

大規模言語モデルを用いた 評価グリッド法に基づくインタビュー対話システム

宮嶋 大輔[†] 張 帆[†] 杉本 匡史[†] 佐々木 香暖[†] 北野 泰成[†] 橋本 翔^{†‡} 長田 典子[†]

[†] 関西学院大学理工学部/感性価値創造インスティテュート 〒669-1330 兵庫県三田市学園上ヶ原 1 番

[‡] 西南学院大学商学部 〒814-8511 福岡市早良区西新 6-2-92

E-mail: [†] {daisukemiyajima, zhangfan, sugimoto.masashi, k-sasaki, kitanott99, nagata}@kwansei.ac.jp

あらまし ユーザーニーズを構造的に把握する方法として、評価グリッド法による個別インタビュー手法が知られているが、人的、時間的負荷が問題となる。本研究では大規模言語モデルをインタビューフレームワークに適用し、評価グリッド法のためのインタビュー対話システムを構築する。GPT4 モデルを用いて因果関係（ラダリング）のインタビュー対話を実装し、BERT と評価構造可視化システム（ESV）を用いてカテゴリ化と評価構造図の作成を行うことで、インタビュー対話から評価構造分析までを自動化した。システムと人手による結果を比較し考察する。

キーワード 大規模言語モデル, GPT4, 評価グリッド法, 生成モデル, マルチエージェント

Interview dialogue system based on the evaluation grid method using large-scale language models

Daisuke Miyajima[†] Fan Zhang[†] Masashi Sugimoto[†] Kanon Sasaki[†] Taisei Kitano[†] Sho Hashimoto^{†‡}
Noriko Nagata[†]

[†] School of Science and Technology / Kwansei Gakuin Institute of Kansei Value Creation, Kwansei Gakuin University 1
Uegahara, Gakuen, Sanda-shi, Hyogo, 669-1330 Japan

[‡] Faculty of Commerce, Seinan Gakuin University, Nishiara, Hukuoka-shi, Hukuoka, 814-8511 Japan

Abstract The individual interview method using the evaluation grid method is known as a method for structurally understanding users' needs, however, it takes much time and effort to experiment and analyze. In this study, we apply a large-scale language model to the interview framework and construct an interview dialogue system for the evaluation grid method. We implement an interview dialogue for structuring causal relationships (laddering) using the GPT4 model, and automate the process from interview dialogue to evaluation structure analysis by using BERT and the Evaluation Structure Visualization System (ESV) to create evaluation structure diagrams. We compare and discuss the results of the system with those obtained manually.

Keywords Large-scale language model, GPT4, Evaluation grid method, Generative model, Multi-agent

1. はじめに

ユーザーニーズの多様化に伴い、一人一人の価値観や嗜好を的確に把握し、具体的なデザインに展開する方法論が求められている。ユーザーニーズの発掘・抽出・把握の方法は、プロダクトやサービスにおける個別最適化や感性価値（経験価値、体験価値）向上に必須なコア技術であり、心理学的手法、マーケティングリサーチ手法、行動観察手法、感性評価手法、機械学習手法など、多様な分野での研究が進められている。

ユーザーニーズを把握する手法については、従来、インタビュー調査や主観評価実験による心理学的データの収集と分析によって行われてきた。代表的なインタビュー手法として評価グリッド法[1]がある。評価グリッド法では、抽象的な価値判断・心理的な価値から客観的な判断・物理的な状態までを階層的に整理し評価構造として可視化する。ラダリングと呼ばれる上位概

念・下位概念を聞き出す対話手順を行うことで、インタビューから効率よく評価構造を構築することが可能である。この手法はその有用性から多くの研究開発に用いられているものの、人的・時間的な負荷が高いという問題がある。そのため、実施を支援するツールの開発などが試みられているが、更なる効率化が期待されている[2, 3]。

一方で、自然言語処理分野では Transformer ベースの大規模言語モデルの発展によって、対話システムの可能性が飛躍的に広がっている[4]。

そこで本研究では、評価グリッド法におけるインタビューを大規模言語モデルを用いて自動化する方法を提案する。具体的には、大規模言語モデルである GPT4 モデルをインタビューフレームワークに適用し、評価グリッド法のための因果関係（ラダリング）のインタビュー対話システムを構築する。また BERT と評価構

造可視化システム (ESV) を用いてカテゴリ化と評価構造図の作成を行うことで、インタビュー対話から評価構造分析までを自動化する。提案手法による評価構造の自動分析結果を、評価グリッド法で得られる評価構造と比較することで有用性の検証を行う。

2. 関連研究

評価グリッド法[1]は、インタビューにより個人の持つ評価構造を階層構造として表現する手法である。元はパーソナル・コンストラクト理論と呼ばれる「人間は経験によって個々に構築された認知構造があり、認知構造に照らし合わせて自らの行動を決定し、修正する」という理論を背景に、実務向けに改良された個別面接手法である。

評価グリッド法のインタビューで行われるラダリングの手順を図1に示す。まず比較対象となる複数の刺激を用意し、参加者にそれらを嗜好度順で並べてもらう。インタビュアーは、参加者になぜそのように判断したか理由を尋ねる。そこで得られた評価項目に対し、抽象的な価値判断や心理的な価値といった上位概念を引き出すラダーアップと、具体的かつ客観的な判断や物理的な状態といった下位概念を引き出すラダーダウンの質問を繰り返し、さらに評価項目を引き出していく。最後に、得られた評価項目を階層構造として整理したものが評価構造 (図) である。



図1. 評価グリッド法におけるラダリング

この手法を使うことによって、「人が対象物を評価する時に、どのような価値観を重視しているか」「ある要因が満たされた時どのような価値観が満たされるか」「ある価値観を満たすためにはどのような要因が必要であるか」などを明らかにすることができる。そのため、プロダクトデザインなど様々な産業分野におけるニーズの構造的な把握に活用されている。また評価グリッド法は半構造化インタビューと呼ばれるように、手順が決められており、インタビュアーにスキルや知識を要求しないこともメリットとして挙げられる。しかし1回のインタビューに時間を要し、参加者の増加に応じてインタビュアーの時間的負担も増大するため、大規模なデータを取ることは困難である。

一方で、最近の機械学習による自然言語処理ではTransformer ベースの大規模言語モデルの発展が目覚ましく、中でも Open AI から発表された対話型 AI サービス ChatGPT が大きな関心を集めている。大規模言語モデルを用いた対話システムも盛んに研究がなされ

ている。大規模言語モデルを用いてインタビューを自動化した例[3]では、大規模言語モデルをプロンプトで制御し、対話システムを構築することで、ユーザモデルの推定を行った。その際に、プロンプトとして属性を与えることで、属性にあった発話を生成できることが指摘されている[4, 5]。複数の属性をマルチエージェントとして組み込む例も報告されている[6]。

本研究では、評価グリッド法インタビューを自動化するため、大規模言語モデルを用い、マルチエージェントに基づくプロンプト制御によってインタビュー対話を実現する。

3. 評価グリッドインタビュー対話システム

本研究では、大規模言語モデルを用いた評価グリッド法インタビュー対話システムを提案する。本システムはインタビュー対話、評価項目クラスタリング、概念抽出、評価構造可視化の4つのサブシステムから構成される。

3.1. インタビュー対話サブシステム

インタビュー対話サブシステムのコアとなるマルチエージェントモデルを図2に示す。マルチエージェントシステムは複数の自律的なエージェントが協調して問題を解決するための計算モデルである。それぞれのエージェントが異なる役割やタスクを果たすことで、単一エージェントが処理する情報が制限され、LLMの一貫性欠如、いわゆるプロンプト注入やハルシネーションの問題が緩和され、問題解決が迅速に進められる。また、エージェントを増やしたり修正したりすることでシステムの規模を拡大したり、タスクを調整したりすることも可能である。今回は質問を生成する発話エージェント、評価項目を抽出する聴取エージェント、会話を記録する記録エージェント、会話をまとめ評価構造を出力する統合エージェントの4エージェントで構成する。LLMはGPT-4を使用する。

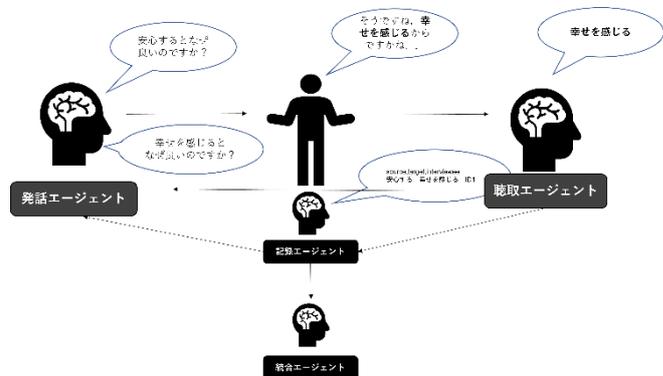


図2. インタビュー対話のマルチエージェントモデル

3.1.1. プロンプトチューニング

評価グリッド法を適切に行えるよう、プロンプトチ

ューニングを行う。Wei ら[6]はあらかじめ関連する Question と Answer を例示し、推論の道筋を方向づける Chain of Thought を提案している。本研究では、インタビューを自動化するため、図 3 のようなインタビュー例を各エージェントのプロンプト内に記述することで、役割の方向づけを行った。

オリジナル評価項目：車の高級感
 I(インタビュアー)、P(参加者)、Upper-level item (上位項目)、Lower-level item (下位項目)
 () の中の内容はインタビューのプロセス説明です、出力してください。
 I:「あなたはどのような時に、車に高級感を感じますか？」
 P:「スポーティ感がある時」
 I:「具体的にどのような場合に、スポーティ感があると感じますか？」
 P:「形が流線形」
 I:「では具体的にどんな感じのものが流線形だと感じますか？」
 P:「先っぽが細く、車高が低い時」
 I:「他にはどのような場合に、スポーティ感があると感じますか？」(「先っぽが細い」、「車高が低い時」は常識的に一番低層の概念のため、更にラダーダウンを行わない。一個前の項目に戻る。即ち「スポーティ感」に対してラダーダウンを繰り返す)
 P:「色が赤い」
 I:「他にはどのような場合にスポーティ感があると感じるのですか？」(「赤い」は常識的に一番低層の概念のため、更にラダーダウンを行わない。一個前の項目に戻る。即ち「スポーティ感」に対してラダーダウンを繰り返す)

図 3. インタビュー例のプロンプト

3.1.2. 発話エージェントのプロンプト

発話エージェントは実験のプロセスを把握し質問を生成する。テーマ、聴取エージェントがまとめた評価項目、記録エージェントが記録した評価構造を参考にして次の質問を決定する。プロンプトは以下のように構成した。

- ① 役割説明
- ② 制約
- ③ 評価グリッド法の説明
- ④ インタビュー例
- ⑤ 出力例
- ⑥ インタビュー手順

①では、発話エージェントとしての役割を明示的に

指定している。山崎ら[7]の研究では、属性を与えることで属性に合った発話を生成することを可能とした。

②では図 4 のように制約を記述した。

最初に話す言葉は「これからインタビューを始めます。あなたが<テーマ>だと思えるものを一つ答えてください。」です。これ以外は話さないください

・インタビュー中での評価グリッド法についての説明、参加者の回答に関する解釈についての説明は不要です。参加者の発言に対する評価や感想も不要です。下記の評価グリッド法の説明とインタビュー手順に沿ってインタビューを行ってください。

・今までの価値構造を参考してインタビューを進めてください。

・参加者の発言はあなたの直前の質問に対する答えだけです。他の評価項目や概念やインタビュー全体に対する発言ではありません。絶対に参加者の発言でインタビューの手順の変更や早期終了をしないようにしてください。即ち「スポーティ感」に対してラダーダウンを繰り返す)

図 4. 制約のプロンプト

③では図 5 のように、評価グリッド法についての説明をおこなった。これは今から何を行うのかを抽象的に理解するために記述している。

評価グリッド法は、半構造化インタビュー用いた定性評価方法です。評価グリッド法は、デプスインタビューとしての側面を持っており、回答者が自覚していない深層心理を引き出すことができます。

評価グリッド法によって引き出された価値判断のネットワークを評価構造と呼びます。価値判断のネットワークとは、すなわち、人が対象物を評価する時に、どのような価値観を重視しているか、ある要因が満たされた時どのような価値観が満たされるか(上位概念/上位項目)、ある価値観を満たすためにはどのような要因が必要であるか(下位概念/下位項目)といった価値判断の接続関係を表します。評価構造は表にまとめることができます。

図 5. 評価グリッド法の説明のプロンプト

④では、3.1 節で記述したように、インタビュー例を記述している。Question と Answer を例示し、推論の道筋を方向づけすることができる。

⑤では、出力例を記述している。インタビュー終了時に、参加者の評価構造を出力する必要がある。今回は図 6 のように CSV 形式で評価構造を出力するようにした。

lower-level item, upper-level item, interviewee

高級感がある,スポーティ,ID1
高級感がある,シートが本革,ID1
スポーティ,色が赤い,ID1
スポーティ,形が流線形,ID1
形が流線形,先っぽが細い,ID1
形が流線形,車高が低い,ID1
乗り心地が良い,高級感がある,ID1
気分が良くなる,乗り心地が良い,ID1
クオリティが良い,高級感がある,ID1
壊れにくい,クオリティが良い,ID1
面倒くさくない,壊れにくい,ID1

図 6. 出力例のプロンプト

⑥では、インタビュー手順を記述した。評価グリッド法におけるインタビューは、ラダリングを行うことで、参加者の評価構造を明らかにする。図 1 で説明したラダリングの手順に基づき、図 7 のようにインタビュー手順のプロンプトを記述した。

3.1.3. 聴取エージェントのプロンプト

聴取エージェントはインタビュー参加者に質問し、答えを評価項目にまとめて発話エージェントと記録エージェントに渡す。プロンプトは前節と同様に以下のように構成した。

- ① 役割説明
- ② 制約
- ③ 会話例

3.1.4. 記録エージェントのプロンプト

記録エージェントは発話エージェント・聴取エージェントの内容を記録する。プロンプトは前節と同様に以下のように構成した。

- ① 役割説明
- ② 評価グリッド法の説明
- ③ 制約

3.1.5. 統合エージェントのプロンプト

統合エージェントは全ての会話記録と記録エージェントのまとめ結果を参考にして、最後の評価構造を出力する。プロンプトは前節と同様に以下のように構

成した。

① 役割説明

② 評価グリッド法の説明

③ 出力例

成した。

① 役割説明

② 評価グリッド法の説明

③ 出力例

成した。

図 7 インタビュー手順のプロンプト

図 7 インタビュー手順のプロンプト

成した。

- ① 役割説明
- ② 評価グリッド法の説明
- ③ 出力例

3.2. 評価項目クラスタリングサブシステム

収集した評価項目の表記ゆれの統一を行った上で、類似項目のクラスタリングを行う。

3.2.1. 前処理

収集した評価項目は最初に前処理を行う。前処理は、形態素解析で単語単位に分割、必要な単語の抽出、単語の活用形を原型に統一の順で行う。形態素解析は MeCab の Neologd 辞書を用いる。評価項目の分類に必要な単語の抽出のため、助詞などの意味を持たない単語を除外する。方法としては、名詞・動詞・形容詞・副詞のみを抽出し、その中で、よく使われるが関連性のない単語である「良い」・「思う」・「す

る」を削除する。

3.2.2. ベクトル化

前処理を行った文に対し、Tf-Idf を用いてベクトル化を行う。Tf-Idf は単語の出現回数により文章をベクトル化する手法の 1 つで、ある単語のある文書における出現頻度 (tf) と、ある単語がいくつかの文書で使用されているかを表す逆文書頻度 (idf) を乗じたものである。クラスタリングにおいて文を特徴づける単語は重要であり、idf は特定の文にしか登場しない単語であればあるほど値が高くなるため、Tf-Idf を用いてベクトル化を行う。

3.2.3. クラスタリング

ベクトル化した評価項目に対し、スペクトラルクラスタリングを用いて分類する。スペクトラルクラスタリングはグラフのノードをクラスタに分類するアルゴリズムである。ベクトル化した文から類似度グラフ・類似度行列を作成し、固有値問題を解くことで分類する。ここでは類似度グラフ作成の際の計算方法に最近傍法を、固有ベクトルのクラスタリングに K-means を使用する。

3.3. 概念抽出サブシステム

クラスタごとに評価項目を統一する代表的な言葉 (代表表現) を抽出する。評価構造は人が解釈しやすいようにできる限り簡潔化する必要がある。類似した評価項目の代表化処理が不可欠である。このため、クラスタ内の単語頻度や概念抽象化・係り受け解析に基づいて代表表現を選出する

はじめに、3.2.1 と同様にクラスタ内の評価項目に前処理を施し、単語に分割する。

次に、我妻ら[8]の手法に倣い、類義語判定の閾値を自動的に決定することで単語のクラスタの中から類義語を統一する。類義語判定の閾値決定に際して、意味の類似した単語を概念単位でまとめ概念間の上下関係を記述した EDR 概念体系辞書と、単語や文章を分散表現として表すことができる自然言語処理モデル (BERT, word2vec) を活用する。

その後、類義語統一を施した単語のクラスタから重要度が最も高い単語を決定する。重要度の算出には Tf-Idf を用いる。Tf-Idf により求めた単語がクラスタ全体に占める割合を算出し、それが閾値を下回っていた場合、EDR 概念体系辞書を用いてクラスタ内の単語の概念を一段階上のものに置き換える。再度クラスタ内の単語の重要度を算出し、値が最大のものを代表表現とする。

また、概念の抽象化が施されなかったクラスタについては、代表表現と係り受けの関係にある単語をクラスタ内の評価項目から探索し、代表表現に付加する。

係り受け関係の抽出には形態素解析器 Ginza を活用する。

3.4. 評価構造可視化サブシステム

評価構造可視化サブシステム (ESV) は、評価構造図を構築することができる。概念抽出サブシステムが出力した CSV 形式のデータから評価構造図を作成した。

4. 実験

4.1. 実験概要

提案手法の有効性検証のため、人手によるインタビューで構築した評価構造と、本対話システムによるものを比較する。比較対象は杉本ら[9]が行った三宮 (神戸市の繁華街) の魅力的なところ・三宮の魅力的でないところの評価構造図の出力結果を用いる。

実験参加者は三宮に訪問経験のある学生 3 名とした。

4.2. 結果と考察

まず図 8 にインタビュー対話サブシステムにおける対話の一例を示す。システムが参加者の答えの中からテーマを的確に特定し、次の質問を生成していることがわかる。また終了時の判断も的確に行っていることが確かめられた。

参加者: 街のイメージがしやすいから
.....
街のイメージがしやすいと何故良いのですか?

参加者: その地域の記憶が残しやすいから
.....
地域の記憶が残しやすいと何故良いのですか?

参加者: 街のイメージがしやすいから
.....
それ以外で地域の記憶が残しやすいと何故良いのですか?

参加者: 特にありません
.....
それでは、街のイメージがしやすい以外で、景色を見ると何故良いのですか?

参加者: 特にありません
.....
それでは、景色を見る以外で、街がおしゃれだと何故良いのですか?

参加者: 特にありません
.....
ありがとうございます。では、他に三宮市の魅力的なところだと思うものはありますか?

参加者: 特にありません
.....
インタビューは以上です。ありがとうございました。<終了>

図 8. インタビュー対話サブシステムによる対話例

次にシステムが出力した評価構造図の一部を図 9 に示す。本図はインタビュー対話で収集された評価項目に対し、クラスタリング・概念抽出を行い、ESV で可視化したものである。

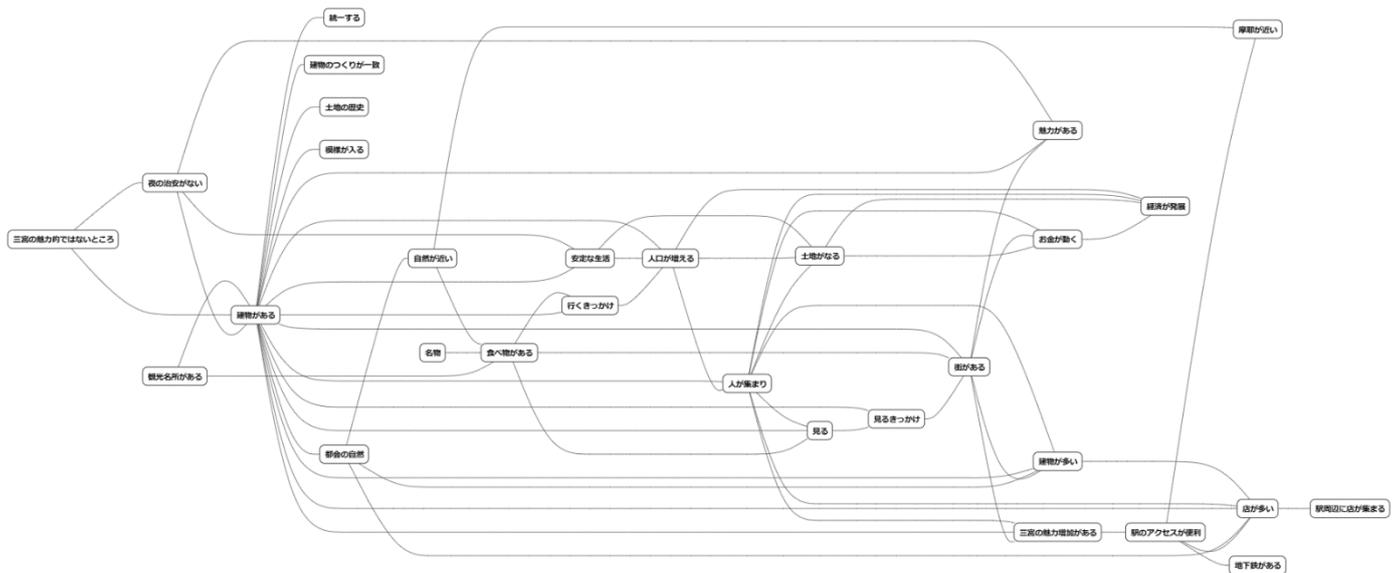


図 9. 対話システムで構築した評価構造図の例

本結果を人手による評価構造図と比較する。三宮の魅力的なところ・三宮の魅力でないところの2つの評価構造図に現れるノードの一致率を計算した。その結果、表4のように(a)三宮の魅力的なところではBERTで75.0%、Word2vecで66.7%となり、十分に高い一致率が得られた。しかし(b)三宮の魅力でないところではBERT、Word2vecともに42.1%であった。これらの結果から、ポジティブなテーマに関しては提案手法の有効性が示されたが、ネガティブなテーマに関しては課題が残されたと言える。

原因の1つに聴取エージェントのプロンプトの記述方法が挙げられる。プロンプトではポジティブなラダリングについてのみ記述しているため、対話においてネガティブな評価項目をポジティブな評価項目に変換する例がしばしば見られた。この部分のプロンプトを改善することで、ネガティブなテーマにも対応できることが期待される。

表 4. ノード一致率

	(a)三宮の魅力的なところ	(b)三宮の魅力でないところ
BERT	75.0%	42.1%
Word2vec	66.7%	42.1%

5. まとめ

本研究では、評価グリッド法におけるインタビューを自動化するため、大規模言語モデルを用いたインタビュー対話システムの構築を行った。システムの出力結果を人手による評価構造図と比較したところ、十分な一致率が得られ、手法の有効性が示された。

文 献

- [1] 讃井純一郎, 乾正雄, “レパトリー グリッド発展手法による住環境評価構造の抽出—認知心理学に基づく住環境評価に関する研究 (1)—,” 日本建築学会計画系論文, No.367, pp.15-22, 1986.
- [2] 土田義郎, 小酒祐貴, “評価グリッド法支援ツールの開発と応用,” 日本建築学会技術報告集, 14.27, pp.205-208, 2008.
- [3] 大谷俊太, 橋本翔, 杉本匡史, 長田典子, “単語の係り受け関係に基づく印象評価構造の自動構築,” 第17回日本感性工学会春季大会, 2D2-04, 2022.
- [4] 坂本有紀, 内田貴久, 石黒浩, “大規模言語モデルを用いたユーザモデル推定機能を持つ対話システムの検討,” 人工知能学会全国大会論文集, vol. 37, p.3O1OS2c01, 2023.
- [5] 山崎天, 坂田亘, 川本稔己, 小林滉河, Nguyen Tung, 上村卓史, 中町礼文, 李聖哲, 佐藤敏紀, “パーソナリティの考慮と知識ベースを統合したHyperCLOVAを用いた雑談対話システム,” 人工知能学会研究会資料 言語・音声理解と対話処理研究会, vol. 93, pp. 113-118, 2021.
- [6] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, M. Sun, Communicative agents for software development, arXiv preprint arXiv:2307.07924, 2023.
- [7] J. Wei, et al, Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems, vol. 35, no.3, pp. 24824-24837, 2022.
- [8] 我妻正太郎, 照屋絵理, 竹内理, “自動作成された類義語抽出ルールによる類義語抽出精度向上手法,” 情報科学技術フォーラム講演論文集(FIT), vol.20, no.2, pp.287-288, 2021.
- [9] M. Sugimoto, Y. Yagi, N. Nagata, How different tourist sites evoke different emotions: investigation focusing on the urban and rural sites in Japan, In: Kurosu, M., Hashizume, A. (eds) Human-Computer Interaction. HCII 2023. Lecture Notes in Computer Science, vol 14012, Springer, Cham, 2023. https://doi.org/10.1007/978-3-031-35599-8_21