電気通信大学情報理工学域
先端工学基礎課程卒業論文

多肢選択式問題における LLM による問題解答過程を 組み込んだ Deep-IRT

2026 年 2 月 27 日

先端工学基礎課程

学籍番号 2120031

廣澤 愛美

指導教員 植野 真臣

令和７年度 先端工学基礎課程卒業論文概要

令和 3 年度　入学	学籍番号　2120031
指導教員　植野 真臣	氏名　廣澤 愛美
題目　多肢選択式問題における LLM による問題解答過程を組み込んだ Deep-IRT	

概要

e テスティング（e-testing）は，ISO/IEC 23988 において，異なる問題項目から構成されるテストであっても同一の測定精度を実現できるコンピュータベーステスト（Computer Based Testing: CBT）として定義されている．この特性により，受検者が異なるテストセットを受検しても，能力値を同一尺度上で公平に比較することが可能となる．この仕組みを支える項目反応理論（Item Response Theory: IRT）は，受検者の潜在能力と項目の特性を独立に推定する理論であるが，推定において能力分布に標準正規分布を仮定するため，実際の受検者集団の偏りを反映できず，予測精度に限界があるという課題が指摘されている．

これに対し，統計的な分布の仮定に依存しない深層学習に基づく手法が提案されている．Tsutsumi らの Deep-IRT は，受検者の潜在能力と項目特性をそれぞれ独立したネットワークを用いて推定することで，パラメータの不変性と解釈性を維持しつつ高精度な予測を実現した．しかし，既存の Deep-IRT 手法は主に正誤の 2 値反応を対象としており，多肢選択式問題（MCQ）における「どの誤解に基づいてどの誤答選択肢を選択したか」という情報を活用できていない．これを解決する統計モデルとして名義反応モデル（Nominal Response Model: NRM）が存在するが，NRM も IRT ベースのモデルであるため，依然として能力分布の仮定に関する制約が残る．

そこで本研究では，Deep-IRT を多肢選択式モデルへと拡張し，能力分布の仮定に依存せず多値の反応予測を可能とする手法（MCD-IRT）を提案する．本手法では，大規模言語モデル（LLM）を用いて問題解答過程や誤答選択理由などのテキスト特徴量を抽出し，これを項目特性として組み込むことで予測精度の向上を図る．評価実験では，実際の反応データを用いて選択確率の予測精度を従来手法と比較し，提案手法の有効性を検証する．さらに，その結果の応用として項目の難易度推定を行い，教育評価における本手法の有用性を示す．

目次

1　まえがき

Web 上でテストを実施することを「e テスティング（e-testing)」と呼ぶ．e テスティングは，ISO/IEC 23988 において，異なる問題項目から構成されるテストの場合でも，同一の測定精度を実現できるコンピュータベーステスト（Computer Based Testing: CBT）として定義されている [1]．この特性により，同一の能力を持つ受検者が異なるテストを受検しても，受検者の能力値を同一尺度上で公平に比較することが可能となる [2, 3]．この仕組みを実現しているのが項目反応理論（Item Response Theory: IRT）である．IRT は，受検者の潜在能力と各項目の特性を独立に推定し，正答確率を予測するテスト理論である [4, 5].

しかし，従来の IRT では能力分布に標準正規分布を仮定するため，実際の受検者集団の分布の偏りや特定の能力層への集中を反映できず，予測精度に限界が生じる課題があると指摘されている．現実の受検者は単一の母集団からランダムにサンプリングされるとは限らず，学校やクラス単位のサンプリングでは分布に偏りが生じやすいためである [6].

そこで Yeung らは IRT の枠組みに深層学習を融合させ，潜在能力と項目特性（難易度や識別力）を，ニューラルネットワークを用いて推定する Deep-IRT を提案した [7]．深層学習は統計的な分布の仮定に依存しないため，従来の IRT と比較すると高い予測精度を示すことが報告されている．しかし，Yeung らのモデルでは潜在能力を反応データと項目の潜在変数の双方を組み合わせて推定していたため，潜在能力の推定値が特定の項目の特性に依存してしまい，推定された値の解釈が困難であった．そこで，Tsutsumi らは受検者の潜在能力と項目特性をそれぞれ独立したネットワークを用いて推定する手法を提案し，パラメータの不変性と解釈性を向上させた [8, 9, 10, 11].

一方で，Tsutsumi らが提案した既存の Deep-IRT 手法は主に正誤の 2 値反応を対象としており，多肢選択式問題（Multiple-Choice Questions: MCQ）における誤答選択肢が持つ情報を活用できていないという課題がある．その問題を解決するモデルとして，Bock による名義反応モデル（Nominal Response Model: NRM）がある [12]．これは MCQ において，どの誤答選択肢を選択したかという情報を活用するモデルであり，誤答選択肢の選択パターンにも受検者の能力に関する情報が含まれていると考え，潜在能力に応じた各選択肢の選択確率をモデル化している．そのため，「能力が低い受検者ほど特定の誤解に基づいた誤答を選びやすい」といった特性の記述を可能にしている．しかし，NRM は解答データの統計的な傾向から選択肢の性質を推定するモデルであり，選択肢の内容が持つ言語的な意味情報を解析に含めることはできない．また，NRM は各選択肢に対して独

立したパラメータを割り当てるため，2 値反応モデルに比して推定すべきパラメータ数が大幅に増加する．DeMars は，この選択肢数の増大がパラメータ推定値の誤差分散を増大させ，安定した推定には大規模なサンプルサイズを要することをシミュレーションを通じて示している [13]．こうした制約により，反応データが十分に存在しない新規項目に対して，各選択肢の特性まで含めた項目特性を把握することが困難であるという課題が生じる．

これに対し，近年の自然言語処理（NLP）技術の発展により，解答データに依存せずテキスト情報から項目の特性を直接評価するアプローチが可能となった．Benedetto ら [14] は，問題文だけでなく選択肢のテキスト情報をモデルに入力することで予測精度が向上することを示しており，選択肢の内容そのものが項目特性を決定づける重要な情報源であることを明らかにしている．さらに Feng らは，問題テキストだけでなく，問題解答過程やテキスト特徴量が各選択肢の正答確率と関係していると考えた [15]．彼らは，擬似的な人間の思考過程を特徴量とすることで，受検者の反応の予測精度が向上することを示した．しかし，問題解答過程のような詳細な記述はテキスト長が長くなる傾向があり，通常の Transformer モデルでは計算コストやトークン長の制限といった課題がある．そこで Feng らは，長文のコンテキストを効率的に処理可能な Longformer[16] を採用している．

そこで本研究では，Tsutsumi らによる「反応データに基づく Deep-IRT」を基盤とし，これを大規模言語モデル（LLM）による言語的特徴抽出機能によって拡張した統合モデルを提案する．本手法は受検者ネットワークにおいてどの誤答を選んだかの反応データを用いた深層学習を行うことで統計的分布に依存せずに受検者能力を推定し，NRM よりも受検者の能力推定精度を向上できることを期待する．

さらに提案手法は項目ネットワークにおいて各選択肢の特徴を捉え，予測精度を向上させるためにテキスト特徴量を用いる．具体的には，LLM を用いて問題文および各選択肢から，正答の問題解答過程および誤答選択理由を説明するテキストを生成する．次に，生成されたこれらの長文テキストから Longformer を用いて，多次元の埋め込みベクトルを抽出し，これを項目の特徴量として扱い，この特徴量を，推定された潜在能力と合わせて多層パーセプトロン（MLP）へ入力する．

深層学習による受検者能力と，選択肢ごとの問題および各選択肢のテキスト特徴量を予測に活用することで，十分な反応データ量がない状況においても，未知課題への反応予測精度の向上が期待できるという仮説を立てた．評価実験により提案手法が従来手法よりも選択確率の予測精度が向上したことを示す．

2　多肢選択式問題における項目反応理論

2.1　項目反応理論 (IRT)

項目反応理論（Item Response Theory: IRT）[4, 5] は，異なる項目から構成されるテストにおいても，受検者の能力を同一尺度上で評価することを目的とした，解釈性の高い測定モデルである．IRT にはいくつかの代表的なモデルがあるが，まず基本的なモデルとして 2 パラメータロジスティックモデル（2-Parameter Logistic Model: 2PLM）が挙げられる．2PLM では，潜在能力 $\theta_i \in (-\infty, \infty)$ をもつ受検者 i が項目 j に正答する確率を次式で表す．

$$P_j(\theta_i) = P(u_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp(-1.7a_j(\theta_i - b_j))} \tag{1}$$

ここで，$a_j \in [0, \infty)$ は項目 j の識別力パラメータであり，受検者の能力差をどの程度識別できるかを表す．また，$b_j \in (-\infty, \infty)$ は項目 j の難易度パラメータである．1.7 はロジスティック関数を正規分布に近似するためのスケーリング定数である．

さらに，多肢選択式問題においては，能力の低い受検者が偶然正答する可能性（当て推量）を考慮した 3 母数ロジスティックモデル（3-Parameter Logistic Model: 3PLM）も広く用いられる．3PLM では，受検者 i が項目 j に正答する確率を次の式で定義する．

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + \exp(-1.7a_j(\theta_i - b_j))} \tag{2}$$

ここで，$c_j \in [0, 1]$ は当て推量パラメータであり，能力が極めて低い受検者が偶然正答する確率を表す．なお，この当て推量の可能性を考慮しない場合（$c_j = 0$）が，上述の 2PLM に相当する．

IRT モデルでは推定において潜在能力が標準正規分布に従うという強い仮定を置くことが多く，実際の受検者集団の分布の偏りや特定の能力層への集中を反映できず，予測精度に限界が生じる可能性も指摘されている [6]．

さらに, これらのモデルは反応を正誤の 2 値として扱うため, 多肢選択式問題（Multiple-Choice Questions: MCQ）における誤答選択肢が持つ情報を活用できないという制約がある．Thissen らは，反応を 2 値に集約することは情報の損失を招くと指摘しており，どの誤答を選んだかという情報は，能力推定の精度向上や項目の質的分析において極めて重要である [17, 12]．したがって，こうした誤答の背後にある受検者の思考プロセスを詳細

に扱うためには，多値の反応を許容するモデルが必要となる．

2.2 名義反応モデル (NRM)

2PLM が正誤の 2 値データを対象とするのに対し，Bock が提案した名義反応モデル（Nominal Response Model: NRM）[12] は，多肢選択式の誤答選択肢など，順序性のない 3 つ以上の選択肢からなる反応を扱うためのモデルである．NRM では，能力値 θ_i をもつ受検者 i が，項目 j において選択肢 k $(k = 1, 2, \ldots, K)$ を選択する確率を次式で表す．

$$P_{j_k}(\theta_i) = \frac{\exp(a_{j_k}\theta_i + d_{j_k})}{\sum_{h=1}^{K} \exp(a_{j_h}\theta_i + d_{j_h})} \tag{3}$$

ここで，a_{j_k} は項目 j の選択肢 k における識別力パラメータ，d_{j_k} は選択肢の選択されやすさを規定する切片パラメータである．

モデルのパラメータを特定するためには，通常，各項目内で以下のような識別条件（制約）が課される．

$$\sum_{k=1}^{m_j} a_{j_k} = 0, \quad \sum_{k=1}^{m_j} d_{j_k} = 0 \tag{4}$$

NRM を用いることで，正解以外の選択肢が持つ情報も能力推定に反映させることが可能となり，より詳細な項目分析や受検者の診断が可能となる．しかし，NRM は各選択肢に対して独立したパラメータを割り当てるため，2PLM に比して推定すべきパラメータ数が大幅に増加する．DeMars は，この選択肢数の増大が推定値の誤差分散を増大させ，安定した推定には大規模なサンプルサイズを要することを示している [13]．また NRM は IRT をベースとしていることから，同様に推定において潜在能力が標準正規分布に従うという強い仮定を置くことが多く，実際の受検者集団の分布を反映できず，予測精度に限界が生じる可能性が指摘されている．

3　提案手法

3.1　Deep-IRT

Tsutsumi らによって提案された Deep-IRT[9, 6, 10, 11] は，深層学習を用いることで従来の項目反応理論（IRT）の制約を解消するテスト理論である．Deep-IRT は，受検者パラメータを抽出する受検者ネットワークと，項目パラメータを抽出する項目ネットワークを独立に構成し，それらを統合することで受検者の反応を予測する点に特徴がある．

ここで，i は全受検者数，j は全項目数を表す．

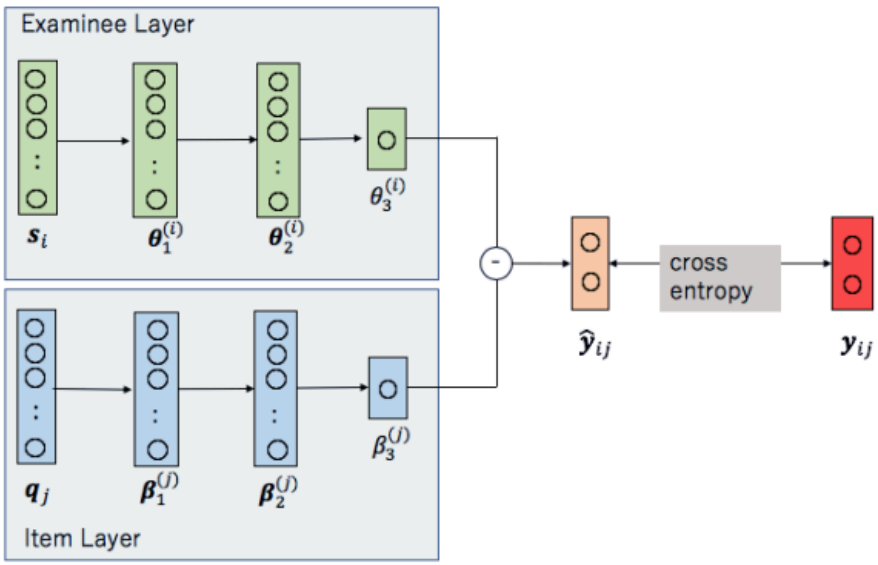

図 1　Deep-IRT おけるモデル図（Tsutsumi ら [9] より引用）

3.1.1　受検者ネットワーク

受検者ネットワークでは i 番目の受験者を識別するため，i 番目の要素のみが 1, 他の要素が 0 の one-hot ベクトル $\boldsymbol{s}_i \in \mathbb{R}^I$ を入力とし，項目ネットワーク（重み $\boldsymbol{W}^{(\theta_n)}$，バイアス $\boldsymbol{\tau}^{(\theta_n)}$）を用いてスカラー値である受検者 i の 能力パラメータ $\theta_N^{(i)}$ を算出する．この多層パーセプトロンモデル (MLP) は N 層のニューラルネットワークから構成される．

$$\boldsymbol{\theta}_1^{(i)} = \tanh(\boldsymbol{W}^{(\boldsymbol{\theta}_1)}\boldsymbol{s}_i + \boldsymbol{\tau}^{(\boldsymbol{\theta}_1)}) \tag{5}$$

$$\boldsymbol{\theta}_n^{(i)} = \tanh(\boldsymbol{W}^{(\boldsymbol{\theta}_n)}\boldsymbol{\theta}_{n-1}^{(i)} + \boldsymbol{\tau}^{(\boldsymbol{\theta}_n)}) \quad (n = 2, 3, ..., N-1) \tag{6}$$

$$\theta_N^{(i)} = \boldsymbol{W}^{(\theta_N)}\boldsymbol{\theta}_{N-1}^{(i)} + \tau^{(\theta_N)} \tag{7}$$

活性化関数として，以下の双曲線正接関数（ハイパボリックタンジェント関数）を用いている．

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{8}$$

3.1.2 項目ネットワーク

同様に，項目ネットワークでは j 番目の項目を識別するため，j 番目の要素のみが 1, 他の要素が 0 の one-hot ベクトル $\boldsymbol{q}_j \in \mathbb{R}^J$ を入力とし, N 層の項目ネットワーク（重み $\boldsymbol{W}^{(\boldsymbol{\beta}_n)}$，バイアス $\boldsymbol{\tau}^{(\boldsymbol{\beta}_n)}$）を用いてスカラー値である項目 j の難易度パラメータ $\beta_N^{(j)} \in \mathbb{R}$ を算出する．この多層パーセプトロンモデル (MLP) は N 層のニューラルネットワークから構成される．

$$\boldsymbol{\beta}_1^{(j)} = \tanh(\boldsymbol{W}^{(\boldsymbol{\beta}_1)}\boldsymbol{q}_j + \boldsymbol{\tau}^{(\boldsymbol{\beta}_1)}) \tag{9}$$

$$\boldsymbol{\beta}_n^{(j)} = \tanh(\boldsymbol{W}^{(\boldsymbol{\beta}_n)}\boldsymbol{\beta}_{n-1}^{(j)} + \boldsymbol{\tau}^{(\boldsymbol{\beta}_n)}) \quad (n = 2, 3, ..., N-1) \tag{10}$$

$$\beta_N^{(j)} = \boldsymbol{W}^{(\beta_N)}\boldsymbol{\beta}_{N-1}^{(j)} + \tau^{(\beta_N)} \tag{11}$$

3.1.3 選択確率の算出と学習

推定された能力値 $\theta_N^{(i)}$ と難易度 $\beta_N^{(j)}$ を用い，出力層（重み $\boldsymbol{W}^{(y)}$，バイアス $\boldsymbol{\tau}^{(y)}$）を介して受検者 i が項目 j に正答（$y_{ij} = 1$）する確率をモデル化する．まず，能力と難易度の差分に基づき，正誤反応に対応するロジット $\boldsymbol{h}^{(i,j)} \in \mathbb{R}^2$ を算出する．

$$\boldsymbol{h}^{(i,j)} = \boldsymbol{W}^{(y)\top}(\theta_N^{(i)} - \beta_N^{(j)}) + \boldsymbol{\tau}^{(y)} \tag{12}$$

これに対しソフトマックス関数を適用することで，正誤反応の予測確率ベクトル $\hat{\boldsymbol{y}}_{i,j} = [P(y_{i,j} = 0), P(y_{i,j} = 1)]$ を得る．

$$\hat{\boldsymbol{y}}_{i,j} = \mathrm{Softmax}(\boldsymbol{h}^{(i,j)}) \tag{13}$$

3.2 テキスト特徴量

項目の特徴量として，問題文や選択肢のテキスト特徴量を用いる手法がある．具体的には文字数，文の長さ，単語数など [18] の基本的なテキスト特徴量を用いる手法や，事前学習済み Transformer モデルである BERT[19] 等を用いた手法, LLM を用いた手法 [20] が提案されている．

3.2.1 事前学習済み Transformer モデル

エンコーダーとして用いる事前学習済み Transformer モデルには BERT 以外にも複数のモデルが存在している．例えば Longformer[16] は BERT よりも長文を解釈する能力に優れているが，計算効率が良いという利点がある．その理由として．言語モデルが長文理解に用いる Attention 機構の違いが挙げられる．全てのトークンに注目する Full Attention の BERT に対し，Longformer においては特定のトークンでのみ全体を参照する Global Attention と，それ以外では隣接トークンのみを参照する Sliding Window Attention を組み合わせている．この仕組みにより，長文に対しても効率的に Attention 計算を行うことが可能となっている．

3.3 多肢選択式問題における LLM による問題解答過程を組み込んだ Deep-IRT

本研究では，多肢選択式問題において選択確率を予測する，Deep-IRT モデル (Multiple Choice Deep-IRT; MCD-IRT) を提案する．提案手法のモデルは Deep-IRT と同様に受検者ネットワークと項目ネットワークを持つ．受検者ネットワークにおいては同様に反応データから個々の能力を推定する．一方，提案手法における項目ネットワークにおいてはテキストを特徴として問題の特徴量を算出する．

この構造より，従来の IRT モデルや NRM モデルと異なり，受検者の能力を正規分布の仮定に依存せずに推定できるため，多様な能力値を表現できる．また従来の Deep-IRT モデルと異なり，問題のテキスト特徴量を使用しているほか誤答を区別するため，問題と選択肢ごとの詳細な違いをモデルの予測に反映でき，予測精度の向上が期待できる．

事前の反応データが存在しない新規項目に対しても，各選択肢の選択確率を予測することが可能となる．

3.3.1 モデルの構成

提案手法である MCD-IRT は，以下の 3 つのネットワークによって構成され，受検者の one-hot vector とテキスト特徴量を入力とし，受検者ごとの各選択肢の選択確率を出力とする．

1) **受検者ネットワーク**: 受検者の過去の反応履歴に基づき，受検者の能力ベクトル $\boldsymbol{\theta}_i$ を算出するネットワークである．
2) **項目ネットワーク**: 問題文などの項目内容から文脈情報を考慮した特徴量ベクトル $\boldsymbol{q}_j$ を抽出するネットワークである．
3) **選択確率予測ネットワーク**: 得られた能力ベクトル $\boldsymbol{\theta}_i$ と項目特徴量ベクトル $\boldsymbol{q}_j$ を入力とし，各選択肢の選択確率を算出するネットワークである．

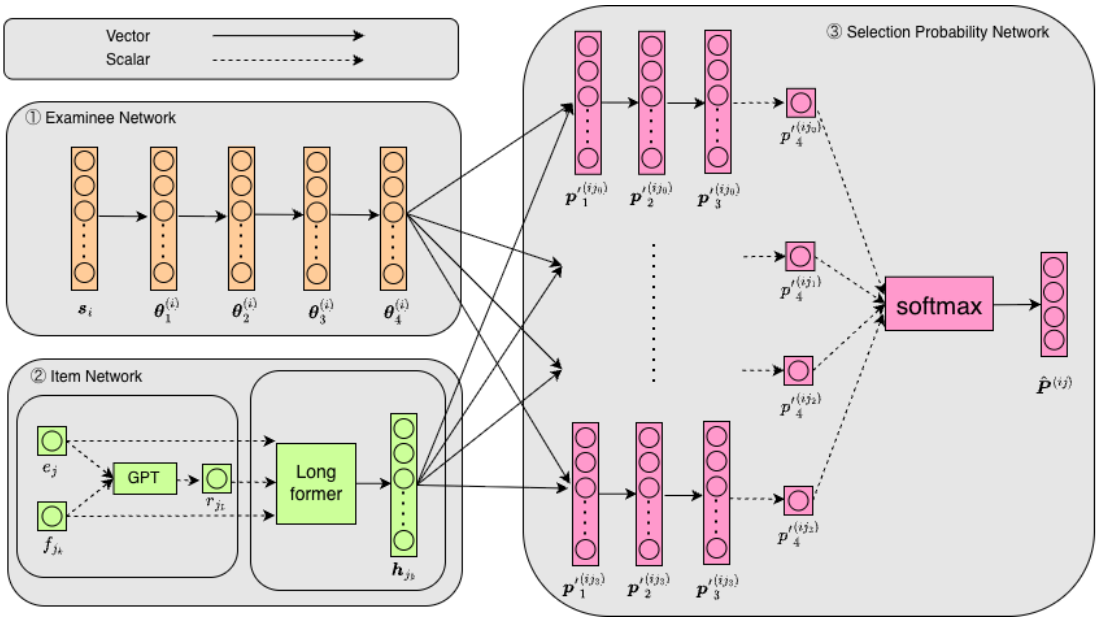

図 2　MCD-IRT のモデル図

3.3.2　項目の特徴量生成 (Item Network)

(i) **解答過程および誤答理由の生成:** 各項目 $j \in \{1, \ldots, J\}$ における問題文を e_j，各項目 j に対する選択肢を f_{j_k} とする．ここで $k \in \{0, 1, 2, 3\}$ は選択肢のインデックスを表す．これらを入力条件とし，生成モデル G（GPT-4o）により，各選択肢に対する解答過程 r_{j_0} または誤答選択理由 r_{j_k} を出力として得る．なお誤答における $k \in \{1, 2, 3\}$ は順不同とする．

$$r_{j_k} = G(e_j, f_{j_k}) \tag{14}$$

なお問題解答過程を出力する際に GPT-4o に与えるプロンプトは，次の二つの指示から構成される．

（a）問題文，質問文，選択肢，および正答といった問題の構成要素を提示すること．

（b）正答の問題解答過程を出力するよう求めること．

(ii) **Longformer による項目特徴量の抽出:** 問題文 e_j，選択肢 f_{j_k}，および解答過程 r_{j_k} を連結したテキストを入力する．これを 事前学習済み Transformer モデルである Longformer に入力し，項目 j の選択肢 k に対する項目特徴量ベクトル $\boldsymbol{h}_{j_k} \in \mathbb{R}^{768}$ を抽出する．

$$\boldsymbol{h}_{j_k} = \text{Longformer}(e_j, f_{j_k}, r_{j_k}) \tag{15}$$

3.3.3 受検者能力 $\boldsymbol{\theta}_4^{(i)}$ の特徴量生成 (Examinee Network)

Deep-IRT に基づき，活性化関数を tanh とする 4 層の多層パーセプトロン（MLP）を用いて，受検者 i の能力値ベクトル $\boldsymbol{\theta}_4^{(i)}$ を算出する．入力として，i 番目の要素のみが 1, 他の要素が 0 の one-hot ベクトル $\boldsymbol{s}_i \in \mathbb{R}^I$ を用いる．

$$\boldsymbol{\theta}_1^{(i)} = \tanh(\boldsymbol{W}^{(\boldsymbol{\theta}_1)}\boldsymbol{s}_i + \boldsymbol{\tau}^{(\boldsymbol{\theta}_1)}) \tag{16}$$

$$\boldsymbol{\theta}_2^{(i)} = \tanh(\boldsymbol{W}^{(\boldsymbol{\theta}_2)}\boldsymbol{\theta}_1^{(i)} + \boldsymbol{\tau}^{(\boldsymbol{\theta}_2)}) \tag{17}$$

$$\boldsymbol{\theta}_3^{(i)} = \tanh(\boldsymbol{W}^{(\boldsymbol{\theta}_3)}\boldsymbol{\theta}_2^{(i)} + \boldsymbol{\tau}^{(\boldsymbol{\theta}_3)}) \tag{18}$$

$$\boldsymbol{\theta}_4^{(i)} = \boldsymbol{W}^{(\boldsymbol{\theta}_4)}\boldsymbol{\theta}_3^{(i)} + \boldsymbol{\tau}^{(\boldsymbol{\theta}_4)} \tag{19}$$

3.3.4 選択確率 $\hat{\boldsymbol{P}}_{ij}$ の算出 (Selection Probability Network)

抽出された項目特徴量と受検者能力を MLP によって統合し，選択肢ごとの選択確率の潜在変数を算出する．能力値ベクトル $\boldsymbol{\theta}_n^{(i)}$ と項目特徴量 $\boldsymbol{h}_{j_k}$ を結合した $\boldsymbol{t}_{ij_k} = [\boldsymbol{\theta}_n^{(i)}; \boldsymbol{h}_{j_k}]$ を入力とし，選択確率ネットワーク（重み $\boldsymbol{W}^{(\boldsymbol{p}'^{(ij_k)})}$，バイアス $\boldsymbol{\tau}^{(\boldsymbol{p}'^{(ij_k)})}$）を用いて各選択肢 k の潜在変数 $p'^{(ij_k)} \in \mathbb{R}$ を算出する．

$$\boldsymbol{p}_1'^{(ij_k)} = \mathrm{LeakyReLU}(\boldsymbol{W}_1^{(\boldsymbol{p}_1')}\boldsymbol{t}_{ij_k} + \boldsymbol{\tau}_1^{(\boldsymbol{p}_1')}) \tag{20}$$

$$\boldsymbol{p}_2'^{(ij_k)} = \mathrm{LeakyReLU}(\boldsymbol{W}_2^{(\boldsymbol{p}_2')}\boldsymbol{p}_1'^{(ij_k)} + \boldsymbol{\tau}_2^{(\boldsymbol{p}_2')}) \tag{21}$$

$$\boldsymbol{p}_3'^{(ij_k)} = \mathrm{LeakyReLU}(\boldsymbol{W}_3^{(\boldsymbol{p}_3')}\boldsymbol{p}_2'^{(ij_k)} + \boldsymbol{\tau}_3^{(\boldsymbol{p}_3')}) \tag{22}$$

$$p_4'^{(ij_k)} = \mathrm{LeakyReLU}(\boldsymbol{W}_4^{(p_4')}\boldsymbol{p}_3'^{(ij_k)} + \boldsymbol{\tau}_4^{(p_4')}) \tag{23}$$

これら 4 つの潜在変数をまとめたベクトルを $\boldsymbol{p}'^{(ij)} = [p'^{(ij_0)}, p'^{(ij_1)}, p'^{(ij_2)}, p'^{(ij_3)}]$ とする．このベクトルにソフトマックス関数を適用し，予測選択確率ベクトル $\hat{\boldsymbol{P}}^{(ij)} = [P^{(ij_0)}, P^{(ij_1)}, P^{(ij_2)}, P^{(ij_3)}]$ を得る．

$$\hat{\boldsymbol{P}}^{(ij)} = \mathrm{Softmax}\left(\boldsymbol{p}'^{(ij)}\right) = \frac{\exp(p'^{(ij_k)})}{\sum_{k'=0}^{3}\exp(p'^{(ij_{k'})})} \quad (k = 0, 1, 2, 3) \tag{24}$$

4　選択確率予測

4.1　実験設定

提案手法である MCD-IRT の有効性を検証するため，従来手法である NRM および Deep-IRT との比較実験を 5 分割交差検証により行った.

4.1.1　評価指標

評価指標には，Accuracy，F1 score，および交差エントロピー誤差（Cross-Entropy）を用いた.

各指標の定義は以下の通りである.

$$\text{Accuracy} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\mathbb{I}(f_{ijk} = \hat{f}_{ijk}}{)} I \times J \tag{25}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

ここで，f_{ijk} は実際に受検者が選んだ選択肢，$\hat{f}_{ijk}$ はモデルが予測した中で最も選択確率が高い選択肢を表す．また，$\mathbb{I}(\cdot)$ は指示関数であり，条件が真であれば 1，偽であれば 0 を返す．さらに，TP（真陽性），FP（偽陰性），FN（偽陽性）を用いて，Precision および Recall は次式で定義される.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{27}$$

4.1.2　パラメータ設定

各手法の実験における設定値は以下の通りである.

- **NRM (Nominal Response Model)**: Bock [12] による多値項目反応理論のモデルに基づき，IRT Pro を用いてパラメータ推定を行った．推定にはマルコフ連鎖モンテカルロ法（MCMC）を採用し，サンプリング回数を 4,000 回，バーンイン数を 2,000 回に設定した.
- **Deep-IRT**: 既存の Deep-IRT モデルは本来 2 値（正誤）予測を想定した構造であるため，本実験では出力層の次元を 4 次元へと変更し，4 択問題の各選択確率を出

力できるように拡張した．学習パラメータは，学習率 0.001，バッチサイズ 32，能力ベクトルの隠れ層のノード数 32，能力ベクトルの出力次元数 2，エポック数 50 とした．

- **MCD-IRT (提案手法)**: 提案手法においても，Deep-IRT と同様のハイパーパラメータ（学習率 0.001，バッチサイズ 32，能力ベクトルの隠れ層のノード数 32，能力ベクトルの出力次元数 2，エポック数 50）を設定し，学習および評価を行った．

4.1.3 データセット

評価データセットには，EEDI [21] を用いた．これは 2020 年度に英国の中学生を対象として行われた 4 択式の数学問題のデータセットである．画像問題を除いた項目と，該当項目の反応データを使用したため，実験で用いた項目数は 327 件，受検者数は 4,424 人である．

4.2 実験結果と考察

表 1 に選択確率予測の比較結果を示す．有効数字は 5 桁とした．

実験の結果，NRM よりも全ての評価指標において提案手法が従来手法を上回る予測精度を達成した．その要因としては，正規分布の仮定に依存せずに受検者の能力を詳細に推定できたことが挙げられる．

次に Deep-IRT と比較して，全ての評価指標において提案手法がより高い予測精度を示した．その要因としては，問題の特徴量を活用することで文脈に応じた予測が可能となったことが要因として考えられる．

表 1　各手法による選択確率予測の比較 (Mean ± SD)

Method	Accuracy	F1 Score	Cross-Entropy
NRM	0.5349 ± 0.0100	0.5334 ± 0.0100	1.1990 ± 0.0265
Deep-IRT	0.5368 ± 0.0065	0.5364 ± 0.0064	1.2411 ± 0.0067
MCD-IRT (提案手法)	**0.5729 ± 0.0020**	**0.5724 ± 0.0020**	**1.0410 ± 0.0062**

5　難易度予測への応用

予測した選択確率は，教育評価において多様な目的で活用できる．例えば Feng ら [15] は, 各選択肢の選択確率の平均から項目の難易度を予測する手法を提案している．そこで，今回の提案手法である MCD-IRT のモデルにおいても，難易度予測を行うように拡張し，Feng らの手法との比較実験を行った．

5.1　既存手法におけるモデル

Feng らは MCQ における難易度予測において, LLM を融合させた手法を提案した．LLM で生成した「問題解答過程」と「誤答選択理由」を項目に関する多次元の特徴量とし, 標準正規分布からサンプリングした仮想的な受検者の能力ベクトルとの相互作用を IRT（NRM）から着想を得た式で計算することで, 受検者の各選択肢の選択確率平均を求め，難易度予測に応用している．以降に具体的な手法について記載する．

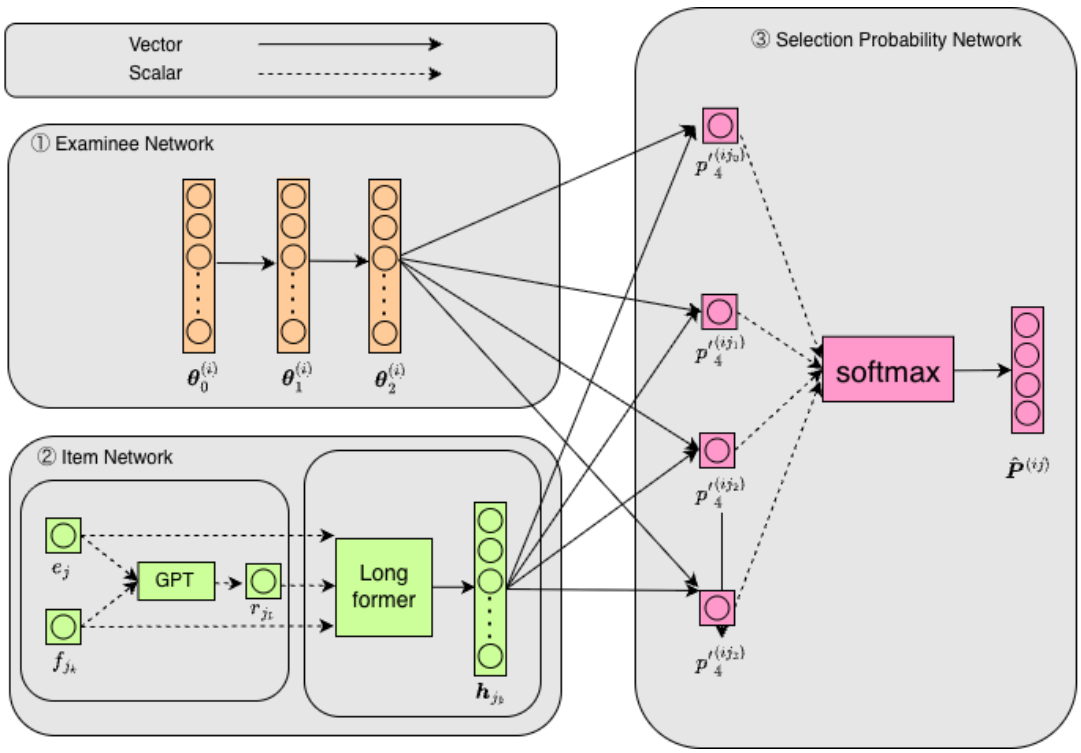

図 3　既存手法におけるモデル図

5.1.1 項目の特徴量生成 (Item Network)

項目の特徴量生成 (Item Network) においては, Feng らの提案から着想を得たため，提案手法と同様である．

5.1.2 受検者能力の射影 (Examinee Network)

標準正規分布からサンプリングした仮想的な受検者 i の潜在能力ベクトル（初期値）を $\boldsymbol{\theta}_0^{(i)}$ とする．これを項目の特徴量生成ネットワークの出力である，Longformer により出力された特徴量と同じ $\boldsymbol{h}_{j_k} \in \mathbb{R}^{768}$ の高次元の空間へ射影するため，2 層の多層パーセプトロン（MLP）を用いて変換を行う．

$$\boldsymbol{\theta}_1^{(i)} = \mathrm{LeakyReLU}(\boldsymbol{W}^{(\theta_1)}\boldsymbol{\theta}_0^{(i)} + \boldsymbol{\tau}^{(\theta_1)}) \tag{28}$$

$$\boldsymbol{\theta}_2^{(i)} = \mathrm{LeakyReLU}(\boldsymbol{W}^{(\theta_2)}\boldsymbol{\theta}_1^{(i)} + \boldsymbol{\tau}^{(\theta_2)}) \tag{29}$$

5.1.3 選択確率 $\hat{\boldsymbol{P}}^{ij}$ の算出 (Selection Probability Network)

抽出された項目特徴量と受検者の能力値を統合し，選択肢ごとの選択確率を算出する．

各選択肢 $k \in \{0, 1, 2, 3\}$ に対し，項目特徴量 $\boldsymbol{h}_{j_k}$ と変換された能力ベクトル $\boldsymbol{\theta}^{(i)}$ の内積により，スコア $p'^{(ij_k)}$ を算出する．この計算は IRT（NRM）から着想を得ている．

$$p'^{(ij_k)} = \boldsymbol{h}_{j_k}^{\top}\boldsymbol{\theta}^{(i)} \tag{30}$$

これら 4 つのスコアをまとめたベクトルを $\boldsymbol{p}'^{(ij)} = [p'^{(ij_0)}, p'^{(ij_1)}, p'^{(ij_2)}, p'^{(ij_3)}]$ とする．これにソフトマックス関数 $\mathrm{Softmax}(\cdot)$ を適用することで，最終的な選択確率ベクトル $\boldsymbol{P}^{(ij)}$ を得る．

$$\hat{\boldsymbol{P}}^{(ij)} = \mathrm{Softmax}(\boldsymbol{p}'^{(ij)}) \tag{31}$$

5.2 実験方法と結果考察

Feng ら [15] は, 各選択肢の選択確率の平均から項目の難易度を予測する手法を提案している．そこで本研究でも同様のモデルを使用し, 提案手法で予測した各選択肢の選択確率の平均を用いて項目の難易度を予測した結果を比較した．具体的な計算は以下の通りである．

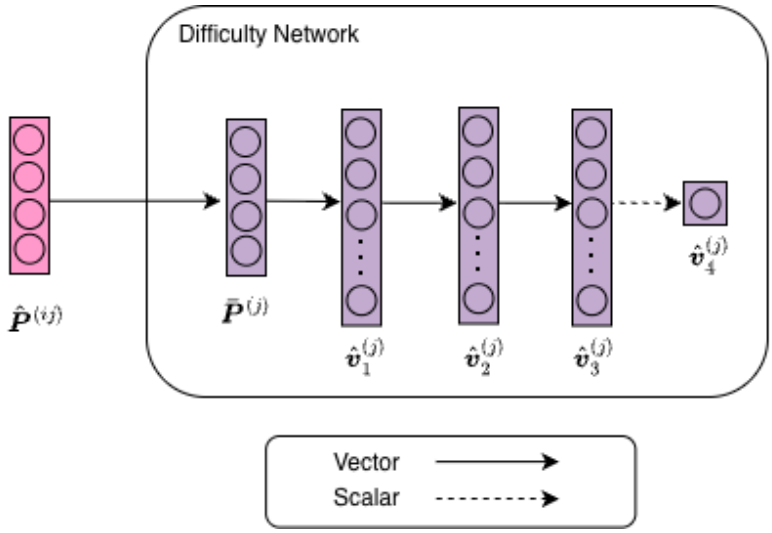

図 4　難易度予測モデル

(i) **項目難易度の予測:** 難易度予測ネットワーク（重み $\boldsymbol{W}^{(\hat{\boldsymbol{v}}_n^{(j)})}$，バイアス $\boldsymbol{\tau}^{(\hat{\boldsymbol{v}}_n^{(j)})}$）において，全受検者の平均選択確率ベクトル $\overline{\boldsymbol{P}}^{(j)} \in \mathbb{R}^4$ を入力とし，予測難易度 $\hat{v}_4^{(j)}$ を算出する．

$$\hat{\boldsymbol{v}}_1^{(j)} = \tanh(\boldsymbol{W}^{(\hat{\boldsymbol{v}}_1)}\overline{\boldsymbol{P}}^{(j)} + \boldsymbol{\tau}^{(\hat{\boldsymbol{v}}_1)}) \tag{32}$$

$$\hat{\boldsymbol{v}}_2^{(j)} = \tanh(\boldsymbol{W}^{(\hat{\boldsymbol{v}}_2)}\hat{\boldsymbol{v}}_1^{(j)} + \boldsymbol{\tau}^{(\hat{\boldsymbol{v}}_2)}) \tag{33}$$

$$\hat{\boldsymbol{v}}_3^{(j)} = \tanh(\boldsymbol{W}^{(\hat{\boldsymbol{v}}_3)}\hat{\boldsymbol{v}}_2^{(j)} + \boldsymbol{\tau}^{(\hat{\boldsymbol{v}}_3)}) \tag{34}$$

$$\hat{v}_4^{(j)} = \tanh(\boldsymbol{W}^{(\hat{\boldsymbol{v}}_4)}\hat{\boldsymbol{v}}_3^{(j)} + \boldsymbol{\tau}^{(\hat{\boldsymbol{v}}_4)}) \tag{35}$$

ここで，$\boldsymbol{W}$ および $\boldsymbol{\tau}$ は難易度予測のための学習パラメータである．

(ii) **損失関数：**モデルの損失関数 L は，難易度予測誤差と KL ダイバージェンスの和の平均として次式で定義される．

$$L = \frac{1}{J}\sum_{j=1}^{J}(L_{v_j} + \alpha L_{\mathrm{KL}_j}) \tag{36}$$

ここで，各項および変数の定義は以下の通りである．

- J: 問題の総数
- α: 2 つの損失の重みを調整するハイパーパラメータ
- L_{v_j}: 難易度予測誤差（平均二乗誤差）

$$L_{v_j} = (v^{(j)} - \hat{v}^{(j)})^2 \tag{37}$$

  - $v^{(j)}$: 第 j 問の IRT などで求めた難易度予測値
  - $\hat{v}^{(j)}$: モデルによる予測難易度
- L_{KL_j}: 全受検者の実際の選択割合と予測選択確率の平均との差異（KL ダイバージェンス）

$$L_{\mathrm{KL}_j} = \sum_{k=0}^{3} P(j_k) \log \left(\frac{P(j_k)}{\overline{P}(j_k)} \right) \tag{38}$$

  - $P(j_k)$: 全受検者が実際に第 j 問の選択肢 k を選んだ割合
  - $\overline{P}(j_k)$: モデルが予測した選択肢 k を選ぶ全受検者の選択確率の平均

実験では, 平均二乗誤差（MSE）および決定係数（R^2），以下に示す全ペア比較における正解率である MATCH によって各モデルの難易度予測精度を評価した．MATCH は，2 つの問題の難易度の相対的な順序をモデルが正しく予測できた割合として定義され，以下の式で算出される：

$$\mathrm{MATCH} = \frac{1}{|\mathcal{Y}|} \sum_{(i,j)\in\mathcal{Y}} \mathbb{I}(\mathrm{sign}(v_i - v_j) = \mathrm{sign}(\hat{v}_i - \hat{v}_j)) \tag{39}$$

ここで，各記号の定義は以下の通りである：

- $\mathcal{Y}$：難易度が異なる全アイテムペア (i, j) の集合
- $|\mathcal{Y}|$：比較対象となる全ペアの総数
- $v^{(j)}, v^{(j')}$：アイテム j, j' における難易度の正解ラベル（Ground Truth）
- $\hat{v}^{(j)}, \hat{v}^{(j')}$：モデルによって予測されたアイテム j, j' の難易度スコア
- $\mathbb{I}(\cdot)$：指示関数（条件が真であれば 1，偽であれば 0 を返す）

本実験でも，Feng ら [15] の手法を再現するため，同論文で指定されたハイパーパラメータおよびモデル構成を採用した．Feng らと同条件で選択確率を求めたのち，MLP 層の学習率として 0.01 を設定した．損失関数における重み関数の α においては，Feng らの設定に従い 0.0886 を採用した．なお実験における有効数字は 4 桁とする．

表 2　各手法による予測難易度の比較

Method	MSE	R^2	MATCH
Feng らの手法	0.394	0.515	0.744
MCD-IRT (提案手法)	**0.317**	**0.610**	**0.810**

実験の結果, 実データから推定した能力値を入力とした提案手法は, ベースラインである既存手法を上回る難易度予測精度を記録した. 従来の手法よりも予測精度の高い選択確率を学習に用いたことが要因と考えられる.

6 むすび

本研究では, Deep-IRT をもとに，LLM を活用した多肢選択式問題のための MCD-IRT を提案した. 先行研究と比較して，提案手法は選択確率と難易度をより高い精度で予測できた.

今後の課題としては，他のデータセットを用いても同様の結果が得られるか検証すること，提案手法のうちの受検者ネットワーク・項目ネットワークそれぞれのみでの予測選択確率の結果と双方を使用した提案手法の結果を比較して検討したい.

参考文献

[1] ISO/IEC. Information technology — A code of practice for the use of information technology (IT) in the delivery of assessments. ISO/IEC 23988, 2007.

[2] 植野真臣. e テスティング: 先端理論と技術. **教育システム情報学会誌**, 26(2):204–217, 2009.

[3] Maomi Ueno. Ai based e-testing as a common yardstick for measuring human abilities. In *2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE, 2021.

[4] Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. Addison-Wesley, 1968.

[5] Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of item response theory*. Sage, 1991.

[6] 木下涼 and 植野真臣. 深層学習によるテスト理論: item deep response theory. **電子情報通信学会論文誌** *D*, 103(4):314–329, 2020.

[7] Chun-Kit Yeung. Deep-irt: Make deep learning based knowledge tracing explainable. *arXiv preprint arXiv:1904.11738*, 2019.

[8] Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep-IRT with independent student and item networks. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, pages 401–407, 2021.

[9] Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep item response theory as a novel test theory based on deep learning. *Electronics*, 10(9):1020, 2021.

[10] Emiko Tsutsumi, Y. Guo, Ryo Kinoshita, and Maomi Ueno. Deep knowledge

tracing incorporating a hypernetwork with independent student and item networks. *IEEE Transactions on Learning Technologies*, 16(5):586–599, 2023.

[11] Emiko Tsutsumi, T. Nishio, and Maomi Ueno. Deep-irt with a temporal convolutional network for reflecting students' long-term history of ability data. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED 2024)*, pages 250–264. Springer, 2024.

[12] R Darrell Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972.

[13] Christine E DeMars. Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, 27(4):275–288, 2003.

[14] Luca Benedetto and et al. Is the answer just a number? investigating the impact of answer text on item difficulty prediction. *Proceedings of the 14th International Conference on Educational Data Mining*, 2021.

[15] Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew S Lan. Reasoning and sampling-augmented mcq difficulty prediction via llms. In *International Conference on Artificial Intelligence in Education (AIED)*, pages 31–45, 2025.

[16] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[17] David Thissen and Lynne Steinberg. A response model for multiple-choice items. *Psychometrika*, 49(4):501–519, 1984.

[18] S. AlKhuzaey, F. Grasso, T. R. Payne, and V. Tamma. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914, 2024.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[20] Radhika Kapoor et al. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*, 2025.

[21] Eedi. Neurips 2020 education challenge, 2020. Accessed: 2025-05-20.