Do we still need Human Assessors? Prompt based GPT-3 User Simulation in Conversational AI

ANONYMOUS AUTHOR(S)

Scarcity of user data continues to be a problem in research on conversational user interfaces and often hinders or slows down technical innovation. In the past, different ways of synthetically generating data, such as data augmentation techniques have been explored. With the rise of ever improving pre-trained language models, we ask if we can go beyond such methods by simply providing appropriate prompts to these general purpose models to generate data. We explore the feasibility and cost-benefit trade-offs of using non fine-tuned synthetic data to train classification algorithms for conversational agents. We compare this synthetically generated data with real user data and evaluate the performance of classifiers trained on different combinations of synthetic and real data. We come to the conclusion that, although classifiers trained on such synthetic data perform much better than random baselines, they do not compare to the performance of classifiers trained on even very small amounts of real user data, largely because such data is lacking much of the variability found in user generated data. Nevertheless, we show that in situations where very little data and resources are available, classifiers trained on such synthetically generated data might be preferable to the collection and annotation of naturalistic data.

CCS Concepts: • **Computing methodologies** → **Natural language generation**.

Additional Key Words and Phrases: datasets, nlp, text generation, conversational ai

ACM Reference Format:

Anonymous Author(s). 2018. Do we still need Human Assessors? Prompt based GPT-3 User Simulation in Conversational AI. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

A major problem in chatbot research, past and present, has been a scarcity of user data, especially for domain-specific Conversational Agents (CAs) focused on specific strategies or outcomes and for CAs that are constructed for non-English speakers [3, 9, 25]. In some of those domains, researchers have little or no data at their disposal and have to fall back on expensive and time consuming data collection approaches, such as user studies followed by manual annotation [10, 12], time consuming collection and annotation of existing texts [18], or use out-of-domain data that might not be perfectly suitable for the task they are trying to solve [11, 13]. Such data collection approaches require extensive manual labour, usually conducted by multiple people to acquire valid results. In all of these cases, both the data itself and the annotations have to be explicitly provided by humans.

There are a number of approaches that have been used to mitigate the costs of data collection [15] and data scarcity in the past, mainly focused on distant supervision, which uses weakly labeled examples gathered using noisy techniques, and data augmentation, which exploits different techniques to create new data from existing examples [6, 14, 21]. In recent years, Natural Language Generation (NLG) approaches have become ever more powerful and are by now capable of outputting very human sounding and seemingly coherent text [4]. This has led to impressive results in other NLP

Manuscript submitted to ACM

domains such as hate speech detection, where small datasets enriched with huge amounts of synthetic data significantly improved classification results [23]. GPT-2 has also been used to simulate users in the context of clarifying questions [22], leading to our idea that such data might be able to play an even larger role in data scarce CA-related domains.

With increasing capabilities of such models, it has even become possible to generate relevant and realistic texts for a task without fine-tuning, solely by supplying a natural language prompt –such as "translate this sentence into English" followed by a non-English sentence– to the language model [5, 19]. There is much research to be done on the potential of prompt-based generation compared to fine-tuning, but it has been shown that prompt-based approaches are a feasible means of data augmentation in fields with very little training data. GPT-3 has been praised for its few-shot and zero-shot learning capabilities [5]. Yoo et al. introduced a few-shot learning data augmentation technique where very few data samples are combined with prompts to generate realistic synthetic data. They consistently outperformed other data augmentation techniques in classification benchmarks [24]. Reynolds and McDonell explored prompt programming for zero-shot learning, showing how important carefully crafted prompts are to achieve relevance of the output and outlining various tweaks that can be applied to prompts to improve results [20]. Here, we ask the provocative question: To what extent can we leave the human out of corpus construction by exploiting such pre-trained language models?

The human-like wording of generated texts raises the question of how such data actually compares to data provided by actual human users or annotators. In the context of responses to clarifying questions, texts written by humans were perceived as more natural but not more useful than GPT-generated texts [22]. Other research found that even after being trained and seeing examples, human evaluators were not able to distinguish GPT-3 created stories, news articles and recipes from those written by humans [7]. Ironically, transformer-based approaches have been shown to be very good at recognising the difference between human-generated and synthetic text [16], hinting at a structural difference between the two that might not be intuitive to humans.

In this work, we use an existing human-labelled German-language dataset associated with a conversational domain where resources are limited, and construct a synthetic dataset using prompt-based GPT-3. We consciously choose prompts in such a way that the labels from the naturalistic dataset apply to the GPT-3 output. We then perform a number of classification experiments to explore the feasibility of replacing or enhancing naturalistic data with prompt-based, non fine-tuned synthetic data and compare the two datasets concerning their language variability and content. In doing this, we shed light on some of the countless questions still surrounding the usefulness of the black box that are (generative) transformer-based architectures.

2 METHODOLOGY

As a naturalistic data basis for our experiments we use an existing corpus from one of the authors' ongoing research projects [1]. The project explores ways of increasing motivation and success of health behaviour changes by administering Motivational Interviewing (MI) via a CA [2]. Motivational Interviewing focuses on making patients aware of their own reasons to change, and increasing their self-efficacy [17]. As such, a CA would mainly ask the user questions and exhibit reflection by rephrasing user utterances. This ELIZA-style setting allows for conversations in which domain knowledge plays a secondary role. The corpus encompasses data from a German weight loss forum and annotations regarding motivational factors of behaviour change. The annotations contain three labels: **Taking Steps (TS)** entails steps that have been taken towards or against change in the past, **Reason (R)** encompasses all reasons, basis, incentives, justifications or motives for or against making a change and **Commitment (C)** revolves around specific commitments

[1] https://anonymous.4open.science/r/GLoHBCD-58D1
[2] source withheld for anonymous reviews

Label	Question	Output
R	What habits would you like to implement in your everyday life?	If I ate less, that would be great. But I don't know how to start. I have no idea how to go about it.
	Why do you want to lose weight?	Well, I don't want to stand on the scales anymore and then get in a bad mood.
TS	What preparations have you made so far to change your eating behaviour?	I have read some books on healthy eating, read up on the internet and had a nutritional consultation with my GP. I also talked to a friend about it and she gave me some tips.
	Have you already taken steps to change certain behaviours in exercise and diet?	Yes, I have done a lot. I went jogging regularly for the first time in many years and did more sport. I also eat a lot more fruit and vegetables.
C	Do you have a plan for what you will do to manage your weight better?	Yes, I have a plan. I want to exercise more and eat less.
	What are your plans for the next few days to implement your great ideas on exercise and nutrition?	I plan to exercise more. I want to spend my time in nature. I will try some new sports. I want to bring more movement into my life. I want to get to know my body and soul better.

Table 1. Example questions posed to GPT-3 to generate the dataset including example outputs for each label, translated from German.

regarding the change for the near future. Standard classification approaches work reasonably well (e.g. BERT achieves an F1-score of 76.96% on an independent test set[3]), however there remains scope for improvement. This context is an appropriate case study since MI can serve as a framework for designing realistic questions to pose to GPT-3 as prompts. We expect, however, that the described process would transfer to any NLP or conversational topic, in which supplying facts to a user is not the predominant focus, such as mental health chatbots or other soft skill focused CAs.

2.1 Synthetic Data Generation

A focus group of information science students identified common topics in the annotated forum data and designed a number of suitable MI-questions for each topic-label combination. Since **R** has a number of sublabels in the original corpus and is thus responsible for 65% of the data, we constructed more questions for this label. For validation, the focus group controlled whether the wording of each questions cohered to MI guidelines and whether a native speaker would be able to correctly understand the question's intention. We collected 142 questions, of which 22 were focused on eliciting **C** and **TS** statements, respectively and 98 on eliciting **R** statements. In Table 1 we show two example questions and corresponding outputs for each label.

We then generated responses by posing the questions as prompts to GPT-3 DaVinci using the Completion Engine. As the original dataset revolves around behavioural change for weight loss, each question was embedded in the following contextualising prompt: *"AI: Hello! I'm here to support you on your journey to a healthier weight.* ***[Question]*** *Human:"*. We set output temperature to 0.5 and max_tokens to 100. In post-processing, generated output was ended after it indicated a change in speakers (i.e. markers such as "Robot:", "Human:") or after the last punctuation mark in the output to avoid half-sentences in the synthetic dataset. We generated 100 outputs for each question.

[3]source withheld for anonymous reviews

Training Set	Size	% R	% TS	% C
user	3,779	64.78	25.62	9.6
mixed	24,508	68.01	18.41	13.59
mixed predicted	22,128	71.09	18.86	10.05
synthetic	20,728	68.59	17.09	14.31

Table 2. Overview of training datasets

2.2 Experiments

The synthetic data was processed the same way as the original and split into sentences. To find out to what degree synthetic data might replace real user data, we combined the original and synthetic dataset in different ways:

- **user** contains only data from the original dataset
- **synthetic** contains only the synthetic dataset. Labels for the synthetic dataset are determined by the intention of the question posed to GPT-3.
- **mixed** combines **user** and **synthetic**
- **mixed predicted** combines **user** and **synthetic**, where **synthetic** labels are classified with a 95% confidence threshold by a baseline classifier trained on **user**.

We split the datasets into training and test sets (stratified 80:20 split, see Table 2 for an overview of the datasets and label distributions). We then fine-tuned $\text{BERT}_{\text{base}}$German-cased across three epochs to each training-set using 10-Fold Cross-Validation. At each fold, in addition to the validation set, we predicted the user test set after the third epoch.

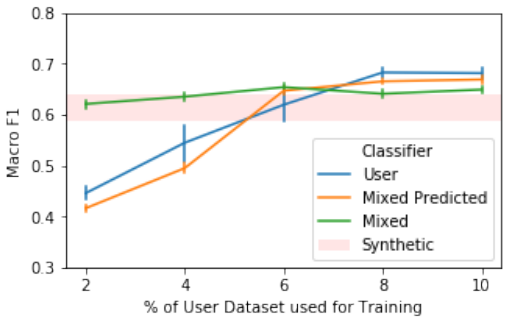
In addition to the classification experiment, we also analysed and compared the user and synthetic training sets on a structural and semantic level. The results from these comparisons, as well as a preliminary reflections on the meaningfulness of the GPT-3 outputs are outlined in Section 4.

3 CLASSIFICATION RESULTS

We did not see any improvements in classification of the user test set when incorporating synthetic data. The synthetic classifier performed 15 percentage points worse than the user classifier on this test set (see Table 1a in Figure 1).

Classifier	User Test Set	Val Set
user	**76.68 (1.17)**	75.04 (3.93)
mixed	73.52 (1.09)	78.52 (1.07)
mixed predicted	76.53 (1.26)	93.99 (0.68)
synthetic	61.72 (1.8)	80.66 (1.14)
stratified random	34.61	
majority	26.20	

(a) Mean Macro-F1 score (Standard Deviation) of classifiers trained on each training dataset on user test data and validation set.



(b) Classification performance when training with different subsets of **user** compared to results range of synthetic classifier (horizontal bar)

Fig. 1. Overview of the classification performance

	User training data	Synthetic training data
# data points	3,778	20,727
# words	58,030	195,552
# unique lemmas	5,584	3,685
mean lemma occurrence	10.39	53.07
# lemmas > 10 x occurrence	594	798
# significant keywords	327	143
Top 10 keywords	surgery (op), one (man), there (da), have (hab), times (mal), the (das), today (heute), is (ist), still (noch), goes (geht)	I (ich), would like (möchte), more (mehr), sport (sport), do (treiben), weight (gewicht), to (zu), will (werde), have (habe), healthier (gesünderen)

Table 3. Comparison of user and synthetic data

However, we note that the synthetic classifier's results on the user test set is still significantly higher than the random and majority baselines, indicating that the generated data does bear similarity to the user data and might be useful to a certain extent, when little data is available.

To explore this notion, we trained further classifiers with small fractions of our initial user training data set (between 2 and 10% at 2% increments). We also recreated the mixed and mixed predicted datasets the same way as the initial datasets, using these fractions. For the mixed predicted dataset, confidence thresholds for prediction had to be reduced to 50% for ensuring that each label was predicted. We find that at least 8% of the original dataset, corresponding to 302 samples, are needed to consistently reach better results than the synthetic classifier on the user test set (see Figure 1b).

4 COMPARISON OF NATURALISTIC AND SYNTHETIC DATA

To compare the user data and synthetic data on a structural level we examined the user and synthetic training sets in more detail, with a focus on language variability. Even though the synthetic dataset is over five times the size of the user dataset, the unique word count is lower by almost 2,000 lemmatized words and only has a third as many Bonferroni-corrected significant keywords at $p < 0.05$ when compared to the other dataset (see Table 3). This shows that a lot of words present in naturalistic data tend to not appear in synthetic data, while the opposite is less common. It also indicates less variability, regarding both content and vocabulary in the synthetic data. From the top keywords of each dataset when compared to the other, it becomes apparent that many keywords in the user dataset seem to be functional (i.e. the, is, there,...), while words in the synthetic dataset seems to bear more meaning (i.e. sport, more, weight,...), indicating that synthetic data seems to utilise less filler words common to natural, human-written language.

	TS	C	R
TS	0.2525	0.2484	0.2325
C		0.2869	0.2454
R			0.251

(a) Within text similarity of user sentences for each label

s \ u	TS	C	R
TS	0.2627	0.2639	0.2440
C	0.2593	0.2873	0.2488
R	0.2398	0.2549	0.2502

(b) Between text similarity of synthetic (s) and user (u) sentences

	TS	C	R
TS	0.3376		
C	0.3427	0.3894	
R	0.3043	0.3294	0.3171

(c) Within text similarity of synthetic data for each label

Table 4. Mean cosine similarity between all sentence embeddings of synthetic and user training sets within and between labels and datasets. Calculated with SentenceTransformers.

	Test Set		
Classifier	Mixed	Mixed Predicted	Synthetic
User	72.43 (1.5)	92.77 (0.5)	71.51 (1.79)
Mixed	**79.93 (0.2)**	84.31 (0.49)	**81.11 (0.29)**
Mixed Predicted	71.6 (0.84)	**94.26 (0.22)**	70.64 (1.05)
Synthetic	78.2 (0.38)	81.69 (0.53)	80.96 (0.24)
stratified random	33.2	33.08	32.69
majority	26.99	27.71	27.13

Table 5. Mean Macro-F1 score (Standard Deviation) in percent for synthetic, mixed and mixed predicted test sets.

To further test the assumption that the synthetic data has less variety than the user data, we created sentence embeddings for each sample in the datasets and compared the cosine similarity within and between labels and datasets using a pretrained sentence transformer model for the German language[4] (see Table 4). On average, synthetic samples were much more similar to each other than the user data. It also became apparent that, while the sentences in the user dataset tended to be most similar to sentences of the same label (Table 4a), this was not necessarily the case for the synthetic data, where R sentences showed more similarity with TS and C sentences than with themselves, and TS sentences were most similar to C sentences (Table 4c). When comparing the synthetic embeddings with the user embeddings, cosine similarities were similar to those within user data. Again, however, we noticed that synthetic sentences did not always have the highest mean similarity with user sentences from the same class. Both synthetic TS and R sentences seemed to be most similar to user C statements, although the difference in mean similarity was very small. A potential reason for this is the GPT-outputs for these labels containing certain words that are common for C in the user data. A more detailed qualitative examination of the outputs or the similarity measures between and across prompts given to GPT might shed light on these results.

The results above might give the impression that the generated outputs are simply not relevant to the classification task at hand, leading to low classification performance of the synthetic classifier on the user test set. To refute this assumption, we also had each classifier predict on the synthetic, mixed and mixed predicted test sets at each fold (see Table 5). From this, we draw a couple of interesting observations:

- Combining synthetic data with user data leads to better classification results on the synthetic test set.
- Classifying synthetic data with a classifier trained on user data is much more reliable than the other way around.
- The mixed predicted test set is easiest to predict for all classifiers, regardless of label ground truth.

These observations show that the synthetically generated data does bear similarity to the user dataset and that the notion of generating labelled data with prompts is feasible to a certain extent. If this were not the case, the prediction of synthetic data with the user classifier would have led to many label changes as compared to the intended classes and a worse performance of the synthetic and mixed classifier on the mixed predicted test set.

5 COST-BENEFIT ANALYSIS

As shown in Section 3, with real data, we need at least 302 user sentences, to reach better classification results on the user test set than the synthetic classifier. While this may not sound like much, collecting this amount of data in the case of our data source still involves multiple days of identifying suitable user posts, crawling and annotation. Only about 16% of forum posts screened to create the user dataset and 30% of the sentences in the resulting user posts contained

[4]https://huggingface.co/Sahajtomar/German-semantic

relevant information for this task[5]. On average, one post was 12 sentences long. Therefore, to obtain 302 relevant sentences, we have to screen

$$\frac{302}{12 \cdot 0.3 \cdot 0.16} = 524$$

posts, or 6288 (relevant or irrelevant) sentences, of which we can then identify $\frac{302}{12 \cdot 0.3} = 83.9$ relevant posts, or 1006.8 sentences to annotate in detail. Assuming 10 seconds to classify a sentence as relevant, it takes 17.47 hours to identify the required 1006.8 sentences. Assuming that it takes 20 seconds to annotate a sentence in a relevant post, the annotation process would take one person 5.6 hours. For label reliability, this person and at least two other annotators would have to label all relevant posts [1, p. 562], leading to a total time expenditure of 16.8 hours for annotation. The total process would then take 17.47 + 16.8 = 34.27 hours. Based on the median hourly wage of $ 28.72 of researchers in the U.S.[6], data collection and annotation for this small amount of data would cost $ 984.23. If similar data were collected in a user experiment, the time expenditure would likely be even higher. In comparison, generating synthetic data for this experiment cost $ 111.88. While we still had to do manual work to create the questions, many of the questions can easily be adapted to different topics by changing or rephrasing them slightly. Future work could also explore ways of automatising this process, for instance by applying rephrasing technology to example questions provided in Motivational Interviewing manuals [8, 17].

6 DISCUSSION

So, do we still need human assessors? This research suggests yes, but with caveats. We have shown that, while synthetic data generated by solely supplying prompts to GPT-3 appears coherent and plausible on a superficial level, it does not exhibit the same language variability as human-written data. Nevertheless, the synthetic classifier did significantly outperform random baselines. After our cost analysis, we conclude that when very few data is at hand and resources are limited, prompt-based data generation may lead to better classification results than collecting and annotating user data.

This work has a number of limitations that we plan to address in future work. Although we were able to elicit GPT-3 output fitting for the labels in our classification task, as can be seen by the high classification performance of the user classifier on the synthetic datasets, it is important to mention that the conversation style of the two datasets was slightly different due to the context in which they were created. While the user data was made up of forum posts, presenting a highly asynchronous form of online conversation that is reliant mainly on "interacting monologues", the GPT-3 output more closely follows a question-answer conversation style as it is expected in CAs. In future work, we plan to compare the performances of both classifiers on conversational user data collected for this context.

Initially, for this exploratory research, we generated 100 outputs for each question. As some questions might yield more varied results than others, we plan on taking a more structured approach to building the synthetic dataset in the future. One way to achieve this would be to identify stopping points, for instance when mean cosine similarity between output embeddings rises above a certain threshold such as the mean cosine similarity plus standard deviation of sentences in our user data. Other approaches include more sophisticated filtering of the synthetic data to weed out irrelevant, repetitive or nonsense output. Another way to increase the variety of generated outputs would be automatising the question generation process as mentioned in section 5 or to increase output temperature. Since the later would likely result in a decrease in relevance, exploring this variety-relevance trade-off might be an interesting avenue of research in itself.

[5] source withheld for anonymous reviews
[6] https://www.bls.gov/oes/current/oes193022.htm

We note that at these early stages of our research, we solely used prompts to generate synthetic text samples, and no GPT fine-tuning at all was involved in our experiments. To further explore ways of mitigating the cost of data collection, it would be interesting to explore to what extent fine-tuning GPT-3 to small shares of the dataset (for instance the 8% needed to outperform the synthetic classifier) or more general texts connected to health before data generation can improve output variability and classification performance. Furthermore, we intend to explore other data augmentation techniques [14] such as back-translation [2] or available paraphrasing technologies[7], which were not considered in this initial research due to scope limitations.

7 CONCLUSION

In this preliminary study, we explored the feasibility of replacing and enhancing user data with synthetic data in a conversational interaction context. To this end, we generated large amounts of synthetic data and set up a number of classification experiments to test the performance of classifiers trained on different combinations of naturalistic and synthetic data. We come to the conclusion that while it is possible to predict naturalistic data with a classifier fine-tuned on synthetic data, the results do not compare to classifiers fine-tuned on even very little naturalistic training data in the specific case of this study. Reasons include a difference in conversational style between user and synthetic data and a lack of content and semantic variability in the generated samples. Nevertheless, this approach could pose useful in certain situations, where resources and data are exceptionally scarce.

REFERENCES

[1] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics* 34, 4 (2008), 555–596.

[2] Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 597–604.

[3] Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. 2020. A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. *ACM Computing Surveys (CSUR)* (2020).

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Paweł Budzianowski and Ivan Vulić. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, Hong Kong, 15–22. https://doi.org/10.18653/v1/D19-5602

[7] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[8] Dawn Clifford and Laura Curtis. 2016. *Motivational interviewing in nutrition and fitness*. Guilford Publications.

[9] Arthur RT de Lacerda and Carla SR Aguiar. 2019. FLOSS FAQ chatbot project reuse: how to allow nonexperts to develop a chatbot. In *Proceedings of the 15th International Symposium on Open Collaboration*. 1–8.

[10] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. "What Can I Cook with these Ingredients?"-Understanding Cooking-Related Information Needs in Conversational Search. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2022), 1–32.

[11] Ahlam Fuad and Maha Al-Yahya. 2022. Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual Transformer Model mT5. *Mathematics* 10, 5 (2022), 746.

[12] Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. Human-Human Health Coaching via Text Messages: Corpus, Annotation, and Analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 246–256. https://aclanthology.org/2020.sigdial-1.30

[13] ARDB Landim, AM Pereira, Thales Vieira, E de B. Costa, JAB Moura, V Wanick, and Eirini Bazaki. 2021. Chatbot design approaches for fashion E-commerce: an interdisciplinary review. *International Journal of Fashion Design, Technology and Education* (2021), 1–11.

[7] https://huggingface.co/ceshine/t5-paraphrase-paws-msrp-opinosis

[14] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325.

[15] David E Losada, David Elsweiler, Morgan Harvey, and Christoph Trattner. 2021. A day at the races: using best arm identification algorithms to reduce the cost of information retrieval user studies. *Applied Intelligence* (2021).

[16] Antonis Maronikolakis, Mark Stevenson, and Hinrich Schütze. 2020. Transformers Are Better Than Humans at Identifying Generated Text. *ArXiv* abs/2009.13375 (2020).

[17] William R Miller and Stephen Rollnick. 2002. Motivational interviewing: Preparing people for change. Book Review. (2002).

[18] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. 42–51.

[19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[20] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[21] Gözde Gül Şahin. 2021. To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP. *Computational Linguistics* (2021), 1–38.

[22] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating Mixed-Initiative Conversational Search Systems via User Simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 888–896. https://doi.org/10.1145/3488560.3498440

[23] Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight Fire with Fire: Fine-tuning Hate Detectors using Large Samples of Generated Hate Speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4699–4705.

[24] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826* (2021).

[25] Munazza Zaib, Quan Z Sheng, and Wei Emma Zhang. 2020. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *Proceedings of the Australasian Computer Science Week Multiconference*. 1–4.