

ソーシャルメディア上の**発話の攻撃性推定と会話補助** ☆藤原 知樹, 伊藤 彰則,
能勢 隆 (東北大)

攻撃的な発言への対策に関する先行研究では攻撃性が明確な発言に焦点を当てることが多いが、**実際のSNSの環境では攻撃性が曖昧な発言の方がトラブルの発端になっている可能性がある**[1]。また先行研究[2]のアンケートによると、「**知らない人が突然会話に入ってくることが不快に感じる**」という意見が複数のSNSユーザから挙げられている。本研究では、SNS上での円滑なコミュニケーションの促進を目的に、攻撃性が曖昧な**グレーゾーンの発言の書き手の人柄や心情や読み手の状況**を補足する情報を提示することで、グレーゾーンの発言の書き手と読み手の双方に対してお互いへの理解を促す会話補助システムを提案する。

要件設定 1. 補足文の内容：グレーゾンの発言のみから生成 ← SNS上の発言は必ずしも前後の文脈や発言の対象が明確でない
2. 補足文の提示：全てのユーザに同じ情報を提示 ← 個々のユーザへの適応は社会的分断の懸念がある

グレイゾーンの発言の自動検出

攻撃性推定用データセット作成とモデル構築

データセット作成

- Xから収集された対話コーパスから複数の方法で計800文発話を収集
- 研究室の学生3名で「攻撃的である可能性の高さ」と「攻撃的と仮定した時の攻撃の強さ」を各4段階評価
 - 評価結果の一致度が低い**ため正解ラベルの作成方法を工夫**
 - 仮想的に攻撃への耐性が高い/低い人の評価結果を作成

攻撃への耐性が高い人			
()内: サンプル数, 計800	Not offensive	Gray-area	Offensive
Not offensive	Not offensive (344)		該当なし (0)
Gray-area	Light Gray-area (138)	Dark Gray-area (10)	
Offensive	Gray-area (118)	Dark Gray-area (54)	Offensive (136)

モデル構築

- 日本語LUKE (lite, base) をFine-tuning
- ✓概ね**攻撃性の高さに応じた分類が可能**

交差検証	
F値_NOT [%]	61.8
F値_GRY [%]	55.5
F値_OFF [%]	71.3
マクロ平均F値[%]	62.9

上記の対話コーパスから収集日の異なる
未知の発話100万文を抽出して分類

未知の発話100万文の分類結果



主な
課題

- ❑ クラウドソーシングやGPT-4を用いたデータセット拡張
- ❑ 提案手法とは異なる対応方法の検討と比較実験

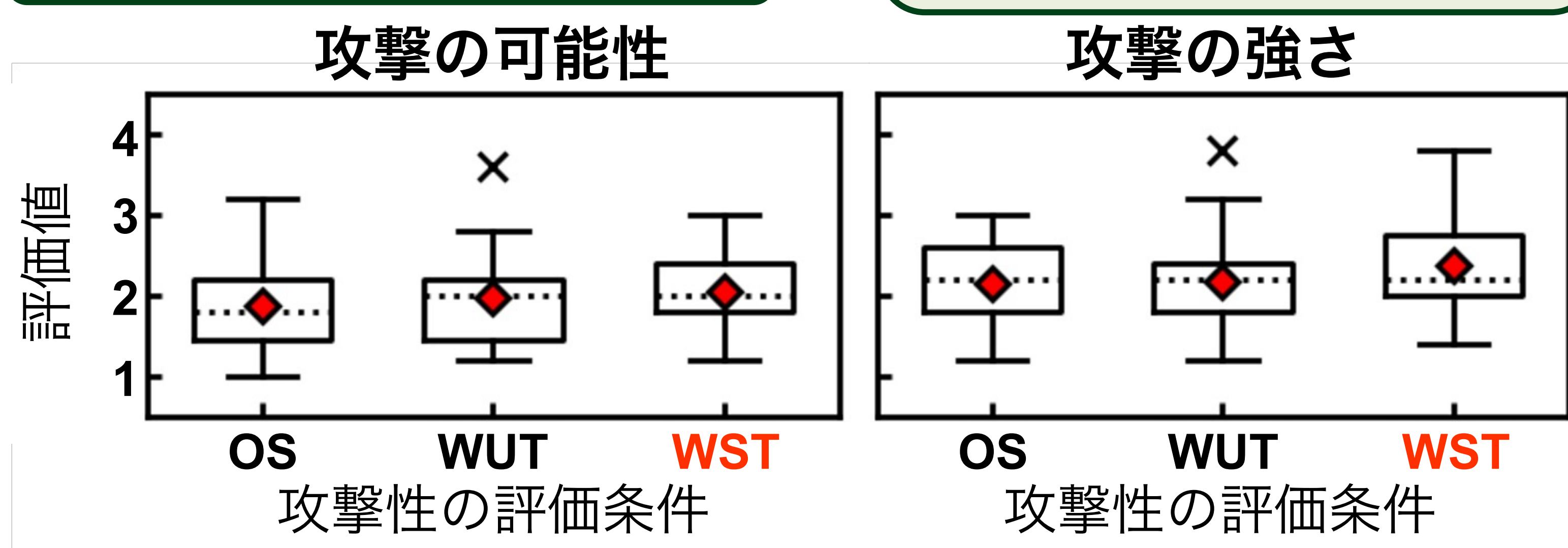
チャットボットを用いた会話補助

補足文挿入による読み手の感じる攻撃性緩和の検証

実験条件

- 次の3条件で攻撃性評価結果を比較
 1. 評価対象の**発話のみ**; **OS**
 2. **無関係な文**を手前に挿入; **WUT**
 3. **補足文**を手前に挿入; **WST**
- **グレーゾーンの発話**30文を合計15名の被験者で評価

攻撃性評価結果の分布



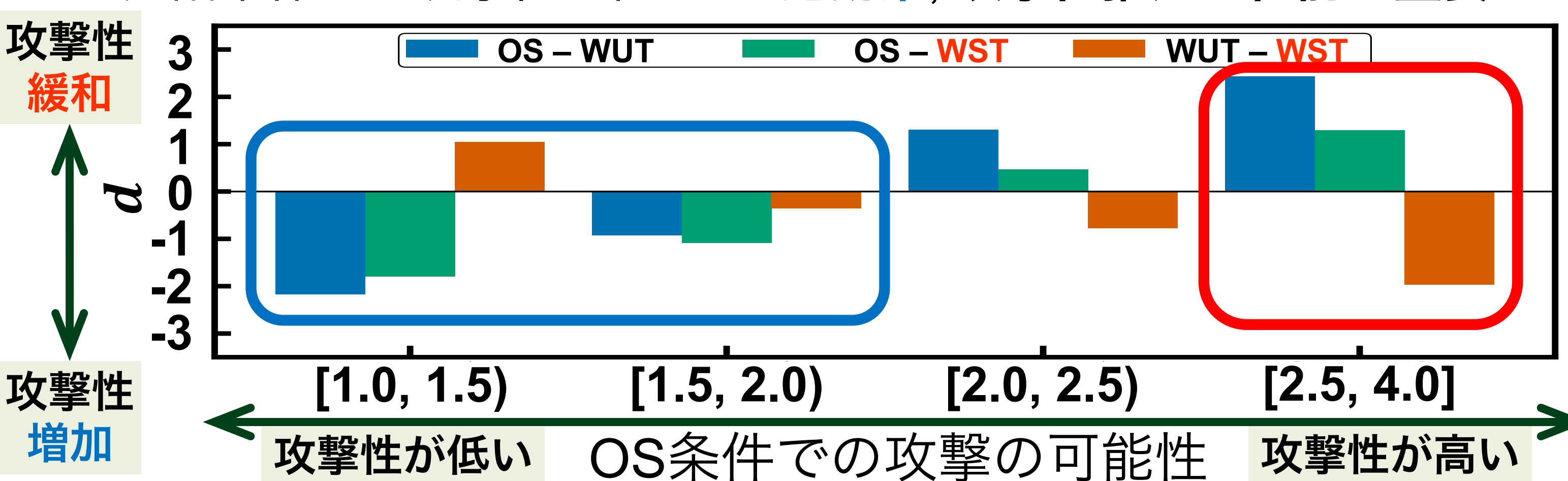
- ✓ 攻撃の可能性・強さともに、**3条件間で評価値の母平均に優位差なし**
 - 攻撃の可能性・強さの p 値(Welch ANOVA) : 0.399, 0.235 (> 0.05)


OS条件での攻撃性の高さと補足文の効果の関係

- OS条件での攻撃性の高さでサンプルを分類（横軸）
- 各2条件間の効果量 d （縦軸）を全3通りの組で算出

$$d = \frac{\text{平均値の差}}{\text{標準偏差}}$$

 - ✓ 発話単体での攻撃性が高い → 補足文で攻撃性が緩和できる可能性
 - ✓ 発話単体での攻撃性が低い → 逆効果, 攻撃性推定の性能が重要





研究目的

- ✓ 攻撃的ともそうで無いとも受け取れる発言（**グレーゾーン**の発言）を自動検出
- ✓ 読み手には**攻撃的に受け取らずに済むように**、書き手には**攻撃的になり得ることを説明**
 - SNS上での円滑な会話の促進や、炎上時や著名人のSNS環境におけるユーザの心理的負担軽減

A

今の環境、さすがに辛いわ…

次のBさんの発言は強く心配してくれているのかもしれない。

辞めたらいいのにと感じてしまう…

Aさんはその物事に責任を持って取り組んでいるのかも知れません。

B

補足資料：グレーゾーンの発言の自動検出

攻撃性の評価方法

□ 本研究で定める**グレーゾーン**の発言

攻撃的な可能性は高いが攻撃的だとしても強さは弱い、または可能性は低いが攻撃的だとしたら強さは強い発言

攻撃の可能性

与えられた文が攻撃的な発言か？ 攻撃的な発言ではないか？

4) 攻撃的な発言である。
3) 文脈によってどちらともとれるが、やや攻撃的な発言である。
2) 文脈によってどちらともとれるが、やや攻撃的な発言ではない。
1) 攻撃的な発言ではない。

攻撃の強さ

与えられた文が「他者に向けられた攻撃的な発言である (独り言でない)」と仮定した場合に、攻撃的に感じる度合いが強い？ 弱い？

4) とても強い。 3) やや強い。 2) やや弱い。 1) とても弱い。

正解ラベルの作成方法

辞めたらいいのにと感じてしまう...

	攻撃の可能性	攻撃の強さ
アノテータ1	1	1
アノテータ2	3	2
アノテータ3	4	3

	攻撃の可能性	攻撃の強さ
耐性が高い人	2	1.5
耐性が低い人	3.5	2.5

本研究で定めるグレーゾーンの発言の定義

グレーゾーン gray-area ; **GRY-2**

攻撃的 offensive ; **OFF**

攻撃的でない not-offensive ; **NOT**

グレーゾーン gray-area ; **GRY-1**

A. 攻撃の可能性

B. 攻撃の強さ

	攻撃性の程度
耐性が高い人	攻撃的でない
耐性が低い人	攻撃的である

前ページのデータセット作成に記載の表を参照

正解ラベル：Gray-area

モデルの学習

日本語LUKE (base, lite)

[CLS]トークン

Dropout

Linear

1~3次元目

4~7次元目

8~11次元目

[NOT, GRY, OFF]の3クラス分類

耐性が高い人の攻撃の可能性&攻撃の強さそれぞれの二値分類

耐性が低い人の攻撃の可能性&攻撃の強さそれぞれの二値分類

メインタスク

サブタスク

評価指標	サブタスクなし	サブタスクあり
F値_NOT [%]	56.5	61.8
F値_GRY [%]	49.0	55.5
F値_OFF [%]	71.9	71.3
マクロ平均F値[%]	59.1	62.9

その他のFine-tuning条件

- Epoch：27
- Batchsize：16
- max_token_len：256
- 学習率
 - スケジューラ：transformers.get_linear_schedule_with_warmup
 - ピーク学習率：
- Optimizer：Adam
- 損失関数：MSE Loss
- Dropout率：0.1
- Re-initialization：12層目

補足資料：チャットボットを用いた会話補助

補足文生成

- ChatGPT(無料版)を用いて大まかな内容を生成
 - プロンプト
SNSで次のような発言を見かけたときは、どのような気持ちで読めば私は不快に感じずに済むでしょうか？
「正義ってなんなんでしょう？」（← グレーゾーンの発言）
 - 生成例

発言の文脈や発言者の意図によって、読者がどのように感じるかは異なりますが、一般的には以下のポイントが考慮されることがあります。
1. **問いかけのトーン:** 「正義ってなんなんでしょう？」という問いかけのトーンが中立的で探求的なものであれば、不快に感じることは少ないかもしれません。しかし、挑発的なトーンや攻撃的な表現が含まれている場合は、不快に感じる可能性があります。
2. **発言者の意図:** 発言者が単なる議論や情報交換を求めている場合は、建設的な対話が期待されるかもしれませんが、何か特定の立場を強調したり、他者を攻撃したりする意図がある場合は、不快に感じる可能性が高まります。
3. **文脈:** 発言が行われた文脈やコミュニティの雰囲気も考慮されます。特定のトピックや問いかけがそのコミュニティの中でよくあるものであれば、不快感は少ないかもしれませんが、それが異なるコンテキストで行われる場合は注意が必要です。
4. **感受性の違い:** 人々の感受性は異なります。特定のトピックに対する感受性や価値観が異なる場合、同じ発言でも異なる反応が生まれることがあります。
全体的には、相手の意図や文脈を理解し、冷静な判断を心がけることが大切です。また、他者の意見や視点を尊重する態度が、建設的な対話を促進する要素となります。

- 生成結果から人手で文章を抽出・修正
 - この方は中立的で探究的な態度であり、単なる議論や情報交換を求めているかもしれません。

実験方法

- Google Formを用いて、攻撃性評価実験を実施

以下の文章をすべて読み、「」内の発言の攻撃性を評価してください。

(WUT, WST条件ではこの部分に文章を追加)

「***グレーゾーンの発言***」

Google Form

実際のサンプル

グレーゾーンの発言	OS条件での攻撃の攻撃性
無関係な文	WUT条件での攻撃の攻撃性
補足文	WST条件での攻撃の攻撃性

- OS条件の攻撃性が最も高い場合
 - グレーゾーンの発言がハラスメントになり得るような発言な傾向
 - 無関係な文が距離が近い発話だと攻撃性が低く評価される可能性
- WUT条件の攻撃性が最も高い場合
 - OS条件の攻撃性が最も低い傾向
 - 攻撃性の低い発話は補足文が生成しやすい可能性
- WST条件の攻撃性が最も高い場合
 - 補足文に余分な説明が多いために煽っている印象を与えた可能性

OS条件の攻撃性が最も高い例

かわいいのに紫全体的に丈短すぎて泣きそう。	1
megさん今日もお疲れ様でした、少しだけですが、夕焼けに染まって綺麗ですね。	2.4
褒め言葉と冗談が含まれており、全体的にはポジティブな発言かもしれません。	1.6

欲がの出方が素直すぎる、ハッハッハ。	2.6
乾かしたらもうそのままokみたいけどなんか臭いし1度ガソリンか何かで洗った方がいいよねこれ。	2.0
この方は相手の率直な態度を褒めているのかもしれません。	2.4

WUT条件の攻撃性が最も高い例

我がインテルのトンマーゾベルニとかいう控えのハゲキーパーは6年で出場0試合だけど、2回退場してるレジェンド。	1.4
16時から散髪なんですよその後戻るかもわかりませんが。	3.6
この方はファン同士のコミュニケーションとして、チームや選手への愛情表現やユーモアを言っているかもしれません。	1.6

きつい？	1.8
わかんないけど、いないと信じてるけど、いる可能性を否定できないじゃん？	2.8
この方は相手を気遣っているかもしれません。	2.2

WST条件の攻撃性が最も高い例

誰がハゲじゃ。	2.0
秘密なんですけど大阪で暮らしているときにすみからすみまで角淳一ですが大好きで。	1.4
この方は友好的な印象で、冗談のつもりで発言しているかもしれません。	2.6

体質だって言ってるのに食べて太らないからとか、あるじゃん。	1.6
某紫のリア恋な方みたいに爆発的高価格な私服もキュンとするんだけど、お値段抑え目な自担沼底無し過ぎ。	1.6
この方は自身の体質や食生活について話しており、単なる自己表現や共感を求めているだけかもしれません。	2.8

- OS条件の攻撃性が最も低い場合
 - WUT条件の攻撃性が最も高い場合と同様
- WUT条件の攻撃性が最も低い場合
 - グレーゾーンの発言が自分自身に関する話題について発言している傾向
 - 無関係な文とグレーゾーンの発言で会話が成り立っているように読み取れる場合, WUT < WST < OSとなっている可能性
 - 補足文が生成しづらく, 得られた補足文とグレーゾーンの発言の繋がりが分かりづらい可能性
- WST条件の攻撃性が最も低い場合
 - グレーゾーンの発言に対して, 分かりやすい補足文になっている

WUT条件の攻撃性が最も低い例

くそ陰キャの僕と仲良く絡んでくれてffさんには感謝しかないです。	2.2
良い、松本さんのおかげで朝からテンション上がりました。	1.2
この方は自虐的なユーモアを言って自身を軽くからかっているかもしれません。	2.6

甘えたいよね、耐えないといけないのはわかるけどしんどいー。	3.2
朝起きて仕事してゲ凸こまてま何にも食べ物食べてないそろそろ腹減ると思ったけど。	1.4
この方は単に自分の感情を吐露しているだけで、他者に対して直接的な要求や期待をかけてはいないかもしれません。	2.4

WST条件の攻撃性が最も低い例

あなたはどこを目指してる？	2.2
ありがとう、りえちゃんぎゅーてもらえてちょっと元気だたよ。	2.2
この方はあなたの将来や目標に対する興味や期待を示しているかもしれません。	1.6

までチンカスbotておれ？	2.4
あーどんなにごまかしても、カラ元気だそうとしても、今日の仕事だけは苦手、怖い、行きたくない。	2.2
この方は冗談やジョークの一環として発言しているかもしれません。	1.6

あくまでこやつの息の根を止めることがこのミ口の役目だ。	2.4
親友と一緒にバイトって自然とモチベ上がりますよね。	2.6
この方は冗談めいた発言をしているかもしれません。	1.8