看護師国家試験一般問題の誤答選択肢自動生成手法の検討

伊藤晃 [1] 荒瀬由紀 [1]
[1] 東京科学大学 情報理工学院
ito.h.0d6f@m.isct.ac.jp
arase@c.titech.ac.jp

概要

多肢選択肢問題では，正解ではないがもっともらしい，回答者を惑わす選択肢である誤答選択肢が重要である．誤答選択肢は人手による作成コストが高く，自動生成による作問者補助が期待されている．本研究では，看護師国家試験の一般問題を対象に，誤答選択肢同士の多様性に着目し，妥当かつより多様な誤答選択肢を生成する手法を提案する．提案手法では，LLM による生成結果の多様性を高めるために temperature を高い値に設定し，複数の temperature で得られた候補を集約する．加えて，LLM-as-a-Judge により非妥当な誤答選択肢を除外し，適切な誤答選択肢を抽出する．本手法により，実際の試験問題と同程度の多様性を持つ誤答選択肢を，高い品質を維持しながら作成できることを示した．

1 はじめに

多肢選択肢問題は，ある問題に対し， 複数の選択肢から 1 つまたは複数の正解を選択する形式の問題である．記述回答式の問題に比べ，採点に要するコストが低く，少ない労力で学習者の習熟度を効果的に測定できるテスト形式として，教育分野で広く用いられている．しかしながら，多肢選択肢問題においては，「問題文」「正解選択肢」「誤答選択肢」の 3 つを作成する必要があり，問題作成におけるコストの面には課題がある．特に「誤答選択肢」は，問題に対して複数用意する必要があり，作成コストが高い．適切な誤答選択肢を自動生成することは，多肢選択肢問題を作成する上で大いに役立つと考えられる．誤答選択肢を自動生成する方法として，以前よりさまざまな手法が提案されてきた [1] が，近年では大規模言語モデル (以降 LLM とする) を用いたアプローチ [2] が多く見られ，日々研究が進んでいる．

> 散瞳薬を用いた眼底検査を受ける成人への説明で適切なのはどれか．
>
> 1. 「角膜を観察します」
> 2. 「検査後に抗菌薬を点眼します」
> 3. 「眼を閉じた状態で検査室に誘導します」
> 4. 「点眼後 30 分で散瞳薬の効果が現れます」
>
> 出典: 第 114 回看護師国家試験 午後 第 52 問

図 1 看護師国家試験の問題例

看護師国家試験は，厚生労働省が実施する，日本で看護師として働くための免許の取得に必要な国家試験である．この試験では，看護師が保健医療の現場に第一歩を踏み出す際に，少なくとも備えるべき基本的な知識及び技能が問われる [3]．看護師国家試験の問題の例を図 1 に示す．図 1 に代表されるように，看護師国家試験では，基本的に 4 択もしくは 5 択の多肢選択肢式の問題が出題される．また，看護師国家試験においては，必修問題，一般問題，状況設定問題の 3 種類の問題区分が存在する．必修問題では，看護師が備えるべき基本的な知識が問われる．一般問題では，看護学に関わる基礎的な内容から専門的な内容まで幅広い知識が問われる．状況設定問題では，実際の患者の症状などの事例が与えられ，それに対する適切な対応を選ぶケーススタディ形式の問題が出題される．このうち，必修問題においては，LLM による誤答選択肢の自動生成および作問における有用性についての研究が城戸らにより行われている [4][5] が，一般問題，状況設定問題については研究が進んでいない．

問題: 慢性副鼻腔炎についての説明で適切なのはどれか.
正解: 眼窩内感染を起こす危険性がある.
実際の誤答選択肢:
1 週間の内服で症状が軽減すれば受診の必要はない.
発症後 1 週は空気感染の危険性がある.
透明の鼻汁が特徴的である.
LLM が生成した誤答選択肢:
鼻腔の粘膜が肥厚する.
鼻腔の粘膜が乾燥する.
鼻腔の粘膜が腫脹する.

図 2　極度に類似した誤答選択肢の例

本研究では看護学の体系的な理解と知識の応用力を問う一般問題を対象とする。看護師国家試験の誤答選択肢生成における先行研究 [4][5] では，様々なモデルで誤答選択肢を生成し，複数モデルに共通するものを優先的に選択する手法が取られた. この手法は，必修問題のように単語および名詞節形式の誤答選択肢に対しては有効であるが，一般問題で多い文形式の誤答選択肢に対しては，複数モデルで同一の文を生成する可能性は低いと考えられる. 看護師国家試験の一般問題には従来とは異なるアプローチを考える必要がある.

また，誤答選択肢同士の多様性についても課題がある. 図 2 は，LLM が極端に類似した誤答選択肢を生成した例である. 先行研究 [4][5] で提案された手法では，生成時のハイパーパラメータである temperature は 0 に設定されていた. しかし同設定で LLM で生成した誤答選択肢を観察したところ，このような極度に類似した誤答選択肢が多く，人間による誤答選択肢よりも多様性が低い.

これらを踏まえ，本研究では，看護師国家試験における一般問題について，誤答選択肢の多様性確保に焦点を当て，妥当かつ実際の試験問題における誤答選択肢と同程度の多様性を持つ誤答選択肢を生成する手法を提案する. 提案手法では，生成結果の多様性を高めるために temperature を高い値に設定し，複数の temperature で得られた候補を集約する. 加えて，非妥当な誤答選択肢を除外し，適切な誤答選択肢を抽出する目的で LLM-as-a-Judge を用いる. 実験の結果，人手の誤答選択肢と同程度の多様性を持つ選択肢が自動作成できることを示した.

2　提案手法

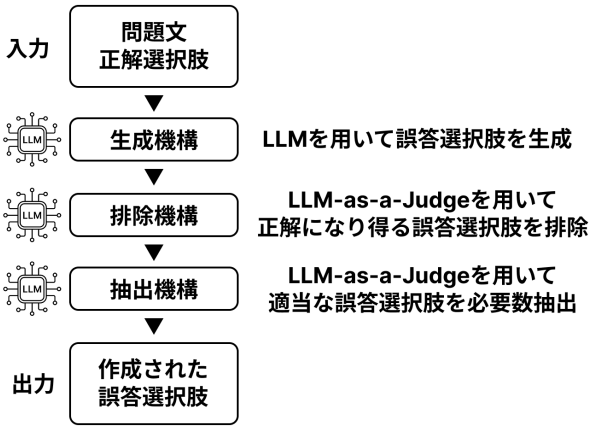

図 3　誤答選択肢の生成フロー

本研究における提案手法の生成フローを図 3 に示す. 本手法では，生成機構，排除機構，抽出機構の 3 つの機構を用いて誤答選択肢の生成を行った. なお，これら全ての機構について，モデルは Llama-3.1-Swallow-8B-Instruct-v0.5 を使用した.

生成機構では，誤答選択肢を Five-Shot，ファインチューニング（SFT）の 2 つの手法でそれぞれ実際の誤答選択肢の数だけ生成させる. このとき，それぞれの手法について 3 種類の temperature を用いて生成させた. 生成して得られたすべての誤答選択肢を集約し，これを次の排除機構に与えた. なお，temperature は (0.2,0.4,0.6) と (0.8,1.0,1.2) の 2 種類の組を試した.

排除機構では，生成機構より与えられた誤答選択肢それぞれについて，問題文に対する正解になってしまう可能性があるか LLM を用いて判定し，正解になりうると判定された誤答選択肢を排除した.

抽出機構では，与えられた誤答選択肢から，LLM が適切と判断したものを実際の誤答選択肢の数だけ抽出した. 適切な誤答選択肢の基準は，システムプロンプトを通して与えた. 基準は Rajiv によるガイドライン [6] を日本語に翻訳したものを基に作成した.

3　実験設定

実際の看護師国家試験を用いた評価実験を行う.

3.1　データセット

厚生労働省より提供を受けた看護師国家試験の過去問題 640 問と，共同研究者より提供を受けた模擬試験問題 300 問をもとに，データセットを構築した. データセットを作成する上で，問題文に図表を含む

	国家試験過去問題	模擬試験問題
訓練データ	107	268
検証データ	107	0
評価データ	236	0

図 4　作成したデータセットの内訳

問題，正解が複数存在する問題，数値を答える形式の問題，選択肢に組み合わせ形式を含む問題などを除外した. 看護師国家試験の過去問題からは 450 問，模試試験問題からは 268 問をデータセットの対象とした. 対象となるデータを図 4 に示すように分割し，訓練・検証・評価データセットをそれぞれ作成した.

3.2　比較手法

ベースラインの生成手法として，Zero-Shot，例を 5 つ与えてから生成させる Five-Shot，SFT の 3 つを設定した. Zero-Shot では，システムプロンプトに生成要件を，ユーザープロンプトに問題文と正解選択肢を与えて生成させた. Five-Shot では，システムプロンプトに加え，問題文，正解選択肢とその応答例を 5 つ与えてから生成させた. 例は問題ごとに訓練データセットから無作為に 5 件抽出した. SFT では，訓練データセットを用いてモデルをファインチューニングした後，Zero-Shot と同様に生成した. なお，計算資源の節約のため，ファインチューニングには QLoRA[7] を採用した. 詳細は付録 A に示す.

全ての手法で Llama-3.1-8B-Instruct[8] を日本語向けにファインチューニングしたモデルである Llama-3.1-Swallow-8B-Instruct-v0.5[9][10][11] を用いた. またベースライン 3 手法の生成においては，生成時における temperature の値を 0 としている.

3.3　評価指標

誤答選択肢の自動評価指標として，城戸らによる先行研究 [4] を参考に 7 指標 (R，P，R_s，P_s，R_cs，P_cs，DV) を設定した.

データセット $\mathbf{Q}$ について，各問題 $Q_i \in \mathbf{Q}$ は問題文 q_i，正解選択肢 a_i，実際の誤答選択肢 G_i からなる. ここで，実際の誤答選択肢の個数 $|G_i|$ は，全ての選択肢が 4 択もしくは 5 択であり，各問題に正解選択肢が 1 つだけ存在することから，$|G_i| \in \{3, 4\}$ を満たす. また，問題文 q_i と正解選択肢 a_i から LLM を用いて生成した誤答選択肢を S_i とする. LLM を用いて生成する誤答選択肢の数 $|S_i|$ は，$|S_i| = |G_i|$ となるようにする. 生成した結果，この条件を満たせなかった問題はデータセットから除外した.

評価指標 R，P は次のように定義される.

$$R = \frac{1}{|\mathbf{Q}|}\sum_{i=1}^{|\mathbf{Q}|}\frac{|S_i \cap G_i|}{|G_i|}, \tag{1}$$

$$P = \frac{1}{|\mathbf{Q}|}\sum_{i=1}^{|\mathbf{Q}|}\frac{|S_i \cap G_i|}{|S_i|}, \tag{2}$$

$$|S_i \cap G_i| = \mathrm{num\_exact\_match}(S_i, G_i). \tag{3}$$

$\mathrm{num\_exact\_match}(S_i, G_i)$ は，入力 S_i, G_i に対し，S_i, G_i 間で完全一致する要素の個数を返す関数とする. これらの指標は，実際の誤答選択肢と LLM が生成した誤答選択肢がどれだけ完全一致するかを評価するため，文を生成する本研究においては厳しい評価指標と言える.

評価指標 $R_\mathrm{s}, P_\mathrm{s}$ は次のように定義される.

$$R_\mathrm{s} = \frac{1}{|\mathbf{Q}|}\sum_{i=1}^{|\mathbf{Q}|}\sum_{j=1}^{|G_i|}\frac{\mathrm{sim}(\mathrm{argmax}_{\mathrm{s}\in\mathrm{S_i}}\mathrm{sim}(\mathrm{s}, \mathrm{g_j}), \mathrm{g_j})}{|G_i|}, \tag{4}$$

$$P_\mathrm{s} = \frac{1}{|\mathbf{Q}|}\sum_{i=1}^{|\mathbf{Q}|}\sum_{j=1}^{|S_i|}\frac{\mathrm{sim}(\mathrm{argmax}_{\mathrm{g}\in\mathrm{G_i}}\mathrm{sim}(\mathrm{g}, \mathrm{s_j}), \mathrm{s_j})}{|S_i|}. \tag{5}$$

$\mathrm{sim}(a, b)$ は文 a と文 b のコサイン類似度を返す関数とする. なお，本論文では，文 a, b の埋め込みを，SentenceTransformer を用いて得ている. SentenceTransformer では，多言語対応のモデルである Multilingual-E5-Large[12] を使用している. これらの指標は LLM が生成した誤答選択肢が，どれほど実際の誤答選択肢に類似した誤答選択肢を生成できているかを示す指標である.

評価指標 R_cs，P_cs は次のように定義される.

$$R_\mathrm{cs} = \frac{1}{|\mathbf{Q}|}\sum_{i=1}^{|\mathbf{Q}|}\frac{\mathrm{MWM}(S_i, G_i)}{|G_i|}, \tag{6}$$

$$P_\mathrm{cs} = \frac{1}{|\mathbf{Q}|}\sum_{i=1}^{|\mathbf{Q}|}\frac{\mathrm{MWM}(S_i, G_i)}{|S_i|}. \tag{7}$$

$\mathrm{MWM}(A, B)$ は，文の集合 A と文の集合 B について，最大重み二部マッチングをした際の重みを返す関数である. 重み関数はコサイン類似度である. これらの評価指標は R_s，P_s と類似しているが，LLM で生成したある誤答選択肢が，実際の誤答選択肢と一対一対応となるため，類似した誤答選択肢がともに高い評価を得ることを抑制できる.

表 1　各生成手法とその評価指標の値

	Valid	R	P	R_s	P_s	R_{cs}	P_{cs}	DV
SFT_0.0	**236**	**0.109**	**0.109**	**0.896**	**0.901**	**0.886**	**0.886**	0.106
Five-Shot_0.0	234	0.097	0.097	0.896	0.898	0.884	0.884	0.118
Zero-Shot_0.0	232	0.102	0.102	0.892	0.895	0.881	0.881	0.109
Proposed_0.2-0.6	**236**	0.076	0.076	0.894	0.895	0.880	0.880	0.121
Proposed_0.8-1.2	235	0.060	0.060	0.889	0.887	0.874	0.874	**0.134**

DV は次のように定義される．

$$\mathrm{DV} = \frac{1}{|\mathbf{Q}|}\sum_{i=1}^{|\mathbf{Q}|}\left(1 - \frac{\sum_{\{(s_j,s_k)|s_j,s_k\in S_i, j\neq k\}} \mathrm{sim}(s_j, s_k)}{|S_i|}\right) \tag{8}$$

DV は，全ての誤答選択肢のペアについて，コサイン距離を平均したものである．よって，この値は誤答選択肢同士の類似度が高いほど低い値となる．したがって，この指標は誤答選択肢同士がどれだけ多様かを示す指標と言える．

4　実験結果

実験結果を表 1 に示す．提案手法においては temperature = (0.2,0.4,0.6) の組で生成させたものを Proposed_0.2-0.6，(0.8,1.0,1.2) の組で生成させたものを Proposed_0.8-1.2 としている．参考として，評価データセットにおける実際の誤答選択肢の DV の値は 0.1325 である．表 1 の Valid は，評価指標の計算に使用された問題の総数を表す．評価対象外となったケースとして，LLM が指定した個数と異なる個数の誤答選択肢を生成したケースや，生成時の出力形式が遵守されなかったケースなどが挙げられる．

表 1 より，提案手法は DV の点でベースライン 3 手法を大きく上回っている．よって，ベースラインと比較して，提案手法はより多様な誤答選択肢を生成できたと言える．加えて，Proposed_0.8-1.2 において，DV は実際の試験問題と同程度の値となっており，生成した誤答選択肢が，実際の誤答選択肢と同程度の多様性を持つと言える．また誤答選択肢の品質を表す評価指標 R，P，R_s，P_s，R_{cs}，P_{cs} についても，多様性と品質にはトレードオフの関係があるが，提案手法はどちらも Five-Shot，SFT と比較して高い水準を保っている．Proposed_0.2-0.6 において，R_s，P_s，R_{cs}，P_{cs} は Zero-Shot と同程度の値となっており，多様性を大幅に向上しながら誤答選択肢の品質を維持できていると言える．Temperature と評価指標の詳細な関係は付録 B に示す．

問題: 心周期に伴う心臓の変化で，収縮期の初期には心室の容積は変わらずに内圧が上昇していく．このときの心臓で正しいのはどれか．
正解: 左心室の内圧は大動脈圧よりも低い．
Five-Shot で作成した誤答選択肢:
左心室の内圧は大動脈圧よりも高い．
左心室の内圧は収縮期を通して一定である．
左心室の内圧は拡張期に上昇する．
Proposed_0.2-0.6 で作成した誤答選択肢:
大動脈弁は開いている．
心室の壁は薄くなっている．
心室の内圧は一定である．

図 5　作成した誤答選択肢の例

作成した誤答選択肢の例を図 5 に示す．図 5 において，Five-Shot で生成した誤答選択肢はすべて「左心室の内圧は」を含んでおり，誤答選択肢同士の多様性が失われてしまっている．一方，提案手法 Proposed_0.2-0.6 で作成した誤答選択肢は「左心室の内圧」以外の項目も含み，多様性が向上していることが確認できる．

5　おわりに

本研究では，看護師国家試験の一般問題を題材として，誤答選択肢同士の多様性の問題に着目した．提案手法では，生成結果の多様性を高めるために temperature を高い値に設定し，複数の temperature で得られた候補を集約した．加えて，非妥当な誤答選択肢を除外し，適切な誤答選択肢を抽出する目的で LLM-as-a-Judge を用いた自動判定を導入した．その結果，本手法により，実際の試験問題と同程度の多様性を持つ誤答選択肢を，高い品質を維持しながら自動作成することに成功した．

一方，本研究により作成した誤答選択肢が，実際の作問現場でどれほど活用可能かについては研究の余地があり，今後の課題としたい．

謝辞

本研究は厚生労働科学研究費補助金 政策科学総合研究事業（臨床研究等 ICT 基盤構築・人工知能実装研究事業）JPMH25AC1001 の交付を受けたものです．本研究は，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施しました．

参考文献

[1] Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 14437–14458, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[2] Yooseop Lee, Suin Kim, and Yohan Jo. Generating plausible distractors for multiple-choice questions via student choice prediction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 23669–23692, Vienna, Austria, July 2025. Association for Computational Linguistics.

[3] 保健師助産師看護師国家試験出題基準 令和 5 年度版, (2026-1 閲覧). https://www.mhlw.go.jp/content/10803000/000958440.pdf.

[4] Yûsei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyo, and Naoko Hayashi. **Automatic Question Generation for the Japanese National Nursing Examination Using Large Language Models**. SciTePress, 2024.

[5] 城戸祐世, 山田寛章, 徳永健伸, 木村理加, 三浦友理子, 佐居由美, 林直子. 言語モデルを用いた看護師国家試験問題の誤答選択肢自動生成. 言語処理学会 第 31 回年次大会 発表論文集, pp. 1074–1079. 言語処理学会, 言語処理学会事務局, 2025.

[6] Rajiv Jhangiani. Guidelines for writing effective distractors for multiple-choice questions, (Accessed 2026-1). https://thatpsychprof.com/wp-content/uploads/2016/12/Guidelines-for-writing-effective-distractors-for-multiple-choice-questions.pdf.

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan et al. The llama 3 herd of models, 2024.

[9] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[10] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[11] Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. Building instruction-tuning datasets from human-written instructions with open-weight large language models, 2025.

[12] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. **arXiv preprint arXiv:2402.05672**, 2024.

付録

A　実験設定の詳細

ファインチューニング時のハイパーパラメータは，LoRA ランク 8，学習率 10^{-4}，学習回数 10 としている．

B　temperature による各種評価指標の変化

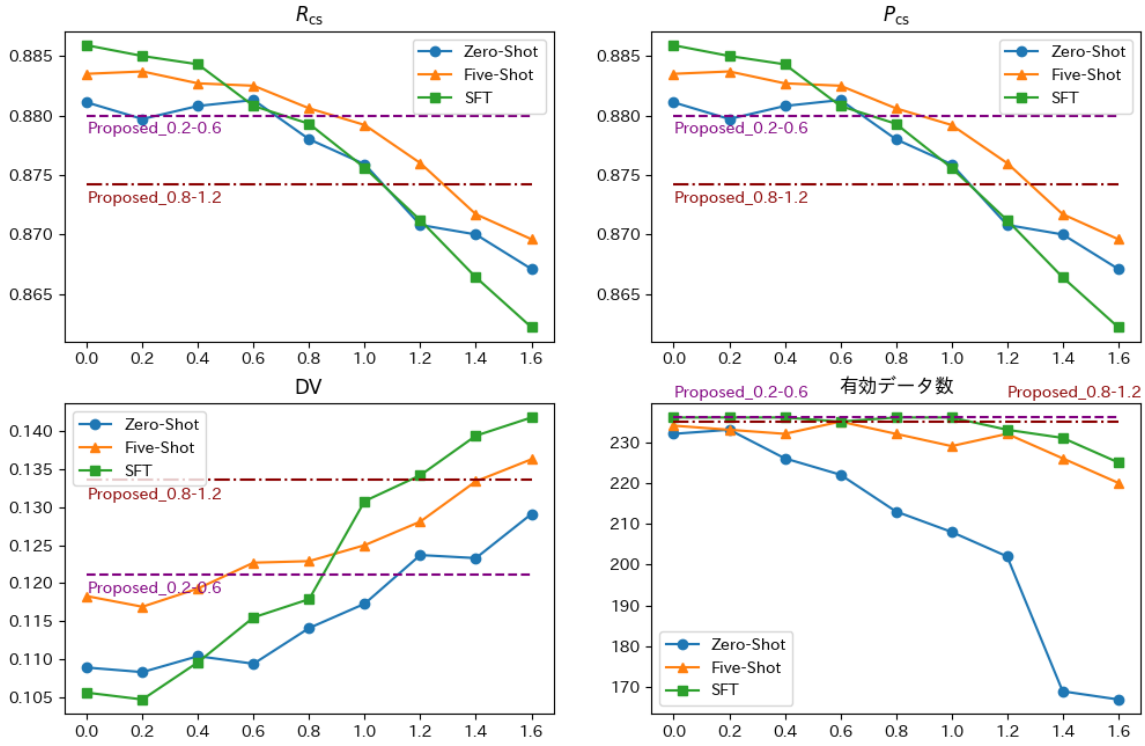

図 6　temperature の変化と R_{cs}，P_{cs}，DV，有効データ数の関係

図 6 に，temperature を変化させた場合における，R_{cs}，P_{cs}，DV，有効データ数の変化を示す．R_{cs}，P_{cs} においては，temperature が上昇するにつれて評価指標の値が減少する傾向にあることが読み取れる．これは，temperature を上げると LLM が実際の誤答選択肢から離れた誤答選択肢を生成してしまうことを示す．一方，DV については，temperature が上昇するにつれて値は上昇することが読み取れる．これは，temperature を上げることで誤答選択肢同士の多様性を高めることができることを示している．

加えて，temperature を上げる手法は，SFT や Five-Shot と相性が良いことに注目する．temperature = 0 において，SFT における DV は，Five-Shot における値よりも低くなっており，ファインチューニングを行うと，誤答選択肢の多様性が下がってしまう問題がある．しかし，有効データ数においては，Five-Shot，SFT では，Zero-Shot に比べ，1.0 より大きい temperature においても高い値を維持している．一般に，temperature を上げることでモデルの出力の分布に自由度が生まれるため，指定した出力形式を遵守しない出力を上げる確率が高まり，利用できるデータが少なくなってしまう問題がある．しかし，Five-Shot で SFT で出力形式の例を与えることで，高い temperature の環境においても正しい形式で出力を行うことができる可能性が高まる．よって，Five-Shot，SFT では，Zero-Shot に比べて高い temperature で生成できることができるので，その分多様性を確保する上で利点があると言える．