

一般発表 | Track 1: 自然言語処理・機械学習基礎

2026年3月2日(月) 13:00 ~ 15:10 | 会場 B

### [8B] 生成AI応用 II

座長: 瀧本 祥章(NTT株式会社) コメントータ: 太田 学(岡山大学) ジュニアコメントータ: 桑田 若菜(兵庫県立大学)

13:00 ~ 13:25

[8B-01] LLM型生成AIによる図形・画像生成と表現度に基づく品質分析

\*田中 克己<sup>1</sup>、井上 沙紀<sup>1</sup>、角谷 和俊<sup>1</sup>、荘司 慶行<sup>2</sup> (1. 関西学院大学、2. 静岡大学)

13:25 ~ 13:50

[8B-02] AI師匠：ユーザの知的探求を支援するAIエージェント

\*小池 美優<sup>1</sup>、森部 想水<sup>2</sup>、牛尼 剛聡<sup>3</sup> (1. 九州大学芸術工学部、2. 九州大学大学院芸術工学府、3. 九州大学大学院芸術工学研究院)

13:50 ~ 14:15

[8B-03] 大規模言語モデルを用いた質問力向上支援システム

\*井上 時音<sup>1</sup>、馬 強<sup>1</sup> (1. 京都工芸繊維大学)

14:15 ~ 14:35

[8B-04] 生成AIを用いた保有技術分析手法

\*後藤 直紀<sup>1</sup>、多田 裕美<sup>1</sup>、柴田 秀哉<sup>1</sup> (1. 三菱電機株式会社)

14:35 ~ 15:00

[8B-05] [技術報告] AIエージェントによる「業務の自動運転化」に向けた技術的挑戦と実践例

\*松村 優也<sup>1</sup> (1. 株式会社LayerX)

# LLM 型生成 AI による図形・画像生成と表現度に基づく品質分析

田中克己<sup>†</sup> 井上沙紀<sup>†</sup> 角谷和俊<sup>†</sup> 荘司慶行<sup>††</sup>

<sup>†</sup> 関西学院大学 〒669-1330 兵庫県三田市学園上ヶ原 1 番

<sup>††</sup> 静岡大学 〒432-8011 静岡県浜松市中央区城北 3 丁目 5-1

E-mail: <sup>†</sup> tanaka\_7339@yahoo.co.jp, {huk50102,sumiya}@kwansei.ac.jp, <sup>††</sup> shojiy@inf.shizuoka.ac.jp

あらまし LLM 型生成 AI を用いて図形・画像の新規生成と生成された図形・画像の品質分析を行う。品質分析は、新規に生成された図形・画像を「表現度」という概念を用いて分析する。「表現度」は、入力テキストのどの部分がどの程度、図形・画像に関与したかを表す指標である。実際に、ChatGPT や Gemini3 を用いて図形・画像を生成し、「表現度」を計算して品質分析実験を行ったのでこれについても報告する。

キーワード LLM, 生成 AI, マルチメディア, 品質分析, 表現度

## 1. はじめに

近年、生成 AI<sup>[1][2]</sup>や画像生成 AI<sup>[3][4]</sup>の開発・普及が著しい。LLM 型生成 AI の ChatGPT<sup>[1]</sup>では、入力テキストから新たなテキスト・図形・画像・映像を生成することが可能になっている。これは、生成 AI が、あらかじめ大量の画像・テキスト対を機械学習しているためである。

所望の図形・画像・映像を自動生成するためには、適切なプロンプトを入力することが肝要である。しかし、現状の生成 AI は、プロンプトの作成に関しての指針やアルゴリズムは提供しておらず、適切なプロンプトを与えられるかどうかは、あくまでユーザーに任されている点が課題である。

本論文では、LLM 型生成 AI によって生成した図形・画像の品質分析を行う。品質分析のために、新規生成されたデータを「表現度」という概念を用いて分析する。「表現度」は、入力テキストのどの部分がどの程度、図形・画像に関与したかを表す指標である。実際に、ChatGPT を用いて生成した図形・画像の「表現度」を計算する実験を行ったのでこれについても報告する。

## 2. テキストからの概念構造抽出に基づく

### 図形の自動生成と品質分析

入力テキストからその概念構造（例：列挙、順序、階層など）を抽出し、その概念構造を表現する構造化図形を自動生成する手法を開発した。

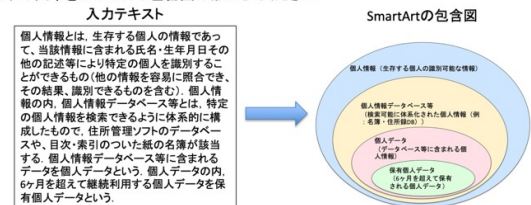
入力テキストからの概念構造抽出は ChatGPT 単体でも可能であるが、抽出した概念構造に対応する図形表現を自動的に選択することが困難であり、これを技術的課題とし、入力テキストから抽出された概念構造を抽出し、適当な SmartArt グラフィックス<sup>1</sup>を自動選定・図形化する GPT を開発した（図 1 参照）。



図 1: SmartArt グラフィックスの図形

図 2 と図 3 は、ChatGPT を用いて SmartArt 図形を自動生成した例である。包含図やバランス図といったグラフィック形式をあらかじめ指定し、入力テキスト全体を ChatGPT-4o で直接図形に変換したもので、入力テキスト全体が包含的または対比的構造を含んでいたため、容易に図形化が行えた例である。

以下の文章を SmartArt の包含図の形にしてください。



入力テキストから SmartArt の包含図を生成

図 2: テキストからの SmartArt 包含図の生成

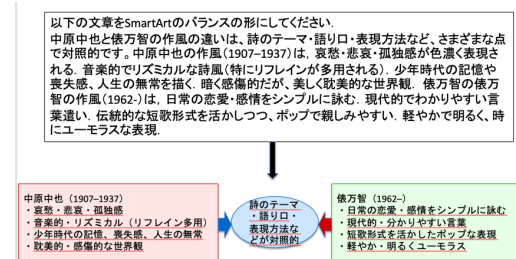


図 3: テキストからの SmartArt バランス図の生成

しかし、一般的には、入力されたテキスト全体を一つの図形に自動変換することは困難である。そこで、

<sup>1</sup> SmartArt グラフィックスは、多様な概念構造に対応する図形表現を集積したもので、リスト・プロセス・階層などがある

本研究では、入力テキストを解析し、図形化可能な部分テキストを抽出し、これに対応する SmartArt レイアウトを生成・提示する GPT (図 4) を構築した。この GPT は、リスト構造、プロセス、循環構造、階層構造、集合関係、マトリックス、ピラミッドの 7 種類の概念構造タイプにもとづき、対応する部分テキストの概念構造を分類する。分類は、文構造や接続詞、表現パターン、列挙マーカーなどの特徴に基づいて行なわれる。表 1 に、各構造タイプに対応する SmartArt グラフィックスを示す。対応関係は本 GPT に知識として組み込まれており、概念構造分類に応じて自動的に提示される。

さらに、本 GPT は分類結果に応じて複数のレイアウト候補を提示し、仮ラベルを付与したサンプル図も生成する。これにより、ユーザーは文章の論理構造を確認しながら、目的に応じて図を選択・編集できる。また、テキスト全体が単一構造に収まらない場合でも、文単位・段落単位で部分構造を抽出し、複数の図候補を提示できるため、複雑な文書でも一部を図解として活用できる (図 4)

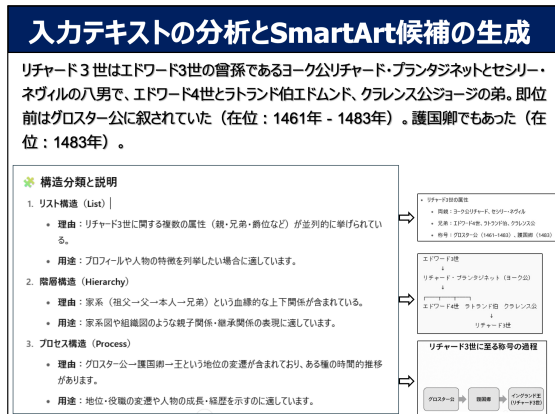


図 4: SmartArt 候補の生成を行う GPT<sup>2</sup>

構造タイプ	概要	SmartArt グラフィックス
リスト構造	並列・列挙的な情報の提示	Vertical Bullet List, Basic List
プロセス(手順)	一方向的な段階的進行	Basic Process, Chevron List
循環構造	継続的・反復的な流れ	Cycle, Continuous Cycle
階層構造	上位・下位の分岐	Hierarchy, Organization Chart
集合関係	包含・共通・交差	Basic Venn, Radial Venn
マトリックス	比較・分類 (2軸)	Grid Matrix, Titled Matrix
ピラミッド	抽象度・重要度の段階的上昇	Basic Pyramid, Segmented Pyramid

表 1: テキストの構造タイプと対応する SmartArt

品質分析は、生成された図形が元のテキストをどの程度表現できているかを示す指標 (「図形によるテキスト表現度」) に基づいて行う。

### 3. 生成 AI によるテキスト-画像の相互変換と変換コンテンツの品質分析

入力テキストからそのテキストの内容を表す画像を自動生成したり、逆に、入力画像からその画像の内容を表すテキストを自動生成する。これにより、テキスト-画像の相互変換が可能となる。

入力テキストからの画像の自動生成や、画像からのテキストの自動生成の実験には、ChatGPT の画像生成モデル DALL-E や Google Gemini3 を用いた。

入力テキストは、惹句 (キャッチコピー)、俳句、短歌、地物名などの短文テキストに加えて、歌詞のような比較的長文のテキストからの画像生成も行った。一方、入力画像からのテキストの自動生成には、風景写真を用いた。

自動生成される画像・テキストの品質分析では、生成に使用した情報と根拠を用いて、生成画像が元のテキストをどの程度表現できているか (表現度) という指標を用いて品質分析を行った。

#### 3.1. テキストからの画像の自動生成と品質分析

図 5 に ChatGPT を用いてテキストから画像を自動生成した例を示す。入力テキストは、図 5 にあるように、俳句テキスト自身に加えて、画像生成を行うために使用した情報と根拠も示すような指示文を追加したものを用いた。

#### テキストからの画像の自動生成

##### 入力テキスト

「梅の実の七つほどつき良き日かな」という俳句から画像を生成してください。画像生成に使用する情報も示してください。

##### 画像生成に使用する情報と根拠

要素 画像に反映する要素  
「梅の実」: 青みを帯びた熟れかけの梅の実、みずみずしい初夏の光と緑  
「七つほど」: 梅の実を7つだけ並べる (過不足なく)  
「つき」: 枝や皿に並んだ自然な状態で梅の実を表現  
「良き日かな」: 穏やかな自然光・涼しげな風景・静けさと幸福感のある空気感

##### 画像生成の構想

背景: 日本家庭の縁側や庭先  
光: 柔らかな初夏の自然光  
主題: 青~黄色の梅の実がちょうど7つ、籠や陶器の皿にのせられている  
周囲: 青葉、竹垣、和風の植栽などが静かに描かれる

##### ChatGPT及びDALL-Eで生成した画像



図 5: テキストからの画像生成

ChatGPT と Google Gemini3 を用いてテキストから画像を生成し、その表現度を計算した例を、図 6、図 7、図 8 に示す。対象としたテキストは、図 6 は惹句 (キャッチコピー)、図 7 は俳句・短歌、図 8 は歌詞・地物名である。

図 6 は、惹句からの画像生成と生成画像の表現度を示しているが、次のような問題点が発見できた。

2 SmartArt ナビゲーター: <https://chatgpt.com/g/686795402e8881919074e6a2e31d897d-smartartnahiketa>

### ● 短かすぎる惹句テキスト

「少し愛して長く愛して」や「ウイスキーが大好きでしょ?」のような短かすぎる惹句テキストから生成された画像は表現度がさほど高くない。

### ● 「対比構造型」の惹句

「へたなカツラをつけるならばげのままでいい」のような対比構造型の惹句からは比較的高い表現度の画像が生成できた。

### ● ChatGPT の優位性

ChatGPT (ChatGPT4o+DALL-E) の方が Gemini3

よりもより表現度の高い画像が生成できた。

図 7 は、俳句・短歌から画像生成しその表現度を計算したものであり次のような問題点があった

### ● 短い俳句テキスト

「雀の子そこのけそこのけお馬が通る」や「寝返りをするぞそこのけきりぎりす」等の俳句からは ChatGPT, Gemini3 とともに、低い表現度の画像しか生成できなかった。

### ● 比喩を用いた短歌

「白銀も金も玉も何せむに優れる宝子にしかめやも」のような比喩を用いた短歌からは「高すぎる」表現度を持つ画像が生成された。原因は、比喩対象のすべてを画像化しているためである。

図 8 では、歌詞・地物名からの画像生成と生成画像の表現度を示している。

#### ● 歌詞全体からの画像生成

歌詞全体から画像生成する場合は、複数の画像を生成する必要があるものと考えられる。

#### ● 地物名からの画像生成

複数の地物名を与えてそのような地物を巡る「観光コース名」と「観光コース画像」を出力させた。コース名やコース画像の評価が今後の課題である。

## 3.2 画像からのテキストの自動生成と品質分析

ChatGPT や Gemini3 には、入力画像からテキストを自動生成する機能がある。この機能を用いて入力画像からテキストを出力し、さらに、このテキストから画像を再生成することで、元の入力画像からの意味を獲得し、獲得した意味を反映した画像を再生成することができる。

図 9 にその例を示す。最初の例は、福知山城周辺の風景画像であり、2 番目の例はバーカウンターの情景画像である。ChatGPT4o による出力テキストを見るとこれらの画像の意味を正確に生成している事が分かる。この出力テキストから画像を再生成した画像を見ると元の画像がより洗練され、画像の意味もより明確に表現していることが分かる。

入力画像 (提供される画像の印象を述べてください)	画像 タイプ	出力テキスト (ChatGPT4o)	出力テキストからの再生成画像 (ChatGPT4o) (この説明から画像を生成してみてください)
	風景画像	この写真は、福知山城周辺の風景を捉えています。空は青く、木々も緑々しく、城の石垣や塔が印象的です。 【生成されたテキスト】 福知山城、青空、遠くには山々が見え、静かな風景が広がっています。城の石垣は古びた感じがして、歴史を感じさせます。空には白い雲が浮かんでいて、穏やかな雰囲気が漂っています。 【生成された画像】 この生成された画像は、元の画像の印象を忠実に再現しています。城の石垣や塔のディテールもよく表現されており、全体的な雰囲気もよく再現されています。	
	バーカウンター	この写真は、バーカウンターの情景を捉えています。様々な種類のアルコール飲料のボトルが並び、グラスも置かれています。照明は暖かみがあり、落ち着いた雰囲気が漂っています。 【生成されたテキスト】 バーカウンター、様々な種類のアルコール飲料のボトルが並び、グラスも置かれています。照明は暖かみがあり、落ち着いた雰囲気が漂っています。 【生成された画像】 この生成された画像は、元の画像の印象を忠実に再現しています。様々な種類のアルコール飲料のボトルやグラスのディテールもよく表現されており、全体的な雰囲気もよく再現されています。	

図 9: ChatGPT による画像からのテキスト生成

## 3.3 画像-テキストの相互変換

一般に、画像を入力とした言語タスクを解く仕組みは VLM (Vision Language Model) と呼ばれる技術によって実現される。VLM は、入力画像を Vision エンコーダによってベクトル化し、言語トークンと同じように画像トークンを扱うことで、画像とテキストをまたいだ相互変換的なタスクを可能にした。

VLM による画像処理は、古典的なキャプションが画像中に描かれている要素やオブジェクトを単語の列挙や文で表すのに対して、より高度で抽象的なタスクにも対応している。すなわち、「この画像には何が写っているか」というような単純なタスクだけではなく、「この画像が見る人に与える印象は何か」や、「この画像が暗示しているものは何か」というような推論を行える。こうしたキャプションと VLM による画像理解では、そこで扱う意味の粒度が異なっていることに注意を要する。

同様のことは、画像生成でも言える。古典的な画像生成 AI は、画像生成用のプロンプトから画像を生成する。このプロンプトは画像に映っているオブジェクトや、画像の持つ画風等の特徴を入力とする。一方で、ChatGPT や Gemini に搭載された画像生成 AI では、はじめにユーザの入力からプロンプトを生成し、そのプロンプトから画像を生成する。つまり、「冷たい印象を受ける画像」のように抽象的なプロンプトであったり、モチーフを使って表現しないといけない画像について、LLM は「冷たいという印象を与える画像には、こういうものが写っており、こういう画風である」という推論から、プロンプトを生成する。こうした仕組みはプロンプト拡張 (Prompt Expansion) [8] と呼ばれ、実際の画像生成はこうして拡張されたプロンプトによって行われる。

これらの直接的なキャプションによる画像生成と、LLM を経由した画像生成では、取扱える意味のレイヤが異なる点に注意を要す。テキスト-画像の相互変換 (4 種類の変換) を、表 2 にまとめる。ここ

で、テキスト同士、画像同士はそれぞれ比較可能であることに注目すると、相互変換によって意味のレイヤを跨いだ生成物の評価が実現できる。

図 10 に実在の短歌を題材に、4 種類の変換をそれぞれ行った結果を示す。元の短歌から直接的に画像生成した場合には「肌」や「血潮」といった語彙に引きずられているのに対し、プロンプト拡張を介した場合は短歌の持つ意味を考慮して男女の画像が生成される。こうして生成された画像に対して、キャプション、解説、短歌への復元をそれぞれ行うと、それぞれ意味のレイヤの異なるテキストが得られる。

このような画像とテキストの意味のレイヤを跨いだ相互変換は、単なる生成物の直接的な利用以外に、生成結果の比較や評価に利用可能である。折り返し生成を行うことで、テキスト同士、画像同士を、既存の手法で比較し、評価できるようになる。

こうした比較は、生成 AI による応用的なタスクにも使える。たとえば、「抽象的なイメージを言語で描写する」といったタスクにおいて、中間的に画像を生成して、それをキャプションと言った処理が考えられる。加えて、「もっとも元の短歌に復元しやすいプロンプト拡張」を行うための学習データを自動で生成するなど、機械学習の学習データそのものを生成 AI に作らせるなどの用途にも応用可能であることが予想される。

#### 4. まとめ

本論文では、LLM 型生成 AI によって生成されるテキスト・図形・画像の品質分析を行った。品質分析には、新規に生成されたデータを「表現度」という概念を用いて分析した。「表現度」は、入力テキストのどの部分がどの程度、図形・画像に関与したかを表す指標である。

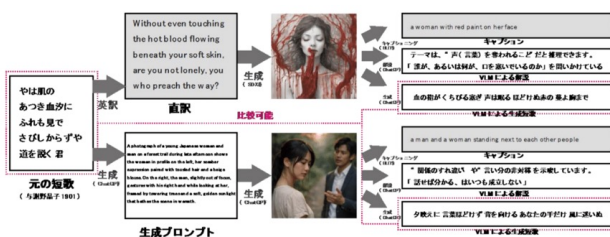


図 10： 4 種類の変換に基づく折り返し生成の例

	直接的な生成	意味変換を介した生成
テキスト → 画像	入力をそのままプロンプトとして画像を生成  StableDiffusion	入力から LLM でプロンプトを生成し、そのプロンプトで画像を生成  ChatGPT, Gemini
画像 → テキスト	画像をキャプション  BLIP 等	画像とテキストプロンプトを VLM に入力する言語タスク  ChatGPT, Gemini での画像理解

表 2： 4 種類の画像・テキストの相互変換

#### 謝辞

本研究は、令和 6 年度科研費基盤研究 (B) (課題番号 22H03905, 課題番号 23K21845), 令和 7 年度科研費基盤研究 (B) (課題番号 25K03229) による。

#### 参考文献


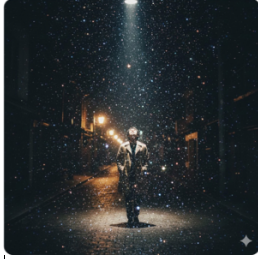
- [1] OpenAI: ChatGPT <https://chatgpt.com>
- [2] Gemini3: <https://chat.aichatapp.ai/jp/app/ai-chat/?model=gemini>
- [3] Midjourney: <https://pollo.ai/ja/im/midjourney>
- [4] Stable Diffusion: <https://deivid.ai/ja/model/stable-diffusion>
- [5] 杉田祥太郎, 野村修, 崔童殿, 田中克己: 生成 AI とテキストの絵画性分析に基づく画像自動生成, 4B-01, DEIM2025, 2025 年 3 月
- [6] 井上沙紀, 角谷和俊, 田中克己: LLM 型生成 AI による図形・画像生成と表現度に基づく品質分析, F-19, DBWS2025 (同志社大学), 2025 年 9 月
- [7] Katsumi Tanaka: Is ChatGPT Enough for Content Generation & Management?, KJDB2025, Ise, Japan, Dec. 2025
- [8] S.Datta et al.: Prompt Expression for Adaptive Text-to-Image Generation, ACL, 3449-3476, 2024

原文	ChatGPT4o+DALL-E	表現度	脚注	Gemini3	表現度	脚注
少し愛して 長く愛して		4			2	長く愛しての部分が表現できていない
女房辭わせて どうするつもり？		5			4	女房の部分が表現できていない
へたなカツラを つけるなら ハゲのままでいい		5	対比構造型		4	

図 6： 惹句からの画像生成と表現度

原文	ChatGPT4o+DALL-E	表現度	脚注	Gemini3	表現度	脚注
雀の子そのけそ このけ御馬が通る 俳句・小林一茶		4			4	
春のめだか 雛の足あと 山椒の実 それのもの 一つかわが子		5	比喻対象をすべて表示		3	足跡が人間のものになっている。
やは肌のあつき血 汐にふれも見でさ びしからずや道を 説く君 短歌・与謝野晶子		5			4	君が「借侶」になってしまっている点が良い

図 7： 俳句・短歌からの画像生成と表現度

原文	ChatGPT4o+DALL-E	表現度	脚注	Gemini3	表現度	脚注
上を向いて歩こう にじんだ星をかぞえて思い出す 夏の日 一人ぼっちの夜 歌詞 上を向いて歩こう		5			4	

入力：地物1, 地物2, ... XXXをめぐる観光コース名を出力してください。また生成された観光コース名の理由を説明してください	出力（観光コース名）	出力画像（英訳から）	表現度
木屋町, 祇園, 先斗町	「京の宵、川と灯りの小径」		5
金閣寺, 銀閣寺, 祇園	「京の三美をめぐる道」		5

図 8：歌詞・地物からの画像生成と表現度

# AI 師匠：ユーザの知的探求を支援する AI エージェント

小池 美優<sup>†</sup> 森部 想水<sup>††</sup> 牛尼 剛聡<sup>†††</sup>

<sup>†</sup>九州大学芸術工学部 〒815-8540 福岡県福岡市南区塩原 4 丁目 9-1

<sup>††</sup>九州大学大学院芸術工学府 〒815-8540 福岡県福岡市南区塩原 4 丁目 9-1

<sup>†††</sup>九州大学大学院芸術工学研究院 〒815-8540 福岡県福岡市南区塩原 4 丁目 9-1

E-mail: †{koike.miyu.962,moribe.sosui.695}@s.kyushu-u.ac.jp, ††ushiana@design.kyushu-u.ac.jp

**あらまし** 現在の推薦行動の多くは、過去の行動から「現在の関心」に基づいた情報を推薦するものであり、ユーザが興味のある未知なる範囲のことを学ぶには適していない。そこで、本研究では、ユーザの熱意や現在の知識、今後の学びに関する意向をもとに、知的探求を支援し、成長を促す AI エージェント「AI 師匠」の開発に取り組む。このエージェントでは、推薦までの思考段階を「プロフィール予測」「興味方向性予測・学習計画設計」「推薦・探求的支援」の 3 つに分ける。これにより、ユーザの漠然とした学習意欲を明確にし、よりユーザの知的探求心を伸ばせる情報を推薦し、ユーザの学習を継続的に支援することを目指している。

**キーワード** LLM、推薦システム

## 1 はじめに

### 1.1 研究背景・目的

近年の推薦技術は、協調フィルタリングや行列分解に代表される手法に基づき発展し、ユーザの過去の行動履歴や現在の嗜好から関心の高いものに基づいた情報を提示する技術として広く利用されている [1], [2]。この手法は、すでに明確な「現在の関心」を有する分野においては有効である一方で、ユーザ自身がまだ十分に理解・言語化できていない新しいことに対する学習意欲や、将来的に形成される可能性のある潜在的な関心、「未来の関心」や「成長する興味」に十分に対応できていない。

特に学習の初期段階においては、「この分野を学んでみたいがどこから手を付ければよいのか分からない」「興味はあるが、難しそうで踏み出しづらい」「自分ではわからないので誰かに聞いてみたいが、誰になんて聞けばいいかわからない」といった心理的・認知的障壁が存在する。このような探索的学習行動の重要性は exploratory search の研究において指摘されており [3], [4]、このような状態では、明示的な検索クエリを入力することが難しく、既存の検索・推薦システムでは十分に対応できないという課題が生じる。

近年では、ユーザとの対話を通じて嗜好や要求を逐次的に推定する会話型推薦システムの研究が進展している [5], [6]。さらに、大規模言語モデル (LLM) を基盤とした推論と行動を統合するエージェント型枠組み [7] や、人間社会を模擬した長期的振る舞いを示す生成エージェント [8] が提案されている。しかし、これらの多くは短期的な意図推定や単発的タスク遂行を主眼としており、学習過程における関心の変化や成長を長期的に捉え、知的探求活動に継続的に伴走する仕組みは未だ十分に確立されていない。

本研究では、このようなユーザの漠然とした熱意や関心、未言語化の学習意欲に寄り添い、ユーザと対話を繰り返しながら

学習の入り口や道筋を提示し、ユーザの知的探求活動を継続的に支援する「師匠」的存在となる AI エージェントの実現を目指す。ユーザの現在の知識や熱意だけでなく、将来の学びに対する迷いや期待を汲み取り、ユーザの興味の成長を予測し、長期的な知的探求を伴走的に支援する点に本研究の特徴がある。

### 1.2 本研究の貢献

本研究の具体的な貢献は以下の 3 点である。

1. ユーザの現在の知識や熱意だけでなく、将来の学びに対する迷いや期待を汲み取り、ユーザの興味の成長を予測し、段階的な思考の枠組みを提案した。
2. 長期的な知的探求を伴走的に支援するためのシステムを考案し実装した。
3. 実際に出力された推薦文について、観点別に評価を行い、有効性を検証した。

### 1.3 本論文の構成

本論文の構成は以下の通りである。第 2 章では本研究の関連研究について触れる。第 3 章では提案手法となる「AI 師匠」の仕組みについて説明を行う。第 4 章では「AI 師匠」の実装に関するプロトタイプシステムについて述べる。第 5 章では評価手法としての実験概要や、実験の評価項目について述べる。第 6 章では第 5 章で行った実験の結果、第 7 章では実験結果の考察を述べる。第 8 章では本研究を経ての今後の展望を述べる。

## 2 関連研究

### 2.1 プロファイル予測に関する研究

近年、大規模言語モデルを用いてユーザ属性や嗜好を推定・拡張する研究が進展している。永沼ら [9] は、LLM を用いたユーザプロフィール拡張により推薦精度が向上することを示し、特にコールドスタートユーザに対する初期理解の有効性を報告している。また、Zhang ら [10] は LLM に基づくユーザシミュ



ユーザのニーズに応じて、各学習段階の目的 (何ができるようになるか (ユーザ視点))、学習内容 (何をどう体験・理解するか (師匠視点)) が含まれる。このプロセスは、目標達成への動機付けに寄与する。

学習計画の提示後、学習計画に関するユーザからのフィードバック (感想や要望など) を受け取る。このフィードバックによって学習計画をよりユーザに合ったものに更新していくことが可能となる。

### 3.2.3 探求支援と適応的推薦

このフェーズでは、前のフェーズで設定された学習計画に対して、適切な情報を推薦する。推薦フェーズでは、今までユーザと対話しながら作ったユーザプロフィールと学習計画のデータを推薦 AI に与えて推薦するものを考える。推薦の際には、楽曲のみならず推薦理由や背景、楽曲の動画リンクも一緒に提示することでユーザの理解を促す。また、ユーザの関心の深化や拡張を促すような情報提示を重視する。ユーザは、推薦されたものについて、感想・質問・次の推薦への要望といったフィードバックを送ることができ、感想や要望に関連した次の推薦や質問の回答を得ることができる。また、発話内容を分析し、関心語を抽出して次の推薦に反映することでよりユーザに合った学習を行うことができる。また、この対話を通じて得られるフィードバックを反映し、必要に応じてユーザプロフィール推定や学習計画設計に戻る柔軟な対応を行い、ユーザの状態に合った学習ができるように設計する。学習計画で定めた段階について満足したら、その旨を入力すると次の学習段階に学習が進む仕組みになっている。

## 4 実 装

本研究では、第 3 章で述べた支援フレームワークを具現化するため、LangChain を用いたプロトタイプシステムを実装した。本章では、システム全体構成と各処理モジュールの役割について述べる。

### 4.1 全体構造

まず、ユーザは学びたいテーマを 1 つ入力する。その後、対話制御モジュールがあらかじめ設計した 6 問の質問を順に提示し、ユーザの回答を収集する。6 問の初期質問は以下の通りである。

- このテーマに関連したもので、これまでに以下のような体験・経験をしたことはありますか？
  - 授業で聴いた
  - 自分で調べた
  - 演奏した
  - 楽譜を読める
  - コンサートに行った
  - ほぼ初めて
- このテーマについて、特に印象に残っている経験・体験を教えてください。
- このテーマ、どんな要素が特に気になっていますか？

- 歌・声
- 楽器演奏
- 作曲
- 歴史・背景
- 理論
- 聴き比べ
- 有名曲を知りたい

- もし説明するとしたら、1. まず全体のイメージを知りたい 2. 具体的な例から入りたい どちらがしっくりきそうですか？
- 次の説明で好みは？
  - 短く要点
  - 物語つき
  - 図解
  - 実例中心
  - 用語定義から
- 逆に、こういう説明の仕方はちょっと苦手かも…というのがありますか？ あれば内容を教えてください。

収集された対話ログはプロフィール推定エージェントへ送られ、興味対象、既有知識、学習動機などの観点からユーザプロフィールを推定する。推定結果は構造化データとして保持すると同時に、ユーザが理解しやすい文章形式に変換して提示される。ユーザから得られた修正意見や感想は、後続の学習計画生成および推薦フェーズで再利用するために保存される。

次に、学習計画生成エージェントは、推定されたユーザプロフィールおよび初期対話で得られた応答内容を入力として、3 段階構成の学習計画を生成する。生成された学習計画についてもユーザからのフィードバックを受け取り、大きな拒否反応が見られた場合には、状態に応じて計画を再生成する仕組みを備えている。

最後に推薦エージェントは、確定したユーザプロフィールおよび学習計画を入力として、教材、作品、学習行動案などを推薦する。本研究では推薦結果をテキストとして提示し、対話的に次の探究行動へ接続する設計とした。

### 4.2 ユーザインタラクション管理と設計上の工夫

本プロトタイプでは、複数フェーズにまたがる対話を一貫して制御するため、対話管理モジュールを中心に各エージェントを連携させている。このモジュールは、ユーザ入力の受付、現在のフェーズ判定、必要なエージェントの呼び出し、内部状態の更新を担当する。

対話管理については、「初期入力」、「質問フェーズ」、「プロフィール提示」、「学習計画生成」、「推薦」、「フィードバック反映」といった状態を順に遷移する構造とした。

各フェーズ内で、LangChain のチェーン構成を用いて適切なプロンプトと LLM を組み合わせた処理を実行する。たとえば質問フェーズでは、事前に定義した質問テンプレートを順番に提示し、ユーザ応答を対話履歴として蓄積する。

ユーザ理解から推薦までを複数のエージェントに分離し、LangChain によって制御するという構成により、各処理段階

を独立して改良可能とし、将来的な拡張や比較実験を容易にしている。

ユーザ応答は json ファイルとして保存・蓄積され、プロフィール推定や学習計画設計の入力として再利用される。また、json 形式にすることで評価実験において状態遷移の追跡や分析を行うことを可能とする。また、フェーズごとに保持すべき情報を整理し、ユーザプロフィール、学習計画、推薦履歴、関心語といった要素をセッション単位で管理する設計とした。

推薦フェーズ以降では、ユーザの自由入力に応じて同一フェーズ内で複数回の推薦や質問応答を繰り返すことができる。さらに、ユーザが学習段階の完了を宣言した場合には、状態を更新して次の学習フェーズへ進む。

### 4.3 プロファイル推定エージェント

本節では、ユーザ理解フェーズを担うプロフィール推定エージェントの実装について述べる。

このエージェントは、初期入力および質問フェーズで得られたユーザの発話ログを入力として受け取り、LLM により構造化されたユーザプロフィールを生成する。LangChain 上では、質問応答履歴をまとめたテキストを入力とし、JSON 形式で出力するチェーンとして構成した。

出力されるユーザプロフィールは、第3章で示した項目に対応するスキーマを持ち、それらの情報を含む。これにより、後続の学習計画生成および推薦処理において、一貫した情報参照が可能となる。

以下に実際に推定したユーザプロフィールを示す。

#### 予測したユーザプロフィール (例: バロック音楽の特徴を聴いて理解したい初心者)

```
"profile_text": "このユーザーは、バロック音楽に対して具体的な例を通じてその雰囲気や特徴を理解しようとする興味を持っているようです。興味の広がりとしては、音楽の背後にある構造や作曲のルールに対する関心が見られます。動機は、カフェでの印象的な経験をきっかけに、さらに深く知りたいという好奇心から来ているようです。現在の知識段階は入門者レベルで、有名な曲を少し聴いたことがある程度ですが、具体的な例から学び進めることを好むことから、適応型 (Accommodating) の学習スタイルが適しているかもしれません。一方、専門用語が多い説明は避けたほうがよさそうです。感情的な関与は、バッハの曲に対する個人的な印象からも強いものが感じられます。好奇心は、音楽の構造や作曲のルールに向いており、次取るべき関わり方としては、短く要点を押さえた説明やたとえ話をを用いた理解しやすいアプローチが有効でしょう。"
```

生成されたプロフィールは、そのままでは専門的で冗長になる可能性があるため、別の LLM チェーンを用いてユーザ提示用の自然言語説明文へ変換する。この説明文とともに、ユーザに対して「予測された内容は適切か」「修正したい点はあるか」

といった確認を行い、フィードバック入力を促す。

ユーザから得られた修正意見や補足情報は、再度 LLM に与えることでプロフィールを更新する処理を行う。この更新後プロフィールを正式な状態情報として保存し、以降の学習計画設計および推薦フェーズで参照する。

これらの処理により、本プロトタイプでは、一度きりの推定に留まらず、対話を通じてユーザ理解を段階的に深める設計とした。

### 4.4 学習計画生成エージェント

本節では、ユーザプロフィールを入力として受け取り、段階的な学習計画を生成する学習計画生成エージェントの実装について述べる。

本エージェントは、前節で確定したユーザプロフィールを入力とし、LLM により初級・中級・発展の三段階から構成される学習計画を生成する。LangChain 上では、ユーザプロフィール JSON を変数として埋め込んだプロンプトテンプレートを用い、計画全体を一括で出力させる構成とした。

生成される学習計画には、各段階ごとにユーザ視点の到達目標、師匠視点での学習内容、想定される活動例を含める。これにより、ユーザが自身の成長過程を具体的に想像できるよう設計した。

以下に実際に生成した学習計画を示す。

#### 生成した学習計画 (例: バロック音楽の特徴を聴いて理解したい初心者)

```
"vector": "バロック音楽の雰囲気や特徴、作曲のルールに対する理解",
"learning_theme": "バロック音楽の基礎理解と構造の探求",
"phases": [
  {
    "id": 1,
    "goal": "バロック音楽の基本的な特徴や代表的な作曲家を理解できるようになる",
    "content": "YouTube で『バロック音楽入門』と題した初心者向け動画を視聴し、バッハやヴィヴァルディなどの有名作曲家の代表作を聴いて雰囲気を感じ取る。具体的な例を通じて、バロック音楽の特徴を短く要点を押さえた説明で学ぶ。"
  },
  {
    "id": 2,
    "goal": "バロック音楽の背後にある基本的な構造や作曲のルールを理解できるようになる",
    "content": "ウェブサイトでバロック時代の作曲技法や音楽理論を紹介する記事を読む。避けたほうが良い専門用語を使わず、たとえ話をを用いた理解しやすいアプローチで学ぶ。"
```

```

},
{
  "id": 3,
  "goal": "バロック音楽の曲を聴いてその構造を分析し、印象を深めることができるようになる",
  "content": "オンラインでバロック音楽の構造分析を行うワークショップや講座を受講し、自分で簡単な曲の構造を分析する。感情的な関与を引き出し、バッハの曲に対する理解を深めるために、具体的な曲の事例を用いて学ぶ。"
}
]

```

学習計画はユーザに提示され、妥当性、興味との一致度、難易度感についてのフィードバックを受け取る。ユーザから大きな拒否や修正要望が示された場合には、その内容を追加入力として再度 LLM に与え、学習計画を更新する処理を行う。

更新された学習計画は、状態管理モジュールにより保存され、推薦フェーズにおける制約条件として利用される。この仕組みにより、本プロトタイプでは、固定的なカリキュラムではなく、対話的に調整される動的学習計画を実現している。

#### 4.5 推薦生成エージェント

本節では、学習計画およびユーザプロフィールに基づいて推薦を行う推薦生成エージェントの実装について述べる。

本エージェントは、確定済みのユーザプロフィールと現在進行中の学習段階、および直前の推薦履歴や嗜好データを入力として受け取り、LLM により推薦対象とその提示文を生成する。LangChain 上では、これらの情報をテンプレート化したプロンプトに埋め込み、推薦結果を自然言語で出力する構成とした。

楽曲推薦においては、単なるタイトル提示ではなく、推薦理由、楽曲背景、注目すべき聴取ポイント、関連する次の探求テーマを含めるよう設計した。さらに、外部検索 API を併用して楽曲の公開情報や YouTube の動画リンクを取得し、推薦文中に付与する仕組みとした。

以下に推薦文の出力例を示す。

##### 実際の推薦文 (例: バロック音楽の特徴を聴いて理解したい初心者)

```

"title": "Brandenburg Concerto No. 3 in G Major, BWV 1048",
"artist": "Johann Sebastian Bach",
"reason": "バッハの代表作の一つで、バロック音楽の特徴である複雑な構造とリズムカルな要素を体験できる。この曲を通じて、バロック音楽の基本的な特徴を具体的に感じ取ることができ、入門者向けの学習目標を満たす。"

```

推薦後には、ユーザからの発話を「感想」「質問」「要望」「拒否」といったカテゴリに分類する。この分類結果に基づき、質問の場合は知識提供チェーンを呼び出し、要望や拒否の場合に

は推薦方針の調整を行う。

また、ユーザの発話から関心語を抽出し、プロフィールの補助情報として保存する。これにより、次回以降の推薦生成時には、これらの語を考慮した候補生成が可能となる。

必要に応じて、ユーザ理解フェーズや学習計画生成フェーズに戻す制御も実装しており、システム全体として循環的にユーザモデルを更新する構造を採用した。

## 5 評価手法

評価手法として、提案手法に基づいたプロトタイプシステムを実装し、LLM で生成した仮想ユーザを対象とした実験を行った。また、評価手法として LLM as a Judge を採用した。

### 5.1 実験

仮想ユーザは GPT-4o を用いて生成した。仮想ユーザのデータには「学びたいこと」、及びユーザプロフィール予測用の初期質問の回答内容を含めている。このデータを用いてユーザプロフィール予測、学習計画設計、推薦を行った。なお今回は、プロフィールや学習計画についてのフィードバック・修正の段階は省き、全て最初に出力されたものを使用することとする。

実験では初回推薦文に関して出力を評価者 AI に読み込ませて、LLM as a Judge によって客観的に評価を行った。「学びたいこと」やユーザ属性がそれぞれ異なる仮想ペルソナを 10 人分設定し、試行を行った。なお、提案手法の途中に出てくる学習計画についての確認の対話と学習計画の修正は省くものとし、最初に出力された学習計画に基づいて推薦されたものを評価対象とする。提案手法、比較手法、評価者 AI 全てにおいて、使用大規模言語モデルは GPT-4o である。

### 5.2 比較手法

入力は提案手法のプロトタイプシステムに入力した仮想ペルソナの情報と同じとする。ユーザの「学びたいこと」のみを入力し、通常 LLM で推薦する場合を比較手法とする。比較手法の構造を図 2 に示す。

### 5.3 評価項目

評価項目は以下の 6 つである。

1. 正確性：曲名、アーティスト、ジャンル特性、背景説明などに明確な事実誤認がないか。
2. 要望遵守性：ユーザの要望（どのような曲を聴きたいか）に沿った推薦になっているか。
3. 論理性：推薦理由が、曲の特徴とユーザの要望を筋道立てて結びつけているか。
4. 情報充足性：曲を選ぶ上で重要な観点（雰囲気、特徴、聴きどころ等）が過不足なく含まれているか。
5. 新しい視点：ユーザにとって新しい気づきや視点を提供しているか。
6. 関心深化：この推薦文によって、さらに音楽を聴きたい・知りたいという関心が高まるか。

## 5.4 評価方法

### 5.4.1 指標を用いたペアワイズ比較

評価者 AI に対し、上記の 6 つの評価項目を評価指標として与え、提案手法および比較手法の出力のうち、どちらが仮想ユーザにより適しているかを項目ごとに二者択一で選択させた。この時、必ず A と B のいずれか一方を選択させ、同一項目における引き分けは生じないものとした。10 名の仮想ユーザに対して各 5 回、全 50 試行を実施した。試行では 6 観点に基づくペアワイズ比較を行い、各観点で選択された手法に 1 点を付与した。仮想ユーザごとに、全 30 比較（5 試行×6 観点）における勝利数の合計および勝率を算出し、これを主たる評価指標とした。

提案手法の勝利数  $W_i$  は式 (1) で定義する。

$$W_i = \sum_{t=1}^5 \sum_{k=1}^6 x_{i,t,k} \quad (1)$$

ここで  $x_{i,t,k}$  は試行  $t$ 、観点  $k$  において提案手法が選択された場合に 1、それ以外の場合に 0 を取る。

勝率  $Score_i$  は式 (2) により算出する。

$$Score_i = \frac{1}{30} \sum_{t=1}^5 \sum_{k=1}^6 x_{i,t,k} \quad (2)$$

また、各観点  $k$  における勝率は式 (3) に基づいて求めた。

$$Score_k = \frac{1}{50} \sum_{i=1}^{10} \sum_{t=1}^5 x_{i,t,k} \quad (3)$$

### 5.4.2 Self-consistency による評価の安定化

各条件について 1 回の生成結果のみを用いるのではなく、確率的生成によるばらつきを抑えるため、同一の仮想ユーザ設定に対して 5 回の試行を行い、得られた評価結果を平均化した。

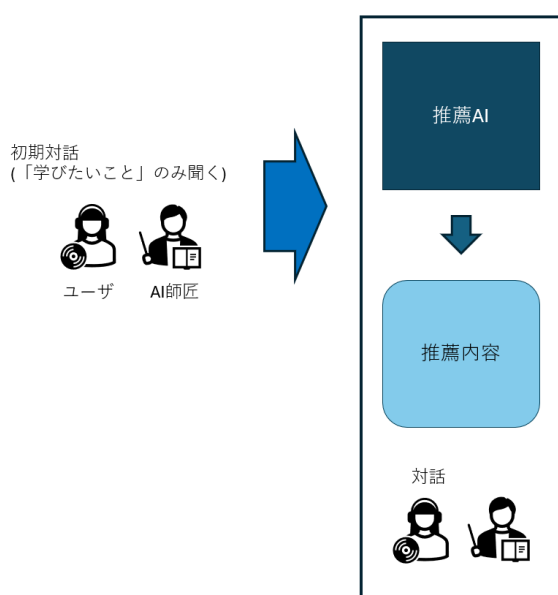


図2 比較手法

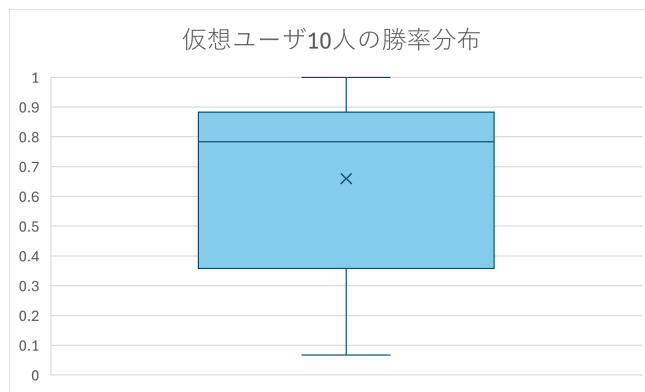


図3 仮想ユーザ 10 人分の勝率分布

表1 仮想ユーザ単位で集約した勝率の平均と Wilcoxon 検定結果

手法	平均勝率	SD	Z	p
通常 LLM	0.34	0.32	-	-
AI 師匠	<b>0.66</b>	0.32	1.48	0.14

これは Self-consistency による安定化を目的としたものである。乱数シードは固定せず、各試行で独立に生成した。

### 5.4.3 有意差検定

本研究では、式 (2) で定義した仮想ユーザ単位の勝率  $Score_i$  を用いて、提案手法と比較手法の性能差を検証した。帰無仮説  $H_0$  は「平均勝率が 0.5 である」と設定した。

統計検定には Wilcoxon 符号付順位検定を用いた。サンプルサイズは仮想ユーザ数に対応する  $N = 10$  である。有意水準は  $\alpha = 0.05$  とした。

## 6 結果

### 6.1 全体性能

まず、仮想ユーザ 10 人の  $Score_i$  分布について図 3 に示す。

なお、勝率の平均値は 0.66、第一四分位数は 0.36、中央値（第二四分位数）は 0.78、第三四分位数は 0.88、最小値は 0.07、最大値は 1 であった。標準偏差は 0.32 である。

また、仮想ユーザ単位で集約した勝率の平均と Wilcoxon 検定結果について表 1 に示す。

今回の実験では提案手法の方が勝率が高かったものの、有意差は認められなかった。

### 6.2 観点別勝率

6 観点別の勝率を、図 4 に示す。

正確性、要望遵守性、論理性、関心深化の観点においては勝率が 0.5 を超えたが、情報充足性、新しい視点の観点では 0.5 を下回った。

## 7 考察

結果としては、提案手法での推薦はユーザごとに結果にばらつきが多くなり、勝率は 0.5 を超えたものの有意差は確認できなかった。また、正確性、要望遵守性、論理性、関心深化の点

## 文 献

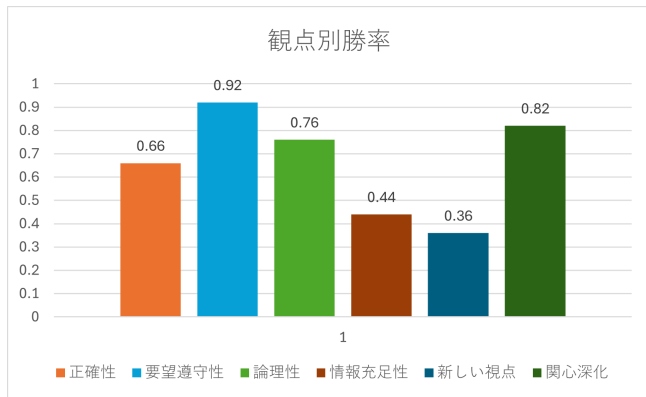


図 4 観点別勝率

では良い評価を得られたが、情報充足性、新しい視点の観点で比較手法の通常 LLM を下回ってしまった。

この結果に関して今回は LLM を用いて生成した仮想ユーザであるため、生成されたユーザデータの質によって結果が大きく左右されたのではと考える。また、今回は推薦結果が推定ユーザプロフィールや学習計画により適しているかを評価したが、ユーザプロフィールや学習計画そのものの評価、及び推薦を繰り返したときの継続的学習という観点での評価ができていない。これに関して、仮想ユーザのみでの実験では限界があるため、今後は実際の人間のユーザを対象とした実験を行ってみたいと考える。

また、評価の際に、情報充足性は曲を選ぶ上で重要な観点（雰囲気、特徴、聴きどころ等）が過不足なく含まれているか、新しい視点はユーザにとって新しい気づきや視点を提供しているかと定義した。そのため、比較手法の通常 LLM の方が、プロフィールや学習計画といった情報を選別する要素がほぼない分、多めの情報を入れる傾向があり評価されやすかったのではと推測する。

## 8 おわりに

本研究では、ユーザの漠然とした熱意や関心、未言語化の学習意欲に寄り添い、知的探求活動を支援する AI チャットボットの構築を目的として、ユーザ分析・学習計画生成・推薦といった役割を担う複数の LLM による段階的な推薦構造を設計・検討した。

その結果、提案したエージェントは、総合評価の平均点や一部の評価観点において高い評価を得たものの、比較手法との間に統計的に有意な差は確認されなかった。

一方で、本研究では生成されたユーザプロフィールおよび学習計画そのものに対する詳細な評価は十分に行えておらず、これらは本研究の限界である。今後は、被験者実験の実施や評価指標の精緻化を通じて、プロフィール推定精度や学習計画の妥当性を検証し、本手法の有効性をより包括的に明らかにしていく予定である。

- [1] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 175–186, 1994.
- [2] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, Vol. 42, No. 8, pp. 30–37, 2009.
- [3] Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, Vol. 49, No. 4, pp. 41–46, 2006.
- [4] Ryan W. White and Resa A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool, 2009.
- [5] Yongfeng Zhang and Xu Chen. Towards conversational recommender systems. In *Proceedings of the SIGIR Conference on Research and Development in Information Retrieval*, 2018.
- [6] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. Conversational recommender systems. *ACM Computing Surveys*, Vol. 54, No. 5, pp. 1–36, 2021.
- [7] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations*, 2023.
- [8] Joon Sung Park, Joseph O’Brien, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2023.
- [9] 永沼翔翼, 石垣龍馬, 酒造正樹, 前田英作. Llm によるユーザプロフィール拡張を用いた推薦システムの検討. 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 100 回 (2024/02), pp. 107–110. 一般社団法人人工知能学会, 2024.
- [10] Zijian Zhang, Shuchang Liu, Zirui Liu, Rui Zhong, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Qidong Liu, and Peng Jiang. Llm-powered user simulator for recommender system. *arXiv preprint arXiv:2412.16984v1*, 2024.
- [11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903v6*, 2023.
- [12] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601v2*, 2023.
- [13] 櫻井崇貴, 白松俊. Llm を活用した目標達成支援のための段階的聴取型情報推薦手法. 人工知能学会全国大会論文集 第 39 回 (2025), pp. 3J6GS505–3J6GS505. 一般社団法人人工知能学会, 2025.
- [14] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158v2*, 2023.
- [15] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, et al. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231v1*, 2024.
- [16] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for “mind” exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023.
- [17] Keith J. Topping. Trends in peer learning. *Educational Psychology*, Vol. 25, No. 6, pp. 631–645, 2005.
- [18] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Talk-

play: Multimodal music recommendation with large language models. *arXiv preprint arXiv:2502.13713v3*, 2025.

# 大規模言語モデル間の討論を用いた 質問力向上支援システムの構築と教育的効果の検証

井上 時音<sup>†</sup> 馬 強<sup>††</sup>

京都工芸繊維大学 〒606-0951 京都府京都市左京区松ヶ崎橋上町

E-mail: <sup>†</sup>b2122501@edu.kit.ac.jp, <sup>††</sup>qiang@kit.ac.jp

2026年2月11日

**あらまし** 近年、大規模言語モデル（LLM）の技術的發展を背景に、教育分野における AI 活用は新たな局面を迎えている。しかし、既存の AI チューターシステムの多くは、学習支援や回答生成に主眼を置いており、学習者が自ら問いを立てる「質問力」の育成支援は十分ではない。本研究では、質問力向上に有効とされている問いづくり技法（QFT）のアイデア発想を支援する SCAMPER 法を活用したマルチ LLM エージェントによる討論評価機構を備えた質問力向上支援システムを提案する。本システムは、学習者に対して SCAMPER のそれぞれの視点に基づき、物事に対する思考や視点を増幅させるとともに、作成された質問に対して複数のエージェントが議論を行い、質の高いフィードバックを提供して、学習者の質問力を向上させる。また、意図的に低品質な質問を用いたテストを実施し、評価の妥当性およびスコアインフレに対する頑健性を保証する。被験者実験の結果、システム利用群は教示のみのグループと比較して、質問の質を示す指標（Graesser 指標および Bloom 階層）において顕著な向上を示した。

**キーワード** LLM, 教育工学, 質問力, SCAMPER, マルチエージェントシステム, QFT

## 1 はじめに

大規模言語モデル（Large Language Models: LLM）の技術的發展を背景に、教育分野における AI 活用は新たな局面を迎えている。Rahman ら [1] が包括的に論じているように、生成 AI は教育・研究分野において、個別の学習ニーズに適応した個人適応型学習や業務効率化の大きな機会を提供する一方で、学術的誠実性への懸念といった新たな課題も提示している。これからの知識基盤社会において真に求められるのは、AI から与えられた知識を単に記憶する能力ではなく、AI を活用しながら自らの思考を論理的に構築し、未知の事象に対して問いを立てる「質問力」である。

国際的な学力調査である PISA2022 の結果 [2] によれば、日本の生徒は数学的リテラシーにおいて世界トップレベルの成績を維持している一方で、「探究型の学習活動」の頻度や、「実生活の課題解決に自信を持つ」生徒の割合は OECD 平均を大きく下回っている。これは、知識の習得（インプット）には長けているものの、そこから自ら問いを立て、探究

するプロセス（アウトプット）が不足していることを示唆している。

問いを立てるプロセスは、単なる情報収集ではなく、既存の知識を組み替え、新たな視点を生み出す創造的な知的活動である。Widihastuti ら [5] は、SCAMPER 法を用いた学習活動が、学生の創造的思考スキル（特に流暢性と柔軟性）を有意に向上させることを実証している。また、Chin と Osborne [3] は、学習者が自ら問いを発することは、既存の知識と新しい情報の乖離を認識させ、深い概念的理解を促すための重要なステップであると論じている。彼らの研究によれば、生徒が発する「問い」の質は、その生徒の学習深度と相関しており、質問生成を促す教育的介入は極めて重要である。

しかし、こうした高度なスキルの育成には、適切な指導とフィードバックが不可欠である。適切な「問いかけ」や「発想の転換」が示されなければ、学習者の思考は深まらず、表面的な理解に留まる危険性がある。従来の多人数講義において、教員が学生一人ひとりの質問生成プロセスに対し、こ

のような質の高い相互作用を提供することは、人的リソースの観点から極めて困難であった。

本研究ではこれらの課題に対し、LLM エージェントが教師の役割を一部代替・拡張することで解決を図る。具体的には、SCAMPER の 7 つのアイデア創造の思考プロセスを活用し、学習者に提示することで創造的な質問生成を促し、さらにマルチエージェント討論 [10] を用いた「視点の正確性」「授業との関連性」「オリジナリティ」「素材選択の適切さ」の 4 基準に基づいて多角的なフィードバックを即座に提供するシステムを構築する。これにより、対面指導の効果を維持しつつ、人的コストを大幅に削減したスケーラブルな支援環境の実現を目指す。

本研究の主な貢献は以下の通りである。

- SCAMPER のアイデア創造の思考プロセスを活用した質問力の向上支援 (3 節)
- スコアインフレを抑制した頑健な評価機構の提案 (3 節)
- ユーザ実験による質問力向上の検証 (4 節)

本稿の構成は以下の通りである。次節では、質問力育成および、LLM の教育利用、質問力向上に繋がる QFT に関する関連研究について述べ、第 3 節にて提案手法である SCAMPER 法を用いた質問生成支援および、マルチエージェント討論による評価の頑健性について説明する。第 4 節では、プロトタイプを用いた被験者実験の結果とその考察を行い、システムの教育的効果を検証する。第 5 節に本研究をまとめる。

## 2 関連研究

### 2.1 質問力育成の重要性と指導上の課題

学習者自身が問いを生成する活動は、知識の定着だけでなく、批判的思考 (Critical Thinking) やメタ認知能力の向上に寄与することが知られている。Chin と Osborne [3] の包括的なレビューによれば、生徒による質問生成は、受動的な学習から能動的な学習への転換点となる。しかし、彼らは同時に、多くの教室環境において生徒からの自発的な質問が著しく少ないことも指摘している。その主な要因として、心理的な障壁や、どのように問えばよいか分からないというスキル不足が挙げられる。したがって、質の高い質問を作成させるためには、教員による適切な適切な質問力を向上させるためのフィードバックや、質問生成に必要な能

力や知識の支援が不可欠である。しかし、個々の生徒の発想に対してきめ細やかなフィードバックを行うには多大な時間を要するため、多人数講義での継続的な実施には人的リソースの限界がある。

### 2.2 LLM を活用した教育支援とその限界

近年、ChatGPT に代表される LLM の教育利用が進んでいる。Baidoo-Anu ら [4] は、生成 AI がカリキュラムの作成や個別指導において多大な恩恵をもたらす可能性がある一方で、その出力の正確性や倫理的課題についても慎重であるべきだと論じている。特に、Lee ら [7] の最新の調査によれば、生成 AI への過度な依存は、ユーザーの批判的思考 (Critical Thinking) への関与を減少させ、自信過剰を引き起こすリスクがあることが指摘されている。AI が単に答えを与えるだけの「回答生成型」の支援では、学習者は受動的になりがちであり、自ら問いを立てる力はむしろ低下する恐れがある。

これに対し、学生の思考を促すために対話を行う相手として LLM を活用する試みもある。しかし、単一の LLM エージェントを用いた場合、AI が学習者の意見に過度に同調してしまったり (Sycophancy)、誤った前提 (ハルシネーション) を修正できずに議論が進んでしまったりする課題がある。教育的なフィードバックにおいては、単一の正解が存在しない場合も多く、公平かつ多角的な視点からの評価が求められる。

### 2.3 マルチエージェントシステムによる討論と評価

単一モデルの限界を克服するアプローチとして、複数の LLM エージェントを協調させるマルチエージェントシステムが注目されている。Xiong ら [10] は、異なる役割を持たせた複数のエージェント同士に議論 (Debate) させることで、単一モデルよりも事実整合性が高く、論理的な結論が導かれることを示した。さらに、Du ら [8] は、複数の LLM エージェントによる討論が、モデルの推論能力と事実性を向上させることを実証している。彼らの研究によれば、単一のエージェントでは見過ごしてしまう誤りを、他のエージェントが指摘することで修正が可能となる。

また、Liang ら [9] は、マルチエージェントによる討論が「拡散的思考 (Divergent Thinking)」を促進することを報告している。これは、本研究が目指す SCAMPER 法を用いた創造的な質問生成と強く合致する知見である。評価の信頼性に関しては、Zheng ら [11] が「LLM-as-a-Judge (審査員としての LLM)」の有効性を検証しており、強力な LLM

**S: Substitute (代用する)**

構成要素の一部を他のものに置き換える視点。

「ルール、材料、あるいは人を置き換えることは可能か？」と問うことで、既存の制約条件を取り払い、代替案の可能性を探索させる。

**C: Combine (組み合わせる)**

異なる目的、アイデア、工程などを結合する視点。

独立して存在する要素同士に関連性を見出し、

統合することで、「相乗効果 (Synergy)」を生み出すことを目的とする。

**A: Adapt (適応させる)**

他の文脈や過去の事例からのアイデアを、現在の課題に適応させる視点。

異分野の解決策を借用することで、現在の課題に対して外部からの新たな洞察を取り入れる。

**M: Modify (修正・拡大・縮小する)**

形状、属性、意味などを変更、あるいは拡大 (Magnify) ・縮小 (Minify) する視点。

対象の属性を極端に変化させることで、通常では見落とされがちな特性や新たな価値を顕在化させる。

**P: Put to other uses (他の用途に転用する)**

本来の目的とは異なる使い方を考える視点。

既存の市場や文脈以外での利用可能性を探ることで、対象の潜在的な価値を再定義することを推奨している。

**E: Eliminate (削減する・取り除く)**

不要な部分、機能、工程を取り除く視点。

複雑性を排除し、対象の核となる機能 (Core Function) に焦点を絞ることで、本質的な問いを生成させる。

**R: Reverse / Rearrange (逆転・再構成する)**

順序、配置、因果関係を逆にしたり、並べ替えたりする視点。

論理プロセスを反転させることで、既存の前提条件の妥当性を疑い、構造的な欠陥や新たなフローを発見する。

図1 SCAMPER の定義と具体例 [13]

(GPT-4 等) を用いた評価は、人間の専門家による評価と高い相関を持つことが示されている。国内においても伊藤ら [14] [15] [16] [17] は、複数 LLM の活用仕組 Ensemble ToT とそれを活用した自己学習支援システム「GET」を提案し、自動採点における信頼性と説明性の向上に取り組んでいる。

本研究では、質問力の向上やマルチエージェントによる議論といった知見を統合し、肯定・否定・中立の立場からの議論を通じて、単一の「正解」が存在しない創造的な質問に対しても、納得感のある公平な評価を行うことを目指す。

**2.4 QFT (問いづくり技法)**

本システムは、学生が自律的に問いを立てるプロセスを支援するため、QFT (Question Formulation Technique) の枠組みを採用する。QFT は、学習者が自ら質問を作成し、改善し、優先順位をつける能力を育成するための体系的な手法である。Mwikya [6] によれば、QFT のプロセスは以下の 4 つの厳格なルールに基づく「質問生成 (Produce Your Questions)」から始まる。

1. できるだけたくさんの質問を出す。

2. 質問について話し合ったり、評価したり、答えを言ったりしない。
3. 質問は発言された通りに書き出す。
4. 意見や主張は疑問文に書き直す。

このプロセスにおいて、判断や評価を一時的に停止することは、拡散的思考 (Divergent Thinking) を促進するために極めて重要である。しかし、初学者が何の手がかりもなくこのルールに従って質問を出し続けることは認知的な負荷が高い。そこで本研究では、次節で述べる SCAMPER 法をこのプロセスのスタートである質問作成において、強力な支援方法として導入し、質問の作成支援を行う。

**2.5 質問の質的評価における理論的背景と LLM による自動評価**

本研究における質問の評価基準を策定し、LLM による自動評価の妥当性を担保するために、教育学における理論的枠組みを参照する。教育目標の分類として Bloom のタキソノミーを参照する。Scaria ら [22] の研究では、LLM が「分析」や「創造」といった高次な思考を促す質問を識別・

評価できることが実証されている。質問の「深さ」を評価する指標として、Graesser ら [24] のカテゴリ分類を採用する。Graesser らは、事実確認のみの「浅い問い」に対し、因果関係やメカニズムを問う「深い問い」を定義し、後者が理解の深化に不可欠であることを示した。

### 3 提案システム

#### 3.1 SCAMPER 法とその自動化による効率化

SCAMPER 法は、既存のアイデアや対象に対し、7つの異なる操作を加えることで、強制的に新しい視点を獲得することを目的としたアイデア発想のチェックリスト法である。Ozyaprak [12] による各文字の定義や質問の具体例は図 1 に示す。

この手法の教育的効果については、Widihastuti ら [5] が実証的な検証を行っている。彼らの研究では、SCAMPER 法を用いた学習活動を取り入れた結果、学生の創造的思考スキル、特に思考の流暢性 (Fluency) と柔軟性 (Flexibility) が有意に向上したことが報告されている。しかし、前述の通り人的コストの問題から、多人数講義での実施は困難であった。本研究では、LLM がこの指導プロセスを自動化することで、教育効果を維持しつつ、効率的な質問力育成環境の構築を目指す。

#### 3.2 システムの概要と利用フロー

本システムは、Web ブラウザ上で動作する学習支援アプリケーションとして構築される。学習者は講義スライドを閲覧しながらシステムを利用し、全 7 回 (SCAMPER の創造思考プロセスの数) の活動を通じて質問作成能力の向上を目指す。具体的な利用フローは、以下の通りであり、図 2 に示す。また、システムの画面フローは図 3 に示す。

1. **視点の提示:** システムは学習者の履歴データベースを参照し、履歴データベースに存在しない視点を未実施の SCAMPER の視点とし、その中からランダムに一つを選択して提示する。同時に、その視点の定義を表示する。
2. **質問作成:** 学習者は提示された視点 (例: 「S: 代用する」) に基づき、スライドの内容に関する質問を作成し、入力する。同時に、その質問の根拠となったスライドページを選択させる。
3. **システムによる評価:** 入力された質問とスライドページに対し、バックグラウンドでマルチ LLM エージェントによる討論が行われ、学習者の質問作成・スライド番号

選択の二つの活動を評価する。

4. **フィードバック:** 討論の結果に基づき、合否判定と詳細なフィードバック (改善点や良かった点) が学習者に提示される。合格であればその SCAMPER の視点は完了となり、次の未実施の SCAMPER の視点での活動へ進むことができる。

#### 3.3 マルチ LLM エージェントによる評価機構

本システムの核心は、作成された質問に対する評価を単一の LLM で行うのではなく、複数のエージェントによる討論 (Debate) を通じて決定する点にある。

##### 3.3.1 エージェントの構成と役割

システムには以下の討論エージェント 3 種類と最終決定エージェント 1 種類の計 4 種類のエージェントが実装されている。

- **討論エージェント (3 種類):**
  - **Evaluator\_Agree (肯定):** 常に学習者の質問の良い点を探し、肯定的な立場から評価を行う。創造性や挑戦的な姿勢を評価する。
  - **Evaluator\_Disagree (否定):** 常に批判的な立場から評価を行う。視点の適用の誤りや、スライド内容との乖離、論理的な飛躍を厳しく指摘する。
  - **Evaluator\_Neutral (中立):** 客観的な立場から、肯定・否定双方の意見の妥当性を判断し、バランスの取れた意見を述べる。
- **最終決定エージェント (Judge):** 討論の進行役を務める。各ターンで発言するエージェントを指名し、最終的に議論の内容を要約して合否判定を下す。

##### 3.3.2 討論プロセス

討論は以下の手順で進行する。

1. 学習者が入力した「質問」とその根拠となる「選択したスライド」が、コンテキストとして全エージェントに共有される。
2. 最終決定エージェントが討論を開始し、最初のエージェント (肯定または否定) を指名する。
3. 指名されたエージェントは、自身の役割 (ペルソナ) に基づき、後述する評価基準について意見を生成し、前に発言したエージェントがつけた得点に同意をするか、得点をつけ反対意見を出すかのどちらかの行動を行う。
4. 最終決定エージェントは、出された意見を踏まえて次

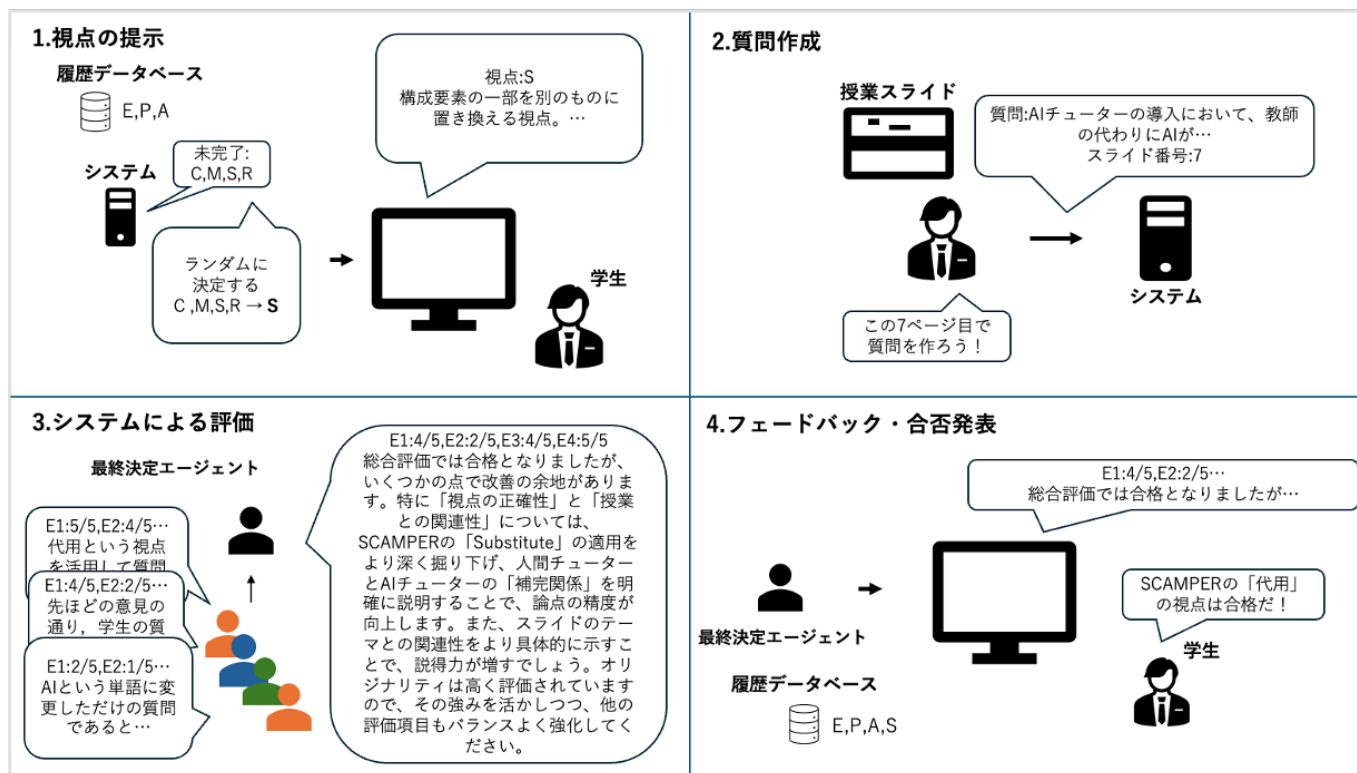


図2 具体的な利用フロー

に発言すべきエージェントを指名する。これを最大30ターン繰り返す。また、討論エージェントが一つの評価基準で討論を行い、四つの評価基準での討論が終了したら、最終決定エージェントが評価をまとめる場合について、評価点数・処理時間がどのように変化するかについても今後実験を行っていく。

- 3種類の討論エージェントが全員同じ点数で同意した場合、または規定ターン数に達した場合、最終決定エージェントは議論を終了し、最終評価を出力する。
- また、最終評価に関して、肯定意見を述べるエージェントの意見により、全ての評価基準において、高得点の評価を出す可能性が懸念されたため、質の低い質問を入力した場合の、最終評価を確認する。

### 3.3.3 評価基準

エージェントは以下の4項目について議論を行い、各項目5点満点(計20点)で採点を行う。合格ラインは12点以上とする。また、実験結果を示す際にはそれぞれの評価基準を番号(E1,E2,E3,E4)で示す。

- 視点の正確性 (Perspective):** 指定された SCAMPER の視点を正しく理解し、適用できているか。単なる語句の置換ではなく、視点の本質を捉えているか。

- オリジナリティ (Originality):** スライドの記述をそのままコピーしただけではなく、学習者自身の独自の考察や視点が含まれているか。
- 授業との関連性 (Relevance):** 講義の本質的な内容に関連しているか。スライドの文脈を無視した突飛な質問になっていないか。
- 素材選択の適切さ (Slide Match):** 質問の根拠として選択されたスライドページは適切か。質問内容とそのページの内容に関連性があるか。

### 3.3.4 評価基準型討論

評価基準型討論は、討論エージェントが4つの評価基準(視点の正確性、オリジナリティ、授業との関連性、素材選択の適切さ)のそれぞれについて、独立した討論セッションを順次展開する手法である。各セッションでは、肯定・否定・中立のエージェントが特定の基準のみに焦点を当てて議論を行い、基準ごとの得点を確定させていき、最終的には、その議論内容を元に最終決定エージェントが得点を決定する。

### 3.3.5 全体俯瞰型討論

全体俯瞰型討論は、まず全エージェントが質問の質について包括的な議論(全体討論)を行い、その議論の内容を踏まえて最終的に最終決定エージェントが各基準に基づいて採点を行う手法である。

## SCAMPER 思考トレーニング (Vision対応)

## Step 1: Combine (組み合わせる)

異なるもの同士を組み合わせる。目的を統合する、アイデアを統合するなど、何かが組み合わされないか考える視点。

## 今日の視点の使い方 (例題)



(a) 視点の提示

## 1. スライドを選択 (課題用: Part 2)



## 2. 質問を作成



(b) 質問作成

## AIによる評価会議



(c) 評価討論

## AIによる評価会議



(d) フィードバック

図3 システムの利用フロー詳細

## 4 実験と考察

## 4.1 予備実験概要

## 4.1.1 実験設定

提案システムの教育的効果および評価アルゴリズムの妥当性を検証するため、被験者実験を行った。被験者は20歳から24歳の学生9名であり、以下の3つのグループ(各3名)にランダムに割り当てた。

- **教示群 (Group 1)**: システムを利用せず、SCAMPER法の概念と各視点の定義について講義形式の教示を受ける。
- **基準別討論群 (Group 2)**: 開発システムを利用する。エージェント討論において、評価基準型討論・採点を行うアルゴリズムを採用する。
- **全体俯瞰討論群 (Group 3)**: 開発システムを利用する。エージェント討論において、全体俯瞰型討論・採点を行うアルゴリズムを採用する。

実験に使用する教材として、京都工芸繊維大学の講義資料である「ソフトウェア工学」[19]のスライドを用いた。提案システムにおいては、各被験者はこれらのスライドを閲覧し、提示されたSCAMPERの視点に基づいた質問作成タスクを計7回行う。

## 4.1.2 手順

実験は以下の手順で行う。

1. **事前テスト**: 全員に共通の短いスライドを提示し、5分間で可能な限り多くの質問を作成させる。
2. **学習フェーズ**: 各グループの条件に基づき、指定された教材スライドの提示、もしくは提案システムの利用による質問作成活動を行う。
3. **事後テスト**: 事前テストとは異なる部分のスライドを提示し、同様に5分間で質問を作成させる。

## 4.2 評価指標

本実験における教育的効果を多角的に測定するため、以下の3つの観点から指標を定義し、算出を行った。以下の、それぞれの指標での判定は Gemini[21] を使用している。

- **質問の量 (Quantity / 流暢性)**: 被験者が各セッション(システム利用前・利用後)において生成した質問の総数をカウントした。
- **質問の質 (Quality / 深層性・認知レベル)**:

表1 グループ別評価指標の平均値と変化量

グループ	項目	質問の量	SCAMPER 種類	Graesser	Bloom
全体俯瞰 (ID 1-3)	利用前	4.33	—	1.69	1.47
	利用後	4.67	3.67	2.73	4.17
	変化量	+0.33	—	+1.04	+2.70
基準別 (ID 4-6)	利用前	4.00	—	1.86	2.10
	利用後	5.00	4.33	2.93	4.98
	変化量	+1.00	—	+1.07	+2.88
教示群 (ID 7-9)	利用前	4.67	—	2.22	2.57
	利用後	5.00	3.67	2.62	3.93
	変化量	+0.33	—	+0.40	+1.36

表2 グループ別 SCAMPER の視点別出現回数

グループ	S	C	A	M	P	E	R
全体俯瞰	2	0	3	5	0	2	1
基準別	3	3	2	2	0	3	2
教示群	1	4	2	2	0	2	1

表3 システム利用群によるアンケート評価

グループ	使いやすさ	向上の実感	手間の少なさ
全体俯瞰	3.67	4.00	4.00
基準別	4.00	4.67	2.67

- **Graesser 指標 (構造的深さ)** : Graesser ら [24] の理論に基づき、質問を「レベル1: 事実確認 (Shallow)」、「レベル2: 中間 (Intermediate)」、「レベル3: 因果・メカニズム (Deep)」の3段階で判定した。
- **Bloom 階層 (認知レベル)** : 教育目標分類 (認知領域) に基づき、質問を「1: 記憶」から「6: 創造」までの6段階で判定した。
- **計算方法** : 被験者ごとの全質問のスコアを合算し、質問数で除した平均値を個人の「質のスコア」とした。
- **質問のバリエーション (Flexibility / 柔軟性)** : 被験者が質問生成に使用した SCAMPER の7つの視点 (S, C, A, M, P, E, R) のうち、ユニークな種類の数をカウントした。

### 4.3 討論アルゴリズムによる評価指標の差異

実験の結果、評価の粒度の違いは、質問の構造的深さと認知レベルの双方において異なる影響を及ぼした。表1に基づく詳細な比較結果を以下に示す。Graesser 指標において、基準別討論群は利用前の 1.86 から 2.93 へと向上 (変化量 +1.07) し、全体俯瞰型討論群は 1.69 から 2.73 へと向上 (変化量 +1.04) した。両群の差は 0.03 ポイントと極めて僅差であり、いずれも教示群 (+0.40) を大幅に上回る結果となった。一方、Bloom 階層においては、基準別討論群 (変化量 +2.88) が全体俯瞰型 (変化量 +2.70) を 0.18 ポイント上回り、本実験において最も高い学習効果を記録した。

### 4.4 SCAMPER の視点別利用傾向と質問の深さ

表2に、各グループにおいて質問生成に使用された SCAMPER の視点別の出現回数を示す。全体俯瞰討論群では「Modify (M)」が5回と最も多く、既存の内容を修正・拡張するアプローチが中心であった。一方、基準別討論群では、「Substitute (S)」「Combine (C)」「Eliminate (E)」がそれぞれ3回ずつ均等に利用されており、特定の思考操作に偏らず、多様な角度から問いを立てている傾向が確認された。教示群においては「Combine (C)」が4回と突出しており、複数の要素を組み合わせる思考が定着しやすい一方で、それ以外の視点の活用は限定的であった。

### 4.5 考察

#### 4.5.1 粒度の違いがもたらす「問い」の変容

Graesser 指標と Bloom 階層の向上幅を統合的に分析すると、アルゴリズムごとに次の特徴が明らかになった。全体俯瞰型は、議論を要約して提示することで、質問の「構造 (Why/How を用いた深さ)」を改善させるには十分な能力を有している。しかし、質問の内容をより「洗練 (認知レベルの向上)」させるためには、基準別討論のように各観点から独立した批判を受けるプロセスがより効果的であると考えられる。これは、Bloom 階層による評価において、個別の基準に特化した精緻なフィードバックが、学習者の思考を「分析・評価・創造」といったより高次の段階へと押し上げたことを示唆している。また、Graesser 指標では、両群とも事後スコアが「深い問い」の閾値である 2.70 を超えた点は特筆すべきであり、討論形式の違いに関わらず、エージェント間の議論を提示する仕組み自体が質問の構造を深める上で極めて有効であることが示された。

#### 4.5.2 思考の多様化（柔軟性）と質問の質の相関

基準別討論群において使用された SCAMPER の視点数が最も多かった事実は、本システムが学習者の思考の柔軟性を拡大させたことを意味する。基準ごとに「オリジナリティ」や「素材選択の適切さ」を評価されることで、学習者は「代用」や「除去」といった、普段使い慣れない高度な思考フレームワークを積極的に採用する動機付けを得たと考えられる。前述の通り、これらの視点をを用いた質問が Graesser 指標のレベル 3（深い問い）を記録しやすい傾向にあったことから、「視点の多様化（種類の増加）」が「問いの深層化（質の向上）」を直接的に牽引しているという論理的関連性が明らかになった。対して全体俯瞰型では「Modify」等の特定の視点に偏る傾向が見られ、思考の柔軟性の拡大においては基準別討論に優位性が認められた。

#### 4.5.3 討論の粒度が認知プロセスと作業負荷に与える影響

アンケート結果が示した「基準別討論群の高い向上実感と高い負担感」という結果は、教育的介入の強度を表している。基準別の個別討論は、Bloom 階層（+2.88）を最大化させる強力な手段であるが、同時に学習者の認知的負荷を高めるトレードオフが存在する。一方で、全体俯瞰討論群においても Bloom 階層の向上（+2.70）と高い負担の少なさ（4.00）が両立されており、エージェント間の議論の要約を閲覧する形式が、学習効果を維持しつつ継続性を担保する効率的な足場かけとして機能していることが示された。

## 5 まとめ

本研究では、SCAMPER 法とマルチ LLM エージェント討論を統合した質問力向上支援システムを構築し、その有効性を実証した。被験者実験の結果、提案システムは学習者が使用する SCAMPER の視点の種類（バリエーション）を増やし、思考の柔軟性を向上させることで、Graesser らの定義する「深い問い」および Bloom 階層における「高次元の問い」の生成を促進することが明らかになった。今後の課題として、基準別討論による詳細なフィードバックは学習効果を最大化させる一方で負担も大きいと、学習者の習熟度や好みに応じて討論の粒度を調整するパーソナライズ機能の実装が、持続的な学習支援において重要になると考えられる。今後は Scaria ら [22] の知見を活かし、より大規模な環境下での自動評価の精度向上と能力定着度の検証を進めていく。

## 謝辞

本研究の一部は科研費（23K28094）による。

## 参考文献

- [1] Rahman, M. M. and Watanobe, Y.: "ChatGPT for Education and Research: Opportunities, Threats, and Strategies", *Applied Sciences*, Vol.13, No.9, 5783 (2023).
- [2] OECD: "PISA 2022 Results (Volume I): The State of Learning and Equity in Education", OECD Publishing (2023).
- [3] Chin, C. and Osborne, J.: "Students' questions: a potential resource for teaching and learning science", *Studies in Science Education*, Vol.44, No.1, pp.1-39 (2008).
- [4] Baidoo-Anu, D. and Owusu Ansah, L.: "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning", *SSRN Electronic Journal*, Vol.7, No.1, pp.52-62. (2023)
- [5] Widiastuti, W. and Rifa'i, A.: "The Effectiveness of SCAMPER Techniques on Creative Thinking Skills", *Research and Evaluation in Education*, Vol.6, No.2, pp.111-120 (2020).
- [6] Mwikya, M. K.: "21st Century Pedagogical Skills: The Question Formulation Technique (QFT) for Classroom Instructions", *International Journal of Research and Innovation in Social Science*, Vol.8, No.3, pp.827-836 (2024).
- [7] Lee, H. P. H., et al.: "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers", *Microsoft Research*, pp.1-22. (2025)
- [8] Du, Y., et al.: "Improving Factuality and Reasoning in Language Models through Multiagent Debate", *arXiv preprint arXiv:2305.14325* (2023).
- [9] Liang, T., et al.: "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate", *arXiv preprint arXiv:2305.19118* (2023).
- [10] Xiong, K., Ding, X., Cao, Y., Liu, T. and Qin,

- B.: "Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate", Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 7572–7590 (2023).
- [11] Zheng, L., et al.: "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", arXiv preprint arXiv:2306.05685 (2023).
- [12] Ozyaprak, M.: "The Effects of SCAMPER Technique Activities in the 8th Grade Simple Machines Unit on Students", Journal of Education and Learning, Vol.10, No.2, pp.129-140 (2021).
- [13] Serrat, O.: "The SCAMPER Technique", Knowledge Solutions, Springer, pp.311-314 (2017).
- [14] 伊藤優希, 馬強: "GET: 複数 LLM を活用した自己学習支援のための自動採点システム", 第 17 回データ工学と情報マネジメントに関するフォーラム (DEIM2025), 3H-03 (2025).
- [15] Yuki Ito, Qiang Ma: "Supporting student self-learning using generative AI", in Proceedings of the 2024 the 16th, pp.97-103 (2024).
- [16] Yuki Ito, Qiang Ma: "Ensemble ToT and Its Application to Automatic Grading", Database and Expert Systems Applications: 36th International Conference, DEXA 2025, Bangkok, Thailand, August 25–27, 2025, Proceedings, Part I, pp.230 - 244 (2025).
- [17] Yuki Ito, Qiang Ma: "ENSEMBLE TOT OF LLMs AND ITS APPLICATION TO AUTOMATIC GRADING SYSTEM FOR SUPPORTING SELFLEARNING", arXiv preprint arXiv:2502.16399(2025).
- [18] 京都工芸繊維大学教材「情報分析概論」, <https://moodle.cis.kit.ac.jp/>, 2025
- [19] 京都工芸繊維大学教材「ソフトウェア工学」, <https://moodle.cis.kit.ac.jp/>, 2025
- [20] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre-Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, William El Sayed: "Mistral 7B", arXiv preprint arXiv:2310.06825 (2023).
- [21] Google: "Gemini 1.5 Flash", <https://gemini.google.com/>.
- [22] Scaria, N., Chenna, S. D. and Subramani, D.: "Automated Educational Question Generation at Different Bloom's Skill Levels using Large Language Models: Strategies and Evaluation", arXiv preprint arXiv:2408.04394 (2024).
- [23] Wang, Z., Valdez, J., Basu Mallick, D. and Baraniuk, R. G.: "Towards human-like educational question generation with large language models", In AIED 2022, pp.153–166 (2022).
- [24] Graesser, A. C. and Person, N. K.: "Question asking during tutoring", Vol.31, No.1, pp.104–137 (1994). <https://doi.org/10.3102/00028312031001104>

# 生成 AI を用いた保有技術分析手法

後藤 直紀<sup>†</sup> 多田 裕美<sup>†</sup> 柴田 秀哉<sup>†</sup>

<sup>†</sup>三菱電機株式会社 情報技術総合研究所 〒247-0056 神奈川県鎌倉市大船5丁目1-1

E-mail: <sup>†</sup>{Goto.Naoki@dh, Tada.Hiromi@cw, Shibata.Hideya@cb}.MitsubishiElectric.co.jp

**あらまし** 企業などの組織が持続的に発展するには研究開発戦略が重要であり、その策定には組織が保有する技術の全体像把握が不可欠である。しかし、特に大規模な組織では、部門ごとに技術開発、成果管理が行われていることが多く、組織横断的な保有技術の把握に時間と労力がかかるという課題がある。この課題に対し、組織が保有する技術文書群に対して、クラスタリング等の文書分類技術を適用し、文書間の類似性等から保有技術の全体像を把握する試みは有効と考えられる。しかし、単純なクラスタリングのみでは、各分類がどのような技術を表すかの理解が難しいことに加え、非階層的なクラスタリングを用いた分析では、分類間の関係性が適切に把握できないという課題がある。そこで本論文では、階層分類と生成 AI を活用した分類見出し作成を組み合わせた保有技術分析手法を提案する。また、提案手法の有効性を主観評価により検証した結果を説明する。

**キーワード** 保有技術の全体像把握、文書分類、生成 AI、開発戦略の策定支援

## 1. はじめに

企業などの組織における研究開発は、組織の持続的な発展や競争力の強化に直結する重要な活動である[1]。従って、各組織が将来を見据えて適切な研究開発戦略を策定することは、運営上の重要課題の一つである。

適切な研究開発戦略を立案するためには、まず組織が保有する技術の全体像を正確に把握することが不可欠である。しかし、大規模な組織では、複数の部門やグループが独立して多様な研究開発を行っており、技術成果が組織内の各所に分かれて管理されている。このような状況では、組織全体の保有技術を横断的に把握するために多大な労力と時間を要する。

このような背景から、技術全体像を効率的に把握可能とする手法の確立が求められている。それにより、戦略策定に必要な情報を迅速かつ正確に収集・整理することが可能となり、結果としてより適切な研究開発戦略の立案につながる。

保有技術全体像を把握するための有効なアプローチの一つとして、技術文書の活用が挙げられる[2]。組織に蓄積された技術文書には、各技術の内容や開発の背景、適用分野などの詳細な情報が含まれている。そのため、それらを分析することで、保有技術の特徴や複数の技術間の関係性を明らかにすることが可能であり、保有技術全体像の理解を促進できる。

文書群を分析する単純な手法としては、k-means 法などの非階層的なクラスタリングを用いた手法が挙げられる。例えば、特許分析において文献[3]では、調査対象とする技術領域の技術構成などを考察するために、調査対象の公開広報をクラスタリングし、関連するもの同士を近い位置にマッピングするような例が示されている。この手法では、同一または類似する技術に関

する文書を同じ分類に割り当て、各分類に含まれる文書数や文書のタイトルの確認、平面上への可視化などを通じて、技術の保有状況を把握できる。しかし、この手法では、分類の階層構造が把握できないこと、及び各分類に割り当てられた文書の内容的特徴の把握に手間がかかることという2つの課題がある。

そこで本論文では、前者の課題を解消するため、階層的なクラスタリングを導入し、分類の階層構造を明確化する。これにより、技術間の関係性を把握しやすくする。また、後者の課題に対しては、生成 AI を活用して分類見出しを作成し、分類の内容的特徴を明示する。これらの手法により、保有技術の全体像を容易に把握することを可能とする。

本論文では、2章で提案手法、3章で提案手法の評価について説明し、4章でまとめを行う。

## 2. 提案手法

本章では、提案手法について説明する。提案手法では、文書の階層分類と分類見出しの作成を行う。

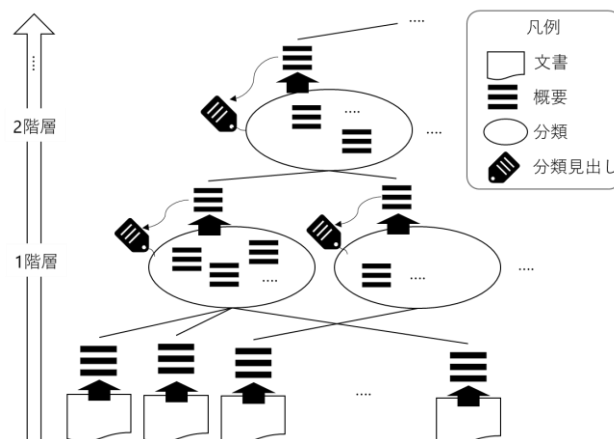


図 1 提案手法の全体像

まず、階層分類は、ボトムアップ的なアプローチで実施する。具体的な流れは、まず文書を分類し、その分類を段階的に統合するというものである。分類見出しの作成では、各分類の内容的特徴を説明する分類概要を作成し、分類概要を基に分類見出しを作成する。提案手法の全体像を図 1 に示す。

## 2.1. 文書の階層分類

本節では、文書の階層分類について説明する。本処理は、以下の 2 つの手順から成る。

1. 非階層的な文書分類
2. 段階的な分類の統合

まず、手順 1 では、非階層的クラスタリングを用いて、文書を分類する。文書を分類する手法として、テキストに対応するベクトルを用いたクラスタリングが挙げられる[4]。しかし、図などのテキストでない形式のデータを多く含む文書の場合、文書に記載されているテキストだけに対応するベクトルを用いた分類では、精度が低下する可能性がある。また、文書ごとに書式が異なる場合に、それが分類精度を低下させることも考えられる。そこで本論文では、テキストデータだけでなく、図などのテキストではないデータも参照して、統一された書式で文書の概要を作成し、それに対応するベクトルを用いて文書を分類する。

以上を纏めると、手順 1 は次のように細分化される。

- 1-1. 文書に対して、マルチモーダルな生成 AI を用いて文書の概要を作成する。
- 1-2. 手順 1-1 で作成した概要に対応するベクトルを作成する。
- 1-3. 手順 1-2 で作成したベクトルを用いて、非階層的なクラスタリングを行い、その結果を基に分類を作成する。

手順 1-3 において、本論文では k-means 法を使用することとし、事前に分類数を指定する。

次に、手順 2 では、手順 1 により作成した分類を段階的に統合することで、階層分類を作成する。この際、手順 1 により作成した分類を最下位層（1 階層）と定義し、ある層の分類を統合して新たに作成される分類を 1 つ上位の層と定義する。

手順 2 の詳細を以下に記す。また、手順 2 のフロー図を図 2 に示す。

- 2-1. 最下位層を現在の層に設定する。
- 2-2. 現在の層の各分類に対して、分類概要を作成する。
- 2-3. 指定した層に到達すれば終了する。そうでなければ手順 2-4 に進む。
- 2-4. 手順 2-3 で作成した概要に対応するベクトルを作成する。

2-5. 手順 2-4 で作成したベクトルを用いて非階層的なクラスタリングを行い、その結果を基に分類を統合する。

2-6. 手順 2-5 で作成した上位層を現在の層に設定し、手順 2-2 へ戻る。

手順 2-2 について補足する。分類概要は、各分類に含まれる文書の概要を基に生成 AI を用いて作成する。具体的には、複数の概要文から共通の話題を抽出すること、可能な限り具体的な内容とすることなどの指示を生成 AI に与えて作成する。

ただし、最下位層より上位の層については、各分類に含まれる文書の概要の代わりに、1 つ下位の層での分類に対する分類概要を使用して当該分類概要を作成する。このようにすることで、概要作成時に参照すべきテキスト量を抑えることができる。また、ある分類の下位の層の分類概要に含まれない内容が当該分類の分類概要に含まれることが起こりにくくなり、階層間で分類概要の記載内容の不整合が生じにくくなることが期待できる。

手順 2-5 では、事前に分類数を指定することとする。

## 2.2. 分類見出しの作成

分類見出しを作成する目的は、ユーザに対して各分類の内容を分かりやすく提示することである。分類概要にも分類内容が記されているが、分類での使用に耐える量の情報を含むため、そのままでは見出しとして過剰である。

そこで本処理では、分類概要を基に各分類に適した簡潔な見出しを生成 AI により作成する。具体的には、分類概要を基に適切な見出しを作成すること、文字数制限などの指示を生成 AI に与えて作成する。

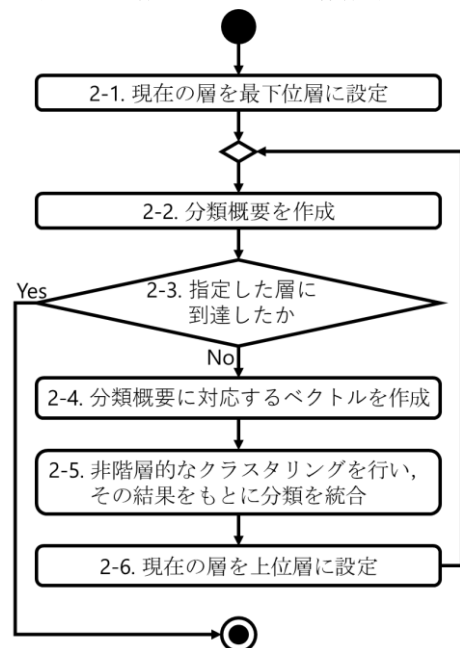


図 2 手順 2 のフロー図

### 3. 評価

本章では、提案手法の有効性を検証する。評価の目的は、技術文書群を提案手法により分類した結果が保有技術を把握する上でどの程度有用であるかを確認することである。評価方法としては、ある部署で作成した技術文書群を対象に、当該部署の所属員による主観評価を採用した。

#### 3.1. 評価項目

本評価では、以下の2つを評価項目とした。

1. 分類精度
2. 分類見出しの自然さ

分類精度は、各分類に含まれる文書群の中で、適切ではない、または違和感がある文書がどの程度含まれているかという観点での評価項目である。また、分類見出しの自然さは、作成される分類見出しが適切ではない、または違和感があるかどうかという観点での評価項目である。

なお、提案手法の有効性を検証するためには、上記2つの評価項目に加えて、分類の階層構造が適切である、または違和感があるかどうかという観点での評価項目もまた必要である。当該項目の評価は今後の課題である。

#### 3.2. 評価方法

本節では、前節で設定した評価項目の評価方法について説明する。評価者は、評価対象の技術文書群を作成した部署の所属員であり、評価者からアンケート形式で意見を収集する形で評価を実施した。

評価項目1では、ある階層での分類結果を評価者に提示した上で、以下の4つの選択肢の中から回答を選択する形式とした。

1. 適切・違和感なし、または数件不適切・違和感あり
2. 半数程度不適切・違和感あり
3. 文書間に関連性なし
4. 判断できない

評価者が回答として選択肢1を選択する場合、当該分類に含まれる文書は概ね適切であると考えていると言える。また、選択肢2, 3を選択する場合、当該分類に含まれる文書は適切ではないと考えている可能性が高い。選択肢2と3を区別した理由については後述する。

評価項目2では、評価項目1と同じ階層の分類見出しについて、以下の2つの選択肢の中から回答を選択する形式とした。

1. 適切・違和感なし、またはほぼ違和感なし
2. 違和感あり、または不適切

評価者が評価項目2の回答として選択肢1を選択する場合、当該分類の見出しは当該分類を概ね正しく表現できていると考えていると言える。一方、選択肢2を

選択する場合、見出しが当該分類を適切に表現できていないと考えている可能性が高い。

評価項目2については、全評価者、全分類を対象とするのではなく、評価項目1で評価者が選択肢1、または2を選んだ場合のみを対象とした。これは、選択肢1, 2を選ぶ評価者は、少なくとも当該分類は存在していると考えていると言えるためである。一方、選択肢3を選択した評価者は、当該分類は分類として成立していないと考えている可能性が高いため、評価項目2の評価対象からは除外する。評価項目1で選択肢2と3を区別したのは、このような評価を実現するためである。

#### 3.3. 評価条件

本評価では、244件の技術文書を評価対象とする。階層分類は3階層で実施し、最上位層である3階層での分類結果(分類数は15個)に対して評価者より意見を収集した。評価者の人数は14名である。

文書の概要、分類概要、及び分類見出しの作成には、Anthropic社のClaude 3.5 Sonnetを使用した。文書の概要に対応するベクトル及び分類概要に対応するベクトルの作成には、Amazon社のTitan Embeddings G1-Textを使用した。

#### 3.4. 評価結果

評価項目1, 2の評価結果をそれぞれ図3, 図4に示す。両図の縦軸と横軸はそれぞれ、各分類と当該分類に対する各選択が選択された割合を示す。

評価項目1について、選択肢1「適切・違和感なし、または数件不適切・違和感あり」が選択された割合を確認すると、分類13, 14, 15の3個の分類において割合が相対的に低く、他の12個の分類での割合が70%以上であることと比較し、その差が顕著である。従って、これらの3個の分類は分類精度に問題がある可能性が高い。分類精度が低い主な原因として、どの分類にも属さない、いわゆる「その他」カテゴリに含まれるような文書が相対的に多く含まれていることが挙げられる。

評価項目2について、選択肢1「適切・違和感なし、またはほぼ違和感なし」を選択した人の割合を確認すると、分類1, 9, 12, 13, 14, 15の6個の分類において割合が70%を下回っており、他の9個の分類と比較して相対的に低いことが分かる。従って、これらの6個の分類については、作成された分類見出しが当該分類を適切に表現できていない可能性が高い。

評価項目2で分類見出しが自然でない判断された6個の分類のうち、分類13, 14, 15については、評価項目1においても分類精度が低いと判断されていた。この分類精度の低さが、見出しの作成精度が低い主要因であると考えられる。残りの分類1, 9, 12に関しては、

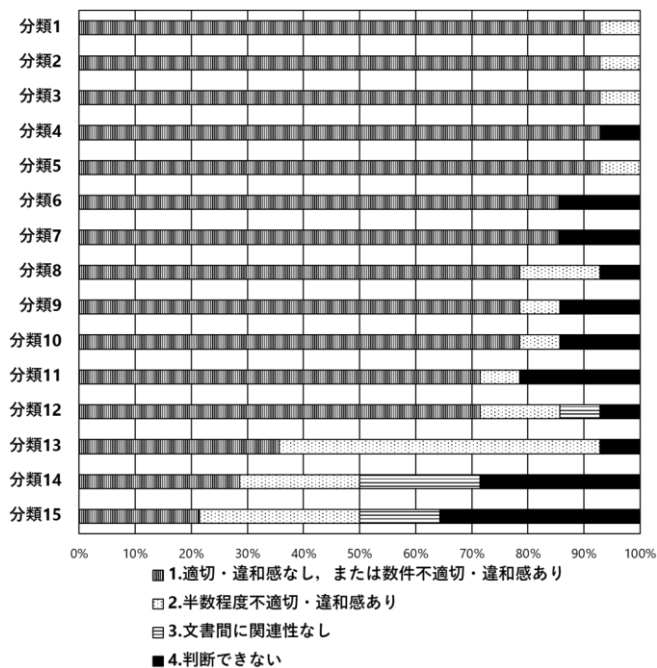


図 3 評価結果（評価項目 1：分類精度）

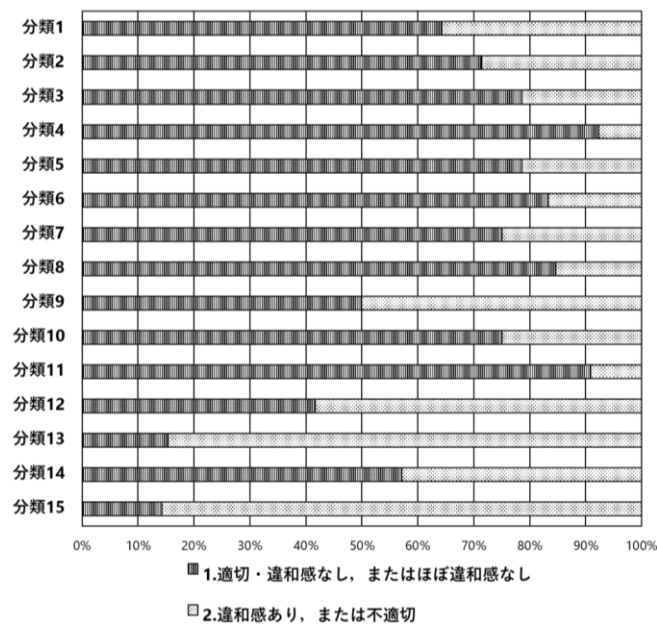


図 4 評価結果（評価項目 2：分類見出しの自然さ）

原因の一つとして、分類概要から分類見出しを作成する際に、重要な特徴を表す語句が抜け落ちていることが挙げられる。これに関しては、分類見出しを作成する際の生成 AI への指示において、文字数制限を緩和することで改善可能な見込みを得た。

また、分類見出しが当該分類を適切に表現できていない原因として、当該分類に含まれる一部の文書に特有の内容が分類概要に記載されていることが挙げられる。例えば、分類概要に「この分類は A と B を説明している」と記載されている場合に、各文書が必ずしも A と B の両方を説明しているわけではなく、A または

B のいずれか一方のみを説明しているというような場合があることを確認した。分類見出しは分類概要を基に作成しているため、分類概要の該当部分が分類見出しにも反映され、分類見出しに対して違和感を覚えやすくなっていると考えられる。

以上のことから、分類概要を適切に作成し、分類見出しの文字数を調整することで、分類見出しの自然さを向上できると考える。適切な分類概要の作成方式の検討については、今後の課題である。

#### 4. まとめ

本論文では、文書の階層分類と生成 AI を活用した分類見出し作成を組み合わせた保有技術分析手法を提案した。そして、提案手法の有効性を検証するために主観評価を実施した。

主観評価の結果、分類精度については、評価対象の 8 割の分類に関して、分類精度が相対的に高く、残りの 2 割の分類に関して、分類精度に問題がある可能性が高いことを確認した。分類精度が低い原因として、本来「その他」扱いすべきような少数カテゴリに含まれる文書を適切に扱えていないことが挙げられる。これに対する改善策の検討は、今後の課題である。

また、分類見出しの自然さについては、評価対象の 6 割の分類に関して、作成された分類見出しに違和感を覚える人の割合は相対的に低く、残りの 4 割の分類に関して、作成された分類見出しが当該分類を適切に表現できていない可能性が高いことを確認した。分類概要を適切に作成し、分類見出しの文字数を調整することで、分類見出しの自然さを向上することができる。適切な分類概要の作成方式の検討については、今後の課題である。

本論文では評価できなかったが、分類の階層構造が適切であるかどうか、または違和感があるかどうかという観点での評価を実施することも今後の課題である。

#### 参考文献

- [1] 塩谷景一, “日本の超大企業の研究開発システムの実態－製造関連企業の事例研究－”, <https://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-DP154-FullJ.pdf>, 2025 年 12 月 2 日閲覧。
- [2] H. Nonaka, A. Kobayashi, H. Sakaji, Y. Suzuki, H. Sakai and S. Masuyama, “Extraction of the effect and the technology terms from a patent document”, *Journal of Japan Industrial Management Association*, Vol. 63, No. 2, pp. 105-111, 2012.
- [3] 特許庁, “経営戦略に資する IP ランドスケープ実践ガイドブック”, [https://www.jpo.go.jp/support/example/ip-landscape-guide/document/index/all\\_guide\\_book.pdf](https://www.jpo.go.jp/support/example/ip-landscape-guide/document/index/all_guide_book.pdf), 2025 年 12 月 2 日閲覧。
- [4] A. Petukhova, J. P. Matos-Carvalho and N. Fachada, “Text clustering with large language model embeddings”, *International Journal of Cognitive Computing in Engineering*, Vol. 6, pp. 100-108, December, 2025.

一般発表 | Track 1: 自然言語処理・機械学習基礎

2026年3月2日(月) 13:00 ~ 15:10 | B会場

**[8B] 生成AI応用 II**

座長: 瀧本 祥章(NTT株式会社) コメントータ: 太田 学(岡山大学) ジュニアコメントータ: 桑田 若菜(兵庫県立大学)

14:35 ~ 15:00

**[8B-05] [技術報告] AIエージェントによる「業務の自動運転化」に向けた技術的挑戦と実践例**\*松村 優也<sup>1</sup> (1. 株式会社LayerX)

発表者区分: スポンサー

種別: 技術報告

インタラクティブ発表: あり

キーワード: AIエージェント、LLM、VLM、RAG、汎用機械学習技術

LayerXでは、LLMやAIエージェント、機械学習をはじめとするAI技術を駆使し、複雑なビジネスプロセスを自律化する「業務の自動運転化」を目指しています。

本セッションでは、この実現に向けて直面している技術的課題と、それらに対する具体的な取り組みについて紹介します。具体的には、企業ごとに異なる多様な業務ルールをいかにしてAIに学習・実行させるのか、また、AIの精度が完全ではない状況において、どのように「人間による介在」をシステムに組み込み、業務を完遂させるのかといった点について、事例を交えて解説します。