# Training Data Augmentation

## AI Transparency Technical Note

March 19, 2024

# Introduction

At Cisco, we believe that Artificial Intelligence (AI) can be leveraged to power an inclusive future for all. We also recognize that, by applying this technology we have a responsibility to mitigate potential harm. That is why Cisco adheres to our Responsible AI Framework, which is based on six principles of Transparency, Fairness, Accountability, Privacy, Security and Reliability (the "Framework Principles"). Cisco translates the Framework Principles into product development requirements, which ultimately form part of the product development lifecycle alongside our Security by Design, Privacy by Design, and Human Rights by Design processes.

Accordingly, Webex connect features that leverage AI are built with transparency, fairness, accountability, privacy, security, and reliability at their core. Each feature powered by AI undergoes an AI Impact (AII) Assessment — a best-in-class review of how the technical underpinnings of the functionality measure against the Framework Principles.

The training data augmentation feature (surfaced in the Intent editor for Task bots in Webex connect bot builder) was built with the Framework Principles at the center of how we deliver the AI-powered technology. This technical note describes more information about the feature and the AI underpinning it. This is an optional beta feature and clients choose to get it disabled for its users if needed.

# Feature Overview

The training data augmentation capability which is available in Webex connect bot builder addresses the "cold start" problem with bot building - developers having to manually add training data to their intents to get the bot working at a reasonable accuracy. The training data consists of different ways in which a user can invoke the same intent. Creating this training corpus manually can be tedious and time consuming. Developers tend to cut corners in this crucial part of bot building by adding only a few variants, or by adding only keywords as variants instead of meaningful sentences. This feature simplifies the process by allowing developers to use AI-powered training data generation by clicking the 'Generate' button in their intents.

# Model Overview

## Model Architecture

This is a capability built by Webex connect by leveraging a third-party Large Language Model (LLM) from Microsoft's Azure OpenAI Service. For more details on Azure Open AI Service and how it handles data, please review Microsoft's transparency note available here.

Training data augmentation currently uses the GPT-3.5 turbo model offered by Azure OpenAI Service. Based on our research and benchmarking exercises, we found this model to provide the best value for this use case in terms of cost, latency & quality. However, we are constantly evaluating several other alternate models – both in-house as well as from 3rd parties – which we are likely to add as options in the product.

### Model Inputs and Outputs

Input to the model is the intent name, intent description, and existing training data along with the appropriate prompt and instructions. Output from the model would be the requested number of training utterances generated by the LLM.

### Usage Guidelines

Support for training data generation is currently limited to task bots. Users can generate a maximum of 20 training utterances in one go. Training data generation is enabled when at least one training utterance is added along with the intent name. It is disabled once the intent has more than 50 training utterances. The quality of generated training data is dependent on the existing training data and the intent description, so it is recommended to provide an appropriate description of the intent and the slots required in the intent.

### Data Sources for Training and Evaluation

For details about the data sources used by the underlying model from Microsoft's Azure OpenAI Service, please review the GPT-3 Model Card published by OpenAI.

## Model Evaluation and Performance

At Cisco, we are constantly evaluating the models used for all our AI features to improve performance of any given feature. Humans with the appropriate roles and permissions are involved in the review, testing, and quality assurance processes and may sample the inputs to and outputs from these models periodically.

## Safety and Ethical Considerations

All our third-party vendors, including Microsoft, undergo rigorous vendor reviews, which include security and safety assessments.

Given the non-deterministic nature of LLMs, the model we use may output toxic, harmful or unsafe content if such content is present in the description of the code requested by developers. Microsoft does provide safeguards such as content filtering which is enabled by Cisco and can help mitigate these issues to some extent. Microsoft also attempts to improve mathematical reasoning with process supervision. However, processes such as abuse monitoring, which require logging the data for verification by humans are turned off by Cisco.

## Fairness

To understand how the model is trained for fairness, please reference Microsoft Azure OpenAI Transparency Note.

## Privacy and Security

Information about how we approach processing of and security around personal data, including data retention periods, etc., can be found in our Privacy Data Sheets, found on the Cisco Trust Portal.

As mentioned above, humans with the appropriate roles and permissions are involved in the review, testing, and quality assurance processes and may sample the inputs to and outputs from these models periodically.

## Updates and Maintenance

All changes to the product are documented in the appropriate artefacts such as the release notes, change logs, user guides, etc. We will update this transparency note as and when a change is warranted due to updates to the underlying model or our data processing.

## References

RAI Principles

RAI Framework

Cisco Trust Portal

Content Filtering

Abuse monitoring

Transparency Note for Azure OpenAI Service

User guide