



NTT西日本における新規サービス創出・競争力強化を見据えたAI技術の研究開発についての取り組み——面接サポートAI開発

近年、リスクリングの普及を背景として、AI（人工知能）を活用した人材評価業務支援へのニーズが高まっています。NTT西日本バリューデザイン部のテックデザイン担当では、専門的な技術スキルをヒアリングし、評価者へ参考情報を提供する面接サポートAIの研究開発を行い、その有効性を検証しました。本稿では、面接サポートAIの研究開発内容および質問の網羅性・具体性を、AIどうしの模擬面接や利用者アンケートをとおして評価した取り組みについて紹介します。

面接サポートAI研究開発施策についての背景

NTT西日本バリューデザイン部は、ビジネス・営業戦略に基づくサービス開発や製品主管などを担っており、その中でも当テックデザイン担当では、研究開発組織としてAI（人工知能）を中心とした最新技術の調査・検証やシステムの試作・開発に取り組んでいます。

このような研究開発の1つとして、今回着目したのが人材のスキル評価にかかわる業務です。近年、DX（デジタルトランスフォーメーション）の進展やリスクリングの普及により、企業には必要な人材を獲得するだけでなく、社員が保有するスキルや専門性を適切に把握し、育成や配置、キャリア形成につなげていくことが求められています⁽¹⁾。そして、このような社員の専門的スキルの把握・評価の流れは、NTT西日本における人材育成の方向性とも合致しています。お客さまや社会からのニーズが高度化・複雑化する中、当社では7つの重点領域を設定し、各領域で高度な専門性を発揮できる人材を計画的・持続的に育成・評価する仕組みを整備しています。その一環として、社員の専門知識や技術力を一定の基準で確認し、基準を満たした社員を認定・評価する「高度スキル人材認定制度」の運用を開始しています⁽²⁾。

こうした環境の変化や制度の整備を踏まえると、人材育成や専門スキルの評価など、人材に関するヒアリング・評

価業務をAIによって支援するニーズは、今後さらに高まると考えられます。実際に、当社内の人事育成関連部署へ各種面接業務での課題をヒアリングしたところ、面接負荷の軽減や面接評価のサポートに対するニーズを把握することができました。

そこで当担当では、人材に関する専門スキルのヒアリング・評価業務へのAI活用の可能性を検証するため、専門的な技術スキルのヒアリング・評価を支援する「面接サポートAI」（本AI）*の研究開発に取り組みました。

面接サポートAI研究開発内容

本AIを開発するにあたり、まずは一般的な面接のプロセスを整理しました。一般的な面接では、応募者の適性や能力を客観的に評価するため、あらかじめ「評価項目」と各評価項目に対応する「評価基準」を設定してから評価する方法が推奨されています⁽³⁾。「評価項目」は、応募者に求める知識・技術・経験（例：〇〇における知識、□□の業務経験）などを複数の観点に分けて設定したものであり、「評価基準」は、それぞれの評価項目について合格に必要な水準（例：〇〇について××まで理解している、□□の業務を×年以上経験している）を示したものです。そして、応募者は面接前に、これまでに経験した案件や実績、保有スキルなどを記載した調書を提出します。面接官は調書の内容を確認し、その内容を踏まえて質疑応答を行ったうえで、各評価項目について評価基準を満たしているかを判断するという流れになります。

加えて、面接において応募者の専門的なスキルや経験を

* 本AIは認定結果や可否を自動判定するものではなく、評価者が判断を行う際の参考情報を生成することで、ヒアリングおよび評価業務を支援することを目的としています。また、各種面接における専門スキル評価業務への適用可能性を検証している段階のものであり、現時点で当社における面接プロセスに導入しているものではありません。

適切に確認・深掘りする際には、STAR法という手法が広く用いられています。この手法は、各Situation（状況）において、Task（課題・役割）・Action（行動・工夫点）・Result（結果・成果）という観点で質問を行い、応募者のスキル・経験を具体的に深掘りしていくものです。当社内の面接官へのヒアリングでも、同様の考え方で面接が行われていることが確認されました。そこで本AIでは、これらの面接手法を参考に、「各取り組みにおける自身の役割」「工夫点」「結果」（3基準）の有無を、応募者からの回答の具体性確認および深掘り質問を実施する際の基準として設定しました。本AIが3基準を確認しながら面接を行うことで、応募者の回答内容に不足する情報を、より実際の面接に近いカタチで引き出す深掘り質問の生成ができるようにしました。

以上を踏まえて検討した本AIによる面接システムの構成を図1に示します。本AIによる面接では、各評価項目・評価基準と、応募者が提出した調書を入力したうえで、各評価項目についての質問を順次行います。応募者から回答が得られると、本AIが3基準の観点から、具体的な情報が十分に含まれているかを確認します。不足がある場合には、その情報を補うための深掘り質問を生成し、十分な情報が得られている場合には次の評価項目へ移行します。これにより、実際の面接同様に、回答内容に応じて必要な情

報を深掘り質問で確認しながら、各評価項目を漏れなく確認できる仕組みとしました。

すべての質問が完了すると、調書と面接時の回答内容を基に面接結果をまとめるとともに、回答内容のうち良かった点と、より高い評価につなげるための応募者へのアドバイスを生成し、評価者による面接評価を支援します。

面接サポートAIの検証内容

本AIのヒアリング性能を検証するにあたり、実際の面接と同様に、評価項目を確認しながら、評価に必要な具体的情報を引き出せるかを確認する必要がありました。そこで、「評価項目を漏れなく確認できているか」を「質問の網羅性」、「評価に必要な具体的情報を引き出す質問ができているか」を「質問の具体性」と定義し、この2点を評価軸として設定しました。そして、これら2点を下記「検証① AIどうしの模擬面接による評価」と「検証② 技術スキル認定面接経験者による評価」の2つの方法で検証しました。加えて、本AIが生成する質問やアドバイスに、年齢・性別などの属性に基づく不公平な偏りが含まれていないかを確認するため、「検証③ BBQに基づくバイアス評価」を実施しました。

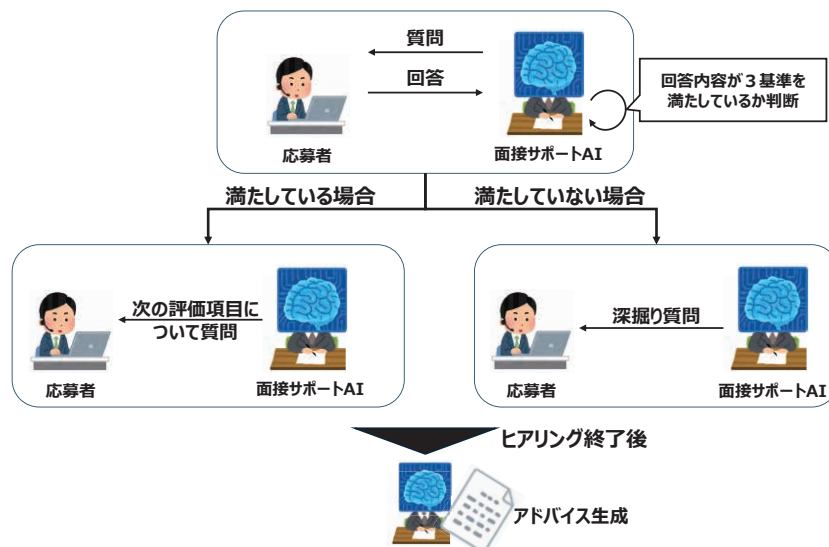


図1 本AIによる面接の構成（イメージ）

■検証① AI どうしの模擬面接による評価

本AIに加えて応募者役のAIを作成し、AIどうしによる模擬面接を実施しました（図2）。応募者役には3基準の情報を与え、すべての質問に対して3基準を「満たす」回答を生成するAIと、「満たさない」回答を生成するAIの2種類を用意し、それぞれ本AIとの模擬面接を100回ずつ実施しました。得られた対話ログは、各評価項目と3基準を入力した評価用AIを用いて分析しました。「質問の網羅性」については、各模擬面接ですべての評価項目について質問できていたかを確認しました。そして、「質問の具体性」については、3基準を満たす回答に対して不要な深掘り質問を行っていないか、反対に3基準を満たさない回答に対して必要な深掘り質問を行えているかを確認し、その回数を集計しました。これにより、本AIが回答内容の具体性を適切に判断したうえで、必要に応じて具体的な情報を引き出す深掘り質問を実施できているかを定量的に評価しました。

■検証② 技術スキル認定面接経験者による評価

AIどうしの検証だけでは把握しにくい質問への違和感や

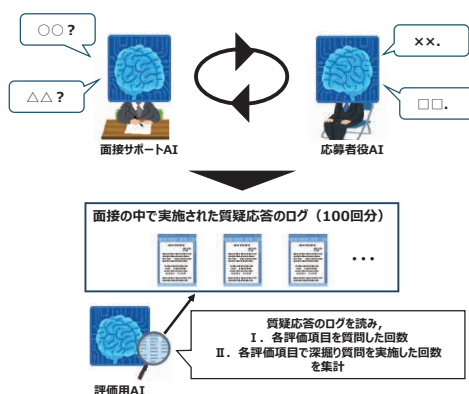


図2 AIどうしによる模擬面接の実施内容（イメージ）

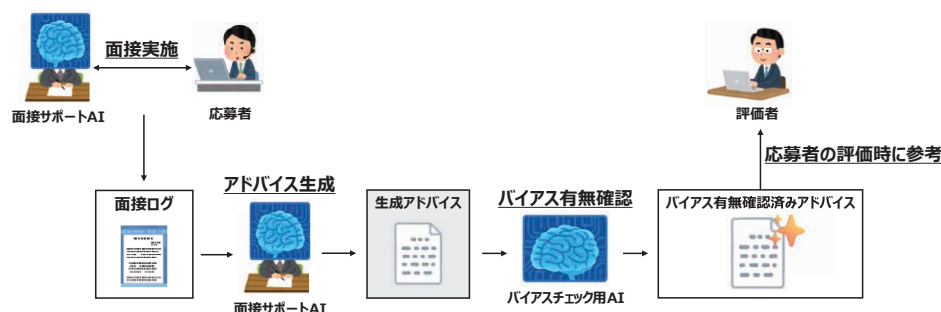


図3 生成アドバイスにおけるバイアス有無の評価（イメージ）

実利用上の課題を確認するため、当社内での技術スキル認定を行う面接の経験者である高度スキル人材認定の合格者10名に本AIとの面接を体験してもらい、アンケートを実施しました。アンケートでは、実際の認定面接との比較を踏まえ、「質問の網羅性」「質問の具体性」「本AIの有用性」の3項目を10点満点で評価してもらおうとともに、自由記述による意見も収集しました。これにより、利用者の観点から本AIを評価するとともに、検証①の結果との整合性についても確認しました。

■検証③ BBQに基づくバイアス評価

本AIが生成する内容について、年齢、性別、人種、身体的特徴などの属性に基づく偏見やステレオタイプが含まれていないかを確認するため、Bias Benchmark for QA⁽⁴⁾（BBQ）に基づくバイアス評価を実施しました（図3）。具体的には、BBQで定義されている9カテゴリを評価観点としてバイアスチェック用AIに与え、本AIが生成した内容について各カテゴリに該当する偏った表現や不適切な記述の有無を判定させました。これにより、倫理面・公平性の観点から、バイアスチェック用AIを、本AIが生成する質問やアドバイスのチェック機能として利用できる可能性を確認しました。

検証結果と考察

■検証① AI どうしの模擬面接による評価の結果

模擬面接の質疑応答を評価用AIに読み込ませ、各質問回数を集計した結果を表1に示します。「質問の網羅性」を確認するため、応募者役AIが3基準を満たす回答を生成

表 1 模擬面接における各評価項目についての質問回数

評価項目	3基準を「満たす」回答を生成するLLMでの面接		3基準を「満たさない」回答を生成するLLMでの面接	
	I の回数	II の回数	I の回数	II の回数
評価項目A	101	0	100	97
評価項目B	100	0	100	91
評価項目C	100	0	100	98
評価項目D	100	0	99	85
評価項目E	99	0	100	99
評価項目F	100	0	101	101
評価項目G	100	0	100	89
			平均	94.3

した場合の質問回数を確認したところ、評価項目Eは99回、評価項目Aは101回となりました。これは、100回の面接のうち99回では評価項目Eを質問できていた一方、残る1回では評価項目Eを質問せず、評価項目Aを重複して質問したことを示しています。また、3基準を満たさない回答を生成した場合の評価項目D・Fについても同様の傾向が確認されました。これは、評価項目A・E間およびD・F間には共通する語句が含まれていたことから、本AIが質問生成時に類似する評価項目を混同したことが原因と考えられます。このため、各評価項目を明確に識別できる固有の語句をプロンプトに含めるなど、項目間の区別を強化することで、質問漏れや重複を抑制できると考えられます。

次に、「質問の具体性」を確認したところ、応募者役AIが3基準を満たす回答を生成した場合には、100回すべての面接で不要な深掘り質問は行われませんでした。一方、3基準を満たさない回答を生成した場合には、各評価項目で平均約94%の割合で深掘り質問が実施されました。深掘りが行われなかった約6%のケースでは、本AIが3基準を満たしていない回答に対しても、回答に具体性があると判断していました。対話ログを確認すると、3基準のうち一部のみが含まれている回答を、本AIが具体的であると判断するケースが確認されました。このことから、3基準の充足状況に対するAIの判断に一定のばらつきが生じたものと考えられます。今後、3基準それぞれの充足状況を個別かつ明確に確認するようプロンプトを具体化することで、判断のばらつきを抑え、深掘り質問の要否をより安定して

判定できると考えられます。

以上から、本AIは「質問の網羅性」では99%、「質問の具体性」では平均94%以上の割合で想定した面接フローを実現できた一方、類似する評価項目の識別や、回答の具体性に対するAIの判断精度には改善の余地が確認されました。

■検証② 技術スキル認定面接経験者による評価の結果

経験者によるアンケート結果を表2に示します。「質問の網羅性」は10点満点中平均7.2点とおおむね良好な評価でした。一方、「質問の具体性」は平均6.4点と比較的低く、改善の余地が確認されました。

本AIでは、3基準を中心に深掘り質問を生成していますが、実際の面接担当者は、3基準に加えて、説明の一貫性や過去の回答との整合性、技術的な内容の妥当性など、応募者の回答状況に応じて、より柔軟に複数の観点を組み合わせながら質問していると考えられます。このように、実際の面接では3基準以外にも本AIへ十分に反映できていない判断要素が存在するため、本AIによる質問内容に一部機械的な印象や違和感が生じた可能性があります。したがって、具体性の判断基準や深掘り質問の生成方法については、実際の面接担当者が回答内容に応じて用いる多様な観点をさらに整理し、反映する余地があると考えられます。一方、「本AIの有用性」は平均8.2点と高く、専門技術スキルのヒアリング・評価を支援するAIについて一定の有用性が示されました。また、自由記述においても、深掘りが不十分であることや質問が機械的であるとの指摘があった一方、質問の網羅性や本AIの有用性を評価する意見が確認されました。

■検証③ BBQに基づくバイアス評価の結果

バイアスチェック用AIを用いて、本AIが出力した各アドバイスに対し、BBQの9カテゴリに基づくバイアスチェックを実施したところ、評価対象とした項目について顕著なバイアスは検知されませんでした。さらに、判定性能を確認するため、評価結果に「若いわりに」など属性に関する偏った表現を意図的に混入したところ、当該表現をバイアスとして適切に検出できました。一般に、現在利用されているAIは偏った意見が含まれるような不適切な表現は生成しにくく、そのような表現の出力を指示しても不適切であ

表 2 アンケートにおける質問内容と回答

No.	評価項目	質問内容	平均評価点数 (10点満点) ※④は記述内容の抜粋
①	質問の網羅性	面接サポートAIからの質問は実際の面接の内容を網羅していましたか.	7.2点
②	質問の具体性	専門的な技術スキルをヒアリングするという観点で、面接サポートAIからの質問は適切だと思いましたか.	6.4点
③	面接AIの有用性	専門的な技術スキルをヒアリングする面接の受け答えの練習用として、今回の面接サポートAIは有用だと思いましたか.	8.2点
④	自由記述欄	面接サポートAIを利用した感想や改善点等をご自由に記入ください.	<p>【肯定意見】</p> <ul style="list-style-type: none"> ・ 調書内容については一通り確認してくれた ・ 的確に要点の確認がされていると感じた ・ 面接の練習としては非常に有用だと感じた <p>【改善意見】</p> <ul style="list-style-type: none"> ・ 本番面接と比べて、具体的なスキル/技術知識については、あまり踏み込まれなかった印象 ・ 深掘質問を通じて当人の考えていることまでを反映させるには課題があると感じた ・ 規定質問に対し忠実に遂行している感じが強い

る旨を返す場合があるため、本AIが出力するアドバイスにも、そのような表現が含まれてしまう可能性は低いと考えられます。そのうえで、万が一アドバイス内容に偏った表現が含まれた場合でも、本手法により検知できる可能性が示されました。

今後は、BBQの観点に加えて、利用シチュエーション独自のNG表現や評価上避けるべき表現を判定条件に追加することで、評価結果の公平性・信頼性を補完するチェック機能としての活用も期待できます。

まとめ

本取り組みでは、専門的な技術スキルのヒアリング・評価の支援を目的とした面接サポートAIを開発し、AIどうしによる模擬面接と面接経験者による評価を通じて、評価項目を網羅した質問や、回答内容に応じた深掘り質問の有効性を確認しました。また、バイアスチェック用AIによるチェックを実施し、AIが生成する評価結果の公平性を確認しました。一方、類似する評価項目を混同して質問する場合や、回答の具体性を十分に判定できず必要な深掘りが行われない場合があり、利用者評価からも深掘り質問の内容

や自然さに改善の余地が確認されました。これらについては、評価項目や判断基準を明確に識別するプロンプトの工夫に加え、実際の面接官が用いる判断基準や着眼点をさらに反映することが、ヒアリング精度の向上に有効と考えられます。バイアス評価についても、評価対象や判定条件を拡充することで、より詳細な公平性・信頼性の確認が可能です。

本取り組みで得られた、AIによる対話制御、専門情報を踏まえた質問生成、回答評価およびバイアスチェックに関する研究開発の知見は、専門的・技術的な情報を扱う新たな対話型AIサービスの検討・開発にも活用されており、NTT西日本におけるAI活用による新たなサービス創出に貢献しています。

■参考文献

- (1) https://www.meti.go.jp/policy/it_policy/investment/dgc/dgc3.0.pdf
- (2) https://www.ntt-west.co.jp/sustainability/wellbeing/human_resources_strategy/
- (3) https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/koyou_roudou/koyou/newpage_56780.html
- (4) A. Parrish, et al.: "BBQ: A Hand-Built Bias Benchmark for Question Answering," ACL 2022, Dublin, Ireland, May 2022.

◆問い合わせ先

NTT西日本
バリューデザイン部 システム開発部門 テックデザイン担当