Project mission

To fully compare generative AI responses to human responses by using NLP techniques on sentiment, emotional consistency, and linguistic complexity. This will help businesses better understand and use generative AI for more complex, caring interactions with customers and employees.

01. **Bridging Human and AI Communication Gaps:** Our primary goal is to bridge understanding between human-driven answers and those generated by AI, highlighting strengths and nuances

02. **Benchmarking AI Consistency and Reliability:** By deeply analyzing emotional and lexical aspects, the project emphasizes measuring consistency and accuracy across AI responses

03. **Driving Data-Driven Emotion and Sentiment Insights:** Using sentiment and emotion metrics to provide insights into how AI tools resonate emotionally compared to human feedback

Response Generation by LLMs ChatBots

We used five large language models (LLMs)—ChatGPT, Google Gemini, Microsoft Copilot, Claude, and Perplexity—to generate responses to questions asked of corporate specialists. These AI-generated responses are then analyzed to compare emotional depth, sentiment, lexical complexity, and other factors against responses from human professionals.

ChatGPT

Developed by OpenAI, ChatGPT is based on the GPT (Generative Pre-trained Transformer) architecture.

Gemini

Developed by Google as part of Google's AI expansion focusing on language understanding and multimodal capabilities.

Copilot

Built on Microsoft's Azure OpenAI Service, Copilot uses GPT-based models to assist users in Microsoft products like Word, Excel, and Teams.

Claude

Developed by Anthropic, emphasizing on reducing harmful outputs and is fine-tuned to provide respectful, balanced responses

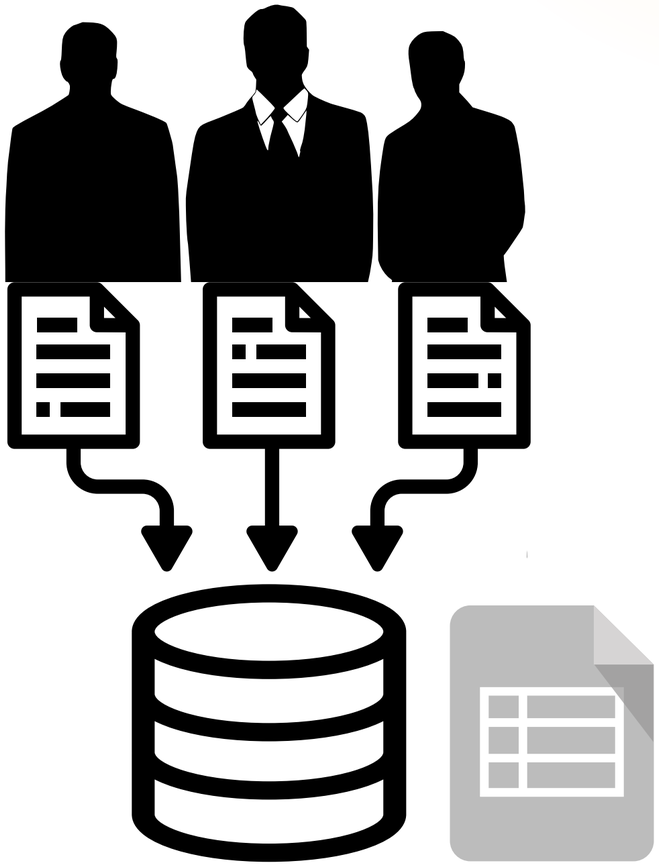
Perplexity

This model is designed with a focus on providing information-rich responses, particularly for search-based tasks.

DATA COLLECTION | EXTRACTION

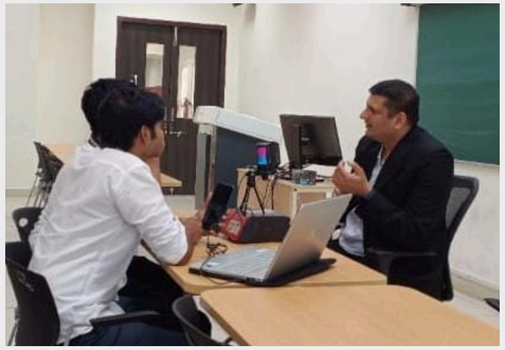
1. First we tried to connect through **LinkedIn**
2. Collect data in different **cooperative visits** in our college, requesting every other cooperate that visits the IIT campus for different festivals, i.e., Sandstone, Tedx, Industrial Visit, etc.
3. Took the **video interviews** of the **industry experts and their openings on some problem-questions.**
4. **Data** is being extracted through **manual and automatic audio-to-text conversion.** Then stacked in **excel file as:**

```
[Serial No. as: QuestionHuman Answer] OpenAI Chatgpt AnswerGoogle Gemini AnswerMicrosoft Copilot AnswerClaude AnswerPerplexity Answer]
```



GLIMPS | INTERVIEWS

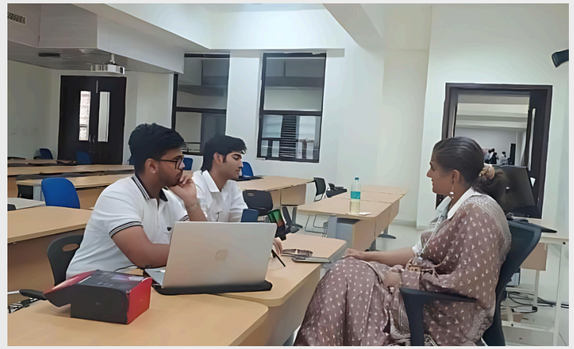
CTO Chat360
Shivam Verma



Vikrant Kulkarni
Director ,Bank of New York

Rahul Barve
Executive VP , ZEE NEWS

Shantamona Bharadwaj
VP Talent Aquisition , DBS



Amrita Singh
CDO,Tata Motors

Aditya Agarwal
Director(Marketing) , LeadSquare

Pankaj Mishra
Director - Autodesk

Natural Language Processing (NLP)

A branch of artificial intelligence (AI) that uses machine learning to help computers understand, interpret, and manipulate human language.

A branch of NLP that looks at how computers can understand what they read. To understand text, NLU uses methods such as sentence analysis, categorizing, logical and grammatical analysis, and separating meanings from words. Through analyzing grammar and context, NLU can help machines figure out what a sentence means.

NLU powers applications that require context-sensitive responses, such as personal assistants, sentiment analysis, and customer service automation. For Example : **LLM chatBots like ChatGPT,Claude**

Preprocessing of Text Responses

```
example_text = "Visit https://example.com for more details! It's really great."
```

CLEANING

Cleaning the text by removing URLs, non-alphanumeric characters, and unnecessary white spaces.

'Visit for more details Its really great'

TOKENIZATION

Breaking the text into words using spaCy and NLTK. We'll use NLTK for tokenization since it's simple for basic tasks.

['Visit', 'for', 'more', 'details', 'Its', 'really', 'great']

NORMALIZATION

Converting all tokens to lowercase for consistency.

['visit', 'for', 'more', 'details', 'its', 'really', 'great']

STOPWORD REMOVAL

Removing common words (stopwords) that do not add value to our analysis.

['visit', 'details', 'really', 'great']

LEMMATIZATION

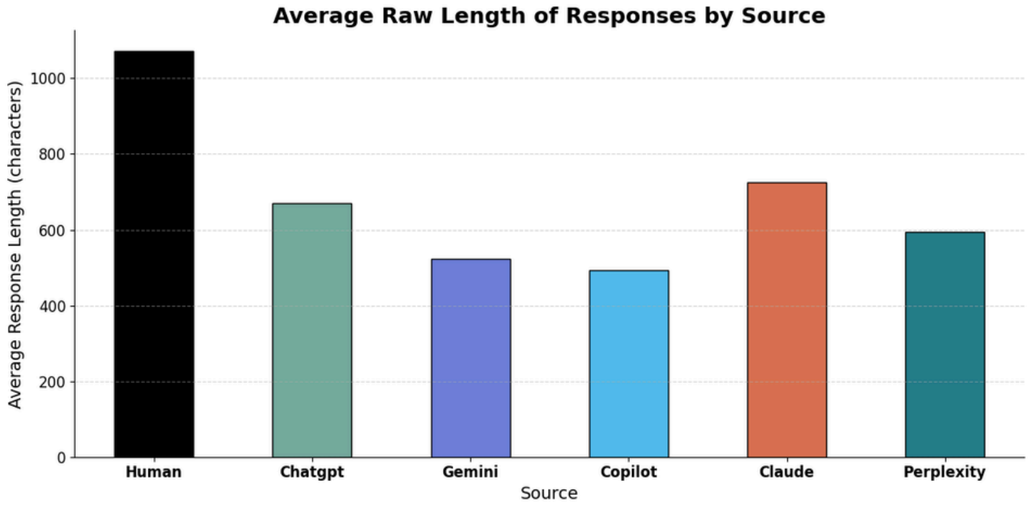
['visit', 'detail', 'really', 'great']

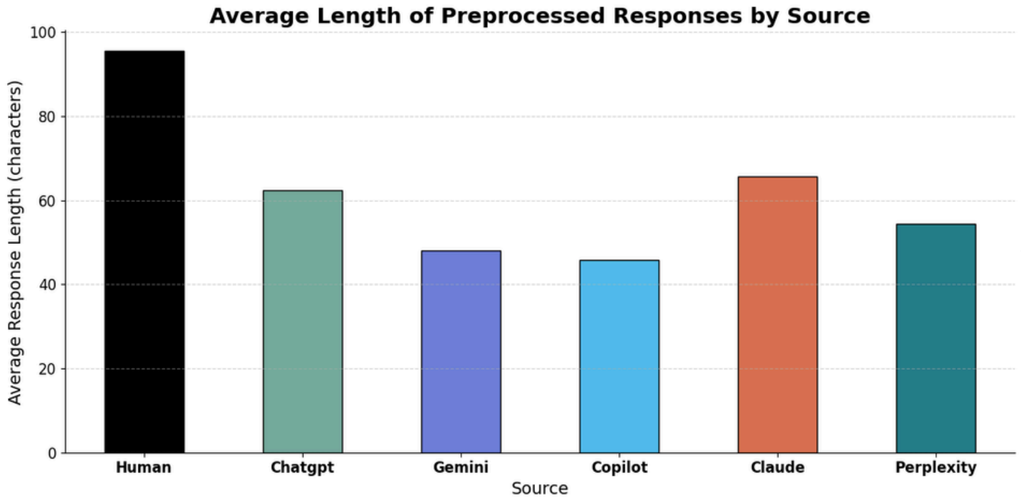
Reducing the words to their base form using spaCy's lemmatizer.

Libraries

NLTK (Natural Language Toolkit) for text processing, spaCy for tokenization and lemmatization, used for performing various Natural Language Processing tasks.

BASIC DESCRIPTIVE STATISTICS





BOX PLOT - RAW DATA

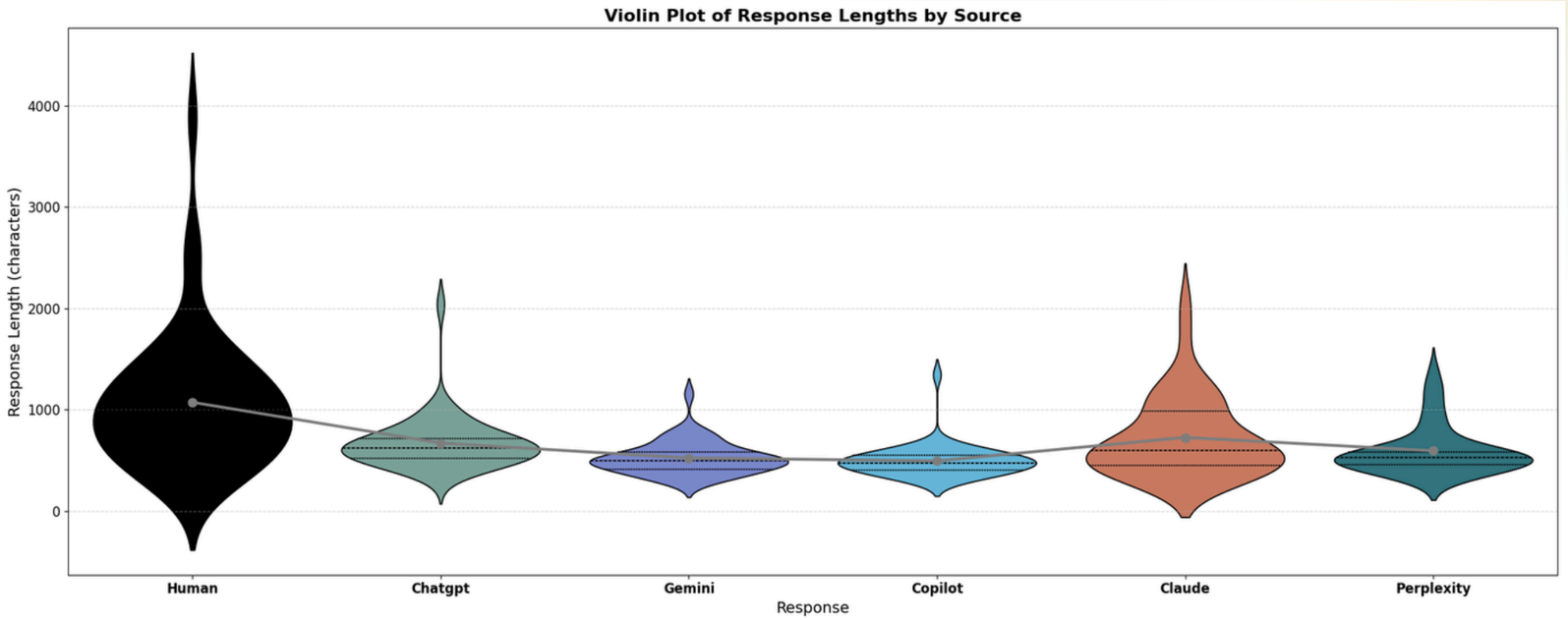
- **Human answers are the longest** on average (1072.8 characters), indicating a more detailed response style.
- **Claude-generated** responses are the **longest among the AI models. Copilot**-generated responses have the **shortest raw length**, indicating **more concise responses** compared to other AI models.

BOX PLOT - PREPROCESSED

- **Human** responses also reduce significantly in length , maintaining a larger length compared to AI responses, which imply **a higher level of context or descriptive language.**
- **AI responses** generally become more **condensed after post-processing**

AI models tend to give short answers, but human answers naturally have more information, even after a lot of text reduction

BASIC DESCRIPTIVE STATISTICS



VIOLIN PLOT

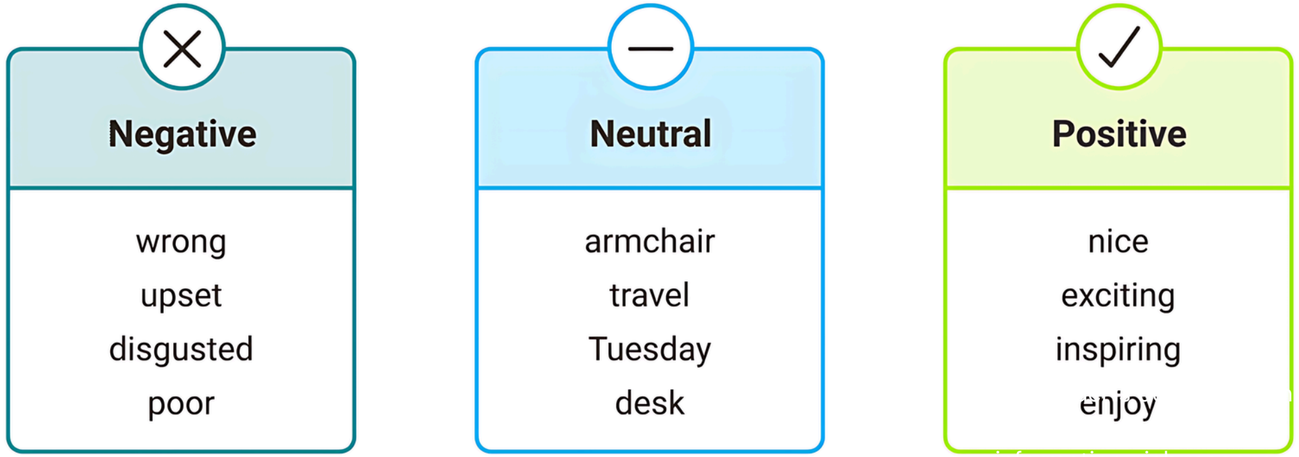
- **Human answers** have a significantly **larger distribution of response lengths**, with a **wider spread** and a notably **higher median length** compared to AI models
- Among the AI models, **ChatGPT** has a **somewhat wider range** of response lengths, **other AI models** (Gemini, Copilot, Claude, and Perplexity) show **smaller and more consistent distributions**, suggesting they **generate more concise** responses
- **Mean points** highlight that human responses are generally longest on average, with ChatGPT being the closest among the AI models .
- The **other AI models** — show a **clustering of shorter average** response lengths. This clustering reflect a more standardized response pattern among these tools, potentially due to optimization forefficiency in answering.

SENTIMENT ANALYSIS IDEATION

It is the process of classifying whether a block of text is positive, negative, or neutral. The goal that Sentiment mining tries to gain is to be analysed people's/any respose in a way that can help businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.) - called Emotional Analysis

IDEATION

Our aim was to compare the sentiment of AI-generated answers with human responses, ideally to see how closely these AI answers match the sentiment expressed by real human experts.



SENTIMENT ANALYSIS PROCESS

Used 2 General and 1 Corporate Specific Model

VADER

Valence Aware Dictionary and Sentiment Reasoner using lib. SentimentIntensityAnalyzer

TextBlob

VADER is generally more tailored for social media-like text, **TextBlob** uses a simpler approach for polarity and subjectivity analysis.

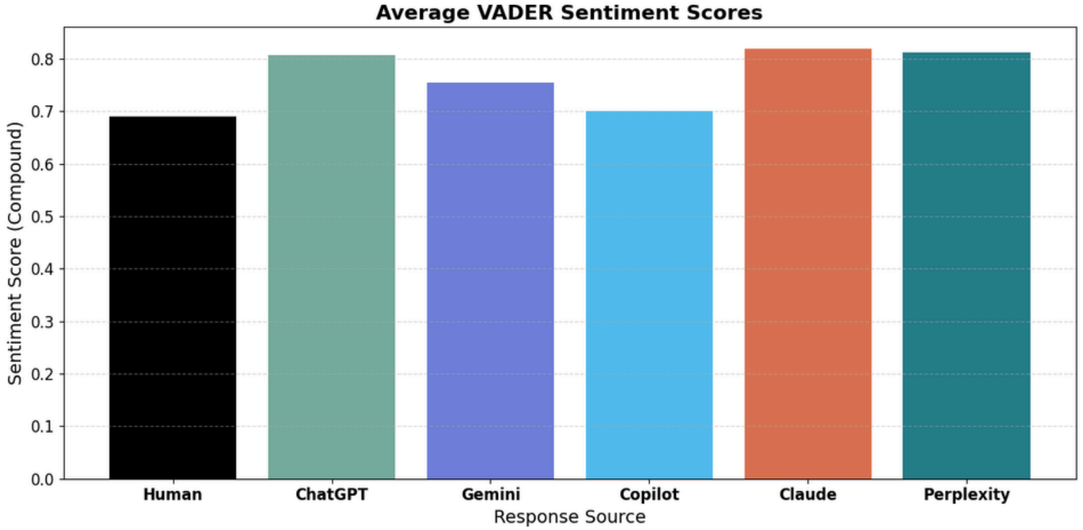
FinBERT

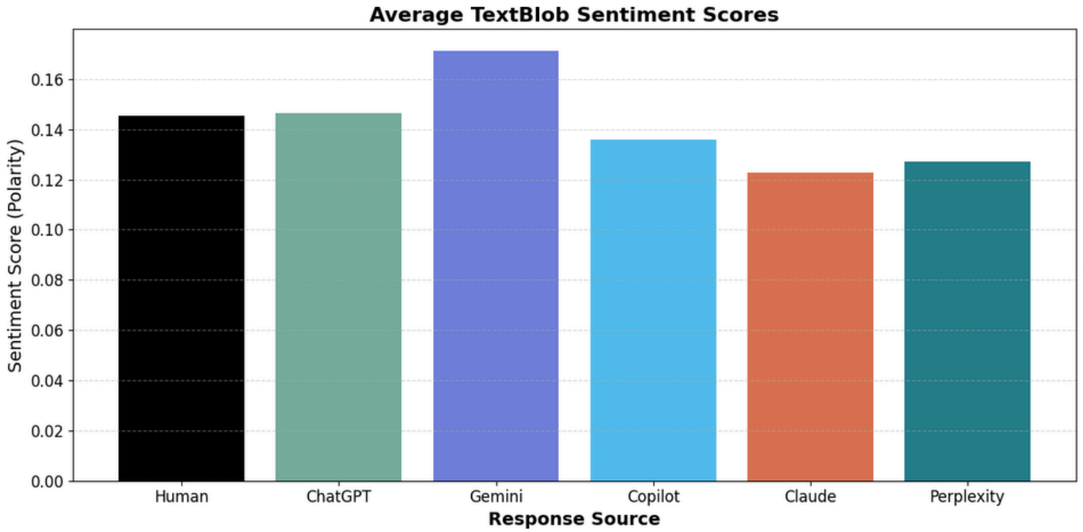
Business-Specific Sentiment: FinBERT is pre-trained on financial news and reports, making it particularly suited for analyzing corporate, financial related text.

Process Flow

- **Data Input**: The cleaned (processed) text (answers or responses) is inputted into the model.
- **Sentiment Detection:** The model processes the text and identifies the underlying sentiments, classifying them into various categories Posisitive , Negative , Neutral .
- **Output**: The results are displayed with the detected sentiment labels and scores, allowing for visualisation

SENTIMENT ANALYSIS RESULTS





BOX PLOT - VADER Sentiment

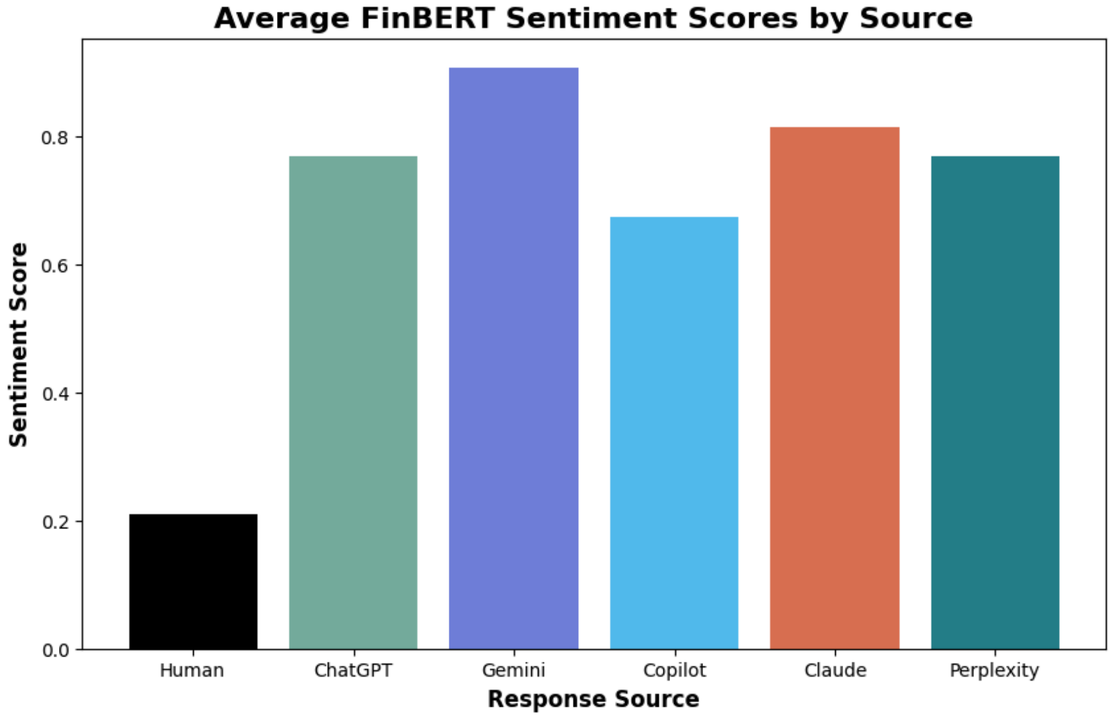
- **Almost all AI**-generated responses, display **higher sentiment scores** than **human** responses.
- **Copilot** have comparatively lower sentiment scores, indicating a more moderate tone
- **Are AI models tend to generate more positive or optimistic content?**

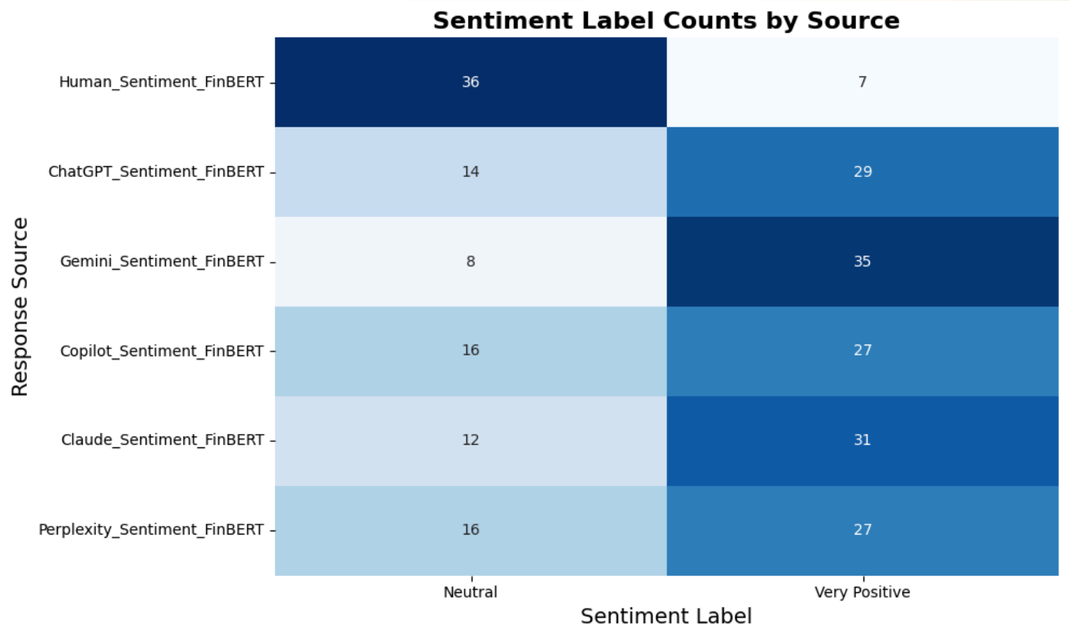
BOX PLOT - TextBlob Sentiment

- **Lower Sentiment Scores Overall:** TextBlob sentiment scores remain close to neutral across all sources.
- **TextBlob and VADER have different methods for interpreting sentiment**

VADER is more responsive to positive language and expressions. TextBlob has difficulty diffusing in this context

SENTIMENT ANALYSIS RESULTS

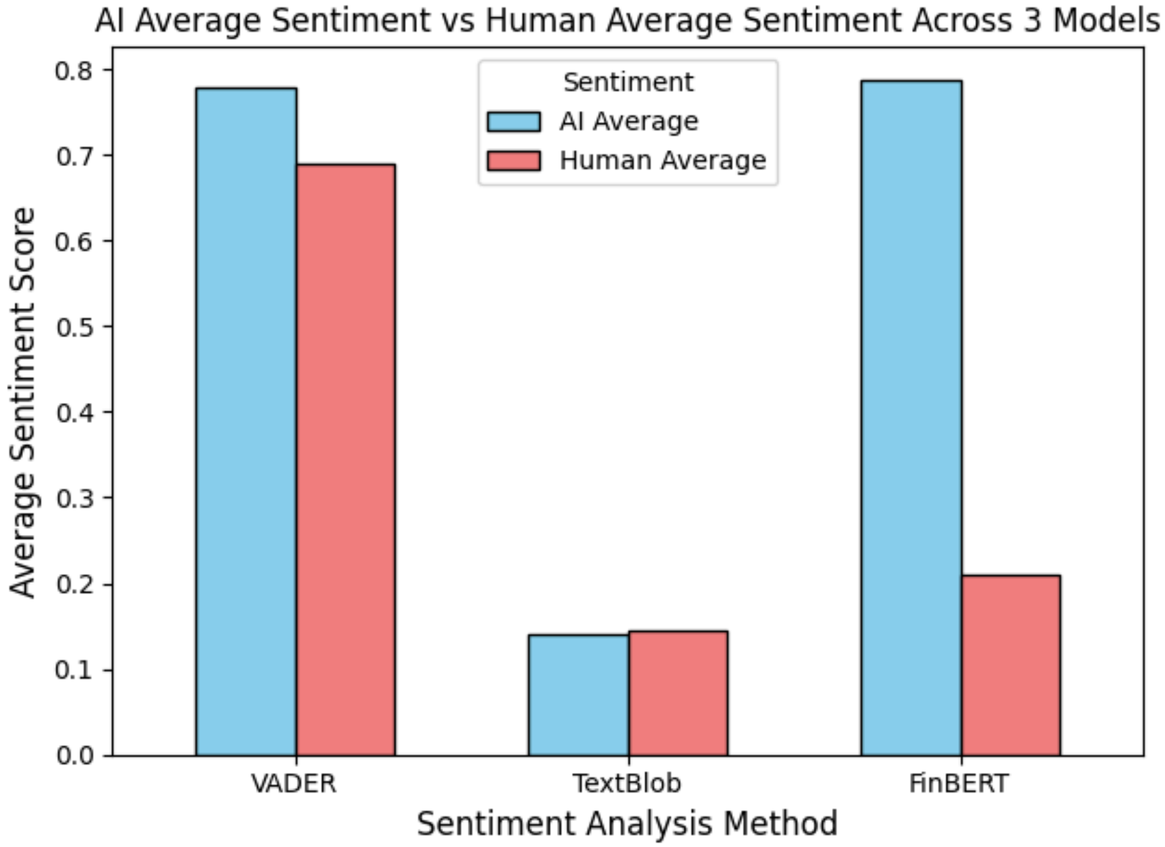




BOX PLOT – FinBERT Sentiment

- **A Neutral sentiment score** for **human answers** aligns with **formal, corporate language,** which tends to be **cautious and neutral.** This seems appropriate, as **corporate responses** often aim to be **neutral or balanced.**
- **ChatGPT (0.75), Copilot (0.7), and Perplexity (0.75)**:Show a neutral to slightly positive sentiment, which is typical for AI models trained to provide responses that are informative but not too negative.
- **Gemini (1.0) and Claude (0.8)**: Show a **more positive sentiment**, suggest that they generate responses that are either overly optimistic or confident. It could be seen as a bias towards positive outputs due to AI Favered Mindset

SENTIMENT ANALYSIS RESULTS



- **While TextBlob and VADER are** useful for sentiment analysis in **general,** they might miss the nuances **of corporate tone.**
- **These general model can,t able to distiguidh difeernce Human vs AI**
- **FinBERT is Ideal for Corporate Language:** It is best suited for this task, offering insights into AI vs. human sentiment and identifying where AI might exaggerate or misinterpret tone

BOX PLOT - MODEL Comparision

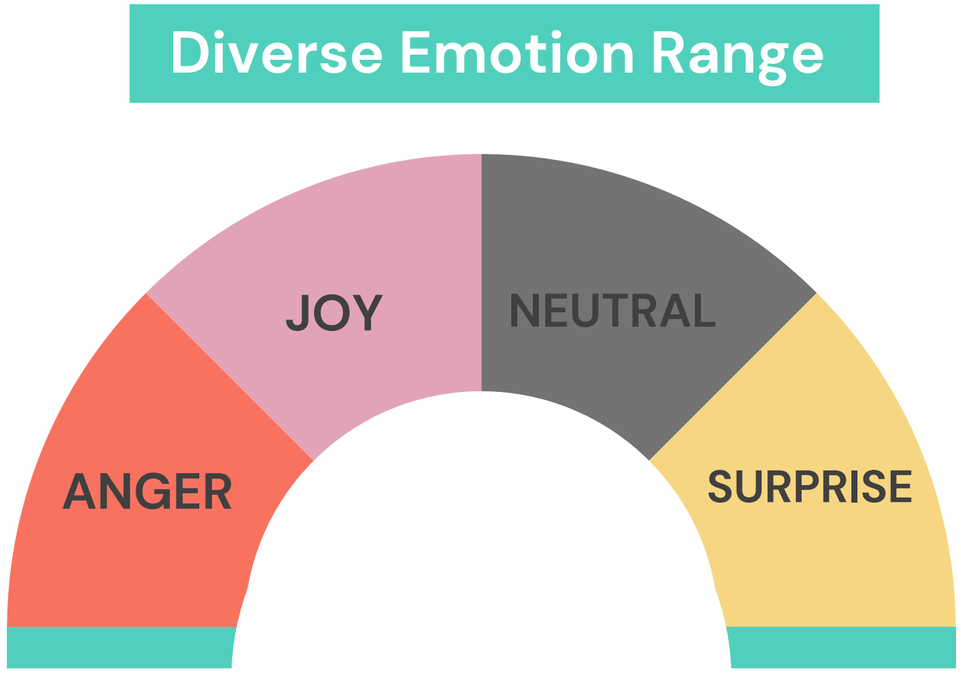
FinBERT's higher variation in sentiment between human and AI responses gives a clearer distinction for deeper analysis of AI model biases.

EMOTIONAL ANALYSIS IDEATION

Sentiment Analysis gives us a high-level view of whether a piece of text expresses a positive, negative, or neutral sentiment. Emotional Analysis,, goes a step further. It helps identify complex emotional states beyond just sentiment polarity. For example, emotional analysis will break down into emotions like joy, excitement, surprise, trust, etc., which gives a more nuanced understanding of the emotional tone

IDEATION

Our aim was to compare the Emotion Range of AI-generated answers with human responses, ideally to see how diverse these AI answers wrt real human experts.



EMOTIONAL ANALYSIS PROCESS

Used 1 Corporate Specific Model

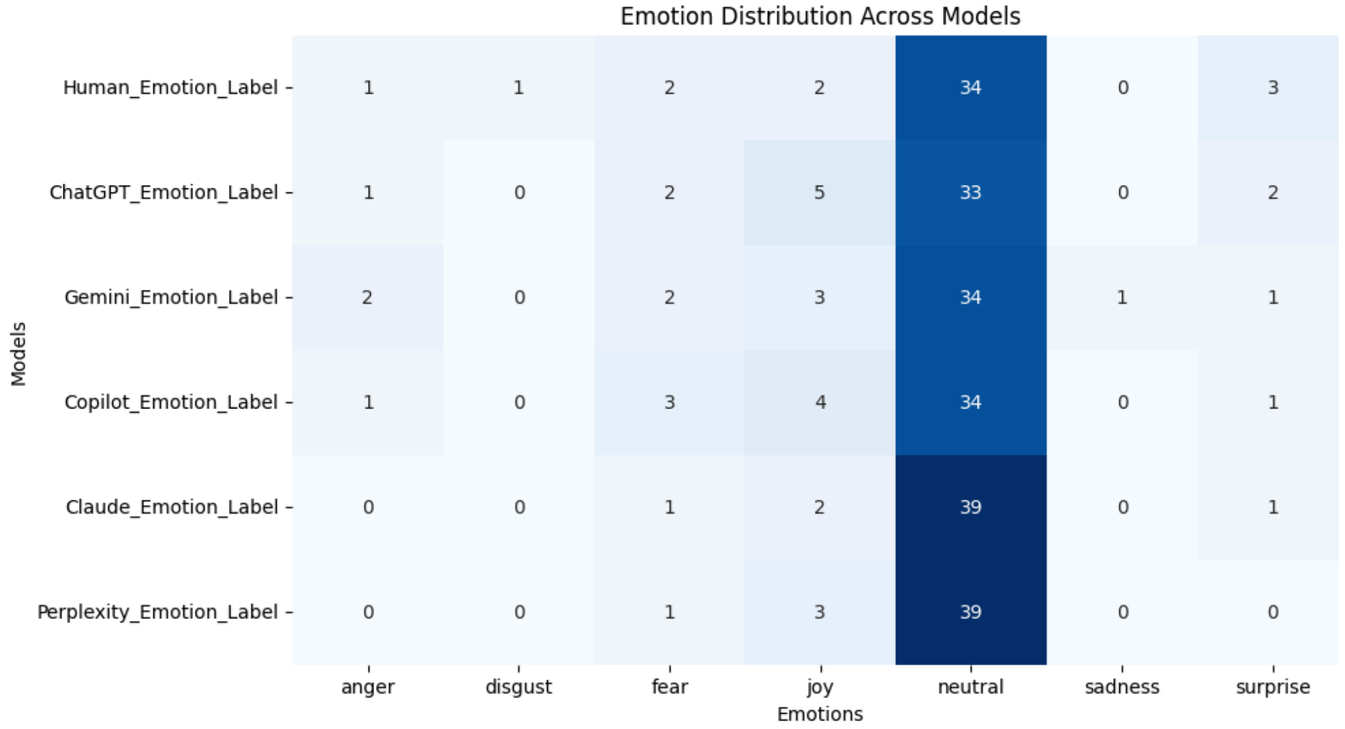
Process Flow

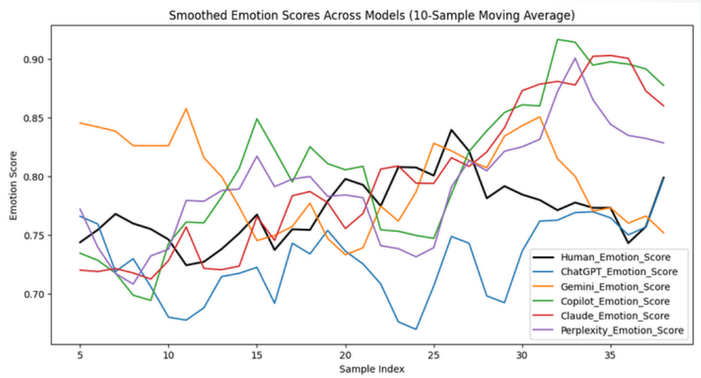
- **Data Input:** The cleaned text (answers or responses) is inputted into the model.
- **Emotion Detection:** The model processes the text and identifies the underlying emotions, classifying them into various categories such as joy, anger, sadness, fear, etc.
- **Output**: The results are displayed with the detected **emotion labels and confidence scores**, allowing for deeper emotional insights into the text.
- EX- {'label': 'joy', 'score': 0.917616605758667}

Model : j-hartmann/emotion-english-distilroberta-base

Transformer-based model :
It perform emotional analysis of tex DistilRoBERTa, is a lightweight version of the RoBERTa transformer architecture, fine-tuned specifically for emotion classification.

EMOTIONAL ANALYSIS RESULTS





BOX PLOT - FinBERT Sentiment

- **Humans have a slightly more distributed pattern across emotions like anger, sadness, and surprise** than some AI models.
- **ChatGPT and Copilot** exhibit **slightly more diversity** in emotional labeling, **resembling human labels more closely,** though not perfectly.
- **Models (Claude, Perplexity)** appear **conservative, heavily favoring "neutral"** and showing **limited diversity in emotion** classification.