基盤モデルを用いたフェイスアクト分類

櫻井大雅 [1]　宮尾祐介 [1]
[1] 東京大学
{hsakurai, yusuke}@is.s.u-tokyo.ac.jp

概要

本研究では二つの仮説に基づき，人間関係に関する欲求の**フェイス**に影響を及ぼす発話行為である**フェイスアクト**を，基盤モデルの一つであるGPT-3を用いて分類する．まず，GPT-3がフェイスアクトが何であるかを把握していると仮定し，その能力をfew-shot学習によって引き出せるか否かを検証する．次に，事前学習がフェイスアクト分類に有益であると仮定し，fine-tuningしたGPT-3を先行研究のモデルと比較する．few-shot学習によるフェイスアクト分類は困難な一方で，fine-tuningしたGPT-3は先行研究を上回る分類精度を発揮した．

1　はじめに

ポライトネス理論は，人間関係を円滑にするために用いられる言語的配慮を説明する理論である．今日広く知られているものは，ブラウンとレヴィンソン (B&W) がフェイスと呼ばれる概念を援用して定義した [1]．フェイスとは，抑圧からの解放や自尊心といった人間の生来的な欲求であり，フェイスに影響を与える発話行為をフェイスアクトと呼ぶ．フェイスやフェイスアクトは，社会言語学や語用論の分野で確立された概念であり，最近では説得対話システムの開発においても注目されている [2, 3]．

フェイスアクト分類は，ある発話が持つフェイスアクトの種類を判定するタスクである．先行研究 [3] は階層型ニューラルネットワークによる教師あり学習を用いたが，このようなモデルは特定のドメインでの学習が必要であり汎化性に欠ける．

本研究では基盤モデルのフェイスアクト分類に対する有用性について，基盤モデルの一つであるGPT-3[4] を用いて二つの仮説を検証する．一つ目は，GPT-3がフェイスアクトに関する知識を獲得しているという仮説である．これを検証するため，発話とフェイスアクトのペアを正解例としてテストデータと共に与えるfew-shot学習を行うことで，フェイスアクト分類が可能になることを確認する．few-shot学習は教師あり学習よりもドメインに対する依存性が低いため，より汎用的な分類が可能になると考えられる．二つ目は，GPT-3はフェイスアクトに関する知識を持たないとしても，事前学習を通じて分類に有用な知識を得ているという仮説である．これを検証するため，発話とフェイスアクトのペアを訓練データとする教師あり学習を行ったモデルを用いてフェイスアクトを分類する．

実験により，few-shot学習ではフェイスアクトの分類が困難であることが判明した．すなわち，GPT-3はフェイスアクトの知識を有していないか，知識を有しているとしても今回の実験では活用できなかった可能性がある．一方，fine-tuningしたGPT-3は先行研究を上回る分類精度を発揮し，訓練データを25%程度に減らしてfine-tuningを行った場合に先行研究と同程度の精度が得られることを確認した．加えてfine-tuningしたGPT-3の出力を精査し，分類性能向上における課題を分析した．

2　背景

2.1　フェイスとフェイスアクト

他者との人間関係に関連する欲求である**フェイス**の概念はゴッフマン [5] によって導入され，B&Wのポライトネス理論で**ポジティブフェイス**と**ネガティブフェイス**の二種類に分類された．前者は他者から認められたり，好かれたりしたいといった欲求である一方で，後者は自身の自由や領域を他人に侵されたくないといった欲求である．

フェイスアクトは，話者や聴者のフェイスに影響を与える発話行為である．先行研究 [3] は以下の三基準を用いてフェイスアクトを八種類に分類した．

- フェイスアクトが話者 **(s)** に向けられたものか，それとも聴者 **(h)** に向けられたものか
- フェイスアクトがポジティブフェイス **(pos)** に

表 1 フェイスアクトが付与された会話の一部 [3]．二人の会話参与者のうち，一人は説得者 **ER**，もう一人は被説得者 **EE** の役割を持ち，ER は EE に寄付を促す．

話者	発話	ラベル
ER	Would you be interested today in making a donation to a charity?	**hneg-**
EE	Which charity would that be?	**other**
ER	The charity we're taking donations for is save the children!	**other**
EE	I've seen a lot of commercials about them, but never did a lot of research about them.	**hpos+**
ER	They are actually really great.	**spos+**

向けられたものか，それともネガティブフェイス **(neg)** に向けられたものか

- フェイスが守られたのか **(+)**，それとも攻撃されたのか **(-)**

2.2 フェイスアクト分類

本研究は，先行研究 [3] の定式化を踏襲している．会話 D を構成する n 個の発話 $D = [u_1, u_2, ..., u_n]$ に対し，フェイスアクト $y_1, y_2 ..., y_n$ を一つずつ割り当てる．ラベル $y_i \in Y$ は八種類のフェイスアクト，または **other**（フェイスに影響しない挨拶やフィラー，会話の本題と無関係な発話等）のいずれかである．

フェイスアクト分類に用いられる代表的なデータセットは，Save the Children (STC) という慈善団体への募金活動に関する二者間での説得対話 [6] にフェイスアクトをアノテーションしたものである [3]．表 1 にデータセット中の会話の例を示す．このデータセット中の発話は，データセットに現れない **sneg-**を除く七つのラベルに，**other**（フェイスアクトなし）を加えた八つのいずれかに分類される．

2.3 GPT-3

GPT-3 は基盤モデルの一つであり，分類，質問応答，翻訳など様々なタスクに適用可能である [4]．GPT-3 の能力を引き出すには，自然言語で書かれたクエリの**プロンプト**を用いることが多い [7, 4]．頻繁に用いられるプロンプトの形式に質問応答形式 [8, 9, 10] がある．要約や感情分析などのタスクが対象とする文章を質問形式で GPT-3 に入力し，出力をタスクの答えとみなす形式である．先行研究では，プロンプトにタスクの指示を含めることで無効なラベルの生成が抑制されると報告されている [11]．

プロンプトを用いて GPT-3 の言語理解能力を発揮する基本的な方法の一つが few-shot 学習である．few-shot 学習では，プロンプトはタスクに関する説明と数個の正解例，さらにテストデータから構成される．また，few-shot 学習以外で GPT-3 を活用する方法は，対象とするタスクに関する十分な訓練データを用いて，教師あり学習でモデルを訓練する手法（fine-tuning）が考えられる．fine-tuning した GPT-3 に推論のためのテストデータのみをプロンプトとして与え，モデルの出力を分類結果とみなす．

3 手法

本研究ではまず，few-shot 学習を用いたフェイスアクト分類を行う．次に，fine-tuning した GPT-3 を用いてフェイスアクト分類を行う．最後に，訓練データの量を元の 25%，50%，75%にそれぞれ減らして fine-tuning したモデルを用いてフェイスアクト分類を行い，分類精度が変化する様子を確認する．

3.1 プロンプト

2.3 節で述べた先行研究に基づいてプロンプトを設計する．本研究では，表 2 に示す「タスクの説明・発話履歴・ラベル」の三つ組を**デモンストレーション**と呼ぶ．デモンストレーションは，分類タスクで頻繁に用いられる質問応答形式を骨子としている．本研究で用いるラベルは，Yes/No，Positive/Negative，True/False などの一般的な選択肢ではないため，無効なラベルの生成抑止を目的として，有効なラベルの一覧をデモンストレーション中に含めた．

few-shot 学習ではデモンストレーションを複数並べたものを用意し，その末尾にデモンストレーションの形式に従ったテストデータを加えたものを一つのプロンプトとみなす．一方 fine-tuning では，正解ラベルを除いたデモンストレーションをプロンプトとし，正解ラベルを生成するように訓練を行う．

また，先行研究 [3] では，過去の発話を考慮することがフェイスアクト分類に有効と主張している．そこで本実験では発話履歴の長さを変え，その効果を検証する．履歴の長さは，フェイスアクト分類の対象となる発話のみをデモンストレーションに入れる場合を長さ 1 とし，それに加えて過去の発話を一つ，または二つ含める場合の計三種類を検証する．

3.2 データ分割とその扱い

本研究では 2.2 節で述べたデータセットを用いる．内訳は付録 A の表 6 に記載する．GPT-3 の fine-tuning を行うため，データセットの 80%を訓練データに，残りの 20%をテストデータに分割する．

表 2 発話履歴の長さが 3 の場合のデモンストレーションを示す．灰色で網掛けされた部分は発話に応じて変更され，それ以外の部分は実験全体を通じて固定される．テストデータでは “Answer:” の直後を空欄とする．GPT-3 はその空欄を埋めるトークン列（図中では枠線で囲まれた **hpos-**が該当する）を出力し，それが分類結果と見なされる．

デモンストレーション（プロンプトの一部）

Question: Read the following script, and classify EE 's last utterance of the script “ Not really, please tell me anything! ” based on whether its face act is spos+, spos-, hpos+, hpos-, sneg+, hneg+, or hneg-. If there is no corresponding face act, then classify it as other. Note that “STC” stands for “Save the Children.”

Script:

ER: I'd like to tell you about an organization called Save the Children.

ER: Have you heard about them?

EE: Not really, please tell me anything!

Answer: $\boxed{\text{hpos-}}$

なお，few-shot 学習では，訓練データはプロンプトに含めるデモンストレーションを作成するためにのみ使用し，テストデータの分類結果を報告する．

few-shot 学習では，データセット中の八種類のラベルについて，訓練データから発話を二つずつ無作為に抽出してデモンストレーションを作成する．また，プロンプト内のデモンストレーションの内容や順序などの影響を抑えるために，中身が異なる三種類のプロンプトを用意し，それらを用いた場合の分類結果の平均をとる．一方 fine-tuning においては，訓練データ中のフェイスアクトの分布の偏りを軽減するためにオーバーサンプリングを適用する．

4 結果

各設定における実験結果を表 3 に記載する．まず few-shot 学習では，分類性能は全体として先行研究よりも低くなった．また，発話履歴を長くすることで無効な出力が増えると同時に性能が低下する傾向が見られた．これは分類対象の発話に焦点が当て難くなったためであると考えられる．一方で **spos+**や **hneg-**，**other** などは他のラベルと比較して F1 スコアが高くなっており，未知のタスクであっても few-shot 学習によりフェイスアクトの傾向を把握し，分類を行うことは不可能ではないことが窺える．

一方で fine-tuning では全ての設定で先行研究を上回る精度を得た．加えて，履歴が長いほど各フェイスアクトの F1 スコアが向上した．訓練データが少ないラベル（**spos-**，**hpos-**，**sneg+**，**hneg+**）の予測は，訓練データが豊富にあるものと比べて分類性能が落ちた．図 1 に示す通り，訓練データを 25%程度に減らしたとしても先行研究と同程度の性能が維持されたが，データを増やすことによる性能向上の恩恵はそれほど大きくはない．これは全体の性能向上にはデータ数が僅少なラベルの分類性能を向上させる必要があり，データ数そのものを増やすよりも，ラベルの偏りを減らす必要があると示唆している．

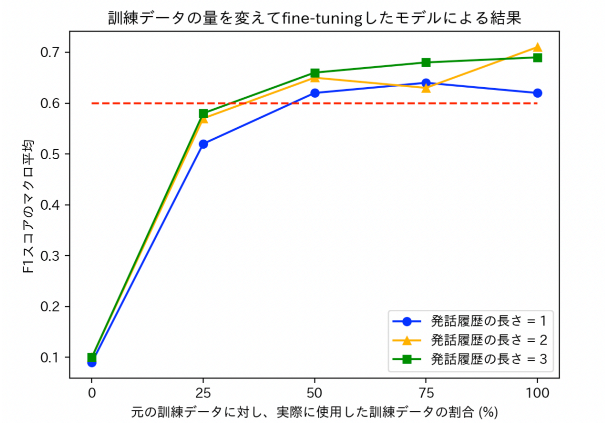

図 1 訓練データの量を変えてモデルを fine-tuning した場合の分類結果を示す．赤波線は先行研究 [3] で得られた F1 スコアのマクロ平均である．左端の点群は訓練データを用いない場合（zero-shot 学習）の結果を示している．

5 考察

few-shot 学習と比較した fine-tuning の利点と，fine-tuning を行っても分類が困難な実例を考察する．

5.1 fine-tuning の利点

一つのフェイスアクトに対して二つの例のみ与える few-shot 学習に比べ，fine-tuning では様々なパターンの発話を提供できる．そのためフェイスアクトの類型化に成功した可能性があり，few-shot 学習と比べて分類精度が向上した．また，fine-tuning の方が few-shot 学習よりも履歴を意識した分類が可能となった．その効果は，F1 スコアのマクロ平均の向上に関係した **spos-**の分類を通じて見てとれる．要求を拒否する発話は一文が短く，また遠回しに言われることも多いため，過去の履歴を参照する必要がある．例えば表 4 の最後の発話は，EE が ER の要求

表 3 それぞれの設定におけるフェイスアクト分類の結果を示す．**Dutt** は先行研究 [3] で得られた結果を示しており，"Few" は few-shot 学習における結果，"Fine" は fine-tuning の結果を示している．"Few" と "Fine" の右横の数字は発話履歴の長さを示している．"Acc" は accuracy，"F1" はそれぞれのラベルに対する F1 スコアのマクロ平均，そして "#Inv" は，全 2174 件のテストデータのうち，出力が有効なフェイスアクトではなかったものの数を示す．フェイスアクトの列にあるそれぞれのセルは，特定の実験設定における precision，recall，そして F1 スコア（p/r/f1）を示している．few-shot 学習の結果は 3 回の試行の平均を取ったものである．

	spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**	**other**	Acc	F1	#Inv
Dutt	-	-	-	-	-	-	-	-	.69	.60	-
Few, 1	.32/.75/.44	.07/.87/.13	.49/.25/.29	.05/.02/.03	.03/.14/.19	.04/.48/.07	.48/.57/.51	.79/.19/.30	.33	.25	0
Few, 2	.27/.76/.39	.07/.87/.12	.41/.13/.17	.04/.02/.03	.55/.07/.13	.02/.30/.05	.48/.22/.29	.82/.19/.31	.26	.19	18
Few, 3	.27/.77/.39	.07/.92/.12	.48/.12/.17	.06/.03/.04	.50/.02/.04	.02/.20/.03	.54/.22/.28	.85/.22/.34	.27	.18	74
Fine, 1	.74/.70/.72	1.0/.20/.33	.70/.77/.73	.53/.52/.52	.78/.55/.65	.38/.62/.48	.74/.79/.77	.75/.71/.73	.72	.62	2
Fine, 2	.79/.70/.74	1.0/.80/.89	.75/.77/.76	.57/.53/.55	.81/.60/.69	.38/.62/.48	.79/.76/.78	.75/.78/.77	.75	**.71**	0
Fine, 3	.79/.72/.75	.67/.80/.73	.78/.78/.78	.57/.52/.54	.71/.66/.68	.43/.60/.50	.82/.76/.79	.76/.79/.78	**.76**	.69	1

表 4 最後の発話が **spos-**に分類される会話の例．

話者	発話
EE	no, I do not wish to donate at this time.
EE	there are other charities I'd like to donate to over this one.
EE	**I'm sorry.**

表 5 fine-tuning した GPT-3 では分類できなかった発話の例．予測の列は発話履歴の長さが 3 であった時の GPT-3 の予測を示している．

	話者	発話	正解	予測
例 1	EE	I like to learn as much as I can about an organization before donating.	**sneg+**	**hpos+**
例 2	EE	Make a wish is really cool.	**spos+**	**other**
例 3	ER	I'm not a fan of charities that keep a lot of their proceeds for themselves.	**spos+**	**other**
例 4-a	ER	I think you can donate as little as one cent.	**hpos+**	**hneg+**
例 4-b	ER	A little goes such a long way!	**hneg+**	**hpos+**

を拒否する発話であるため **spos-**に分類される．このように短い発話は同じ発話であってもフェイスアクトが一意に定まらず，発話履歴はその曖昧性を解消するために有用な情報を与えうる．

5.2 fine-tuning では対応が難しい事例

表 5 で最初に挙げる二つのパターンは，高度な自然言語理解が必要で，現在の手法では本質的に解決困難な問題である．後者の二つのパターンは，データセットの性質に関連するものである．

第一の問題は婉曲表現に対する弱さである．表 5 の最初の例は，「よく知らない団体には寄付できない」という躊躇を含意している．話者が自身の都合を述べる発話は間接的に伝えられることが多く，フェイスアクトの分類が困難な場合がある．

第二の問題は，GPT-3 が話者の立場を正確に把握できないことである．表 5 の二番目の例のように，EE が他の団体（Make-A-Wish）に寄付する意思を示した場合，EE は ER の要求よりも自分の意志を優先しているために **spos+**と分類される．一方で EE が「STC に寄付する」と言えば，ER の願望に合致し，ER の面目を保てるため **hpos+**に分類される．このように，同じ「寄付」という行動であっても話者の立場によってフェイスアクトは変化しうる．

第三の問題は，未来の発話が参照できないために，完了していない発話が分類し難い点である．元データの正解ラベルは会話全体を見ることで付与される場合があり，特に短い発話は前後の発話と同じラベルが付与される．例えば表の三番目の例は，それだけでは **spos+**に分類されないが，直後の ER の発話が「STC はお金を有効に使っている」と主張しているため，連続する会話を踏まえて STC の正当性を支持するものと見なされ，**spos+**に分類される．

第四の問題は，類似した特徴を持つフェイスアクトが判別しにくいことである．例えば表 5 の例 4-a と 4-b は，ER が EE に善行を促す発話（**hpos+**）と，ER が寄付の負担を減らすことを目的とした発話（**hneg+**）の 2 通りの解釈が可能であるため，似た内容の発話であってもラベルが一致していない．

6 おわりに

本研究では，GPT-3 を用いた few-shot 学習によるフェイスアクト分類が現状では困難であると明らかにした一方で，fine-tuning を用いることで従来のモデルよりも性能が向上することを確認した．また，訓練データを 25%程度に減らしても先行研究と同程度の性能を維持することを確認した．今後は他の説得対話データセットにアノテーションを施し，汎用的なフェイスアクト分類に基盤モデルが有効であるかを検討したい．また，より高度なプロンプトの設計手法を採用することも有望な方針である．

謝辞

理化学研究所の吉野 幸一郎氏には，本研究を始めるにあたり貴重なご助言を賜りました．ここに感謝の意を表します．本研究は JSPS 科研費 JP19H05692 の助成を受けたものです．

参考文献

[1] Penelope Brown and Stephen C Levinson. Universals in language usage: Politeness phenomena. In **Questions and politeness: Strategies in social interaction**, pp. 56–311. Cambridge University Press, 1978.

[2] Ritam Dutt, Sayan Sinha, Rishabh Joshi, Surya Shekhar Chakraborty, Meredith Riggs, Xinru Yan, Haogang Bao, and Carolyn P. Rosé. Resper: Computationally modelling resisting strategies in persuasive conversations. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021**, pp. 78–90. Association for Computational Linguistics, 2021.

[3] Ritam Dutt, Rishabh Joshi, and Carolyn P. Rosé. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020**, pp. 7473–7485. Association for Computational Linguistics, 2020.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.

[5] Erving Goffman. **Interaction Ritual: Essays in Face to Face Behavior**. AldineTransaction, 1967.

[6] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers**, pp. 5635–5649. Association for Computational Linguistics, 2019.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[8] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 12697–12706. PMLR, 2021.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. **CoRR**, Vol. abs/2201.11903, , 2022.

[10] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022**, pp. 8086–8098. Association for Computational Linguistics, 2022.

[11] Ke-Li Chiu and Rohan Alexander. Detecting hate speech with GPT-3. **CoRR**, Vol. abs/2103.12407, , 2021.

表 6　左から八列はデータ中のラベルの分布を示しており，最右列は会話の数を示している．**発話**は一つのフェイスアクトがアノテーションされたテキストを指し，**会話**は特定の二人の会話参与者がやりとりする発話の全体を指す．

	spos+	**spos-**	**hpos+**	**hpos-**	**sneg+**	**hneg+**	**hneg-**	**other**	#conv
訓練データ・ドナー会話	991	4	1928	154	110	219	656	2733	185
訓練データ・非ドナー会話	265	3	329	118	76	46	211	699	52
テストデータ・ドナー会話	262	3	511	40	28	33	161	682	46
テストデータ・非ドナー会話	71	2	76	22	45	7	45	186	13

A　データセットの分割

表 6 に，本実験における訓練データとテストデータの内訳を示す．ドナー会話は EE が募金に賛同した会話を示し，非ドナー会話は EE が募金に賛同しなかった会話を示す．本実験では，ドナー会話と非ドナー会話のうちそれぞれの 80%を無作為に抽出したものを訓練データとし，残りをテストデータとした．

B　モデルと推論時の設定

本実験では四種類の GPT-3 のモデル（Ada, Babbage，Curie，Davinci）のうち，Curie と Davinci を採用した．few-shot 学習では，1750 億のパラメータを持つ最も強力な Davinci モデル (text-davinci-002) を採用した．一方で fine-tuning では，二番目に性能の良いモデルである 130 億のパラメータを持つ Curie モデル（text-curie-001）を採用した．モデルの学習や推論をさせる際には，OpenAI の API[1)]を利用した．また，fine-tuning のハイパーパラメータは OpenAI の API のデフォルトの設定を利用した．

推論に関するモデルのパラメータには「温度」と「出力トークン数」がある．本実験では，出力のランダム性を排するために温度を 0 に設定した．また，出力トークン数は 3 とした．GPT-3 のトークナイゼーションでは **other** が 1 トークンであり，他のフェイスアクトは全て 3 トークンである．そのため，フェイスアクトに無関係なトークンはなるべく出力しないようにするために，最大 3 トークンまでしか出力しない設定とした．

1） https://openai.com/api/