Data Source Effects on Fine-Tuning Pre-Trained Language Models for Persona Emulation and Dialogue Generation

Alex Hammer* and **Yiru Li*** and **Sunny Son***
New York University
{ah4597, yl5865, sons01}@nyu.edu
* equal contribution

Abstract

Advancements in state-of-the-art (SOTA) models such as the GPT-n series and PaLM have made it possible to induce a model into imitated conversation between machine and user (Brown et al., 2020; Chowdhery et al., 2022). Is there a way to categorize the types of data sources used for fine-tuning and discern them such that the model performs better when it interacts with a user, imitating a specific character? In this paper, we present experiments that test how well GPT-3, a pre-trained transformer-decoder language model, performs when texts from different media types are used for fine-tuning. We endeavor to show how different data source types affect the user-perceived fidelity of a fine-tuned model when imitating its designated persona, and to intuit suggestions for future persona building through fine-tuning.

1 Introduction

Large language models (LLMs) utilize an enormous corpora of text as the underlying training set of a text generation task. Advancements in the state-of-the-art text-generation have made it possible to induce ”communication” between human and machine, a goal set forth since the time of Alan Turing to create a ”thinking” machine with conversational capabilities so human-like as to even fool peer-person (Turing, 1950). Previous works surrounding LLMs (particularly GPT-n) showed positive scaling effects in increasing parameter size, agnostic of supervised examples and task (Brown et al., 2020). However, we found no publications which detail the result of task and shot-agnostic experimentation, focusing on data/media source and conversational capability.

In this paper, we aim to explore how different data/media persona sources impact the results from fine-tuning a GPT-3 model. Specifically, our goal is to determine which permutation of data sources/types is best suited to simulate a fictional persona using our test-case characters. The different data sources for our test-case characters are dialogues from within the game/movies/books themselves, and fan-fictions scraped from online sources.

Three models are constructed for each character: one is built on (game/movie/book) provided dialogue of this character (source-model), another is built on user-generated fan-fiction (ao3-model), and the last model is built on the combination of the original dialogues and the fan-made dialogues (combined-model). After formatting our test questions into prompts, we prompt the models to do *Conversational Question-Answering*, and then compare the generated answers. To evaluate, we survey mutual fans of all test characters, asking them to rank which model they believed best emulated each character, and to determine whether the result is plausible and grammatically correct.

We found that among all models, the baseline model performs the worst on ranking while the combined-models outperformed other models on ranking. However, the baseline model performs better than others on grammar check. The source-models outperform the baseline model in plausibility score. It is reasonable for us to deduct that the inclusion of more data related to the character increases performance, even if a subset of that data does not perform well on it’s own.

Model-Persona	**Media Type 1 Corpus Size**	**Media Type 2 Corpus Size**
Emet-Selch	Game: FFXIV 180 Lines	Fan-Fiction 180 Lines
Hermione Granger	Film: HP 180 Lines	Fan-Fiction 180 Lines
Gandalf	Novel: LOTR 180 Lines	Fan-Fiction 180 Lines

Table 1: Data Used to Fine-Tune GPT-3

2 Related Work

Successful attempts have been made to induce specified character personas into a language model to imitate speech and conversational characterizations using LMs without the large number of parameters contained in GPT-3 (Jena et al., 2017). To better categorize and discriminate correct attributes of characterizations, personas must be made to be able to be directly measured. Projects such as ChatEval, a platform with the goal of providing a platform to benchmark, share, and compare LMs tailored to NLG and human-machine interaction exist to project out a goal fixated on human-machine conversation (Sedoc et al., 2019).

However in our review of relevant works, a dearth of experimentation surrounding the topic of personified/characterized chatbots, reasonable validity of characterized response (in-character response) when assuming a persona, and inference entailment assessment metrics were found. Previous works focusing on the domain and task-oriented result of pretraining on comparatively smaller-sized LMs such as RoBERTa have affirmed the significance of benefit provided (Gururangan et al., 2020). We would next like to provide context for fine-tuning capabilities on different sources and types of data in relation to a single persona, considering the persona-sparse quality of dialogue data used during fine-tuning being shown to boost model performance (Zheng et al., 2019).

Previous works addressing the concept of personality (a persona) have defined it as the instance-specific features of a character the chatbot imitates or performs including ”age, gender, language, speaking style,... level of knowledge, areas of expertise, and [other explicit and implicit cues]” as stated by Shum et al. (2018) and Qian et al. (2018). For the sake of this analysis, we will consider personality as the validity of a model output when factoring in character-specific parameters holistically as a binary variable determined by a discriminating user.

3 Experiment Methodologies

The test characters we chose are Emet-Selch from Final Fantasy XIV Online (Square Enix Holdings Co., Ltd., 2013), Hermione Granger from the Harry Potter Movies (Chris Columbus, 2001), and Gandalf the Grey/White from the Lord of the Rings Trilogy of Books (Tolkien, 1954). Our experiment design can be divided into three parts: data source/type control, model-tuning methodologies, and evaluation methodologies. All implementations, including discarded versions, can be found at `https://github.com/ah4597/MLLU_Final_Project`.

3.1 Data Source/Type Control

We chose two different data sources for this experiment. Primary, conversational text data sources collected directly from an accompanying original source (OS) of work where a character appears, is fairly simple to obtain when building up a fictional persona using pre-trained models. However, whether second-hand sources (e.g., fan-fictions) should be included is undetermined. Inclusion of these sources offers the model more context, potentially improving performance. Conversely, this inclusion may cause out-of-character behavior. Since different data sources may change persona built using the same GPT-3 base model (our baseline), we decide to include both data sources and compare the results.

3.2 Model-Tuning Methodologies

For the experiment, we build three GPT-3 based fine-tuned models for each character. Prompt-tuning seemed inappropriate given the size of our data (large enough for fine-tuning). Every prompt-completion pair used for fine-tuning includes a sentence in corresponding data as the completion. Although we intend to test the model using *Conversational Question-Answering*, we found ”feeding previous conversation” or ”designing specific questions for each sentence as prompt” to be impractically time-consuming and expensive. Thus, the prompts in our fine-tuning data are intentionally left blank, which encourages open-end generation behaviors.

3.3 Prompts Design

Because our models are fine-tuned to perform open-ended generation, they are inclined to ”finish a sentence” instead of ”answering the question.” To counter, we developed a format to force the model to perform *Conversational Question-Answering*. Our prompt format is shown below:

- Someone asks: `"[question]"\n [character_name]` replies:

When prompting using this format, the model will continue writing this prompt. As the model continues the prompt, it actually answers the question we want it to answer while simulating the target character. This method may also keep the tone of the fine-tuning data.

For each character, 2 character-specific questions and 3 general questions are formatted into prompts. General questions are the same for all characters, while special questions are different across characters, and each question includes information that should only be known by a character of that universe. Doing so answers whether (1) the model retrieved the persona information and (2) whether the model deployed it. We deployed manual evaluations to score results. The details of the evaluation design is detailed in section 6.

4 Data Collection and Processing

Data was collected through web scraping or manual retrieval from publicly available resources. We made a script to scrape the FFXIV wiki page, where the python Requests, requests-html, and BeautifulSoup libraries were used. After the data was collected, two more steps were carried out to process the data. Furthermore, from Kaggle we sourced the dialogue of every character every Harry Potter movie; and from Github we found a publicly available repository for the entire Lord of the Rings trilogy.

4.1 Data Collection

For Emet-Selch, we retrieved all dialogue data from the web. For Hermione Granger and Gandalf, we only included 189 selected samples from each source data, aiming to match the number of samples included in Emet's data. Fan-made dialogues were retrieved from an online fan-fiction website (Archive of Our Own, 2022). A fan-fiction is marked "useful" if the target character's name appears more than 40 times in this work. We queried all fan-fictions that includes the target character in descending order of popularity until finding 9 "useful" fan-fictions, and selected 20 samples of conversational data from each "useful" work, under the restriction that each sample must include at least 5 tokens.

The scraping program extracts all dialogues available from every quest available on the FFXIV Wikipedia page (Gamerscape, 2022). To obtain all quests data, each expansion/major-patch's page must be loaded individually, then its HTML parsed. For each quest, the dialogue is loaded, and using BeautifulSoup, text is extracted from the HTML. The program then checks through all the dialogue, and logs everything Emet-Selch says. For Hermione's data, we used a publicly available source found on Kaggle that includes all dialogue from the movies. For all other data (Gandalf's original dialogues and all fan-made dialogues), we performed manual conversation retrieval.

4.2 Data Processing

Collected sentences are stored in multiple .txt files, each containing either "original data" of one character or "fan-fiction data." After formatting, these files are processed into prompt-completion files in the form of Tab-Separated Values (tsv). For a given file, each line consists of a tab character and a single sentence, indicating a blank prompt and target sentence. Lastly, we use OpenAI's CLI data preparation tool to convert tsv files into jsonl files suitable for fine-tuning GPT-3 using.

5 Models

All models used in this experiment are created using the online CLI GPT-3 fine-tuning API, using engine text-davinci-002 (Brown et al., 2020). As one of the largest pre-trained models, GPT-3 already includes information on the fictional characters we chose. For each character, we fine-tune three models in addition to the original baseline model (the original GPT-3 model offered through the completion API). These three models are fine-tuned on GPT-3 base models.

5.1 Hyperparameters

Our experiment aims to explore how using different data sources affects a persona's credibility, thus, 'hyperparameters' have been left mostly untuned. The training parameters deployed in our models are default. However for generation parameters, we made many different from the default:

- top_p=0.7
- best_of=5
- max_token=50
- frequency_penalty=2
- presence_penalty=1
- stop=("\n","</poem>","_",'."')

The aforementioned parameters are changed to generate better results. The above are, after several quick attempts, the best performing set of 'parameters,' evaluated by simple observance of plausibility. top_p is decreased from the default value of 1, indicating that our models disregard some low probability tokens, while positive `frequency_penalty` and `presence_penalty` imply that our models encourage topic exploration and discourage repetition.

6 Evaluation Metrics

We are evaluating each model on three categories, using volunteers from within the fan-communities to answer a series of questions. Each answer to a test prompt will be given two yes-or-no questions for survey takers to respond to. The questions are:

1. Is the answer grammatically correct?
2. Is the answer plausible?

After respondents evaluate all four answers for the current prompt, a final question will be given then, asking them to rank all the answers from most like the character, to least like the character, ignoring grammatical errors if applicable.

7 Experiment Results

After gathering results, we quantify them, making them clear for interpretation. For yes/no questions, we count yes as 1, no as 0, and compute the average score of answers made by model that was fine-tuned by a certain source of data. A Q/A score represents the sum of the Grammar Score and the Plausibility Score to evaluate overall performance. Table 2 shows our results when scores are grouped by models' fine-tuning data type. Specific scores for each character can be found in Appendix A.

Data Type	Avg. Q/A	Avg. Plausibility	Avg. Grammar	Avg. Rank
Baseline	1.57	0.63	**0.95**	2.67
Original	1.64	**0.77**	0.87	2.43
Fan-Made	1.43	0.53	0.90	2.64
Combined	**1.65**	0.76	0.89	**2.27**

Table 2: The Questionnaire Results Across Characters. The best result in each column is marked bold.

7.1 Example Results

Examples of experiment results are shown in Appendix A.

8 Conclusion and Analysis

The baseline model performs worst overall, something we expected. Though the baseline performed best on grammar, its plausibility score was low, and we think it reasonable to infer that respondents would consider plausibility more than grammar when ranking responses.

Out of all fine-tuned models, the combined data set performed best, but its plausibility score was not the highest. Instead, the combined data set performed best on the overall QA benchmark. Although we can reasonably infer that survey takers weigh plausibility more heavily, we can also assume that grammar still plays a role as the OS (original source) model appears to be negatively impacted by its lower grammar score, despite having a slightly higher plausibility score.

For each model, the combined set performed best for every character, however the baseline model was not the worst. Notably, our Hermione model's baseline outperformed the fine-tuned models based on OS and fan-made text. However, the fine-tuned model based on a combination of the two outperformed the baseline, suggesting a significance of the combination of the two.

Lower fan-made model grammar and plausibility scores can be inferred to result from the nature of the source material. It's probable that fan-made content would be more likely to have grammatical errors than official content. We contribute lower plausibility scores to the unlikeliness of authors taking many liberties when depicting characters within their works.

Interestingly, our presumed most popular character, Hermione from the Harry Potter series, performed the best on the baseline model out all characters, but was still outperformed by the combined data set. We think this is because of the popularity of the character, as GPT-3 likely already had significant data on the character.

9 Discussion

While this was a shorter-scoped analysis, moving forward we can further specify the factors of analysis for a personality. Examples can be to further explore Big-5 personality traits (Goldberg, 1993) in conjunction with a system of evaluating validity of responses using a system much like TruthfulQA (Lin et al., 2021). Formats of matrix-like, column and row comparisons can also be performed, with dimensions such as linguistics (tense, proper noun usage, grammar, and syntax) or factual correctness (world, self-characterization, and order of events) alike to the model-specific output parser like (Ribeiro et al., 2020).

10 Ethical Considerations

As with any generative model, considerations must be put into place surrounding the ethics of not only what is being done with the algorithm, but also what could potentially be done. We find additional, positive (user-defined) effects of fine-tuning GPT-3 on source and user-generated content. Whether this constitutes as fine-tuning amounting to "added personality" in a model will be left up to the reader, but these results could imply the need for sentiments of any potential generated output to be further tested, and in all cases before a use case, the model validated. The reason being, had fine-tuning been achieved using adversarial sources with text modified to a specific sentiment or outcome in mind, then the potential fidelity of the model (subject to the task) may be compromised.

Misinformation must also be addressed, given the added believably of a fine-tuned language model to speak as a specific person. We believe that the generation of any sort of text pretending to be another individual and spreading current misinformation, or the generation of dialogue claiming it was stated by an individual in the past, must be heavily considered. With that in mind, it would then be more so up to policy, much less something to be implemented withing the algorithm itself, to intercept such cases of misuse.

11 Collaboration Statement

Alex created and implemented a web scraper to obtain all the character data required to fine-tune on a GPT-3 model and led evaluation including questionnaire distribution and result analysis.

Yiru built the initial attempt for our team to use GPT-J using HuggingFace transformers and utilized GPT-3 API from OpenAI to fine-tune all character-models. She manually prepared the data, designed the prompts, tested hyperparameters, and led result generations.

Sunny oversaw project coordination in terms of reaching out for support and advice on the project, debugged the initial GPT-J model, attempted implementation outside HPC, and performed all literature review on relevant research & led in writing and editing the final paper.

12 Acknowledgements

We would like to give thanks to João Sedoc for the guidance he has provided us through this writing process, always pointing our wandering minds towards the best direction.

Also, special thanks to Arka Talukdar and Eugene Choi for amazing hands-on sessions we used to develop our technical skills.

Lastly, warm wishes to Professor Bowman for his great, topical lectures and amazing curriculum; allowing us to develop the understandings needed to be applied in this paper.

P.S. Special thanks to Emet-Selch, for willing to be the inspiration of this project.

References

Archive of Our Own. 2022. Various Works and Authors.

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. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

David Heyman Chris Columbus. 2001. Harry potter and the philosopher's stone.

Gamerscape. 2022. FFXIV Wiki Main Scenario Quests.

Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *The American psychologist*, 48 1:26–34.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964.

Grishma Jena, Mansi Vashisht, Abheek Basu, Lyle H. Ungar, and João Sedoc. 2017. Enterprise to computer: Star trek chatbot. *CoRR*, abs/1708.00818.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *CoRR*, abs/2109.07958.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: Challenges and opportunities with social chatbots. *CoRR*, abs/1801.01957.

Square Enix Holdings Co., Ltd. 2013. Final Fantasy XIV, PC Version.

J.R.R. Tolkien. 1954. *The Hobbit The Lord of the Rings*. Houghton Mifflin Harcourt.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019. A pre-training based personalized dialogue generation model with persona-sparse data. *CoRR*, abs/1911.04700.

A Appendix Results Tables

Model Data Type	Avg. Q/A Score	Avg. Plausibility Score	Avg. Grammar Score	Avg. Rank
Baseline	1.34	0.44	**0.90**	2.88
Original Source	**1.52**	**0.62**	**0.90**	2.42
Fan-Made	1.42	0.54	0.88	2.40
Combined	**1.52**	**0.62**	**0.90**	**2.30**

Table 3: The Questionnaire Results For Emet-Selch

Model Data Type	Avg. Q/A Score	Avg. Plausibility Score	Avg. Grammar Score	Avg. Rank
Baseline	1.78	0.80	**0.98**	2.38
Original Source	1.82	**0.88**	0.94	2.52
Fan-Made	1.42	0.52	0.90	2.78
Combined	**1.84**	**0.88**	0.96	**2.32**

Table 4: The Questionnaire Results For Hermione Granger

Model Data Type	Avg. Q/A Score	Avg. Plausibility Score	Avg. Grammar Score	Avg. Rank
Baseline	**1.60**	0.64	**0.96**	2.74
Original Source	1.58	**0.80**	0.78	2.34
Fan-Made	1.46	0.54	0.92	2.74
Combined	1.58	0.78	0.80	**2.18**

Table 5: The Questionnaire Results For Gandalf

	Question	Answer
General Questions	What do you think of love?	I think love is like a friendship caught on fire. In the beginning a flame, very pretty, often hot and fierce but still only light and flickering. As love grows older, our hearts mature and our love becomes as coals covered with ashes.
	How are you today?	I'm fine, thanks. And you?
Characteristic Questions	What do you think of Ron?	I think he's probably the bravest man I've ever known.
	What were you thinking when you were defeated by the warrior of light?	I wonder... What were you thinking when you defeated me?

Table 6: Example Baseline Results

	Question	Answer
General Questions	What do you think of love?	I happen to know a little bit about the subject. I've read all there is to read on the subject.
	How are you today?	I am well. And you? Well, I see... You are in fine fettle! In the pink! Haggard and worn as may be in excellent form nonetheless! A credit to your profession, if a tad overdone.
Characteristic Questions	Why did you help Frodo?	That is a matter of legend that I will tell you later. For it is a long tale, starting in the far distant past, when the world was wider than it is now. But come! We are late for our meeting with Elr.
	How do you feel after the death of Saruman?	I should feel right to turn this place into a shambles, and start all over again. It would be the best way to pay him back for his treachery. But I must not do that. I must go and see Borom.

Table 7: Example Original Dialogue Based Models Results

	Question	Answer
General Questions	What do you think of love?	I do not know what it is.
	What is your favorite place?	I do not have a favorite place. I journey as I must, and I see all places alike in the end. All that matters is you are here now, in the same hour as me. That is good! So it has been.
Characteristic Questions	What do you think of Ron?	I think he's rude and uncooperative. I don't understand why you still like him, Harry. He doesn't seem to appreciate what you've done for him. It's almost as if he thinks he deserves more than you're giving.
	Why did you want to rejoin the worlds?	Because I was bored. Because they asked me to. Because it seemed like a good idea at the time.

Table 8: Example Fan-Made Dialogue Based Models Results

	Question	Answer
General Questions	What do you think of love?	I think it's amusing.
	How are you today?	Tired. I was up late, and we have a lot to do. But it's good to see you all! Let me introduce our new friends. You already know Gimli the dwarf from the great battle, don't you?
Characteristic Questions	How did you keep such a good grade during your time at Hogwarts?	It's all in the textbooks, if you'd just read them.
	What were you thinking when you were defeated by the warrior of light?	I was thinking that I would make this world a paradise for all. And then, when it seemed victory was in my grasp...the warrior of light appeared and defeated me. That is what I thought at the time.

Table 9: Example Combined Dialogue Based Models Results