14	0.75	0.69

は修正した。モデルの学習は交差エントロピー誤差を最小化することで行った。

### 3 評価実験

本章では、提案手法の有効性を評価するために、筆者らが作成した 4 つのシナリオに基づく実験を行う。以降の 3.1 節, 3.2 節で詳述する評価実験では、それぞれのシナリオ文を指定した ChatGPT に、受検者の質問文を一文ずつ与え、その質問に対する回答を評価した。作成したシナリオの詳細や、実験の設定は付録 A.2 に詳述した。

#### 3.1 過剰回答の検出精度の評価

まず、提案手法により過剰回答を検出できるかを評価する実験を行った。作成した 4 つのシナリオに基づく回答生成をそれぞれ 5 回ずつ行い、計 495 回の対話を生成した。ただし、ここでは回答にフィードバックは適用していない。生成した ChatGPT の各回答が過剰であるかを著者らが評価し、それらを正解値として Recall と Specificity を用いて提案手法の精度評価を行った。また、質問-シナリオ間関連性予測モデルにおいて、2.4 節のファインチューニングを行わないモデルとの性能比較も行った。

シナリオごとの 5 回の試行の平均精度を表 1 に示す。本研究において、Specificity は、実際は過剰な回答でないにも関わらず、過剰と判断した割合に対応する。表 1 より、提案手法では、シナリオやファインチューニングの有無に依らず総じて正確に過剰でない回答を識別できていることが確認できる。Recall は、実際は過剰な回答であるが、過剰でないと判断した割合に対応し、本研究の目的である過剰な回答に対するフィードバックを行うために特に重

要である。Recall の結果から、ファインチューニングを行ったモデルを用いた提案手法では、シナリオ 3 を除く全てのシナリオで正確にそれらを検出できていることが確認できる。

#### 3.2 過剰回答のフィードバックによる改善の有無の評価

ここでは、過剰回答のフィードバックにより最終的な回答に改善が見られたかを評価する実験を行った。具体的には、過剰回答を検出した場合に、実際に ChatGPT の回答に 2.3 節で説明したフィードバックを適用し、その最終出力において、過剰なシナリオが削減されているかどうかを評価した。この実験では、それぞれ 4 つのシナリオに基づく回答生成を 3 回ずつ行い、生成した計 297 回の対話で評価した。過剰なシナリオを含んだ全体の回答の内、69% の割合で過剰なシナリオを削減した回答の再出力ができていた。これらの詳細を表 2 に示す。例として、フィードバックにより過剰な回答が改善された回答を付録 A.1 に示す。

### 4 まとめ

本研究では SP としての回答をする対話 AI を実現するために、質問-シナリオ間関連性予測モデルと、AI 回答-シナリオ間関連性予測モデルと呼ぶ 2 種類の BERT モデルを用いて ChatGPT の回答に含まれる過剰なシナリオを検出し、それらを抑制する手法を提案した。提案手法では、検出した過剰なシナリオを抑制するフィードバックを経て、OSCE の場面において適切である必要なシナリオのみを参照した回答の再出力を可能にした。今後は、実際の被験者実験を通して、提案手法の有効性を評価したい。

## 謝辞

本研究は JSPS 科研費 19H05663, 21H00898 の助成を受けたものです。

## 参考文献

- [1] R. M. Harden. What is an OSCE? **Medical Teacher**, Vol. 10, No. 1, pp. 19–22, 1988.
- [2] 伴真太郎. 客観的臨床能力試験. **医学教育**, Vol. 26, No. 3, pp. 157–163, 1995.
- [3] Iris Schleicher, Karsten Leitner, Jana Juenger, Andreas Moeltner, Miriam Ruessler, Bernd Bender, Jasmina Sterz, Karl-Friedrich Schuettler, Sarah Koenig, and Joachim Gerhard Kreuder. Examiner effect on the objective structured clinical exam—a study at five medical schools. **BMC Medical Education**, Vol. 17, No. 1, p. 71, 2017.
- [4] Masaki Uto. A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. **Behaviormetrika**, Vol. 48, No. 2, pp. 425–457, 2021.
- [5] Masaki Uto. Accuracy of performance-test linking based on a many-facet Rasch model. **Behavior Research Methods**, Vol. 53, No. 4, pp. 1440–1454, 2021.
- [6] Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. **Behaviormetrika**, Vol. 47, No. 2, pp. 469–496, 2020.
- [7] Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, and Maomi Ueno. Integration of Prediction Scores From Various Automated Essay Scoring Models Using Item Response Theory. **IEEE Transactions on Learning Technologies**, Vol. 16, No. 6, pp. 983–1000, 2023.
- [8] T Liu, J Luo, H He, J Zheng, J Zhao, and K Li. History-taking instruction for baccalaureate nursing students by virtual patient training: A retrospective study. **Nurse Educ Today**, Vol. 71, pp. 97–104, 2018.
- [9] A.J. Kleinheksel. Transformative Learning through Virtual Patient Simulations: Predicting Critical Student Reflections. **Clinical Simulation in Nursing**, Vol. 10, pp. 301–308, 2014.
- [10] R Dickerson, K Johnsen, A Rajj, B Lok, J Hernandez, A Stevens, and DS Lind. Evaluating a Script-Based Approach for Simulating Patient-Doctor Interaction. **Proceedings of the International Conference of Human-Computer Interface Advances for Modeling and Simulation**, Vol. 1, pp. 79–84, 2005.
- [11] Thomas Parsons, P Kenny, and Albert Rizzo. Virtual human patients for training of clinical interview and communication skills. **International Conference on Disability, Virtual Reality and Associated Technologies**, p. 9, 2008.
- [12] Julia El Zini, Yara Rizk, Mariette Awad, and Jumana Antoun. Towards A Deep Learning Question-Answering Specialized Chatbot for Objective Structured Clinical Examinations. In **2019 International Joint Conference on Neural Networks (IJCNN)**, pp. 1–9, 2019.
- [13] Daniela S M Pereira, Filipe Falcão, Andreia Nunes, Nuno Santos, Patrício Costa, and José Miguel Pêgo. Designing and building OSCEBot® for virtual OSCE - performance evaluation. **Med Educ Online**, Vol. 28, No. 1, p. 16, 2023.
- [14] Kellen R Maicher, Adam Stiff, Marisa Scholl, Michael White, Eric Fosler-Lussier, William Schuler, Prashant Serai, Vishal Sunder, Hannah Forrester, Lexi Mendella, Mahsa Adib, Camille Bratton, Kevin Lee, and Douglas R Danforth. Artificial intelligence in virtual standardized patients: Combining natural language understanding and rule based dialogue management to improve conversational fidelity. **Med Teach**, p. 7, 2022.
- [15] Roth Benjamin Fink Maximilian Rojewiec, Robin. **Intent Recognition in Doctor-Patient Interviews**. No. 702–709. European Language Resources Association, Marseille, France, 2020.
- [16] Han Wei Ng, Aiden Koh, Anthea Foong, and Jeremy Ong. Real-Time Hybrid Language Model for Virtual Patient Conversations. In Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova, editors, **Artificial Intelligence in Education**, pp. 780–785, Cham, 2023. Springer Nature Switzerland.
- [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv:1810.04805**, pp. 4171–4186, 2019.
- [18] Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. **Scientific Data**, Vol. 9, No. 1, p. 313, 2022.
- [19] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. **Proceedings of the 2018 EMNLP Workshop Blackbox(NLP): Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, 2018.
- [20] Yi Sun, Yu Zheng, Chao Hao, and Hangping Qiu. NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task-Next Sentence Prediction. **International Conference on Computational Linguistics**, p. 18, 2022.
- [21] 伴真太郎, 鈴木富雄, 青松棟吉, 西城拓也, 阿部恵子, 桑嶋愛. よくわかる医療面接と模擬患者. 名古屋大学出版会, 2015.

## A 付録

### A.1 ChatGPT の過剰回答

ChatGPT は受験者からの質問では言及されていないシナリオまで情報を参照し、過剰な回答をすることがある。過剰な回答の例と、それらの回答を提案モデルによって検出し、フィードバックすることで改善された回答の例を表 3 に示す。

表 3: ChatGPT の過剰な応答をフィードバック (FB) し、改善された対話サンプル

<b>Doctor :</b>	In the back pain that you are having, does the pain radiate or move to any other part of your body?
<b>Patient(Pre-FB) :</b>	Yes, the pain in my lower back radiates down my right leg. It starts from the lower back and travels all the way down to my foot. The pain can be quite intense and it feels like a shooting or tingling sensation. Sometimes, I also experience numbness and reduced sensation in my lower limbs. It's quite uncomfortable and makes it difficult for me to move around or perform certain activities.
<b>Patient(Post-FB) :</b>	Yes, the pain in my lower back radiates down my right leg.
<b>Doctor :</b>	When did the symptom start?
<b>Patient(Pre-FB) :</b>	The symptoms started three weeks ago when I caught a cold. Initially, it was just a regular cough and phlegm, but it hasn't improved over time.
<b>Patient(Post-FB) :</b>	The symptoms started three weeks ago when I caught a cold.

### A.2 実験設定

本研究では、文献 [21] の SP 用のシナリオを基に、呼吸器、心臓、2つの筋骨格に関する症状をそれぞれの主訴とする 4つのシナリオを作成した。作成したシナリオは、表 4 に示したような患者の年齢、家族背景などのペルソナ情報や、抱えている症状の詳細などのシナリオ文のリストと、医者との質問文のリストが用意されている。医者との質問文は文献 [18] の医療面接を基に作成した。

本研究では、指定したシナリオに沿って回答する ChatGPT として gpt-3.5-turbo を採用した。gpt-3.5-turbo のパラメータの一つである temperature は 0.7 に設定した。ChatGPT には {"role": "system", "content": "j 番目のシナリオ文"} の形式でシナリオ文を一文ずつ指定し、 $j = 1, 2, \dots, J$  まで入力した上でプロン

表 4: ChatGPT に与えられるシナリオの例

1. I'm Yamaguchi Aoi.
2. I'm 50 years old.
3. I have a persistent cough and phlegm since catching a cold three weeks ago.
4. The color of the phlegm is transparent.
5. Two weeks ago, I visited a nearby hospital and was prescribed cold medicine, but it didn't improve my condition even after taking it.
6. Yesterday, I revisited my regular doctor, and they took a chest X-ray, but they couldn't find anything conclusive, so they referred me here for further tests.
7. For the past week, I've been feeling a slight shortness of breath even just climbing stairs at school.
8. At night, when lying down, the symptoms worsen, making it difficult to sleep, but sitting down provides some relief.
9. Six months ago, when I went mountain climbing, I had a similar experience. I felt short of breath and had a cough, but I thought it was due to exhaustion and didn't seek medical attention.
10. My bowel movements haven't changed, but I feel like my urine flow has decreased since catching the cold.
11. My feet are swollen, making it difficult to put on shoes.
12. Until the age of 40, for 20 years, I smoked one cigarette a day.
13. I drink alcohol occasionally, only a small bottle on holidays.
14. My menstrual cycle has become irregular, and the flow has decreased.
15. I initially thought it was just a common cold, but since the cough hasn't improved and I feel slightly short of breath, I'm worried it might be something more serious, like pneumonia.
16. My father passed away from colon cancer, and my mother from uterine cancer. My brother died suddenly last year at the age of 52 from a heart attack.

プトには以下のように指示した。

“You'll act a role of patient. The patient's scenarios are described in detail. Note that the generated response should always begin with an utterance.”