修士論文

お題の類似性による回答作成手順の抽出と LLM による自動評価を用いた Tree-of-Thoughts による回答自動生成

高橋 巧実

2026 年 1 月 30 日

岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域

鈴木研究室

Title1

本論文は岐阜大学大学院自然科学技術研究科に
修士（工学）授与の要件として提出した修士論文である.

高橋 巧実

指導教員：
鈴木 優 教授

お題の類似性による回答作成手順の抽出と LLM による自動評価を用いた Tree-of-Thoughts による回答自動生成*

高橋 巧実

内容梗概

本研究では，大喜利における人間の回答作成手順を LLM に入力することによって，回答の自動生成を目指す．我々は，人間は面白さに寄与する要素を入れ込んで回答を作成しているが，大喜利データを用いた指示学習では LLM はそのように訓練されないため，面白い回答を生成できるとは限らないと考えた．しかし，人間の高評価回答から生成した回答作成手順の全てを入力するには，計算量が大きくなるという課題がある．そこで，Tree-of-Thoughts を参考にした段階的な回答生成を提案する．お題どうしの類似度による回答作成手順の絞り込みと LLM による回答の自動評価によって，計算量を減少させる．類似するお題どうしは同じ回答作成手順が有効であると仮定し，文章分散表現の類似度をもとに使用する回答作成手順を抽出した．LLM によって回答を自動評価できると仮定し， LLM を用いて次段階の入力となる回答を選択する．実験では，回答作成手順を LLM に入力して生成した回答に対して評価者の主観による面白さ評価を行った．回答作成手順の適用によって面白い回答を生成できるかどうかを，生成した回答の面白さと既存手法によって生成した回答の面白さを比較することによって検証した．お題どうしの類似度によって対象のお題と相性の良い回答作成手順を絞り込めるかどうかを，文章分散表現の類似度に基づいて抽出した回答作成手順を適用して生成した回答に対する面白さ期待値を確認することによって検証した．LLM によって回答を自動評価できるかどうかを，LLM による自動評価に基づいて抽出した回答に対する面白さ期待値

*岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域 修士論文, 学籍番号: 1241752810, 2026 年 1 月 30 日.

を確認することによって検証した.

キーワード

大喜利，大規模言語モデル，生成 AI

目次

図目次

表目次

第 1 章　はじめに

本研究の目的は，大喜利の回答自動生成ならびに LLM (Large Language Model; 大規模言語モデル) による創造的タスクにおける回答作成手順の構造化・再利用ができるかどうかを検証することである．大喜利とは，与えられたお題に対して面白く回答をする演芸の一つである．

LLM はさまざまな創造的なタスクにおいて，創造的な出力を生成することができる．しかし，LLM の生成プロセスはブラックボックスであるため，LLM の生成プロセスは体系化されていない．生成プロセスが体系化されていないと，別の入力における良い出力を導くまでの手順を体系的に再利用できないという課題がある．

この課題は大喜利においても同様である．大喜利のお題と回答を使用して教師あり学習した LLM は，ある程度面白い回答を生成することがあるとわかっている．しかし回答の生成において，人間が回答作成時に重要視している面白さ要素が，LLM の生成プロセスにおいても同様に重要視されているとは限らない．我々は，面白い回答を生成するためには，回答生成時に面白さ要素を考慮する必要があると考えた．

そこで本研究では，人間が大喜利においてお題から回答を導くまでの考え方を構造化する．そして，別のお題における構造化した考え方を再利用することによって面白い回答を生成することを目指す．また，回答の生成を通じて，創作タスクにおける回答作成手順の構造化・再利用ができるかどうかを検証する．

本稿では CoT(Chain-of-Thought; 思考の連鎖プロンプト) による手法と，ToT(Tree-of-Thoughts) を参考にした手法の，2 種類の手法を提案する．CoT による手法では，人間による回答を導く考え方を創作具体例として構造化する．創作具体例を LLM に入力することによって回答を生成する．創作具体例は，お題と回答，および考え方を参考にしてお題から回答を導く過程からなるテキストである．創作具体例は CoT 形式のプロンプトである．CoT とは，特定のタスクにおける入力と出力および入力から出力を導く思考をプロンプトで例示することによって，LLM のタスク精度向上を行うテクニックである．そのため生成プロセスを思考として表現できるタスクでは，入力から出力を導く考え方を CoT によって構造化できるといえる．我々は，大喜利が CoT によって構造化可能であると考え，創作具

体例の入力によって面白い回答が生成できると仮定した．

我々は「こんな○○はいやだ」形式のお題に対する創作具体例を五つ作成する．創作具体例を入力することによって，人間の回答を導く考え方を参考にした回答生成を LLM に促す．創作具体例を LLM に入力して生成した回答には，人間による回答を導く考え方が反映されるようになり，回答に面白さ要素が含まれるようになると考えた．

ToT を参考にした手法では，人間がお題から回答を導くまでの考え方を，回答作成手順として構造化する．回答作成手順とは，人間がお題から回答を導くまでの考え方をいくつかの手順に区切ったものである．我々は，別のお題において回答を導くまでの考え方をいくつか組み合わせて再利用することによって，人間は回答を作成していると考えた．また，大喜利ではお題に回答作成手順を適用して途中回答を作成し，さらに回答作成手順を適用することによって次の途中回答を作成することによって，回答を作成できると考えた．

そこで我々は，人間の高評価回答をもとに回答作成手順を生成し，高評価回答を導く考え方を回答作成手順として構造化した．回答作成手順を段階的に適用することによって，人間と同様に別のお題において回答を導く考え方をいくつか再利用した回答生成を LLM に促す．しかし，回答作成手順の総数が増えるにつれて，組み合わせ数も膨大になるという課題がある．そこで我々は，ToT を参考にした回答作成手順の探索による回答生成手法を提案する．ToT とは．段階的な回答生成と LLM 自身による自動評価を組み合わせた LLM の推論フレームワークである．ToT では，LLM が出力した複数の思考を LLM 自身が評価を行い，評価が高い回答をもとに次段階の回答を生成することによって，複雑なタスクにおける精度向上が期待できる．そのため ToT の適用範囲は LLM によって回答の評価が可能であり，回答が段階的に洗練されるタスクであるといえる．我々は，大喜利は ToT の適用範囲であると考えた．

ToT を参考にした手法では，一つの段階において複数の回答作成手順を適用することで複数の回答を生成する．生成した回答に対して LLM による自動評価を行い，評価が高い回答のみをもとに次段階の回答を生成する．生成と評価を何回か繰り返すことによって，段階的に洗練された回答を生成する．

ToT を参考にした手法では，いくつかの仮定を導入することによって回答作成手

順の組み合わせ数削減を行っている．我々は，類似するお題どうしは同じ回答作成手順を適用することによって面白い回答を生成できると仮定した．そこで，対象のお題と文章分散表現のコサイン類似度が大きいお題から生成した回答作成手順のみを LLM に入力する．また，LLM によって回答の面白さを評価できると仮定した．各段階の途中回答に対して LLM による自動評価を行い，評価が高い途中回答のみから次の途中回答を生成する．回答作成手順の役割は適用順によって決まると仮定した．各段階の段階数と，生成の元となったお題と回答における適用順が一致する回答作成手順のみを適用して回答を生成する．タイトルが同じ回答作成手順は内容も同じであると仮定した．タイトルによって回答作成手順のグループ化を行い，同一の回答作成手順として扱う．

本研究の実験では，CoT による手法と ToT を参考にした手法それぞれにおいて導入した仮定を検証する．CoT による手法に対する実験では，創作具体例の入力によって面白い回答を生成できるかどうかを検証する．創作具体例を LLM に入力して生成した回答と，既存手法である通常プロンプトの入力および大喜利 LLM によって生成した回答を面白さにおいて比較することによって検証する．また，LLM のパラメータ数が大きいほど，CoT を入力することによって精度が向上する傾向があることがわかっている．そこで，創作具体例においても同様の傾向があるかどうかを検証する．同じ創作具体例をパラメータ数が異なる二つの LLM に入力し，生成した回答の面白さを比較することによって検証する．

実験の結果，四つの創作具体例を LLM に入力して生成した回答の面白さが，通常プロンプトの入力によって生成した回答の面白さよりも有意に低くなった．一つの創作具体例を LLM に入力して生成した回答の面白さが，大喜利 LLM によって生成した回答の面白さよりも有意に低くなった．そのため，創作具体例の入力によって面白い回答を生成できるとはいえなかった．また，全ての創作具体例において生成した回答の面白さは，使用する LLM のパラメータ数の違いによる有意差はなかった．そのため創作具体例には，LLM のパラメータ数が大きいほど精度が向上する傾向はないことがわかった．

ToT を参考にした手法に対する実験では，回答作成手順を適用して回答生成を 3 段階分行った．お題と相性の良い回答作成手順を適用できた場合を想定し，途中回答に対して評価者による評価を段階ごとに行うことによって，最も評価が良い途

中回答から次段階の途中回答を生成した．回答作成手順の段階的な適用によって面白い回答が生成できるかどうかを検証する．最終的に生成できた回答と既存手法によって生成した回答の面白さを比較することによって検証する．類似するお題どうしは同じ回答作成手順を適用することによって面白い回答を生成できるかどうかを検証する．お題どうしの文章分散表現の類似度によって抽出した回答作成手順を入力して生成した回答に対する面白さ期待値と，ランダム抽出した場合の面白さ期待値とを比較することによって検証する．LLM によって面白い回答を抽出できるかどうかを検証する．LLM の自動評価によって抽出した回答に対する面白さ期待値と，ランダム抽出した場合の面白さ期待値を比較することによって検証する．

実験の結果，生成した回答に対する面白さの最大値は，既存手法で生成した回答の面白さと比較して同じか大きかった．そのため，お題と相性の良い回答作成手順を適用できた場合，回答作成手順の段階的な適用によって面白い回答が生成できることがわかった．お題どうしの文章分散表現の類似度によって抽出した回答作成手順を入力して生成した回答に対する面白さ期待値が，ランダム抽出した場合の面白さ期待値を上回る事例がいくつか確認できた．そのため，類似するお題どうしは同じ回答作成手順を適用することによって面白い回答を生成できる場合があることがわかった．LLM の自動評価によって抽出した回答に対する面白さ期待値が，ランダム抽出した場合の面白さ期待値を上回る事例をいくつか確認できた．そのため，LLM によって面白い回答を抽出できる場合があることがわかった．

本研究の貢献は以下のとおりである．

- LLM の創作的タスクにおいて，回答作成手順の構造化という新しい視点の導入
- お題どうしの類似度と LLM 自動評価による回答作成手順の探索手法の提案
- 回答作成手順の構造化・再利用を大喜利への適用と検証

本論文の構成は以下の通りである．2 章では基本的事項について述べる．3 章では関連研究ついて述べる．4 章では提案手法ついて述べる．5 章では評価実験ついて述べる．最後に 6 章では本論文のまとめと今後の課題について述べる．

第 2 章　基本的事項

本章では本研究で用いた手法や指標について説明する．

2.1　大喜利

大喜利とは，与えられたお題に対して面白みのある回答をする演芸である．お題と回答の形式として，テキスト，発話，イラストやこれらの組み合わせ，対話などが挙げられる．本研究では，お題と回答の形式を共にテキストのみに限定している．

2.2　言語モデル

言語モデルとは，文章の生起確率をモデル化したものである．言語モデルはトークン化されたテキストを入力すると，入力のトークン列に続くトークンの出現確率を予測する．トークンとは，LLM にテキストを入力するために分割するときの単位であり，一つまたは複数の文字からなる．予測した出現確率が高いトークンを入力のトークン列に追加し，予測を繰り返すことによって文章を生成できる．

2.3　機械学習

機械学習とは，人工知能を実現する手段の一つである．人工知能とは，人間の知的行動をコンピュータが自動的に行うことである．機械学習の他に，ルールベースなどが人工知能を実現する手段として挙げられる．機械学習の長所は，データを与えることによって分類や回帰，生成といったタスクにおけるパターンをコンピュータが自動的に発見できる点である．

2.3.1　ニューラルネットワーク

ニューラルネットワークとは機械学習器の一つである．ニューラルネットワークは生物の脳を模した構造を持ち，層状に配置されたノードと活性化関数が主な構成

要素である．ノードは脳神経細胞であるニューロン，活性化関数はニューロンの発火現象に相当する．ノードは前の層にあるノードから入力を受け取る．受け取った入力を活性化関数に入力することによって，次の層にあるノードへの出力を計算する．活性化関数に非線形関数を用いることによって，ニューラルネットワーク全体が複数の非線形関数の合成関数になるため，大きな表現力を持つ機械学習器を作成できる．また，多層のニューラルネットワークを用いた機械学習手法を深層学習と呼ぶ．

Transformer

Transformer[1] とは，Google が 2018 年に発表したニューラルネットワークのアーキテクチャである．Attention 機構によるトークン間の関係に基づいた出力を行う Encoder-Decoder モデルである．これまで自然言語処理タスクに用いられていた RNN(Reccurent Neural Network; 回帰型ニューラルネットワーク) には，入力トークンに対して逐次処理を行うため並列処理ができないという課題があった．Transformer によって並列処理ができるようになったため，よりパラメータ数が大きいモデルを作ることができるようになり，大量のデータを用いた学習が可能になった．

LLM

LLM(Large Language Model; 大規模言語モデル) とは，文章生成タスクに特化したニューラルネットワークによる言語モデルである．LLM は，多数のパラメータを持ち，大量の文章データを用いて訓練されている．ChatGPT や Gemini のような大規模言語モデルの精度が向上した背景には，Transformer の登場によってインターネットから収集した大量のテキストデータを学習できるようになったことが挙げられる [2]．大量のテキストデータを用いた事前学習に加え，指示学習や RLHF(Reinforcement Learning from Human Feedback; 人間からのフィードバックによる強化学習) で訓練される．LLM に入力する文章はプロンプトと呼ばれる．プロンプトを通じて指示や例示を行うことで，タスクに沿った文章生成ができる．本研究では Meta が公開している Llama3[3] および Google が公開している Gemma3[4] を使用している．

埋め込みベクトル

埋め込みベクトルとは，データを特徴を表した固定長のベクトルである．埋め込みベクトルのほとんど全ての成分は非ゼロである．文章埋め込みベクトルに Sentence-BERT がある．Sentence-BERT[5] とは，BERT[6] という言語モデルに Pooling 層を連結した構造のモデルである．対照学習によって，文章の意味を考慮したベクトルを出力するよう訓練される．対照学習では，Siamese-Network という同じモデルが二つ連結した構造のモデルを用いて，意味が近い文章同士は類似度が大きいベクトルを，意味が異なる文章同士は類似度が異なるベクトルを出力するように調整される．本研究では，Google が公開している文章埋め込みモデルである EmbeddingGemma[7] を使用している．EmbeddingGemma も Sentence-BERT と同様に，改良版の Encoder を積み重ねたモデル構造であり，対照学習によって訓練される．高性能モデルのパラメータを利用した初期化と蒸留による，表現力の向上や多言語対応を達成している．

2.4 In-Context Learning

In-Context Learning とは，コンテキスト入力によってパラメータの再学習なしで LLM を特定のタスクに適応させることである．LLM にはコンテキストにおけるタスクの説明や指示，出力例など読み取り，目的のタスクに適応してテキストを出力する能力があることがわかっている [8]．LLM の適用能力によって In-Context Learning が可能になっている．In-Context Learning の例として，プロンプト・エンジニアリングや Few-Shot 学習が挙げられる．プロンプト・エンジニアリングとは，LLM への入力文であるプロンプトを調整することによって，LLM のタスク精度向上を狙うことである．Few-Shot 学習とは，特定のタスクにおける質問と解答の例をいくつか LLM に入力することによって，目的の質問に対する解答を LLM に出力させることである．

2.4.1 Chain-of-Thought

CoT(Chain-of-Thought；思考の連鎖プロンプト) とは，LLM に最終的な答えを出力するまでの思考過程を生成させることで，LLM のタスク精度向上を狙う技術である．Few-s Shot 形式の CoT[9] では，思考過程の例示を行う．特定タスクの入力と出力の例を複数提示する few-shot プロンプトにおいて，入力と出力の間に思考過程を例示する．これによって，例示した思考過程に沿った出力ができるようになるため，精度向上を狙えるようになる．また，プロンプトにある指示の末尾に「Let's think step by step.(ステップバイステップで考えてください．)」という文言を追加することによって，LLM に思考過程の生成を促す Zero-shot 形式の CoT[10] もある．

2.4.2 Tree-of-Thoughts

Tree-of-Thoughts[11] とは，段階的な推論と LLM による自動評価を組み合わせた LLM による推論の枠組みである．解答を生成するまでの中間ステップを探索することによって，探索や戦略的先読み，および初期における決定が重要であるタスクにおける精度向上が可能になる．中間ステップを複数生成し，LLM 自身が中間ステップの評価を行う．評価に基づいて，中間ステップを深さ優先や幅優先などのアルゴリズムによって探索する．数学パズル，クロスワードパズルにおいて，ToT は CoT や通常の指示プロンプトによる生成よりも精度が高かった．また，短い文章の創作タスクにおいて，ToT は CoT や通常の指示プロンプトによる生成よりも一貫性の高い文章を生成できた．

2.5 評価指標

2.5.1 t 検定

t 検定とは，二群間の平均値の差に対する統計的な検定手法である．標本をそれぞれ X と Y，X と Y の標本平均をそれぞれ $\bar{X}$，$\bar{Y}$，標本数を m，n，標本分散を

S_x，S_y であるときの統計検定量 T は，式 2.5.1 の通りである．

$$T = \frac{\sqrt{m+n-2}\left(\bar{X}-\bar{Y}\right)}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)\left(m{S_x}^2+n{S_y}^2\right)}} \tag{2.5.1}$$

T は自由度 $m+n-2$ の t 分布に従う．

2.5.2 Kappa 係数

Kappa 係数とは、評価者間の評価の一致度合いを表す指標である．評価者が二人の場合は Cohen の Kappa 係数を，三人以上の場合は Fleiss の Kappa 係数 [12] を使用する．本研究では，Fleiss の Kappa 係数を使用している．

Fleiss の Kappa 係数 κ は式 2.5.2 の通りである．

$$\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e} \tag{2.5.2}$$

このとき，N は評価の対象となるデータの数である．n は評価者の数である．k は一つのデータに対し付与される評価の種類数である．$i = 1, 2, \ldots, N$，$j = 1, 2, \ldots, k$ である．n_{ij} は i 番目のデータに j 種類目の評価を付与した評価者の数である．最初に，j 番目の評価を付与した人の割合 p_j を計算する．p_j は以下のように計算する．

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}, 1 = \sum_{j=1}^{k} p_j$$

p_j は評価の種類別割合であるため，全ての j で和を取ると 1 となる．

次に P_i を計算する．P_i は，i 番目のデータに対する評価が一致する評価者ペアの総数である．P_i は以下のように計算する．

$$
\begin{aligned}
P_i &= \frac{1}{n\,(n-1)} \sum_{j=1}^{k} n_{ij}\,(n_{ij}-1) \\
&= \frac{1}{n\,(n-1)} \sum_{j=1}^{k} \left({n_{ij}}^2 - n_{ij}\right) \\
&= \frac{1}{n\,(n-1)} \left[\sum_{j=1}^{k} \left({n_{ij}}^2\right) - n \right]
\end{aligned}
$$

i 番目のデータに対する評価のばらつきが均等になるほど，P_i は 0 に近づく．また，全ての評価者が一種類の評価に集中した場合，P_i は 1 になる．

$\bar{P}$ と $\bar{P_e}$ は以下のように計算する．

$$
\begin{aligned}
\bar{P} &= \frac{1}{N} \sum_{i=1}^{N} P_i \\
&= \frac{1}{Nn\,(n-1)} \left[\sum_{i=1}^{N} \sum_{j=1}^{k} \left({n_{ij}}^2\right) - Nn \right]
\end{aligned}
$$

$\bar{P}$ は P_i の全データの相加平均である．

$$
\bar{P_e} = \sum_{j=1}^{k} {p_j}^2
$$

$\bar{P_e}$ は p_j を 2 乗したもの全ての評価種類の相加平均である．

κ は-1 から 1 の値を取り，1 に近づくほど評価者どうしの評価値が一致していることを示す．0 は評価者間の評価に一貫性がない状態であり，-1 に近づくほど評価者どうしは全く一致しないように評価をしていることを示す．

第 3 章　関連研究

大喜利回答の自動生成に取り組んでいる研究が複数ある．竹内ら [13] は, Seq2Seq モデルの転移学習によって大喜利回答システムを作成している．機械学習による大喜利回答生成システムの実現のためには，お題と回答からなる大喜利データが大量に必要であるという課題がある．そこで，会話コーパスを学習した Seq2Seq モデルを転移学習することによって，少量の大喜利データで回答生成システムを作成している．転移学習では，最初に Twitter から取得した約 150 万件の会話コーパスを学習したのち，1931 対のお題と回答の組データを学習する．学生と教員による四段階評価の結果，最高評価が最大で 31% 占める回答が生成できたことがわかっている．

Zhong ら [14] は，LLM に創発的な回答生成を促す学習フレームワーク CLoT(Creative Leap-of-Thought) を提案している．大喜利などの創造的タスクでは飛躍的思考で必要であるが，CoT のような従来の手法では LLM の飛躍的思考が十分に発揮されないという課題がある．そこで人間の飛躍的思考に着想を得たフレームワークである CLoT によって，創造的タスクにおける LLM のタスク精度向上を目指す．CLoT では二段階の学習を行う．一段階目では，ヒントとして人間の回答に含まれる単語を与えられた状態で回答生成タスクと回答識別タスクを学習する．一段階目の学習で，LLM は連想能力を獲得する．二段階目では，LLM が生成した回答と人間の回答のどちらがより面白いかを LLM 自身が識別する．人間の回答よりも面白いと識別された LLM の回答をデータセットに追記し，回答生成タスクと回答識別タスクを再学習する．二段階目の学習で，LLM は探索能力を獲得する．また，Zhong らは実験のために Oogiri-Go データセットを構築している．ユーザ評価の結果，CLoT で学習した LLM が生成した回答は，ベースラインと比較して面白い傾向にあることがわかっている．

株式会社わたしはは，大喜利 LLM* を公開している．Meta 社の LLM である Llama2-13b に対して大喜利データを用いて指示学習することによって，大喜利 LLM を作成している．使用した大喜利データは，自社で収集した約 693 万件のお題と回答の組データである．大喜利番組の人気投稿者による評価の結果，生成した

*https://huggingface.co/watashiha/Watashiha-Llama-2-13B-Ogiri-sft

回答のうち約 16.3% に一定の面白みがあることがわかっている．

また，中村ら [15] は大喜利の回答における面白さを構成する要素を調査している．従来におけるテキストの機械的特徴量は，面白さを表しているとはいえないという課題がある．そこで面白さを相対的にわかりやすい構成要素に分解することによって，面白さを表す機械的特徴の作成における指針を示している．面白さを構成する特徴量として新しさ，わかりやすさ，お題との関係性に着目し，クラウドソーシングを用いて回答の特徴量を収集した．収集した特徴量を用いて回答の面白さ回帰モデルを作成したところ，各特徴量が面白さをある程度説明していることがわかった．また，各特徴量を表す機械的特徴量として Embedding，平均単語難易度，LDA を用いて，回答の面白さ自動推定を試みている．Embedding はお題との関係性に対する機械的特徴量であり，お題と回答における単語ベクトルの平均値どうしのコサイン類似度である．平均単語難易度とはわかりやすさに対する機械的特徴量であり，クラウドソーシングによって構築されたデータベースから抽出する．LDA は新しさに対する機械的特徴量であり，お題と回答どうしのトピックベクトル類似度である．機械的特徴量を用いた面白さ回帰モデルは，ベースラインよりも悪い予測精度を示した．

第 4 章　提案手法

本章では，提案手法について説明する．本研究では，大喜利のお題となる文を入力すると回答を出力するシステムを作成する．Chain-of-Thought[9] による手法と，Tree-of-Thoughts[11] を参考にした手法の二つを提案する．

4.1　Chain-of-Thought

CoT(Chain-of-Thought; 思考の連鎖) による手法では，創作具体例を LLM に入力することによって回答を生成する．我々は，大喜利には経験則が存在すると考えた．経験則とは，大喜利の回答を導き出すための発想や思考の方針である．我々は，経験則を参考にすることによって人間は面白い回答を作成できると考えた．なぜなら，経験則には回答に面白さに寄与する要素を入れ込む働きがあると考えたためである．そこで，LLM に経験則を参考にした回答生成を促すことで，面白い回答の生成を目指す．しかし我々は，LLM に経験則を直接入力することは難しいと考えた．なぜなら，経験則には言語化できないニュアンスが含まれるためである．よって，経験則を条件や規則として LLM に入力しても，LLM は経験則を参考にした回答生成ができないと考えた．そこで我々は，経験則を創作具体例として具体化する．創作具体例には，大喜利のお題，経験則を参考にお題を導く手順および手順によって導かれた回答が含まれる．創作具体例を LLM に入力することによって，LLM に創作具体例と同じ手順に従った回答生成を促し，経験則が反映された面白い回答を得る．

4.1.1　創作具体例

創作具体例は，CoT 形式のプロンプトである．CoT とは，従来の Few-Shot プロンプトで LLM に例示していた入力と出力に加えて入力から出力を得るまでの思考を例示することによって，タスクにおける精度向上を目指すテクニックである．創作具体例は，入力がお題，出力が回答，思考が経験則を参考にして回答を作成する手順である CoT 形式のプロンプトであるといえる．例えば，「こんな○○はいや

だ」というお題に対する「良いことを悪いことになるように変更する」という経験則があるとする．この経験則は，最初にお題に関係する良いことを考え，次に考えた良いことを踏まえて悪いことを考えることによって回答を作成する．この経験則を創作具体例として LLM に入力すると，LLM は最初にお題に関係する良いことを出力し，出力した良いことを踏まえた悪いことを最終的な回答として出力する．

我々は大喜利の解説記事*を参考にして，「こんな○○はいやだ」形式のお題に対する五つの経験則を考えた．以降，それぞれの経験則を経験則 1〜5 と呼ぶ．

1. **あるあるの変更**: あるあるとは，人やものや出来事に関係する性質や言動の説明のうち，多くの人が経験をしていたり共感ができる事象である．この経験則では，お題に含まれる単語のあるあるに対して，表現の言い換えや単語の追加などを行う．例えば学校のあるあるが「校庭に犬が入ってくる」であるとすると，単語を追加した「校庭に国家の犬が入ってくる」を回答とする．
2. **正の性質から負の性質**: お題に関連するポジティブな性質や出来事を説明する文章に対して，表現の言い換えや単語の追加を行いネガティブになるようにする．例えば，学校のポジティブな性質が「給食が美味しい」であるとき，この性質に単語を追加して，「給食が美味しいが水が別売りで 800 円」とネガティブな内容にする．
3. **五感からの発想**: お題に関係する単語を，聴覚，嗅覚，味覚，触覚，視覚の側面から列挙する．列挙した単語から一つ選択し，選んだ要素に着目してお題に対するネガティブなことを言う．例えば，聴覚面で学校に関係する単語として「チャイム」であると，この単語に対して着目してネガティブなこと言った「チャイムの音が黒板を引っ掻く音に似ている」を回答とする．
4. **要素の発想**: お題に関係する単語を，目につくもの，場所，時間，季節，人物の側面から列挙する．列挙した単語から一つ選び，選んだ要素に着目してお題に対するネガティブなことを言う．例えば，人物という側面で学校に関係する単語が「校長先生」であると，この単語に着目してネガティブなことを言った「児童が育てたミニトマトを校長先生が勝手に食べている」を回答とする．

*https://note.com/akaminesouri/n/n275ba94ed6bf

お題
こんな学校はいやだ
思考
学校で起きる良いこととして、「給食が揚げパン」が考えられます。
この良いことが悪いことになるように変更すると、「給食の揚げパンの粉が別売り」となります。
回答
揚げパンの粉が別売り

図 4.1　回答作成事例の例

5. **当たり前**: お題に含まれる単語に対して，同じ性質を持つ別の単語にも共通する普遍的な悪いことを言う．回答は，できるだけ単語数が少ない簡潔な内容にする．例えば，学校と同じ「施設」という性質を持つ別の単語にいえる悪いことである「家から遠い」を回答とする．この回答は，単語数が少なく，内容も簡潔である．

創作具体例はいくつかの回答作成事例で構成される．回答作成事例は，与えられたお題とは別のお題，そのお題に対する回答を経験則に従って作成する過程，及び回答で構成される三つのテキストの組である．「こんな学校はいやだ」というお題に対する「正の性質から負の性質」という経験則を表す回答作成事例を図 4.1 に示す．我々は，創作具体例を与えることによって LLM に経験則に従った回答生成をさせる．我々は，回答作成事例が CoT における Few-Shot 例になると考えた．なぜなら，回答作成事例のお題と回答の作成過程と回答が，CoT における入力と思考と出力になると考えたためである．

今回我々は，二つの回答作成事例で構成される創作具体例を作成した．なぜなら，CoT では Few-Shot 数が多いほど出力の精度が上がる傾向があることが報告されている [9] ためである．よって，創作具体例における回答作成事例の数を増やすことによって，より経験則が反映された回答の生成を目指す．我々は各創作具体例ごとに，お題が「学校」と「遊園地」である場合の二つの回答作成事例を作成した．

4.1.2 プロンプト

CoT による手法において LLM に入力するプロンプトは，三つの要素で構成される．一つ目は，指示文である．指示文には，与えられる文章が大喜利のお題であることの説明と，お題に対して面白く回答するように指示する内容が含まれる．二つ目は，創作具体例である．三つ目は，入力となるお題である．

4.1.3 出力

CoT による手法において LLM が出力するテキストは，二つの要素で構成される．一つ目は，思考である．思考には，プロンプトで与えた創作具体例を参考にして回答を作成する過程が含まれる．この過程には，与えた創作具体例のもとになった経験則が反映されている．二つ目は，回答である．この回答は，思考に含まれる過程を経て導かれている．思考の過程には経験則が反映されているため，その過程を経て導かれた回答にも創作具体例のもとになった経験則が反映されているといえる．よって，創作具体例を入力することによって，経験則が反映された回答を生成できるといえる．

4.2 Tree-of-Thoughts

ToT を参考にした手法では，探索木において回答作成手順の適用と LLM による自動評価繰り返すことによって，段階的に洗練された回答を生成する．本手法の概要図を図 4.2 に示す．回答作成手順とは，大喜利における途中回答から次の回答を導くための手順である．我々は，人間は別のお題における回答の導き方を複数個組み合わせることによって，回答を作成していると考えた．そこで，人間の高評価回答から回答作成手順を抽出する．回答作成手順を入力することによって，人間による別のお題における回答作成を参考にしながら途中回答を生成できると考えた．しかし，お題と回答作成手順には相性があり，全ての回答作成手順を入力しても面白い回答を作成できるとは限らないと考えた．そこで，類似するお題どうしは同じ回答作成手順が有効であると仮定し，対象のお題と類似するお題にける回答作成手順を適用する．また，LLM による自動評価を行い，面白い途中回答のみをもとに次

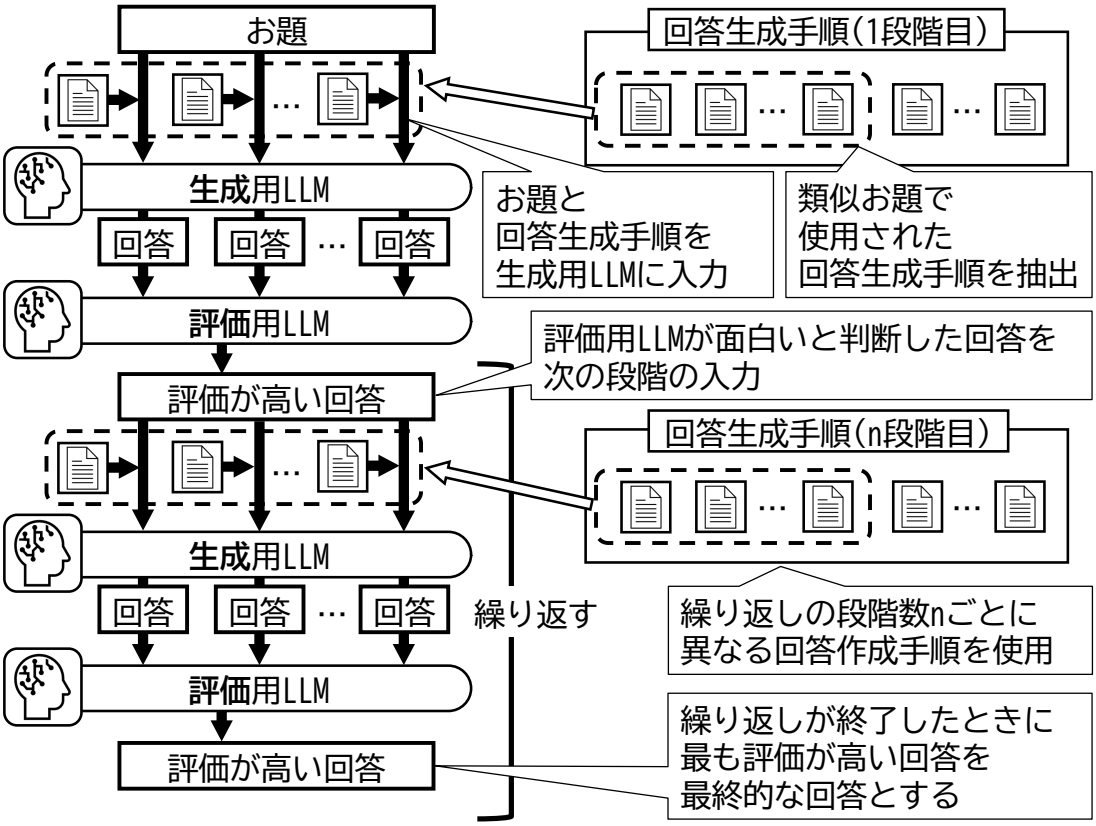

図 4.2　ToT を参考にした手法の概要図

段階の途中回答を生成することによって，段階的に洗練された回答を生成する．

本手法と ToT の異なる点は，途中回答の生成における LLM への入力である．ToT では，同一のプロンプトを LLM に入力することによって各段階における全ての途中回答を生成する．本手法では，探索木の各段階において，複数種類の回答作成手順を LLM に入力することによって途中回答を生成するという点が，ToT とは異なる．探索木では，LLM のサンプリングによる文章生成ではなく，回答作成手順をもとに途中回答のバリエーションを担保する．そのため，探索木の各段階で LLM に大喜利における回答の導き方を網羅的に考慮した途中回答の生成を促すことができる点が，ToT と比較した本手法の利点である．また，ToT における探索木では深さ優先探索と幅優先探索のどちらも可能であるが，本手法では幅優先探索を行う．

ToT を参考にした手法によって，対象のお題 $\hat{q}$ に対する回答 $\hat{r}$ を生成する手順を以下に示す．探索木の各ノードは途中回答を表し，深さ d のノードは d 個の回答作成手順を適用した結果である．各ノードからは，適用順が $d+1$ である回答作成手順に対応する子ノードを生成する．本手法では，大喜利データと回答作成手順を用いて回答を生成する．大喜利データ $\{Q, R\}$ は，お題と回答からなる組 N 個の集合である．お題の集合を Q，回答の集合を R とする．お題 $q \in Q$ と，q 組みを成す回答 $r \in R$ を LLM に入力して生成した回答作成手順列を $p_{q,r} = \{p_{q,r_j} | j = 1, 2, \ldots, M_{q,r}\}$ とする．ただし，$M_{q,r}$ は $p_{q,r}$ の長さである．一つの段階で生成する回答の件数を b，作成する探索木の深さを D とする．

Step 1 $\hat{q}$ と Q に含まれるすべてのお題との文章分散表現の類似度を計算する．$\hat{q}$ と $q \in Q$ との文章分散表現の類似度を $s(q)$ とする．$s(q)$ が大きい上位 b 個のお題の集合を Q_{sim} とする．

Step 2 $q_{\mathrm{sim}} \in Q_{\mathrm{sim}}$ と，大喜利データにおいて q_{sim} と組みを成す回答 r_{sim} を LLM に入力して生成された回答作成手順列 $p_{q_{\mathrm{sim}}, r_{\mathrm{sim}}}$ を抽出する．$p_{q_{\mathrm{sim}}, r_{\mathrm{sim}}}$ の集合を P_{sim} とする．

Step 3 段階数 $d = 1$ にする．LLM を用いて，$\hat{q}$ における主題を抽出する．入力 r_{input} を抽出した主題にする．

Step 4 $p_{\mathrm{sim}} \in P_{\mathrm{sim}}$ における d 番目の回答作成手順 $p_{\mathrm{sim}d}$ と r_{input} を LLM に入力して回答 $r_{p_{\mathrm{sim}d}}$ を生成する．$r_{p_{\mathrm{sim}d}}$ の集合を R_d とする．

Step 5 LLM を用いて，$r_d \in R_d$ に対する面白さ評価値 $e(r_d)$ を算出する．$e(r_d)$ が最も大きい r_d を $r_{\mathrm{best}d}$ とする．

Step 6 d が D と等しいならば，$\hat{r} = r_{\mathrm{best}d}$ とする．
そうでなければ，$r_{\mathrm{input}} = r_{\mathrm{best}d}$ とし，$d = d + 1$ としたのち，Step 4 に戻る．

4.2.1 回答作成手順

回答作成手順について説明する．回答作成手順とは大喜利において，ある途中の回答から別の回答を導くまでの考え方である．我々は，お題に対して回答作成手順

お題「ナルシスト裁判長の特徴を教えてください」
回答「入場曲がある」

回答作成手順1

タイトル	誇張表現
入力	裁判長
思考	(前略)裁判長は一般的に厳粛な雰囲気の中で入場します。 ナルシスト裁判長の場合、この入場の際に、 自己顕示欲を満たすような派手な演出を行うと考えられます。
出力	派手な演出で法廷に登場する

回答作成手順2

タイトル	具体化
入力	派手な演出で法廷に登場する
思考	派手な演出として、音楽を用いることは一般的です。 特に、ナルシスト裁判長の場合、(中略)自分のために 作曲された入場曲を使用することが考えられます。
出力	自分のための入場曲がある

回答作成手順3

タイトル	言い換え
入力	自分のための入場曲がある
思考	(前略)より簡潔で口語的な表現に言い換えます。
出力	入場曲がある

図 4.3　回答作成手順列の例

をいくつか適用することによって，人間は回答を作成していると考えた．図 4.3 に，お題「ナルシスト裁判長の特徴を教えてください」から回答「入場曲がある」を導くまでに適用される回答作成手順列を例示する．この例では三つの回答作成手順を

順番に適用することによって，お題から回答を導くことができるといえる．

回答作成手順は，五つの要素で構成される．一つ目はタイトルである．タイトルは，回答作成手順の内容を端的に表している．二つ目は入力である．入力は，回答作成手順を適用する前段階の回答である．三つ目は思考である．思考は，入力から出力を導き出すまでの思考の過程を表す文である．四つ目は出力である．出力は，入力に対して回答作成手順を適用したときに得られる回答である．

4.2.2 事前準備

大喜利データの用意

大喜利データとは，大喜利のお題とそのお題に対する回答からなる組データの集合である．我々は大喜利のお題と回答の組データを，大喜利投稿 Web サイト[†] [‡] [§] から手動または自動で取得した．大喜利投稿 Web サイトから取得した回答には，人間による評価が付与されている．面白い回答の生成ができる回答作成手順を生成するために，各お題において最も評価が高い回答のみを使用した．つまり，大喜利データにおけるお題には，ただ一つの回答のみが対応している．

人間の高評価回答を元にした回答作成手順の生成

お題と回答の組データを LLM に入力して，回答作成手順列を生成させる．入力したお題と回答の組データを (q_i, r_i)，生成した回答作成手順列を $P_i = \{p_{i,j} | j = 1, 2, \ldots, M_i\}$ とする．回答作成手順列には，複数の回答作成手順が順番に含まれる．ただし，$p_{i,j}$ は P_i における適用順が j 番目の回答作成手順，M_i は P_i の長さである．q_i に対して $p_{i,j} \in P_i$ を適用順 j と同じ順番でに適用することによって，r_i を導くことができるといえる．

[†] https://chinsukoustudy.com/
[‡] https://oogiri-chaya.com/
[§] https://watage.site/asobiba/

回答作成手順のグループ分け

生成した回答作成手順は，適用順とタイトルによってグループ分けを行う．最初に，適用順によってグループ分けを行う．我々は，適用順が同じ回答作成手順は同じ役割を持つと仮定した．大喜利における回答の導き方は，最初に回答の大まかな内容を決定したのち，回答の細部を調整するという一筋の構造になっていると考えた．したがって，回答作成手順列における最初の回答作成手順は，大まかな内容を決定する役割があるといえる．逆に，回答作成手順列において適用順が後である回答作成手順は，大まかな内容がすでに決定されている途中の回答に対して細部を調整する役割があるといえる．この仮定は，大喜利のように回答が段階的に洗練されるタスクでは自然であると考えられる．しかし，回答の導き方が発想を並列に生成して比較・統合するような構造となっているタスクでは，この仮定は成り立たない可能性がある．その場合，適用順と役割の対応が崩れるため，探索木における各段階の意味付けが曖昧になることが考えられる．本研究でこの仮定は未検証であるため，今後検証しなければならない．

次に，タイトルによるグループ分けを行う．タイトルが完全一致またはタイトルが他方のタイトルに含まれる回答作成手順を，同じ種類の回答作成手順として扱う．これによって，回答作成手順の重複を防ぐ．我々はタイトルが同じ回答作成手順は，内容も同じであると仮定した．大喜利における回答の導き方はいくつかの単純な手順に分割できると考えた．また，他のお題における手順を適用することで，面白い回答を導けることがあると考えた．そのため，手順の内容を端的に表すタイトルによって，回答作成手順をグループ分けできるといえる．しかし，回答の導き方を単純な手順に分割できないタスク，他の入力において回答を導く手順を再利用できないタスクでは，この仮定は成り立たないといえる．その場合，入力に対して回答を導ける手順が探索木に存在しなくなることが考えられる．本研究でこの仮定は未検証であるため，今後検証しなければならない．

4.2.3　文章分散表現に基づいたお題どうしの類似度計算

対象となるお題と類似するお題で使用されている回答作成手順を用いて，回答を生成する．お題とお題の類似度を，お題の文章分散表現のコサイン類似度で計算す

る．文章分散表現は文の意味合いが反映されたベクトルである．そのため我々は，主題や構造が類似するお題どうしの文章分散表現は類似すると考えた．また，今回は文章分散表現の類似度計算にコサイン類似度を用いた．文章分散表現の類似度計算にコサイン類似度を用いるのは一般的ではあるため，本研究では暫定的に使用している．そのため，コサイン類似度を用いる必然性について今後検証する必要がある．

4.2.4　回答作成手順を適用した回答生成

我々は，類似するお題どうしは，同じ回答作成手順を適用することで面白い回答を生成できると仮定した．我々は，大喜利では異なるお題どうしであっても，回答を導くまでの考え方に共通点が存在する場合があると考えた．類似するお題において面白い回答を導くことができた考え方を参考にして，人間は回答を作成しているといえる．そのため，主題や形式が類似する別のお題における回答作成手順を適用することによって，面白い回答を生成できるといえる．しかし，回答を導く考え方が他の入力では通用しないタスクでは，この仮定が成り立つとはいえない．その場合，他の入力における回答の導き方を参考に途中回答を展開する探索木では，最終的に良い回答を生成することができないと考えられる．本稿では，5.2 においてこの仮定を検証する．

類似お題集合に含まれるお題を元に生成された回答作成手順列のうち，適用順が d である回答作成手順を適用して，回答を生成する．我々は，回答生成の段階 d が小さいときは回答の全体を決定するような回答作成手順を適用すべきであると考えた．逆に，d が大きくなるにつれて回答の細部を調整するような回答作成手順を適用すべきであると考えた．そのため，回答生成の段階 d と回答作成手順の適用順を一致させる必要があると仮定した．

また，回答作成手順を適用した回答生成では，LLM に入力するプロンプトにはタイトルが同じ回答作成手順をともに記載する．我々は，同じタイトルを持つ回答作成手順どうしは内容も同じであると考えた．なぜなら，回答作成手順の生成において，タイトルは回答作成手順の内容を端的にあらわす文章にするように LLM に例示しているためである．回答作成手順には，入力と出力および入力から出力を導

き出すまでの過程である思考が含まれる．そこで，同じタイトルを持つ回答作成手順をプロンプトに記載する．これによって，共通の思考に沿って入力から出力を導く Few-Shot の CoT[9] になるため．回答作成手順の内容の通りに出力を生成できると考えた．

4.2.5 LLM による回答自動評価

LLM による回答自動評価について説明する．LLM の自動評価は，以下の三種類の方法を提案する．

1. 面白さの絶対評価
   LLM にお題とお題に対する回答一つを入力し，回答が面白いかどうかを二値評価させる．面白い場合は評価の値を 1，面白くない場合は評価の値を 0 とする．
2. 面白さの相対評価
   LLM にお題とその段階の回答を 5 個ずつ入力し，最も面白い回答を選択させる．選択した回答をさらに 5 個ずつ入力し，もう一度最も面白い回答を選択させる．これを繰り返すことによって，最も面白い回答を一つ選択する．
3. 面白さに寄与する要素の有無による評価
   LLM にお題とお題に対する回答一つを入力し，面白さに寄与する要素があるかどうかを二値評価させる．我々が考えた七つの面白さに寄与する要素を用いて，自動評価を行う．用いる面白さに寄与する要素は，明確な関連性，意外な関連性，文法，適切性，意図，知識，他の要素である．要素それぞれに対する評価値を，ある場合を 1，ない場合を 0 とする．要素ごとの評価値の和を，回答の評価とする．

我々は，LLM は回答の面白さを評価できると仮定した．大喜利の回答評価では，個人の主観に加えて，客観的な観点が存在すると考えた．LLM は大量のテキストデータを学習しているため，モデル内部に大喜利やお笑いに関する情報が含まれている可能性がある．このとき，大喜利における客観的な評価観点に基づいて回答を評価できると考えた．しかし，回答の評価が完全に個人の主観によるタスクでは，

この仮定が成り立つとは限らない．その場合，探索木において高評価回答をもとに次段階の回答を生成できなくなり，段階的に洗練された回答を生成することができなくなる．本稿では，第 5 章でこの仮定を検証する．

第 5 章　評価実験

本章では，第 4 章で導入した仮定について検証を行う．節 5.1 では，CoT による手法で導入した仮定について検証を行う．CoT による手法における仮定は次の二つである．

仮定 1-1 創作具体例を LLM に入力することによって既存手法よりも面白い回答を生成できる．

仮定 1-2 創作具体例を入力するモデルのパラメータ数の大きいほど，面白い回答を生成できる．

実験で我々は，創作具体例を LLM に入力することによって回答を生成する．5 種類の創作具体例とパラメータ数が異なる二つの LLM を組み合わせて，回答を生成する．また，既存手法である通常のプロンプトおよび大喜利 LLM を用いて回答を生成する．生成した回答の面白さおよび 7 項目の面白さに寄与する要素に対して，ユーザ評価を行う．生成した回答の評価を比較することによって，仮定を検証する．創作具体例を入力して生成した回答と，通常のプロンプトを入力して生成した回答および大喜利 LLM で生成した回答の評価値を比較することによって，仮定 1-1 を検証する．また，同じ創作具体例をパラメータ数が異なる LLM に入力して生成した回答の評価値を比較することによって仮定 1-2 を検証する．

節 5.2 では，ToT を参考にした手法で導入した仮定について検証を行う．ToT を参考にした手法における仮定は次の三つである．

仮定 2-1 LLM に回答作成手順を入力して段階的に回答を生成することによって，既存手法よりも面白い回答を生成できる．

仮定 2-2 対象となるお題と回答作成手順のもとになったお題の文章分散表現の類似度によって，対象となるお題に対して面白い回答を生成できる回答作成手順を抽出できる．

仮定 2-3 LLM による自動評価によって，面白い回答を抽出できる．

実験で我々は，複数の回答作成手順を LLM に入力することによって，途中回答を生成する．途中の回答に対する面白さ評価を，ユーザ評価によって行う．ユーザ評

価による面白さ評価が最も高い途中回答と，次段階の回答作成手順を LLM に入力することによって，次段階の途中回答を生成する．これを 3 段階分繰り返す．各段階において出力された回答と，既存手法で生成した回答の面白さを比較することによって，仮定 2-1 を検証する．各段階において，対象のお題の類似度に基づいて抽出した回答作成手順を適用して生成した回答に対する面白さ評価の期待値を確認することによって，仮定 2-2 を検証する．各段階において LLM の自動評価に基づいて抽出した回答に対する面白さ評価の期待値を確認することによって，仮定 2-3 を検証する．

5.1 Chain-of-Thought

この実験では，創作具体例を入力することによって，面白い回答が生成できるかどうかを確認する．この実験における検証事項は二つある．一つ目は，創作具体例によって既存手法よりも面白い回答が生成できるかどうかである．創作具体例を入力して生成した回答と既存手法によって生成した回答の面白さを比較することによって検証する．既存手法には，LLM による通常の生成と大喜利 LLM の二種類を使用する．LLM による通常の生成では，プロンプトで大喜利回答タスクの指示とお題の入力のみを行って，回答を生成する．これは，LLM を用いた最も単純な大喜利の回答生成手法である．大喜利 LLM では，大喜利データを使用した教師あり学習によって訓練されたモデルにお題を入力することによって回答を生成する．ニューラルネットワークを大喜利のお題と回答の組データで訓練するのは，大喜利の自動生成に関する研究における主流のアプローチ [13][14] *である．提案手法で生成した回答が，既存手法で生成した回答よりも面白いかどうかを検証するために t 検定を実施する．また，回答の面白さに対する Kappa 係数を計算することによって，評価の傾向を手法ごとに分析する．

二つ目は，回答の面白さが使用するモデルのパラメータ数に依存するどうかである．CoT では，使用するモデルのパラメータが大きいほど，タスクの精度が向上する傾向にあることがわかっている [9]．そのため，創作具体例の入力による大喜利の回答生成にも同様の傾向があるかどうかを確認する．

*https://huggingface.co/watashiha/Watashiha-Llama-2-13B-Ogiri-sft

同一の創作具体例をパラメータ数の異なるモデルに入力したとき，パラメータ数の大きいモデルが生成した回答が小さいモデルが生成した回答よりも面白いかどうかを検証するために t 検定を実施する．

5.1.1 実験設定

使用 LLM

本研究では，三つのモデルを使用している．llama-3-youko-70b-instruct[†]と llama-3-youko-8b-instruct[‡]は，提案手法において創作具体例を入力する LLM 及び既存手法における通常の LLM として使用した．これは Meta 社が公開している Llama3[3] を rinna 株式会社が日本語用にファインチューニングしたモデルである．この二つのモデルは学習データが同じであり，パラメータ数のみが異なる．そのため我々は，これらのモデルを使用することで，回答の面白さがモデルのパラメータ数に依存するかどうかを確認できると考えた．以降はこれらのモデルをそれぞれ 70B モデル，8B モデルと呼ぶ．もう一つの使用するモデルは，Watashiha-Llama-2-13B-Ogiri-sft[§]である．このモデルは，株式会社わたしはが公開している大喜利 LLM である．Meta 社が公開している Llama2-13B[16] を，株式会社わたしはが収集した約 693 万件の大喜利データを使用して訓練することによって作成した．このモデルは，大喜利データを学習した LLM であるため，大喜利の自動生成に関する研究における主流のアプローチであるといえる．そのため，既存手法の一つとして用いた．以降はこのモデルを SFT モデルと呼ぶ．

いずれのモデルを使用する場合も，サンプリングを有効にして回答を生成した．システムプロンプト及びプロンプトのフォーマットは，各モデルの公開ページから引用している．生成時の設定は，Temperature が 0.8，Top-P が 0.9，Top-K が 50 であり, その他の設定は各モデルの既定値及び Python の LLM ライブラリである transformers の既定値を使用している．いずれのモデルも使用する際にパラメータの量子化は行っていない．

[†]https://huggingface.co/rinna/llama-3-youko-70b-instruct
[‡]https://huggingface.co/rinna/llama-3-youko-8b-instruct
[§]https://huggingface.co/watashiha/Watashiha-Llama-2-13B-Ogiri-sft

評価項目

この実験では，一つの回答に対して以下の八項目において評価を行う．各評価項目は「はい」または「いいえ」の二値で評価する．知識に関しては評価者が判断できない場合を考慮して「わからない」を加えて評価を行い，「いいえ」として集計した．

1. **面白さ**: 回答が面白いかどうかを評価者の主観で評価する．大喜利とは与えられたお題に対して面白みのある回答をする演芸であるため，より面白い回答を生成できる手法がより良い手法であるといえる．
2. **明確な関連性**: 明確な関連性がある単語が，回答に一つ以上含まれるかどうかを評価する．我々は関連性のある単語を，お題と双方向の連想ができる単語であると定義した．双方向の連想とは，お題を提示されたときにその単語が連想できるかつ，その単語を提示されたときに元の単語が連想できることである．我々は明確な関連性がある単語を，関連性があるかつ，評価者がそのお題で大喜利をするときに自力で思いつくことができる単語であると定義した．我々は明確な関連性がある回答は共感を得られやすいと考え，面白さに寄与する要素になると考えた．
3. **意外な関連性**: 意外な関連性がある単語が，回答に一つ以上含まれるかどうかを評価する．我々は意外な関連性がある単語を，関連性があるかつ，評価者がそのお題で大喜利をするときに自力で思いつくことができない単語であると定義した．関連性がある単語の定義は，明確な関連性のときと同様である．我々は，意外な関連性がある回答ではお題の深掘りができていると考えたため，面白さに寄与する要素になると考えた．
4. **文法**: 回答の文章が文法的に正しいかどうかを評価する．我々は，回答の文章が文法的に正しいことは受け手に回答を理解させるために必要であるため，面白さに寄与する要素であると考えた．
5. **適切性**: 回答がお題に合致しているかどうかを評価する．例えば，「こんな学校はいやだ」というお題に対する「先生がみんな優しい」という回答があるとする．お題はいやなことを聞いているが，回答は良いことを言っているため，この回答は適切性がないと判断する．我々は，回答がお題に合致していることは大喜利が成立するための必要最低限の条件であるため，面白さに寄

与する要素であると考えた.

6. **意図**: 回答者が回答で伝えたい意図が，受け手に伝わっているかどうかを評価する．我々は回答に含まれる意図が評価者に伝わることによって，回答者がどのように面白みを持たせようとしたかがわかると考えたため，面白さに寄与する要素であると考えた.
7. **知識**: 回答を理解する上で特定の知識が必要かどうかを評価する．特定の知識とは，学問で出現する専門用語や，漫画に登場する人物の名前など，理解するために特定の集団への所属やコンテンツの視聴などが必要な事象である．我々は回答の理解に特定の知識が必要になることによって回答が内輪ネタとなるため，面白さに寄与すると考えた.
8. **他の要素**: 回答に，他の要素が含まれるかどうかを評価する．他の要素とは，お題に関係しない要素に関係する表現である．例えば，「こんな学校はいやだ.」というお題に対する「遊具で遊ぶのにファストパスが必要」という回答があるとする．この回答には，お題には関係しない遊園地に関係する「ファストパスが必要」という表現が含まれているため，他の要素があると判断する．我々は，回答に他の要素が含まれることによって評価者がお題から想像しづらい内容になるため，面白さに寄与すると考えた.

5.1.2 実験手順

実験は以下の手順で行う.

Step 1 創作具体例 1〜5 それぞれを 70B モデルと 8B モデルそれぞれに入力して，回答を生成する.

Step 2 Step 1 にて生成した回答に対して，5.1.1 節で示した 8 項目ごとに，評価者による評価を行う.

Step 3 Step 2 にて付与した評価値に対して，創作具体例と既存手法ごと，及び使用モデルごとに検定を行う.

実験では五つの「こんな○○はいやだ」形式のお題を用いる．○○として，スーパーマーケット，レストラン，動物園，神社，駅を選択した．これらのお題は，各

創作具体例に含まれるお題とは異なる．以降はそれぞれのお題をお題 1〜5 とする．

Step 1 では，一つのお題に対して 13 種類の方法で回答を生成する．5 種類の創作具体例を 70B モデルと 8B モデルのそれぞれに入力する．創作具体例が含まれないベースプロンプトを 70B モデルと 8B モデルのそれぞれに入力する．ベースプロンプトを LLM に入力することによって生成した回答を，LLM による通常の生成による回答としてベースラインの一つにする．また，ベースプロンプトを SFT モデルに入力する．SFT モデルで生成した回答を，大喜利 LLM で生成した回答としてベースラインの一つにする．それぞれの方法において，五つのお題それぞれに五つの回答を生成する．

Step 2 にて回答の評価を行う評価者は，鈴木研究室の学生 8 名である．評価者ごとに回答を提示する順番をシャッフルすることによって，順序効果を排除する．各評価項目の評価値は，その項目ごとに「はい」と答えた評価者の割合である．

Step 3 では，t 検定を実施する．両側検定を行う．有意水準を 0.05 に設定した．また，等分散性を仮定している．本実験は創作具体例とモデルとの複数の組み合わせで検証を行う多重比較である．そのため，各検定では比較回数に応じた Bonferroni 補正を行う．

また，評価値の平均に対する検定に加え，使用モデル，創作具体例ごとに Kappa 係数を計算する．Kappa 係数とは，評価者間の評価の一致度を表す指標である．この実験では評価者が 3 人以上いるため，Fleiss の Kappa 係数を計算する．Kappa 係数は-1 から 1 の値を取り，1 に近づくほど評価者間の評価が一致していることを示す．0 は評価者間の評価に一貫性がない状態であり，-1 に近づくほど評価者どうしは全く一致しないように評価をしていることを示す．

5.1.3 結果・考察

生成した回答に対する評価値の平均を，表 5.1 に示す．全てのモデルと創作具体例の組み合わせにおいて，文法と適切性に対する評価値の平均が 0.8 を上回った．また，70B と創作具体例 1 の組み合わせ以外は，意図に対する評価値の平均が 0.7 を上回った．我々は，適切性と文法および意図は，全ての組み合わせにおいて高水準であったと考える．適切性と文法は，大喜利の回答として最低限の基準を満たす

かどうかを表す評価項目であるといえる．また，意図は回答の内容が伝わるかどうかを表す評価項目であるといえる．よって，全てのモデルと創作具体例の組み合わせで，受け手に内容が伝わる大喜利の回答を生成することが可能であると言える．

また，全てのモデルと創作具体例の組み合わせにおいて，知識に対する評価値の平均が 0.1 を下回った．我々は，知識は全ての組み合わせにおいて低水準であったと考える．そのため，創作具体例の入力および既存手法では知識が含まれる回答を生成することが困難であるといえる．我々は，創作具体例の改善や RAG[17] による外部情報の参照を取り入れた回答生成が，知識が含まれる回答の生成に必要であると考える．

既存手法との比較

創作具体例を入力して生成した回答と，通常生成による回答の各評価項目における t 検定の p 値を表 5.2，Cohen の d を表 5.3 に示す． モデルと創作具体例

表 5.1　生成した回答に対する評価値の平均

モデル	創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
70B	1	0.045	0.940	0.125	0.845	0.915	0.660	0.035	0.330
	2	0.060	0.900	0.140	0.875	0.880	0.820	0.005	0.155
	3	0.055	0.615	0.160	0.830	0.950	0.745	0.055	0.280
	4	0.085	0.790	0.250	0.930	0.915	0.890	0.025	0.240
	5	0.010	0.705	0.095	0.965	0.965	0.945	0.005	0.055
	なし	0.165	0.975	0.195	0.925	0.880	0.835	0.070	0.390
8B	1	0.085	0.860	0.110	0.870	0.885	0.850	0.020	0.290
	2	0.035	0.900	0.160	0.865	0.895	0.800	0.035	0.035
	3	0.025	0.325	0.120	0.785	0.950	0.795	0.010	0.235
	4	0.025	0.725	0.175	0.870	0.855	0.855	0.015	0.180
	5	0.005	0.665	0.140	0.980	0.990	0.970	0.005	0.020
	なし	0.055	0.925	0.200	0.940	0.920	0.835	0.020	0.240
SFT	なし	0.090	0.735	0.175	0.775	0.890	0.705	0.045	0.180
全体		0.057	0.774	0.157	0.881	0.915	0.823	0.027	0.202

の組み合わせ 10 個を 70B モデルまたは 8B モデルの通常生成と比較するため，Bonferroni 補正は 10 である．実験の結果，創作具体例 1,2,3,5 を 70B モデルに入力して生成した回答は，70B モデルの通常生成による回答よりも面白さ評価値が有意に低かった．また，その他の創作具体例とモデルの組み合わせで生成した回答も，通常生成による回答よりも面白さ評価値が低い結果となった．結果から，創作

表 5.2　通常生成との検定における p 値

モデル	創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
70B	1	0.0003	0.1700	0.0734	0.1172	0.3292	0.0125	0.1947	0.4323
	2	0.0031	0.0211	0.2000	0.3505	1.0000	0.7958	0.0020	0.0010
	3	0.0025	0.0000	0.3850	0.1008	0.0372	0.1676	0.5539	0.1119
	4	0.0235	0.0050	0.3223	0.9173	0.3481	0.3180	0.0553	0.0504
	5	0.0000	0.0017	0.0153	0.4438	0.0458	0.0394	0.0020	0.0000
8B	1	0.3294	0.2147	0.0148	0.1203	0.4280	0.8309	1.0000	0.5457
	2	0.3989	0.4918	0.3360	0.1071	0.5640	0.5985	0.4715	0.0004
	3	0.1741	0.0000	0.0917	0.0108	0.4910	0.5754	0.5213	0.9398
	4	0.1967	0.0015	0.5552	0.1574	0.1787	0.7755	0.7580	0.3660
	5	0.0151	0.0014	0.1741	0.1562	0.0120	0.0186	0.2481	0.0003

表 5.3　通常生成との検定における Cohen の d

モデル	創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
70B	1	-1.090	-0.394	-0.518	-0.451	0.279	-0.734	-0.372	-0.224
	2	-0.881	-0.674	-0.368	-0.267	0.000	-0.074	-0.927	-0.995
	3	-0.903	-1.384	-0.248	-0.473	0.606	-0.396	-0.169	-0.458
	4	-0.662	-0.833	0.283	0.030	0.268	0.285	-0.555	-0.568
	5	-1.588	-0.943	-0.711	0.218	0.580	0.599	-0.927	-1.662
8B	1	0.279	-0.356	-0.715	-0.447	-0.226	0.061	0.000	0.172
	2	-0.241	-0.196	-0.275	-0.465	-0.164	-0.150	0.205	-1.069
	3	-0.390	-2.165	-0.487	-0.750	0.196	-0.160	-0.183	-0.021
	4	-0.370	-0.951	-0.168	-0.406	-0.386	0.081	-0.088	-0.258
	5	-0.713	-0.960	-0.390	0.407	0.739	0.689	-0.331	-1.108

具体例を LLM に入力しても通常生成より面白い回答を生成できないといえる．

我々は，創作具体例の入力によって面白い回答を生成できなかった理由の一つは，複数の回答生成で同一の思考過程を LLM が生成するためであると考えた．創作具体例 1 を 70B モデルに入力して生成したお題 4 に対する回答の一部を以下に示す．

回答 1-1 賽銭箱に小学生しか入れていない
回答 1-2 賽銭箱に小説しか入れられない

以上に示した回答はいずれも「賽銭箱に小銭しか入らない」という同一のあるあるが元になっている．我々は，お題 4 において創作具体例 1 が LLM に，このような面白さにつながらない同一の思考を生成させたと考えた．そのため，面白くない回答が複数生成されたと考える．我々は手法の改善には，同一の思考を生成しづらい創作具体例の作成や，重複する思考の生成を防ぐ機構の組み込みが必要であると考える．

創作具体例を入力して生成した回答および通常生成による回答と，大喜利 LLM による回答の各評価項目における t 検定の p 値を表 5.4，Cohen の d を表 5.5 に示す． モデルと創作具体例および通常生成の組み合わせ 12 個を大喜利 LLM と比較

表 5.4　大喜利 LLM との検定における p 値

モデル	創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
70B	なし	0.0606	0.0019	0.6816	0.0638	0.8523	0.1046	0.4530	0.0047
	1	0.1738	0.0078	0.2728	0.3376	0.6195	0.5904	0.7603	0.0410
	2	0.3968	0.0352	0.4748	0.1847	0.8630	0.1334	0.1501	0.6877
	3	0.3331	0.2411	0.7484	0.4786	0.2178	0.6214	0.7517	0.1194
	4	0.8887	0.5594	0.2170	0.0333	0.6266	0.0145	0.5032	0.3979
	5	0.0087	0.7796	0.0909	0.0127	0.1745	0.0015	0.1501	0.0182
8B	なし	0.3128	0.0161	0.6079	0.0253	0.5640	0.1243	0.3989	0.3979
	1	0.8930	0.1421	0.1362	0.2185	0.9315	0.0785	0.3989	0.1684
	2	0.0871	0.0329	0.7541	0.2484	0.9308	0.2254	0.7544	0.0046
	3	0.0365	0.0002	0.2981	0.9071	0.3014	0.2754	0.2284	0.3749
	4	0.0417	0.9108	1.0000	0.2344	0.5676	0.0687	0.3069	1.0000
	5	0.0050	0.4918	0.4842	0.0041	0.0360	0.0003	0.1501	0.0029

するため，Bonferroni 補正は 12 である．実験の結果，創作具体例 5 を 70B モデルと 8B モデルに入力して生成した回答は，大喜利 LLM が生成した回答よりも面白さ評価値が有意に低かった．また，その他の創作具体例とモデルの組み合わせで生成した回答も，大喜利 LLM による回答よりも面白さ評価値が低い結果となった．結果から，創作具体例を LLM に入力しても大喜利 LLM より面白い回答を生成できないといえる．

我々は，大喜利 LLM の面白さに対する評価値の平均が高い理由の一つは，多様な形式の回答を生成できるためであると考える．大喜利 LLM が生成した回答を以下に示す．

回答 1-3 御朱印に「笑」と書かれる

回答 1-4「いらっしゃいませ、一名様ですか？」「はい」(中略)「お会計の方はいかがなさいますか？」「えっ？」「ありがとうございました」

回答 1-3 と回答 1-4 は，大喜利 LLM で生成したお題 4 とお題 2 に対する回答である．回答 1-3 はこの実験で生成した回答の中で面白さに対する評価が最も高い回答である．回答 1-3 の特徴的な点は，「笑」という効果的なフレーズが含まれると

表 5.5　大喜利 LLM との検定における Cohen の d

モデル	創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
70B	なし	0.544	0.930	0.117	0.537	-0.053	0.468	0.214	0.838
	1	-0.390	0.786	-0.314	0.274	0.141	-0.153	-0.087	0.594
	2	-0.242	0.613	-0.204	0.381	-0.049	0.432	-0.414	-0.114
	3	-0.277	-0.336	-0.091	0.202	0.353	0.141	0.090	0.449
	4	-0.040	0.166	0.354	0.620	0.139	0.717	-0.191	0.241
	5	-0.773	-0.080	-0.488	0.732	0.390	0.955	-0.414	-0.692
8B	なし	-0.289	0.706	0.146	0.653	0.164	0.442	-0.241	0.241
	1	-0.038	0.422	-0.429	0.353	-0.024	0.509	-0.241	0.396
	2	-0.494	0.621	-0.089	0.330	0.025	0.347	-0.089	-0.842
	3	-0.609	-1.132	-0.298	0.033	0.296	0.312	-0.345	0.253
	4	-0.592	-0.032	0.000	0.341	-0.163	0.527	-0.292	0.000
	5	-0.833	-0.196	-0.199	0.853	0.610	1.094	-0.414	-0.889

ころである．回答 1-4 の特徴的な点は，セリフが連続する形式となっているところである．このような特徴は，提案手法で生成した回答には見られなかった．そのため我々は，大喜利 LLM は多様な形式の回答を生成できると考える．反対に，大喜利 LLM で生成した回答は，明確な関連性と適切性および意図の三つの評価項目に対する評価値が提案手法や通常の LLM と比較して低い傾向にあった．特に回答 4 は，明確な関連性と適切性および意図に対する評価値が低い．明確な関連性と適切性および意図は，回答のわかりやすさを表す評価項目であるといえる．そのため，大喜利 LLM は受け手にとってわかりづらい回答を生成する傾向にあるといえる．我々は提案手法の改善策として，創作具体例と大喜利 LLM の組み合わせによって，回答形式の多様さとわかりやすさを両立した生成手法の実現を考えている．

モデルのパラメータ数の大きさでの比較

創作具体例を 70B モデルと 8B モデルに入力して生成した回答の各評価項目における t 検定の p 値を表 5.6，Cohen の d を表 5.7 に示す． 各創作具体例および通常生成において，70B モデルと 8B モデルを比較するため，Bonferroni 補正は 6 である．実験の結果，ベースプロンプトを入力した場合，70B モデルで生成した回答が 8B モデルで生成した回答よりも面白さに対する評価値の平均が有意に高かった．創作具体例 4 を入力した場合，70B モデルと 8B モデルで回答の面白さに対する評価値の平均に比較的大きな差はあったが，有意差はなかった．また，創作具体例 2,3,5 を入力した場合も，70B モデルで生成した回答が 8B モデルで生成した回答よりも面白さに対する評価値の平均が高かったが，有意差はなかった．結果から，

表 5.6　パラメータ数の大きさでの検定における p 値

創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
なし	0.0016	0.1185	0.9060	0.7609	0.2946	1.0000	0.0313	0.0504
1	0.1670	0.1029	0.6295	0.5853	0.4770	0.0090	0.4965	0.6325
2	0.3150	1.0000	0.6309	0.8422	0.7659	0.7223	0.0962	0.0070
3	0.2178	0.0069	0.3731	0.4928	1.0000	0.4647	0.0226	0.4417
4	0.0190	0.4239	0.1787	0.2139	0.2078	0.5441	0.5508	0.3660
5	0.5609	0.7100	0.2802	0.6458	0.4359	0.4770	1.0000	0.2278

ベースプロンプトを入力する場合，よりパラメータ数の大きいモデルを使用した方が面白い回答が生成されやすいといえる．また，創作具体例を入力する場合は，使用するモデルのパラメータ数で生成される回答の面白さが変わるとは言えない．

我々は，よりパラメータ数が大きいモデルほどプロンプトにおける抽象的な指示に従うことができるため，面白い出力ができることが理由であると考えた．創作具体例 4 を 70B モデルと 8B モデルに入力して生成したお題 3 に対する回答をいかに示す．

回答 1-5 飼育員が全員、動物の声真似がうまく、動物たちと会話しているようだが、飼育員同士では人間の言葉を一切使わず、動物の声を真似てコミュニケーションを取っていた

回答 1-6 飼育係が動物の名前を忘れてしまい、呼び方を適当にしている

回答 1-5 が 70B モデル，回答 1-6 が 8B モデルによって生成された．面白さに対する評価値は，回答 1-5 が 0.375，回答 1-6 が 0.125 である．いずれの回答も，飼育員 (飼育係) という同じ意味の単語が含まれている．そのため，両方のモデルが「お題から単語を列挙する」という思考は同様にできているため,「ネガティブなことを言う」という部分で面白さの差が生まれたといえる．我々は,「ネガティブなことを言う」が思考における抽象度が高い箇所であったと考える．また，ベースプロンプトにおける指示も単純な大喜利タスクの説明であるため，同様に抽象度が高いと考える．我々は提案手法の改善には，モデルのパラメータ数を考慮して創作具体例の抽象度を調整する必要があると考える．

表 5.7　パラメータ数の大きさでの検定における Cohen の d

創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
なし	0.945	0.450	-0.034	-0.087	-0.300	0.000	0.627	0.568
1	-0.397	0.470	0.137	-0.155	0.203	-0.770	0.194	0.136
2	0.287	0.000	-0.137	0.057	-0.085	0.101	-0.480	0.797
3	0.353	0.798	0.254	0.195	0.000	-0.208	0.667	0.219
4	0.687	0.228	0.386	0.356	0.361	0.173	0.170	0.258
5	0.166	0.106	-0.309	-0.131	-0.222	-0.203	0.000	0.346

Kappa 係数

生成した回答に対する評価値の Kappa 係数を，表 5.8 に示す．この実験全体では，明確な関連性に対する評価値の Kappa 係数が 0.5 を上回り，最も高い評価項目であった．また，文法，適切性，意図および他の要素に対する評価値の Kappa 係数は 0.2 以上と中程度であった．このことから，今回の実験である程度評価が一致する評価項目を作成できたといえる．面白さ，意外な関連性および知識に対する評価値の Kappa 係数は 0.1 を下回った．我々は，面白さ，意外な関連性および知識は，評価の際に個人の主観や感覚に依存することが，Kappa 係数が低くなった理由であると考える．

また，創作具体例や使用モデルによって Kappa 係数が異なった．大喜利 LLM が生成した回答は他の回答と比較して，面白さに対する評価値の Kappa 係数が高い．このことから，大喜利 LLM は多くの受け手が面白いと感じる回答を生成することがあるといえる．逆に，通常のプロンプトを 70B モデルに入力して生成した回答は，面白さに対する評価値は高いにもかかわらず Kappa 係数は-0.001 と低い．

表 5.8　生成した回答に対する評価値の Kappa 係数

モデル	創作具体例	面白さ	明確な関連性	意外な関連性	適切性	文法	意図	知識	他の要素
70B	1	0.019	0.037	0.001	0.013	0.017	0.179	0.133	0.218
	2	0.063	0.079	0.063	0.118	0.188	0.119	-0.005	0.198
	3	0.107	0.454	0.001	0.139	0.008	0.117	-0.03	0.100
	4	0.017	0.466	0.170	0.078	0.054	0.154	0.033	0.272
	5	-0.010	0.687	0.078	0.598	0.598	0.272	-0.005	0.079
	なし	-0.001	0.150	0.013	0.557	0.067	0.238	0.013	0.183
8B	1	0.054	0.300	-0.036	0.179	0.207	0.339	0.052	0.389
	2	0.006	0.032	0.022	0.199	0.217	0.152	0.091	0.006
	3	-0.026	0.534	0.188	0.310	0.579	0.259	0.134	0.098
	4	0.033	0.244	0.025	0.280	0.222	0.349	0.081	0.148
	5	-0.005	0.490	0.086	0.052	0.134	0.264	-0.005	0.417
	なし	0.128	0.571	0.134	0.562	0.431	0.420	0.319	0.264
SFT	なし	0.128	0.571	0.134	0.562	0.431	0.420	0.319	0.264
全体		0.072	0.507	0.072	0.282	0.215	0.288	0.096	0.269

このことから，通常のプロンプトと 70B モデルの組み合わせでは，一部の受け手が面白いと感じる回答を多く生成する傾向にあるといえる．

5.2 Tree-of-Thoughts

5.2.1 実験手順

実験は，以下の手順で行う．

Step 1 お題の主題を入力として，適用順が 1 回目である回答作成手順を全て適用して，LLM に回答を生成させる．

Step 2 Step 1 で生成した回答を評価者に評価させる．

Step 3 Step 2 で最もよかった回答を入力として，適用順が 2 回目である回答作成手順を全て適用して，Step 1 と同様のことを行う．これを，適用順が 3 回目の場合も行う．

Step 4 お題どうしの類似性に基づいて抽出した回答作成手順を適用して生成した回答に対する面白さ期待値が，ランダム抽出した回答作成手順を適用して生成した回答に対する面白さ期待値よりも大きいかどうかを確認する．また，LLM による自動評価に基づいて抽出した回答に対する面白さ期待値が，ランダム抽出した回答に対する面白さ期待値よりも大きいかどうかを確認する

本実験では回答生成と自動評価を 3 回繰り返すことによって，段階数が 3 である提案手法を想定する．これは，使用した大喜利データから生成した回答作成手順列のうち半数の 54.1% の長さが 3 であったためである．また，生成した回答の中で最も評価が高かった回答を，次段階における回答生成の入力としている．これは，前段階までで最も良い回答生成および LLM 自動評価ができた場合を想定して，各段階において仮定を検証するためである．

評価者は鈴木研究室の学生 11 名である．各回答に対して，面白いまたは面白くないの二値評価を評価者それぞれの主観にて行った．面白いと評価した評価者の人数を評価者の総数で割った数値を，各回答の最終的な面白さとする．また，主観評価である評価値の信頼性を検証するためにカッパ係数を算出し，評価の一致度合いを観察する．

我々は，小さいグループの回答作成手順は特定のお題に特化しているため，他のお題における回答生成に適用しても面白い回答を生成できないと考えた．そこで，グループの大きさが 5 個以上の回答作成手順のみを回答作成手順に適用する．ただし，対象のお題と類似度が最も大きいお題から抽出された回答作成手順は，グループの大きさによらず適用した．1 段階目では 39 個，2 段階目では 24 個，3 段階目では 13 個の回答作成手順を適用して，回答生成を行っている．本実験では合計 392 件の回答に対して評価を行っている．

LLM には一度に 5 個の回答作成事例を入力する．対象のお題との類似度が大きいお題がもとになった回答作成手順を必ず入力する．同じグループからサンプリングした 4 個の回答作成事例を入力する．サンプリングの際，シード値は 42 に固定している．

事前準備における回答作成手順の生成，提案手法における回答の生成及び自動評価に用いた LLM は，Google 社が公開している gemma3:27b[4] の Q4_K_M 量子化モデル¶である．生成時のパラメータは公開されているモデルカードを参考に，Top-k を 64,Top-p を 0.95 に設定した．Temperature は LLM 推論ソフトウェア Ollama の既定値である 1.0 に設定した．最大コンテキスト長には gemma3:27b の最大のコンテキスト長である 131072 を設定した．

文章分散表現の計算には Google 社が公開している埋め込みモデル EmbeddingGemma[7] を使用した．この埋め込みモデルに文章を入力することによって，768 次元の文章分散表現を得る．

本実験で使用したお題は，我々が大喜利投稿サイトや大喜利番組などを参考にして決定した以下の五つである．実験で使用したお題は，回答作成手順の元となった大喜利データには含まれない．

お題 1「疲れた仕事行きたくない」をポジティブに言い換えてください

お題 2「続いては、カルガモ親子の大冒険です」よりももっとほっこりするニュースを教えてください

お題 3 こんなスーパーマーケットはいやだ

お題 4 ギャルみたいな裁判官、どんなの？

¶https://ollama.com/library/gemma3:27b

お題 5 マグマに落とされて一言

五つのお題それぞれに対して，1 段階目から 3 段階目それぞれ全ての回答作成手順を適用して回答を生成する．本実験における既存手法として通常プロンプトの LLM への入力による回答生成および大喜利 LLM である Watashiha-Llama-2-13B-Ogiri-sft[‖]による回答生成を選択した．本手法で生成した回答と既存手法で生成した回答の面白さを比較することによって，仮定 2-1 を検証する．

五つのお題と 1 段階目から 3 段階目それぞれにおいて，お題どうしの類似度に基づいて回答作成手順を抽出する．抽出する回答作成手順の件数を変化させたとき，抽出した回答作成手順を適用して生成した回答の面白さ期待値の推移を観察する．抽出した回答作成手順を適用して生成した回答の面白さ期待値と，ランダムに回答作成手順を抽出した場合の回答の面白さ期待値を比較することによって，仮定 2-2 を検証する．

五つのお題と 1 段階目から 3 段階目それぞれにおいて，LLM による自動評価に基づいて回答作成手順を抽出する．抽出する回答の件数を変化させたとき，抽出した回答の面白さ期待値の推移を観察する．抽出した回答の面白さ期待値と，ランダムに回答を抽出した場合の回答の面白さ期待値を比較することによって，仮定 2-3 を検証する．

5.2.2 結果・考察

既存手法との比較

各段階における回答の面白さ評価値の最大値と，既存手法による回答の面白さ評価値を図 5.1 に示す．各グラフはお題ごとの本手法による回答および既存手法による回答の面白さを表す．各グラフの 1 段階，2 段階，3 段階は各段階における回答の面白さ評価の最大値である．通常 LLM は，通常のプロンプトを LLM に入力して生成した回答の面白さ評価である．大喜利 LLM は，大喜利 LLM が生成した回答の面白さ評価である．通常 LLM および大喜利 LLM は既存手法である．全てのお題において，各段階で生成した回答の面白さの最大値は既存手法で生成した回答

[‖]https://huggingface.co/watashiha/Watashiha-Llama-2-13B-Ogiri-sft

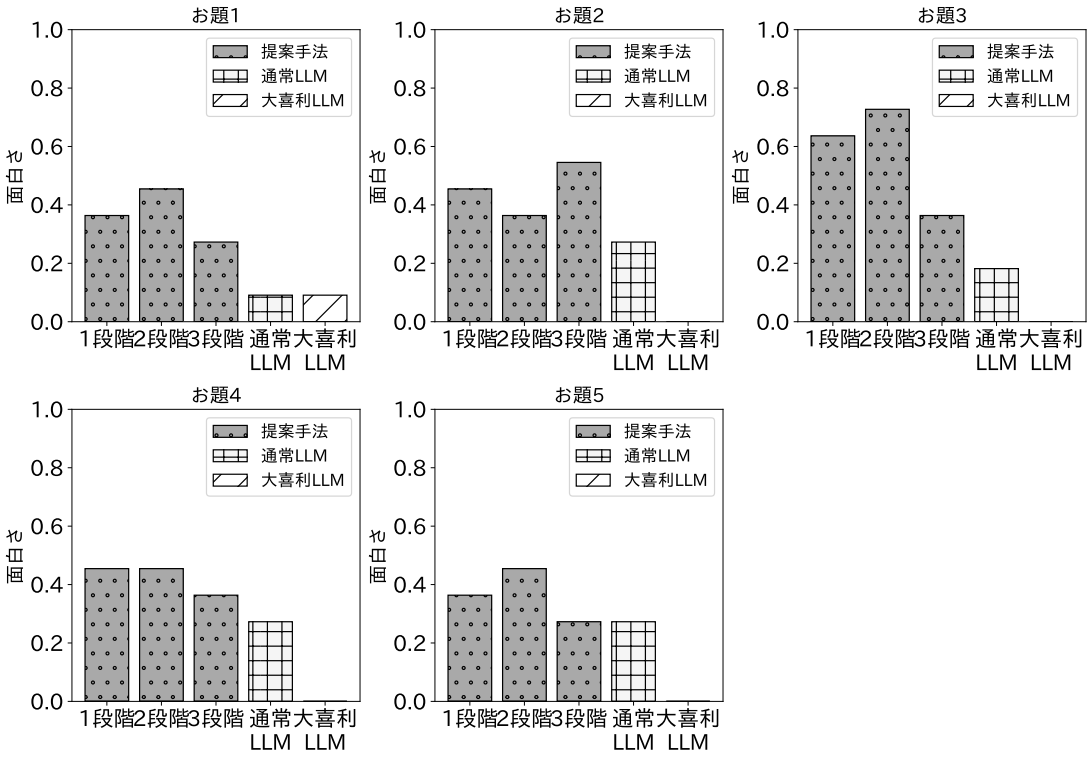

図 5.1　各段階における回答の面白さの最大値と既存手法による回答の面白さ

の面白さよりも大きいか同じである．よって，適切な回答作成手順を適用できた場合，仮定 2-1 は成り立つといえる．

お題ごとの各段階における回答の面白さ評価の箱ひげ図を 5.2 に示す．いずれのお題も各段階における回答の面白さ中央値は，0.2 未満である．このことからいずれの段階においても，LLM に入力しても面白い回答を生成できない回答作成手順がほとんどを占めているといえる．そのため，本手法で段階的に洗練された回答を生成するには，各段階において途中回答の評価を精度よく行う必要があるといえる．

また，お題によって回答作成手順の最適な適用回数が異なる．お題 2 以外の四つのお題において，2 段階目の方が 3 段階目よりも回答の面白さ最大値が大きい傾向にあった．このことから今回の実験対象であるお題には，回答作成手順の適用回数を 2 回目にした方が良いお題が多いと言える．以下の三つの回答は，本手法で生成したお題 3 に対する回答である．

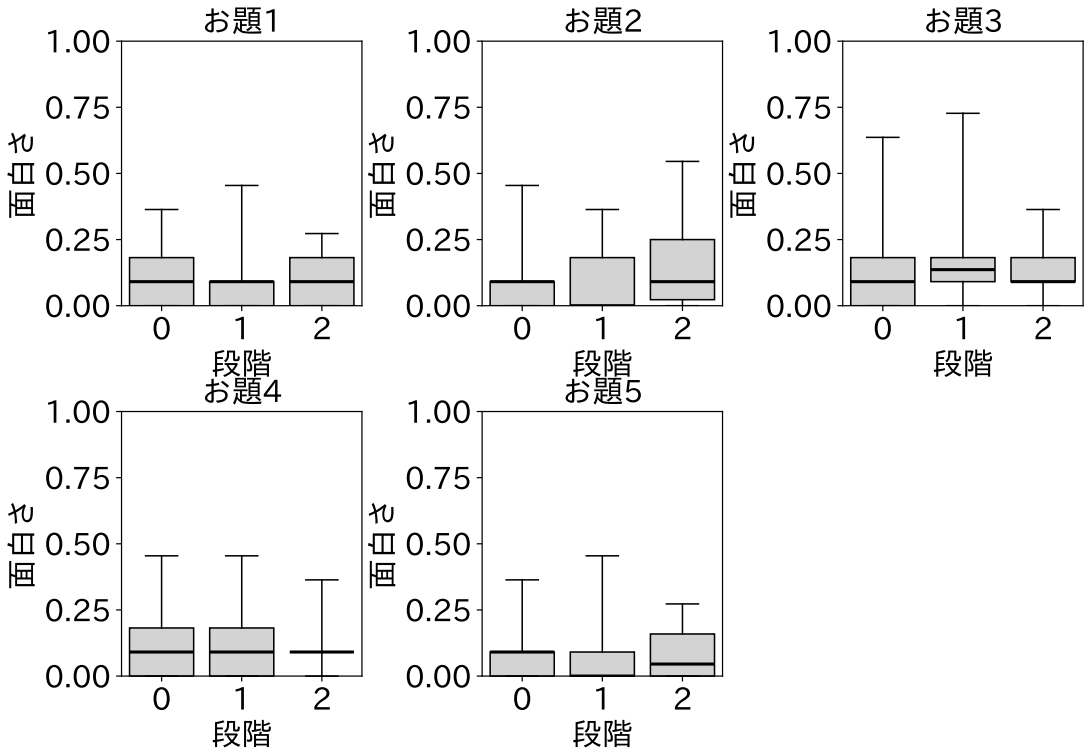

図 5.2　各段階における回答の面白さ最大値の箱ひげ図

回答 2-1 食材が全部動いている
回答 2-2 客が食材に選ばれる
回答 2-3 今日のおすすめはあなたです

回答 2-1，回答 2-2，回答 2-3 は，それぞれ 1 段階目から 3 段階目における最も面白い回答である．回答 2-1 の面白さは 0.636，回答 2-2 の面白さは 0.727，回答 2-3 の面白さは 0.364 である．回答 2-3 の内容は「今日のおすすめ (の食材) はあなたです」という，回答 2-2 をもとにしたお題である．回答 2-3 の面白さが低い理由は，「客が食材である」という内容が省略されたためであると考えられる．回答 2-3 が 3 段階目における面白さが最大の回答である．我々は，お題 3 の 3 段階目には，2 段階目の途中回答に適用することによって面白い回答を生成できる回答作成手順が存在しなかったと考える．手法の改善案として我々は，前段階の面白い回答も LLM による自動評価の対象とすることによって，お題ごとに回答作成手順の最適な適用

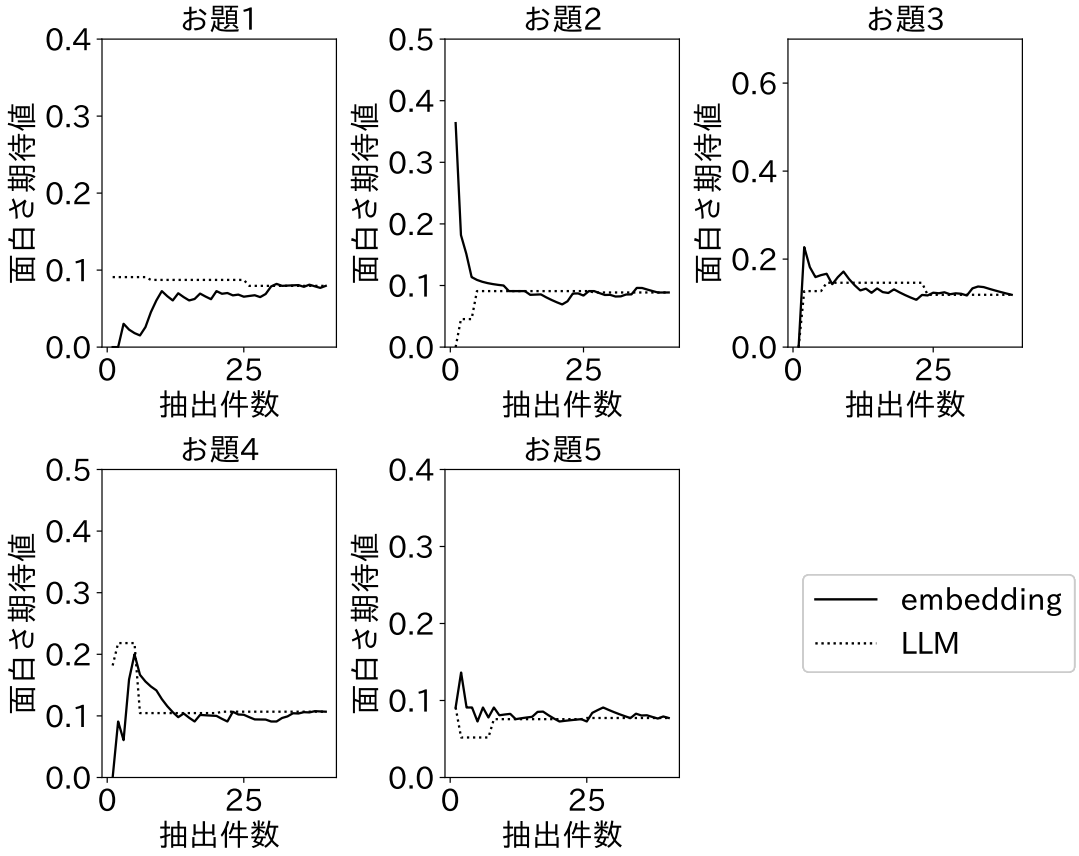

図 5.3　1 段階目におけるお題どうしの類似度に基づいて抽出した回答作成手順の件数と，抽出した回答を適用して生成した回答の面白さ期待値

回数での探索終了を考えている．

お題どうしの類似度による回答作成手順の抽出

お題どうしの類似度による回答作成手順の抽出件数と面白さ期待値の関係を図 5.3，図 5.4，図 5.5 に示す． これらのグラフにおける横軸は，文章分散表現の類似度および LLM による類似度によって抽出する回答作成手順の件数である．embedding が文章分散表現による類似度，llm が LLM による類似度の場合を表す．これらのグラフにおける縦軸は，抽出した回答作成手順を LLM に入力して生成した回答の面白さ相加平均である．embedding が文章分散表現による類似度，llm が

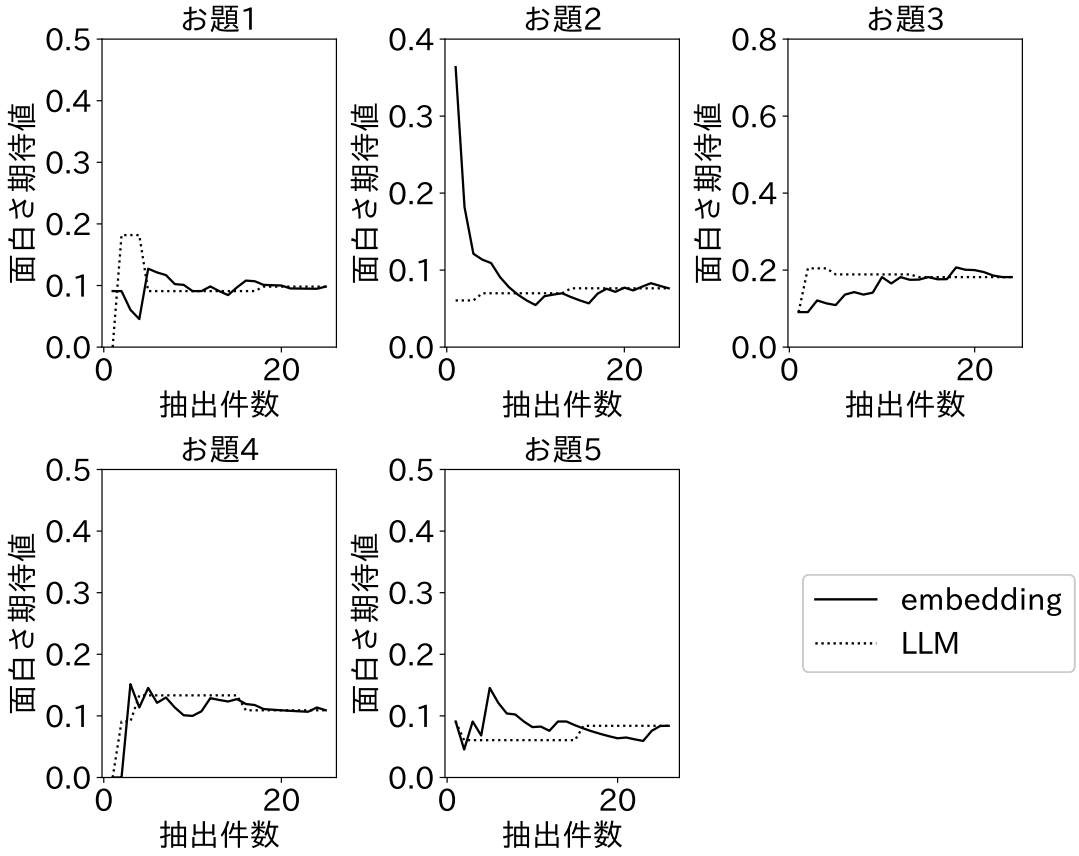

図 5.4　2 段階目におけるお題どうしの類似度に基づいて抽出した回答作成手順の件数と，抽出した回答を適用して生成した回答の面白さ期待値

LLM による類似度の場合を表す．図 5.3 のお題 3 を例にすると，文章分散表現による類似度が大きい上位 2 件の回答作成手順を LLM に入力したとき，生成される 2 個の回答に対する面白さの平均値が 0.23 であるということである．つまりこのグラフは，ある類似度をもとに回答作成手順をいくつか抽出したとき，抽出した回答作成手順を入力して生成される回答に対する面白さの期待値を表している．面白さ期待値が高い位置で推移する類似度ほど，その類似度をもとに面白い回答を作成できる回答作成手順を抽出できていることを意味する．また，グラフの終端は生成された全ての回答に対する面白さの平均である．そのためグラフの終端は，ランダムに抽出した回答作成手順を入力して生成される回答に対する面白さの期待値を表

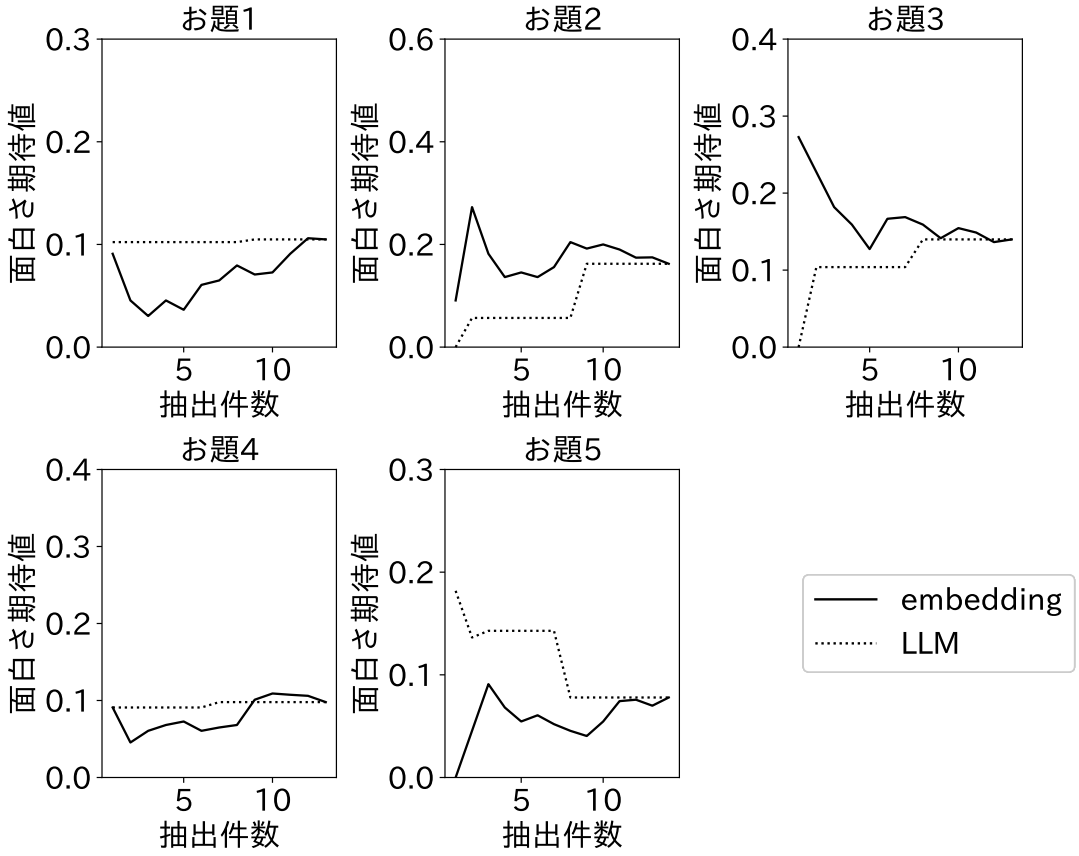

図 5.5　3 段階目におけるお題どうしの類似度に基づいて抽出した回答作成手順の件数と，抽出した回答を適用して生成した回答の面白さ期待値

す．このグラフを観察することによって仮定 2-2 を検証する．

文章分散表現に基づく抽出では，お題 2 の 1 段階目と 2 段階目において，上位 1 件を抽出したときにランダム抽出よりも面白さ期待値が約 4 倍高い．また, お題 4 の 1 段階目とお題 3 の 3 段階目において，ランダム抽出よりも面白さ期待値が約 2 倍ほど高い箇所がある．このことから，文章分散表現に基づいて抽出した回答作成手順を LLM に入力することによって，面白い回答を生成できることがあるといえる．

我々は，面白い回答を生成できる回答作成手順を文章分散表現に基づいて抽出できた理由の一つは，文章分散表現によって主題が似ているお題における回答作成手

順を抽出できたためであると考える．お題 2 と文章分散表現が最も類似するお題は「「今日のチャラいニュース」トップニュースは？」であった．二つのお題にはどちらも「ニュース」という単語が含まれるため，文章分散表現の類似度が大きくなったといえる．2 段階目においてこのお題から抽出できた回答作成手順は「経済ニュース風の表現」であった．1 段階目の「ペンギンが南極でプロポーズ」という途中回答に，この回答作成手順を適用して「ペンギンのプロポーズ、急増」という回答が得られた．この回答はお題 2 の 2 段階目において最も面白い回答であった．このことから，文章分散表現をもとにしたお題どうしの類似度からお題 2 の主題にふさわしい回答作成手順を抽出できたといえる．

我々は，反対に面白い回答を生成できる回答作成手順を文章分散表現に基づいて抽出できなかった理由の一つは，文章分散表現ではお題形式の類似度を計算できないことがあるためであると考える．お題 3 の 2 段階目における最も面白い回答は「食材が全部動いている」であった．この回答は,「こんなサウナは嫌だ、どんなサウナ?」と同じ「連想の拡張」という回答作成手順から導かれている．お題 3 と「こんなサウナは嫌だ、どんなサウナ?」は両方とも「こんな (施設や場所) はいやだ」形式のお題であるといえる．我々はこの二つのお題どうしの類似度は大きいことが望ましいと考える．しかし，お題 4 と「こんなサウナは嫌だ、どんなサウナ?」との文章分散表現の類似度は比較的小さかった．そのため「連想の拡張」は 40 件中 33 番目に抽出される回答作成手順であった．よって，文章分散表現ではお題形式の類似度が計算できないことがあるため，仮定 2-2 が成り立たない場合があるといえる．

我々は手法の改善案の一つに，ルールベースによるお題形式の分類を考える．大喜利には「こんな○○はいやだ」,「○○みたいな○○」のような代表的なお題形式があると考えられる．代表的なお題形式には,「こんな」,「いやだ」,「みたいな」のように，形式を特徴づける文字列が含まれるといえる．そこで代表的なお題形式を正規表現として保存したリストを参考にすることによって，お題形式に基づいたお題どうしの類似度計算ができると考える．

LLM による回答自動評価

1 から 3 段階目における LLM 自動評価による回答抽出と面白さ期待値の関係を図 5.6，図 5.7，図 5.8 に示す． これらのグラフにおける横軸は，LLM による自動

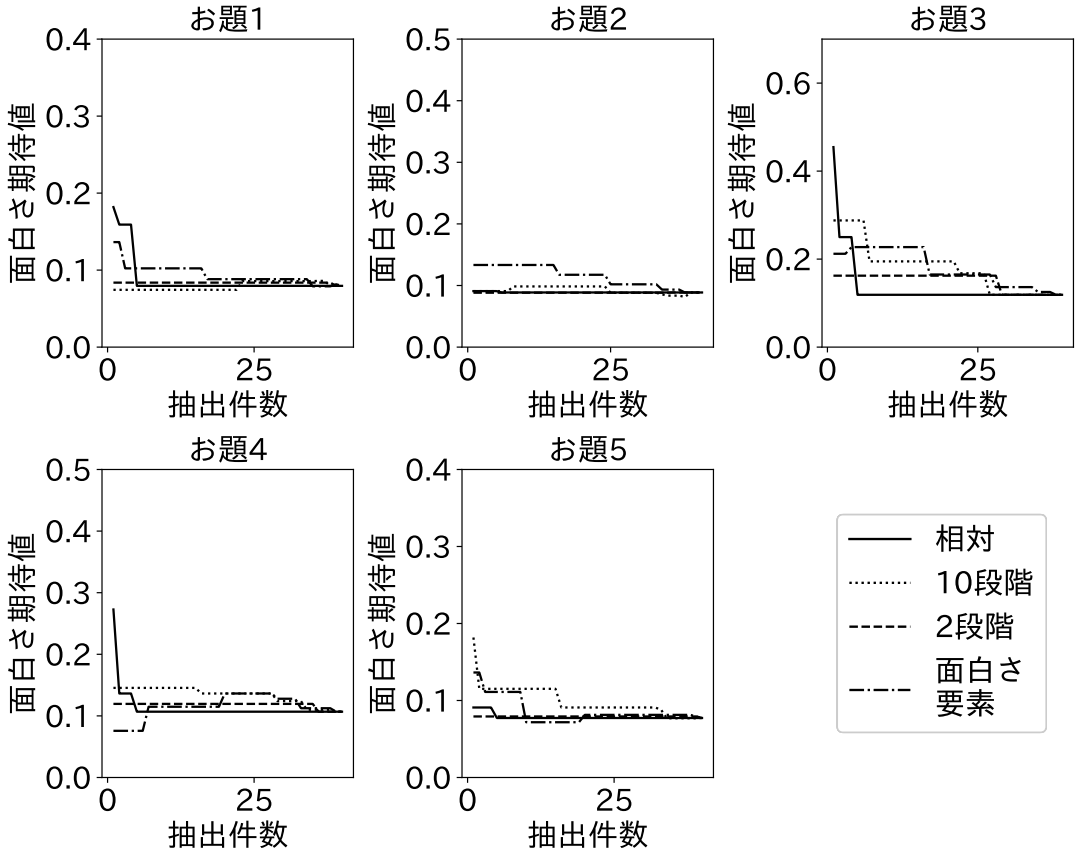

図 5.6　1 段階目における LLM 自動評価による回答の抽出件数と面白さ期待値

評価をもとに抽出する回答の件数である．これらのグラフにおける縦軸は，抽出した回答の面白さ相加平均である．図 5.6 のお題 3 を例にすると，LLM による相対評価が大きい上位 3 件の回答に対する面白さの平均値が 0.25 であるということである．つまりこれらのグラフは，ある LLM 自動評価に従って回答を順番に抽出したとき，抽出した回答に対する面白さ期待値の推移を表している．面白さ期待値が高い位置で推移する LLM 自動評価ほど，その評価をもとに面白い回答を抽出できていることを意味する．また，グラフの終端は生成された全ての回答に対する面白さの平均である．そのためグラフの終端は，ランダムに抽出した回答に対する面白さの期待値を表す．このグラフを観察することによって仮定 2-3 を検証する．

相対評価によって抽出した回答の期待値がランダム抽出した回答の期待値の 3 倍

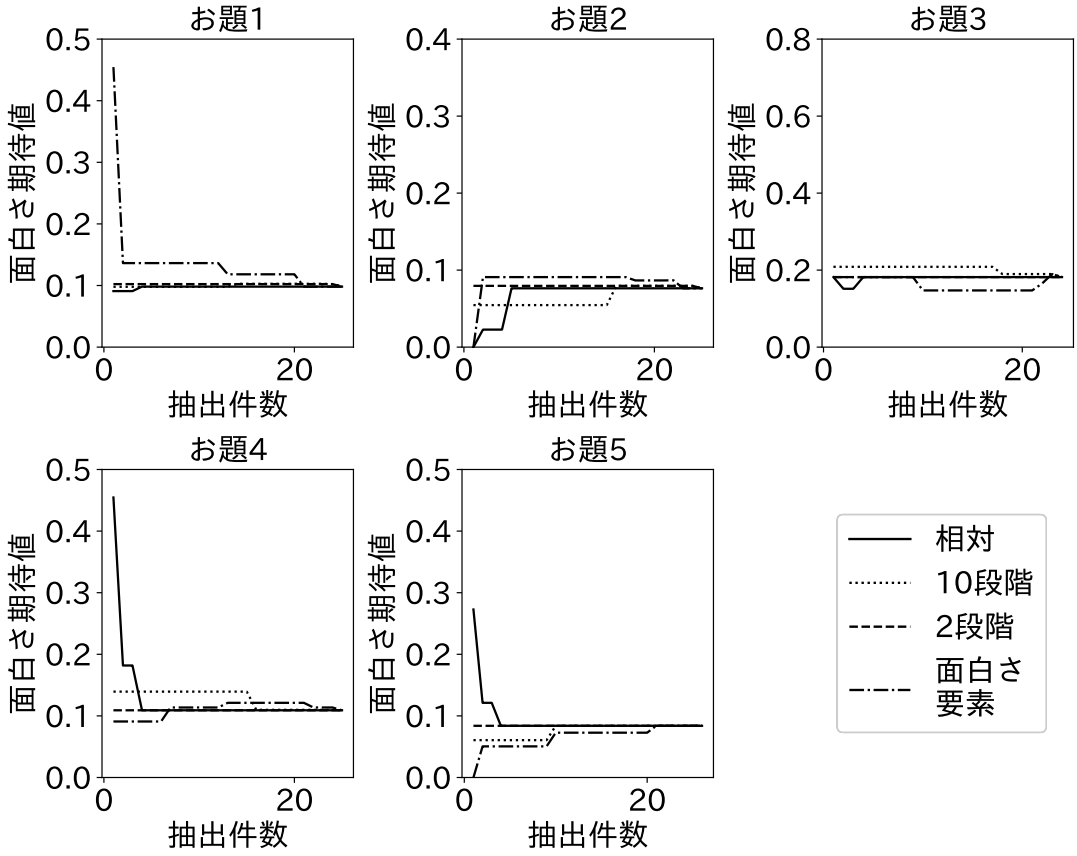

図 5.7　2 段階目における LLM 自動評価による回答の抽出件数と面白さ期待値

以上となる箇所が，お題 3 の 1 段階目やお題 4 の 2 段階目において認められる．また，ランダム抽出と比較して期待値が 2 倍以上となる箇所が，お題 1 の 1 段階目と 3 段階目，お題 2 の 3 段階目，お題 4 の 1 段階目と 3 段階目，お題 5 の 3 段階目に見られた．相対評価では LLM が最も面白いと判断した回答にのみ最大の評価を付与する．このことから LLM はいくつかの回答を比較する場合，面白い回答に対して面白いと評価する傾向がお題によって見られるといえる．

10 段階評価によって抽出した回答の期待値がランダム抽出した回答の期待値の 2 倍以上となる箇所が，お題 3 の 1 段階目と 3 段階目，お題 5 の 1 段階目において認められる．しかし，その他のほとんどのお題と段階では，ランダム抽出した回答の期待値と近い値を推移している．このことから，LLM が 10 段階評価を行うと

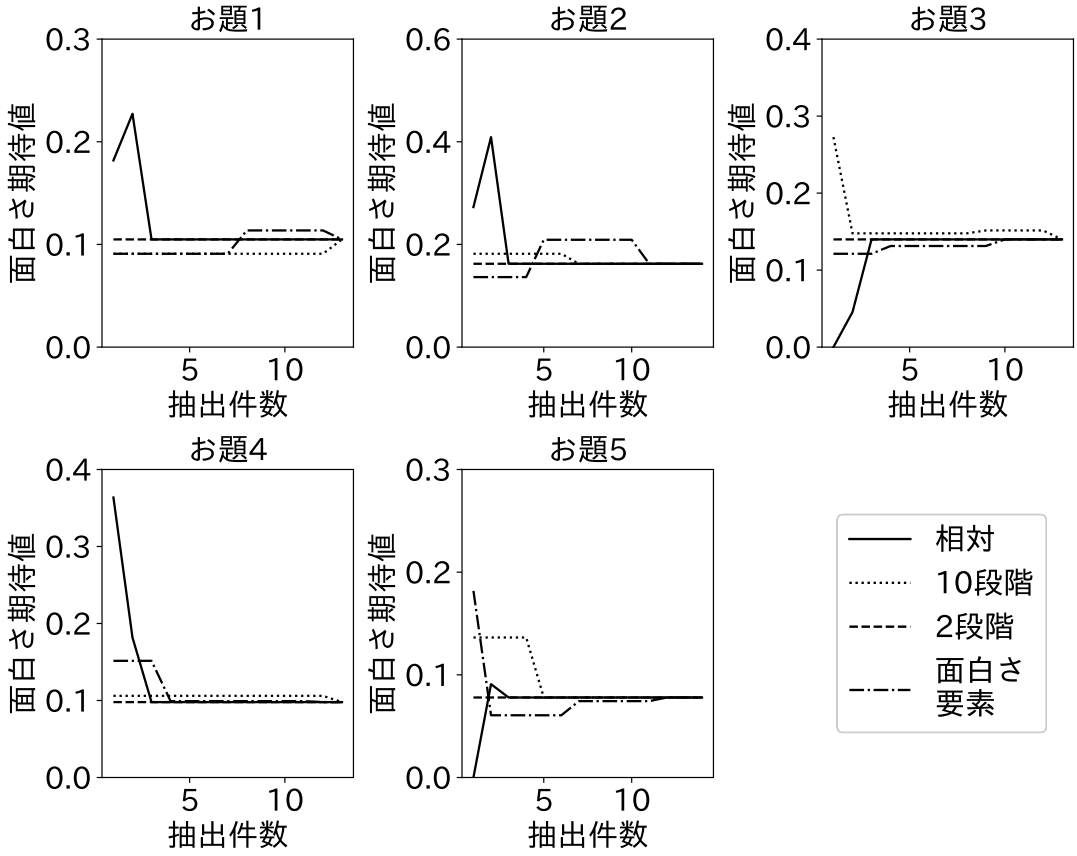

図 5.8　3 段階目における LLM 自動評価による回答の抽出件数と面白さ期待値

き，多くの場合で回答の面白さによって評価をつけられない傾向にあるといえる．また，面白さ期待値が完全な横ばいに近い推移をしている箇所がある．これは，ほとんど全ての回答に対して同じ評価をしているためである．

2 段階評価によって抽出した回答の期待値は，ほとんど全ての場合でランダム抽出した回答の期待値と近い値を推移している．このことから，LLM は本当に面白い回答だけに対して面白いと評価し，それ以外に面白くないと評価をつけることができないといえる．

面白さ要素での自動評価によって抽出した回答の期待値は，ほとんど全ての場合でランダム抽出した回答の期待値と近い値を推移している．ところが，お題 1 の 2 段階目においては，他の LLM 自動評価がいずれもランダム抽出に近い推移をして

いるのに対して，抽出した回答に対する面白さの期待値がランダム抽出と比べて 3 倍以上高い箇所がある．以下の三つは，お題 1 の 2 段階目における回答である．

回答 2-4 今日も生きて、社畜として全うできる！
回答 2-5 今日も地球が私を必要としてる
回答 2-6 ゾンビのように出勤する

回答 2-4 の面白さは 0.455 であり，お題 1 の 2 段階目において最も面白かった回答である．回答 2-5 の面白さは 0.273，回答 2-6 の面白さは 0.182 である．我々は，回答 2-4 の面白いところは「社畜」という言葉を使用している点であると考える．面白さ要素での自動評価は，回答 2-4 が最大値の 7，回答 2-5 と回答 2-6 が 6 であった．面白さ要素での自動評価において，LLM は回答 2-4 に対して，知識の項目に 1 を付与している．LLM の出力を見たところ，「社畜」という言葉が知識の項目における評価の理由であることがわかった．そして，回答 2-5 の面白さ要素での自動評価が 6 であるのは，知識の項目が 0 であったためだとわかった．よって，面白さ要素での自動評価にける知識の項目で，回答 2-4 が持つ「社畜」の面白さを評価できたといえる．また，回答 2-6 の面白さ要素での自動評価が 6 であったのは，お題 1 のポジティブな言い換えに合致していないと LLM が判断したことによって，適切性の項目が 0 になったためであるとわかった．よってお題 1 の 2 段階目において，回答の面白さを決める方向性と，LLM が判断する面白さに寄与する要素の有無と重なったといえる．以上の理由から，お題 1 の 2 段階目において抽出した回答に対する面白さの期待値が高く推移したと我々は考える．このことから，回答の面白さを決める方向性と LLM が面白さを判断する基準の方向性が重なる場合，仮定 2-3 が成り立つといえる．

Kappa 係数

実験 5.2 における回答に対する面白さ評価の Kappa 係数を表 5.9 に示す．実験全体の Kappa 係数は 0.074688 であった．Kappa 係数は 1 に近づくほど評価者間での評価が一致する傾向にあり，0 に近づくほど一致しない傾向にある指標である．そのため，本実験における面白さ評価は，評価者間で一致しづらい傾向にあるといえる．我々は評価者間の面白さ評価が一致しない理由の一つは，評価者ごとに面白

さの感じやすさ異なるためであると考えた．最も多くの回答に対して面白いと判断した評価者は，392 個中 131 個の回答に対して面白いと評価していた．最も少ない回答に対して面白いと判断した評価者は，392 個中 3 個の回答に対して面白いと評価していた．このことから，本実験では評価者間の笑いやすさに大きな差があったといえる．我々は，評価の重要性が評価者の笑いやすさによって異なると考えた．評価者間の笑いやすさの差がお笑いや大喜利に対する評価の集計における課題であると考える．そのため，評価者ごとの笑いやすさを考慮した面白さ評価の集計方法が必要であると考える．

表 5.9　回答に対する面白さ評価の Kappa 係数

		お題					
		1	2	3	4	5	全体
段階数	1	0.025326	0.050815	0.132120	0.070976	0.018256	0.067849
	2	0.038978	0.061680	0.093519	0.064626	0.154008	0.096054
	3	0.001979	0.121364	0.011789	0.002326	0.023944	0.049559
	全体	0.026962	0.082420	0.102498	0.058372	0.065679	0.074688

第 6 章　おわりに

本研究では，大喜利の回答生成手法として，CoT(Chain-of-Thought；思考の連鎖プロンプト) による手法と，ToT(Tree-of-Thoughts) による手法の二つの手法を提案した．CoT による手法で我々は人間が大喜利の回答を導く考え方を，CoT 形式のプロンプトである創作具体例として構造化した．創作具体例を LLM に入力することによって，人間が回答を導く考え方と同様の過程に従った回答生成を LLM に促した．

実験の結果，創作具体例を LLM に入力して生成した回答は既存手法で生成した回答よりも面白さが低い傾向にあることがわかった．また，創作具体例を入力して生成した回答の面白さは，使用する LLM のパラメータ数の大きさに依存しないことがわかった．実験の結果から，大喜利は思考という形式で構造化できないタスクであったといえる．そのため回答作成手順の構造化・再利用では，タスクごとに最適な構造化形式を検討する必要があることがわかった．

実験の結果，CoT による手法では，知識が含まれる回答や多様な形式の回答が生成しづらい傾向にあった．これはタスクにおける出力に対する評価観点の一部が考慮できない構造化形式であったことが理由であると考えられる．また，使用モデルによって抽象的な指示に対する出力の品質に差があった．タスクにおける評価観点や使用モデルの性能を考慮して，回答を導く考え方の構造化形式を精査する必要がある．

ToT を参考にした手法で我々は，人間が大喜利の回答を導く考え方を回答作成手順として構造化できると仮定した．回答作成手順とはお題から回答を導くときの考え方をいくつかの手順に分割したものである．複数の回答作成手順を組み合わせて段階的に適用することによって，段階的に洗練される回答生成を LLM に促した．また，回答作成手順の組み合わせ数削減のために，いくつかの仮定を導入した．類似するお題どうしは同じ回答作成手順が有効であると仮定し，お題どうしの類似度に基づいて抽出した回答作成手順を適用して回答を生成した．LLM によって回答の面白さを評価できると仮定して，評価の高い途中回答のみをもとにして次段階の回答生成を行った．回答作成手順の役割が適用順によると仮定し，段階数と一致する回答作成手順のみ適用した．また，タイトルが同じ回答作成手順は内容が同じで

あると仮定し，タイトルによるグループ化を行った．

実験の結果，回答作成手順を複数回適用することによって，既存手法よりも面白い回答が生成できることがわかった．このことから，大喜利における回答を導く考え方を回答作成手順として構造化できるという仮定が支持されたといえる．

特定のお題において，お題どうしの文章分散表現類似度に基づいて抽出した回答作成手順を適用すると，面白い回答を生成できた．このことから，入力どうしの類似度に基づいて再利用する回答作成手順を抽出することに一定の有効性があることがわかった．我々は，特定の場合に文章分散表現がお題と回答作成手順との相性を表す特徴となってことが理由であると考える．

特定のお題と段階において，LLM の自動評価によって面白い回答を抽出できることがわかった．このことから，LLM による自動評価に基づいて抽出した回答をもとにした次段階の回答生成に，一定の有効性があることがわかった．我々は，LLM の評価観点と特定のお題における回答の面白さ評価の方向性が一致したことが理由だと考えた．

今後の展望を述べる．本研究では，ToT を参考にした手法で導入した仮定「回答作成手順は適用順によって役割が決まる」,「同じタイトルを持つ回答作成手順は内容が同じ」が未検証である．これらの仮定を検証することによって，回答作成手順の効率的な再利用に対するさらなる考察が必要である．また，ToT を参考にした手法の多様な形式のタスクへの一般化が挙げられる．この手法は，段階的に洗練される回答，LLM によって回答が自動評価可能であること，他の入力における出力を導く考え方が流用可能であることといった性質をタスクに仮定していることが課題である．一般化によって，より少ない仮定においても回答作成手順の再利用が可能な手法の提案が必要であると考える．また，ToT を参考にした手法の改善策に，入力どうしの類似度計算手法のタスク別設計や，入力ごとに最適な評価観点に基づいた自動評価が挙げられる．これらの改善によって，本実験で特定の場合でお題どうしの類似度や LLM 自動評価にみられた傾向を，より多くのお題や段階で見られるようになるといえる．

謝辞

本研究を進めるにあたって，指導教員である鈴木優教授には多くのことをご教授いただきました．学部と大学院を通じて大喜利というテーマに取り組めたのは，指導教員が鈴木先生だったからこそだと思います．3 年半という長い間，お世話になりました．感謝申し上げます．ありがとうございました．

研究室の事務補佐員である駒澤さんと井尾さんには，各種申請や事務作業の際に大変お世話になりました．感謝申し上げます．ありがとうございました．

鈴木研究室の皆さんのおかげで，楽しい研究室生活を送ることができました．特に自分と同期である川上君と城所君には研究以外の多くの場面で支えてもらいました．また，すでに鈴木研究室を卒業された先輩方には現在も大変良くしていただいています．感謝申し上げます．今後ともよろしくお願い申し上げます．

私の生活を支えてくれた家族と，仲良くしてくれている友人に感謝申し上げます．ありがとうございました．

本研究で使用しました言語モデルである Gemma3，EmbeddingGemma および Llama3 を公開してくださいました Google 社，Meta 社に感謝申し上げます．また，Llama3 の日本語版モデルである llama-3-youko-70b-instruct および llama-3-youko-8b-instruct を公開してくださいました rinna 株式会社，大喜利 LLM である Watashiha-Llama-2-13B-Ogiri-sft を公開してくださいました株式会社わたしはに感謝申し上げます．

創作具体例を作成するにあたって参考にしました大喜利解説記事の著者である赤嶺総理さんに感謝申し上げます．また，回答作成手順を生成するもととなった大喜利データの取得元である大喜利サイトである大喜利総合サイト，大喜利茶屋および遊び場大喜利に感謝申し上げます．

Passion．魂．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.

[2] Seyed Mahmoud Sajjadi Mohammadabadi, Burak Cem Kara, Can Eyupoglu, Can Uzay, Mehmet Serkan Tosun, and Oktay Karakuş. A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. *Electronics*, Vol. 14, No. 18, 2025.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. *arXiv e-prints*, p. arXiv:2407.21783, July 2024.

[4] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report, 2025.

[5] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] Henrique Schechter Vera, Sahil Dua, Biao Zhang, et al. Embeddinggemma: Powerful and lightweight text representations, 2025.

[8] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan,

and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24824–24837. Curran Associates, Inc., 2022.

[10] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 22199–22213. Curran Associates, Inc., 2022.

[11] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[12] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, pp. 378 – 382, 1971.

[13] 竹内海地, 謝孟春, 中嶋崇喜, 森徹. 転移学習を用いた大喜利回答システムの構築. 2022 年度 情報処理学会関西支部 支部大会 講演論文集, Vol. 2022, , 09 2022.

[14] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13246–13257, June 2024.

[15] 中川裕貴, 村脇有吾, 河原大輔, 黒橋禎夫. クラウドソーシングによる大喜利の面白さの構成要素の分析. 言語処理学会第 25 回年次大会 (NLP2019), 2019. B3-2.

[16] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[17] Lewis Patrick, Perez Ethan, Piktus Aleksandra, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

発表リスト

[1] 高橋巧実，鈴木優『大喜利創作技法の抽出と Chain-of-Thought プロンプト設計による回答生成手法の基礎検討』WebDB 夏のワークショップ 2025, 2025.
[2] 高橋巧実，鈴木優『創作具体例の入力による経験則に従った大喜利の回答自動生成』東海関西データベースワークショップ 2025，2025
[3] 高橋巧実，鈴木優『Tree-of-Thoughts を用いた最適な回答作成手順の探索による大喜利自動生成 』DEIM2026 第 18 回データ工学と情報マネジメントに関するフォーラム, 2026.