ChatGPT が考える日本語ジョークの面白さ：人間との比較

中川 隼三郎　野元 裕樹
東京外国語大学
{nakagawa.junzaburo.v0, nomoto}@tufs.ac.jp

概要

本稿ではChatGPT (GPT-4o)と人間が日本語のジョークをどのように評価するかを比較する．人間が考えたジョークとChatGPTが生成したジョーク各9個（1つを除き全て対話形式）について，「面白さ」「不快さ」「わかりやすさ」の3つの観点での評価を調べた．その結果，ChatGPTの評価は人間より「面白さ」「わかりやすさ」において甘く，「不快さ」において厳しめであること，英語ジョークに関する類似研究とは逆に，ChatGPTの生成した日本語ジョークの評価は人間によるものより低くなることが分かった．その背景には人間にはないレベルの客観性，構成的意味計算能力の不十分さがあると主張する．

1 はじめに[i]

近年，様々な生成系AIが登場し，急速な勢いで進化を続けている．機械翻訳，文章要約などにおいて人間と同等あるいはそれ以上の優れた能力を発揮する生成系 AI であるが，苦手分野があることも知られている．例えば，ChatGPTに関して，感情の理解や共感の欠如，文化的背景の理解の欠如，創造性の限界といった弱点が指摘されている[2, 3]．ジョーク・ユーモア（以下「ジョーク」とする）の理解はそのような弱点が影響するだろうタスクの一つである．対話においてジョークを聞いた時，人間は「面白い」「不快だ」といった感情を抱く．感情には程度があり，ジョークの内容によってそれは異なる．ジョークを聞いた時に人間が持つこのような感情を生成系 AI どの程度有しているのであろうか？本稿ではChatGPT (GPT-4o) と人間に18の日本語のジョークを「面白さ」「不快さ」「わかりやすさ」の3つの観点での評価してもらい，その結果を比較する．生成系 AI のジョーク理解を扱う研究は英語のジョークについては存在するものの，日本語のジョークについては管見の限り見当たらない．

2 関連研究

[4] は，ChatGPTに対し，英語のジョークの生成および説明を行わせることによって，ChatGPTのジョークに関する能力の検証を行なっている．生成については，生成された全ジョークの90%以上が25種類のジョークの使いまわしであり，上位4位の頻度のジョークが全体の50%以上を占めると指摘している．ジョークはWhy …? Because ….という自問自答形式となっている．説明に関しては，ほぼ全てのジョークで有効な説明が得られたという．

[5] は，ChatGPT と人間が考えた英語のジョークを，クラウドソーシングにより集めた米国人200名に評価してもらい，ChatGPTと人間のジョーク生成能力を比較している．その結果，ChatGPTは感情を持たないにも関わらず，ChatGPTの考えたジョークの方が高い評価を受けたという．

ジョークを扱う言語資源としては，英語にはUR-FUNNY というデータセットが存在する [6]．UR-FUNNYは，テキスト，音声，視覚から成るマルチモーダルなデータセットであり，funnyとnon-funnyそれぞれ8,257の事例が収録されている．筆者の調べた限りでは，これに類似したデータセットは日本語ではまだ存在しない．

3 方法

本研究では人間が考えたジョーク9とChatGPT (GPT-4o) が考えたジョーク9の計18の日本語ジョークをChatGPT自身と人間に評価してもらう．

3.1 調査対象とした日本語ジョーク

表 1 に調査対象とした日本語ジョークをまとめる．IDの(a), (b), (c)は後述するカテゴリーを示す．

[i] 本研究の大部分は第一著者の卒業論文に基づく [1].

表 1　調査対象とした日本語ジョーク

ID	ジョーク
H1 (a)	「幸せなら手を叩こうっていうけどさ，不幸せな奴は何を叩けばいいの？」 「...与党...」
H2 (a)	「もしもし，健太くん？あの，お父さんいる？」 「いらなーい！」[7]
H3 (a)	医者　「ううむ，スミスさん．あなたは妊娠しているみたいですね」 スミス夫人「ああ神様，なんて素晴らしい．私，妊娠したんですね？」 医者　「妊娠しているように見えると言ったんです．減量しましょう」[8]
H4 (b)	ある職場にて 先輩「新人くん，今日は腹を割って話そうじゃないか」 新人「切腹しろってことですか？」
H5 (b)	ある部活の引退試合にて 「今日は先輩の引退だし，花持たせなきゃなぁ」 「え！？すみません，今すぐ花束買ってきます！」
H6 (b)	妻「あなたがこんなに貧乏だって分かっていたら絶対結婚しなかったわ」 夫「君は僕の全てだって結婚前に何度も言ったじゃないか！」[8]
H7 (b)	妻は僕に「あなたは 100 万人に 1 人の男よ」と言った． ある日，妻のスマホを覗き見したら確かにその通りだった．[8]
H8 (b)	「一日一個のりんごは医者を遠ざけるっていうけど，本当かな？」 「ああ，よく狙って投げればね」[8]
H9 (c)	医師「旦那さんは絶対安静が必要です．ここに睡眠薬がありますので，飲んでください」 妻　「いつ旦那に飲ませたらいいですか？」 医師「いいえ，あなたがこれを飲むんです」[9]
C1 (a)	「ねぇ，カフェで働いてる友達が辞めたらしいよ．」 「なんで？」 「もう我慢の限界だったんだってさ．毎日，マメに働くのが疲れたんだって！」
C2 (a)	「最近，友達がパン屋を辞めたんだって．」 「どうして？」 「仕事がつらすぎて，もうパンパンだったらしい．」[ii]
C3 (b)	「最近，漁師の友達が仕事を辞めたんだって．」 「どうして？」 「上司がいつも『足を引っ張るな』って言うけど，カニを獲る仕事だから無理だってさ！」
C4 (b)	「昨日，友達が急に道で倒れたんだ！」 「え，大丈夫だったの？」 「うん，本人いわく『足元をすくわれた』らしいけど，転んだ原因はバナナの皮だったよ．」
C5 (b)	「昨日，友達が『頭が上がらない』って言うから何があったのか聞いたんだ．」 「それで，何があったの？」 「ただ単に首を寝違えただけだったよ！」
C6 (b)	「昨日，友達が急に電話で泣き出したんだ．」 「どうしたの？」 「本人いわく，『胸が張り裂けそう』だったらしい．でもよく聞いたら，ボタンシャツのボタンが全部飛んだだけだったよ．」
C7 (b)	「医者の友達が『目が回るほど忙しい』って言ってたんだ．」 「大変そうだね．それでどうなったの？」 「次に会ったら，自分で三半規管の検査してたよ．」
C8 (b)	「この前，友達が『猫の手も借りたい』って言うから，ペットショップで猫を借りてきたんだ．」 「で，どうなったの？」 「部屋が毛だらけになっただけで，全然役に立たなかったよ．」
C9 (b)	「この前，友達が『犬も歩けば棒に当たる』って言うから，試しに犬の散歩をしてみたんだ．」 「どうだった？」 「棒どころか，電柱に当たったよ．こっちがね．」

[ii] 「パン」と「パンパン」は別の語なので，ChatGPT は多義性を利用したジョークの生成に失敗している．

評価基準
「面白さ：「１非常につまらない」「２ややつまらない」「３どちらでもない」「４やや面白い」「５とても面白い」
不快さ：「１全く不快ではない」「２あまり不快ではない」「３どちらでもない」「４やや不快」「５非常に不快」
わかりやすさ：「１非常にわかりやすい」「２ややわかりやすい」「３どちらでもない」「４ややわかりづらい」「５非常にわかりづらい」
この採点基準で満点を取れるようなジョークを考えてください．なお，そのジョークは単語の多義性・慣用表現・ステレオタイプを利用したものでお願いします．

図 1　ジョーク生成に用いたプロンプト

H7 以外は全て対話形式になっている．H1～H9 は人間が考えたもので C1～C9 は ChatGPT 自身が考えたものである．H1 は理想郷計画【Utopia project】という YouTuber が 2024 年 11 月 13 に YouTube shorts に投稿した動画より取得した[iii]．H3 および H6～H9 は，英語のジョークを日本語に翻訳したもので（H9 は筆者訳），H4 および H5 は筆者が独自に考えた．C1～C9 は図 1 に示したプロンプトにより生成した．

これらのジョークは背景にある言語的・非言語的事象により，(a) 多義性，(b) 慣用表現，(c) ステレオタイプの 3 つに分類することができる．

(a) 多義性　話し手・聞き手双方が多義語を意義 s1 で用いているという共通信念が，想定外の別の意義 s2 を用いた発話により覆されることを利用[iv]．例えば，H1 では「叩く」を最初は「打つ」という意味で使用しているが，次に出てくる「叩く」は「批判する」という意味で使用している．

(b) 慣用表現　話し手・聞き手双方が慣用表現を慣用的意味で用いているという共通信念が，想定外に字義通りの意味を用いた発話により覆されることを利用．例えば，H4 では先輩は「腹を割って話す」という慣用表現を「本心を打ち明ける」という慣用的意味で用いているのに，新人は字義通りの意味で用いている．

(c) ステレオタイプ　話し手・聞き手双方が社会における役割・立場に紐づいたステレオタイプを前提として会話しているという共通信念が，ステレオタイプを前提としない発話により覆されることを利用．例えば，H9 では医師が共通基盤にあると想定する「女性はおしゃべりだ」というステレオタイプが実は妻とは共有されていないために，誤解が生じている．当該ステレオタイプを知らないと，「夫を安静にさせるためにはおしゃべりな妻を眠らせる必要がある」という医師の説明は理解できない．

3.2 ジョークの評価方法

ChatGPT によるジョークの評価に関しては，図 2 の形式のプロンプトを用いた．C1～C9 は満点を取れるようにという指示を与えて ChatGPT 自身に生成させたものであるが，それらのジョークを再度 ChatGPT に評価させた．

[ジョーク]
上記のジョークを，「面白さ」「不快さ」「わかりやすさ」の 3 つの観点において，評価をお願いします．またジョークのどこが面白いポイントとなっているのかの解説もお願いします．
（図 1 に**太字**で示した評価基準）

図 2　ジョークの評価に用いたプロンプト

人間によるジョークの評価は，図 1 に**太字**で示した評価基準を用い，Google フォームにより調査した．回答者は 20 代大学生 31 名である．

4 結果と考察

表 2 および表 3 はそれぞれ人間，ChatGPT が考えた日本語ジョークに対する評価をまとめたものである．人間による評価は平均値を示す（より詳細な統計量は付録を参照）．また，ChatGPT と人間の評価に 1 ポイント以上の差があるセルの値は太字により強調する．

まず，「面白さ」「わかりやすさ」は全てのジョークに関して，ChatGPT が人間よりも高い評価をしている．そのうち，「わかりやすさ」は全てのジョークが満点 5 の評価となっている．一方，「不快さ」に関しては，ChatGPT は人間が考えたジョークのうち 3 つについて人間よりも 1 ポイント以上高い「不

[iii] https://youtube.com/shorts/MWTumxRu9tI?si=JCBhb3kdoaMrY8ml

[iv] 多義は構造的曖昧性に基づくものもでもよいが，表 1 にはそのようなジョークは含まれていない．

表 2 人間が考えた日本語ジョークの評価

ID	面白さ		不快さ		わかりやすさ	
	C	人	C	人	C	人
H1 (a)	4	3.84	**3**	**1.81**	**5**	**3.45**
H2 (a)	4	3.29	**1**	**2.26**	**5**	**3.84**
H3 (a)	4	3.45	2	2.13	5	4.13
H4 (b)	**4**	**2.61**	1	1.65	5	4.19
H5 (b)	**4**	**2.71**	**3**	**1.58**	5	4.10
H6 (b)	4	3.06	2	1.71	**5**	**2.94**
H7 (b)	**4**	**2.52**	**3**	**1.68**	**5**	**2.16**
H8 (b)	**5**	**3.29**	1	1.55	**5**	**3.26**
H9 (c)	4	3.19	2	1.97	**5**	**2.74**

表 3 ChatGPT が考えた日本語ジョークの評価

ID	面白さ		不快さ		わかりやすさ	
	C	人	C	人	C	人
C1 (a)	**5**	**3.10**	1	1.35	**5**	**3.97**
C2 (a)	**5**	**2.39**	1	1.48	5	4.10
C3 (b)	**5**	**2.51**	1	1.45	**5**	**2.84**
C4 (b)	**5**	**2.03**	1	1.39	**5**	**2.77**
C5 (b)	**5**	**2.19**	1	1.48	**5**	**3.42**
C6 (b)	**5**	**2.42**	1	1.65	**5**	**3.48**
C7 (b)	**5**	**2.61**	1	1.35	**5**	**3.61**
C8 (b)	**5**	**2.06**	1	1.42	5	4.10
C9 (b)	**5**	**2.77**	1	1.39	**5**	**3.23**

快さ」「わかりやすさ」に関しては人間よりも評価が甘く，「不快さ」に関しては人間よりも評価が厳しめであると言える．ChatGPT が「わかりやすさ」で高い評価をする理由の一つは客観性だろう．普通，人間はある言語形式に一つの意味（あるいは構造）を結び付けると，他に存在する別の意味には注意が向けられなくなってしまう．これはだまし絵を見る時に生じるのと同じ心理的作用である．多義性や慣用表現に基づくジョークを理解するには，その別の意味に気付くというハードルを乗り越えなければならない．しかし，ChatGPT の場合，おそらく複数の意味に平等に注意を当てることができるため，そのようなハードルが存在しないのだろう．

次に，ChatGPT 自身が生成したジョークに目を向けてみる（表 3）．「不快さ」に関しては顕著な差がない一方，「面白さ」「わかりやすさ」では人間の評価との間に顕著な差が確認された．人間は ChatGPT が生成したジョークの多くを面白くない，分かりづらいと感じていることが見て取れる．特に，「面白さ」では C1 以外の全てのジョークで ChatGPT 自身と人間の評価に 2 ポイント以上の差がある．この点は，英語の類似研究の結果と逆である．2 節で見たように，[5] は，英語ジョークに関して，人間が ChatGPT の考えたジョークを人間の考えたものより高く評価したと報告している．

ChatGPT の日本語ジョークを人間が面白く感じない理由はいくつか考えられる．第一に，多くのジョークはともすると「不快さ」につながるようなきわどさを孕んでいるが，ChatGPT のジョークにはそれがない．例えば，人間が考えた H1 には社会風刺的な要素が含まれる．この点は ChatGPT が「不快さ」に対して厳しめであることとも整合する．第二に，日本語慣用表現の字義通りの意味がきちんと理解できていないことがある．転んだ時に「足元をすくわれる」(C4)，寝違えた時に「頭が上がらない」(C5) と言うことはほぼないだろう．字義通りの意味は各要素の意味を構成して得られるが，構成的意味計算は ChatGPT には難しいようである．

5 おわりに

本稿では，日本語ジョークを対象に ChatGPT と人間の評価の違いを明らかにした．その中には ChatGPT 自体が生成したジョークも含まれ，英語ジョークを対象とした先行研究とは逆に，ChatGPT の考えた日本語ジョークは人間による評価が人間の考えたジョークよりも低いということを示した．

最後に，本研究には少なくとも 3 点の欠点が存在する．第一に，調査対象のジョーク数がわずか 18 と少ない．第二に，ジョークのカテゴリーが偏っている．ほとんどが慣用表現が関与するもので，ステレオタイプが関与するものは 1 つしかない．また多義性は，語彙的曖昧性に基づくものだけで，構造的曖昧性に由来するものがない．これらの問題は，UR-FUNNY [6] に相当する日本語ジョークのデータセットが構築されると解決が容易になるはずである．

第三の問題は，人間による評価実験の回答者が全て大学生で属性に偏りがあることである．[5] がそうしたようにクラウドソーシングを用いてより多くの多様な属性の回答者を対象とすれば，人間の日本語ジョーク評価のより包括的な理解につながるはずである．

参考文献

[1] 中川隼三郎. 2025.『AI と人間の境界線～ChatGPT の得意分野と苦手分野を通して見る「人間を人間たらしめているもの」とは～』東京外国語大学卒業論文.

[2] Abujaber, Ahmad. A., Alaa Abd-alrazaq, Ahmad R. Al-Qudimat and Abdulqadir J. Nashwan. 2023. A strengths, weaknesses, opportunities, and threats (SWOT) analysis of ChatGPT integration in nursing education: A narrative review. *Cureus* 15(11): e48643. https://doi.org/10.7759/cureus.48643

[3] Kalla, Dinesh, Nathan Smith, Fnu Samaah and Sivaraju Kuraku. 2023. Study and analysis of Chat GPT and its impact on different fields of study. *International Journal of Innovative Science and Research Technology* 8(3): 827–833. https://ssrn.com/abstract=4402499

[4] Jentzsch, Sophie and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 325–340. Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.29

[5] Gorenz, Drew and Norbert Schwarz. 2024. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. *PloS ONE* 19(7): e0305364. https://doi.org/10.1371/journal.pone.0305364

[6] Hasan, Md Kamrul, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2046–2056. Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-1211/

[7] 田中雅敏. 2017.「ジョークの語用論的考察と異文化理解への応用」『東洋法学』61(2): 321–334.

[8] JAPAB JOURNALS. 2020.「イギリスの選り抜きジョーク集」『ONLINE ジャーニー』. https://www.japanjournals.com/life/english-jokes.html（最終アクセス日:2025 年 1 月 9 日）

[9] 中野清治 (2002)「ジョークの中の人間模様」『高岡短期大学紀要』 17: 139–148. https://doi.org/10.15099/00007400

A 付録

以下の表はそれぞれ, 人間が考えた日本語ジョーク（H1~H9）, ChatGPT が考えた日本語ジョーク（C1~C9）に対する人間による評価の詳細をまとめたものである.

表 4 人間が考えた日本語ジョークに対する人間による評価

ID	面白さ				不快さ				わかりやすさ			
	平均値	中央値	最頻値	SD	平均値	中央値	最頻値	SD	平均値	中央値	最頻値	SD
H1	3.84	4	4/5	0.74	1.81	1	1	1.06	3.45	4	3/4/5	1.27
H2	3.29	4	4	0.99	2.26	2	1	1.27	3.84	4	4	0.95
H3	3.45	4	4	0.87	2.13	2	1/2	1.10	4.13	4	4	0.86
H4	2.61	2	2	1.18	1.65	1	1	1.09	4.19	4	5	0.78
H5	2.71	3	4	1.04	1.58	1	1	1.01	4.10	4	4	0.89
H6	3.06	3	4	1.36	1.71	1	1	1.19	2.94	3	4	1.22
H7	2.52	2	1	1.23	1.68	1	1	1.09	2.16	2	1	1.19
H8	3.29	3	4	1.03	1.55	1	1	0.91	3.26	3	4	1.32
H9	3.19	3	4	1.03	1.97	1	1	1.31	2.74	4	4	1.19

表 5 ChatGPT が考えた日本語ジョークに対する人間による評価

ID	面白さ				不快さ				わかりやすさ			
	平均値	中央値	最頻値	SD	平均値	中央値	最頻値	SD	平均値	中央値	最頻値	SD
C1	3.10	3	2	1.18	1.35	1	1	0.90	3.97	4	4	0.97
C2	2.39	2	2	1.13	1.48	1	1	1.01	4.10	4	5	0.86
C3	2.51	2	2	1.07	1.45	1	1	0.98	2.84	3	2	1.27
C4	2.03	2	1	0.98	1.39	1	1	0.90	2.77	3	1	1.43
C5	2.19	2	1	1.09	1.48	1	1	0.91	3.42	4	4	1.21
C6	2.42	2	2	1.09	1.65	1	1	1.06	3.48	4	4	1.16
C7	2.61	2	4	1.21	1.35	1	1	0.82	3.61	4	4	1.10
C8	2.06	2	1	1.17	1.42	1	1	0.94	4.10	4	4	0.86
C9	2.77	3	2	1.26	1.39	1	1	0.87	3.23	3	3	1.36