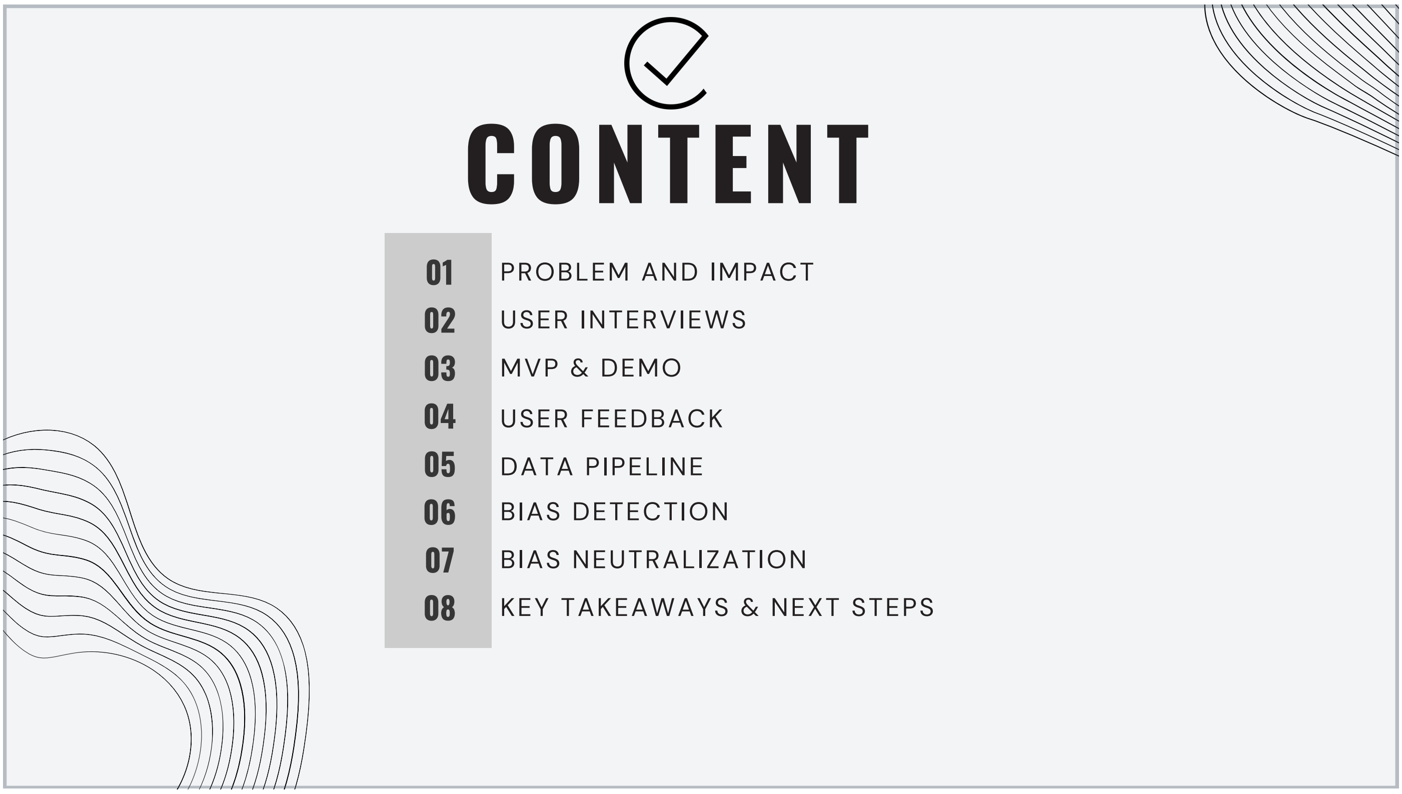
BIASCHECK

minimizing subjectivity in language

Etienne Ndidi Rachel Tripoli Sinan Urediler Violet Davis

Fall 2024 Capstone Final Presentation

Good evening everyone! Tonight myself, Violet, Etienne, and Rachel will be providing an overview of our product, BIASCheck. Our mission is to minimize subjectivity in language.



CONTENT

01 PROBLEM AND IMPACT

02 USER INTERVIEWS

03 MVP & DEMO

04 USER FEEDBACK

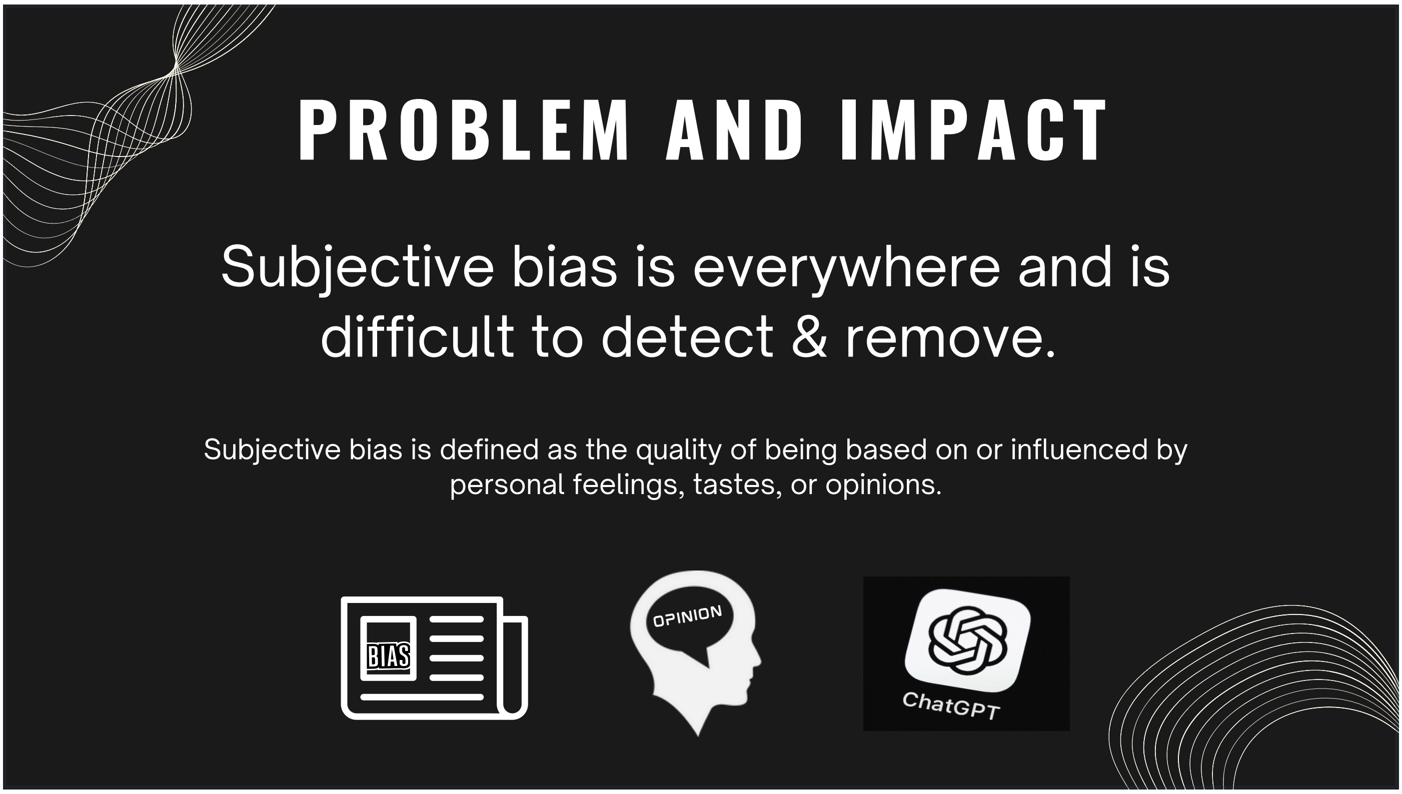
05 DATA PIPELINE

06 BIAS DETECTION

07 BIAS NEUTRALIZATION

08 KEY TAKEAWAYS & NEXT STEPS

These are the areas we will be covering for tonight.



PROBLEM AND IMPACT

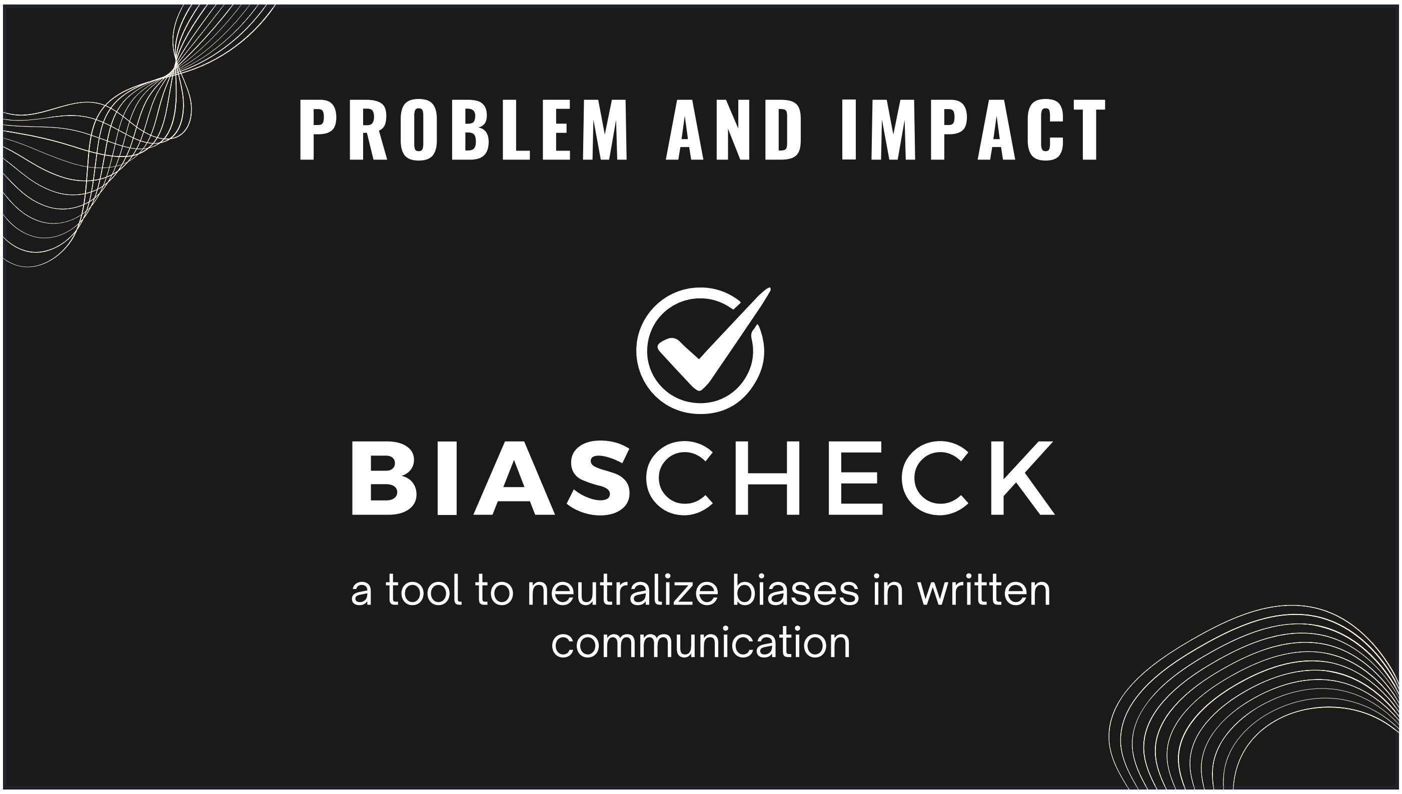
Subjective bias is everywhere and is difficult to detect & remove.

Subjective bias is defined as the quality of being based on or influenced by personal feelings, tastes, or opinions.

BIAS
OPINION
ChatGPT

Subconscious subjective biases are often embedded in everyday communications, even when a writer aims to be neutral.

Therefore, the problem we are focusing on is minimizing subjective biases in written language.

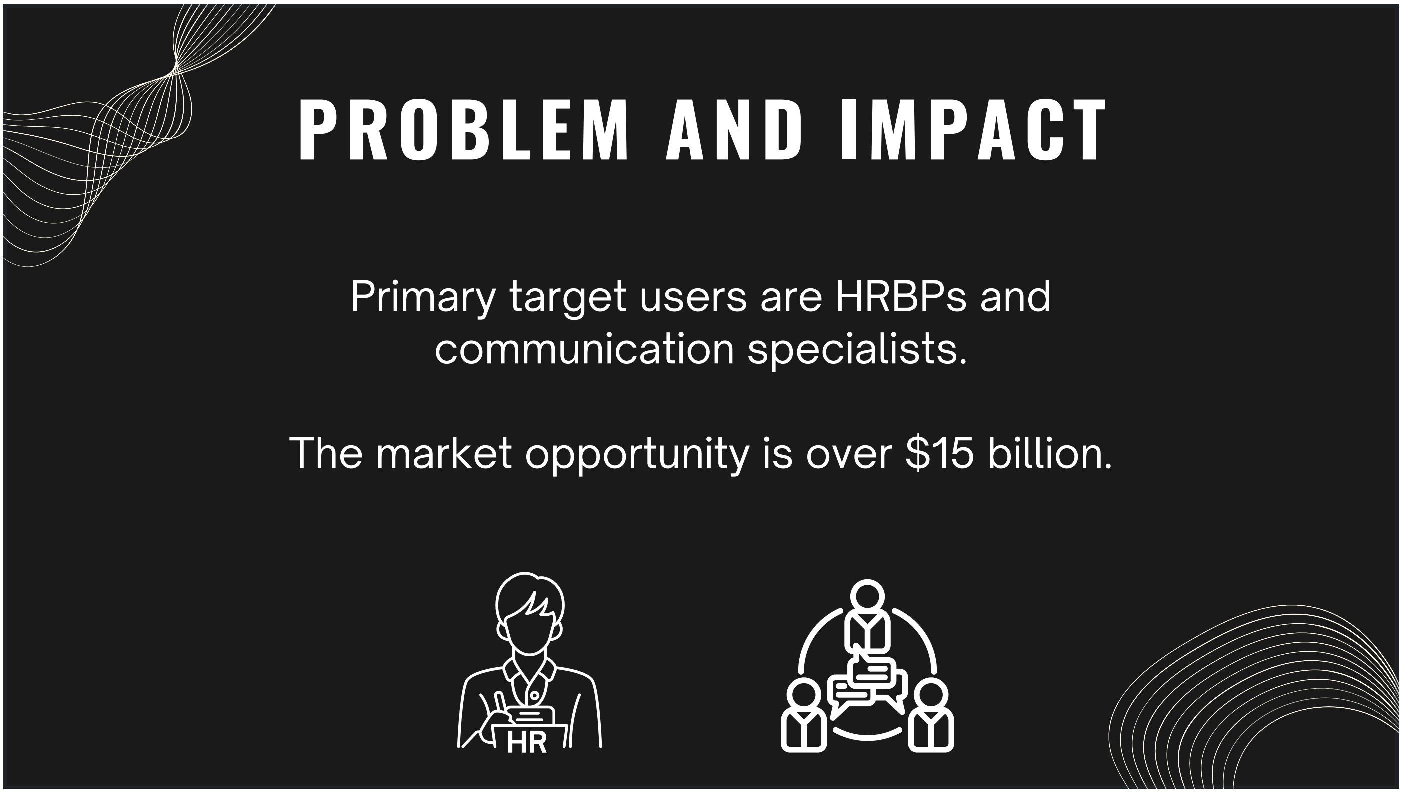


PROBLEM AND IMPACT

BIASCHECK

a tool to neutralize biases in written communication

Our product aims to bridge the gap between current solutions and the nuance of these subjective biases to help writers become more objective in their communications.

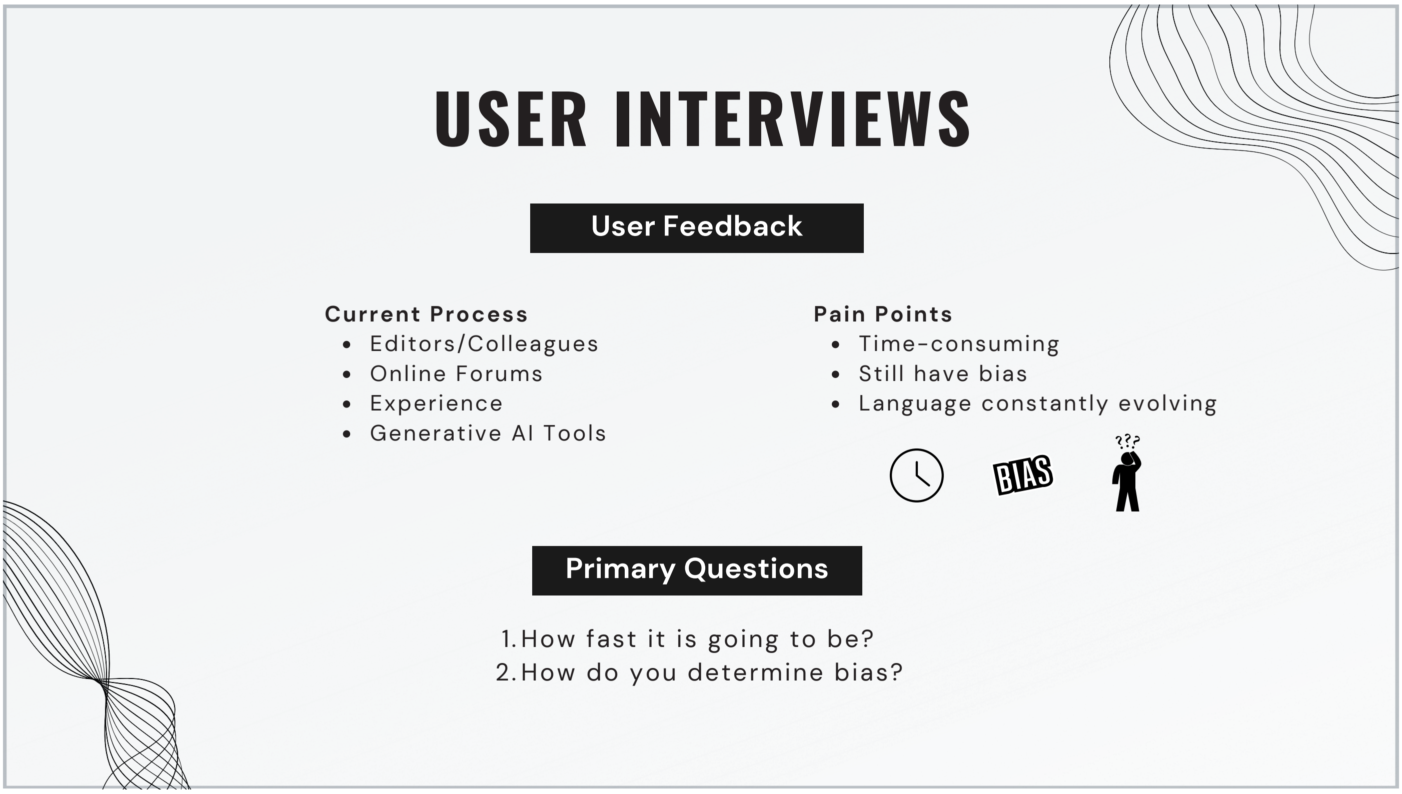


PROBLEM AND IMPACT

Primary target users are HRBPs and communication specialists.

The market opportunity is over $15 billion.

Our primary target users are Human Resources Business Partners and communication specialists seeking a bias mitigation tool. The market opportunity is in this sector is vast as current existing solutions are centered primarily around less nuanced biases, and there is a lack of user-facing products.



USER INTERVIEWS

User Feedback

Current Process

- Editors/Colleagues
- Online Forums
- Experience
- Generative AI Tools

Pain Points

- Time-consuming
- Still have bias
- Language constantly evolving

Primary Questions

1. How fast it is going to be?
2. How do you determine bias?

We did multiple rounds of interviews and interviewing with domain experts and potential users indicated to us the importance of consistent, neutrally presented communications in their fields.

Current processes to check for biases are time consuming and often leave text with some remaining biases. As a result, the primary questions resulting from our interviews are:
1. How fast is it going to be?
2. How do you determine bias?

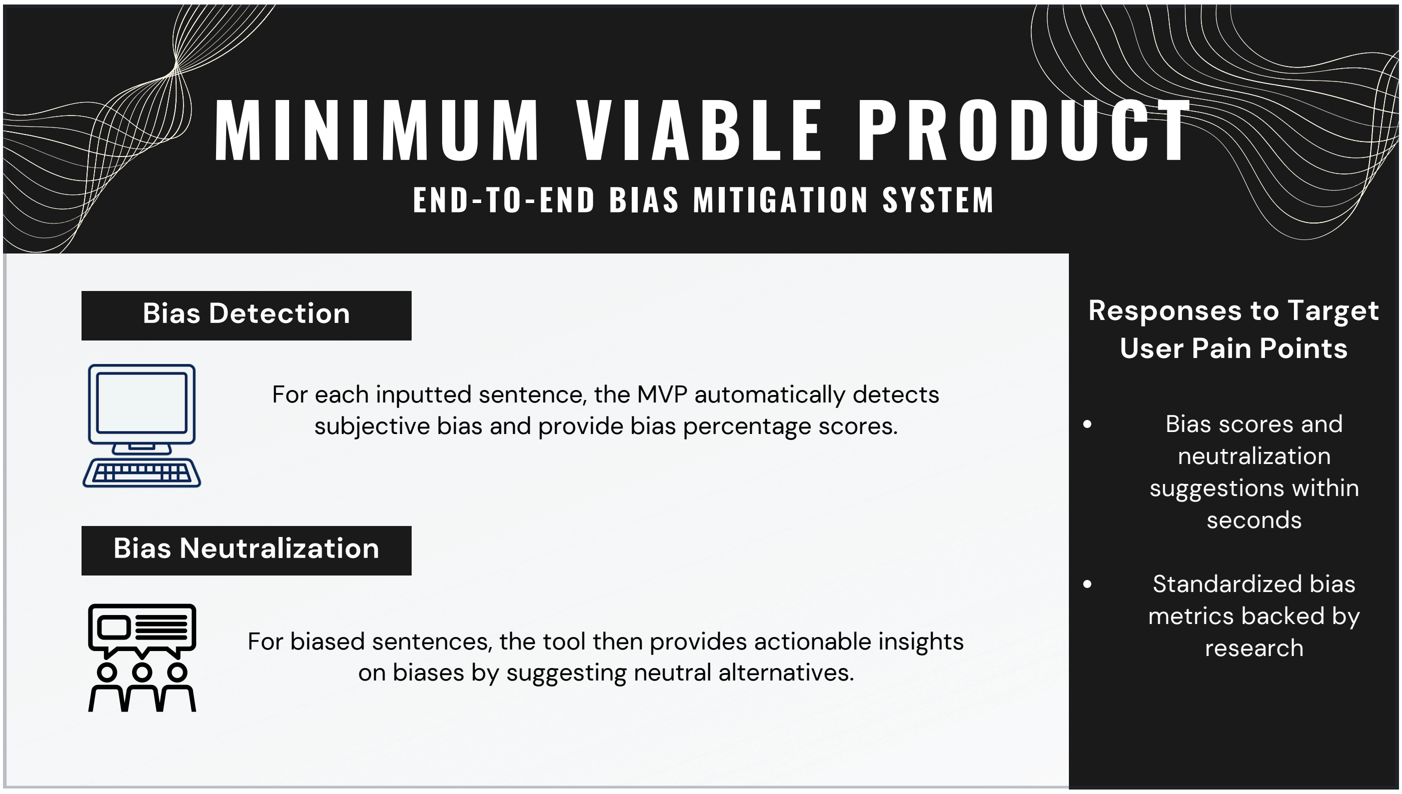
I want to go through some highlights of the interviews:

-Users pointed out that it is extremely important to have consistent, neutral point of view in written communications in their fields.

-Human resources business partners currently rely on editors/colleagues, online resources such as HR forums, experience, and generative AI tools to mitigate bias.

-Based on all user interviews, pain points from current process is time-consuming, still leaving them with some bias and hard to keep up with changes in language.

-Lastly, some users indicate they would use our tool multiple times a week, others indicate they would use our tool multiple times a day



MINIMUM VIABLE PRODUCT

END-TO-END BIAS MITIGATION SYSTEM

Bias Detection

For each inputted sentence, the MVP automatically detects subjective bias and provide bias percentage scores.

Bias Neutralization

For biased sentences, the tool then provides actionable insights on biases by suggesting neutral alternatives.

Responses to Target User Pain Points

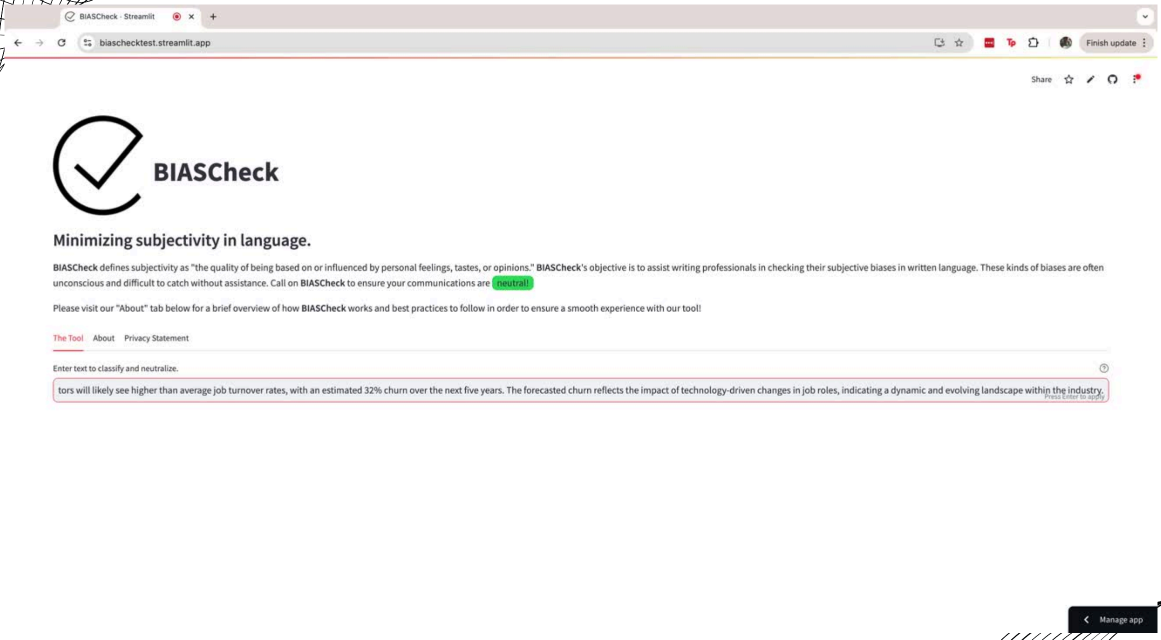
- Bias scores and neutralization suggestions within seconds
- Standardized bias metrics backed by research

The MVP is an end-to-end bias mitigation system designed to analyze written text for subjective, non-neutral biases.

The first aspect of our tool is Bias Detection: Using a classification model, the tool will provide bias scores and highlight sentences with detected bias. The second aspect of our tool, bias neutralization, will suggest neutral alternatives of the inputted text for the user.

Our solution answers to the main pain-points of our users. Instead of their current time consuming process of waiting for colleagues feedback and browsing hr forums, our tool provides our user bias metrics automatically. As well, our users are often relying on other people with their own subjective biases and generative ai tools which are themselves biased. Our tool, however, is backed by standardized measures of subjective bias that are research-based.

DEMO



BIASCheck

Minimizing subjectivity in language.

BIASCheck defines subjectivity as "the quality of being based on or influenced by personal feelings, tastes, or opinions." **BIASCheck**'s objective is to assist writing professionals in checking their subjective biases in written language. These kinds of biases are often unconscious and difficult to catch without assistance. Call on **BIASCheck** to ensure your communications are neutral!

Please visit our "About" tab below for a brief overview of how **BIASCheck** works and best practices to follow in order to ensure a smooth experience with our tool!

The Tool About Privacy Statement

Enter text to classify and neutralize.

tors will likely see higher than average job turnover rates, with an estimated 32% churn over the next five years. The forecasted churn reflects the impact of technology-driven changes in job roles, indicating a dynamic and evolving landscape within the industry.

Press Enter to apply

This is an example of how our app functions. The landing page brings the user directly to the tool. The stationary portion at the top of the webpage provides the user with BIASCheck's purpose along with how we define subjectivity. As the example I input is processing, I'll navigate to our brief privacy statement tab where we explain we collect zero inputted data from the site. We recommend to the user to return to the page periodically to check for updates. On the About tab, we have a brief overview of what BIASCheck is and how the product evaluates the inputted text. We also provide links to our main website and to the dataset we used for training. Returning to our example, the first sentence has a score of less than 50% and therefore has returned as neutral, so no neutralized sentence is outputted. The second sentence has a score greater than 50% and was therefore determined to be biased. The neutralized sentence is displayed below. The last sentence was found to be neutral, so again no neutralized sentence is displayed. Below all of these, the user will see an overall average bias score for the entire input. Finally, we have a short disclaimer at the bottom of the page to cover that BIASCheck explicitly covers subjective biases only. This disclaimer is also featured at the end of our About page.

USER 1 TESTING EXAMPLES

User: Christina
HR business partner auditing statements made by employees in her client group.

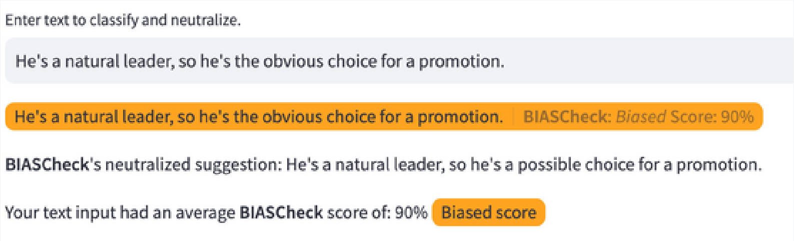
Feedback

Positive:

- neutralized text offered clear differentiation from original input text
- ease of navigation of the interface

Improvements:

- wants it to be integrated as a chrome plugin



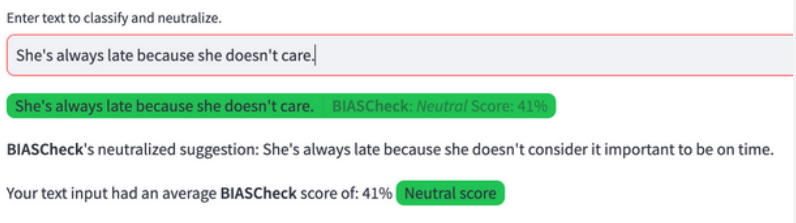
Enter text to classify and neutralize.

He's a natural leader, so he's the obvious choice for a promotion.

He's a natural leader, so he's the obvious choice for a promotion. | BIASCheck: *Biased* Score: 90%

BIASCheck's neutralized suggestion: He's a natural leader, so he's a possible choice for a promotion.

Your text input had an average **BIASCheck** score of: 90% Biased score



Enter text to classify and neutralize.

She's always late because she doesn't care.

She's always late because she doesn't care. | BIASCheck: *Neutral* Score: 41%

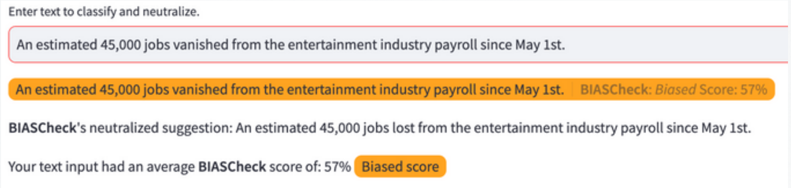
BIASCheck's neutralized suggestion: She's always late because she doesn't consider it important to be on time.

Your text input had an average **BIASCheck** score of: 41% Neutral score

Our first user for testing was an HR business partner who used the tool to audit statements made by employees in her client group. She enjoyed how easy the process was and liked how the neutralized text was clearly different from the original text. She would like to see the product offered as an integrated Chrome plug-in.

USER 2 TESTING EXAMPLES

User: Josed
HR professional creating a market trends deck for a presentation to client-facing business partners.



Enter text to classify and neutralize.

An estimated 45,000 jobs vanished from the entertainment industry payroll since May 1st.

An estimated 45,000 jobs vanished from the entertainment industry payroll since May 1st. BIASCheck: *Biased* Score: 57%

BIASCheck's neutralized suggestion: An estimated 45,000 jobs lost from the entertainment industry payroll since May 1st.

Your text input had an average **BIASCheck** score of: 57% Biased score

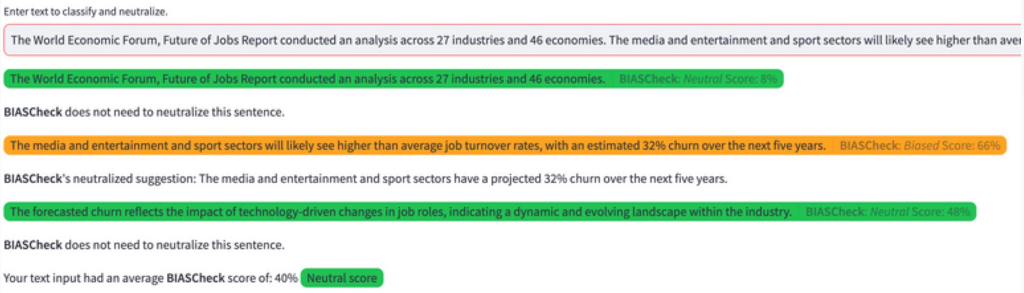
Feedback

Positive:

- efficiency and promptness of reply
- overall score component

Improvements:

- an explanation of why neutral text is "better"



Enter text to classify and neutralize.

The World Economic Forum, Future of Jobs Report conducted an analysis across 27 industries and 46 economies. The media and entertainment and sport sectors will likely see higher than aver

The World Economic Forum, Future of Jobs Report conducted an analysis across 27 industries and 46 economies. BIASCheck: *Neutral* Score: 8%

BIASCheck does not need to neutralize this sentence.

The media and entertainment and sport sectors will likely see higher than average job turnover rates, with an estimated 32% churn over the next five years. BIASCheck: *Biased* Score: 66%

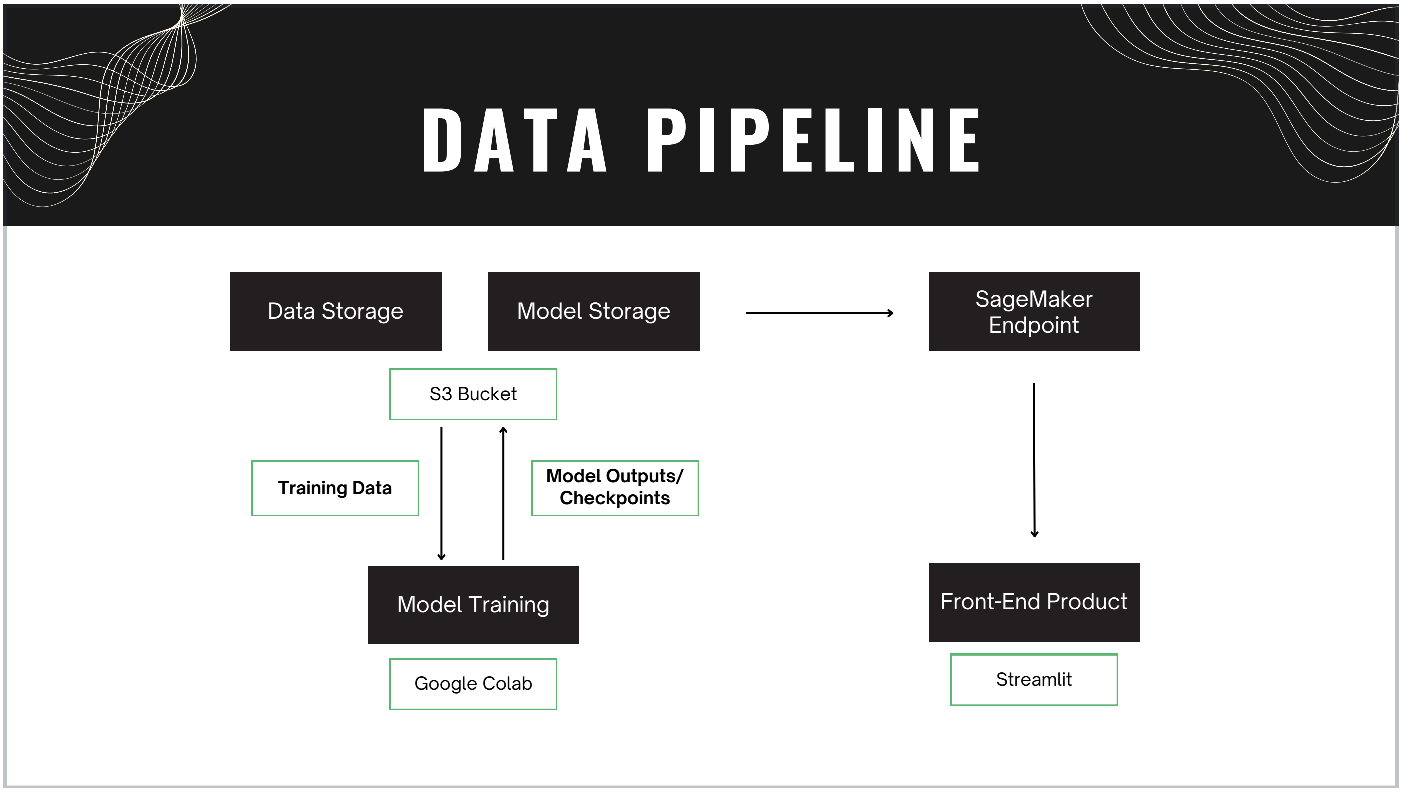
BIASCheck's neutralized suggestion: The media and entertainment and sport sectors have a projected 32% churn over the next five years.

The forecasted churn reflects the impact of technology-driven changes in job roles, indicating a dynamic and evolving landscape within the industry. BIASCheck: *Neutral* Score: 48%

BIASCheck does not need to neutralize this sentence.

Your text input had an average **BIASCheck** score of: 40% Neutral score

Our second user was an HR professional who was prepping a deck to present to business partners. We used the second example on the lower right in our demo. Josed liked the efficiency of the product and the overall score component, but felt like he was missing an explanation for why we considered the neutralized text to be "better".



DATA PIPELINE

Data Storage

Model Storage

SageMaker Endpoint

S3 Bucket

Training Data

Model Outputs/ Checkpoints

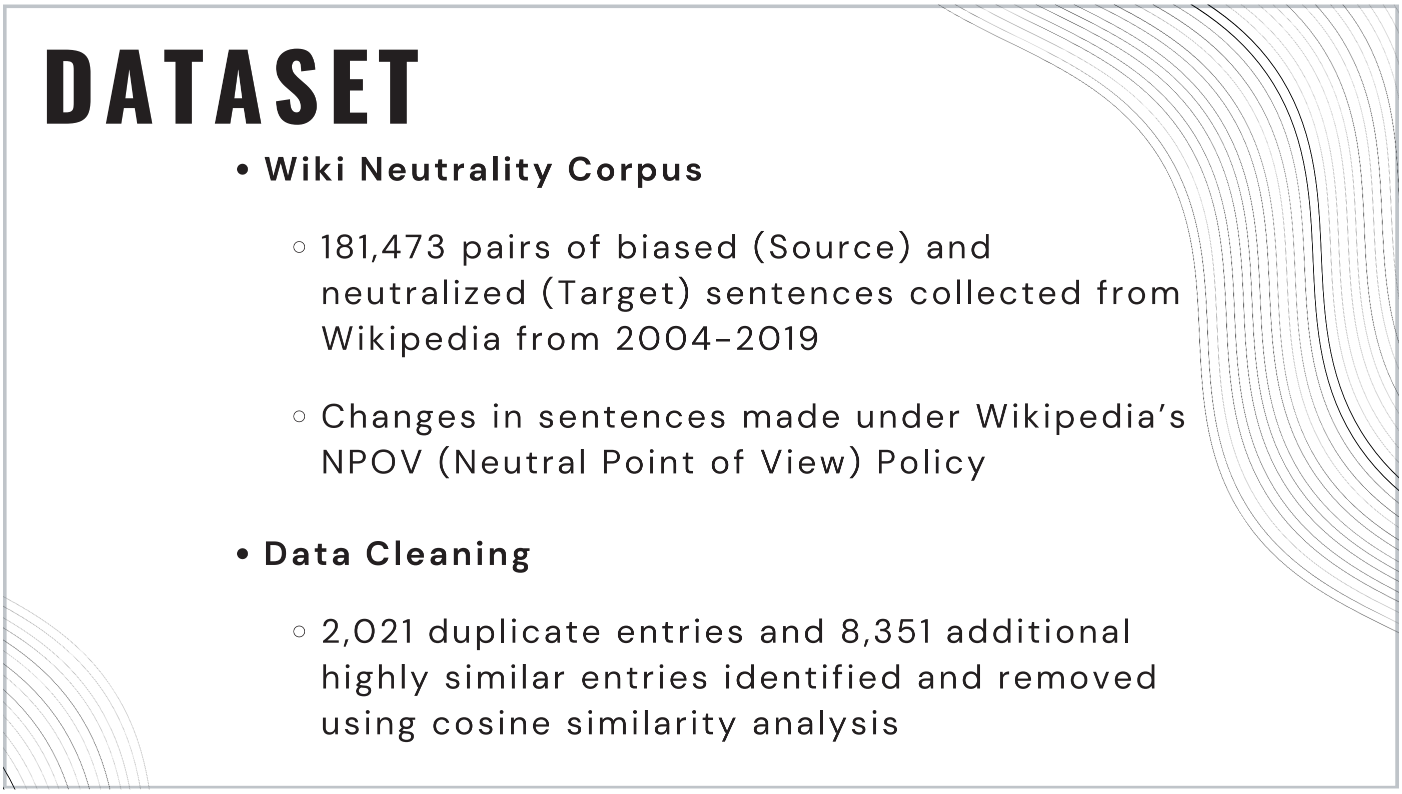
Model Training

Google Colab

Front-End Product

Streamlit

Our data pipeline involved training our models in Google Colab, and then migrating them over to AWS for deployment in Streamlit. Pictured is our idealized version of our pipeline. In BIASCheck's current deployment state, we were able to successfully create a real-time interaction for our users for the classification task. Our neutralization task is still being run locally, but we would hope to work out the kinks in that pipeline as part of our next steps. I'll now hand it off to Etienne who will introduce our dataset and bias detection task.

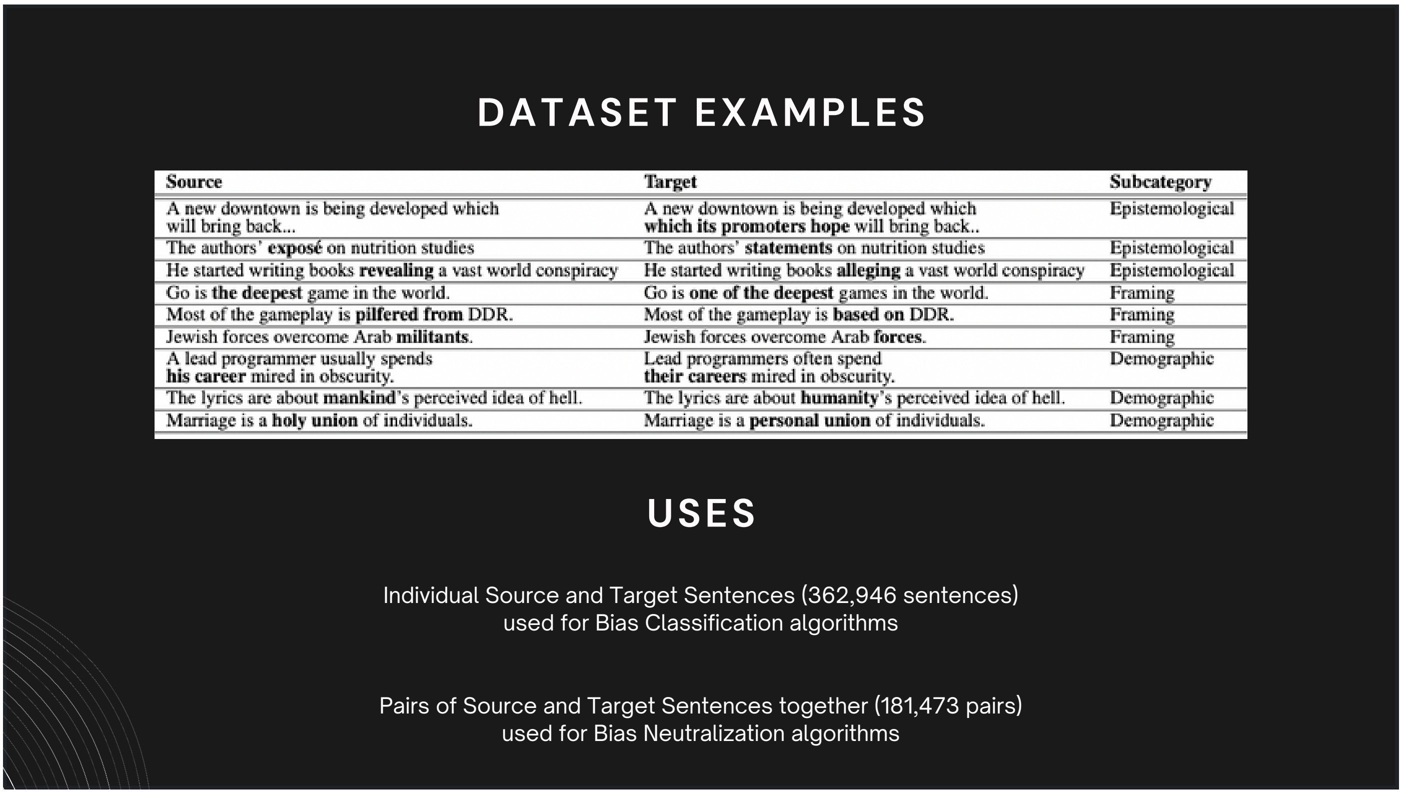


DATASET

- **Wiki Neutrality Corpus**
  - 181,473 pairs of biased (Source) and neutralized (Target) sentences collected from Wikipedia from 2004-2019
  - Changes in sentences made under Wikipedia's NPOV (Neutral Point of View) Policy
- **Data Cleaning**
  - 2,021 duplicate entries and 8,351 additional highly similar entries identified and removed using cosine similarity analysis

Our primary dataset is the Wiki Neutrality Corpus (WNC in short). This dataset is a collection of changes in sentences made under Wikipedia's neutral point of view policy and contains over 180,000 data points consisting of pairs of sentences.

While the data was relatively clean, in our EDA we found numerous duplicates and highly similar entries. In this context we used cosine similarity to remove records that were similar based on proximity.



DATASET EXAMPLES

Source	Target	Subcategory
A new downtown is being developed which will bring back...	A new downtown is being developed which **which its promoters hope** will bring back..	Epistemological
The authors' **exposé** on nutrition studies	The authors' **statements** on nutrition studies	Epistemological
He started writing books **revealing** a vast world conspiracy	He started writing books **alleging** a vast world conspiracy	Epistemological
Go is **the deepest** game in the world.	Go is **one of the deepest** games in the world.	Framing
Most of the gameplay is **pilfered from** DDR.	Most of the gameplay is **based on** DDR.	Framing
Jewish forces overcome Arab **militants**.	Jewish forces overcome Arab **forces**.	Framing
A lead programmer usually spends **his career** mired in obscurity.	Lead programmers often spend **their careers** mired in obscurity.	Demographic
The lyrics are about **mankind**'s perceived idea of hell.	The lyrics are about **humanity**'s perceived idea of hell.	Demographic
Marriage is a **holy union** of individuals.	Marriage is a **personal union** of individuals.	Demographic

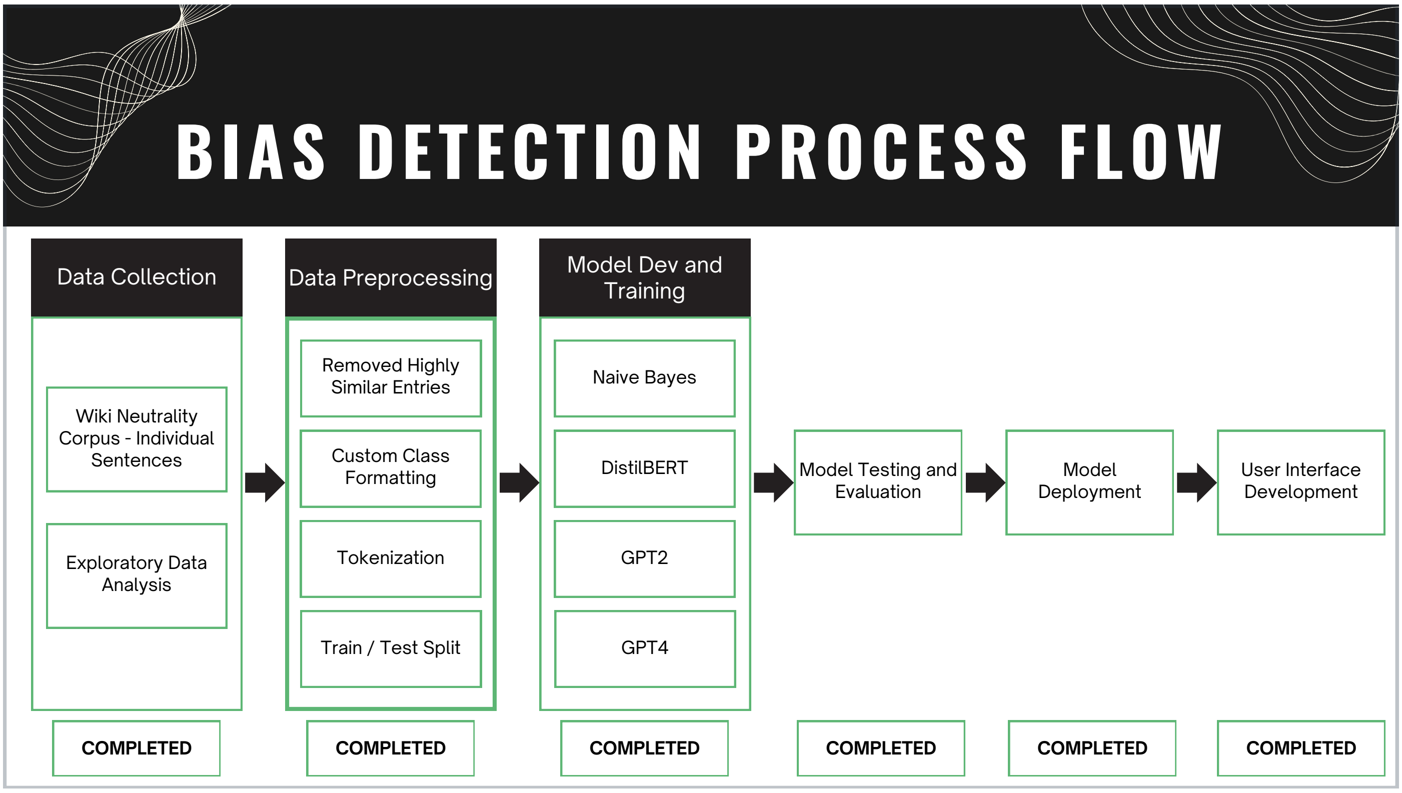
USES

Individual Source and Target Sentences (362,946 sentences)
used for Bias Classification algorithms

Pairs of Source and Target Sentences together (181,473 pairs)
used for Bias Neutralization algorithms

Here you can see some examples of what our dataset includes. There are examples of framing bias as well as demographic biases such as gendered and religious language.

We labeled the Source and Target sentences as biased and neutral, respectively, for the binary classification task. For the neutralization task we left the Source/Target pairs together.



BIAS DETECTION PROCESS FLOW

Data Collection

- Wiki Neutrality Corpus - Individual Sentences
- Exploratory Data Analysis

COMPLETED

Data Preprocessing

- Removed Highly Similar Entries
- Custom Class Formatting
- Tokenization
- Train / Test Split

COMPLETED

Model Dev and Training

- Naive Bayes
- DistilBERT
- GPT2
- GPT4

COMPLETED

Model Testing and Evaluation

COMPLETED

Model Deployment

COMPLETED

User Interface Development

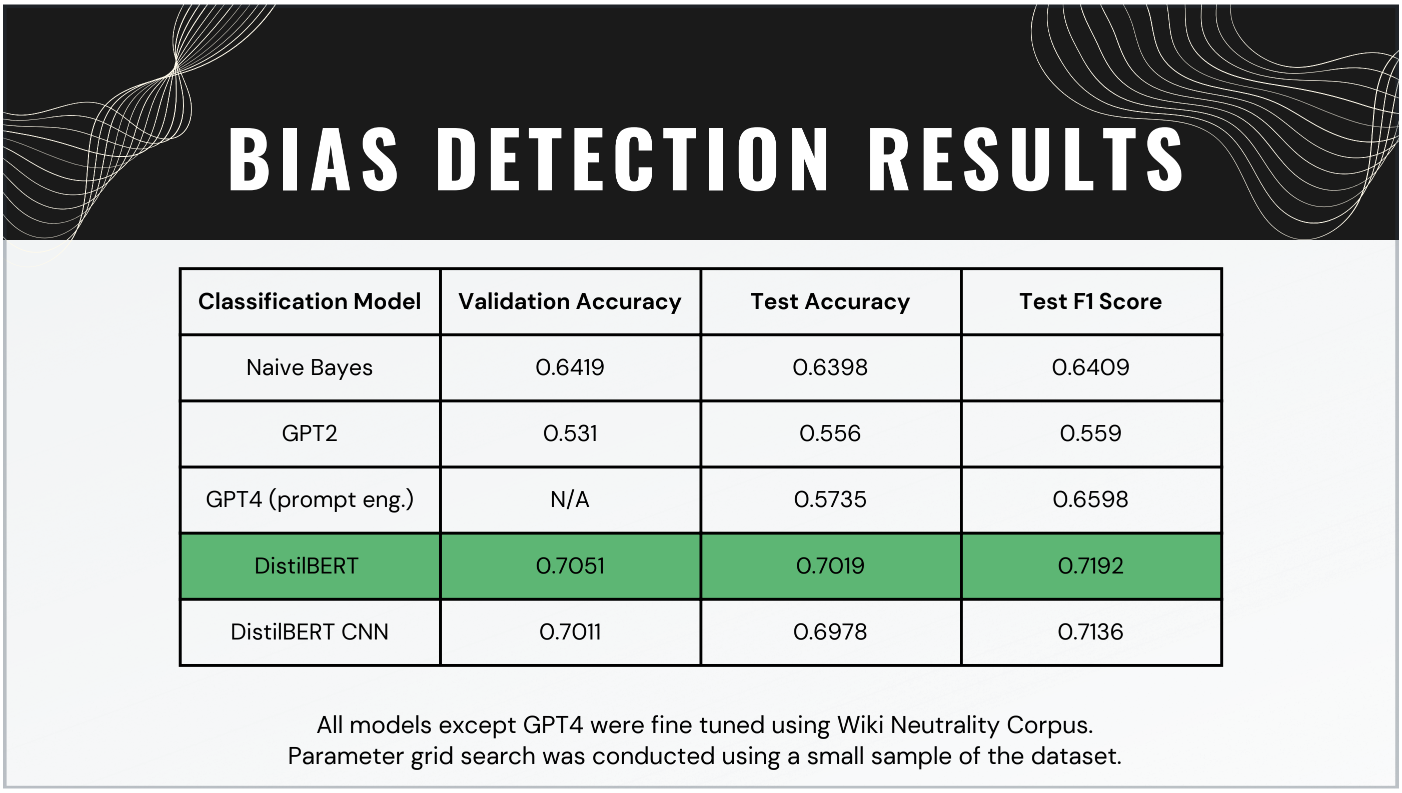
COMPLETED

1) For Data collection we
retrieved data from Wiki Neutrality Corpus data from Kaggle, and in our EDA we were fortunate to find there was no significant label imbalance since raw bias text counts were equal to the target counts.

2) On preprocessing we re-formatted our classes for modeling purposes, and also eliminated low quality text material, such as data envelop information.

For data engineering, we considered various tokenizers such as Spacy for appropriate tokenization and removed piping information.

3) For model development we developed Naive Bayes and GPT2 models (not the generative version) as baseline for binary classification or bias detection, thereby also introducing transformers.

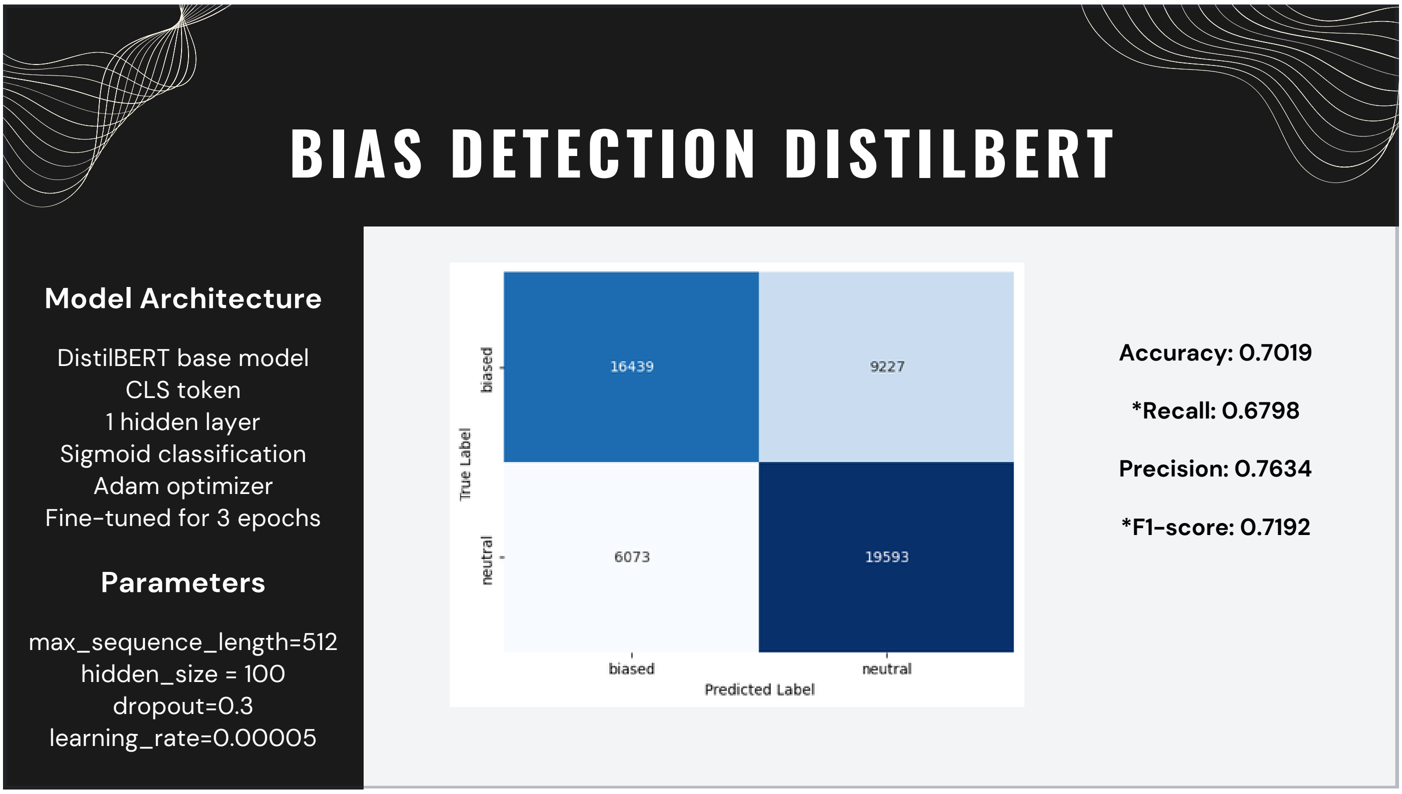


BIAS DETECTION RESULTS

Classification Model	Validation Accuracy	Test Accuracy	Test F1 Score
Naive Bayes	0.6419	0.6398	0.6409
GPT2	0.531	0.556	0.559
GPT4 (prompt eng.)	N/A	0.5735	0.6598
DistilBERT	0.7051	0.7019	0.7192
DistilBERT CNN	0.7011	0.6978	0.7136

All models except GPT4 were fine tuned using Wiki Neutrality Corpus.
Parameter grid search was conducted using a small sample of the dataset.

2) Modelling used Validation and Test Accuracy to qualify DistilBert as the best model for detection with an F1 score of around 72% for test and validation.



BIAS DETECTION DISTILBERT

Model Architecture

DistilBERT base model
CLS token
1 hidden layer
Sigmoid classification
Adam optimizer
Fine-tuned for 3 epochs

Parameters

max_sequence_length=512
hidden_size = 100
dropout=0.3
learning_rate=0.00005

True Label \ Predicted Label	biased	neutral
biased	16439	9227
neutral	6073	19593

Accuracy: 0.7019

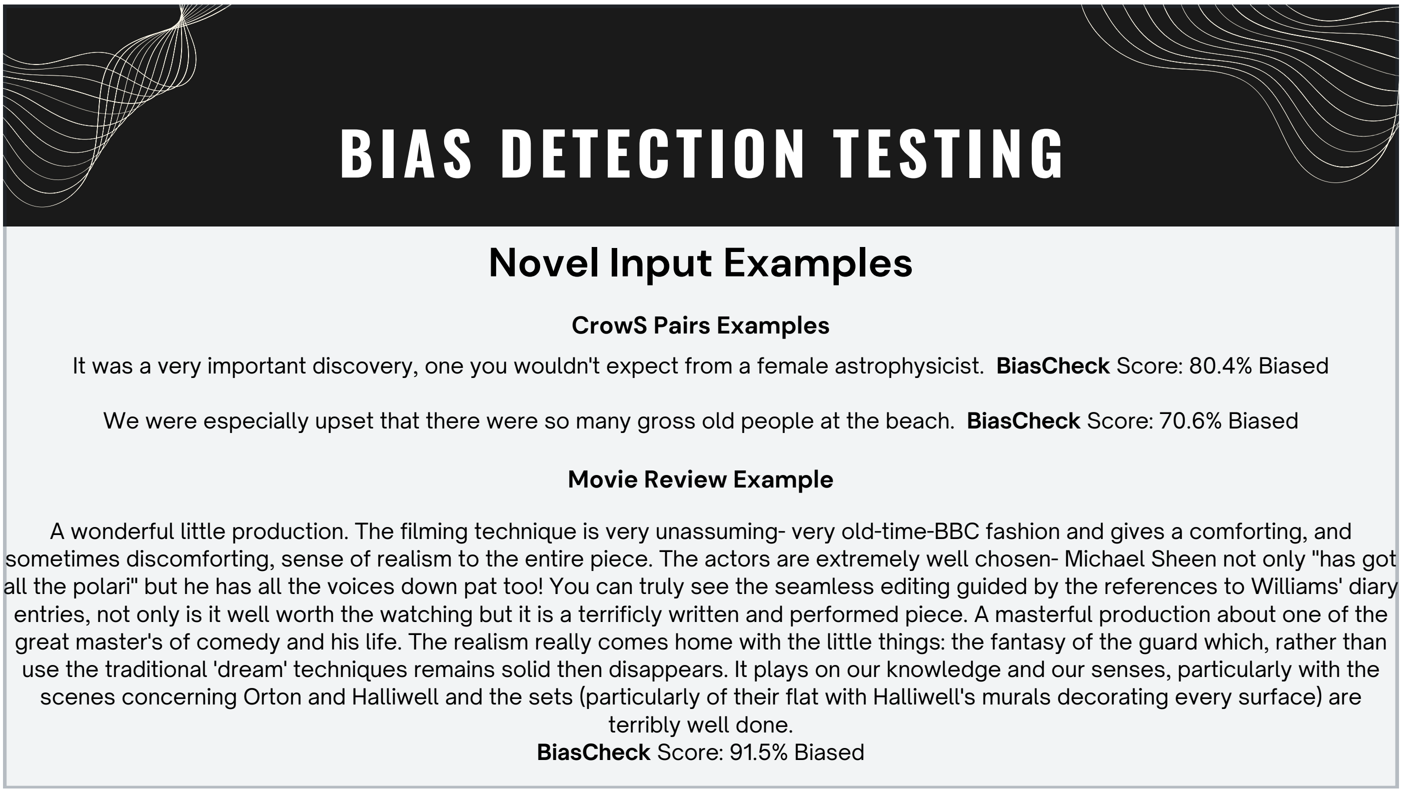
***Recall: 0.6798**

Precision: 0.7634

***F1-score: 0.7192**

We took a closer look under the hood at our best classification model, as this is the model we will be using in our product. the simple DistilBERT model actually slightly outperformed the distilbert cnn. it is CLS-Token-based with one hidden layer, a max seq length of 512 to be able to capture even very long sentences, and was finetuned for 3 epochs.

We cared most about recall and f1 score as we wanted to ensure our models catch biased text. We did try optimizing f1-score through changing the threshold, but that ultimately ended up lowering our recall Score so we did not use it.
(Now turning it over to my teammate Violet to talk about testing)



BIAS DETECTION TESTING

Novel Input Examples

CrowS Pairs Examples

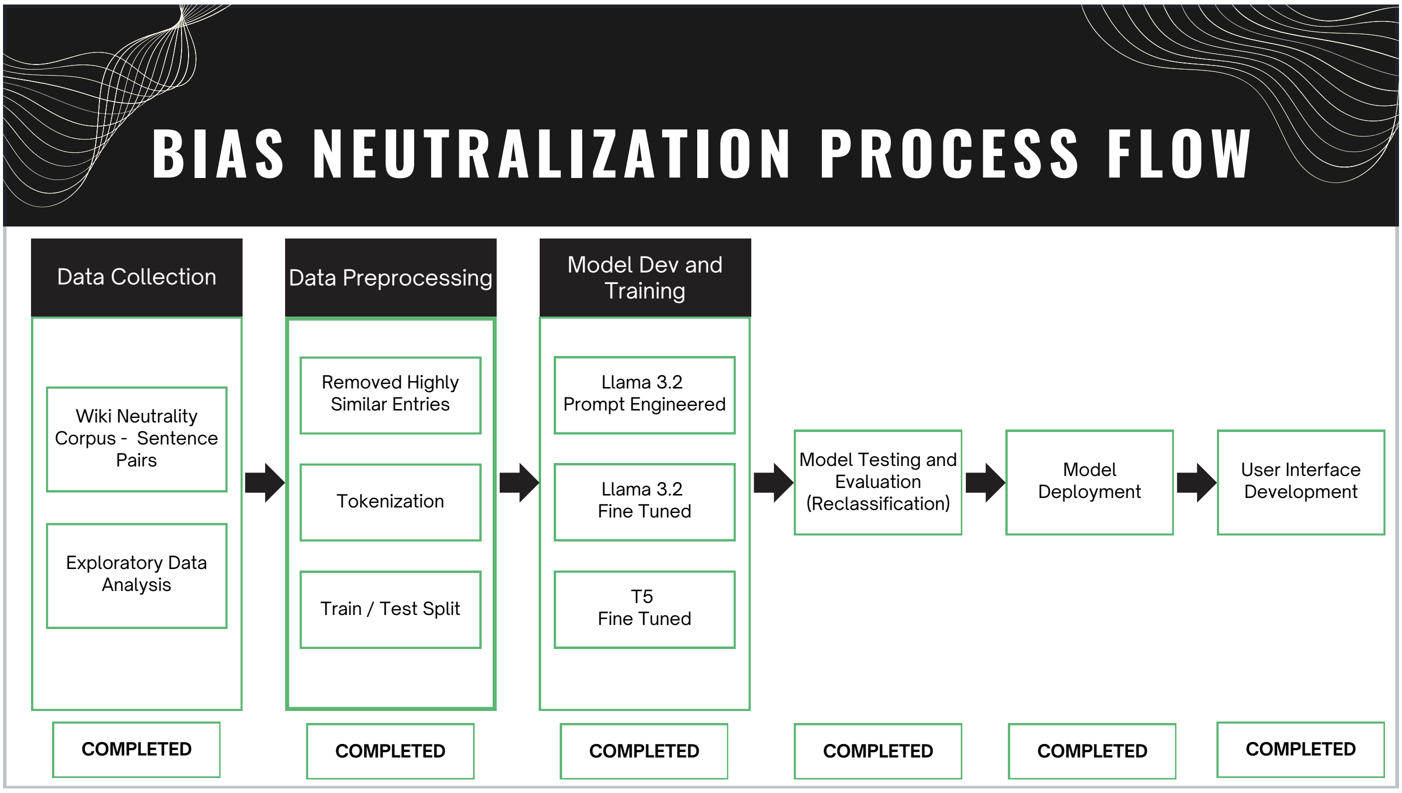
It was a very important discovery, one you wouldn't expect from a female astrophysicist. **BiasCheck** Score: 80.4% Biased

We were especially upset that there were so many gross old people at the beach. **BiasCheck** Score: 70.6% Biased

Movie Review Example

A wonderful little production. The filming technique is very unassuming- very old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of realism to the entire piece. The actors are extremely well chosen- Michael Sheen not only "has got all the polari" but he has all the voices down pat too! You can truly see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the watching but it is a terrificly written and performed piece. A masterful production about one of the great master's of comedy and his life. The realism really comes home with the little things: the fantasy of the guard which, rather than use the traditional 'dream' techniques remains solid then disappears. It plays on our knowledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of their flat with Halliwell's murals decorating every surface) are terribly well done.
BiasCheck Score: 91.5% Biased

Beyond the testing we conducted on our dataset, we also wanted to feed our model novel input examples to see how it responded. We tested on examples from the CrowS Pairs data set which contains stereotyped sentences and those received highly biased scores. We also passed through subjective movie reviews which also gave biased scores.



BIAS NEUTRALIZATION PROCESS FLOW

Data Collection

Wiki Neutrality Corpus - Sentence Pairs

Exploratory Data Analysis

COMPLETED

Data Preprocessing

Removed Highly Similar Entries

Tokenization

Train / Test Split

COMPLETED

Model Dev and Training

Llama 3.2 Prompt Engineered

Llama 3.2 Fine Tuned

T5 Fine Tuned

COMPLETED

Model Testing and Evaluation (Reclassification)

COMPLETED

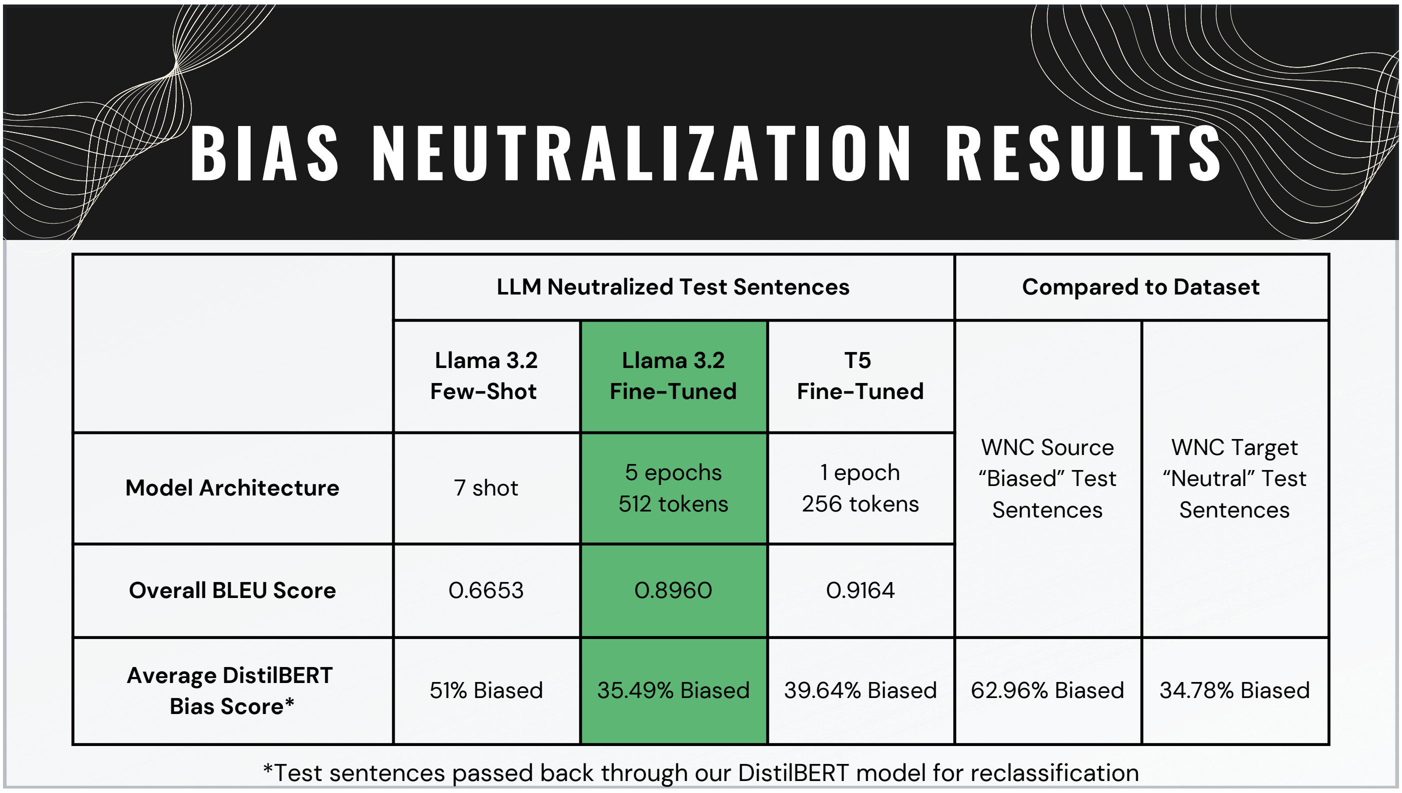
Model Deployment

COMPLETED

User Interface Development

COMPLETED

For our bias neutralization task, we have a similar process flow to bias detection. We developed three models, a llama prompt engineered and a llama fine tuned as well as a t5 fine tuned model. We evaluated these models through testing and reclassifying.

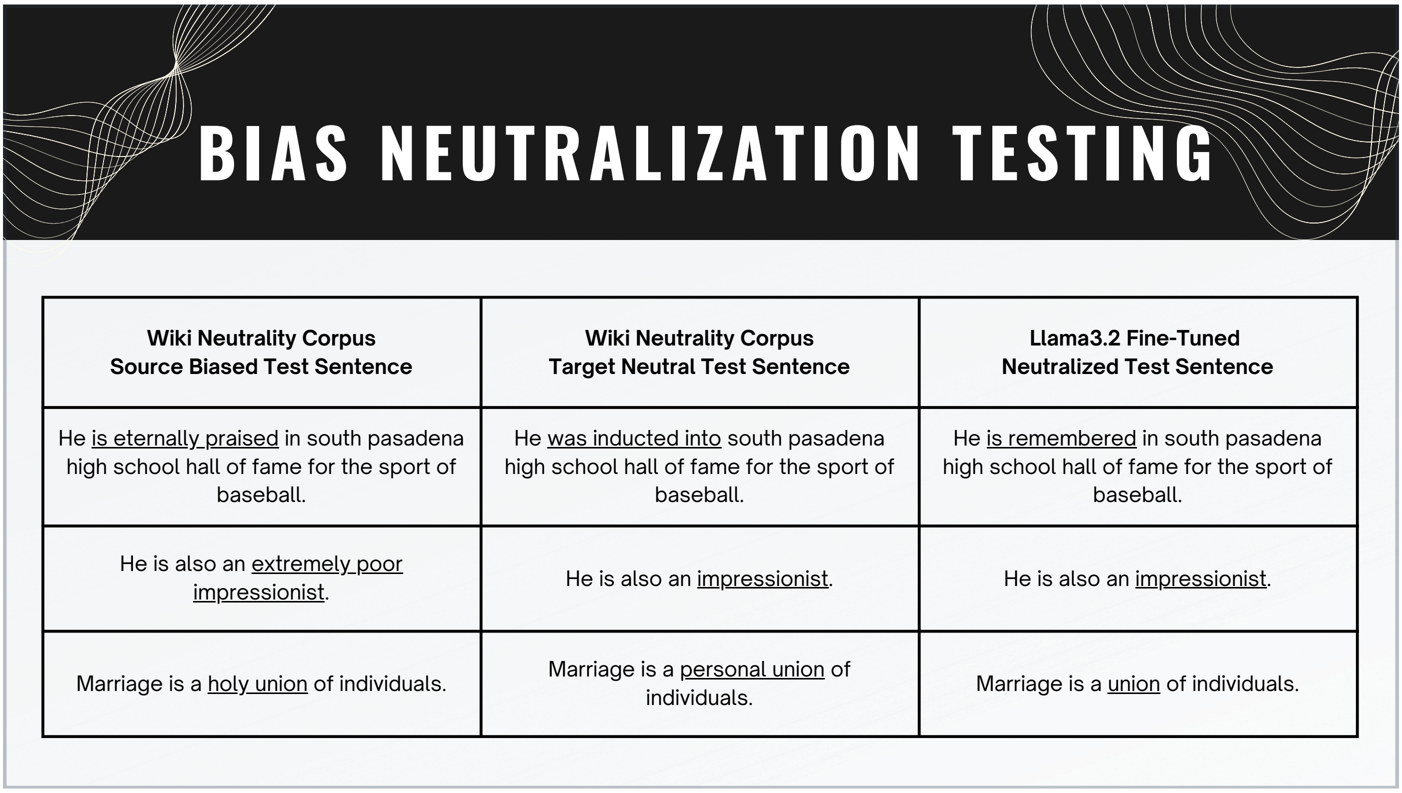


BIAS NEUTRALIZATION RESULTS

	LLM Neutralized Test Sentences			Compared to Dataset	
	Llama 3.2 Few-Shot	Llama 3.2 Fine-Tuned	T5 Fine-Tuned	WNC Source “Biased” Test Sentences	WNC Target “Neutral” Test Sentences
Model Architecture	7 shot	5 epochs 512 tokens	1 epoch 256 tokens		
Overall BLEU Score	0.6653	0.8960	0.9164		
Average DistilBERT Bias Score*	51% Biased	35.49% Biased	39.64% Biased	62.96% Biased	34.78% Biased

*Test sentences passed back through our DistilBERT model for reclassification

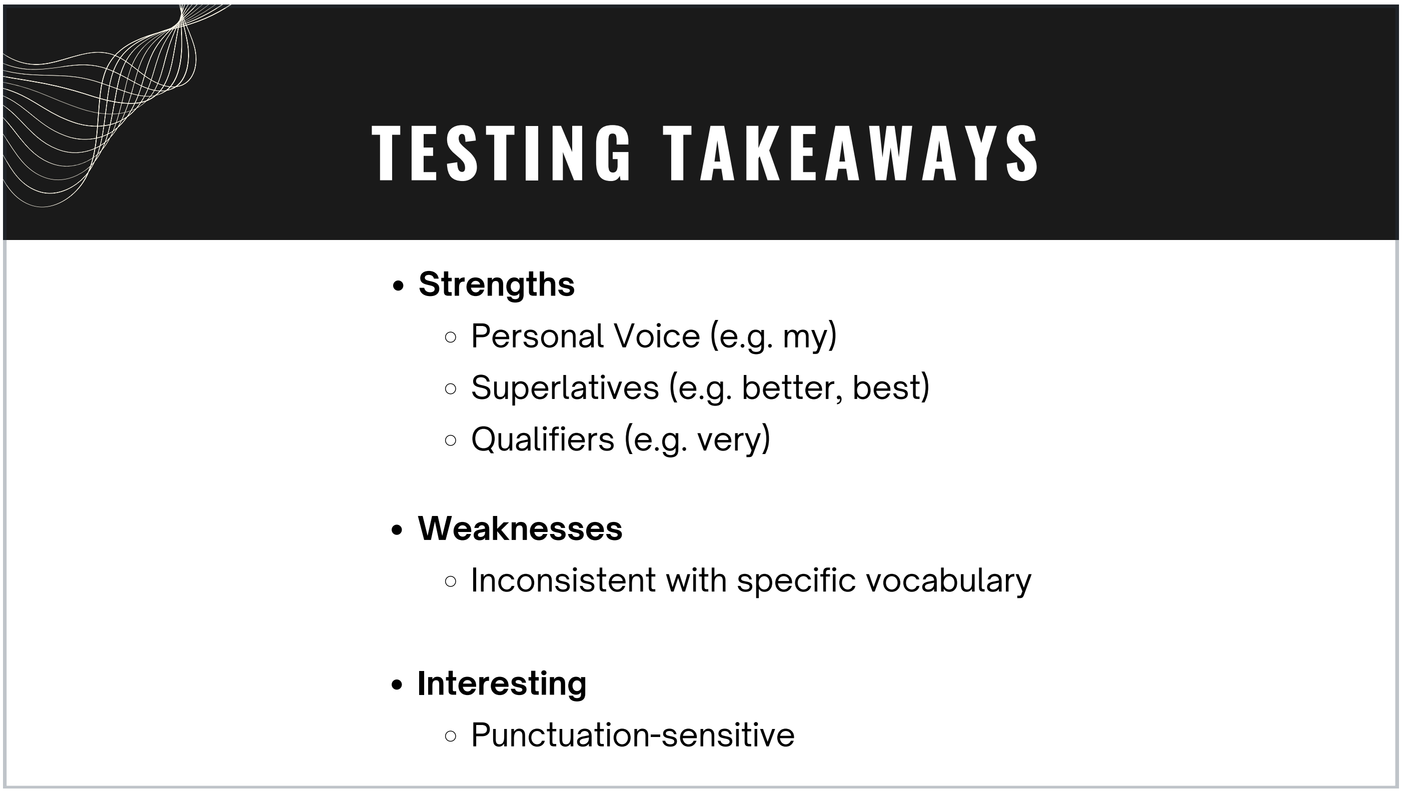
To test our neutralization models, we ran our test dataset through each of our neutralization models and compared those outputs to the target neutral sentences in the Wiki Neutrality dataset. We measured overall bleu scores and average distilbert bias scores, which were calculated by passing back the neutralized sentences through our distilbert classifier. We also measured the average distilbert score of our datasets source and target sentences for a baseline. Our fine tuned llama model performed the best, with an overall bleu score of 0.896 and an average bias score of 35.49%, which is very close to our target baseline of 34.78%.



BIAS NEUTRALIZATION TESTING

Wiki Neutrality Corpus Source Biased Test Sentence	Wiki Neutrality Corpus Target Neutral Test Sentence	Llama3.2 Fine-Tuned Neutralized Test Sentence
He is eternally praised in south pasadena high school hall of fame for the sport of baseball.	He was inducted into south pasadena high school hall of fame for the sport of baseball.	He is remembered in south pasadena high school hall of fame for the sport of baseball.
He is also an extremely poor impressionist.	He is also an impressionist.	He is also an impressionist.
Marriage is a holy union of individuals.	Marriage is a personal union of individuals.	Marriage is a union of individuals.

We conducted additional analysis of our best neutralization model by examining our neutralized outputs and comparing them to our target sentences. We see that even though our model doesn't always exactly match the target sentence, it still neutralizes the bias. for example, with the second entry our model matches the target exactly. In the third example, the target changes holy union to personal union, and our model just drops holy to say marriage is a union.



TESTING TAKEAWAYS

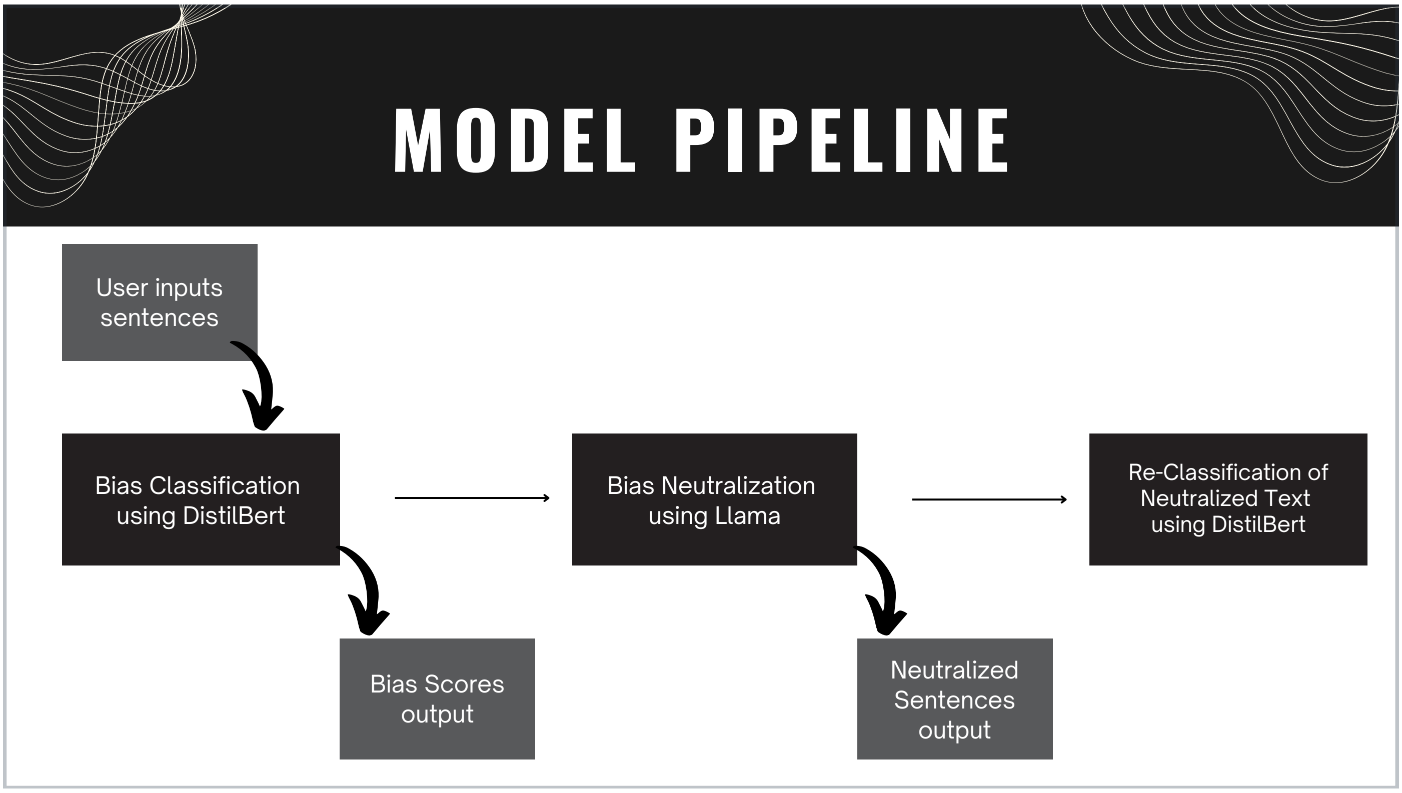
- **Strengths**
  - Personal Voice (e.g. my)
  - Superlatives (e.g. better, best)
  - Qualifiers (e.g. very)
- **Weaknesses**
  - Inconsistent with specific vocabulary
- **Interesting**
  - Punctuation-sensitive

We wanted to highlight some overall takeaways discovered from user and edge case testing.

our model is great at detecting and neutralizing subjectivity when it is framed in the personal voice or through superlatives or qualifiers

however we did notice some inconsistency. for example, when sentences are just over the bias threshold they are sometimes modified and sometimes left as is, especially if the bias is tied to a specific vocab word and not a superlative or qualifier. that is an area for refinement, where we could potentially add a secondary dataset

an interesting note is that our tool is punctuation sensitive. when sentences end with an exclamation point, they are much more likely to be classified as biased



MODEL PIPELINE

User inputs sentences

Bias Classification using DistilBert

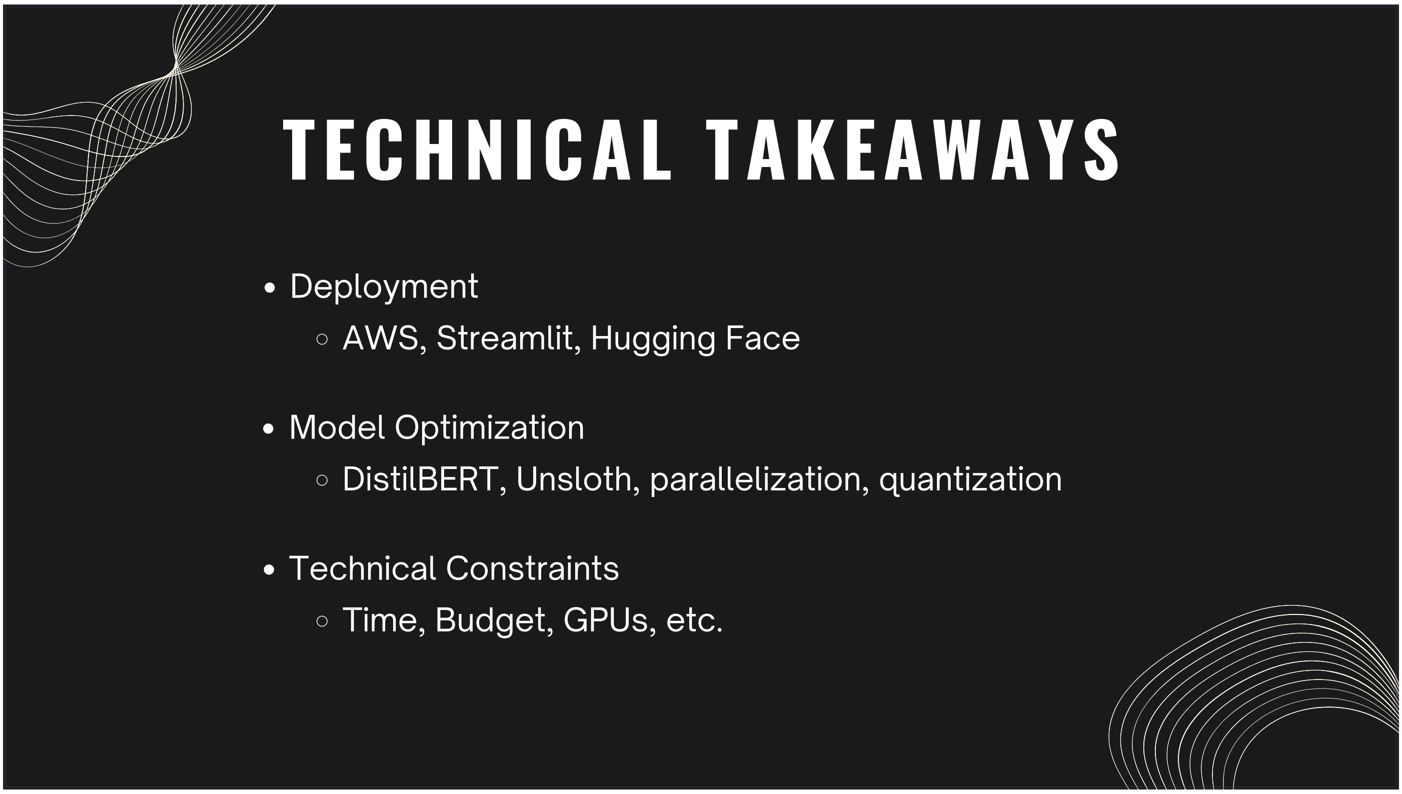
Bias Scores output

Bias Neutralization using Llama

Neutralized Sentences output

Re-Classification of Neutralized Text using DistilBert

Our overall model process flow is as follows: The user inputs text and our DistilBert model classifies it as biased or neutral. The bias scores are output to the user. If the text is classified as biased, the text is pushed through our neutralization model and neutralized sentences are output. To internally evaluate the validity of our neutralization task, we reclassify the neutralized text through our Distilbert model.



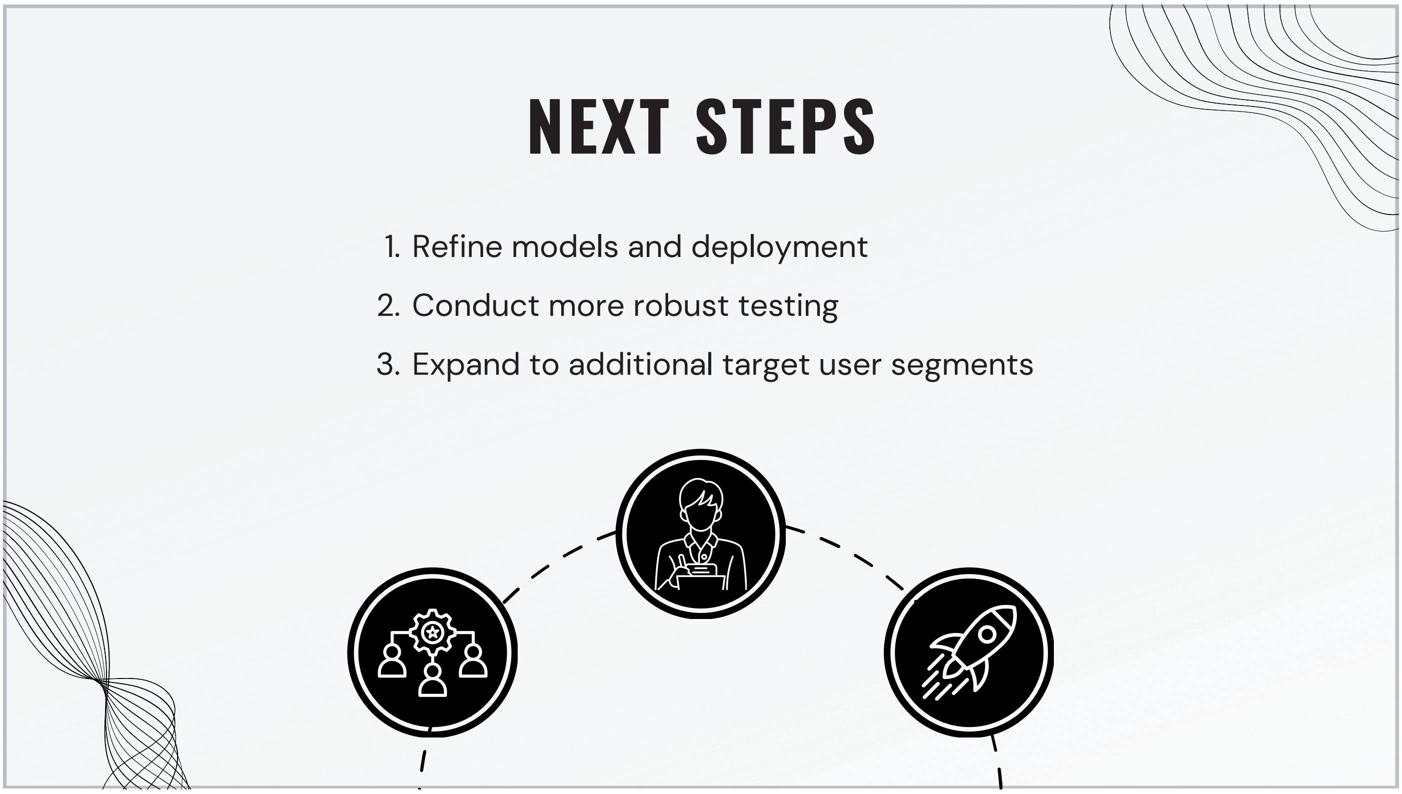
TECHNICAL TAKEAWAYS

- Deployment
  - AWS, Streamlit, Hugging Face
- Model Optimization
  - DistilBERT, Unsloth, parallelization, quantization
- Technical Constraints
  - Time, Budget, GPUs, etc.

We have key technical takeaways:

In terms of deployment, we gained significant experience in developing a deployment system using AWS, Streamlit, and Hugging Face. This was an area where our team had little prior experience, so while there are still some bumps that we would like to smooth out in the future, we gained ample data engineering experience.

We also explored many model optimization techniques including using light weight distilled transformer models, unsloth which is an open-source efficiency tool which we implemented when fine tuning our llms, parallelization, quantization. these techniques helped optimize our training process but we still ran into technical constraints with time, budget, and gpus.



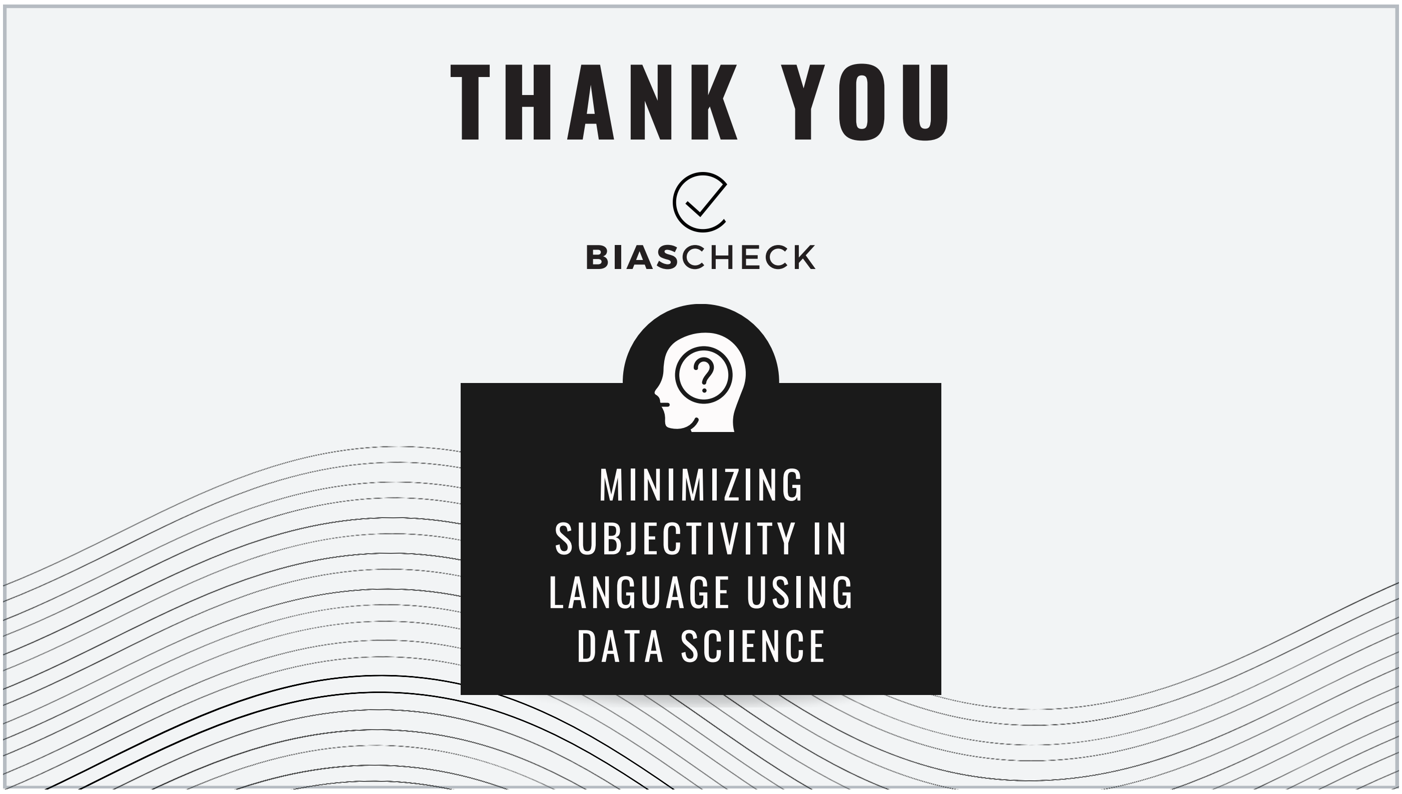
NEXT STEPS

1. Refine models and deployment
2. Conduct more robust testing
3. Expand to additional target user segments

Our next steps for further development are to refine our models and optimize our deployment

As well, we would like to conduct more comprehensive and rigorous testing with additional edge case testing and more formal qualitative user testing sessions

Finally, as a future state, we would expand to broader user segments including advertising, journalism, and politics as ultimately, we want to adapt our tool to meet the needs of anyone who wants to ensure objectivity and neutrality in written text.



THANK YOU

BIASCHECK

MINIMIZING SUBJECTIVITY IN LANGUAGE USING DATA SCIENCE

In summation, with our outlined technical approach of combining a bias classification and bias neutralization, we aim to deliver on our BIAScheck promise: to minimize subjectivity in language using data science. Thank you! We will now take any questions.