

2025年度 卒業論文

生成 AI の性能比較

大東文化大学

経営学部経営学科

氏名 小林 龍生

所属研究室 白井 康之 ゼミ

学籍番号 22161380

# 目次

<b>1.はじめに.....</b>	<b>4</b>
1-1.背景.....	4
1-2.生成 AI について.....	4
<b>2.調査方法.....</b>	<b>5</b>
2-1.調査対象.....	5
2-2.調査項目.....	5
2-3.評価方法.....	6
2-3-1.文章生成の可否および生成形式に関する能力の評価方法 .....	6
2-3-2.文脈理解力の評価方法 .....	7
2-3-3.翻訳の精度の評価方法 .....	7
<b>3.調査結果.....</b>	<b>7</b>
<b>3-1.文章生成の可否および生成形式に関する能力.....</b>	<b>8</b>
3-1-1.ChatGPT の調査結果.....	8
3-1-2.Claude の調査結果.....	10
3-1-3.Gemini の調査結果.....	11
<b>3-2.文脈理解力 .....</b>	<b>13</b>
3-2-1.ChatGPT からの出題.....	13
3-2-2.Claude からの出題.....	17
3-2-3.Gemini の調査結果.....	19

3-3.翻訳の精度 .....	23
3-3-1.ChatGPT の調査結果.....	23
3-3-2.Claude の調査結果 .....	24
3-3-3.Gemini の調査結果 .....	25
3-4.各生成 AI への自己評価.....	26
3-4-1. ChatGPT への自己評価 .....	26
3-4-2. Claude への自己評価 .....	26
3-4-3. Gemini への自己評価.....	27
4. 各生成 AI における課題.....	27
4-1. ChatGPT の課題.....	27
4-2. Claude の課題.....	28
4-3. Gemini の課題 .....	28
4-4. 課題の総括.....	29
5.おわりに.....	29
5-2.総合ランクづけ（本研究に基づく） .....	30
5-3. 今後の課題.....	30
6.参考文献.....	32

## 1.はじめに

近年、大規模言語モデル（Large Language Models, LLMs）の急速な発展により、生成 AI は社会の様々な場面で活用されるようになってきている。特に 2022 年末の ChatGPT の登場以降、対話型 AI は一般ユーザーにも広く普及し、ビジネス、教育、研究開発など多岐にわたる分野で実用化が進んでいる。

現在、生成 AI 市場においては、OpenAI の ChatGPT（GPT シリーズ）、Anthropic の Claude、Google の Gemini などが主要な生成 AI として存在している。これらのモデルは、それぞれ異なる設計と技術的アプローチを持ち、独自の強みを主張している。しかし、実際の性能や適用場面における優位性については、比較研究が不足している。

### 1-1.背景

ChatGPT は史上最速で普及したアプリケーションとなった。野村総合研究所（2023）によれば、2022 年 11 月 30 日の公開から 4 日で利用者 100 万人、2 ヶ月で 1 億人を突破し、TikTok（9 ヶ月）や Instagram（2 年 4 ヶ月）を大きく上回る普及速度を記録した。同調査では日本国内の利用動向も分析されており、2023 年 4 月時点での職業別利用率は学生（21.6%）と教職員（20.5%）が最も高く、教育分野での早期導入が進んでいることが明らかになっている。[1]

そのため、ChatGPT をはじめとする生成 AI を取り上げていき、現在で利用率が高い教育分野における生成 AI の性能について研究していく。

### 1-2.生成 AI について

生成 AI（Generative Artificial Intelligence）とは、機械学習、とりわけ深層学習に基づくモデルを用いて、新規のデータやコンテンツを自動生成する人工知能技術のことである。

生成 AI は、大規模なテキスト、画像、音声、映像などのデータを学習し、その統計的特徴を抽出することで、人間が作成したものと同等、もしくはそれに近い自然なコンテンツを生成できる点に特徴がある。代表的な技術として、自然言語処理分野における大規模言語モデル (Large Language Models: LLM)、画像生成に用いられる拡散モデル (Diffusion Models)、および生成的敵対ネットワーク (Generative Adversarial Networks: GANs) などが挙げられる。

## 2.調査方法

生成 AI の比較にあたっては、実証的調査の方法を用いる。実証的調査として、同一の入力条件 (プロンプト) を用いた比較実験を行っていく。具体的には、同一テーマに基づく文章生成課題を各生成 AI に実行させ、その出力結果を収集した。これにより、生成結果の違いを直接的に比較する。

### 2-1.調査対象

本研究では、生成 AI の調査対象として ChatGPT (OpenAI)、Claude (Anthropic)、Gemini (Google) の 3 つを選定した。これらは、いずれも大規模言語モデルを基盤とし、文章生成を中心とした高度な自然言語処理機能を備えている点で共通している。また、一般利用者が容易にアクセス可能であり、学習・業務の両面で利用実績が多いことから、比較対象として適切であると判断した。

2025 年 12 月時点での無料基盤モデルを使用する。ChatGPT は GPT-4o、Claude は Claude 3.5 Sonnet、Gemini は Gemini 2.5 Flash を使用する。

### 2-2.調査項目

ChatGPT、Claude、Gemini の三つの生成 AI を比較するにあたり、以下の 3 つの調査項目を設定した。

1. 文章生成の可否および生成形式に関する能力

文章生成の可否および生成形式を測るために、各生成 AI が説明文、要約文、論述文など、異なる形式の文章生成にどの程度対応可能であるかを評価する項目である。具体的には、指示された文章形式を正確に理解し、目的に沿った構成および文体で文章を生成できているかを確認した。

## 2. 文脈理解力

文脈理解の適切さを測るために、生成 AI が入力されたプロンプトの意図や文脈を正しく理解し、前後関係を踏まえた一貫性のある文章を生成できているかを評価する項目である。指示内容から逸脱した記述や、文脈上不自然な展開が見られないかを確認した。ここでは、各生成 AI になぞなぞを作成してもらい、そのほかの生成 AI に解いてもらい判断した。

## 3. 翻訳の精度

翻訳の精度を測るために、英語から日本語への翻訳課題を設定し、意味内容が適切に保持されているか、誤訳や不自然な表現が生じていないかを評価した。特に、専門用語や文脈依存的な表現の翻訳の正確さに着目した。

以上の調査項目を用いることで、文章生成および言語処理能力の観点から、ChatGPT、Claude、Gemini の特徴を多面的に比較・分析する。

## 2-3.評価方法

### 2-3-1.文章生成の可否および生成形式に関する能力の評価方法

以下の観点を用いて、ChatGPT、Claude、Gemini それぞれに相互評価をし、リストアップとともに自己評価も用いて判断をする。

#### 【評価観点】

1. 内容の正確性 (5 点)
2. 文脈理解の適切さ (5 点)

3. 構成の明瞭さ (5 点)

4. 表現の自然さ (5 点)

5. 指示遵守度 (5 点)

#### 【出力形式】

- ・各観点の点数
- ・総合評価 (100 字程度)

### 2-3-2.文脈理解力の評価方法

文脈理解力は、以下の観点から評価する。

- ・答えがなぞなぞの条件を満たしている場合：「正答」
- ・条件の一部を誤解している場合：「部分的理解」
- ・文脈を誤認し、成立しない答えを提示した場合：「誤答」

これらをそれぞれの生成 AI に評価してもらい、その評価を記載する。

### 2-3-3.翻訳の精度の評価方法

手順としては、①日本語→英語、②英語→日本語を行い、文章にどれほどのずれがあるかを評価する。

## 3.調査結果

調査方法に基づき、ChatGPT、Claude、Gemini の 3 つの生成 AI について得られた結果を示し、比較・分析を行う。

### 3-1.文章生成の可否および生成形式に関する能力

文章生成の可否および生成形式に関する能力を測るために、

「生成 AI とは何か」というテーマについて、

- ①50 字の要約文
- ②200 字の説明文
- ③400 字の論述文

の 3 種類の文章をそれぞれ作成してください。

というプロンプトについて調査をそれぞれ行った。

#### 3-1-1.ChatGPT の調査結果

- ① 50 字の要約文

生成AIとは、大量のデータを学習し、新たな文章や画像などのコンテンツを自動生成する人工知能である。

- ② 200 字の説明文

生成 AI とは、過去に蓄積された膨大なデータを基に学習し、文章、画像、音声などの新たなコンテンツを自動的に生成する人工知能である。従来の AI が分類や予測を主な目的としていたのに対し、生成 AI は人間の創造的活動を支援・代替できる点に特徴がある。近年では深層学習技術の発展により、高精度な文章生成や画像生成が可能となり、教育、ビジネス、研究など幅広い分野で活用が進んでいる。

- ③ 400 字の論述文

生成 AI とは、機械学習、とりわけ深層学習技術を用いて、大量のデータからパターンや文脈を学習し、新たな文章や画像、音声などを生成する人工知能である。代表例として、文章生成 AI や画像生成 AI が挙げられ、これらは人間の創造的作業を補助する技術として注目されている。生成 AI の普及は、業務効率化や表現の多様化といった利点をもたらす一方で、誤情報の生成や著作権侵害、倫理的課題といった問題も引き起こしている。その

ため、生成 AI を社会に適切に導入するには、技術的理解だけでなく、利用ルールや評価基準の整備が不可欠である。今後は、生成 AI の利便性とリスクの双方を踏まえた上で、人間と AI が協働するあり方が求められる。

#### Claude からの評価

評価観点	
内容の正確性(5点)	5
文脈理解の適切さ(5点)	5
構成の明瞭さ(5点)	5
表現の自然さ(5点)	5
指示遵守度(5点)	5
合計点(25点満点)	25

🗨️ 総合評価

#### ChatGPT 【25点/25点】★優秀

全観点で満点を獲得。文字数制限を厳守し、段階的に情報を深化させる構成が秀逸。生成 AI の利点と課題をバランスよく提示し、論理的かつ実用的。読みやすさと専門性のバランスが優れており、幅広い読者層に対応可能。

#### Gemini からの評価

評価観点	
内容の正確性(5点)	5
文脈理解の適切さ(5点)	5
構成の明瞭さ(5点)	5
表現の自然さ(5点)	5
指示遵守度(5点)	5
合計点(25点満点)	25

総合評価（100字程度）

この出力は、内容の正確性、論理構成、表現の自然さ、全ての観点で極めて優秀です。特に、50字、200字、400字という字数制限を完全に遵守しており、指示に対する形式制御能力が非常に高いことが確認できました。生成 AI の文章生成能力を測るテストとして最高の評価に値します。

## 3-1-2.Claude の調査結果

### ①50 字の要約文

生成 AI は、テキスト、画像、音声などを自動的に生み出す人工知能技術で、大量のデータから学習したパターンを基に新しいコンテンツを創造する。

### ②200 字の説明文

生成 AI とは、テキスト、画像、動画、音声など様々なコンテンツを自動生成できる人工知能技術である。従来の AI が分類や予測を行うのに対し、生成 AI は新しい創作物を生み出す点が特徴だ。膨大なデータを学習し、そのパターンや特徴を理解することで、人間の指示に応じて文章を書いたり、絵を描いたり、プログラムコードを作成したりできる。ChatGPT や Stable Diffusion などが代表例で、ビジネスや教育、創作活動など幅広い分野で活用が進んでいる。

### ③ 400 字の論述文

生成 AI とは、機械学習、とりわけ深層学習技術を用いて、大量のデータからパターンや文脈を学習し、新たな文章や画像、音声などを生成する人工知能である。代表例として、文章生成 AI や画像生成 AI が挙げられ、これらは人間の創造的作業を補助する技術として注目されている。生成 AI の普及は、業務効率化や表現の多様化といった利点をもたらす一方で、誤情報の生成や著作権侵害、倫理的課題といった問題も引き起こしている。そのため、生成 AI を社会に適切に導入するには、技術的理解だけでなく、利用ルールや評価基準の整備が不可欠である。今後は、生成 AI の利便性とリスクの双方を踏まえた上で、人間と AI が協働するあり方が求められる。

### ChatGPT からの評価

評価観点	
内容の正確性(5点)	5
文脈理解の適切さ(5点)	5
構成の明瞭さ(5点)	5

表現の自然さ(5点)	4
指示遵守度(5点)	5
合計点(25点満点)	24

総合評価（約 100 字）

生成 AI の定義から利点・課題までを正確かつ段階的に記述しており、要約文・説明文・論述文それぞれの目的に適した構成となっている。学術的にも妥当で、指示遵守度が高い一方、表現面ではやや定型である。

Gemini からの評価

評価観点	
内容の正確性(5点)	5
文脈理解の適切さ(5点)	5
構成の明瞭さ(5点)	5
表現の自然さ(5点)	5
指示遵守度(5点)	3
合計点(25点満点)	23

総合評価（100 字程度）

この出力は、内容の正確性、論理構成、表現の自然さの点で極めて優秀です。説明文と論述文は字数指示を厳密に遵守していますが、要約文で指示字数をオーバーしており、形式制御の精度にわずかな一貫性の欠如が見られます。全体としては、高品質な文章生成能力が確認できます。

### 3-1-3.Gemini の調査結果

#### ① 50 字の要約文

大量データから学習し、テキスト・画像・音声など、新しいコンテンツを自律的に創造する AI 技術。

#### ② 200 字の説明文

生成 AI (Generative AI) とは、機械学習、特に深層学習 (ディープラーニング) を基盤とし、学習した大量のデータに含まれるパターンや構造を理解することで、そのデータには

存在しなかった新しいコンテンツを生成できる AI 技術です。テキスト、画像、音声、動画、プログラムコードなど、多様な形式のデータを創造的に出力します。代表的なモデルとして、文章生成の LLM（大規模言語モデル）や、画像生成の Diffusion Model などがあり、ビジネスやクリエイティブ分野での活用が急速に進んでいます。

### ③ 400 字の論述文

生成 AI とは、従来の AI が識別・予測・分類といった分析的なタスクを得意としていたのに対し、\*\*「創造」\*\*という新しい領域を可能にした AI 技術の総称です。

その核となるのは、GAN（敵対的生成ネットワーク）や Transformer、Diffusion Model といった深層学習（ディープラーニング）モデルです。これらのモデルは、インターネット上の膨大なテキストや画像データから複雑な特徴や文脈、統計的なルールを抽出し、人間が指示するプロンプトに基づいて、学習データセットにはなかったオリジナルのコンテンツを自律的に創出します。

具体的には、自然な対話が可能な大規模言語モデル（LLM）による文章作成や要約、指示通りの画像を瞬時に描く画像生成 AI、作曲や音声クローニング、さらにはプログラムコードの自動生成まで、その応用範囲は多岐にわたります。この技術革新は、コンテンツ制作のあり方や、ビジネスにおける作業効率を根本から変えるパラダイムシフトを引き起こしており、AI と人間の\*\*協働（ヒューマン・イン・ザ・ループ）による新たな価値創造が期待されています。一方で、著作権や偽情報（フェイク）\*\*といった倫理的・社会的な課題への対応も不可欠となっています。

### ChatGPT からの評価

評価観点	
内容の正確性(5点)	5
文脈理解の適切さ(5点)	5
構成の明瞭さ(5点)	4
表現の自然さ(5点)	4
指示遵守度(5点)	5
合計点(25点満点)	23

総合評価（約 100 字）

生成 AI の技術的背景から社会的影響までを高い精度と論理性で整理しており、学術用途

に適した完成度の高い出力である。特に論述文では、専門性と可読性のバランスが優れている。

Claude からの評価

評価観点	
内容の正確性(5点)	5
文脈理解の適切さ(5点)	5
構成の明瞭さ(5点)	4
表現の自然さ(5点)	4
指示遵守度(5点)	5
合計点(25点満点)	23

Gemini 【23点/25点】★良好

技術的な深さと正確性が際立つ。GAN、Transformer、Diffusion Model など具体的なモデル名を明記し、専門性の高い説明を実現。ただし、論述文での情報密度が高すぎる点と、太字強調の多用がやや過剰で、読みやすさをわずかに損なう。

## 3-2.文脈理解力

文脈理解力を測定するための手法として、生成 AI になぞなぞを作成させ、その提示された問題に対して正しい解答を導き出せるかを検証した。なぞなぞは、表面的な語句の意味理解だけでなく、前提条件や暗示された情報を適切に解釈する能力が求められる課題である。そのため、本調査では、生成 AI が問題文の文脈をどの程度正確に把握し、論理的推論を通じて妥当な解答に到達できるかを評価する指標として、なぞなぞ課題を採用した。

### 3-2-1.ChatGPT からの出題

以下では、ChatGPT になぞなぞを作成してもらい、それを Claude、Gemini に解かせ、相互評価を行う。

なぞなぞ

私は

持つと増え、

分けると減り、

隠すと強く、

語ると弱くなる。

目には見えず、

形もないが、

人の行動を大きく左右する。

私は何でしょう？

答えは

「秘密」です。

理由：

持つと増える：秘密を抱えるほど、別の秘密が生まれやすい

分けると減る：誰かに話すと、秘密ではなくなる

隠すと強い：隠している間は力（影響力）を持つ

語ると弱い：話した瞬間に価値が下がる

目に見えず形がないが、人の行動を左右する：まさに秘密の性質

3-2-1-1.claude の答え

このなぞなぞの答えは「秘密」ですね。

それぞれのヒントが秘密の性質を表しています：

持つと増え - 秘密を抱えると心の負担が増える

分けると減り - 誰かに打ち明けると心が軽くなる

隠すと強く - 秘密にしておくと言響力や重みが増す

語ると弱くなる - 話してしまうともう秘密ではなくなり、力を失う

そして秘密は目に見えないものですが、人の行動や判断に大きな影響を与えますね。

とても巧みに作られたなぞなぞです!

### ChatGPT からの評価

#### 文脈理解力の評価結果

判定：正答

#### 判断理由

提示された回答は、答えとして「秘密」を正しく導いており、さらに各ヒントと秘密の性質との対応関係を的確に説明できている。

「持つと増え」「分けると減る」を心理的負担の増減として解釈しており、文脈に即している

「隠すと強く」「語ると弱くなる」を秘密性と影響力の関係として正しく理解している

「目に見えないが行動に影響する」という抽象的条件についても、秘密の本質として適切に説明している

単に正解語を示すだけでなく、問題文の比喩的・抽象的条件を総合的に理解した説明がなされているため、文脈理解は十分であると判断できる。

### Gemini からの評価

#### 文脈理解力の評価

ご提示いただいたなぞなぞとその答え（「秘密」）、そして解説（ヒントとの対応）を総合的に見ると、\*\*「正答」\*\*の条件を完璧に満たしていると評価できます。

## 3-2-1-2. Gemini の答え

これは\*\*秘密（ひみつ）\*\*ですね！

### 🤖 答えの理由

持つと増え、分けると減り：秘密を多く持てば持つほど（知れば知るほど）増えますが、誰かに話して分け合う（共有する）と、その秘密の力や価値は減っていきます。

隠すと強く、語ると弱くなる：隠している間は「秘密」として強力ですが、語って公になるとその力は弱まります。

目には見えず、形もないが、人の行動を大きく左右する：概念であり、物理的な形はありませんが、秘密の存在は人の言動や決断に大きな影響を与えます。

### ChatGPT からの評価

#### 文脈理解力の評価結果

判定：正答

#### 判断理由

回答は、なぞなぞの答えとして「秘密（ひみつ）」を正しく導き出しており、さらに各ヒントについても文脈に即した説明がなされている。

\*\*「持つと増え、分けると減る」\*\*を、秘密の共有・非共有による価値や影響力の変化として解釈している

\*\*「隠すと強く、語ると弱くなる」\*\*について、秘密性の有無による力の強弱を適切に説明している

\*\*「目に見えず、形もないが人の行動を左右する」\*\*という抽象的条件についても、概念的性質として正確に理解している

すべての条件を矛盾なく満たしており、比喩的表現を含む問題文の文脈を総合的に理解した回答であるため、「正答」と評価できる。

Claude からの評価

総合評価

完全な正答。文脈理解力は非常に高く、なぞなぞの意図を完璧に読み取っています。

### 3-2-2.Claude からの出題

以下では、Claude になぞなぞを作成してもらい、それを ChatGPT、Gemini に解かせ、相互評価を行う。

面白いなぞなぞをいくつか作ってみますね!

**なぞなぞ 1(中級)** 「毎日会っているのに、一度も顔を見たことがない。でも、その人がいないと私は何もできない。これは誰?」

<答え:自分自身(鏡を見ない限り、自分の顔は直接見られない)>

#### 3-2-2-1. ChatGPT の答え

**\*\*答え：空気（くうき）\*\***

**\*\*理由：\*\***

私たちは毎日ずっと空気と「一緒に」いるのに、顔（姿）を見ることはできません。

しかし空気がなければ、呼吸もできず、生きることも、結果的に何もできません。

シンプルだけど本質を突く、定番で良質ななぞなぞですね。

Claude からの評価

評価結果

提示された答え：空気（くうき）

評価：誤答

結論

「空気」という答えは、なぞなぞの本質的な仕掛けである「人物を指す言葉遣い」を見落としており、条件を部分的にしか満たしていないため、誤答と評価します。

ただし、発想としては創造的であり、条件の大部分を満たそうとする試みは認められます。

Gemini からの評価

文脈理解力の評価

答え「空気」は、「自分自身」や「鏡」と比較して、「その人がいないと私は何もできない」という条件を、生命活動の根幹に関わるものとして最も本質的に満たしています。

「誰？」という問いに対する答えとしては、前の二つ（自分自身、鏡）よりも比喩的な解釈が必要になりますが、なぞなぞとして成立しています。なぞなぞの条件をすべて満たしているため、これも\*\*「正答」\*\*として評価できます。

### 3-2-2-2.Gemini の答え

これは\*\*鏡\*\*です。

\* \*\*毎日会っているのに、一度も顔を見たことがない。\*\*: 鏡はあなたの姿を映しますが、鏡自体の「顔」（裏側など）をあなたは見ていません。

\* \*\*でも、その人がいないと私は何もできない。\*\*: 鏡がないと、自分の\*\*顔や身だしなみ\*\*をチェックすることが難しくなり、何かをするときに困ります。

ChatGPT からの評価

文脈理解力の評価結果（なぞなぞ1）

判定：部分的理解

判断理由

回答として提示された「鏡」は、なぞなぞの一部の条件とは対応しているものの、問題文全体の文脈を完全には満たしていない。

「毎日会っているのに、一度も顔を見たことがない」

→ 鏡自体の「顔」を見ていないという解釈は成り立つため、この点は条件を部分的に満たしている。

「いないと私は何もできない」

→ 身だしなみ確認など特定の行為には当てはまるが、「何もできない」という強い表現を十分に説明できていない。

なぞなぞ全体としては、存在そのものに依存している対象を問う構造であり、「鏡」では依存関係が限定的である。

そのため、文脈の核心を一部捉えてはいるが、条件を総合的・抽象的に解釈し切れていないと判断し、「部分的理解」と評価する。

Claude からの評価

結論

「鏡」という答えは、なぞなぞの条件を大幅に誤解・曲解しており、特に「その人」という人物限定の表現と、「何もできない」という極端な依存性の条件を満たしていません。

評価：誤答

ただし、創造的な発想を試みた点は認められますが、文脈の本質的な理解には至っていません。

### 3-2-3.Gemini の調査結果

以下では、Gemini になぞなぞを作成してもらい、それを ChatGPT、Claude に解かせ、相互

評価を行う。

### ? 難しいなぞなぞ ?

私は声を持っていますが、話すことはできません。耳を持っていますが、聞くことはできません。いつもそこにはありますが、触れることはできません。私が誰であるかを知るには、あなた自身の助けが必要です。

さて、私は一体誰でしょう？

### 💡 なぞなぞの答え 💡

なぞなぞの答えは...

あなたの「エコー（反響／こだま）」です。

### 解説

「私は声を持っていますが、話すことはできません。」

あなたの声を受け取り、それを返します（声を持っている）。自分で意味のある言葉を話すわけではありません。

「耳を持っていますが、聞くことはできません。」

あなたの音を聞き取る能力があるわけではなく、ただ音波を反射しているだけです。

「いつもそこにはありますが、触れることはできません。」

音の現象であり、物質ではないため、触れることはできません。

「私が誰であるかを知るには、あなた自身の助けが必要です。」

エコーを発生させるには、あなたが声を発する必要があります。

### 3-2-3-1.ChatGPT の答え

答えは—— 「こだま (エコー)」 です。

理由：

声を「持っている」ように聞こえるが、自分で話しているわけではない

耳があるように返事をするが、実際に聞いているわけではない

いつもそこに“存在”する現象だが、触れることはできない

そして何より、あなたが声を出してはじめて存在が分かる（あなた自身の助けが必要）

とても詩的で、哲学的な良いなぞなぞですね。

### Claude からの評価

評価結果提示された答え：こだま (エコー)

評価：正答 ✓ 評価理由この答えは、なぞなぞのすべての条件を完璧に満たしています。

### Gemini からの評価

結論

私の文脈理解力、および提示した解答の論理的な整合性は、\*\*「正答」\*\*として評価されます。

### 3-2-3-2.Claude の答え

これは面白い謎かけですね!

答えは\*\*「こだま(やまびこ)」\*\*だと思います。

理由を説明しますと:

- \*\*声を持っているが話せない\*\*：こだまは声を返しますが、自分からは話しません

- \*\*耳を持っているが聞けない\*\*：音を受け取りますが、実際に「聞く」わけではありません

- \*\*いつもそこにあるが触れない\*\*：音の反響現象なので、物理的に触れることはできません

- \*\*あなた自身の助けが必要\*\*：あなたが声を出さないと、こだまは存在しません

または、もう一つの可能性として\*\*「あなた自身の声」\*\*という答えもあり得ますね。自分の声は自分では正確に聞けませんし(録音された声と違って聞こえる)、他の人の助けがないと客観的には認識できません。

## ChatGPT からの評価

文脈理解力の評価結果 (難しいなぞなぞ)

判定：正答

判断理由

提示された主回答「こだま (やまびこ)」は、なぞなぞの条件をすべて一貫して満たしている。

声を持っているが話せない

→ こだまは他者の声を反響させる現象であり、自発的に話すことはない

耳を持っているが聞けない

→ 音を受け取るが、知覚や理解を行う主体ではないという比喩的解釈が成立している

いつもそこにあるが触れない

→ 音の反響という非物質的現象であり、触れられない点と一致する

あなた自身の助けが必要

→ 声を出す行為がなければ、こだまは生じない

また、補足として挙げられている「あなた自身の声」についても、一定の文脈的妥当性はあるが、条件との対応関係の明確さ・一貫性の点で、「こだま」の方がより適合度が高い。

以上より、問題文の比喩表現と因果関係を総合的に理解した回答であり、文脈理解は十分であると判断し、「正答」と評価する。

Gemini からの評価

✅ 文脈理解力の評価

ご提示いただいた二つの解答「こだま（やまびこ）」と「あなた自身の声」は、いずれもなぞなぞの全ての条件を論理的な解釈で満たしています。なぞなぞが持つ比喩や擬人化の文脈を正確に理解されているため、\*\*「正答」\*\*として評価いたします。

### 3-3.翻訳の精度

翻訳の精度を測定するための手法として以下の、

『桃太郎』は、江戸時代から語り継がれる有名な日本の昔話の一つです。桃から生まれた男の子の桃太郎は、成長するとお爺さんとお婆さんから黍団子をもらって鬼退治に出発します。鬼ヶ島に向かう途中、犬・猿・雉を家来にし、全員で協力して鬼を退治するというお話です。[2]

を使用して、翻訳が正しく行えているかを検証した。

手順としては、①日本語→英語、②英語→日本語を行い、文章にどれほどのずれがあるかを評価する。

#### 3-3-1.ChatGPT の調査結果

##### ①日本語→英語

“\*Momotaro\* is one of Japan’s most famous folktales, passed down since the Edo period. A boy named Momotaro is born from a peach, and when he grows up, he receives millet dumplings from an old man and an old woman and sets out to defeat the ogres. On his journey to Ogre Island, he gains a dog, a monkey, and a pheasant as his companions, and together they cooperate to defeat the ogres.”

## ②英語→日本語

『桃太郎』は、江戸時代から語り継がれてきた日本で最も有名な昔話の一つです。桃から生まれた男の子・桃太郎は、成長するとお爺さんとお婆さんから黍団子をもらい、鬼退治に出発します。鬼ヶ島へ向かう途中で、犬・猿・雉を仲間にし、力を合わせて鬼を退治します。

## 総合評価

本事例では、意味保持率は非常に高く、翻訳精度は良好であると評価できる。一方で、英語を経由することにより、文化的・語感的なニュアンス（上下関係や表現の硬さ）が一部簡略化される傾向が確認された。

## 3-3-2.Claude の調査結果

### ①日本語→英語

"Momotaro" is one of Japan's famous folktales that has been passed down since the Edo period. Momotaro, a boy born from a peach, grows up and receives millet dumplings from an old man and woman before setting out to defeat ogres. On his way to Onigashima (Demon Island), he recruits a dog, a monkey, and a pheasant as his retainers, and together they all cooperate to defeat the ogres.

### ②英語→日本語

「桃太郎」は江戸時代から語り継がれてきた日本の有名な昔話の一つです。桃から生まれた男の子・桃太郎は成長し、おじいさんとおばあさんからきびだんごをもらって鬼退治に出発します。鬼ヶ島へ向かう途中で、犬、猿、雉を家来にし、みんなで協力して鬼を退治します。

## 評価結果（簡潔）

翻訳往復の結果、原文と再翻訳文の間で物語の主要な意味内容は保持されていることが確認された。一方、語彙や表現には一部言い換えが見られたが、いずれも日本語と

しての自然さを高めるための変化であり、意味の逸脱は認められない。以上より、本翻訳の精度は高いと評価できる。

### 3-3-3.Gemini の調査結果

#### ①日本語→英語

Momotaro is one of the most famous Japanese folktales, passed down since the Edo Period.

The story is about a boy named Momotaro, who was born from a giant peach. Once he grows up, he receives kibi dango (millet dumplings) from the old man and old woman and sets off on a journey to defeat the ogres (oni).

On his way to Onigashima (Ogre Island), he befriends a dog, a monkey, and a pheasant, making them his retainers. They all work together to defeat the ogres.

#### ②英語→日本語

桃太郎は、江戸時代から伝わる最も有名な日本の民話の一つです。

この物語は、巨大な桃から生まれた桃太郎という名の少年についての話です。彼は成長すると、おじいさんとおばあさんからきびだんごをもらい、鬼を退治するための旅に出ます。

鬼ヶ島へ向かう途中、彼は犬、猿、雉（きじ）と友達になり、彼らを家来（けらい）にします。彼らは皆で協力して鬼を打ち破ります。

#### 評価結果（簡潔）

翻訳往復の結果、物語の主要な意味内容および構造は再翻訳文において適切に保持されていることが確認された。一方で、「巨大な桃」「友達になる」など、一部の語彙や表現に言い換えが見られたが、いずれも内容理解に影響を与えるものではない。以上より、本翻訳は意味保持の観点から高い精度を有すると評価できる。

## 3-4.各生成 AI への自己評価

### 3-4-1. ChatGPT への自己評価

#### 1. 文章生成の可否および生成形式

本調査において、ChatGPT は 50 字・200 字・400 字という異なる文字数制約をすべて厳密に遵守し、内容の段階的深化を伴う文章生成を安定して実現できた。要約文では定義を簡潔に示し、説明文では特徴と活用例を補足し、論述文では利点と課題を対比的に論じる構成となっており、形式制御能力と構成力の高さが確認された。

#### 2. 文脈理解力

なぞなぞ課題においては、比喩表現や暗示的条件を総合的に解釈し、適切な解答と論理的説明を提示できた。一方で、解釈の幅が広い問題では、文脈の核心よりも一般的連想を優先する傾向が見られ、誤答と評価される事例も確認された。総じて文脈理解力は高いが、問題意図の厳密な読み取りには改善の余地がある。

#### 3. 翻訳の精度

日本語—英語—日本語の往復翻訳において、物語の構造と意味内容は高い水準で保持されていた。文化的ニュアンスや語感については簡略化が生じる場合があるものの、情報の欠落や誤解を生む変化は見られず、実用的観点から高い翻訳精度を有する評価できる。

### 3-4-2. Claude への自己評価

#### 1. 文章生成の可否および生成形式

Claude は、生成 AI の定義から応用例までを網羅的かつ自然な日本語表現で記述できた。説明文・論述文では指示字数を遵守し、読みやすさと正確性の両立が達成されている。一方、要約文において字数制御が完全ではなく、形式面での一貫性に課題が残った。

#### 2. 文脈理解力

なぞなぞ課題では、問題文に含まれる比喩や擬人化表現を的確に捉え、論理的整合性の高い解答を導くことができた。特に、抽象概念を問う問題に対して、条件との対応関係を明示しながら説明できており、高い文脈理解力を有していると評価できる。

### 3. 翻訳の精度

往復翻訳の結果、物語の主要要素や因果関係は保持されており、意味の逸脱は認められなかった。語彙選択においては自然な日本語表現を優先する傾向が見られ、結果として可読性の高い再翻訳文が生成されている。翻訳精度は全体として高水準である。

## 3-4-3. Gemini への自己評価

### 1. 文章生成の可否および生成形式

Gemini は、GAN や Transformer など具体的な技術用語を用い、専門性の高い文章生成を行うことができた。特に論述文では情報量と技術的正確性が際立っている。一方で、情報密度が高くなりすぎる傾向や、強調表現の多用により、読みやすさがやや低下する点が課題として挙げられる。

### 2. 文脈理解力

なぞなぞ課題では、比喩的・抽象的な条件を柔軟に解釈し、複数の解答可能性を提示するなど、高い推論能力を示した。ただし、問題文の「人物性」や表現上の制約よりも概念的解釈を優先する傾向があり、評価者によって正誤判断が分かれる場面も見られた。

### 3. 翻訳の精度

翻訳においては、固有名詞への補足説明や文化的背景の明示を行い、理解しやすさを重視した出力が特徴的であった。一部に語彙の追加や言い換えが見られるものの、意味内容や物語構造への影響はなく、意味保持の観点から高い翻訳精度を有すると自己評価できる。

## 4. 各生成 AI における課題

### 4-1. ChatGPT の課題

ChatGPT は、文章生成における形式制御能力、文脈理解の安定性、翻訳精度のいずれにおいても高い評価を得ており、総合的な性能は非常に高い。しかし一方で、文脈理解課題、とくになぞなぞのような解釈の幅が広い問題において、問題文の核心的意図よりも一

般的・汎用的な連想に基づく解答を選択する傾向が確認された。その結果、一部の課題では評価者間で「誤答」と判定される事例が生じた。

このことから、ChatGPT の課題としては、創造的な発想力と問題文に内在する制約条件の厳密な読み取りとのバランスが挙げられる。今後は、問題文中の表現（例：「誰」という人物限定表現や、「何もできない」といった強い条件）に対する重み付けをより精緻に行うことで、文脈理解の精度向上が期待される。

## 4-2. Claude の課題

Claude は、自然で読みやすい文章表現と、抽象的・比喩的文脈の理解において高い能力を示した。なぞなぞ課題では、問題文の意図を的確に捉え、条件との対応関係を丁寧に説明する姿勢が評価されている。また翻訳においても、日本語としての自然さを重視した出力が特徴的であった。

一方で、文章生成課題においては、字数制限という形式的制約の遵守に一貫性を欠く場面が見られた。特に要約文では、内容の充実を優先するあまり、指定された字数を超過する傾向が確認されている。この点から、Claude の課題は、内容の豊かさと形式遵守との両立にあるといえる。学術用途や試験的評価においては、形式要件を厳密に満たす能力がより強く求められるため、今後の改善が望まれる。

## 4-3. Gemini の課題

Gemini は、生成 AI の技術的背景に関する専門性の高い知識を積極的に用い、情報量の多い文章生成を行う点に大きな強みがある。論述文では、GAN や Transformer、Diffusion Model など具体的なモデル名を挙げることで、学術的深さを備えた説明を実現していた。

しかしその反面、情報密度が過度に高くなる傾向があり、文章全体の可読性や構成の明瞭さがやや損なわれる場面が見られた。また、文脈理解課題では、比喩的解釈を広く許容する姿勢から、問題文の表現上の制約（人物性・依存関係の強さなど）を相対的に軽視する傾向があり、評価者間で正誤判断が分かれる結果となった。

したがって Gemini の課題は、専門性の高さを維持しつつ、読み手や設問意図に応じて情報量と解釈範囲を適切に調整する能力にあると考えられる。

## 4-4. 課題の総括

以上より、3つの生成 AI はいずれも高い性能を有しているが、それぞれに異なる課題が存在することが明らかとなった。

ChatGPT は「安定性の高さ」と引き換えに、文脈解釈の厳密さに課題を残すこと、Claude は「表現の自然さ」と引き換えに、形式制御の一貫性に課題を抱えること、Gemini は「専門性の深さ」と引き換えに、可読性と解釈の収束性に課題が見られることである。

これらの結果は、生成 AI の性能評価において「正確性」だけでなく、「設問意図への適合性」や「利用目的との相性」を考慮する必要性を示唆している。

## 5. おわりに

### 5-1. まとめ

本研究では、ChatGPT・Claude・Gemini の3つの生成 AI を対象に、文章生成能力、文脈理解力、翻訳精度の観点から評価を行った。まず、各生成 AI に相互評価を行わせることで、他モデルから見た強みと弱みを明らかにした。さらに、その結果を踏まえて各生成 AI の自己評価を行い、相互評価と自己認識の一致点および相違点を整理した。これにより、生成 AI の性能を単一の評価視点に依存せず、多角的に把握することが可能となった。

文章生成能力に関しては、ChatGPT は字数制約や生成形式の遵守に優れ、要約文・説明文・論述文といった異なる文章形式を安定して生成できた。一方、Claude は自然で読みやすい文章表現に強みを持ち、論理の流れが滑らかな文章を生成する傾向が見られた。Gemini は専門用語や技術的要素を多く含む文章を生成でき、情報量と専門性の高さが特徴的であった。

文脈理解力の評価として用いたなぞなぞ課題では、3者ともに比喩や暗示を含む文章を一定程度理解できたが、解釈の方向性には差があった。ChatGPT は一般的な連想を基に解答する傾向があり、Claude は問題文の条件を丁寧に整理した論理的解釈を行った。

Gemini は抽象的・概念的な解釈を行い、複数の可能性を提示する場面が多く見られた。これらの結果から、文脈理解力は単純な正誤だけでなく、評価者の基準によって評価が左右される側面があることが示唆された。

翻訳精度については、いずれの生成 AI も物語の構造や意味内容を大きく損なうことなく翻訳できており、実用レベルでは十分に高い精度を有していると評価できる。ただし、ChatGPT は意味保持を優先し簡潔な翻訳を行う傾向があり、Claude は自然な日本語表現を重視する傾向、Gemini は補足説明を加えることで理解を助ける傾向が見られた。

以上より、生成 AI の性能は一様ではなく、利用目的に応じて適切な AI を選択する必要があることが明らかとなった。

## 5-2.総合ランクづけ（本研究に基づく）

本研究の評価結果を総合すると、以下の順位づけが妥当であると考えられる。

第 1 位：ChatGPT

（形式遵守力・安定性・総合バランスに優れる）

第 2 位：Claude

（文脈理解力と自然な日本語表現が高評価）

第 3 位：Gemini

（専門性は高いが、可読性と形式面に課題）

ただし、この順位は本研究で設定した評価項目と課題内容に基づくものであり、用途や評価基準が異なれば順位が変動する可能性がある。

## 5-3. 今後の課題

今後の課題として、第一に評価者が人間ではなく生成 AI である点による評価の主観性が挙げられる。第二に、なぞなぞや翻訳といった課題の難易度や内容によって評価結果が左右される可能性がある点である。第三に、評価指標を定量化（点数化）することで、より

客観的かつ再現性の高い比較が可能になると考えられる。今後は人間評価との併用や評価課題の多様化を行うことで、生成 AI 比較の妥当性をさらに高める必要がある。

## 6.参考文献

[1] 野村総合研究所（2023）「日本の ChatGPT 利用動向（2023 年 4 月時点）」

[https://www.nri.com/jp/knowledge/report/20230526\\_1.html](https://www.nri.com/jp/knowledge/report/20230526_1.html)

[2] 昔話『桃太郎』のあらすじ・内容解説・感想 | おすすめ絵本

<https://folktalesjpn.com/folktale/0009/>