Data Augmentation for Intent Classification with Generic Large Language Models

Anonymous ACL submission

Abstract

Data augmentation alleviates the problem of data scarcity when training language models (LMs) by generating new examples based on the existing data. A successful approach to generate new samples is to fine-tune a pretrained LM on the task-specific data and then sample from the label-conditioned LM. However, fine-tuning can be difficult when task-specific data is scarce. In this work, we explore whether large pretrained LMs can be used to generate new useful samples without fine-tuning. For a given class, we propose concatenating few examples and prompt them to GPT-3 to generate new examples. We evaluate this method for few-shot intent classification on CLINC150 and SNIPS and find that data generated by GPT-3 greatly improves the performance of the intent classifiers. Importantly, we find that, without any LM fine-tuning, the gains brought by data augmentation with GPT-3 are similar to those reported in prior work on LM-based data augmentation. Experiments with models of different sizes show that larger LMs generate higher quality samples that yield higher accuracy gains.

1 Introduction

A key challenge in creating task-oriented conversational agents is gathering and labeling training data. The realistic training data resulting from actual human interaction with the agent does not exist until the system is launched. Prior to the launch, data gathering options include manual authoring and crowd-sourcing. Both of these options are tedious and expensive. *Data augmentation* methods aim to alleviate the data acquisition issue by automatically generating more examples based on available ones.

A particularly promising trend in recent research on data augmentation for natural language processing is using large pretrained language models (LMs) (Peters et al., 2018; Devlin et al., 2018) for this purpose. The general paradigm of these approaches is to fine-tune a LM on the task-specific data and then to generate new examples from a label-conditioned LM (Wu et al., 2018; Kumar et al., 2019, 2021; Anaby-Tavor et al., 2020; Lee et al., 2021). A potential issue with this approach is that when the task-specific data is scarce, fine-tuning a large LM on very few available examples can become the bottleneck.

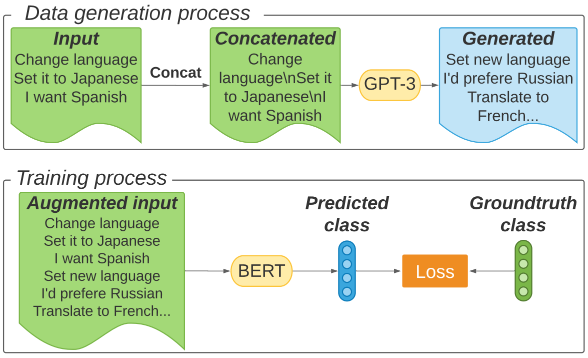


Figure 1: **Method.** We first generate new samples by prompting GPT-3 with a few examples from each class. Then we fine-tune a BERT classifier with the augmented dataset composed of the initial samples plus the GPT-3 generated samples.

In this work we investigate if large LMs can be used for generating new examples without any task-specific fine-tuning. We are inspired by GPT-3 results (Brown et al., 2020), which show that giant LMs can perform text classification when prompted with just a few examples. To use GPT-3 for data generation instead, we apply the example extrapolation approach by Lee et al. (2021). Namely, to create the prompt we concatenate several example utterances for a class (see Figure 1). Crucially, unlike Lee et al. (2021) we do not fine-tune the LM, and instead investigate if useful additional data can be generated by an off-the-shelf pretrained LM.

We focus this study on the task of predicting the user's intent, i.e. what the user of the task-oriented chatbot wants to accomplish. We show that in few-shot intent classification setups based on CLINC150 (Larson et al., 2019) and SNIPS (Coucke et al., 2018) datasets, the data generated by GPT-3 greatly improves performance of an intent classifier. Importantly, the improvements that our data generation method brings are comparable to those reported by Lee et al. (2021) who fine-tune a T5 model on held-out classes with many examples per class. Our intrinsic evaluation (Kumar et al., 2021) with an oracle classifier, shows that when given enough examples in the prompt, GPT-3 indeed latches onto the examples' intent and reliably generates more examples of the same intent. Our experiments with models of different sizes show that the LM's data generation abilities get better as the LM gets bigger.

2 Method

We consider the task of training an intent classifier. An intent is a type of request that the conversational agent supports; e.g. the user may want to change the language of the conversation, play a song, transfer money between accounts, etc. Collecting many example utterances that express the same intent can be difficult and expensive. In this paper, we experiment with an extremely simple method to augment the training data available for an intent. Namely, as shown in Figure 1, we select K examples for an intent, concatenate them with newlines and feed the resulting string as a prompt to a large generic LM, such as e.g. GPT-3 (Brown et al., 2020).

3 Experimental Setup

3.1 Datasets

We use CLINC150 (Larson et al., 2019) and SNIPS (Coucke et al., 2018) intent classification datasets in our experiments. CLINC150 has 23,700 example utterances, out of which, 1200 utterances belong to a special *out-of-scope* (OOS) class. The rest of the dataset covers 10 domains each consisting of 15 distinct intents. The dataset is balanced, with 100 training examples per intent, 20 (100 OOS) for validation, and 30 (1000 OOS) for testing. SNIPS is a dataset collected from the Snips personal voice assistant. The training set contains 13,084 utterances, whereas, the test and validation sets contain 700 utterances each. These utterances cover 7 different intents.

3.2 Setup

We simulate data sparsity in two ways. First, to produce comparable results to the example extrapolation approach (Ex2, Lee et al. (2021)), we consider the *partial few-shot* setting. In this setup, for a *few-shot subset* of intents, we truncate training data to K examples per intent.[1] For CLINC150, we use different domains as the few-shot subsets, whereas for SNIPS, we use the different intents. Second, similar to Vulić et al. (2021), we experiment with a more challenging *full few-shot* setting, in which we only keep K examples per intent for all the intents. When data augmentation is performed, we augment the few-shot intents to have N examples, where N is the median number of examples per intent of the original data.

To precisely describe the training and test data in all settings we will use D_{part} to refer to dataset parts, i.e. train, validation, and test. We use D_F for few-shot data and D_M to refer data-rich intents (in the partial few-shot setting). This notation is defined for all parts, therefore, $D_{part} = D_{\{F,part\}} \cup D_{\{M,part\}}, \forall part \in \{train, val, test\}$. When GPT-3 is used to augment the training data we generate $N - K$ examples per intent and refer to the resulting data as $\tilde{D}_{F,train}$. We experiment with four different-sized GPT-3 models[2] to obtain $\tilde{D}$: Ada, Babbage, Curie, and Davinci. In order, Ada is the smallest model (nearly 350M parameters) and Davinci is the largest (nearly 175B parameters).[3]

3.3 Training and Evaluation

We fine-tune BERT-large (Devlin et al., 2018) on the task of intent classification by adding a linear layer on top of the `[CLS]` token (Wolf et al., 2019).

Partial few-shot. In this setup, we train $\mathcal{S}$ intent classifiers, choosing a different few-shot subset of intents every time to obtain D_F. We then average the metrics across these $\mathcal{S}$ runs. For CLINC150, $\mathcal{S} = 10$ corresponding to the 10 different domains, whereas for SNIPS, $\mathcal{S} = 7$ corresponding to the 7 different intents. We evaluate our method on the following three scenarios introduced by Lee et al. (2021): (i) **Baseline**: models are trained without data augmentation on $D_{\{F,train\}} \cup D_{\{M,train\}}$. (ii) **Upsampled**: $D_{\{F,train\}}$ is upsampled to have N examples per intent. Then models are trained on upsampled

[1] We use the truncation heuristic provided by Lee et al. (2021): https://github.com/google/example_extrapolation/blob/master/preprocess_clinc150.py

[2] https://beta.openai.com/docs/engines

[3] https://blog.eleuther.ai/gpt3-model-sizes/

$D_{\{F,train\}} \cup D_{\{M,train\}}$. (iii) **Augmented**: models are trained on $D_{\{F,train\}} \cup \tilde{D}_{\{F,train\}} \cup D_{\{M,train\}}$.

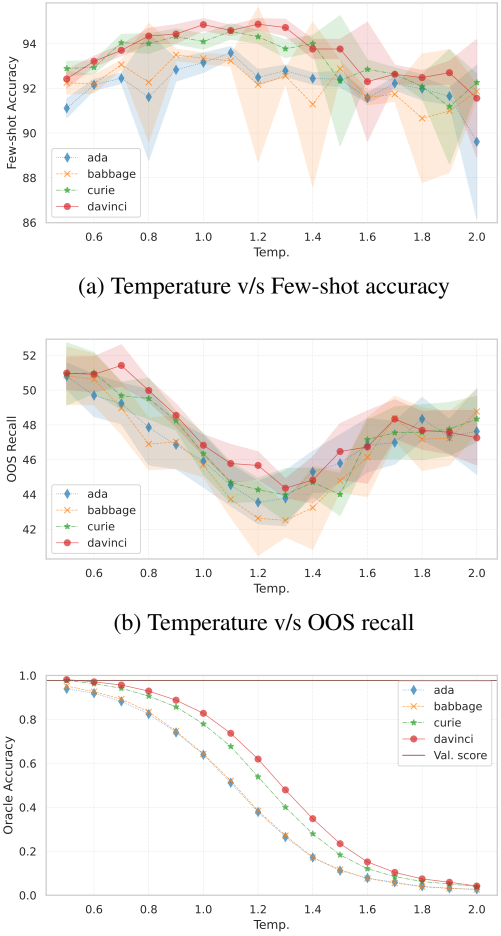
For each scenario, we report the 1) overall in-scope accuracy on the complete test set D_{test}, i.e intent classification accuracy excluding OOS samples in the test set, 2) out-of-scope recall (OOS recall) on D_{test} that we compute as percentage of OOS examples that the model correctly labeled as such, and 3) few-shot classification accuracy of the models on $D_{\{F,test\}}$. We also train an oracle $\mathcal{O}$ on $D_{train} \cup D_{test}$ and use it to measure the fidelity of samples generated by GPT-3. We define fidelity as the accuracy of the oracle classifier on the generated samples. A higher value denotes that the generated samples are more faithful to original data distribution.

Full few-shot. In this setup, we treat *all* the intents as few-shot and evaluate our method on the following three scenarios: (i) **Baseline**: all the intents are truncated to $K = 10$ samples per intent. (ii) **Augmented**: models are trained on $D_{\{F,train\}} \cup \tilde{D}_{\{F,train\}}$. (iii) **Augmented+Relabeled**: we use the oracle $\mathcal{O}$ to relabel the generated samples from GPT-3 and then train the intent classifier on relabeled $\tilde{D}_{\{F,train\}} \cup D_{\{F,train\}}$. The purpose of this experiment is to estimate what further gains can be achieved if the data generated by GPT-3 were labeled by the human.

For each scenario in this setup, we report (i) the overall in-scope classification accuracy, and (ii) out-of-scope recall. For both partial few-shot and full few-shot settings we report means and standard deviations over 10 repetitions of each experiment.

4 Experimental Results

Table 1 shows the results of our experiments on the CLINC150 and SNIPS datasets. By augmenting the dataset with GPT-3 generated samples, the few-shot accuracy improves by up to 2.7% on CLINC150 and 18.12% on SNIPS when compared to the baseline setting. We observe that using larger GPT-3 models improves the classifier performance across all metrics. Specifically, we show that the examples generated by *davinci*, the largest GPT-3 model, bring the largest boost to both few-shot and overall accuracies (this conclusion is significant at $p < 0.01$ level according to the standard T-test). Finally, our method achieves competitive results compared to Ex2 (Lee et al., 2021), both in terms of absolute accuracies and the relative gains



(a) Temperature v/s Few-shot accuracy

(b) Temperature v/s OOS recall

(c) Temperature v/s Fidelity

Figure 2: **Partial few-shot validation performance for different GPT-3 temperatures.** (a) few-shot accuracy, (b) OOS recall, (c) oracle accuracy, for classifiers trained on augmented sets generated by GPT-3 models of different sizes and with different temperatures.

brought by data augmentation. Note that Ex2 uses T5-XL (Roberts et al., 2020) with nearly 3 billion parameters as its base intent classifier, while our method uses BERT-large with only 340 million parameters.

Table 2 shows results of our full few-shot experiments. Data augmentation by GPT-3 is also helpful in this scenario, especially when larger engines are used. The Ex2 method is not applicable because there is no data-rich intents to train the example extrapolator on. We compare instead to (Vulić et al., 2021) who consider a similar full few-shot setting with $K = 10$ examples and report similar results. Notably, augmenting the training data with Davinci brings an improvement of the same magnitude as that achieved by Vulić et al. (2021), who change the classification approach. Lastly, relabeling the

			CLINC150			SNIPS	
			Overall		Few-shot	Overall	Few-shot
	Model	Classifier	Inscope Acc.	OOS Recall	Acc.	Inscope Acc.	Acc.
Baseline (Lee et al., 2021)	-	T5	97.4	-	93.7	95.2	74.0
Upsampled (Lee et al., 2021)	-	T5	97.4	-	94.4	95.9	80.0
Augmented (Lee et al., 2021)	Ex2	T5	97.4	-	95.6	97.8	94.0
Baseline (ours)	-	BERT	96.28 (0.06)	39.14 (0.82)	91.36 (0.47)	95.47 (0.45)	78.38 (3.34)
Upsample (ours)	-	BERT	96.20 (0.05)	40.21 (0.59)	90.93 (0.19)	95.29 (0.37)	79.28 (2.05)
Augmented (ours)	Ada	BERT	96.09 (0.06)	33.30 (1.07)	92.20 (0.37)	97.39 (0.23)	95.16 (0.44)
Augmented (ours)	Babbage	BERT	96.15 (0.04)	33.17 (0.83)	92.41 (0.35)	97.34 (0.11)	94.30 (0.78)
Augmented (ours)	Curie	BERT	96.36 (0.07)	34.90 (0.86)	93.43 (0.39)	97.37 (0.24)	94.90 (0.74)
Augmented (ours)	Davinci	BERT	**96.45** (0.07)	35.55 (0.80)	**94.06** (0.26)	**97.67** (0.18)	**96.50** (0.54)

Table 1: **Partial few-shot results on CLINC150 and SNIPS datasets.** Refer to Section 3.3 for more details.

Model	**Aug.**	**Relabel**	**Inscope Acc.**	**OOS Recall**
BERT+MLP♠			89.88	-
BERT+Sim♠			91.80	-
Baseline			90.28(0.49)	50.18(1.14)
Ada	✓		90.32 (0.28)	19.90 (2.65)
Babbage	✓		91.10 (0.24)	19.05 (1.23)
Curie	✓		92.32 (0.35)	19.05 (1.23)
Davinci	✓		93.52 (0.30)	24.00 (3.25)
Ada	✓	✓	95.64 (0.06)	**82.77** (0.39))
Babbage	✓	✓	96.32 (0.02)	80.15 (0.75)
Curie	✓	✓	**96.66** (0.08)	72.28 (0.84)
Davinci	✓	✓	96.18 (0.01)	77.65 (0.75)

Table 2: **Full few-shot results on CLINC150.** The third section shows results with data augmentation, and section four shows results for augmentation and relabeling of samples by an oracle. ♠ denotes the numbers are taken from Vulić et al. (2021)

generated data by the oracle gives a big boost to accuracies for all engines, confirming our hypothesis that human inspection of the generated data could be fruitful. Relabeling also has a large impact on OOS recall, which is due the fact that much of the generated data was labeled as OOS by the oracle.

4.1 Analysis

Figure 2 shows how validation few-shot accuracies and OOS recall vary when different generation temperatures are used for GPT-3. We observe that for all engines the generated data is most helpful with temperature around 1.0, although lower temperatures result in higher OOS recall. We also observe that the fidelity of the generated samples decreases as we increase the temperature (i.e. higher diversity, see Figure 2c). This suggests better fidelity does not always imply better quality samples as the language model may produce less diverse utterances at lower temperatures. In Appendix A, we perform a human evaluation, reaching similar conclusions as when using an oracle to approximate fidelity.

We refer the reader to Table 3 for examples of generated utterances.

Domain	Input examples	Generated examples
Banking	send 2000 dollars between chase and rabobank accounts	transfer between two accounts
	move money from one account to another	need to send half a million dollars from a bank to a broker firm
	money transfer request	to send some money from dtrusts to b of a
Home	take carrots off my list for shopping	i'm out of kleenex will you add that to the shopping list
	i'm out of bananas; add to shopping list	take batteries off my shopping list
	add sprite to my shopping list	my shopping list has no item on it that begins with "c" please
Small talk	what is life's meaning	can you tell me life's meaning
	what's the point of this dumpster fire known as life	should we try to figure out why we exist or we can just dance around in the rain and live for the moment and not worry about life and what it is
	whats your take on the meaning of life	how do you ask ....

Table 3: **Qualitative results.** For each class we show some input samples and we cherry picked some generated samples. Green samples are considered good ones and red ones are considered bad ones.

5 Conclusion

We propose to prompt large pretrained language models to perform data augmentation in the few-shot intent classification regime. Experiments on CLINC150 and SNIPS show that GPT-3-generated examples significantly improve the performance of few-shot intent classifiers without finetuning and that larger models produce more useful additional data. Our oracle experiments suggest that further gains can be achieved if the generated data were labeled by a human. In future work we will experiment with data generation by other language models, as well as with approaches to identify the generated examples that require human relabeling.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do Not Have Enough Data? Deep Learning to the Rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

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. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. *arXiv:1805.10190 [cs]*. ArXiv: 1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2019*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2021. Data Augmentation using Pre-trained Transformer Models. *arXiv:2003.02245 [cs]*. ArXiv: 2003.02245.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. 2019. A Closer Look At Feature Space Data Augmentation For Few-Shot Intent Classification. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural Data Augmentation via Example Extrapolation. *arXiv:2102.01335 [cs]*. ArXiv: 2102.01335.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018*. ArXiv: 1802.05365.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2018. Conditional BERT Contextual Augmentation. *arXiv:1812.06705 [cs]*. ArXiv: 1812.06705.

Appendix

A Human Evaluation

In Figure 2 we evaluate the fidelity of the samples generated by GPT-3 with respect to the original set of sentences used to prompt it. Fidelity is approximated by the classification performance of an "oracle" intent classifier trained on the whole dataset ($D_{train} \cup D_{test}$) and evaluated over the generated samples. In order assess whether the oracle predictions are comparable to those of a human, we perform a human evaluation study.

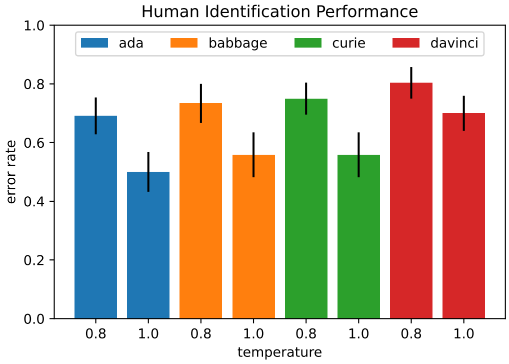
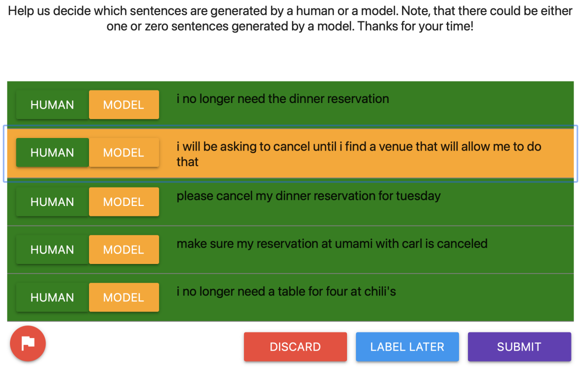


Figure 3: **Human evaluation.** Error rate of human evaluators at the task of finding whether any sentence in a group of 5 was generated by GPT-3 or not. Each color represents a different GPT-3 engine. Higher error rate indicates that humans could not correctly identify generated samples and thus it also indicates higher fidelity. The standard error is displayed as a vertical line on top of each bar.



Help us decide which sentences are generated by a human or a model. Note, that there could be either one or zero sentences generated by a model. Thanks for your time!

HUMAN MODEL i no longer need the dinner reservation

HUMAN MODEL i will be asking to cancel until i find a venue that will allow me to do that

HUMAN MODEL please cancel my dinner reservation for tuesday

HUMAN MODEL make sure my reservation at umami with carl is canceled

HUMAN MODEL i no longer need a table for four at chili's

Figure 4: **Human evaluation tool.** Example of a question for the human evaluators. Human evaluators are asked to flag which example is GPT-3 generated if any among the 5 presented ones.

We consider that a model produces sentences with high fidelity if a human is unable to distinguish them from a set of human-generated sentences belonging to the same intent. Therefore, for each intent in the CLINC150 dataset, we sample five random examples and we randomly choose whether to replace one of them by a GPT-3 generated sentence from the same intent. We generate sentences with each of the four GPT-3 models considered in the main text with two different temperatures (0.8 and 1.0). The sentence to replace is randomly selected. Finally, the five sentences are displayed to a human who has to choose which of the sentences is generated by GPT-3, if any.

The task is presented to human evaluators in the form of a web application (see Figure 4). We placed a button next to each sentence in order to force human evaluators to individually consider each of the examples. Once annotated, the evaluator can either *submit*, *discard*, or leave the task to *label later*. We used a set of 15 voluntary evaluators from multiple backgrounds, nationalities, and genders. Each evaluator annotated an average of 35 examples, reaching a total of 500 evaluated tasks.

For each model and temperature, we report the error rate of humans evaluating whether a task contains a GPT-generated sample. We consider that evaluators succeeds at a given task when they correctly find the sentence that was generated by GPT or when they identify that none of them was generated. Thus, the error rate for a given model and temperature is calculated as #failed / total_evaluated.

Results are displayed in Figure 3. We find that human evaluators tend to make less mistakes when the temperature used to sample sentences from GPT-3 is smaller. This result is expected since lowering the temperature results in sentences closer to those prompted to GPT-3, which are human-made. We also observe that models with higher capacity such as `Davinci` tend to generate more indistinguishable sentences than lower-capacity models such as `Ada`, even for higher temperatures. These results are in agreement with the "oracle" fidelity results introduced in Figure 2.