Humans create more novelty than ChatGPT when asked to retell a story

Fritz Breithaupt, Ege Otenen, Devin R. Wright, John K. Kruschke, Ying Li & YiyanTan.

Scientifc Reports, (2024) 14:875.

2024-08-06

M2 櫛田

Introduction

- コミュニケーションの重要な側面として、物語や短い物語によって、人間は過去、予測、架空の出来事などを伝えてきた。
  - 本論文では、人間による物語の伝達が、大規模な言語モデルである ChatGPT による伝達とどのように異なるかを探る。
  - ただし、現時点での私たちのアプローチは、ChatGPTを人間の心の代理として使用することではなく、ChatGPTを利用して人間の独自性を明らかにすることに焦点を当てる。
    - つまり、物語を体験して記憶する人間のプロセスが、任意のプロンプトに応じて言語を生成するChatGPTの驚異的な能力と具体的にどのように異なるかを明らかにすること。(Yang et al., 2023)
- 物語や出来事を語り直す活動は、本質的に人間の文化や文明と結びついている。(Dunbar, R. I. M, 1996; Kalish, M. L., Grifths, T. L. & Lewandowsky, S., 2007)
- 物語コミュニケーションの特徴は、聞き手が最初に物語とその感情の流れを、あたかもその場にいるかのように体験し、その後に思い出したり語り直したりすることである。(Nabi, R. L. & Green, M. C., 2015)
  - この体験は、物語の出来事の時間的順序が、段階的なプロセスにおける個人の体験に似た形で体験されることを意味する。

Introduction

- 物語はさまざまな種類の情報を伝えるが、物語が連鎖的に伝わることでどのような変化が起こるのかが問題となる。
  - Bartlett, F. C. (1995)は、物語のコミュニケーションの場合、連鎖的に語り継がれると、物語が調整されて短くまとまりのある物語になるという強いパターンを強調した。
- 異なる語り手は、異なる物語の側面 (Neisser, 1982)に焦点を当て、それぞれの先入観に従って新しいアイデアや概念を語りの中に挿入する可能性がある。
  - 私たちの研究では、語り直しにおける継続的な創造性を観察する。
- 物語の連続的再生産に関するこれまでの分析では、ほとんどの研究が、伝達される物語状況の事実情報 (誰が誰に、どこで、いつ、なぜ行うか)に焦点が当てられてきた。
  - 特に、社会的情報はより高い程度で保持される。(Mesoudi et al., 2006; Tafani et al., 2006; Stubbersfeld et al., 2015; Jimenez & Mesoudi., 2021)

Introduction

- 語り直しのもう一つの中心的な側面は、あまり研究されていないが、物語の感情的な側面である。(Wagoner, 2017; Reagan et al., 2016)
  - Bartlett(1995)は、「ムード」という言葉で、感情的な側面が伝達において中心的な役割を果たすことを示唆している。
    - 語り直し者や ChatGPT が伝達の過程で物語の感情的な評価を変えるタイミングを観察することが重要であり、それによって聴衆の反応や行動が変わる可能性がある。

- 強い感情を伴う物語の語り直しでは、強い感情を伴わない物語よりも多くの命題が正しく保持される。(Stubbersfeld et al., 2015; Stubbersfeld et al., 2017)
  - 語り直しを通して感情が存続するかどうかについても文化的な違いがある。
  - Eriksson et al.(2016)は、嫌悪に焦点を当てた物語はアメリカ人によって安定して伝達された（つまり、元のテキストの前置詞がより多く保持された）のに対し、幸せな物語はインド人によってより安定して伝達されたことを発見した。
  - 同時に、物語の全体的な感情評価は、物語が大幅に縮小され、物語の事実が変更された場合でも、幸福、悲しみ、当惑(Breithaupt et al., 2022; He et al., 2023)、驚き(Breithaupt., 2018)などが、高い忠実度で伝達される。
    - つまり、命題は変化するかもしれないが、少し幸せな話は少し幸せな話のままであり、非常に幸せな話は非常に幸せな話のままである。

ChatGPT

- ChatGPT4を使用した研究では、学部生は文学作家Lovecraftの散文と、著者のスタイルである程度トレーニングを受けた後に書いたChatGPTテキストを確実に区別することができなかった。(Garrido-Merchan, 2023)
- Chu & Liu (2023)は、ChatGPT によって書かれたとラベル付けされたストーリーが、人間によって書かれたとラベル付けされた同じストーリーよりも没入感は低いが説得力は高いと評価されていることを比較した。
- ChatGPT には多くの機能があるが、いくつかの制限に注意する必要がある。
  - Acerbi & Stubbersfield (2023)は、ステレオタイプの情報(否定的、社会的、脅威)に焦点を当てたものに関して、ChatGPT による語り直しは人間の語り直しと同様のバイアスに悩まされることを示した。
- 心の理論 (ToM) タスク (キャラクターの内面状態に関する推定の精度を測定するもの)で、Brunet-Gouet et al. (2023)は、ChatGPT3 はほとんどのシナリオで自発的で信頼性の高いToM予測を行わないようだと示唆している。
  - ChatGPT は一般的な人間の能力を大きく下回るパフォーマンスを発揮した。

Overview of study

- 本研究では、人間と ChatGPT による物語の語り直しの比較として、先行研究で概説された 2 つの傾向、つまり言語の安定性と語り直しにおける感情の保存に焦点を当てる。
  - 物語のベースとして、Breithaupt et al. (2022)の研究1の116の物語を使用した。
    - 人間のデータセットでは、Amazon Mechanical Turk (AMT) の最初の参加者セットに、約120～160語の幸せ、やや幸せ、やや悲しい、悲しい物語を書くように依頼した。
      - また、happyやsadなどの感情を明示的に表す言葉を避けるように指示した。
      - 次に、348 人の他の参加者にこれらの物語を語り直しするように依頼した。
      - 参加者は、感情、影響など物語の特定の側面に注意を払うように指示されていない。
      - 物語は語り直しの過程で短くなる傾向があったため、長さのバランスをとるために、同じ物語を3回語り直した。
      - 意味不明な言葉などの欠陥のある物語の連鎖を除外した後、3 つの語り直し (それぞれ異なる参加者による) を含む116のオリジナル物語と語り直しが作成された。
      - 537人の参加者が、感情やその他の基準で物語を評価した。

Overview of study

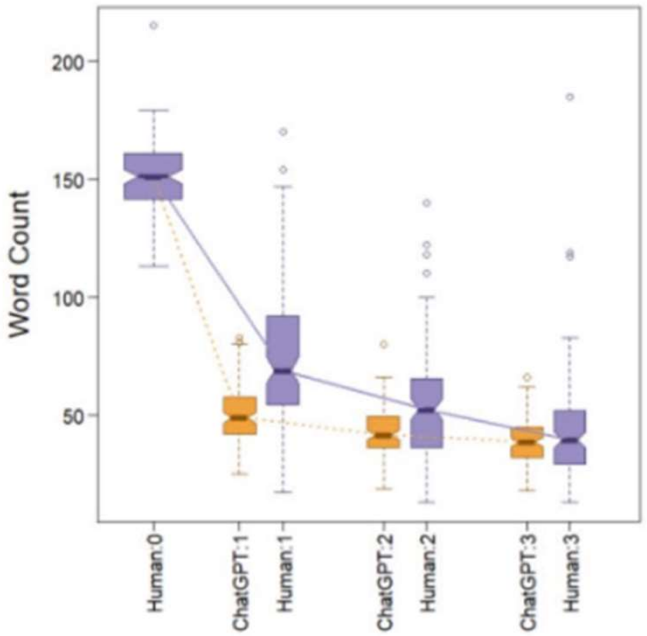
- ChatGPT3 に、同一の開始ストーリーを使用し、語り直しタスクのために人間の参加者に与えたものと同一の指示をした。
  - 語り直しは、ChatGPT が元のストーリーや他のバージョンにアクセスできないようにするために、ChatGPT の親会社である Open AI の異なるアカウントから作成された。
- Prolific で 531 人の参加者を採用し、1回の語り直しの反復から6つのストーリーを評価した。

Results

- ChatGPTと人間の語り直しはどちらも、元の物語よりも著しく短くなった。語り直しでは、物語中の重要な要素 (男性の病気、助けの必要性、妻の頻繁な不在、孫の世話など)が保持されている。
- それでも、ChatGPTと人間の語り直しは異なるパターンを示した。
  - ChatGPTによる3つの語り直しはどれも驚くほど似ていた。
  - 対照的に、人間の語りは繰り返しの間に形を変えた。
    - 人間による語り直しはそれぞれ、以前の物語バージョンに対する新たな異なる解釈を提示しており、異なる視点や斬新な発明も含まれる。

Word count

- ChatGPTと人間はどちらもストーリーを大幅に短縮した(図1)が、ChatGPTによる最初の語り直しは人間による最初の語り直しよりも大幅に短くなった。
  - 回帰分析 (人間とChatGPTに別々の分散パラメータを持つ負の二項分布を使用 (詳細は https://osf.io/jr2py/))により、
    - ChatGPTと人間の両方が語り直し1〜3で統計的に有意な単語数の減少を示した。($p < 0.001$)
    - ChatGPTによる単語数の減少は人間よりも語り直し1〜3で大幅に緩やかだった。($p < 0.001$)
    - 人間の単語数の変動はChatGPTよりも大幅に大きい。($p < 0.001$)



Part-of-speech differences

- NLPパッケージ (nltk) (Loper, E. & Bird, S. Nltk., 2002)を使用して、動詞、副詞、名詞、形容詞、代名詞、前置詞、否定に焦点を当てて、すべてのストーリーのすべての単語の品詞にラベルを付けた。
  - 図1(論文参照)は、ChatGPT と人間による語り直しにおける各品詞の割合を示す。(カウントの比率は、二項回帰またはベータ二項回帰を使用して分析された。)
    - 一般的に、語り直し 1 から 3 の間ではわずかな変化しかなかった。
    - しかし、ChatGPT と人間の間には顕著な違いがあった。
      - 人間は ChatGPT よりも多くの動詞、副詞、代名詞を使用し、人間は ChatGPT の 2 倍の否定を使用した(1.01% 対 0.47%)。
      - 一方、ChatGPT は人間よりも多くの名詞と形容詞を使用し、ChatGPT は人間よりもわずかだが有意に多くの前置詞を使用した。

Part-of-speech differences

- 人間による動詞と副詞の比較的高い使用は、副詞と感情性の関係を考慮すると、人々が行動と感情に焦点を当てていることを示唆している。(Dragut, E. & Fellbaum, C. T., 2014)
- ChatGPTによる名詞と形容詞の比較的高い使用は、ChatGPTが語り直し中に実在物に焦点を当てていることを示唆している。
- 人間による語り直しにおける否定の相対的普及は興味深い。
  - 人間と大規模言語モデルはどちらも否定の使用に困難を抱えている。
    - 人間は否定に労力を要すると感じている。(Kaup et al., 2006)
    - 大規模言語モデル (例: ChatGPT) は否定を理解するのに苦労している。(Hosseini et al., 2021)
  - 否定には認知的労力が必要であるにもかかわらず、人間はすべての語り直しにおいて ChatGPT よりも否定の割合が高い。

Part-of-speech differences

- 人間とChatGPTによる否定の使用の違いを説明できる特定の文脈があるかどうか疑問に思った。
  - 一般的に、文脈と期待は人間の処理コストに影響する(Dale, R. & Duran., 2011)。
  - Nordmeyer & Frank (2015)は、人間は不在などの適切な文脈で否定をよりうまく処理することを発見した。
    - そのため、悲しい話は、否定がより適切な文脈を提供するのではないかと考えた。
      - 幸せな話と悲しい話で、文法的な否定の頻度の違いを確認したが、結果は、人間（1.13% vs 0.88%）とChatGPT（0.58% vs 0.36%）の両方が、悲しい話で幸せな話よりも否定を多く使用しており、統計的に有意な交互作用はないことを示した。

Concepts: creativity, stability, and decay

- テキスト内の概念を表現するために、我々は*synsets*を使用した。
  - *synsetsとは*、 *WordNet* (2010)に「概念的・意味的・語彙的関係によって相互にリンクされた」明確な概念的認知構成。
  - WordNet はツリー構造に階層的に編成されている。
    - これにより、単語が出現する特定の文脈で同義語、下位語、上位語を見つけることができる。
    - たとえば、文脈によっては、「slice」という単語は「apple」の正しい伝達としてカウントされますが、「tasty」の正しい伝達としてはカウントされない。
  - 我々はsynsetを分析して、概念が伝達、忘却、発明される速度を測定した。
- 図2は、最初の語り直しではChatGPTが人間よりも平均して少ないsynsetを使用していることを示した。
  - 語り直し1 ～ 3ではChatGPTのsynset数の減少は人間よりも緩やかである。
  - 図2は、ChatGPTのsynset密度 (単語あたりのsynset数) が人間よりも高いことも示している。
    - 言い換えれば、人間はChatGPTよりもsynsetあたりの単語数が多い (平均)。
    - 語り直し2と3では特に、ChatGPTでは人間よりもsynsetが語り直し間で生き残る可能性が高い。つまり、ChatGPTは使用するsynsetに比較的「固執」している。
    - さらに、特に語り直し2と3では、人間はChatGPTよりも新しい synset を作成します。

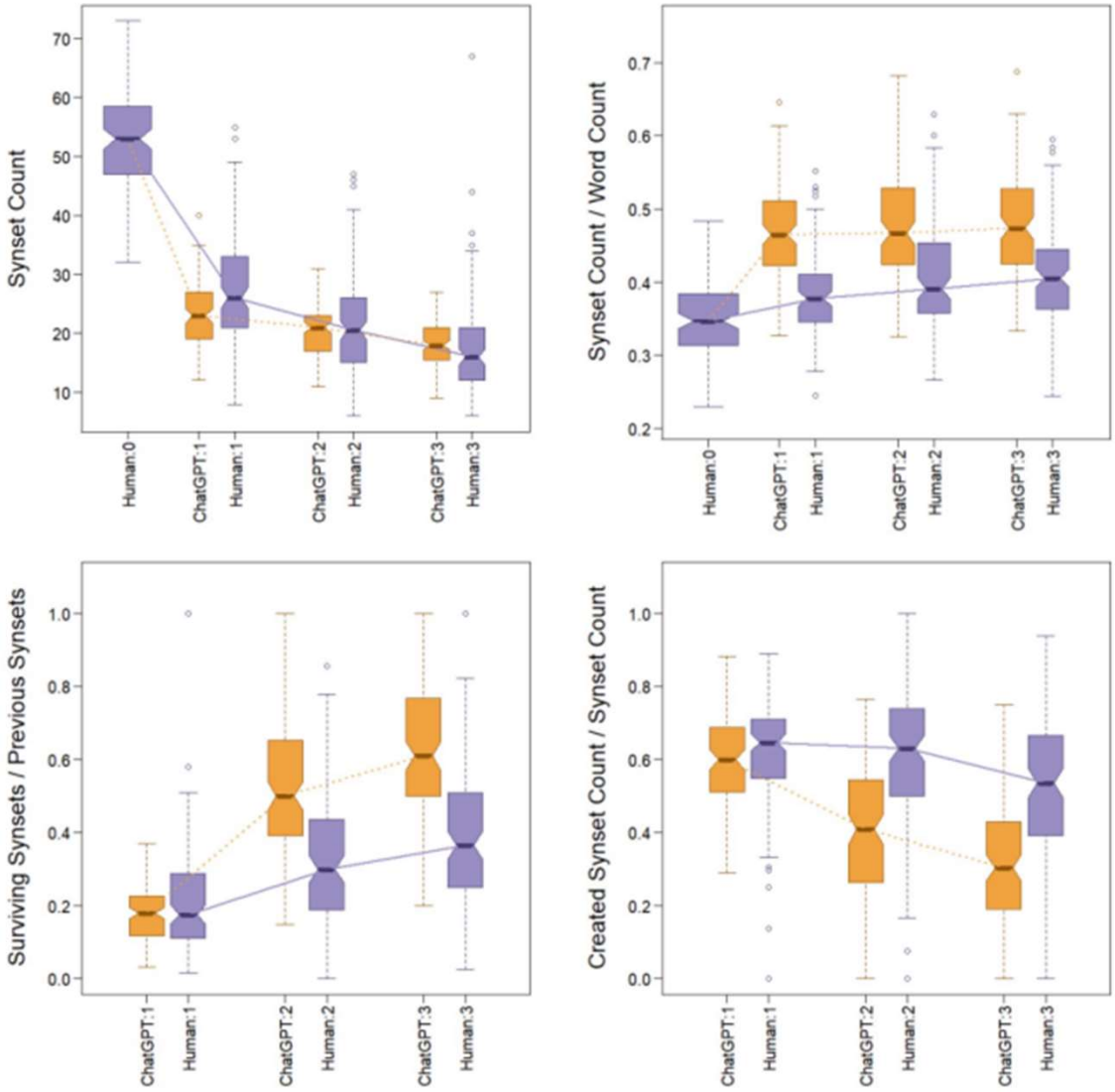


Figure 2. Synset data. Upper left: Counts of synsets. Upper right: Synset density, that is, synsets per word. Lower left: Synset survival, i.e., proportion of parent synsets that appear in retelling. Lower right: Proportion of novel or created synsets.

Concepts: creativity, stability, and decay

- 類似の傾向は、*lemmas*（特定のsynsetsよりも単語レベルに近い）と*ルート上位語 (root hypernyms)*（特定のsynsetsよりも高い概念レベルにあり、「fruit」は「apple」の正しい伝達として数えられることが多い）を分析するときにも発生する。

Age of acquisition

- 幼少期に獲得した単語はより安定した表現を持つため、より容易にかつ正確に思い出すことができる。(Juhasz, 2005)
  - このため、早期に獲得した単語は話者にとっては発音しやすく、聞き手にとっては理解しやすいため、言語コミュニケーションにおいて競争上の優位性を持つ可能性があります。
  - 早期に獲得した単語が好まれる傾向は、過去2世紀にわたる対人コミュニケーションと言語進化の両方で見られ (Ying et al., 2022)、前者のプロセスにおける認知的容易さが、複数世代の言語話者を通じて後者を生み出したと考えられる。
- ChatGPTは人間と同じ認知的制約を受けないため、まず、人間はChatGPTと比較して、物語を語り直すときに早期に習得した単語を保持する可能性が高いかどうかをテストした。
  - 各単語は、前の物語から保持されているかどうかでコード化された。
    - AOAは他のさまざまな共変量（例：頻度、単語の長さ、具体性）とともに、ロジスティック回帰における保持の予測因子として扱われました。
    - AOAの評価はKuperman et al. (2012; Brysbaert, M. & Biemiller., 2017)によって収集され、参加者に単語を習得したと思う年齢 (年) を思い出すように依頼した。

Age of acquisition

- AOA(獲得年齢)は ChatGPT の保存を予測しないが、人間では保存を負に予測する (つまり、AOA が高いほど保存が低くなる) ことがわかった。
- プロデューサー (人間/ChatGPT) とAOAの交互作用は統計的に有意だった。

- 物語を語り直すことは、前の物語の古い言葉を保存するだけでなく、新しい言葉を追加することも含まれる。
  - より全体的な視点から、各物語の平均AOA評価を計算し、セッションとプロデューサーごとにプロットした。
    - 図3に示すように、人間は語り直しを通じてAOAを元のレベルに保つ傾向がある。
    - ChatGPTは平均AOAを元のレベルよりも大幅に増加させ、その後、語り直しを通じてその高いAOAを維持していることがわかった。
    - 他の共変量では、人間とChatGPTの間に統計的に有意な差は見つからなかった。

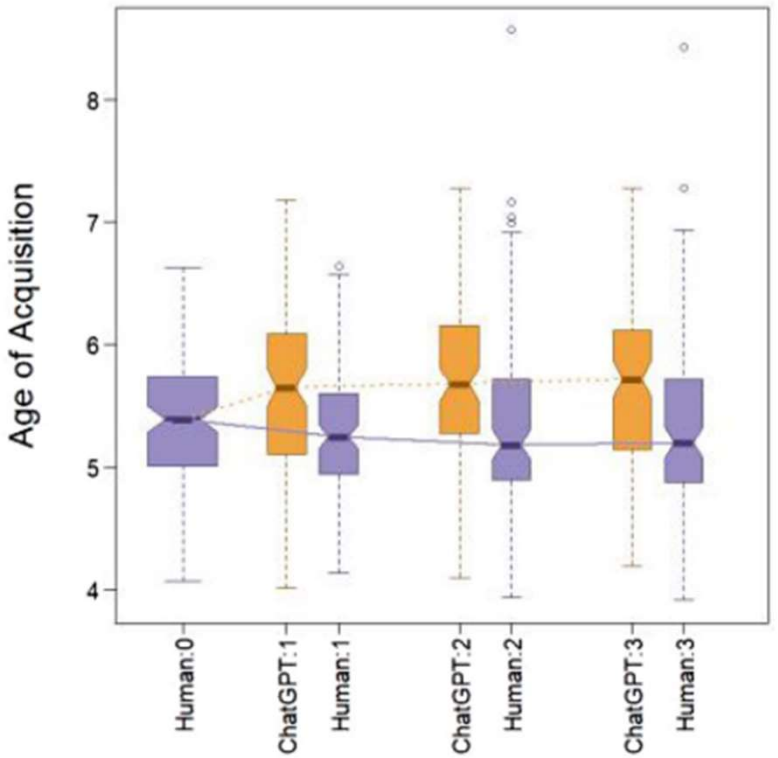


Figure 3. Age of acquisition.

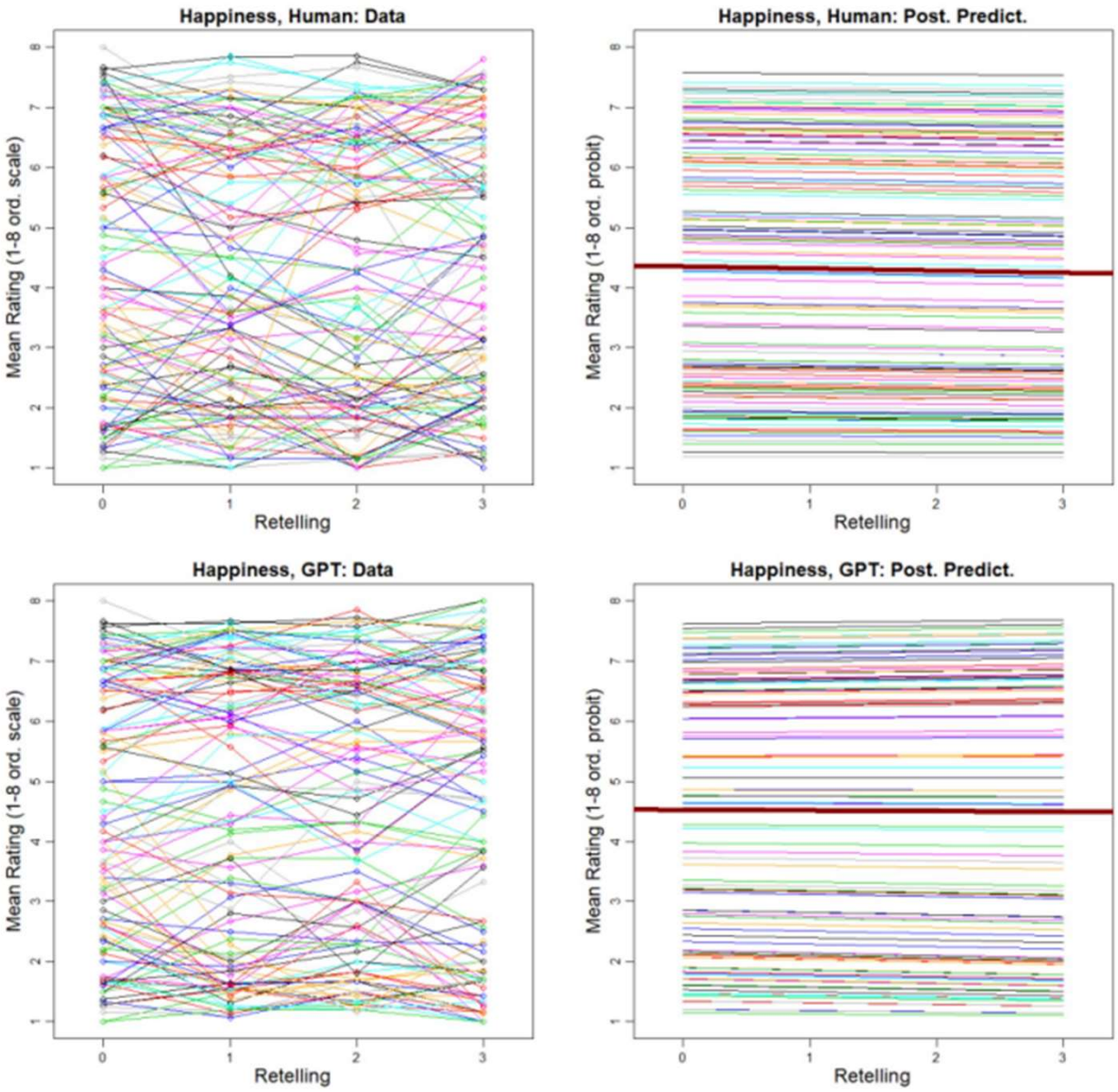
Affect preservation

- Breithaupt et al. (2022)で概説されている方法に基づいて、人間の評価者にすべての語り直しを評価するよう依頼した。
  - Breithaupt et al. (2022)が 492 人の参加者から収集した評価データセットを利用し、次の 4 つの項目を使用した。
    - (i) この状況はどれくらい幸せか? (ii) この状況はどれくらい悲しいか? (iii) 物語を読んでいる間、その中の出来事が容易に想像できた (Green, M. C. & Brock (2000)から引用)。(iv) 物語は私に感情的な影響を与えた。
    - ChatGPT の語り直しでは、元の研究と同じ 4 つの項目と評価手順を使用した。

- 私たちは、語り直し全体の傾向が全体的または典型的な傾向を表す線形の「背骨」を仮定し、個々の物語がそれを中心に収束または発散するという独自のモデル(Breithaupt et al., 2022)を使用して評価を分析した。
  - 線形傾向の傾きは、評価 (例: 幸福) が語り直し全体で上がるか、下がるか、または横ばいになるかを示す。
  - 線形傾向の「圧縮」は、個々の物語が背骨に向かって収束 (または発散) する度合いを示す。

Affect preservation

- 図4は、*happiness*の評価結果を示す。
  - あらゆる種類の評価について、人間による語り直し全体の傾向と ChatGPT による語り直し全体の傾向は驚くほど似ていることがわかった。
  - また、語り直し間での変化はほとんどなく、すべての傾きと圧縮は基本的にゼロであった(つまり、実質的な同等性は ± 0.1 に設定され、ゼロとほとんど変わらない) 。
    - この安定性は、図1で示されたストーリーの長さの大幅な短縮と対照的であり、注目に値する。


Happiness, Human: Data
Mean Rating (1-8 ord. scale)
Retelling
Happiness, Human: Post. Predict.
Mean Rating (1-8 ord. probit)
Retelling
Happiness, GPT: Data
Mean Rating (1-8 ord. scale)
Retelling
Happiness, GPT: Post. Predict.
Mean Rating (1-8 ord. probit)
Retelling

Discussion

- ChatGPT3と人間は、連鎖的に物語を語り直す際に異なる特徴を示した。
  - ChatGPTは、人間の認知能力を欠きながらも説得力のある人間の言語を作成できるモデルであるため、情報を伝え直す際の人間の認知の独自性を理解するためのリファレンスとして使用できる。
    - ChatGPTを参照しないと、人間の語り直しの創造性や持続性をどのように評価すればよいか不明瞭になる可能性があるが、ChatGPTを参照することで、人間の語り直しは確率的大規模言語モデルよりも著しく創造的であり、持続性が低いことがわかる。
  - 人間にとって非物語テキストと物語テキストはまったく異なる可能性があり、この違いが ChatGPT によってうまく捉えられていないことを示唆している可能性がある。

Discussion

- ChatGPT による凝縮された語り直しは、元の物語の感情的な核を高度に保存しており、幸せな物語と悲しい物語の両方で、また感情の強さの程度を問わず、わずかな違いしかない(図4)。
  - ChatGPT は、感情的な重要性を伴う物語の核となる状況に焦点を当てることで、印象的な感情の安定性を実現しているようである。
  - ChatGPT が最初の語り直しで核となるバージョンに到達すると、それ以降の反復では、新しい単語はほとんど作成されず、主に単語が同義語に置き換えられた。
- 対照的に、人間は連続した語り直しを通じて多くの変化を蓄積した。
  - 人間は、3回の語り直しすべてにおいて、新しいsynset（概念）の作成率が約55～60％を示した。
    - この率は、ChatGPTの相対的な安定性とは対照的に、特に高いことが際立っている。それにもかかわらず、人は、幸せな物語や悲しい物語、あらゆる程度の感情の強さにおいて、物語の感情的な核心を保持していた(図4) 。

Discussion

- 反復ごとに大部分の言葉が変更されるにもかかわらず、人間の語り手の感情的核心は安定しているようである。
- 我々は、人間の語り直しは 2 つの力のバランスをとるプロセスであると考えている。
  - 第一に、人々は、基本的なレベルのカテゴリ、人生の早い段階で習得した言葉、認知的に処理しやすい言葉、個人の経験や事前知識に基づいた新しい言語で、常に新しい適応した物語を創り出す。
  - 第二に、人々は物語の全体的な感情的側面を、物語の再構築と再創造のガイドとして使用する。
  - この組み合わせは、語り直しでは、以前のバージョンにはなかった言葉、文脈、出来事、状況であっても、物語の全体的な感情的核心に合うものが選択されることを意味している。
- 対照的に、ChatGPT は、反復を通じて元の物語の感情的な効果をもたらすことができる概念/ synsetの同一性または安定性に依存しているようである。

Discussion

- ChatGPT や同様の大規模言語モデルを研究で人間の代理として使用できるかについて、感情の安定性と社会的偏見(Acerbi & Stubbersfeld., 2023)の保存はこの可能性を示唆しているかもしれないが、創造性、品詞、獲得年齢、および潜在的な視点の変化の違いがこの可能性を著しく妨げている。
  - 現時点では、ChatGPT は個々の人間のパフォーマンスの独自性を測定するための標準を提供している。
- 私たちの研究は、安定していて広く使用され、以前の研究で広く研究されてきた ChatGPT3 を使用して実施された。
  - 大規模言語モデルは、今後も急速に進化し続けると思われる。
  - したがって、私たちの研究のポイントは、ChatGPT の現在の機能を説明することではなく、人間と ChatGPT の対比を利用して人間の認知を明らかにすることである。
    - これらの発見が LLM の改善に役立つ可能性もある。