日本語創造性ベンチマークの構築

福田創 [1]　小川隼斗 [1]　堀尾海斗 [1]　河原大輔 [1]　柴田知秀 [2]
[1] 早稲田大学　[2]SB Intuitions 株式会社
so.fukuda@akane.waseda.jp　cookie3120@ruri.waseda.jp
kakakakakakaito@akane.waseda.jp　dkw@waseda.jp
tomohide.shibata@sbintuitions.co.jp

概要

大規模言語モデル（LLM）の創造性を評価するために，Japanese Creativity Questions (JCQ)，Divergent Association Task (DAT)，そして Story Alteration Task (SAT) という 3 つのベンチマークを構築する．JCQ では，LLM を用いて創造性を包括的に評価する．一方，DAT と SAT では，埋め込みを用いて，創造的能力の一面を測定する．さらに，JCQ と DAT，および JCQ と SAT の間の相関を分析する．JCQ は網羅的な評価ができるが，比較的時間とコストがかかる．一方，DAT と SAT は網羅性が低いが，迅速に評価できる．

1　はじめに

創造性は，人類の進歩と発展を支えてきた重要な能力である．芸術表現や科学的発見，社会問題の解決など，創造的思考は人間の活動の中核を担ってきた．近年，大規模言語モデル（LLM）の発展により，AI システムにも文章生成や問題解決において，人間の創造的活動を支援・拡張する可能性が見出され，活発な研究が行われている [1, 2, 3, 4]．人間と LLM の両者にとって，創造性は，複雑化する現代社会の課題に対応し，新たな価値を生み出すための本質的な要素となっている．

人間の創造性を評価するためのテストとして，Torrance Test of Creative Thinking (TTCT) がある．これは言語テストと図形テストからなる自由記述式のテストであり，例えば，「電球の通常でない使い方をできるだけ多く列挙してください．」といった問題がある．回答を評価する際は，「流暢性」，「柔軟性」，「独創性」，「精緻性」の 4 つの指標がよく使われる．これらの 4 指標は，他の多くの創造性研究でも一般的に採用されている [5, 6, 7]．TTCT は心理学の分野で広く用いられており，多くの人々の創造性を測ることができる優れたテストであるとされている [8]．

また，Divergent Association Task (DAT) という創造性テストがあり，人間を対象とした研究が行われている [9]．DAT は，できるだけ意味的に離れた無関係な単語を列挙するタスクであり，単語間の意味的距離が大きいほど高いスコアが得られる．また，この研究では Alternative Uses Task (AUT) という創造性テストも行われている．これは，例えば「新聞」や「電球」のような一般的な物の使い方をできるだけ多く列挙させるタスクである．DAT のスコアと AUT における「流暢性」および「独創性」のスコアが相関することが示されている．

英語では，TTCT の言語テストを基にしたテストを作成し，OpenAI の GPT-4 を評価者として LLM の創造性を測定した研究 [10] が存在する．日本語においては，LLM の創造性を評価するためのベンチマークは現在のところ存在しない．

本研究では，日本語における LLM の創造性を目的に応じて多角的または効率的に測るため，3 つのベンチマークを構築し，いくつかの LLM を評価する．1 つ目は Japanese Creativity Questions (JCQ) である．Zhao らの研究 [10] に倣い，TTCT の言語テストを基に作成する．7 つのタスクから構成し，評価には 4 つの指標を用いる．2 つ目は DAT である．3 つ目は Story Alteration Task (SAT) である．これは物語を改変させ，元の物語とどれだけ異なるかを測るテストである．JCQ の評価では強力な LLM に評価させる．一方，DAT と SAT の評価では埋め込みを用いる．JCQ は創造性を網羅的に評価できるが，評価には時間とリソースが必要となる．一方，DAT と SAT は埋め込みを活用することで，創造性の特定の側面を迅速かつ手軽に測定できる利点がある．これにより，必要に応じて網羅性を重視した評価と迅速性を重視した評価を使い分けて，LLM の創造性を測ることができる．

表 1　JCQ のタスクの定義と問題例

タスク	定義	問題例
非通常使用	一般的な物体の珍しい使い方や多様な使い方を考えるタスク.	電球の通常でない使い方をできるだけたくさん挙げてください.
結果	普通ではない，または仮説的な状況における結果や影響を予測するタスク.	もしも世界中で 24 時間インターネットが使えなくなったら、社会や日常生活にどのような影響が生じるでしょうか？
仮定	仮定の，しばしば空想的なシナリオとその含意を考えるタスク.	物を消滅させる力を手に入れました. あなたなら何を消しますか？できるだけ多くのアイデアを挙げてください.
状況	与えられた状況に対応するタスク.	もしも重力の向きが反転したら、あなたはどうやって地上で生き残りますか？
一般的問題	多くの人にとって身近で日常的な問題に対し，解決策を生み出すタスク.	冷蔵庫の中身を効率的に管理する方法を提案してください.
改善	既存の物やアイデアを改良したり，修正したりするタスク.	一般的なベッドをより快適にする方法をできるだけ多く挙げてください.
想像的物語	与えられたプロンプトで物語を作るタスク.	「月の裏の図書館」というタイトルで物語を作ってください.

表 2　JCQ における 4 つの指標の定義

指標	定義
流暢性	与えられた質問に応じて，関連するアイデアを数多く生み出す能力. 本質的にはアイデアの量を測定する.
柔軟性	アイデアを生み出すことができるカテゴリーの多様性. 代替案を考えたり，あるクラスや視点から別のクラスや視点に移行したり，与えられた問題や課題にさまざまな角度からアプローチしたりする能力.
独創性	生み出されたアイデアの独自性. 独自なアイデアとは，普通とは異なる珍しい，または型破りなアイデアのこと.
精緻性	アイデアを発展させ，洗練させ，装飾する能力. 細部を付け加え，ニュアンスを発展させ，基本的なコンセプトをより入り組んだ，あるいは複雑なものにすることを含む.

2　日本語創造性ベンチマークの構築

2.1　Japanese Creativity Questions (JCQ)

JCQ は，Zhao らの研究 [10] に倣い，創造性を包括的に測ることを目的として作成した. OpenAI の GPT-4o，o1-preview，Anthropic の Claude 3.5 Sonnet と対話しながら，Zhao らの研究 [10] で使用された 7 つのタスクでそれぞれ 100 問ずつ，合計 700 問の日本語問題を作成した. タスクの定義と問題例を表 1 に示す. 回答例を付録の表 11 に示す.

評価は GPT-4o のような強力な LLM を用いて行う. 強力な LLM を用いた評価の有効性は既に示されている [11]. 具体的にはモデルの回答を「流暢性」，「柔軟性」，「独創性」，「精緻性」の 4 つの指標において 1〜5 のスケールで評価する. 各指標は Zhao らの研究 [10] に倣い，表 2 のように定義する.

2.2　Divergent Association Task (DAT)

DAT は，Olson らの研究 [9] で用いられた，できるだけ意味の離れた単語を 10 個挙げるテストである. 列挙した単語が意味的に離れているほど創造性が高いとみなす. 回答例を付録の表 12 に示す.

評価にはモデルが列挙する 10 個の単語それぞれの埋め込みを用いる. それらの全組み合わせのコサイン距離（1 – コサイン類似度）の平均を 1 回の試行のスコアとする. この試行を複数回行わせ，その平均スコアをモデルのスコアとする.

2.3　Story Alteration Task (SAT)

SAT は，物語を特定の指示で書き換えるテストである. 書き換えた物語が元の物語と異なるほど創造性が高いとみなす. 回答例を付録の表 13 に示す.

評価には元の物語の埋め込みとモデルが出力する物語の埋め込みを用いる. 2 つの埋め込みのコサイン距離を求め，複数の物語における平均をモデルのスコアとする.

3　LLM の創造性評価実験

構築した 3 つのベンチマークを用いて 5 つの LLM の創造性を評価する.

3.1　実験設定

以下の 5 つのモデルに回答させる. 温度は 1 とする.

- gpt-4o-2024-08-06[1)] (GPT-4o)
- claude-3-5-sonnet-20241022[2)] (Claude 3.5 Sonnet)
- calm3-22b-chat[3)]
- llm-jp-3-13b-instruct[4)]

1）https://platform.openai.com/docs/models#gpt-4o
2）https://docs.anthropic.com/en/docs/about-claude/models#model-names
3）https://huggingface.co/cyberagent/calm3-22b-chat
4）https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

表 3　JCQ 結果：モデルとタスク

	非通常使用	結果	仮定	状況	一般的問題	改善	想像的物語
GPT-4o	**3.97**	3.69	3.83	3.28	3.48	**4.01**	3.25
Claude 3.5 Sonnet	3.73	3.42	3.80	3.08	**3.61**	3.80	2.93
calm3-22b-chat	3.84	3.92	**3.91**	**3.73**	3.45	4.00	**3.50**
llm-jp-3-13b-instruct	3.08	**3.92**	3.52	3.69	3.00	3.64	3.01
Swallow-8B-Instruct	3.28	3.33	3.39	2.80	3.08	3.45	2.54

表 4　JCQ 結果：モデルと指標

	流暢性	柔軟性	独創性	精緻性	平均
GPT-4o	4.10	**4.28**	2.73	3.47	3.64
Claude 3.5 Sonnet	**4.29**	4.04	2.73	2.87	3.48
calm3-22b-chat	4.16	4.18	**2.87**	**3.86**	**3.76**
llm-jp-3-13b-instruct	3.74	3.79	2.65	3.45	3.41
Swallow-8B-Instruct	3.91	3.45	2.34	2.79	3.12

表 5　JCQ 結果：タスクと指標

	流暢性	柔軟性	独創性	精緻性	平均
非通常使用	4.50	4.13	2.92	2.78	3.58
結果	4.00	4.31	2.67	3.64	3.65
仮定	4.58	4.43	2.64	3.11	3.69
状況	3.30	4.03	2.57	3.38	3.32
一般的問題	3.98	3.85	2.01	3.46	3.32
改善	4.71	4.51	2.72	3.17	3.78
想像的物語	3.22	2.36	3.12	3.49	3.05

- Llama-3.1-Swallow-8B-Instruct-v0.1[5)] (Swallow-8B-Instruct)

JCQ では，gpt-4o-2024-08-06 を用いて評価する．評価プロンプトを付録の表 16 に示す．

DAT では，モデルの平均スコアを求めるための試行回数を 10 回とする．英単語が混ざった場合や，単に「単語 1」などと出力した場合は，その試行をやり直す．プロンプトを付録の表 14 に示す．埋め込みモデルは GLuCoSE-base-ja-v2[6)] を用いる．

SAT では，113 個の童話を元の物語とする．これらは童話サイト [12] から選択した童話を，gpt-4o-2024-05-13[1)] で 200〜400 文字程度に要約したものである．書き換えの指示は，童話を現代風に作り変えることを設定する．プロンプトを付録の表 15 に示す．埋め込みモデルは simcse-ja-bert-base-clcmlp[7)] を用いる．

3.2　結果

3.2.1　Japanese Creativity Questions (JCQ)

各モデルとタスクにおける全指標の平均スコアを表 3 に示す．全体的に改善タスクが得意で，創造的物語タスクが苦手であるなどの特徴があった．

5)　https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1
6)　https://huggingface.co/pkshatech/GLuCoSE-base-ja-v2
7)　https://huggingface.co/pkshatech/simcse-ja-bert-base-clcmlp

表 6　DAT 結果

	スコア
GPT-4o	0.528 ± 0.007
Claude 3.5 Sonnet	**0.530 ± 0.012**
calm3-22b-chat	0.507 ± 0.006
llm-jp-3-13b-instruct	0.480 ± 0.022
Swallow-8B-Instruct	0.502 ± 0.008

表 7　SAT 結果

	スコア
GPT-4o	0.526 ± 0.023
Claude 3.5 Sonnet	**0.579 ± 0.025**
calm3-22b-chat	0.458 ± 0.027
llm-jp-3-13b-instruct	0.219 ± 0.027
Swallow-8B-Instruct	0.193 ± 0.026

各モデルと指標における全タスクの平均スコアを表 4 に示す．モデル間の流暢性の差と独創性の差に対して，精緻性の差が大きいなどの特徴があった．

各タスクと指標における全モデルの平均スコアを表 5 に示す．想像的物語タスクにおける柔軟性，一般的問題タスクにおける独創性が，他のタスクと比較して突出して低いなどの特徴があった．

3.2.2　Divergent Association Task (DAT)

モデルごとのスコア（95%信頼区間）を表 6 に示す．GPT-4o と Claude 3.5 Sonnet といった強力であるとされている 2 つのモデルのスコアが高かった．

3.2.3　Story Alteration Task (SAT)

モデルごとのスコア（95%信頼区間）を表 7 に示す．Claude 3.5 Sonnet のスコアが突出して高かった．2 番目にスコアが高かったのは GPT-4o であり，DAT と同様に，強力であるとされる 2 つのモデルのスコアが優れていた．

4　分析

4.1　JCQ の評価における GPT-4o と人の相関

JCQ に対する回答の一部を人手で評価した．各タスク 15 個ずつ，合計 105 個の回答に対して，大学生 3 人で協議しながら GPT-4o と同様に評価した．

表 8　JCQ の評価における GPT-4o と人の相関

	流暢性	柔軟性	独創性	精緻性	平均
非通常使用	1.000	0.222	0.208	0.613	0.570
結果	0.688	0.668	0.696	0.745	0.791
仮定	0.964	0.623	0.733	0.683	0.755
状況	0.299	0.619	0.551	0.174	0.707
一般的問題	0.814	0.640	0.539	0.494	0.639
改善	0.868	0.552	0.346	0.730	0.426
想像的物語	0.488	0.340	-0.213	-0.076	0.397
全て	0.683	0.577	0.525	0.546	0.654

表 9　JCQ と DAT の相関

	流暢性	柔軟性	独創性	精緻性	平均
非通常使用	0.916	0.901	0.327	-0.072	0.854
結果	-0.278	-0.319	-0.239	-0.384	-0.450
仮定	0.939	0.742	0.441	-0.174	0.630
状況	-0.658	-0.056	-0.175	-0.557	-0.414
一般的問題	0.864	0.873	0.204	0.241	0.897
改善	0.914	0.763	0.652	-0.530	0.547
想像的物語	-0.057	-0.137	0.738	0.410	0.180
全て	0.882	0.577	0.307	-0.200	0.358

GPT-4o による評価との相関（Pearson）を表 8 に示す. 全体的には相関していたが，一部タスク・指標において相関がなかった. 特に，想像的物語タスクにおける相関が弱かった. GPT-4o は人と同様に物語の創造性を評価できていないと考えられる.

4.2　JCQ と DAT の相関

JCQ と DAT の相関（Pearson）を表 9 に示す. JCQ のそれぞれのタスクにおける各指標について，各モデルのスコアと DAT におけるモデルのスコアの相関を計算した. 一部タスクの「流暢性」,「柔軟性」において強い相関があった. 特に，非通常使用タスクの「柔軟性」と DAT に強い相関があったが，人に対して行われた Olson らの研究 [9] においても非通常使用タスクと同様のタスクである AUT の「柔軟性」と DAT に相関があったことから，この点において人と LLM で同じ傾向となっている. しかし，同研究 [9] において AUT の「独創性」と DAT に相関があったのに対し，本研究の LLM の場合は非通常使用タスクの「独創性」と DAT の相関は弱かった. LLM と人のタスク間における相関の傾向は必ずしも一致しないと言える.

4.3　JCQ と SAT の相関

JCQ と SAT の相関（Pearson）を表 10 に示す. JCQ のそれぞれのタスクにおける各指標について，各モデルのスコアと SAT におけるモデルのスコアの相関を計算した. 一部タスクの「流暢性」,「柔軟性」,「独創性」において強い相関があり，全体的に DAT よりも JCQ との相関が強かった.

表 10　JCQ と SAT の相関

	流暢性	柔軟性	独創性	精緻性	平均
非通常使用	0.606	0.992	0.736	0.114	0.899
結果	0.126	-0.200	0.214	-0.076	-0.017
仮定	0.678	0.945	0.824	0.260	0.897
状況	-0.221	0.368	0.320	-0.117	0.058
一般的問題	0.627	0.978	0.625	0.573	0.981
改善	0.601	0.966	0.939	-0.230	0.812
想像的物語	0.331	0.237	0.960	0.741	0.556
全て	0.908	0.855	0.725	0.170	0.712

5　おわりに

本研究では，LLM の創造性を測るため JCQ，DAT，SAT の 3 つのベンチマークを構築した. それぞれのベンチマークには網羅性と手軽さの観点で一長一短がある. JCQ は 7 つのタスクと 4 つの指標を用いており創造性を網羅的に評価できるが，評価に LLM を用いるため，他の 2 つのベンチマークと比較して時間とコストがかかる. DAT は 1 つのプロンプトしかないため網羅性が低いが，埋め込みを用いて迅速に評価できる. SAT は元の物語を用意する必要があるが，埋め込みを用いた手軽な評価を行える. また，網羅性は物語を書き換えるという 1 つのタスクしかないため JCQ より低いが，複数の物語を用いるため DAT よりは高い.

また，JCQ の評価における GPT-4o と人の相関を分析した. 想像的物語タスクなどの一部タスク・指標を除き，全体的には相関していた. 相関が弱かった一部を除けば，JCQ の結果は信頼できると言える.

さらに，JCQ と DAT，JCQ と SAT の相関を分析した. DAT と SAT は一部タスク・指標において JCQ と相関し，全体的に SAT の方が JCQ との相関が強かった. SAT よりも DAT の方が手軽であるため，JCQ との相関の強さは手軽さとのトレードオフになっている.

創造性を適切に評価することは，LLM の能力を理解し活用していく上で重要な意味を持つ. 本研究で提案した 3 つのベンチマークは，目的に応じて LLM の創造性を効率的に測定する手段を提供するものである. これにより，LLM の創造的能力の現状を把握し，タスクや用途に応じて適切なモデルを選択することが可能となる. また，創造性の評価基準を確立することは，今後の LLM の改善の方向性を示す上でも重要な指針となるだろう.

謝辞

本研究は SB Intuitions 株式会社と早稲田大学の共同研究により実施した.

参考文献

[1] Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. **AI & society**, 2024.

[2] Takaaki TANAKA, Shun OTSUBO, Kotaro ITO, Takuya HATAKEYAMA, Yuji ANZAI, Tomoaki NAGASAKA, Takashi MATSUI, and Nobuyuki ISHIKAWA. Research on ideation applications using llm-based multi-agent systems and idea evaluation methods. **Proceedings of the Annual Conference of JSAI**, pp. 4G3GS205–4G3GS205, 2024.

[3] Kengo WATANABE, Takashi KAWAMURA, Reo KOBAYASHI, Kzuma ARI, Akifumi ITO, and Satoshi KURIHARA. Interactive story generation system: Enhancing creative writing with a llm informed by narrative structure analysis. **Proceedings of the Annual Conference of JSAI**, pp. 1T3OS32a05–1T3OS32a05, 2024.

[4] Jiayang Li, Jiale Li, Yunsheng Su, Stavroula Ntoa, Helmut Degen, Stavroula Ntoa, and Helmut Degen. A map of exploring human interaction patterns with llm: Insights into collaboration and creativity. In **Artificial Intelligence in HCI**, Vol. 14736 of **Lecture Notes in Computer Science**, pp. 60–85. Springer, Switzerland, 2024.

[5] Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. arXiv, 2024. abs/2401.12491.

[6] S A Handayani, Y S Rahayu, and R Agustini. Students' creative thinking skills in biology learning: fluency, flexibility, originality, and elaboration. **Journal of Physics: Conference Series**, Vol. 1747, No. 1, pp. 12040–, 2021.

[7] Eunsook Hong, Yun Peng, Harold F. O'Neil Jr, and Junbin Wu. Domain-general and domain-specific creative-thinking tests: Effects of gender and item content on test performance. **The Journal of creative behavior**, Vol. 47, No. 2, pp. 89–105, 2013.

[8] Kyung Hee Kim. Can we trust creativity tests? a review of the torrance tests of creative thinking (ttct). **Creativity Research Journal**, Vol. 18, No. 1, 2006.

[9] Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. Naming unrelated words predicts creativity. **Proceedings of the National Academy of Sciences**, Vol. 118, No. 25, p. e2022340118, 2021.

[10] Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, Qi Guo, Ling Li, and Yunji Chen. Assessing and understanding creativity in large language models. arXiv, 2024. abs/2401.12491.

[11] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. arXiv, 2023. abs/2401.12491.

[12] おはなしいっぱい♪おうたいっぱい♪昔話童話童謡の王国,（2024-6 閲覧）. https://www.douwa-douyou.jp/index.shtml.

A 付録

表 11 JCQ の回答例

問題	電球の通常でない使い方をできるだけたくさん挙げてください.
モデル	claude-3-5-sonnet-20241022
回答	電球の創造的な使用方法をいくつか提案させていただきます： 装飾・クラフト関連 1. ミニチュアテラリウムの容器 2. クリスマスオーナメント 3. 一輪挿しの花瓶 （後略）

表 12 DAT の回答例

モデル	gpt-4o-2024-08-06
回答	1. 本 2. 海 3. 山 4. 鳥 5. 音 6. 花 7. 雨 8. 笑顔 9. 石 10. 夢

表 13 SAT の回答例

プロンプト	条件に従い以下の物語を作り変えてください. # 条件 - 現代風の物語を創造的に考える - 文字数は 400 文字程度 - 改行はしない - あなたが考えた物語のみを出力する # 物語 昔、心優しい漁師の浦島太郎は子供たちにいじめられている亀を助けました.（中略）竜宮城での楽しい日々は、実は何百年も経っていたのでした.
モデル	claude-3-5-sonnet-20241022
回答	大手 IT 企業で SE として働く浦島拓也は、道端で故障した自動運転車を目撃した.（中略）メタバースでの幸せな時間は、現実では長期の引きこもり生活だったのだ.

表 14 DAT プロンプト

```
# 指示
できるだけ互いに異なる意味や用途を持つ単語を 10 個考え、以下の形式で出力してください.

# ルール
1. 単語はそれぞれ 1 語のみとします
2. 名詞のみ使用可能です (物、対象、概念など)
3. 固有名詞は使用できません (特定の人物や場所など)
4. 専門用語は使用できません
5. 説明は不要です

# 出力形式
1. 単語 1
2. 単語 2
3. 単語 3
4. 単語 4
5. 単語 5
6. 単語 6
7. 単語 7
8. 単語 8
9. 単語 9
10. 単語 10
```

表 15 SAT プロンプト

```
条件に従い以下の物語を作り変えてください.
# 条件
- 現代風の物語を創造的に考える
- 文字数は 400 文字程度
- 改行はしない
- あなたが考えた物語のみを出力する
# 物語
{ 元の物語 }
```

表 16 JCQ 評価プロンプト

```
質問に対する回答を読み、4 つの観点からそれぞれ 5 段階で評価してください.

# 注意事項
- 回答全体を通読してください
- 各基準の説明をよく読み、独立に評価してください
- 評価に迷った場合は、より低い評価を選択してください
- 出力形式に従い、評価結果のみを出力してください

# 出力形式
流暢性: [1-5]
柔軟性: [1-5]
独創性: [1-5]
精緻性: [1-5]

# 質問
{ 質問 }

# 回答
{ 回答 }

# 流暢性: 質問と関連する異なるアイデアの量を評価してください. 重複や言い換えは 1 つとしてカウントしてください.
1. 1-2 個のアイデア
2. 3-4 個のアイデア
3. 5-6 個のアイデア
4. 7-8 個のアイデア
5. 9 個以上のアイデア

# 柔軟性: 回答に示された視点、カテゴリー、またはアプローチの多様性を評価してください.
1. 単一の視点
2. 2 つの異なる視点
3. 3 つの異なる視点
4. 4 つの異なる視点
5. 5 つ以上の異なる視点

# 独創性: 回答に含まれるアイデアがどれだけユニークであるかを評価してください.
1. 誰もが思いつく極めて一般的なアイデア
2. よく見られる一般的なアイデアだが、わずかな工夫がある
3. やや珍しい発想や意外性のあるアイデア
4. 斬新で独創的な発想のアイデア
5. 極めて独特で革新的なアイデア

# 精緻性: アイデアの詳細さや展開の深さを評価してください.
1. アイデアが単純で詳細な説明がない
2. 基本的な説明は含まれているが深い展開がない
3. ある程度の詳細な説明や展開がある
4. アイデアが詳細に説明され、よく展開されている
5. アイデアが非常に詳細で、複雑な展開がなされている
```