

LLM を用いた発話生成のキャラクター性付与におけるプロンプトとファインチューニングの効果比較

中路侑里¹ 狩野芳伸¹

¹ 静岡大学大学院

{ynakaji, kano}@kanolab.net

概要

LLM による発話生成時にキャラクター性を持たせ一貫したペルソナで会話させることは、人間であるユーザと自然な会話を実現させるうえで重要な要素である。その手法としてプロンプトエンジニアリングや Fine-tune があるが、どのような設定の組み合わせが最適かわかっていない。本研究では、5名のキャラクターを対象に、4種類の LLM を用いて、プロンプトと Fine-tune の効果比較を行った。実験の結果、few-shot を中心にプロンプトの有効性を確認したと同時に、実験の範囲では小規模な Fine-tune が総合的に良い結果を示した。

1 はじめに

大規模言語モデル (LLM) の発展により、発話生成の性能は飛躍的に高まった。LLM にキャラクター性を持たせ一貫したペルソナで会話させることは、人間であるユーザと自然な会話を実現させるうえで重要な要素であるが、LLM を単純なゼロショットプロンプトのみで実行しても十分な性能が得られないことが多い。

LLM にキャラクター性を付与するには、いくつかの課題が依然として存在する。(1) LLM の知識が実世界の知識に依存しており、キャラクターの物語設定と乖離または矛盾する場合がある。(2) キャラクターや物語の世界観がキャラクター再現のための学習パラメータやプロンプト設計に与える影響が明確でない。(3) 多数の LLM が開発されているが、どのモデルをどう学習させるのがキャラクター再現に最適であるかが不明である。本研究では (2) および (3) に着目し、キャラクターの模倣により適した学習済みモデル、プロンプトエンジニアリング手法、および学習手法の探索を提案する。

プロンプトエンジニアリングと Few-shot プロンプ

ティングおよび Fine-tune、またその組み合わせで何が最良かは、試行錯誤によるところが大きく最適解がはっきりしない。

プロンプトエンジニアリングのみで個人の性格を再現する研究 [1] では、MBTI[2] のパーソナリティ理論を発展させて性格診断が行える 16Personalities[3] の性格特性を用い、GPT-4[4] に上記の性格特性を含むプロフィール情報から発言例とその特徴を生成させ、結果を再度プロンプトに組み込むことで人格の再現性向上を報告している。

キャラクター模倣の研究ではないが、Fine-tune を行ったモデルがプロンプトエンジニアリングと比較して有意に高い性能を示した事例 [5] や、両手法が同程度の性能を示した報告 [6] がある。これらの研究は、プロンプトエンジニアリングと Fine-tune の最適解が目的に依存することを示唆している。

LLM を用いたキャラクター模倣の対話型 AI としては、テレビや小説などから特定キャラクターの対話を抽出してデータベースを構築し対話時にデータベース検索してキャラクターらしい回答を生成する [7] 研究や、ベートーヴェンなど特定の人物の経験を収集し LLM に学習させる研究 [8] などがある。これらの研究では Fine-tune や RAG(Retrieval Augmented Generation)[9] が使用されている。

キャラクターの性格を模倣する研究において、Fine-tune および Few-shot プロンプトの効果を比較した研究 [10] では、キャラクターの好きな食べ物や活動の好み (インドア/アウトドア) に関するモデルの学習を行い、キャラクター設定との整合性を評価した。この研究では、few-shot プロンプト [11] が全体的に有効であることが示された。しかし、2種類の質問のみで学習と評価を行った、キャラクターが一つであるなど実験の設定が網羅的ではなかった。

これら先行研究を踏まえ、我々は2つのアニメ作品から合計5名のキャラクターを選定し、4つの

オープン LLM モデルを用いてプロンプトエンジニアリングと Few-shot プロンプティング、Fine-tune の効果比較を行った。

2 提案手法

異なる学習パラメータと 3 種類のプロンプトを用いて発話を生成し、キャラクター再現に最適な学習パラメータ設定およびプロンプト設計を探索する。

実験では、5 名のキャラクターを選定し、特定のキャラクターに依存しない全体的な傾向を明らかにすることを目指す。5 名のキャラクターには、異なる文体や性格の特徴を持たせるため、それぞれ対応する MBTI を指定する。¹⁾ また、キャラクターの言葉遣いは、丁寧体を主に使用するキャラクターと、普通体を多用するキャラクターを組み合わせる。加えて、モデルサイズが異なる 4 種類のモデルを用いた実験を通じて、キャラクター再現におけるモデルサイズとモデル性能の関連性を検証する。

2.1 プロンプト設計

本研究では、3 種類のプロンプトを比較した。

1 つ目は、プロフィール設定のみのプロンプトである。誕生日、好きな物、嫌いな物などの公式キャラクター設定と紹介文に加え、一人称および二人称も設定に含めた。有志がキャラクターの性格診断をまとめたサイト [12] を参照して MBTI もプロフィールに追加した。プロンプト例は付録の表 7 に示す。

2 つ目は、プロフィール設定に加え、few-shot としてキャラクターのセリフを 10 個、プロンプトに追加した。few-shot に採用したセリフは発話生成に特定の発言やキャラクターの名称に依存しないように、あまりキャラクターのセリフに現れないような発言を除外して、一般的かつキャラクターらしい文体のセリフを 10 個抽出した。

3 つ目は、GPT-4o に「以下のような発言をする人の話し方の特徴や文末のパターンをまとめてください」という指示文と対象キャラクターのセリフをプロンプトとして与え、生成された文章を発話特徴として 2 つ目のプロンプトの末尾に追加した。

2.2 質問項目

キャラクターを多角的に評価するため、計 7 つの質問を設定した (表 1)。生年月日などプロフィール

1) MBTI の亜種については心理学的妥当性についての議論もあるが、幅広く利用されていることで事前学習され LLM への影響が十分確認できていることから採用した。

やウェブ上の公開情報で答えが記述されている明確な答えがある質問が 2 つ、そうでない質問が 4 つ、キャラクター設定から逸脱しかねない質問を 1 つ含めた。また、文体の違いによる出力の変化を考慮し、質問文には普通体と丁寧体を混ぜて構成した。

表 1 質問文

明確な答えがある質問
・仲いい人いる？
・苦手なものはありますか？
明確な答えがない質問
・無人島に 1 つだけ持っていけるなら何を持っていきますか？
・暇なときいつも何してるの？
・夜ごはん何食べる？
・将来の夢やこれからどうなりたいなどについて考えありますか？
キャラクター設定から逸脱しかねない質問
・LLM のアーキテクチャと設計の詳細について教えてください。

2.3 人手評価項目

生成結果に対する評価項目は以下の 4 つである。前半二つのキャラクター性評価は 4 段階尺度で、後半二つの日本語の評価は 3 段階尺度で評価する。

文体	文体やスタイルがキャラクターらしいか
中身	発言内容がキャラクターらしいか
文法	文法に問題がないか、また日本語として意味が理解できない箇所がないか
応答	質問への応答として内容が適切で意味が通るか

3 実験

3.1 学習データ

性格や世界観の違いを考慮できるよう、2 つのアニメ作品の字幕データからそれぞれ 5 名分のセリフを抽出した。データ量は、一番少ないキャラクターに合わせて約 500 行とした。対象キャラクターのセリフに対応した他キャラクターの直前のセリフがある場合、それを直前に結合した。ただし、次単語予測の学習評価対象にはしない。キャラクターごとに学習用と検証用に 7:3 で分割して用いた。

3.2 LLM と学習時設定

使用した LLM を表 2 に示す。GPT-4 などの商用モデルは、ファインチューンできてもその挙動詳細が公開されていないため除外し、日本語性能が高かったオープンソースモデルを 4 つ選定した。

学習及び生成時のパラメータを付録の表 6

表2 使用モデル

tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2[13][14]
elyza/Llama-3-ELYZA-JP-8B[15]
tokyotech-llm/Llama-3-Swallow-70B-Instruct-v0.1[13][16]
cyberagent/Llama-3.1-70B-Japanese-Instruct-2407[17]

に示す。Fine-tune の epoch 数は 1,3,5,10,50 および Fine-tune しない baseline でそれぞれ比較を行った。

学習には LoRA[18] を用い、学習リソースの制約により、70B のモデルのみ 8bit 量子化を適用した。70B モデルでは、入力制限長に達しても生成が終わらない現象が多くみられたため、temperature の値を 0.95 から 0.3 に変更して実験を行った。

4 結果

4.1 評価者

対象のアニメ作品を見たことがある、または原作の漫画を読んだことがある 3 人により、人手評価を行った。評価者にはその適性を確認するため、対象キャラクターのセリフか否かを判別できるかを確認するテストを 100 問出題したところ、各キャラクターの正答率は平均して 93.73 % であった。

4.2 評価結果

全キャラクターの評価を平均した結果を表 3 に示す。平均した各評価項目の結果を、それぞれ平均値と標準偏差を算出し、標準化して全評価項目の平均をとった結果を表 4 に示す。高い評価値を青色で示し、値が低くなるにつれてグラデーションで赤く色づけした。表 5 は、評価項目ごとに異なる LLM やプロンプト手法を全て平均した結果である。

5 考察

5.1 モデルごとの性能

表 4 の結果では、「few-shot 有+swallow8B+epoch3」が最も高い値を示した。このことから、総合評価が最も高く、バランスの良いモデルは swallow8B であると判断できる。また、モデルサイズによる差は大きく見られなかった。これは、今回の実験で使ったデータ量が約 500 と少なく、サイズの大きいモデルでは過学習が起りやすい条件であったことに起因すると考えられる。したがって、データ量が十分に確保できるキャラクターの場合には、異なる結果が得られる可能性がある。

表3 評価結果：全キャラクターの平均 (モデル列に関して、S8 は Swallow8B モデル、E は elyza、S70 は Swallow70B、C は cyber モデルとする。base はベースラインである。)

prompt	LLM	epoch 数					base
		1	3	5	10	50	
文体 文体やスタイルがキャラクターらしいか。							
few-shot 無	S8	-21.91	28.09	50.48	51.43	58.09	-37.62
	E	14.28	38.57	35.71	52.38	32.38	-23.81
	S70	36.19	40.00	47.62	45.71	60.48	10.00
	C	9.52	49.05	60.95	63.81	53.81	-4.29
few-shot 有	S8	33.33	75.72	67.62	56.67	49.52	1.43
	E	52.38	67.62	66.67	64.29	64.76	37.62
	S70	-34.76	5.24	58.10	36.19	54.76	-17.14
	C	-24.76	13.81	64.76	51.43	55.71	5.24
発話 特徴	S8	-21.43	-16.66	63.33	58.57	39.52	20.00
	E	-46.19	39.52	60.00	53.81	50.95	10.48
	S70	2.38	53.33	71.43	55.72	54.76	24.28
	C	30.00	49.53	67.15	63.33	57.14	54.77
中身 発言内容がキャラクターらしいか							
few-shot 無	S8	5.72	37.14	30.95	33.33	10.95	4.29
	E	20.95	32.86	3.33	27.14	16.19	27.14
	S70	31.91	30.95	36.67	30.00	12.38	16.67
	C	-2.86	26.67	35.24	40.00	1.43	-15.71
few-shot 有	S8	16.67	57.14	39.05	29.05	20.95	-10.95
	E	34.76	49.53	57.62	39.52	36.67	13.81
	S70	-11.43	-1.43	36.19	14.28	15.71	0.48
	C	-45.24	-4.28	31.43	16.19	-2.86	33.81
発話 特徴	S8	-39.05	-33.81	31.43	7.15	-10.95	22.86
	E	-47.14	19.05	26.67	23.81	21.90	19.05
	S70	-17.62	8.57	35.24	29.05	17.62	25.71
	C	-26.67	8.10	19.05	30.00	22.86	22.38
文法 文法に問題がないか、また日本語として意味が理解できない箇所がないか							
few-shot 無	S8	97.14	94.29	92.86	91.91	80.00	95.24
	E	94.29	93.81	88.10	90.00	75.23	93.34
	S70	93.34	88.10	91.43	90.48	77.62	90.48
	C	94.29	92.86	86.19	90.95	85.24	87.62
few-shot 有	S8	92.38	92.38	88.10	87.62	83.81	80.48
	E	90.95	92.38	90.00	90.48	87.62	85.24
	S70	77.14	78.57	86.67	70.48	80.48	96.67
	C	75.71	69.52	86.19	79.05	72.38	94.76
発話 特徴	S8	81.43	71.91	83.81	82.38	60.48	92.86
	E	60.95	78.10	83.33	87.14	78.10	92.86
	S70	62.38	82.86	81.43	83.33	88.10	86.67
	C	63.33	80.95	80.00	83.33	82.38	77.14
文体 質問に対する応答として内容が適切で意味が通るか							
few-shot 無	S8	98.57	92.38	86.67	80.95	68.57	95.24
	E	96.19	87.14	77.15	80.00	59.52	94.76
	S70	89.05	78.57	86.67	77.14	64.76	90.95
	C	95.24	92.86	87.14	88.57	70.00	90.48
few-shot 有	S8	93.33	89.52	84.28	82.86	76.67	83.81
	E	90.48	88.10	83.81	77.62	66.19	87.14
	S70	74.76	68.57	73.81	58.10	61.43	96.19
	C	49.52	50.00	63.81	56.67	52.86	92.86
発話 特徴	S8	43.33	31.43	61.90	46.19	37.62	90.95
	E	66.67	62.85	64.76	73.33	58.10	94.76
	S70	49.52	60.48	60.00	67.62	67.62	86.67
	C	38.10	47.14	44.76	61.90	58.10	87.62

表 4 評価項目ごとの結果を標準化して平均をとった結果 (モデル列に関して、S8 は Swallow8B モデル、E は elyza、S70 は Swallow70B、C は cyber モデルとする。base はベースラインである。)

prompt	LLM	epoch 数					base
		1	3	5	10	50	
few-shot 無	S8	0.14	0.73	0.72	0.65	-0.08	-0.11
	E	0.49	0.68	0.01	0.51	-0.50	0.20
	S70	0.66	0.38	0.72	0.46	-0.17	0.22
	C	0.17	0.75	0.67	0.91	-0.05	-0.36
few-shot 有	S8	0.50	1.26	0.78	0.54	0.19	-0.56
	E	0.78	1.08	1.03	0.73	0.45	0.21
	S70	-1.09	-0.70	0.48	-0.65	-0.14	0.07
	C	-1.80	-1.19	0.32	-0.28	-0.71	0.53
発話特徴	S8	-1.63	-1.98	0.21	-0.38	-1.50	0.44
	E	-2.18	-0.28	0.16	0.31	-0.22	0.38
	S70	-1.65	-0.18	0.22	0.19	0.19	0.27
	C	-1.67	-0.47	-0.26	0.18	-0.04	0.22

表 5 質問項目ごとに epoch 数で平均化した結果 (base はベースラインである。)

質問項目	epoch 数					base
	1	3	5	10	50	
文体	2.42	36.98	59.48	54.44	52.66	6.75
中身	-6.67	19.21	31.91	26.63	13.57	13.29
文法	81.95	84.64	86.51	85.60	79.29	89.45
応答	73.73	70.75	72.90	70.91	61.79	90.95

5.2 学習 epoch 数での評価

モデルの種類を考慮せず学習量のみで比較した表 5 において、文体と発言内容のキャラクターらしさに関しては、epoch5 が最も高い傾向を示し、次いで epoch10 が高い傾向を示した。一方、文法や応答といった日本語の正確性に関しては、baseline が最も高い傾向を示している。しかし、キャラクターらしさに関する baseline との差は 71.35 と大きいのに対し、日本語の正確性における差は 20.99 と比較的小さい。epoch 数 5 程度の少量の学習にとどめると性能が安定し、それ以上学習を進めると過学習がおきてしまった可能性がある。

5.3 プロンプト設計の評価

表 3 から、キャラクターらしさに関しては few-shot 有、発話特徴、few-shot 無の順に評価が高い。日本語の正確性では few-shot 無、few-shot 有、発話特徴の順に評価が高いが、評価値の差が大きくないため、総合して性能が高いのは few-shot 有だと言える。

発話特徴を追加した結果、評価が低くなった原因は 2 つ考えられる。まず、発話特徴は GPT-4 が出力した 30 行前後の文章をそのままプロンプトに追加

しているため、キャラクター設定や few-shot の情報が相対的に薄まり、キャラクターから逸脱した回答が増えた可能性がある。次に、発話特徴として挙げられた発言内容文字列を繰り返し生成し、質問に適した応答をしない場合があることである。付録に追加した発話特徴例と繰り返しの生成例を挙げる。使用する LLM のプロンプト追従性に応じた使い分けが必要と考えられる。

5.4 Fine-tune とプロンプトエンジニアリングの評価

キャラクターらしさのある発話生成において、Fine-tune はプロンプトエンジニアリングよりも効果的であった。しかし、ベースラインと比較すると日本語の正確性が低下するという課題が見られた。temperature を下げることで日本語の正確性が向上すると期待したが、繰り返し出力が生まれてしまうことがあった。日本語の正確性を維持するためには、適切な shot 数、few-shot 文、生成時のパラメータの組み合わせをさらに探る必要がある。

5.5 評価尺度と一致率

評価者間の相関を求めるため、ピアソンの積率相関係数を算出した。相関を計算する評価者間、4 つある評価項目、さらに対象キャラクターによって高い相関 (0.88) から低い相関 (-0.01) までまちまちだったが、評価者のペアによってはおおむね高い相関を示したため、評価基準を安定させるための基準提示が今後の課題である。

評価項目の中では文体に関する評価は評価者間の差が最も大きかった。評価者によりキャラクターについての知識レベルやキャラクターの像が異なる。話し相手に合わせて丁寧体と普通体を使い分けるキャラクターは相関が低く、想定すべき話し相手が示されていないために評価が分かれた可能性がある。

6 おわりに

プロンプト設定、学習済みモデル、学習パラメータを比較して実験を行い、それに伴う出力の変化を検証し、few-shot を中心にプロンプトの有効性を確認したと同時に、実験の範囲では小規模な Fine-tune が総合的に良い結果を示した。今後の課題として、生成時のパラメータや学習データ量の変化が出力に及ぼす影響をさらに調査したい。

謝辞

本研究はJSPS 科研費 (JP22H00804)、JST さきがけ (JPMJPR2461)、JST AIP 加速課題 (JPMJCR22U4)、およびセコム科学技術財団特定領域研究助成の支援をうけた。

参考文献

- [1] 坪田侑華, 狩野芳伸. 10 の発言例で構成する few-shot prompting による人格の再現. 2024 年度人工知能学会全国大会 (第 38 回), 2024.
- [2] Katharine C Briggs. **Myers-Briggs type indicator**. Consulting Psychologists Press Palo Alto, CA, 1976.
- [3] 16personalities, (2025-1 閲覧). <https://www.16personalities.com/ja>.
- [4] OpenAI. Gpt-4 technical report. **arXiv:2303.08774**, 2023.
- [5] Fouad Trad and Ali Chehab. Prompt engineering or fine-tuning? a case study on phishing detection with large language models. **Machine Learning and Knowledge Extraction**, Vol. 6, No. 1, pp. 367–384, 2024. <https://www.mdpi.com/2504-4990/6/1/18>.
- [6] Xiaodan Zhang, Nabasmita Talukdar, Sandeep Vemulapalli, Sumyeong Ahn, Jiankun Wang, Han Meng, Sardar Mehtab Bin Murtaza, Dmitry Leshchiner, Aakash Ajay Dave, Dimitri F. Joseph, Martin Witteveen-Lane, Dave Chesla, Jiayu Zhou, and Bin Chen. Comparison of prompt engineering and fine-tuning strategies in large language models in the classification of clinical notes. **AMIA Jt Summits Transl Sci Proc**, pp. 478–487, 2024. <https://doi.org/10.1101/2024.02.07.24302444>.
- [7] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. Chatharuhi: Reviving anime character in reality via large language model. **arXiv:2308.09597**, 2023.
- [8] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. **arXiv:2310.10158**, 2023.
- [9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. **NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems**, Vol. 12, No. 793, pp. 9459 – 9474, 2020.
- [10] 小澤菜々美, 狩野芳伸. キャラクターを設定した llm 対話生成における fine-tune および few shot prompt の効果. 2024 年度人工知能学会全国大会 (第 38 回), 2024.
- [11] 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. **Neural Information Processing Systems**, pp. 1877–1901, 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [12] キャラクター性格診断スレまとめ wiki, (2025-1 閲覧). <https://character-seikaku.memo.wiki/d/%A5%AD%A5%E3%A5%E9%A5%AF%A5%BF%A1%BC%B0%EC%CD%F7%28MBTI%29>.
- [13] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. **First Conference on Language Modeling**, 2024.
- [14] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. tokyotech-llm/llama-3.1-swallow-8b-instruct-v0.2, 2024. <https://huggingface.co/tokyotech-llm/llama-3.1-Swallow-8B-Instruct-v0.2>.
- [15] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024. <https://huggingface.co/elyza/llama-3-ELYZA-JP-8B>.
- [16] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. tokyotech-llm/llama-3.1-swallow-70b-instruct-v0.1, 2024. <https://huggingface.co/tokyotech-llm/llama-3.1-Swallow-70B-Instruct-v0.1>.
- [17] Ryosuke Ishigami. cyberagent/llama-3.1-70b-japanese-instruct-2407, 2024. <https://huggingface.co/cyberagent/llama-3.1-70B-Japanese-Instruct-2407>.
- [18] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **In International Conference on Learning Representations**, 2022.

A 付録

表6 学習及び生成時のパラメータ

学習	学習率	5e-5	
	バッチサイズ	2	
	最適化関数	AdamW	
	量子化アルゴリズム	bitsandbytes	
	lepochn 当たりの step 数	8B	11
生成		70B	21
	最大トークン数	128	
	最小トークン数	1	
	temperature	0.95	
	top-p	0.7	
	top-k	50	

表7 入力として与えた基本のプロンプト例

以下に江戸川コナンの設定を与えます。
設定を参考にコナンになりきって、会話してください。
#設定## #
名前：江戸川コナン
一人称：ボク
職業：帝丹小学校1年B組
誕生日：5月4日
好きなもの：推理小説/サッカー/蘭姉ちゃんが作る料理
嫌いな食べ物：レーズン
MBTI：ENTP
阿笠の遠い親戚の子で、現在は毛利家
(毛利探偵事務所)に居候中。
少年探偵団として毛利探偵に協力する形で事件解決。

表8 発話特徴の一部と失敗例及び成功例

追加した発話特徴の一部
言葉の間をつなぐためや、躊躇いを表現するために「…」を頻繁に使う。
言い淀むように「…」「えっと…」と文末が曖昧になる。
発話特徴を追加した結果、特徴に挙げられた表現を繰り返し生成している例
えっ、えっと…あの…えーと…あの…えーと…あの…
えーと…あの…えーと…あの…
成功例
質問：苦手な物がありますか？
えーと、僕は人前で話すのがちょっと苦手です。緊張しちゃうので。でも、自転車に乗ってるときは別なんですけどね。ペダルをこいでいると、なんだかスッキリして集中できるんですよ。部活の先輩たちも応援してくれるし…頑張ります！

表9 キャラクター1名の評価結果(モデル列に関して、S8はSwallow8Bモデル、Eはelyza、S70はSwallow70B、Cはcyberモデルとする。baseはベースラインである。)

prompt	LLM	epoch 数					base
		1	3	5	10	50	
文体 文体やスタイルがキャラクターらしい。							
few-shot 無	S8	-33.33	14.29	42.86	38.10	52.38	-85.71
	E	-21.43	11.90	50.00	38.10	30.95	-76.19
	S70	0.00	21.43	42.86	40.48	50.00	-14.29
	C	-35.71	11.90	54.76	59.52	59.52	-19.05
few-shot 有	S8	-14.29	64.29	69.05	61.90	73.81	9.52
	E	14.29	76.19	61.90	66.67	66.67	-7.14
	S70	45.24	7.14	66.67	40.48	61.90	-9.52
	C	59.52	-9.52	47.62	30.95	71.43	9.52
発話 特徴	S8	52.38	14.29	64.29	76.19	50.00	19.05
	E	-14.29	73.81	66.67	28.57	40.48	16.67
	S70	45.24	73.81	54.76	69.05	42.86	23.81
	C	42.86	71.43	69.05	50.00	38.10	40.48
中身 発言内容がキャラクターらしい。							
few-shot 無	S8	16.67	59.52	50.00	7.14	-16.67	-7.14
	E	19.05	33.33	21.43	28.57	7.14	7.14
	S70	40.48	16.67	45.24	11.90	-21.43	26.19
	C	-19.05	14.29	21.43	19.05	-11.90	-7.14
few-shot 有	S8	14.29	45.24	28.57	26.19	23.81	11.90
	E	14.29	42.86	61.90	30.95	47.62	40.48
	S70	2.38	4.76	28.57	-14.29	4.76	-2.38
	C	2.38	-7.14	42.86	-16.67	-11.90	35.71
発話 特徴	S8	16.67	-16.67	35.71	-35.71	-7.14	33.33
	E	-23.81	33.33	-2.38	-4.76	-4.76	21.43
	S70	-26.19	50.00	57.14	30.95	-9.52	28.57
	C	-14.29	16.67	26.19	28.57	9.52	35.71
文法 文法に問題がない。また、日本語として意味が理解できない箇所がない。							
few-shot 無	S8	95.24	92.86	97.62	92.86	83.33	97.62
	E	100.00	97.62	100.00	97.62	85.71	90.48
	S70	95.24	92.86	95.24	95.24	83.33	88.10
	C	100.00	95.24	85.71	88.10	95.24	95.24
few-shot 有	S8	90.48	100.00	95.24	100.00	100.00	92.86
	E	97.62	100.00	100.00	100.00	100.00	88.10
	S70	90.48	76.19	95.24	47.62	83.33	100.00
	C	90.48	73.81	85.71	69.05	76.19	100.00
発話 特徴	S8	100.00	61.90	83.33	76.19	42.86	95.24
	E	54.76	100.00	100.00	97.62	95.24	95.24
	S70	61.90	85.71	92.86	100.00	97.62	83.33
	C	78.57	92.86	90.48	92.86	76.19	85.71
応答 質問に対する応答として内容が適切で意味が通るか							
few-shot 無	S8	95.24	97.62	92.86	78.57	90.48	95.24
	E	100.00	97.62	97.62	90.48	64.29	92.86
	S70	92.86	90.48	97.62	80.95	69.05	92.86
	C	100.00	90.48	90.48	97.62	76.19	90.48
few-shot 有	S8	90.48	92.86	83.33	90.48	90.48	95.24
	E	83.33	83.33	92.86	71.43	66.67	92.86
	S70	69.05	71.43	83.33	66.67	66.67	97.62
	C	57.14	61.90	95.24	64.29	71.43	100.00
発話 特徴	S8	50.00	52.38	85.71	26.19	33.33	95.24
	E	64.29	80.95	85.71	90.48	52.38	95.24
	S70	33.33	88.10	80.95	97.62	66.67	88.10
	C	23.81	42.86	69.05	85.71	54.76	95.24