自由記述回答から選択肢設問を生成するモデルの構築と MCA 参照空間への射影による生成内容の解釈

根本颯汰 [1] 土井智暉 [2] 花田智洋 [3] 谷中瞳 [2] 彌冨仁 [1] 藤本一男 [3,4]
[1] 法政大学大学院 理工学研究科 応用情報工学専攻
[2] 東京大学大学院 情報理工学系研究科 コンピュータ科学専攻
[3] 国立研究開発法人 情報通信研究機構（NICT）サイバーセキュリティ研究所
[4] 津田塾大学 数学・計算機科学研究所
sota.nemoto.5s@gmail.com, iyatomi@hosei.ac.jp
{hanada, kazuo.fujimoto}@nict.go.jp
{doi-tomoki701, hyanaka}@is.s.u-tokyo.ac.jp, fujimoto@tsuda.ac.jp

概要

自由記述回答から選択肢設問の生成は LLM の発展に伴い注目されており, 調査票の改善と回答者や分析者の負担を減らすことができる. 本稿では, データ漏洩のリスクを考慮して, Open-source LLM を組み合わせた自由記述回答から選択肢設問を生成するモデルを構築し, その生成設問を多重対応分析 (MCA) による参照空間に射影し, 解釈を行った. 以上を通して, 構築したモデルの機能的有効性を確認した. また, 類似度が高く, 多様性が低いとされる生成設問でも MCA による幾何学的分析によって解釈可能であることを確認できた.

1 はじめに

調査票 (本稿では従来からの紙の調査票に限らず Web ページによる調査票も含めて調査票と呼ぶ) は, 調査対象に対するセンサーの役割を果たす. それゆえ, 調査票の設計に不備があった場合, データの収集後の統計処理によっては, 収集されてない情報の「回復」は原理的に不可能である. 調査票は, 構造化された選択肢回答部分 (close-end) と非構造化部分である自由記述回答部分 (open-end) によって構成されている. テキスト処理が未発達の段階では, 採取した自由記述回答の分析は困難であったが, 昨今の NLP の発展の中で自由記述部分の活用がひらけてきた.

調査票の構造化部分は, ある時点でのスナップショットとしての調査仮説に対応しているが, そこに拾いきれない要素は必ず存在するため, 調査票の設計では常に更新作業が求められる. 求められる新

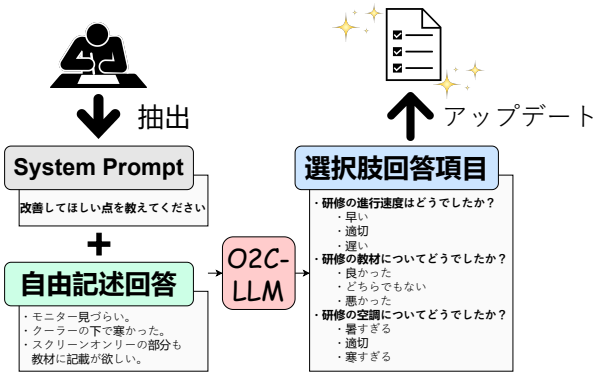

図 1 自由記述回答から選択肢設問生成の流れ

たな視点は, 一般的に外的要素からもたらされるが, 調査結果自身のなかにもヒントはある. それが, 自由記述部分である. こうした調査票設計の現場における要請のサポートとして, 我々は次回以降の調査票の改善のために初回の回答結果に含まれた自由記述回答を元に選択肢設問を作成する機能を開発した.

本稿では, 調査票改善に向けた選択肢設問生成モデルの構築と, その生成した選択肢設問の評価と分析を行う. まず, 自由記述回答から選択肢設問を生成するモデル Open-end2Close-end LLM (O2C-LLM) を構築する. 我々はデータ漏洩のリスクを考慮して, GPT [1] などの Closed-source LLM ではなく Llama [2] などの Open-source LLM を採用した. 次に, O2C-LLM の生成した設問を評価, 分析した. 評価において, 我々は 1) Format violation, 2) 生成した設問の多様性, 3) 自由記述回答と選択肢設問の紐付け の 3 つの観点から評価した. また, 分析において, MCA を用いて採取された選択肢回答の集計表から回答者空間を生成し, その空間に生成した設問を射影することで, ど

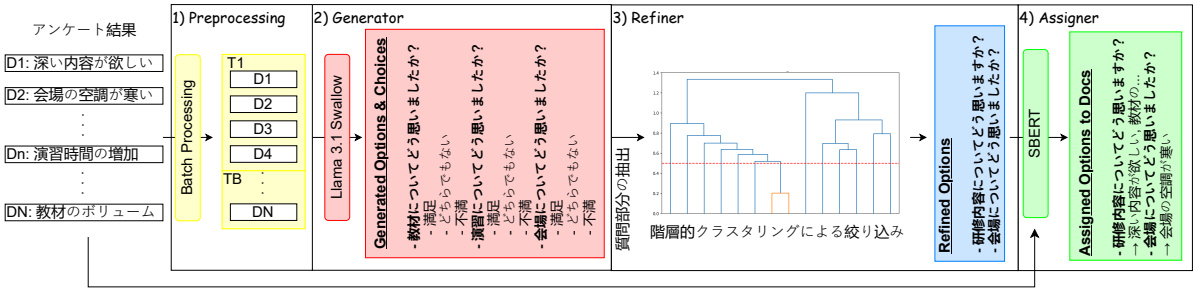

図 2　Open-end2Close-end LLM の図

のような回答者がどのような選択肢設問を生成するかという傾向を分析可能にした. その結果, 生成設問としては類似性があっても, MCA によって生成された参照空間での幾何学的配置を考慮することで特異性を確認できることを示し, システムとして探索的なアプローチが可能であることを示す.

2　Open-end2Close-end LLM

我々の選択肢設問生成タスクで使用したモデル Open-end2Close-end LLM (O2C-LLM) を図 2 に示す. O2C-LLM は 1) Preprocessing, 2) Generator, 3) Refiner, 4) Assigner の 4 つの Phase から構成される.

Preprocessing　は処理を高速化するために自由記述回答群 $\boldsymbol{d}_K = [d_1, \cdots, d_k, \cdots, d_K]$ (K は回答数) を任意のバッチサイズ B ごとに結合してミニバッチ自由記述回答群 $\boldsymbol{T}_I = [\boldsymbol{t}_1, \cdots, \boldsymbol{t}_i, \cdots, \boldsymbol{t}_I]$ (I はイテレータ数, $\boldsymbol{t}_i = [\boldsymbol{d}_i, \cdots, \boldsymbol{d}_{i\times b}, \cdots, \boldsymbol{d}_{i\times B}]$) を得る.

Generator　は Preprocessing から得られたミニバッチ自由記述回答 $\boldsymbol{t}_i$ と few-shot 事例を追加したプロンプト p_i を Open-source LLM に入力して選択肢設問 $\boldsymbol{g}_i^{(M)} = [g_i^{(1)}, \cdots, g_i^{(m)}, \cdots, g_i^{(M)}]$ (M はバッチごとに生成された設問数) を生成させる. (A.1 参照)

Refiner　では Generator から得られた設問群 $\boldsymbol{G}_I^{(M)} = [\boldsymbol{g}_1^{(M)}, \cdots, \boldsymbol{g}_i^{(M)}, \cdots, \boldsymbol{g}_I^{(M)}]$ から質問文のみをルールベースで抽出し, Sentence-BERT [3] で埋め込み表現を取得して, 階層的クラスタリングで低頻度かつ類似した質問を統合する. 階層的クラスタリングを採用することで距離に基づいた閾値処理によって設計者が任意の数の設問が生成できる. これによってより洗練された設問 $\boldsymbol{r}_L = [r_1, \cdots, r_l, \cdots, r_L]$ に絞り込むことができる. (L は Refiner によって絞り込まれた後の設問数)

Assigner　では O2C-LLM の結果をより分析しやすくするため, Sentence-BERT [3] による埋め込み表現の類似度に基づいて, 自由記述回答群 $\boldsymbol{d}_K$ と生成した設問群 $\boldsymbol{r}_L$ の紐付けを行う.

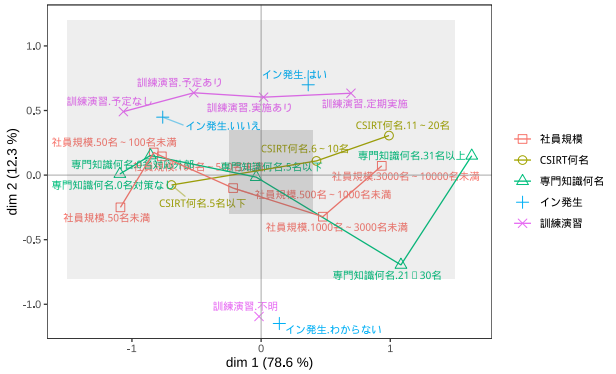

図 3　所属組織のセキュリティ対応水準による MCA 変数空間. 編みかけに生成設問を射影. 内側に図 4, 5 (閾値 1.5, 1.0), 外側に図 6 (閾値 0.5)

3　実験

3.1　データセット

我々が用いたデータは, NICT ナショナルサイバートレーニングセンターが実施する CYDER [1] の 2020 年度 A コース受講者の 2,001 レコードからなるデータである. 自由記述の回答率は 60%である. このデータセットの構造は, 選択肢回答部分 (20 問) と自由記述回答部分 (5 問) によって構成される. 情報セキュリティ対策のトレーニングだが, インシデントレスポンスを中心に構成されるため, この調査票のテーマは限定的かつ明確である.

3.2　生成設問の評価指標

Format violation　我々はプロンプトで指定したフォーマット違反を評価するために, 以下の 2 つの評価指標を採用した. **Format Acc.** は生成設問 $\boldsymbol{G}_I^{(M)}$ が箇条書きの形式 (A.2 参照) を遵守している割合を

1)　CYDER（Cyber Defense Exercise with Recurrence：実践的サイバー防御演習）とは, サイバー攻撃を受けた際のインシデント対応をロールプレイ形式で体験できる演習 [4], [5]

表 1　O2C-LLM の各観点ごとの評価

	Format violation		Generator performance			Assigner performance			
Model	Format Acc. ↑	Length Acc. ↑	Distinct-1 ↑	Distinct-2 ↑	Similarity ↓	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	Similarity ↓
few-shot O2C-LLM	0.831	0.964	0.475	0.824	0.665	0.313	0.128	0.295	0.680

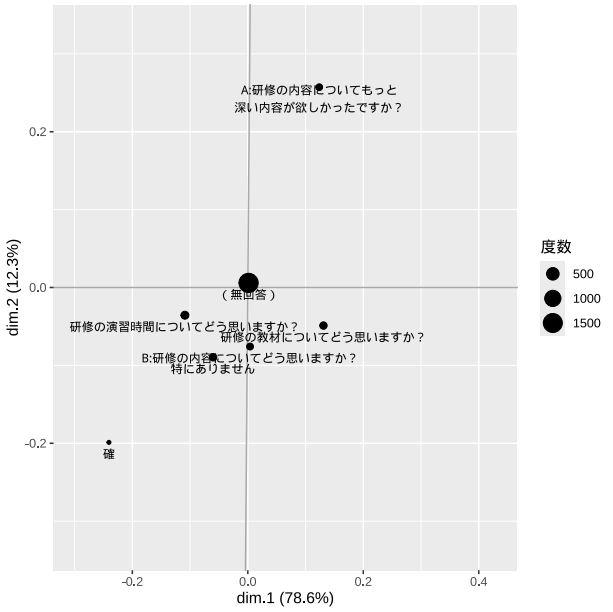

図 4　生成設問 (閾値 1.5) の変数空間内配置

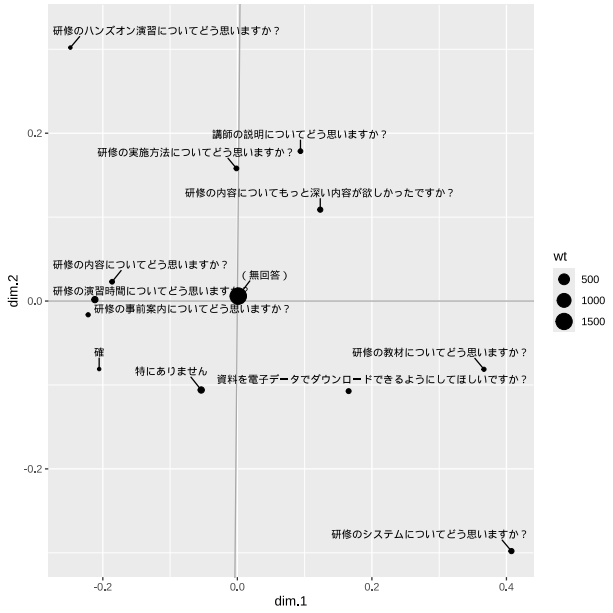

図 5　生成設問 (閾値 1.0) の分布

指す. **Length Acc.** は生成設問 $\boldsymbol{G}_I^{(M)}$ が設問の適切な文字数である 20 単語以下 [6] を満たす割合を指す.

Generator performance　我々は生成設問 $\boldsymbol{G}_I^{(M)}$ 間の多様性を評価するために, 以下の 2 つの評価指標を採用した. **Distinct-1, 2** [7] は生成設問 $\boldsymbol{G}_I^{(M)}$ の N-gram ($N \in \{1, 2\}$) におけるユニークな単語数と総単語数の比率である. **Similarity** は Sentence-BERT [3] による生成設問間の類似度を算出する.

Assigner performance　我々は自由記述回答 $\boldsymbol{d}_K$ と生成した選択肢設問 $\boldsymbol{r}_L$ 間の類似度を評価するために, 以下の 3 つの評価指標を採用した. **ROUGE-1, 2** [8] は N-gram における一致率を指し, **ROUGE-L** [8] は Longest Common Subsequence (LCS, 最長共通部分列) に基づいた一致率を指す. また, Generator の時と同様に **Similarity** も使用した.

3.3　MCA による変数空間と生成設問の射影

我々はこれまで自由記述部分と選択肢回答部分の連携した分析方法の開発を行なっている [9, 10, 11]. この手法に基づいて, O2C-LLM が生成した設問について多重対応分析（MCA）を用いた幾何学的な解釈を行った [12]. MCA とは, カテゴリカルデータである回答選択肢の回答パターンを反映した参照空間に, 自由記述からテーマ抽出などの方法で生成した変数をプロット可能にする. これにより, 生成した参照空間の解釈を豊富化するのみならず, 自由記述部分から生成された設問に対する評価も可能になる.

MCA による参照空間への追加変数のプロット　選択された変数のデータ表に対して MCA を行うと, 回答者空間と変数空間が生成される. 変数空間には, 投入された変数カテゴリの相互の関係が幾何学的にプロットされ, 回答者空間ではその回答選択肢に対応した個体の位置がプロットされる [12]. 例えば, 生成された空間に対してデモグラフィック変数を射影し, 空間構造の解釈が可能になる. また MCA ではそれぞれの空間にその空間生成には寄与しなかった変数を「追加変数」として射影できる [13].

今回我々は, 選択肢設問が生成した変数空間に対し, その追加変数として生成設問を割り当て, 選択肢回答が形成する幾何学的位置を用いた解釈をした.

4　結果

4.1　O2C-LLM の性能

選択肢設問の生成に対するスコアを表 2 に示す. Format violation では先行研究 [14] で懸念されるほど

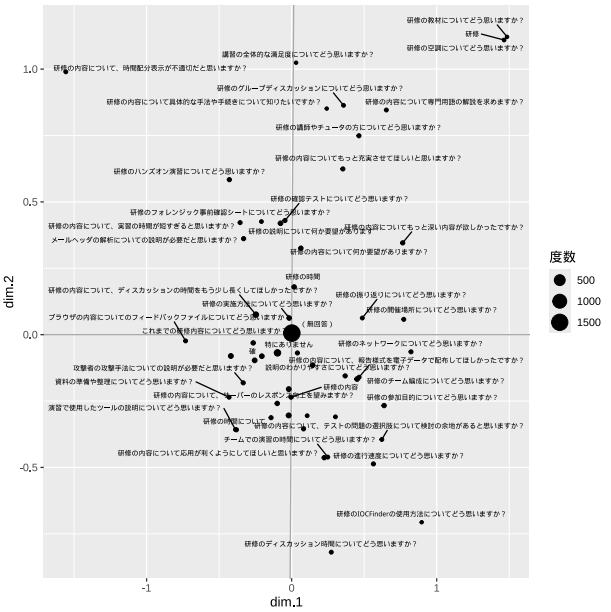

図 6　生成設問 (閾値 0.5) の分布

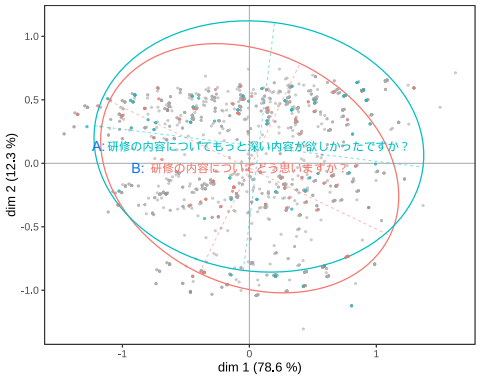

図 7　生成設問 (閾値 1.5) の設問 A, B を生成した個体の集中楕円

のフォーマット違反は発生せず, 一方で Generator や Assigner の Similarity の値は高く, Distinct や ROUGE の値は低い結果となった. つまり, 生成設問同士は類似度が高く, 多様性の低いことがわかった.

4.2　MCA による生成設問の射影と分析

MCA による参照空間は受講者の所属組織のセキュリティ対策に関連する変数によって生成された (図 3 参照) . この空間の 1 軸は, 78.6%, 2 軸は, 12.3% の分散を体現し, 1–2 軸で選択された変数による分散の 90.9%が表現されるため, 分析は 1–2 軸平面で行うことが正当化される. 変数ごとに回答カテゴリを結ぶことによって, 空間の構造が明らかになる. 1 軸は左から右にかけて所属組織の規模に対応し, セキュリティ対策 (配置されている CSIRT の人数, 訓練など) が組織規模に関連して分布した. 2 軸の下には, インシデント：わからない, 訓練演習：不明 があり, 上には, インシデントはい/いいえ, 訓練演習への自覚, がプロットされた. この軸はセキュリティ対策状況に対する自覚度合いを示す軸となった.

この空間に追加変数として, O2C-LLM による生成設問をプロットした. 図 4 は図 3 の編みかけの部分に位置づく. 我々は Refiner で言及した階層的クラスタリングの閾値, 1.5, 1.0, 0.5 ごとの生成設問ごとに検討した. (図 5, 6) 閾値 1.5 は 6 問, 1.0 は 11 問, 0.5 は 53 問が生成され, それらは階層構造を形成する.

生成設問と参照空間へのプロットの実例を用いて, 文字列として似ているが, 幾何学的な位置関係から独自性を検討した例を以下に示す.

- まず閾値 1.5 の生成変数をプロットする.(図 4)
- そこで設問 A「研修の内容についてもっと深い内容が欲しかったですか？」と設問 B「研修の内容についてどう思いますか？」の 2 つの生成設問に注目する.
- A は B の下位設問として包含可能なのか, それともある種の特異性を持ったものなのかを参照空間の幾何学的位置から検証する.
- この 2 つの変数空間上の位置を典型性検定と, 同質性検定を行い, 2 軸方向で差異が認められたことを確認する [12].

以上の検証から, この設問生成は有効であると判断できる (詳細は A.4 参照). 実際には閾値 1.0 や 0.5 の下位設問を参照することで, 生成設問を調査票の構造を考慮にいれた有効性の評価が可能になる.

5　おわりに

自由記述回答から選択肢設問を生成するモデル O2C-LLM を構築し, 性能の評価と MCA による分析を行った. O2C-LLM は一定のフォーマットを遵守した設問生成ができた一方で, 設問同士の類似度が高く, 多様性が低い結果となった. しかし, 我々は MCA による分析を通じて, 評価指標に反してでも新たな設問として設置すべきケースがあることが確認できた. この点を我々は現状の生成テキストの評価時の課題点とみなし, MCA を交えた新たな実用的な評価指標の設置と更なる選択肢設問生成の改善を将来の展望とする.

謝辞

本研究の一部は JSPS 科研費学術変革領域研究（B）「ナラティブ意識学」JP24H00809 の支援を, また, JSPS 科研費基盤研究 (C)JP20K02162 の支援を受けたものである.

なお, 本研究にあたり CYDER 実施の中心センターである NICT ナショナルサイバートレーニングセンターの園田道夫センター長にはデータ利用を含め, さまざまな便宜を図っていただきました. また, 中川哲也氏, 阿部則夫氏にはアンケートデータの構成, CYDER のシナリオ, コースプログラムに関する質問に対応していただきました. 記して感謝いたします.

MCA の実行環境は、NICT 所内 VM で動作している、R4.4.1[15]、RStudio2024[16] 上で、GDAtools2.1[17] を用いている。

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[3] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019.

[4] NICT. CYDER（cyber defense exercise with recurrence：実践的サイバー防御演習）, 2022. https://cyder.nict.go.jp/.

[5] 中川哲也, 溝渕智規, 阿部則夫. CYDER オンライン演習への挑戦. 月刊 J-LIS Jan 2025:12-21, 2025.

[6] A Williams. How to··· write and analyse a questionnaire. **Journal of orthodontics**, Vol. 30, No. 3, pp. 245–252, 2003.

[7] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. **arXiv preprint arXiv:1510.03055**, 2015.

[8] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.

[9] 藤本一男, 大畑和也. 多重対応分析とアスペクトベース感情分析を組み合わせた受講者満足度調査データ分析手法の開発. 言語処理学会 第 29 回年次大会 発表論文集, pp. 255–260, 2023.

[10] 根本颯汰, 藤本一男. クラスタリングによる自由記述回答の要約と選択肢回答空間に射影による解答群間の連関の可視化. 言語処理学会 第 30 回年次大会 発表論文集, pp. 456–460, 2024.

[11] 根本颯汰, 藤本一男. 4-2 構造化データ解析で選択肢回答空間と機械学習による自由記述部分のクラスタリング結果を結合する − CYDER 受講者アンケートの分析手法の開発と今後の課題 −. 情報通信研究機構研究報告, Vol. 70, No. 2, pp. 135–146, 2025.

[12] Brigitte Le Roux, Henry Rouanet, 訳：大隅昇, 小野裕亮, 鳰真紀子. Multiple Correspondence Analysis(多重対応分析). SAGE publisher（オーム社）, 2010(2021).

[13] Michael J. Greenacre, 訳：藤本一男. Correspondence analysis in practice Third edition（対応分析の理論と実践）. Chapman & Hall/CRC interdisciplinary statistics series. CRC Press, Taylor & Francis Group（オーム社）, Boca Raton, 2017(2020).

[14] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. Topicgpt: A prompt-based topic modeling framework. **arXiv preprint arXiv:2311.01449**, 2023.

[15] R Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[16] Posit team. **RStudio: Integrated Development Environment for R**. Posit Software, PBC, Boston, MA, 2024.

[17] Nicolas Robette. **GDAtools: Geometric Data Analysis 2.1**, 2024.

A Appendix

A.1 Generator のプロンプト

Prompt:
System: あなたは誠実で優秀な日本人のアンケートの設計者です。
User: 以下の自由記述回答結果を踏まえて、次回のアンケート実施のために設置すべき選択肢設問を作成してください。生成する選択肢設問は 20 単語以内にしてください。
Examples:
[自由記述回答結果]
{few shot 用の自由記述回答 }

Assistant:
{few shot 用の選択肢設問と選択肢 }

User:
[自由記述回答結果]
{ 選択肢設問を生成したい自由記述回答群 t_b}

Assistant:

A.2 生成する選択肢設問の例

- これまでの研修内容についてどう思いますか？
  - とても満足
  - 満足
  - どちらでもない
  - 不満
  - どちらでもない
  - とても不満
- 内容を復習する機会が必要ですか？
  - 必要
  - どちらでもない
  - 必要ない
- 最後のテストで何らかの不具合が発生しましたか？
  - 発生した
  - 発生しなかった

表 2 典型性検定、同質性検定の検定統計量と p 値

典型性検定(Typicality test)	dim.1		dim.2	
	検定統計量	p値	検定統計量	p値
A:研修の内容についてもっと深い内容が欲しかったですか？	1.005	0.315	-0.477	0.036
B:研修の内容についてどう思いますか？	0.033	0.973	2.091	0.513

帰無仮説：検討している部分集合は、全体集合の分布と同じである（典型である）

同質性検定(Homogenity test)	dim.1		dim.2	
	生成設問A:	生成設問B:	生成設問A:	生成設問B:
A:研修の内容についてもっと深い内容が欲しかったですか？	1	0.485	1	0.053
B:研修の内容についてどう思いますか？	0.485	1	0.053	1

帰無仮説：二つの部分集合の分布は同質である。

A.3 O2C-LLM 実装の詳細

我々は Generator における LLM として Llama 3.1 Swallow 8B [2] を使用した. また, この Generator で使用した few-shot は事前に一度 Generator によって生成された結果を使用した. その時に使用した自由記述回答群は評価時には取り除いてある.

A.4 生成設問を変数空間上の幾何学的配置で分析

MCA で生成された変数空間を参照空間として, O2C-LLM の生成変数 (閾値 1.5) のうちの以下の設問 A と B の分布を比較する.(表 2)

- A: 研修の内容についてもっと深い内容が欲しかったですか？
- B: 研修の内容についてどう思いますか？

典型性検定では, 全体の分布に対して下位集団 (ここでは生成設問 A や B) の分布が「典型的」であるかどうかを検定している. 第 1 軸では p 値は 0.315 であるため, 帰無仮説 (その軸に関して当該の下位集団は典型的である) は棄却されない. また, 生成設問 B は 0.973 と 1 に限りなく近いため, 全体の分布と同じであることが示唆される. 一方第 2 軸については, 図 7 からも想定されるように全体の分布とは異なる傾向が確認されており, 実際に p 値も 0.036 と小さい. 結果として, この生成設問にある「もっと深い内容」について掘り下げていく意義がある.

同質性検定は, 二つの下位集団同士が同じ分布かどうかの検定である. これも第 2 軸では, 5%水準に拘泥するならば帰無仮説を棄却できない値であるが (p=0.053), 非類似である目処をたてられる値である.

2) https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2