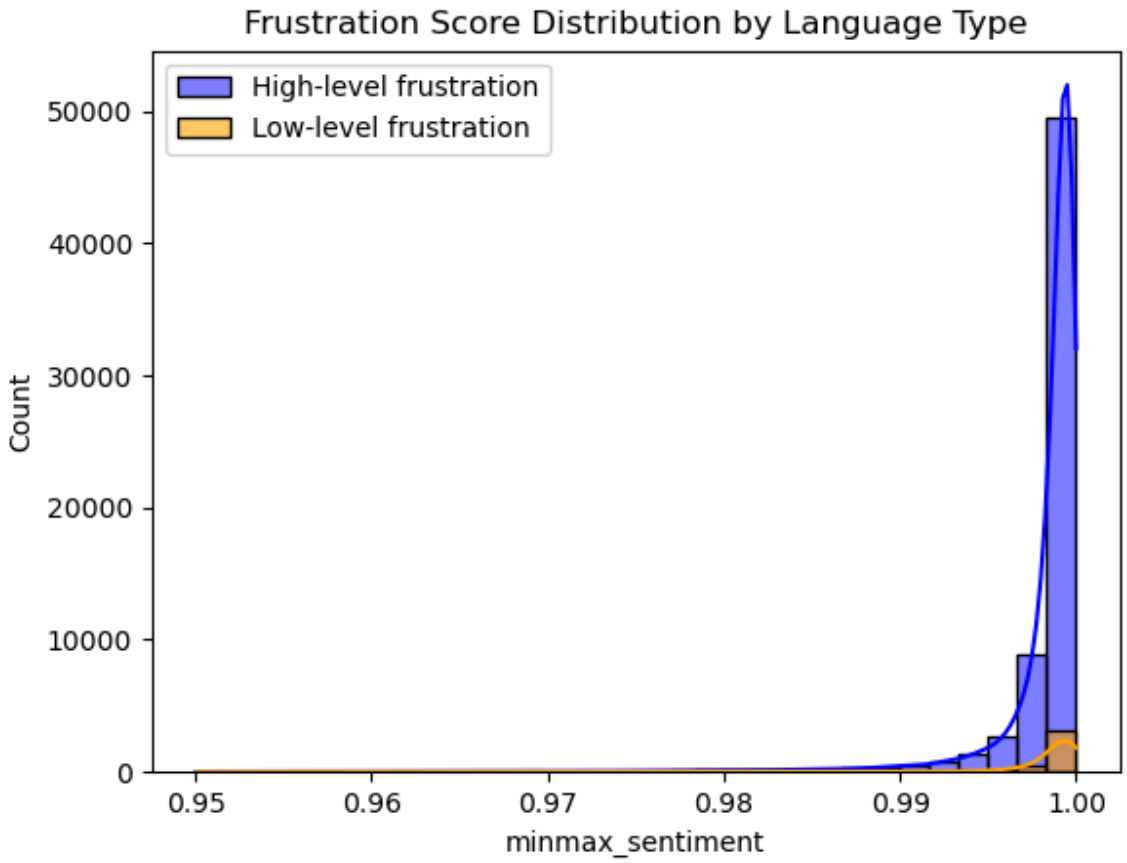
Visualizations, Statistical Tests and ML Models

Hypothesis 1

Visualization



We picked this visualization because histograms are ideal for displaying distributions, and we wanted a quick comparison between the high- and low-level frustration distributions. The main challenge with this visualization was deciding which range of minmax_sentiment values to display. Because our distribution was so concentrated between .99 and 1.00, we truncated the visualization to display values between 0.95 and 1.00, even though the support is technically [0,1]. Thus the obvious alternative would have been to display the entire support set; however, this ends up being less communicative despite technically being more accurate.

For those familiar with the setup and methodology of our project, this visualization should not require text. Of course, for those who want to learn more about our project, we have an abstract and analysis section with details relevant to the results displayed here.

Statistical test: Welch's t-test

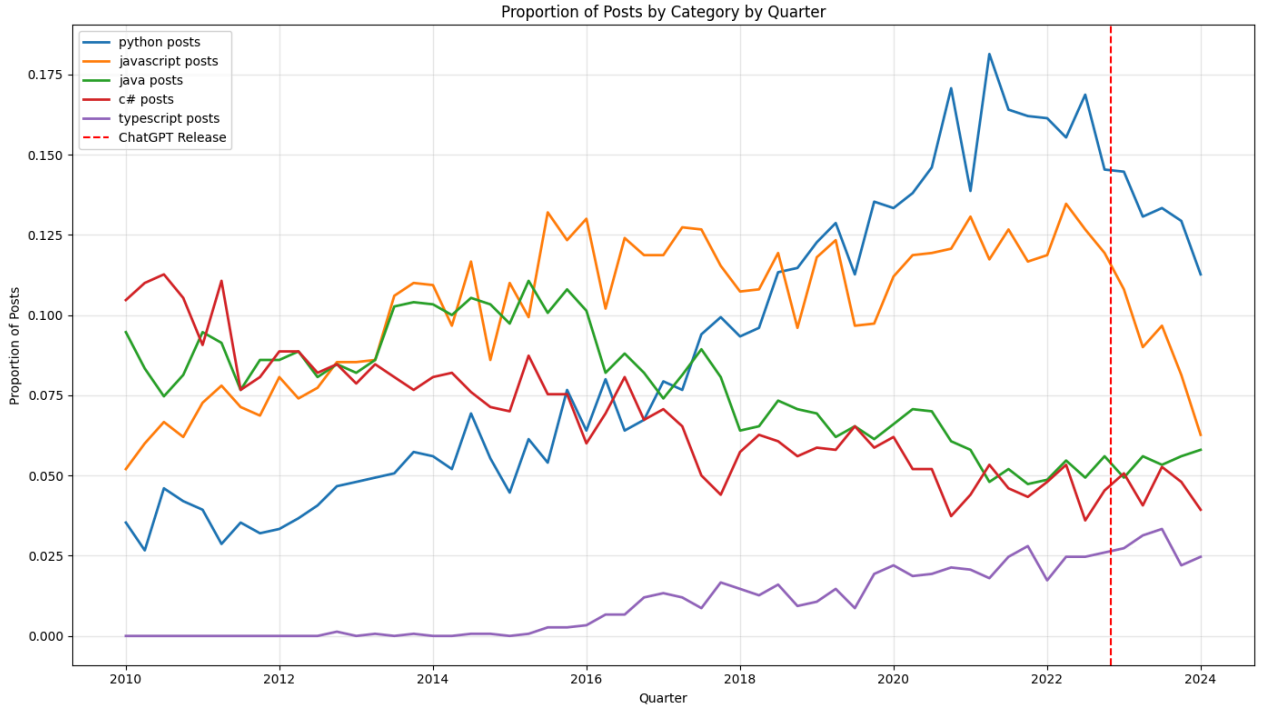
We chose Welch's t-test to test the significance of the differences between the two distributions. Welch's t-test compares the means of two independent distributions with different variances, making it ideal for determining the degree to which the low-level frustration distribution differed from that of the high-level frustration. This differs slightly from the usual two-sample t-test, which we considered using, but which would not be suitable because it assumes that each distribution has equal variances (which a quick inspection of the visualization will show is not the case).

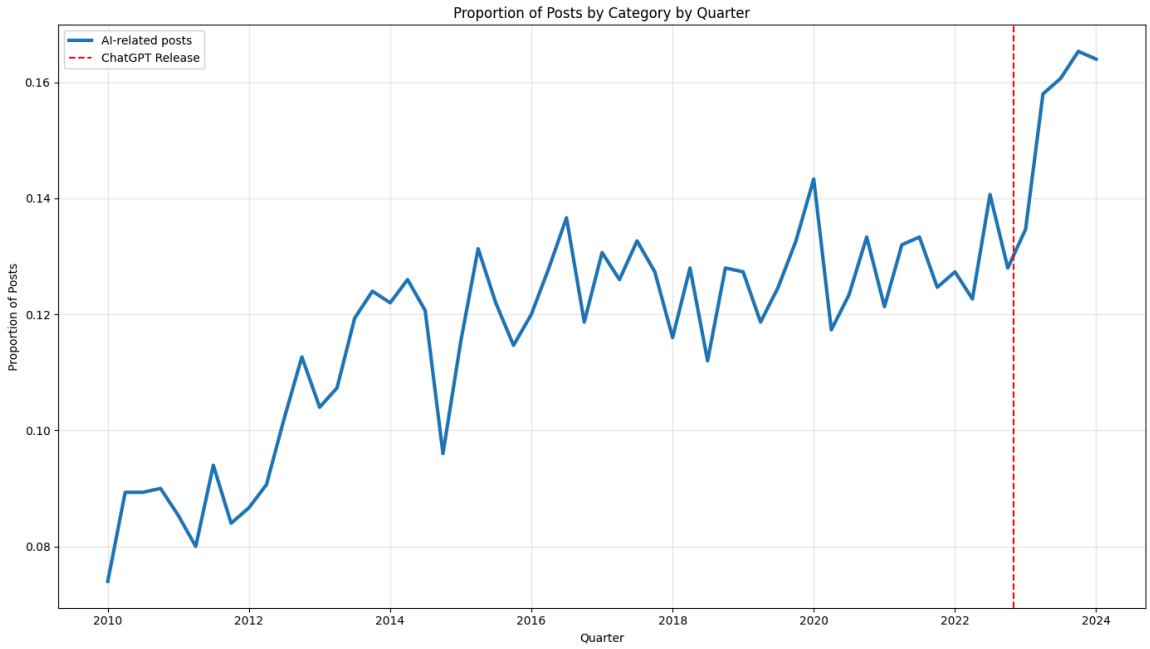
Machine Learning Model: Sentiment Analysis

We used a standard sentiment analysis model, Hugging Face's DistilBERT Sentiment Analysis Model, to get the sentiment of each post. This model was trained on a sentiment analysis of movie reviews, which is likely why the distribution of scores was so skewed — the textual data it was applied to, StackOverflow posts, looks much different from the data on which it was trained.

Hypothesis 2

Visualization





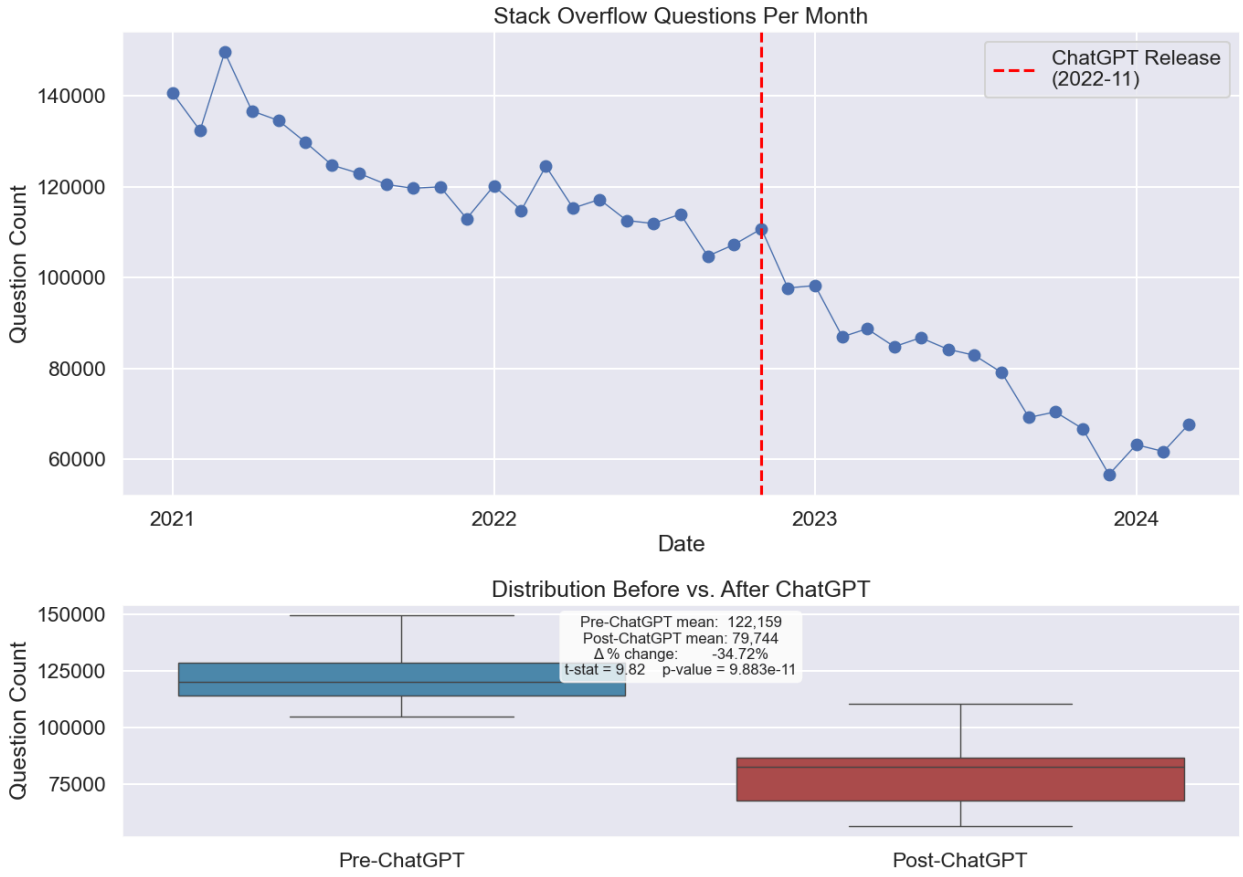
We picked this representation because line charts are effective at showing discrete changes and trends over time. Each vertex on the chart represents the proportion of posts from a given category by quarter. While pie charts are often a more useful alternative for displaying static proportional information, the changing proportions through time make line charts more suitable for conveying temporal trends (we can quickly see, for instance, that AI-related posts increases substantially after ChatGPT's release, while posts about Javascript and Python decreased). We didn't face any challenges in designing this visualization, and the visualization will not require text to provide context, again assuming that the viewer has the general context of Hypothesis 2.

Statistical test: Chi-squared test

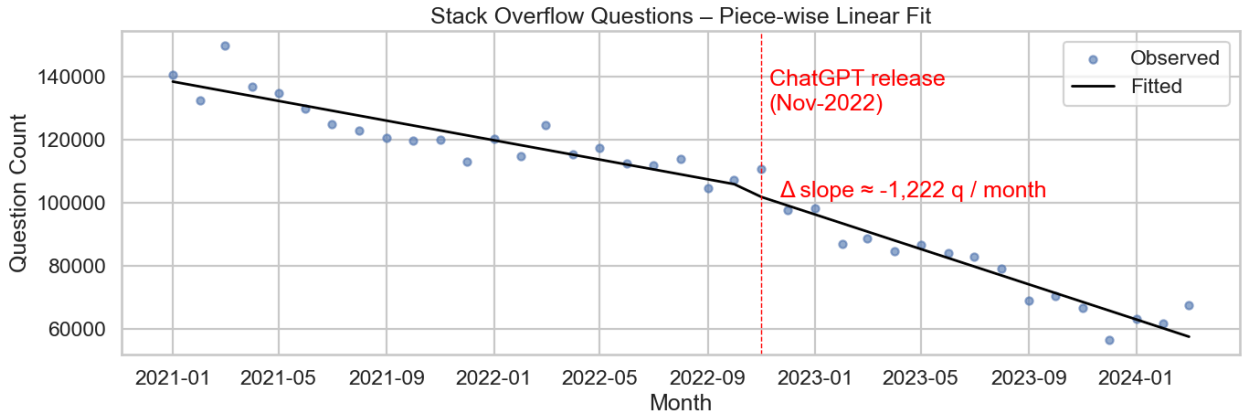
We chose the Chi-squared test here because we wanted to determine whether the distribution of two categorical variables (AI Related or Not AI Related) differed significantly before ChatGPT and after ChatGPT. This data can be easily represented in a (2-by-2) contingency table, which is what the Chi-squared test is built to analyze. An alternative test might have been the two-sample t-test, since we wanted to compare the distributions of two variables. However, t-tests are more suitable for continuous distributions, while this distribution was categorical/discrete (with two categories).

Hypothesis3

Visualization



We picked the first representation, the line chart, for the same reasons we picked it in Hypothesis 2: it's ideal for showing discrete changes in quantities over time periods. We picked the box plot for the second visualization, because it allows for a quick comparison of the range, quartiles, and median of two distributions. One potential alternative would be to use a histogram to show the distribution of question counts, but we thought that the box plot would be more illustrative because it shows the degree to which the medians and ranges vary without displaying more information than necessary. We didn't encounter any challenges in visualizing these results, and our visualization will not require additional text, conditioned on an understanding of the project.



For our third visualization for Hypothesis 3, we chose a scatter plot with a piecewise linear line-of-best-fit. This allowed us to show the changes in trends before and after ChatGPT's release in November 2022. While we might have chosen to use just one line-of-best-fit (to communicate the overall trend) or to simply state the difference in slope between pre-and-post ChatGPT data, we decided that two lines-of-best-fit along with a scatter plot was the richest and most illustrative way to demonstrate the change in trend corresponding to ChatGPT's release. We didn't encounter any challenges visualizing these results, and our visualization will not require text to provide context, as long as the viewer is generally familiar with the project.

Statistical test: Welch's t-test

We applied Welch's t-test to compare the means of StackOverflow posts in the two years before ChatGPT's release and in the two years following. The t-test was suitable as a means of evaluating whether there is a significant difference between the average monthly posts before and after ChatGPT's release and we could directly compare the mean monthly posts. Importantly, however, there is already a clear decreasing trend before ChatGPT's release in StackOverflow's monthly posts, so the likely t-test overestimates the significance in decrease that could be from ChatGPT's impact. To account for this pre-existing decrease over time, we also modelled the decline using linear regression, checking the difference in slope before and after ChatGPT's release.

Machine Learning Model: Regression

OLS linear regression was applied to analyze the trend in the number of StackOverflow questions over time, particularly focusing on the change in slope before and after the release of ChatGPT. Because the data is roughly piecewise linear, this was a natural choice to analyze the difference in trends pre- and post-ChatGPT.