研究指導 中澤 真 教授

GPTを用いた論理破綻を指摘可能な面接練習支援システムの開発と評価

大堀 大翔

1. はじめに

入試や就職活動では選考方法として面接試験が多く用いられるが，面接練習不足によって実力を発揮できないケースもある．株式会社文化放送キャリアパートナーズの調査[1]によれば，就職活動準備においてもっとも不安を感じる要因は「面接対策」であり，回答の36.2%を占める．それは1人での面接練習が難しく，また練習機会が不足していることが理由とされている．そのため1人での面接練習を可能にする支援システムの需要が高まっている．

このような支援システムには，面接官役が果たす2つの機能が求められている[2]．1つは面接練習機能であり、面接官と被面接官（以下ユーザと表記）が質問と回答をやり取りできる必要がある．2つ目は論理判定機能である．経団連の調査[3]によると，採用の観点から特に期待される能力の上位に「論理的思考力」が位置しており，72.1%の企業がこれを挙げている．したがって回答には一定の論理性が求められるため，面接練習時にこの機能により回答が論理的でない場合に指摘できる必要がある．

従来の面接練習用のアプリやサービス[4][5]では，ユーザの論理的でない回答への指摘やアドバイスを得ることができない．清水ら[6]は重要語を使用し，質問に対する回答の論理破綻を検出していたが，論理破綻箇所の指摘や改善策のフィードバックを送ることが難しい．

そこで本研究ではGPT-4[1]のAPI[2]を用いた支援システムに面接練習機能と論理判定機能の2つを組み込み，論理破綻を指摘し，改善案のフィードバックを可能とする面接練習支援システムを開発する．

2. 面接練習支援システムの現状と課題

2.1 面接練習支援システムの概要

入試や就職活動では選考方法として面接試験が多く用いられる．しかし1人での面接練習は難しい．また練習機会が不足していることから被面接者は面接対策に不安を感じている[1]．

そこで1人での面接練習を可能とするために面接官が果たす2つの機能を持った面接練習支援システムが必要である[2]．第一の機能は，面接官の質問とユーザの回答をやり取りする面接練習機能である．この機能ではシステムが面接官役を担い，ユーザに対して投げかけて回答を受け取る．受け取った回答に基づき，関連質問や内容を深堀りする質問を生成する．これによりユーザは実際の面接と同様に，流れのある文章のやりとりを練習できるようになる．第二の機能は，質問文と回答文の対応が論理的であるかを判定する論理判定機能である．あらかじめ決められた基準によってユーザの回答を論理判定し，ユーザへ指摘やアドバイスを返す．これにより人同士の面接練習時のように面接官役からフィードバックを受け取りつつ，回答の論理性を高めることが可能となる．

しかしこれらを実現したシステムは既存では少ない．「steach」[4]や「模擬面接シミュレーター」[5] などの面接練習用アプリやサービスが企業から提供されているが，論理的でないユーザの回答に対して指摘やアドバイスを提供する機能は備えていないという課題がある．これに対応するため，質問に対する回答の論理性を判定可能にする面接練習支援システムに関する研究が進められている．

2.2 面接における論理破綻

面接試験では，被面接者に対して「論理的思考力」が求められており[2]，面接官は被面接者の回答を通じてその論理的思考力を評価している．したがって，評価を高めるための練習が必要となり，論理破綻した場合に指摘を受けられるような練習環境が必要になる．なお，本研究では「質問の主題に逸脱した回答」，「敬語や口語表現が適切に使われていない回答」，「根拠や理由付けまたはエピソードが欠けている回答」のいずれかに該当する回答を，論理破綻したものと定義する．これは後述するGPT[3]の対応力の検証実験から導かれたものである．

清水ら[6]は，回答内の重要語と単語の品詞に着目し，文同士の関係性を推定することで，論理破綻を判定するシステムを提案した．この研究ではTF-IDF[4]を利用し質問文と回答文の重要語を抽出し，重要語同士の類似度によって論理破綻を指摘している．しかし先行研究では論理破綻かどうかの判定のみをしているため，どこが論理破綻しているのか，どう改善すべきかについてユーザはフィードバックを受け取ることができない．

そこで本研究ではGPTを用いて論理破綻箇所の指摘と改善案のフィードバックを可能とするシステムの開発を目指す．

2.3 GPTを用いた面接練習支援システム

本研究ではGPTを用いた面接練習システムを開発する．このGPTモデルにより，人間のようなテキストの生成

[1] https://openai.com/research/gpt-4
[2] プログラムを通じて，ソフトウェアの機能を共有する仕組み
[3] Generative Pre-trained Transformer
[4] 各文書中に含まれる各単語の重要度を表す尺度

や，GPTとはChatGPT[5]などの生成系AIアプリケーションの基礎となっているニューラルネットワークモデルの一種である．会話形式での面接上のやり取りが可能となった．する．GPTとはChatGPTなどの生成系AIアプリケーションの基礎となっているニューラルネットワークモデルの一種である．このGPTモデルにより，人間のようなテキストの生成や，会話形式での面接上のやり取りが可能となった．

本研究のシステムではGPTのAPIを利用し，GPTに想定する面接官の情報，面接練習を行うユーザの情報，面接練習をする上で必要な詳細な条件を入力することで面接官としての役割を与える．これによりそのGPTとの面接練習が可能になる．同様にGPTを介して，面接官役の質問文，ユーザの回答，および論理破綻を判定するための条件を別のGPTに入力することで，面接における対話のやり取りが適切かどうかを客観的に判定する仕組みを構築する．

これらの実装により，既存の面接練習支援システムが直面する課題を克服し，論理破綻箇所の指摘や改善案のフィードバックができる面接練習の実現が可能となる．

2.4 GPT のみの面接練習の現状と課題

GPTを用いた面接練習支援システムを開発するにあたり，面接練習支援システムに求める面接練習機能の対応力が通常のGPTにあるかを検証するための予備実験をした．実験内容は面接官役のGPTの質問に対し，ユーザが実際の面接練習で想定される7つの回答パターンで回答した際のGPTの対応について検証する．

ユーザが回答すると想定した7つの回答パターンを表1に示した．7つのうち適切な回答パターンはNo1と2の2つであり，不適切なものはNo3から7の5つである．GPTの対応としてユーザの回答が適切な場合は指摘せず，話を広げるか次の質問に行くことを正しい対応とし，正しい回答を指摘することを誤った対応とする．ユーザの回答が不適切な場合は，該当箇所の指摘後に同じ質問やアドバイスを送ることを正しい対応とし，指摘せず次の質問に進むことや不適切な回答を広げることを誤った対応とする．実験では各項目について3回ずつ試行して，GPTが正しい対応を行えているかを評価する．

初めはユーザの回答に対してGPTに指摘してほしい項目についての情報を与えず，面接官として自認する条件のみを与え実験した．その実験結果が表1の実験1である．ここではNo.3から7の回答について正しい対応ができないという結果になった．

そこで次の実験2ではGPTにユーザの回答を評価する条件として，実験1で正しい対応ができていなかった「単語のみの回答」，「質問の主題から逸脱した回答」，「敬語や口語表現が使われていない回答」，「根拠や理由付けまたはエピソードがない回答」について指摘するように設定した．結果は表1の実験2の通りである．指摘すべき項目を指示したにもかかわらず，No.5から7のタイプの回答について正しい対応ができないという結果になった．

そこで本研究ではこの3つの論理破綻について指摘することを可能とした面接練習支援システムの開発を目的とする．

表 1 GPT の対応力を検証する実験結果

No.	回答パターン	実験1 正解数	実験2 正解数
1	質問に対して適切な回答	3/3	3/3
2	質問に対して適切な回答 (ローマ字)	3/3	3/3
3	質問と全く関係ない 単語のみの回答	0/3	3/3
4	質問とカテゴリーが似た 単語のみの回答	0/3	3/3
5	質問の主題から逸脱した回答	0/3	0/3
6	敬語や口語表現が 使われていない回答	0/3	0/3
7	根拠や理由付けまたは エピソードがない回答	0/3	0/3

3. 面接練習支援システムの構成

3.1 使用するツール

本研究のシステムはPython3を使用し，Windowsのデスクトップアプリケーションとして開発した．その理由はこのシステムをWindows11搭載のパソコンとインターネット接続ができる環境であれば，場所を選ばずに面接練習を可能とするためである．またシステムの面接練習機能と論理判定機能についてはOpenAI社のGPT API[6]のGPT-4モデルを活用することで実現した．この2つの機能のGPTはそれぞれが独立し，異なる役割を持っているため，以後面接用GPT，論理判定用GPTとする．

3.2 面接用 GPT

面接用GPTには前提条件としてGPTを面接官とする役割を与えるための情報と面接進行の条件を入力し，仮想の面接官と面接練習の場面を準備する．

まず面接用GPTに面接官として機能するための指示文を入力する必要がある．例えば，「あなたは優秀な面接官です．以下の条件に従って面接練習を行ってください．」と入力することでGPTが自身を面接官と自認するようになる．

次に面接進行の条件を面接用GPTに入力する．例えば，「あなたの発言：面接官のみの発言をしてください．」や「質問の頻度：質問はひとつずつ行ってください．私の回答を待ってから再度発言するようにしてください．」など面接練習をする上で常識的なものについても詳しく設定する．これにより，面接用GPTでの面接練習を再現できる．

続いて面接用GPT内の面接官像を具体化するために，面接官の属性情報(国籍，性別，名前，年齢)を入力された条件よりGPTが自動生成させる．ただし，国籍は日本人に限定し，年齢は40歳から60歳の間から選択することを条件としている[7]．また，名前は国籍と性別に基づいて生成することも要求している．

[5] https://chat.openai.com/

[6] https://openai.com/product

さらにGPTが倫理観の欠如した不適切な質問をしないために，面接官の不適切な質問をGPTに設定する．厚生労働省の大阪労働局が発表する「就職差別につながるおそれのある不適切な質問の例」を基に以下の6つの質問を不適切な質問とする[8].

1. 本籍に関する質問
2. 住居とその環境に関する質問
3. 家族の構成や職業・地位・収入に関する質問
4. 資産に関する質問
5. 思想・信条・宗教や支持政党に関する質問
6. 男女雇用機会均等法に抵触する質問

このような要件をあらかじめ入力しておくことで，システムがユーザを不快にさせる質問をすることを未然に防ぐことができる．

最後にユーザが志願する企業の業種や大学の専攻についての情報や志望動機を入力することで，個々のユーザに合わせた面接の質問へとカスタマイズすることが可能になる．これにより，実際の面接を想定した実践的な練習が可能になる．

3.3 論理判定用 GPT

論理判定用GPTは面接用GPTから出力された面接の質問とそれに対するユーザの回答を抽出し，回答の論理性を客観的に判定する．論理判定用GPTには適切な判定を出力させるために様々な情報を入力する．

まず論理性の判定結果の出力形式を設定する．本研究ではJSON[7]形式で，判定評価値と判定理由を出力する．JSON形式にした理由はデータが比較的軽量で，テキストデータを処理する際にも改行やカンマなどの影響を受けないからである．

次に論理性判断の評価値（以下判定評価値と表記）を5段階で設定する．判定評価値が4または5の回答を適切に回答できているものし，1から3の評価値となった回答を不適切とする．不適切な回答に対しては，判定理由を表示することで，改善すべき方向性をユーザに示唆する．

最後に予備実験の結果に基づき，論理判定すべき項目を表 2のように設定する．これらの情報をあらかじめ論理判定用GPTに入力することで適切な論理破綻の指摘が可能になる．

表 2 判定項目一覧

判定項目	判定内容
①	質問に対する回答が丁寧語や敬語を 使用した口語表現での文章となっているか
②	質問に対する回答が質問の主題に答えているか
③	質問に対する回答が根拠や理由またはエピソード を含む説得力のある文章であるか

続いて，論理判定用GPTの動作について図 1に示す．

まず面接官の質問とそれに対するユーザの回答のペアを入力する(ア)．次に入力された二つの文章を基に論理判定用GPTがユーザの回答について判定項目①から客観的に評価する(イ)．判定評価値が1から3の場合は，不適切な回答と判定して，同じ質問を表示して再回答を促す(カ)．判定評価値が4または5の場合は，適切な回答と判定して次の判定項目に遷移する．判定項目②，③についても①と同様の手順で進める(ウエ)．最終的に，質問に対する回答が全判定項目で4から5の評価値を得られた場合には，その回答を面接用GPTへ送信する(オ)．その後，面接用GPTは受け取った回答に基づき次の質問を生成し，面接練習機能の動作を継続する．なお，判定項目2や3に進んだ場合であっても，再回答の内容が一度通過した判定項目を満たさなくなる可能性を考慮し，すべての判定を項目1から再度行う．

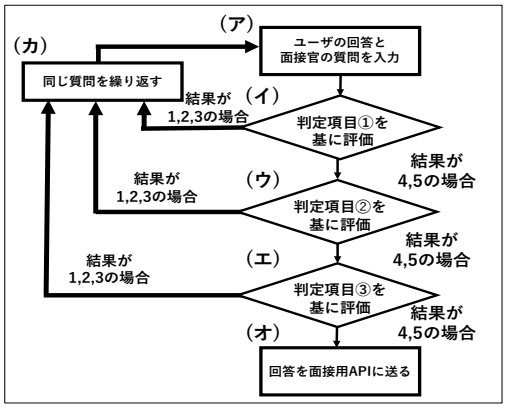

図 1 論理判定用 GPT の動作

3.4 システム構成

本研究のシステム構成を図 2に示す．まず面接用GPTに質問を生成させるための前提条件を入力する(A)．面接官の情報やユーザの情報，就職や進学を目指している企業や大学の情報などを前提条件として入力することでシステムの個別最適化を可能とした．次に面接用GPTから出力された面接の質問ユーザに表示し，回答を促す(B)．ユーザが入力した回答を受け取った後，その回答とそれに対応する質問のペアを論理判定用GPTに送る(C)．そして論理判定用GPTから出力された判定評価値に基づき判定する(D)．判定評価値が1から3であれば，同じ質問と指摘内容をユーザに提示し，再度回答を求めて③に戻る．すべての判定項目の評価値が4または5であれば，ユーザの回答を面接用GPTへ渡して次の質問へと遷移する．

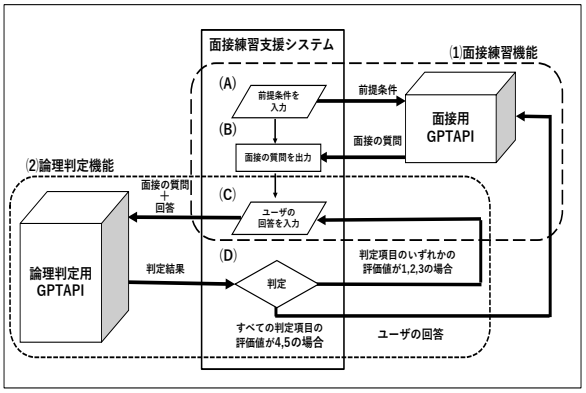

図 2 システム構成

[7] 軽量なテキストベースのデータ交換用フォーマット

4. 論理判定用 GPT の検証と考察

4.1 検証実験の手順

本研究における論理判定用GPTの判定精度を検証する. まず質問文と回答文のデータセットを100組用意する. データセットの構成は, 質問文が10パターン, 回答文が質問に対する回答を判定する項目を分散させた10通りの回答の組み合わせとする. 精度検証に必要な質問文と回答文は就活生のためのサイト「賢者の就活」[9]を参考に作成した.

次にすべてのセットに対し, 回答が質問に対して適切であるかを3種類の判定項目それぞれについて, 手動で判定評価値をラベル付けした. このラベルが正解値となる.

最後に質問文と回答文100組のデータセットを読み込み, 先述の項目によって判定する. なお, 1組のセットにつき判定項目①から③によって回答を判定するためデータ数は300となる. 繰り返し処理によってすべての判定結果とその理由を出力するように組みなおし, 新しく精度検証用プログラムを作成した.

4.2 検証結果と考察

実際に精度検証用プログラムを実行し, 出力結果を分析する. GPTが出力した判定結果とそれぞれのデータに付与した正解値の組み合わせのクロス集計表を

表 3に示した. さらに, 正解率を以下の式にしたがって算出する.

表 3 システムの全判定項目の論理判定結果

正解値 / システム	(1-3)	(4-5)
出力(1-3)	117	42
出力(4-5)	33	108

$$(\textit{全体の正答率}) = \frac{(\textit{システムの出力と正解値とが一致する回答数})}{(\textit{全体の回答数})}$$

検証の結果, システム全体の正答率は0.750, 判定結果1-3の正答率は0.780, 判定結果4-5の正答率は0.720となり, ある程度の水準の正答率を達成できた.

次に先と同様に判定項目ごとの正答率を求め, 各判定項目による正答率の偏りを確認する. 判定項目ごとの正答率は, ①が0.800, ②が0.920, ③が0.530という結果になった.

この結果から①の「質問に対する回答が丁寧語や敬語を使用した口語表現での文章となっているか」と②の「質問に対する回答が質問の主題に答えているか」について, どちらも高い精度で指摘が可能であることがわかる. 一方で③の「質問に対する回答が根拠や理由またはエピソードを含む説得力のある文章であるか」については正答率が6割に満たず, 指摘することが難しいという結論に至った.

判定項目③の正答率が低い原因として, 文章中の理由や根拠として本来適さないユーザの感想や共感についても, GPTが同様に捉えてしまうことが挙げられる. そのため, 判定項目③の指摘を可能にするには, 感想や共感のような感情に基づく内容を理由や根拠とみなさないようにし, これらの定義をGPTへ詳細に設定することで改善できる可能性がある.

また本研究では論理破綻箇所の指摘やそれに対する改善案をGPT-4によって生成し, ユーザに送ることで先行研究が直面していた課題を克服することが可能となった.

5. むすびに

本研究では, 2種類のGPTを連携させて論理破綻を指摘する面接練習支援システムの開発をした. 3つの判定項目による全体の論理破綻の判定として0.750の正答率を達成した. これにより, 既存のシステムと同程度の精度を保ちつつ, ユーザに論理破綻箇所の指摘や改善案のフィードバックができる面接練習システムを実現した. なお, 判定項目③の内容については, GPTに条件を入力する際の情報に感情に基づく内容について提示することで正答率の向上を図ることができる. これを実現することを今後の課題とする.

参考文献

[1] 株式会社文化放送キャリアパートナーズ, “就活準備の不安、面接対策が 3 割超え”, https://prtimes.jp/main/html/rd/p/000000015.000090419.html, (参照:2023-2-5).

[2] 竹内将人ほか, “対話ロボットを用いた面接練習でのフィードバックによる心理的影響の検討”, マルチメディア,分散,協調とモバイルシンポジウム 2022 論文集, vol.2022, p364-368, 2022.

[3] 日本経済団体連合会, ”採用と大学改革への期待に関するアンケート結果”, https://www.keidanren.or.jp/policy/2022/004_kekka.pdf , (参照 2024-02-05).

[4] 株式会社ジェイック, “面接練習アプリ steach, https://jaic-steach.jp/, (参照 2024-02-01).

[5] マイナビ, ”模擬面接シミュレーター”, マイナビ 2024, https://job.mynavi.jp/conts/2024/mensetsu/index_m.html , (参照 2024-02-01).

[6] 清水康平ほか, “実用的な模擬面接システムのための質疑応答における論理破綻検出”, 信学技報, vol.118, no.408, p.63-68, 2018.

[7] 採用係長, ” 中小企業における採用業務の体制に関する実態調査”, 採用係長の採用アカデミー, https://saiyo-kakaricho.com/wp/survey_recruitment_structure/, (参照 2024-02-06).

[8] 厚生労働省大阪労働局, ”就職差別につながるおそれのある不適切な質問の例”, https://jsite.mhlw.go.jp/osaka-roudoukyoku/hourei_seido_tetsuzuki/shokugyou_shoukai/hourei_seido/kosei/futeki.html, (参照 2024-02-01).

[9] 賢者の就活, ”質問の的確な答え方と回答例 50 選”, 合同会社アイプレス, https://syukatsu-kaigi.jp/dashboard, (参照 2024-02-01).

[10] 掌田津耶乃, Python/JavaScript による Open AI プログラミング, ラトルズ, 2023

[11] えんぞう, ”ChatGPT、GPT-4 の API リクエストパラメータ”, Zenn, https://zenn.dev/en2enzo2/articles/cc90b56c80e3f3, (参照 2024-02-01).

[12] 藤野, ”ChatGPT の API を Python から使う”, 藤の手帳, https://fuji-pocketbook.net/chatgpt-api-python/#toc, (参照 2024-02-01).