# The Language of Creativity: Evidence from Humans and Large Language Models

William Orwig, Emma R. Edenbaum, Joshua D. Greene, Daniel L. Schacter. (2024).

*The Journal of Creative Behavior*, Vol. 58, Iss. 1, pp. 128–136

2024-05-07

M2 櫛田

# Introduction

- アイデアの創造的な質をどのように評価するかという問題は、創造性研究における長年のトピックである(Amabile, 1982; Reiter-Palmon, Forthmann, & Barbot, 2019)。
- 発散的思考(DT) (Guilford, 1967)とは、多様なタイプの情報を組み合わせることによって創造的なアイデアを生み出す能力を指す。
  - Alternative Uses Task(AUT)はDTの最も一般的な測定法である。参加者に物体を提示し、その物体の通常とは異なる創造的な使用法を生み出すよう求める。
  - AUTを評価するために、生成された回答を数える流暢さ(fluency)や、異なるカテゴリの回答を数える柔軟性(flexibility)といった指標に依存してきた。しかし、これらの指標は互いに相関性が高く、創造的なアイデアの全容を捉えるには不十分である(Acar, Ogurlu, & Zorychta, 2022; Acar & Runco, 2015)。
- エピソード記憶のDTへの寄与が強調されている。
  - エピソード検索を高める手続きである短いEpisodic Specificity Induction(ESI)を受けた後、参加者は対照群と比較して、 AUTでより多くの新規の使用法を確実に生成する(Madore, Addis, & Schacter, 2015; Madore, Jing, & Schacter, 2016)。
  - 現在の理論では、意味記憶とエピソード記憶の両プロセスが、創造的なアイデアの生み出しに寄与することが示唆されている(Benedek, Beaty, Schacter, & Kenett, 2023)。

# Introduction

- 自然言語処理における画期的な進歩に伴い、生成言語モデルは世界的な注目を集め、その創造的な能力をテストしようとする研究が増えている。

- OpenAIのGPT-3がAUTの応答を生成する能力をテストした研究(Stevenson, Smal, Baas, Grasman, and van der Maas, 2022)では、流暢さ、柔軟性、意味的距離の観点から人間の応答と性能を比較した。
  - AIが近いうちに人間レベルのパフォーマンスを達成する可能性を示唆している。

- DTタスクの自動採点のために、大規模言語モデルをどのように微調整できるかを実証し、創造性のより微妙な評価を可能にした研究もある(Organisciak, Acar, Dumas, and Berthiaume, 2023)。
  - 創造的なコンテンツの生成と評価の両方における大規模言語モデルの多面的な有用性を強調している。
  - 本研究の主な目的は創造的なテキストの特徴を特定することであるため、人間のストーリーに加え、大規模言語モデルをデータ源として活用する。

# Introduction

- 本研究では、計算言語学的ツールを用いて、創造性の特徴をさらに定義する。
  - AUTのような従来のDT評価の限界を克服するために、確立されたストーリーライティング課題(Johnson et al., 2022; Prabhakaran, Green, & Gray, 2014)を用いて短編ストーリーの創造性の質を評価する。
- 我々は、ストーリーが意味的に多様なアイデアをどの程度結びつけるかを評価するために、創造性理論と分布意味論を活用して、divergent semantic integration(DSI)スコアを計算する(Johnson et al., 2022)。
  - 我々は、意味的に離れた概念を組み合わせることに加えて、ストーリーが知覚的な詳細をどの程度組み込んでいるかが創造性の予測因子であると仮定する。
- 人間のテキストと大規模言語モデルによるテキストの創造的な質に実質的な違いがあるかを判断するために、人間とコンピュータが生成したストーリーの主観的な創造性評価を比較する。
- 本研究を開始した後にGPT-4が登場したため、GPT-4によって生成されたストーリーの独立したサンプルでこれらの知見を再現し、現在の言語モデルが創造的な文章を書く課題で人間レベルの性能を達成するかどうかを検証する。

# Methods

- 人間とGPT-3から、創造的なショートストーリーを収集した。
- 人間の参加者(n=50)はProlificを通じてオンラインで募集した。
  - サンプルサイズは同様のアプローチを用いた過去の研究と一致している。
  - 年齢は18~35歳(M = 27.71 ± .09; 29 females)

- 5文創作物語課題(Johnson et al., 2022; Prabhakaran, Green, & Gray, 2014)のコンピュータ版を用いた。
  - 参加者は3つの単語からなるプロンプトを与えられ、5文程度の短い物語を書く際に3つの単語をすべて含めるように求められた。
  - プロンプトは合計6つ与えられ、ランダムな順番で提示された。
  - プロンプトの半分は概念的に関連し、意味的に類似した単語(ex: stamp-letter-send, week-year-embark, and belief-faith-sing)で構成され、
    残りの半分は概念的に異質で、意味的に離れた単語(ex: gloom-payment-exist, organ-empire-comply, and statement-stealth-detect)で構成されていた。
  - 参加者50人からそれぞれ6つの回答を集め、合計300のストーリーを得た。

# Methods

- OpenAIのAPIを使用し、人間に実施した各プロンプトのストーリーをGPT-3に要求した。
  - Stevenson et al. (2022)は、モンテカルロサンプリングを実施し、どのプロンプトとパラメータの設定が最も有効な回答に繋がり、創造性スコアが最も高いかを決定した。
    - 彼らの結果に基づき、プロンプトを実行した。
  - GPT-3から6つのプロンプトそれぞれについて50の回答を収集し、合計300のストーリーを得た。

- 複製サンプルとして、GPT-4からもプロンプトごとに20の回答を収集し、120のストーリーを得た。

# Creativity Assessment

- ストーリーの収集後、Prolificで募集したオンラインの独立した評価者(n=240)にその創造性の質を評価するよう課した。
  - 評価者は18〜55歳(M = 34.58 ± .61; 154 females)
  - 各ストーリーについて12個の評価を収集した。

- 評価者は各ストーリーを読み、1(非常に創造的でない)から5(非常に創造的)までの尺度で創造性のスコアをつけるよう促された。
  - ストーリーの長さや英語の質にこだわりすぎず、全体的な創造性の質を考慮するよう指示された。
  - 各評価者には、無作為に選ばれた30本のストーリーが提示された。
    - ストーリーが人間とAIから生成されたものであることは知らされなかった。

- すべての評価の平均を、各ストーリーの創造性に関する人間の総合評価として計算した。
  - 人間による創造性の評価には高い信頼性が認められた。
    - 平均ICCは0.85で、95%信頼区間は0.76から0.92だった(F (28, 3731) = 7.38, p < .001)。

# Creativity Assessment

- さらに、GPT-4の高度な自然言語処理能力を活用し、ストーリーの創造性の質を人間のように評価するために、GPT-4からも創造性の評価を収集した。
  - 人間の評価者に提供されたのと同じタスク指示を使用して、各ショートストーリーをGPT-4に入力し、1~5スケールで単一の創造性スコアを提供した。

# Assessment

- 次に、ショートストーリーの全サンプルについて、創作の質に関する人間の評価とGPT-4の評価の相関関係を検証する。

## 意味的多様性(Semantic Diversity)

- DSIを用いて、ストーリーがどの程度多様なアイデアを結びつけているかを評価した。
  - DSIとは意味的多様性(Johnson et al., 2022)の確立された評価指標である。
  - DSI値の理論的範囲は0から1であるが、ほとんどのスコアは0.70から0.90の間に収まる傾向がある。
  - DSIのスコアが高いほど、物語がより多様なアイデアを繋いでいることを示す。

## 知覚的詳細(Perceptual Details)

- ストーリーはLinguistic Inquiry and Word Count program (LIWC)に入力された。
  - LIWCは、統合辞書によって定義されたさまざまな心理学的カテゴリーに該当する単語を自動的にカウントする。
- ここでは、知覚のプロセスを表す単語(e.g., “observe,” “heard,” “feeling,” “touch”)を含む、知覚プロセスカテゴリの結果を報告する。
  - 知覚的詳細のスコアが高いほど、各ストーリーにおいて知覚的プロセスの記述が多いことを示す。

# Assessment

## Statistical Analyses

- 創造性、意味的多様性、知覚的詳細の関係を調べるために、多重線形混合効果モデルを行った。
  - 最初のモデルでは、DSIと知覚的詳細の回帰パラメータを創造性の予測因子として推定し、プロンプトと単語数をコントロールした。
  - 次に、DSIと知覚的詳細の交互作用の可能性を検証するために、2つ目のモデルを当てはめた。
  - 人間とGPT-3のストーリーが創造的な質において異なるかどうかを決定するために、条件（人間対GPT-3）を創造性のカテゴリー的予測因子とした3番目のモデルを当てはめた。
  - そして、GPT-4によって生成された新しいストーリーセットでこれらの分析を再現した。
  - さらに、GPT-4から創造性の評価を収集し、創造性に対する人間とGPTの評価の間に対応関係があるかどうかを検証した。
  - 最終段階として、GPTによる創造性の評価を用いて回帰分析を繰り返した。
    - 分析に先立ち、予測変数はzスコア変換を用いて標準化し、意味のある比較ができるようにした。

# Results

- DSIと創造性の間には、人間(r = .56)とGPT-3ストーリー(r = .56)のサンプル内で強い正の相関が観察された。(Figure 1a)
- 人間(r = .16)とGPT-3 (r = .25)が作成したストーリーでは、知覚的詳細と創造性の間に緩やかな相関が観察された。(Figure 1b)
- 我々の分析は、DSIと知覚的詳細の両方が創造性を予測することを示している。

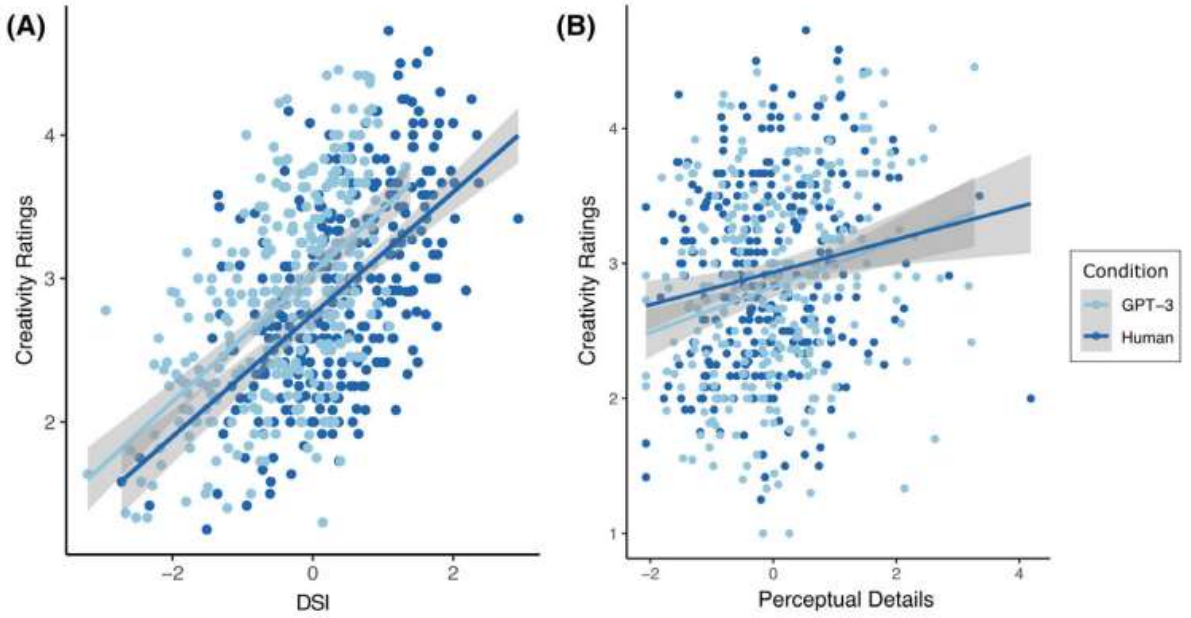


FIGURE 1. Semantic diversity (A) and perceptual detail (B) correlate with creativity ratings.

# Results

- 全標本において、
  - DSIは創造性を強く予測することが観察された($\beta$ = .23, 95% CI [0.18, 0.27], t = 9.55, p < .001)。
    - 単語数とプロンプトをコントロールした場合、DSIが1標準偏差異なると創造性の評価で0.23ポイント異なると予測した。
  - 知覚的詳細と創造性の間に正の相関が観察された($\beta$ = .12, 95% CI [0.08, 0.17], t = 5.40, p < .001)。
    - 知覚的詳細が1標準偏差異なると、0.12ポイント創造性評価が異なると予測した。
  - このモデルは、観察された創造性評価の分散の約半分を説明する($R^2$ = .49)。

- 人間のストーリーのサンプルの中で、
  - DSIは創造性を非常に予測することが観察された($\beta$ = .32, 95% CI [0.24, 0.39], t = 8.63, p < .001)。
  - 知覚的詳細と創造性の間に正の相関が観察された($\beta$ = .11, 95% CI [0.04, 0.17], t = 3.19, p = .002)。
  - このモデルは、観察された創造性評価の分散の約半分を説明する($R^2$ = .47)。

# Results

- GPT-3ストーリーのサンプルにおいて、
  - DSIは創造性を高度に予測することが観察された($\beta$ = .18, 95% CI [0.11, 0.25], t = 4.90, p < .001)。
  - 知覚的詳細と創造性の間に正の相関が観察された($\beta$ = .09, 95% CI [0.03, 0.15], t = 2.92, p = .004)。
  - このモデルは創造性評価の分散の約54%を説明する($R^2$ = .54) 。
- 2つ目のモデルでは、DSIと知覚的詳細の間に有意な交互作用が観察された($\beta$ = .04, 95% CI [0.01, 0.08], t = 2.16, p = .03) 。
  - DSIと知覚的詳細が共にあると、どちらか一方だけよりも大きく創造性への効果があることを示唆している。
- 3つ目のモデルでは、人間とコンピュータが創作したストーリーの間に有意な差は観察されなかった($\beta$ = .07, 95% CI [0.02, 0.16], t = 1.59, p = .11) 。
  - このモデルは観察された分散の約39%を説明する($R^2$ = .39)。
  - 人間によって書かれたストーリーは、同じ長さとプロンプトのGPT-3のストーリーより創造性スコアが0.07ポイント高い。

# Results

- GPT-4ストーリーでも、創造性の評価はDSI($r = .46$)、知覚的詳細($r = .44$)の両方と強く相関している。
  - DSI($\beta = .16$, 95% CI [0.10, 0.22], $t = 5.04$, $p < .001$)と知覚的詳細($\beta = .11$, 95% CI [0.04, 0.18], $t = 2.97$, $p = .004$)の両方が創造性を強く予測する。
  - DSIと知覚的詳細の間に有意な交互作用は観察されなかった($\beta = .03$, 95% CI [0.09, 0.02], $t = 1.14$, $p = .26$)。
  - 人間と比較して、GPT-3、GPT-4のストーリーは創造性において低いスコアを示す傾向があったが、有意ではなかった。

- 人間とLLMの創造性評価の相関を調べたところ、ストーリーの全サンプルを通して、創造性についての人間とGPT-4の評価の間には、強固な正の相関が観察された($r = .65$) 。

# Results

- 創造性のGPT評価について、全サンプルにおいて、DSIはGPTの創造性評価を高度に予測した($\beta$ = .25, 95% CI [0.18, 0.31], t = 7.82, p < .001)。
  - 知覚的詳細とGPTの創造性評価との間に有意な正の相関が観察された($\beta$ = .11, 95% CI [0.05, 0.17], t = 3.73, p < .001) 。
  - このモデルはGPT創造性評価の分散の約28%を説明する($R^2$ = .28)。

  - 人間のストーリーのサンプルでは、DSI($\beta$ = .30, 95% CI [0.21, 0.39], t = 6.58, p < .001)と知覚的詳細($\beta$ = .12, 95% CI [0.04, 0.20], t = 2.92, p = .004)の両方が有意な予測因子であった。
  - GPT-3ストーリーのサンプルでは、DSI($\beta$ = .15, 95% CI [0.07, 0.24], t = 3.52, p < .001)は予測したが、知覚的詳細($\beta$ = .06, 95% CI [0.03, 0.15], t = 1.38, p = .17)には有意な関連は見られなかった。
  - 交互作用モデルでは、DSIと知覚的詳細の有意な交互作用は見られなかった($\beta$ = .05, 95% CI [0.01, 0.10], t = 1.62, p = .11)。
  - 人間のストーリーは、コンピュータが生成したストーリーと比較して、有意に創造的であると評価された($\beta$ = .14, 95% CI [0.02, 0.25], t = 2.36, p = .02)。

# Results

- GPT-4で作成されたストーリーのサンプルでも同様の結果が観察された。
  - GPTの創造性評価は、DSI(r = .60)と知覚的詳細(r = .30)の両方で正の相関がある。
  - DSI($\beta$ = .44, 95% CI [0.34, 0.54], t = 8.67, p < .001)は創造性の評価を有意に予測したが、知覚的詳細($\beta$ = .09, 95% CI [0.02, 0.21], t = 1.61, p = .11)との有意な関係は見られなかった。
  - DSIと知覚的詳細の交互作用は観察されなかった($\beta$ = .02, 95% CI [0.11, 0.08], t = .37, p = .71)。
  - 人間のストーリーと比較して、GPT-3のストーリーは創造性のGPT評価で有意に低く($\beta$ = .14, 95% CI [0.25, 0.02], t = 2.36, p = .02)、GPT-4のストーリーは有意に高かった($\beta$ = .76, 95% CI [0.60, 0.91], t = 9.74, p < .001) 。
    - GPTによる創造性の評価は、人間よりもGPT-4のストーリーに有利であることを示唆している。

# Discussion

- 本研究では、創造的な文章を構成する特徴に焦点を当て、人間と大規模言語モデルによって書かれたショートストーリーにおける意味的多様性と知覚的詳細の役割を明らかにした。
  - 意味的多様性は創造性の重要な予測因子であり、創造性評価の分散の半分以上を説明することが示された。
  - 知覚的詳細と創造性の間に正の相関が観察され、特定の知覚的詳細を含めることが創造的な物語の重要な特徴であることが示された。
  - これらの結果は、意味記憶とエピソード記憶の両方が、創造的な文章を書く過程に共同で寄与している可能性を示唆している。

- 創造的発想における記憶の役割を理解するために提案された枠組み(Benedek, Beaty, Schacter, & Kenett, 2023)は、意味記憶とエピソード記憶の両方が創造的プロセスにおいて重要な役割を果たすことを示唆している。
  - 我々の発見は創造的な物語における知覚的詳細の生成において、エピソード記憶過程が果たす役割の可能性を示唆している。

# Discussion

- 本研究では、ショートストーリーの文脈において、GPT-4と人間の創造性評価との間に強い対応関係があることを発見した。
  - 意味的多様性と知覚的詳細の両方がGPT-4の創造性評価の重要な予測因子であることがわかった。
  - GPT-4は自身のストーリーを人間のストーリーよりも有意に創造的であると評価することがわかった。
    - この結果は、創造性評価におけるLLMの有効性を検証するだけでなく、AIがデジタル時代の創造性の測定に与える影響を浮き彫りにする。

- 我々の結果は、創造的な文章を書くには、意味的に多様な概念を知覚的記述と統合することが必要であると示している。
  - さらに本研究は、GPT-3とGPT-4が、人間が生み出すストーリーに匹敵する創造性を生み出すことができるという証拠を提供する。
  - 今後の研究では、クリエイティブライティングに寄与する明確なエピソード的特徴と意味的特徴を分離し、人間とAIの協働の可能性を探ることを目指すべきである。