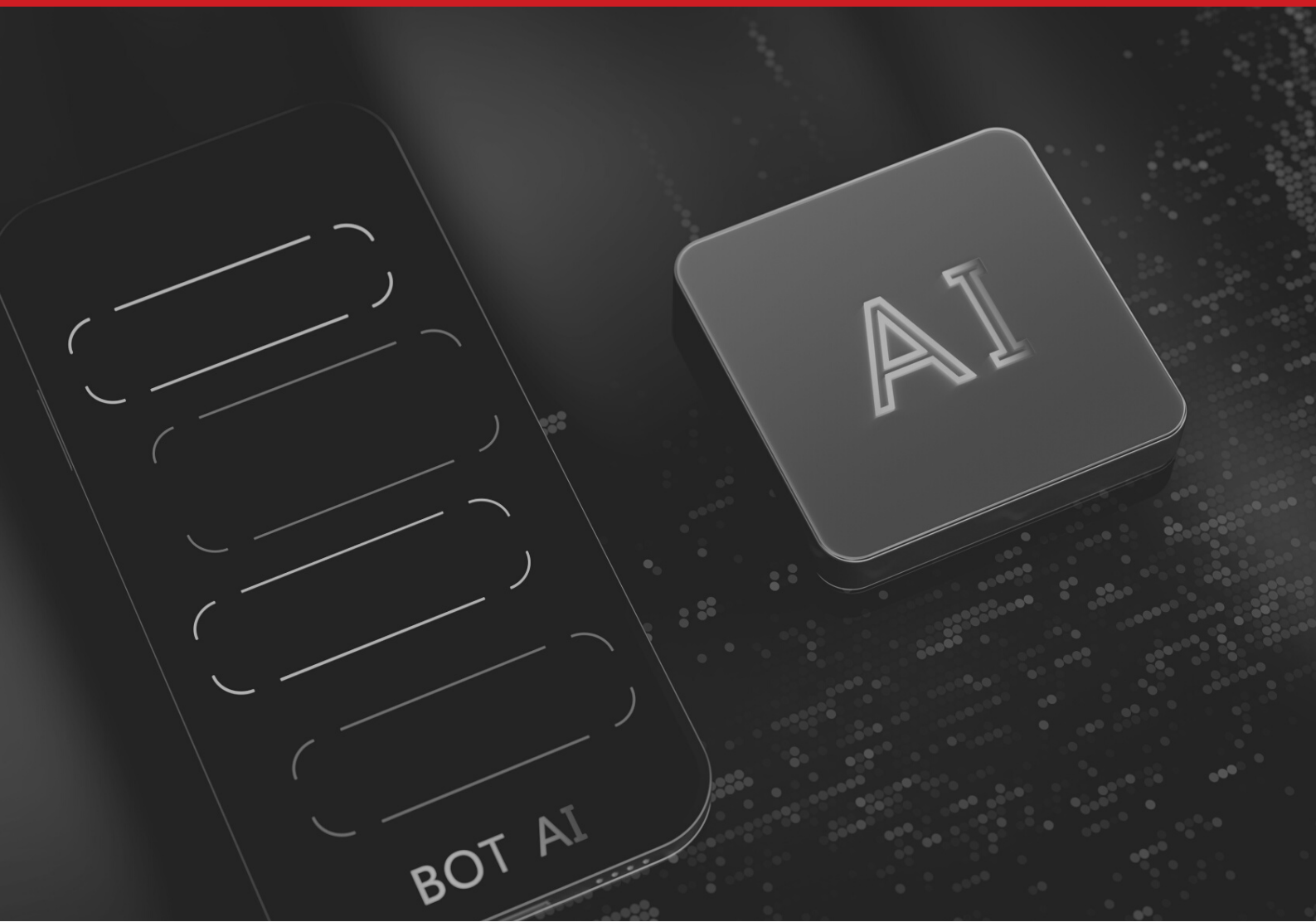
bitext we help AI understand humans

Fine-Tuning GPT-3 for Intent Detection

Optimizing Intent Recognition in Chatbots with Generative Large Language Models



Introduction

Generative Large Language Models (LLMs) were proved to achieve high performance when used to solve different business problems, ranging from generating well-written articles to classify the sentiment of a sentence. Furthermore, training these models on custom data is a direction that is worth exploring because it allows leveraging the general knowledge that is stored in these models and adapt it to custom and real-life use-cases.

One interesting application is the recognition of the user's intent, which is often embedded as an essential language understanding component in Chatbots. Designing and developing the capability of accurately recognizing the user's intent is a complex procedure that requires studying each specific use-case and the expected audience. It is also a repetitive process where it is essential to study the system in production and update the training data to reflect new patterns in users' conversations.

The performance of the intent recognition component is highly dependent on providing the machine learning model with high-quality user utterances with their actual respective intents, so the model can learn to map patterns in language to intent labels. Ensuring that the training data reflect the tone and common language used by the expected audience and ensuring that it covers the different ways a user may express their intent is very important.

At Bitext, we worked on using our proprietary text generation technology to develop high-quality data to train and customize state-of-the-art LLMs for intent recognition. In this work, we showcase using Bitext's Customer Service Open Dataset, that we have published publicly on our GitHub account[1], to customize OpenAI GPT-3 for intent recognition. Our experiments show that using highly-curated language data leads to over 90% accuracy out-of-the-box. Next sections provide details on our used data, the experiments, and the results.

[1] https://github.com/bitext/customer-support-intent-detection-evaluation-dataset

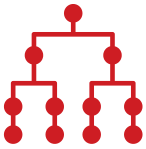
Data

Our Customer Service Open Dataset covers 27 user's intents like 'place an order', 'get refund', and 'contact customer service'. The dataset covers both very distinct intents like 'check payment methods' and 'recover password' and fine-grained closely-related intents like 'get refund' and 'track refund'.

This diversity in intents' closeness is important to challenge the machine learning model and to develop a more robust model. To study the effect of training data size on model's performance, we created several subsets of the data with 1, 100, 250, 500, 1000, 5000 utterances per intent, and trained different models with each subset.

We prepared our data by replacing intent labels with numbers, which is a common practice when training multi-class classification models.

Experiments



We worked with OpenAI's ada model. OpenAI offers at the moment the following versions of GPT-3 for fine-tuning: ada, babbage, curie and davinci. The ada model is often recommended for classification tasks because it is faster and cheaper than the other models while providing comparable performance. We fine-tuned several instances of the ada model with the following data and parameters:

(1) using 1 utterance per intent for 20 epochs and with a batch size of 1 (2) using 100 utterances per intent for 12 epochs and with a batch size of 4 (3) using 250 utterances per intent for 4 epochs and with a batch size of 8 (4) using 500 utterances per intent for 4 epochs and with a batch size of 16 (5) using 1000 utterances per intent for 4 epochs and with a batch size of 32 (6) using 5000 utterances per intent for 12 epochs and with a batch size of 256.

To ensure accurate results, we also developed both a comprehensive validation set of ~ 1000 examples and a test set of ~ 1100 examples. This additional data were developed both manually and using GPT-3.5 after prompting it with 10 utterances per intent from our customer service data. For generated data, we reviewed it manually to exclude sentences with hallucinated or irrelevant content. - As a separate note, we find that ~ 9% of the generated data contains some kind of hallucinated or irrelevant content. - When generating the data, we prompted the model with different contexts to make our data as diverse as possible, including generating utterances from polite and angry customers, and using different styles of writing like sarcastic, creative, and ordinary writings. Also, we generated both lengthy and short sentences.

Results

After training, we report the accuracy of each model on our custom test set. The reported accuracy scores refer to the percentage of sentences in the test set that the model was able to classify into their respective correct intents. The following table summarizes the accuracy scores:

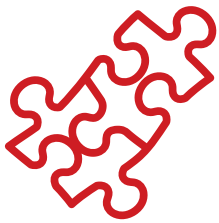
Data	Accuracy (%)
1 utterance per intent	23.5
100 utterances per intent	91.89
250 utterances per intent	92.17
500 utterances per intent	91.25
1000 utterances per intent	91.71
5000 utterances per intent	88.79

We see that GPT-3 achieves low performance when training with only 1 utterance per intent, achieving 23.5% accuracy. When training with 100 utterances per intent using Bitext's Customer Service data, the accuracy goes up to ~ 91%. We notice also that the performance reaches ~ 92% when training with 250 utterances per intent.

Adding more utterances without any customization doesn't lead to improvement in performance. When training with 5000 utterances per intent, we notice that the performance goes down to ~ 88.8% accuracy. This last model was trained for 12 epochs and we expect that training it with more epochs is expected to increase its accuracy, but we left that for upcoming experiments.

The accuracy scores we got is from using our data out-of-the-box and without further customization. Doing a cycle of evaluating the model in a live environment and re-training it with updated user utterances is expected to further enhance the performance.

Conclusion



We worked on exploring how to optimize the performance of the intent recognition components in Chatbots using generative LLMs. In this work, we did several experiments to customize OpenAI GPT-3 with our Customer Service Open Dataset. To test the effect of the size of training data on the model's performance, we trained with different subsets of our data ranging from 1 to 5000 utterances per intent.

We found that training GPT-3 with our data achieves over 90% accuracy out-of-the-box. Training with 250 sentences per intent achieved ~ 92% accuracy. A future direction is to deploy a live Chatbot with the obtained intent recognition capability, then doing a cycle of evaluation and re-training by including users' feedback. By doing this, we expect the accuracy to rise up to over 95%.

Select Customers