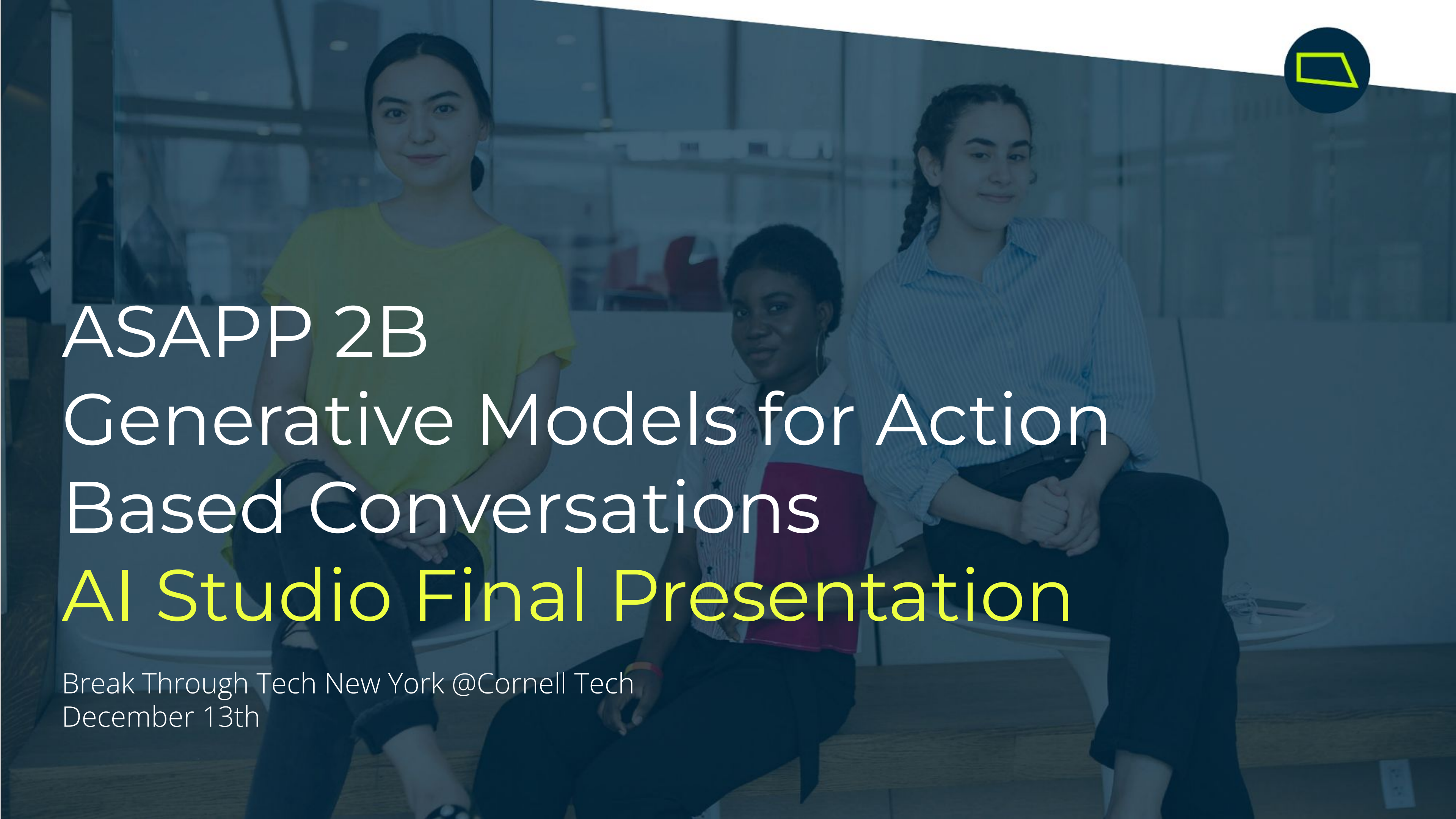


ASAPP 2B
Generative Models for Action Based Conversations
AI Studio Final Presentation

Break Through Tech New York @Cornell Tech
December 13th

Introductions

Meet Our Team!

Maria Reccoppa
Molloy University

Grace Li
Barnard College

Lana Saadeddin
Montclair State University

Bianca Kim
Baruch College

Damaris Campos
Queens College

Our AI Studio TA and Challenge Advisors

Anvita Bhagavathula
AI Studio TA

Paloma Sodhi
Challenge Advisor

Presentation Agenda

1. AI Studio Project Overview
2. Data Understanding & Preparation
3. Modeling & Evaluation
4. Insights and Key Findings
5. Final Thoughts
6. Questions

AI Studio Project Overview

Build a customer service chatbot that can anticipate agent responses and suggest next steps

Our Goal

How can we build a Large Language Model to help customer service agents respond more effectively and increase productivity to different requests?

Business Impact

- **Customer Satisfaction** - reduce time agents spend on each ticket, allowing them to handle more queries
- **Operational Efficiency** - quicker and more helpful responses can lead to higher customer satisfaction
- **Error Reduction** - reduce the chances of human error by suggesting accurate responses

Our Approach

Investigate the structure and format of the Action Based Conservation Dataset

Build an understanding of the problem and the business purpose

September

October

Create a training dataset for the model and perform data processing to ensure training data is in the correct format

November

Finetune DistilGPT2 and conduct preliminary evaluations

Define next steps to improve agent response predictions

December

Finalize presentation and summarize findings from our experimentation

Resources We Leveraged

NumPy

Hugging Face

pandas

Data Understanding & Data Preparation

Dataset Overview and Understanding

- JSON file of ABCD data (Action-Based Conversations Dataset)

```
['agent','ok, unfortunately because it has been more than 90 days we
cannot accept the return. Would there be anything else I can help you
with?'],
['customer', 'What if I ask really, really nicely?'],
['agent', "I can escalate to my manager if you'd like"],
['agent', "I'd just need your phone number."],
['customer', '(977) 625-2661'],
['action', 'Details of (977) 625-2661 have been entered.'],
['action', 'The manager has been notified.'],
```

- Contains a set of customer-agent conversations and what actions the agent takes

Data Selection and Cleaning

Approach 1: Not Including Action Tags

['agent', "I'd just need your phone number."],
['customer', '(977) 625-2661'],
~~['action', 'Details of (977) 625-2661 have been entered.'],~~
~~['action', 'The manager has been notified.'],~~
['customer', "I'll look forward to hearing from them."]

Approach 2: Including Action Tags

['agent', "I'd just need your phone number."],
['customer', '(977) 625-2661'],
['action', 'Details of (977) 625-2661 have been entered.'],
['action', 'The manager has been notified.'],
['customer', "I'll look forward to hearing from them."]

All utterances in a conversation are concatenated into one line of a .txt file

Modeling & Evaluation

Model Training

Trained for the task of Causal Language Modeling: using previous utterances to predict the next agent response in a conversation

	Model 1: DistilGPT2	**Model 2: GPT2-medium**
Fine Tuning Dataset 1: W/ Action	DistilGPT2 finetuned w/ action	GPT2-medium finetuned w/ action
Fine Tuning Dataset 2: W/O Action	DistilGPT2 finetuned w/o action	GPT2-medium finetuned w/o action

Testing and Evaluation

Approach 1: Concatenating Consecutive Utterances

```
['agent', 'Hi!'],
['agent', 'How can
I help you?'],
['customer', 'Hi!
I need to return
an item, can you
help me with
that?'],
['agent', 'sure,
may I have your
name please?'],
```

```
['agent', 'Hi! How
can I help you?'],
['customer', 'Hi!
I need to return
an item, can you
help me with
that?'],
['agent', 'sure,
may I have your
name please?'],
```

Approach 2: Keeping Each Utterance Separate

```
['agent', 'Hi!'],
['agent', 'How can
I help you?'],
['customer', 'Hi!
I need to return
an item, can you
help me with
that?'],
['agent', 'sure,
may I have your
name please?'],
```

```
['agent', 'Hi!'],
['agent', 'How can
I help you?'],
['customer', 'Hi!
I need to return
an item, can you
help me with
that?'],
['agent', 'sure,
may I have your
name please?'],
```

Model Comparison

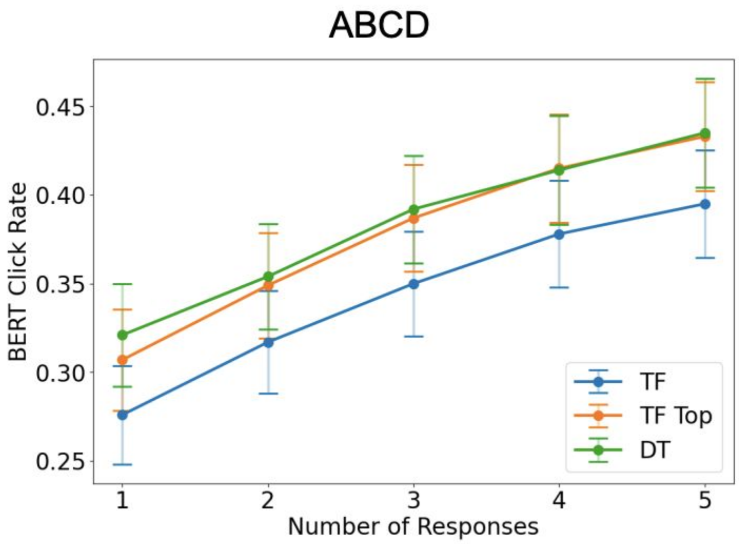
model	concat sequential utterance	non-concat sequential utterance
GPT2-medium w/ action	0.226	**0.262**
GPT2-medium w/o action	**0.230**	0.259
DistilGPT w/ action	0.152	0.196
DistilGPT w/o action	0.147	0.194

Note: concatenated utterance evaluations **should be** performed on a training data with concatenated utterances in the training data. Part of the next steps!

Model Comparison on Non-Concat

Context	True Agent Response	DistilGPT w/ action	DistilGPT w/o action	GPT2-medium w/ action	GPT2-medium w/o action
" …customer: Sure, it's Crystal Minh and my account ID is SETPNFNIRL"	Can I please have your username, email address and order number?	"I'm sorry to hear that Can I please have your username, email address, and order ID?"	"I'm sorry to hear that Can I please have your username, email address, and order ID?"	"And what is the shipping status of your order?"	"Thank you, Crystal Can I also have your username, email address, and order ID?"
F1 Scores:	1	0.539	0.539	0.048	0.507

Contextualizing Our Results



Sodhi et. al. "On the Effectiveness of Offline RL for Dialogue Response Generation"

Training Method	BERTScore
Offline RL	0.404
Fintuning LLM	0.262

Our Hypothesis

1. **DistilGPT vs. GPT2-medium**

   ➔ A bigger model will perform better than a smaller one.

2. **Concatenated Agent Utterances vs. Non-Concatenated**

   ➔ Concatenated agent utterances increases the context length for generations which will increase errors in generation

3. **Finetuning w/ Action vs. w/o Actions**

   ➔ Including Actions will provide more details for the LLM to predict the next sequence of characters

Key Findings

DistilGPT2

- Promising accuracy in predicting agent responses
- High BERT scores for precision, recall, and F1
- Explored larger model size for potential benefits in performance

+

GPT2-Medium

- Larger model size (355M parameters) compared to DistilGPT2 led us to higher results
- High results without action tag and concat is .230 and with tag and non-concat with .262

→

Overall

- A bigger model performs better by 0.06
- Incorporating the action tag boots performance about 0.02-0.03
- Reassessing concatenation performance on a new fine-tuning dataset
- Results are inconclusive; however, our hypothesis finds support in following research: "Why Exposure Bias Matters: An Imitation Learning Perspective of Error Accumulation in Language Generation"

Insights

Recall of goal:

Objective was to build a Large Language Model to aid customer service agents in responding more effectively and increasing productivity in handling various requests.

Impact on Customer Service

- Improved accuracy means our responses are more relevant and higher in quality, aligning with ASAPP's mission statement to boost agent productivity by our results:
    - customer service by offering more precise solutions for different queries

Final Thoughts

What We Learned

Data Prep

Importance of data cleaning, tokenization

Select Features

Power of feature engineering in NLP

Fine-tuning

General ML process, hyperparameters

Eval Metrics

Choice of metrics (accuracy, precision, recall and F1 score)

Challenges

Difficulties in handling unstructured text data, Git version Control

Considerations

How can we improve our NLP model?

Potential Next Steps

Further enhance model to add more data and plug them

FRESH NEW DATA

PLUG NEW MODEL

Explore further fine-tuning / pre-trained models and utilize

Combine predictions from multiple models for better performance

ENSEMBLE METHODS

USER FEEDBACKS

Incorporating user feedback into the model
Aligned in real-world scenarios

Questions?