

人間と LLM の“面白さ”の感性は一致するのか？

坂部立¹ 金輝燦² 小町守¹

¹一橋大学 ²東京都立大学

{dm240011@g.hit-u, kim-hwchan@ed.tmu, mamoru.komachi@r.hit-u}.ac.jp

概要

計算機によるユーモアの研究は、対話システムなどの自然言語処理ツールの高度化に不可欠とされている一方で、大規模言語モデル (Large Language Model; LLM) のユーモア生成や評価に関してはまだ十分に検討されていない。本研究では、図 1 のように大喜利を題材として新たに高品質な日本語の大喜利データセットを構築し、LLM の大喜利生成能力と大喜利評価能力を検証した。

1 はじめに

近年、LLM の飛躍的な発展により、自然言語処理の多様なタスクを高精度で遂行可能となりつつある。しかし、Greengross ら [1] が示したように人間の知性とユーモアセンスは密接に関係している一方で、先行研究では LLM がジョークの検出・説明については信頼に欠けると述べており [2, 3]、またジョーク生成の難しさを指摘している [4]。

日本には、お題に対して回答の面白さを競う競技である大喜利という演芸がある。Zhong ら [5] は大喜利を題材とし、Bokete などのサイトからデータを収集し、日本語・中国語・英語のマルチリンガルな大喜利データセットである Oogiri-GO Dataset を構築した。彼らはこのデータセットを用いて様々な LLM の大喜利回答生成能力と回答の相対評価能力を分析している。結果としては既存の LLM は大喜利回答を生成・相対評価する能力はどちらも低いことを示している。一方、LLM が生成した大喜利回答の人手での評価や、LLM による大喜利回答の絶対評価、日本語単体での評価は行われていない。また、Oogiri-GO Dataset には 2 つの主要な問題点がある。第一に、Bokete などのサイトは投票時に既存の得票数が表示されるため、純粋な面白さの評価が反映されない可能性がある。第二に、テキストのお題が画像として入力されるため、VLM の OCR 能力が結果に影響を与え、LLM の真の大喜利能力を正確

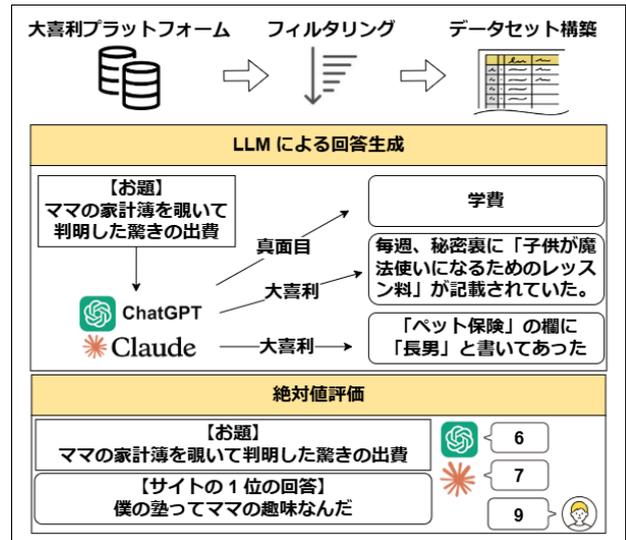


図 1 本研究の概要図

に測れない可能性がある。

そこで、本研究ではこれらの問題点を解消するために新たに日本語のデータセットを構築し、LLM の日本語大喜利生成・評価能力を測った。まず、回答および投票の公平性が確保されており、テキストのお題はテキストとして取得できる大喜利サイトである“大喜利茶屋”¹⁾を使いデータセットを構築した。次に、作成したデータセットを用いて、真面目な回答と大喜利回答を生成させ、回答を定性的に分析した。さらに回答の絶対評価を行い、人手での評価とも比較することで、人間と LLM の“面白さ”の感性が一致するかどうかを検証した。

本研究の主な貢献は以下のとおりである。

- 公平性の担保された大喜利サイトから、日本語の大喜利データを収集・フィルタリングし、高品質な大喜利データセットを新たに構築した。
- 構築したデータセットに対して、LLM による大喜利回答の生成を行い、日本語大喜利における生成特性を定性的に分析し、人間の大喜利回答との違いやユーモア生成の可能性を示した。

1) <https://oogiri-chaya.com>

- LLM とクラウドソーシングで絶対値評価を行い、LLM と人間の評価基準がどの程度一致しているかを分析し、LLM が人間と異なる基準で大喜利回答を高く評価する傾向を示した。

2 関連研究

大喜利を題材とした研究として、中川ら [6] は本研究とは異なり、大喜利の面白さの構成要素をクラウドソーシングにより評価付けし、その結果を用いて回帰モデルを訓練することで有効な特徴量として“関係性”、“わかりやすさ”、“新しさ”が面白さに最も寄与していると示した。

Horvitz ら [7] は、LLM が風刺ニュースの見出しのユーモア性を取り除くことができる一方で、ユーモアのある風刺見出しの生成は困難であることを示した。さらに、LLM を用いた大喜利の研究として、Zhong ら [5] は生成については、提案手法のモデルを含んだ 6 種類の VLM によって生成された 6 件の大喜利回答から最も創造的でユーモラスな回答を選択するタスクを行っている。評価については“真の回答”、“BLIP-2 [8] による画像のキャプション”、“他の画像のキャプション”の 3 つの選択肢から“真の回答”を選ばせる 3 択タスクにおいて、GPT-4 は正解率が 19.3% に留まり、提案手法においても 41.8% 程度の精度しか得られておらず、既存のモデルは大喜利の相対評価能力が低いことを示している。

3 大喜利データセット

本研究では、大喜利サイトには大喜利茶屋を選んだ。このサイトは、回答および投票時に他者の回答や投票内容が非公開となる仕組みを採用しており、回答と投票の公平性が担保されている。大喜利のフォーマットはテキストのみのお題に対してテキストで回答するもの、テキストに加えて画像を含むお題に対してテキストで回答するものの 2 種類存在する。

大喜利茶屋から 2021 年 11 月 21 日から 2024 年 7 月 17 日までの期間のデータをクロールした。クロールしたデータに対して、「お題に対する回答数が多い場合は多様な回答から選ばれた高品質な回答が得られる」と考え、回答数が 100 件未満のお題をフィルタリングした。次に、収集したデータのお題と回答に対して GPT-4o-2024-05-13 を用いて差別的表現、過度に暴力的な表現、性的に露骨な表現、その他社会的に不適切と考えられる表現が含ま

	【サイトの 1 位の回答】 最低賃金の人にインタビューしてるんですが
テキストのみのお題 この後店員に殴られるおじさん「すみません、○○○」	【別のお題の 1 位の回答】 また僕の知らない野菜だ
	【GPT-4o 真面目回答】 すみません、返品をお願いします
	【GPT-4o 大喜利回答】 すみません、著少々お待ちください。今練習中です。
	【Claude 3.5 Sonnet 大喜利回答】 すみません、『5年のご愛顧、ありがとうございます』の横断幕、今から持って帰ります
画像付きのお題 画像で一言 	【サイトの 1 位の回答】 一步間違えば、の方やります
	【別のお題の 1 位の回答】 ピザみたいに取り出される
	【GPT-4o 真面目回答】 危険。
	【GPT-4o 大喜利回答】 パタフライ効果ってこういうこと!?
	【Claude 3.5 Sonnet 大喜利回答】 まだまだ若いと思って、クッションを思いっきり投げる銀行の強盗通知書

図 2 LLM により真面目回答と大喜利回答を生成した後のデータセットの例。

れるお題と回答に対して、ラベルを割り振った。そのラベルが付与されたお題と回答を人手で確認し、不適切と判断した場合は除去した。さらに、人手確認時に作成した不適切語リストに基づくルールベースでの追加フィルタリングを実施した。

これらの処理を経て、最終的なデータセットはテキストのみのお題 425 件、画像付きのお題 126 件の計 551 件となった。

4 実験

本研究では、作成したデータセットを用いて大喜利生成能力と大喜利評価能力の 2 つの観点から LLM の大喜利能力を分析する。前者はお題と同時にユーモアを含まない回答、または、ユーモアを含む回答を生成させるためのプロンプトを与え、プロンプトごとにどのような回答が生成されるかを分析する。後者は、お題に対する回答を 1 つ入力し、0 から 10 点で採点させる**絶対値評価タスク**を行う。

4.1 実験設定

モデル LLM としては GPT-4o-2024-08-06 (以下 GPT)、Claude 3.5 Sonnet v2 (以下 Claude) を用いた。

大喜利生成 2 種類のプロンプトを作成し LLM の回答の生成能力を測る²⁾。まず、ユーモアを含まない真面目な回答を生成させるために、お題の回答としてユーモアを含まないような指示を含むプロンプトを作成した。次に、ユーモアを含む大喜利回答を生成させるために、中川ら [6] による分析を参考

2) プロンプトは付録 A.1 に示す。

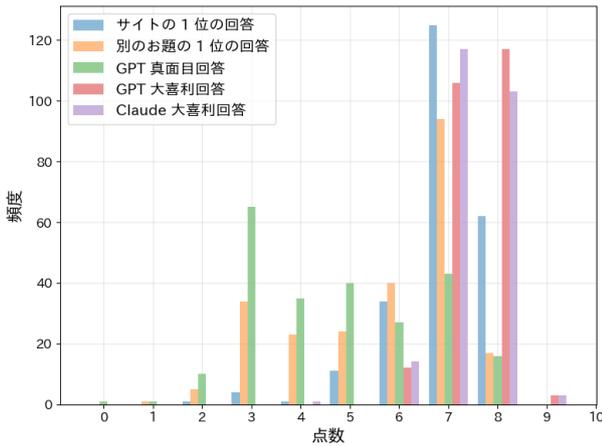


図3 GPT-4oによる絶対値評価の点数分布 ($n = 238$).

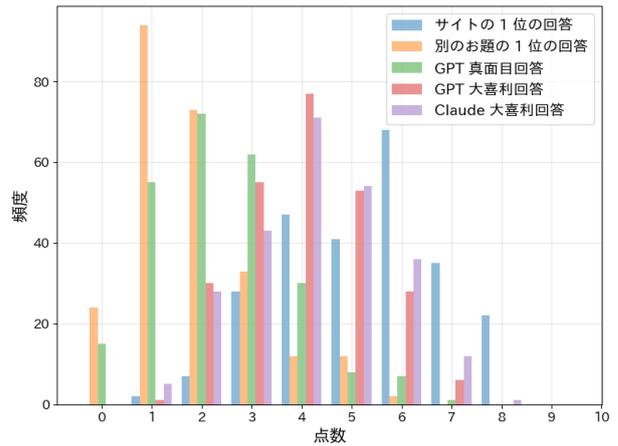


図5 人手による評価の点数分布 ($n = 250$).

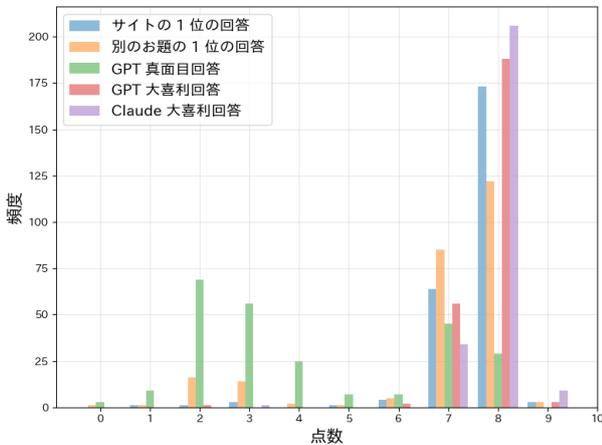


図4 Claude 3.5 Sonnetによる評価の点数分布 ($n = 250$).

に“関係性”，“わかりやすさ”，“新しさ”を考慮したプロンプトを構築した。構築したデータセットの各お題に対して上記の2種類のプロンプトで回答を生成させ、プロンプトにより回答の面白さを制御できるかを定性的に分析した。

大喜利評価 上で生成させた回答を用いて、絶対評価として各回答がLLMにどの程度好まれるのかを定量的に示すため、各お題に対して“サイトの1位の回答”，“GPT真面目回答”，“別のお題の1位の回答”，“GPT大喜利回答”，“Claude大喜利回答”の5つの回答を1つずつ入力し、0から10点で点数を付けさせる絶対値評価タスクを行った。図2に各回答の例を示す。

また、人間とLLMの大喜利回答の評価の一致度を測るために、同様のタスクをクラウドソーシングにより人手でも行った。上記の5つの回答が50件ずつ含まれるように、テキストのみのお題200件と画像付きのお題50件の計250件のサブセットを

5つ作成し、それぞれ5人の評価者が絶対値評価を行った。³⁾各回答に対して5人の平均をその回答の点数とした。

4.2 実験結果

4.2.1 大喜利生成

真面目な回答を生成させるプロンプトを与えた場合、概ねユーモアを含まない真面目な回答が生成された。一方で大喜利回答を生成させるプロンプトを与えた場合には、ユーモアを含む面白い回答が得られることもあった。付録B.1に面白い回答の例を示す。この結果は、プロンプトを工夫することで回答の面白さを制御できる可能性を示唆している。

しかし、現状のプロンプトで生成された大喜利回答にはいくつかの課題が見られる。まず、お題に関連する単語の連想はできているものの、内容が飛躍が大きく、お題との“関係性”を十分に考慮できていない回答が多く見られた。次に、LLMは長文の回答を生成しやすい傾向⁴⁾にあり、“わかりやすさ”の観点で改善の余地がある。これらの課題を解決するためには、プロンプトの試行錯誤や追加学習の実施が必要であると考えられる。

4.2.2 大喜利評価

LLMによる評価 LLMの絶対値評価のヒストグラムを図3、4に示す。⁵⁾“サイトの1位の回答”は7

3) サブセットの詳細な作成手順は付録A.2に示す。

4) “サイトの1位の回答”の平均文字数は13.3字だった一方で、“GPT大喜利回答”，“Claude大喜利回答”の平均文字数はそれぞれ33.0字、30.4字と、人間よりも長文で回答する傾向があった。

5) GPTは12件で点数を付けられなかった。

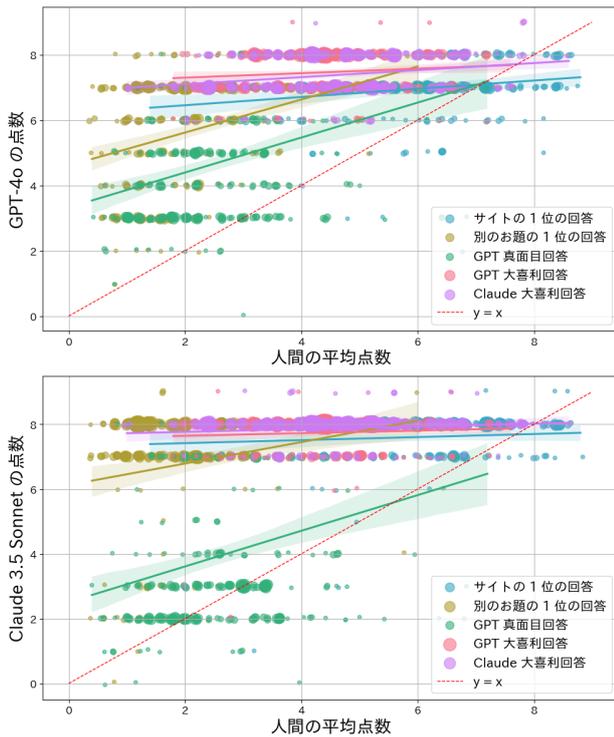


図6 人間とLLMが付与した点数の散布図。

点以上の割合が高く、ユーモアを認識できていると言える。“GPT 真面目回答”は3点程度の低い点数が多い一方で7点付近の割合も高く、真面目な回答の分類には課題があることを示している。また、“別のお題の1位の回答”を高く評価する傾向も見られ、生成同様お題との“関係性”を考慮せず、回答単体の面白さのみを評価している可能性がある。加えて、両LLMともにLLMの大喜利回答に対して“サイトの1位の回答”よりも高い点数を付ける割合が高く、LLMは自身大喜利回答を好む傾向が見られた。

人手による評価 人手評価のヒストグラムを図5に示す。“サイトの1位の回答”は6点付近の割合が高いが分散も大きく、ユーモアの主観性がうかがえる。“GPT 真面目回答”は2点付近が多く、GPTは真面目な回答を概ね生成できることを示している。また、人間は“別のお題の1位の回答”を、“GPT 真面目回答”よりも低く評価しており、これはLLMの評価とは異なる傾向である。人間はお題との“関係性”を重視し、LLMは回答自体のユーモアを重視していると推察される。さらに人間は、LLMの大喜利回答については“サイトの1位の回答”には劣るものの高得点を与えている場合があり、特にClaudeの大喜利回答をより評価する傾向が確認できた。⁶⁾

6) 人間による250件の評価において、Claudeの大喜利回答は、サイトの1位の回答を上回った件数が70件、GPTの回



図7 1サブセットのピアソンの相関係数 (n = 238)。

LLMと人間の評価の比較 人間とLLMが付与した点数の散布図を図6に示す。全体のピアソンの相関係数はGPTが0.49、Claudeが0.38であり、GPTの方が人間との一致度は高かった。また、“別のお題の1位の回答”、“GPT 真面目回答”の相関係数は、それぞれGPTが0.36、0.40、Claudeが0.21、0.32であり他の回答より高い値を示した。お題によっては“別のお題の1位の回答”、“GPT 真面目回答”を人間も面白く感じることもあり、それらを高く評価してしまうことで相関係数が高まった可能性がある。

次に、評価者とLLMのピアソンの相関係数を図7に示す。評価者同士と比べ、LLM同士は相関が高く、LLMに共通する採点傾向があることがうかがえる。また、評価者平均とLLMの相関を見ると、GPTが0.57、Claudeが0.44であり、やはりGPTの方が人間との一致度が高い。加えて、絶対値評価を基にランキングを作成した場合でもGPTの方が人間との一致度は高かった(付録B.2参照)。

5 おわりに

本研究では、高品質な大喜利データセットを構築し、人間とLLMの大喜利生成・評価能力の一致度について分析した。その結果、LLMの大喜利生成能力に関しては、総じて人間の大喜利回答には及ばない傾向にあるものの、人間の評価が高い回答を生成できる例も一部に認められ、LLMによるユーモア生成の可能性が示唆された。LLMの大喜利評価能力に関しては、人間の大喜利回答を評価する能力において人間とLLMで一致度は低いことがわかった。

答を上回った件数133件であった。GPTはそれぞれ61件、108件であった。

謝辞

本研究の一部は JST さきがけ JPMJPR2366 の支援を受けたものである。

参考文献

- [1] Gil Greengross and Geoffrey Miller. Humor ability reveals intelligence, predicts mating success, and is higher in males. **Intelligence**, Vol. 39, No. 4, pp. 188–192, 2011.
- [2] Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. You told me that joke twice: A systematic investigation of transferability and robustness of humor detection models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13701–13715, Singapore, December 2023. Association for Computational Linguistics.
- [3] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [4] Sophie Jentzsch and Kristian Kersting. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In Jeremy Barnes, Orphée De Clercq, and Roman Klinger, editors, **Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis**, pp. 325–340, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let’s Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation . In **2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 13246–13257, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.
- [6] 中川裕貴, 村脇有吾, 河原大輔, 黒橋禎夫. クラウドソーシングによる大喜利の面白さの構成要素の分析. 言語処理学会 第 25 回年次大会 発表論文集, 2019.
- [7] Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. Getting serious about humor: Crafting humor datasets with unfunny large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 855–869, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In **Pro-**

真面目回答生成のプロンプト
(<画像>) 大喜利のお題: <お題> このお題に対して真面目に回答してください。
大喜利回答生成のプロンプト
以下に大喜利のお題を提示します。以下の制約に従い、面白い回答を1つ生成してください。 # 制約 1. お題の内容を踏まえた上で、お題に対するボケとしてわかりやすく、新しさを感じるような回答を生成してください。 2. お題は復唱せず、大喜利の回答以外の内容は生成しないでください。
(<画像>) お題: <お題>
絶対値評価のプロンプト
以下に大喜利のお題に対する回答を提示します。以下の制約に従い、面白さの点数を付けてください。 # 制約 1. お題に対する回答に対して、以下の形式で0から10の面白さの点数を付けてください。 点数: [0から10の間で回答に対する点数を付けてください] 2. 点数以外は何も記述しないでください。
(<画像>) お題: <お題> 回答: <回答>

図8 生成・評価のプロンプト例

A 実験設定の補足

A.1 プロンプト

回答生成と絶対値評価のプロンプトを図8に例示した。

A.2 人手評価用のサブセットの作成手順

人手評価用のサブセットは、サイトの1位の回答と最下位の回答の得票数差の分布を考慮してテキストのみのお題200件と画像付きお題50件の計250件を抽出した。この250件のお題それぞれに対し、5つの回答のうち1つのみを含み、かつ各回答タイプが50件ずつ含まれるように5つのサブセットを作成した。各サブセットに対して5人の評価者による評価を実施することで、サブセット内での評価者間の一致度と、サブセット間での回答タイプによる評価の差異を分析可能とした。

B 実験結果の補足

B.1 回答例

表1, 2, 3に回答を例示した。表2, 3は、LLMの大喜利回答の評価者平均が“サイトの1位の回答”を超えた事例である。

B.2 絶対値評価を基にランキングを作成した場合の一致度

絶対値評価を基にランキングを作成しQWKを算出した結果、人間とGPT-4o間が0.45、人間とClaude 3.5 Sonnet間が0.33、GPT-4oとClaude 3.5 Sonnet間が0.19となった。これはランキングにおいても人間との一致度はGPT-4oが優位であることを示すとともに、LLM間の低い一致度(0.19)は各LLMが異なるユーモアセンスを持つ可能性を示唆している。

表1 人間とLLMの評価が異なる回答の例。H=人間の平均, G=GPT, C=Claude.

お題	サイト1位	H	G	C
ゲーマーだけが住む島、どうなった?	ああああさんがたくさんいる	6.8	2	3
文字が無い世界について教えてください	京都の山を燃やすだけの行事がある	8.2	5	3
雷を食べきって一言	裁かれないよって人	1.4	7	8
闇鍋で起こったハプニング	カレーが負けた	2.8	8	8

表2 人間が高く評価したGPTの大喜利回答の例

お題	サイト1位	GPT大喜利
巨大迷路の全てのルートを通ってみてわかったこと	ズルをするお婆ちゃんもいる	迷路の出口よりも、途中で何度も現れるアイスクリーム屋が一番の人気スポットだった。
アホアホ科学館でやっている展示会	あったかい水「お湯」	風速100メートルで髪型がどうなるか選手権
凶暴なマラソン	輸血所がある	スタート直後からゴールテープが逃げる。

表3 人間が高く評価したClaudeの大喜利回答の例

お題	サイト1位	Claude大喜利
ネチネチハンバーガーショップ	関東では「ネッチ」、関西では「ネチバ」	ハンバーガー食べてると「そういえば、君の前の彼氏も同じような食べ方してたよね」って店員がささやいてくる
忍者試験に受からなかった人あるある	音立てて帰る	履歴書の特技欄に「姿を消す」と書いたら面接官に見つかってしまった
演劇のポスターに書かれていた思わず観たくなる一言	演技じゃない不仲あります	開演30分前に差し入れた1万円札は2万円になって戻ってきます