**Task 5**

**Data file path in GitHub: Tasks/task/Task5\_Extended\_Statistics\_Comparison.xlsx**

Task 5: Analyse the linguistic features using widely used tool called Linguistic Inquiry and Word Count (LIWC) to analyze the linguistic features of Github answers and Github Copilot answers.

We plan to use the LIWC-22 tool to analyze the proportion of emotional and linguistic features in AI\_answers and human\_answers on GitHub. First, we discussed which types of words might be related to emotions and linguistic patterns. Based on our discussion, we selected a total of 41 categories of words for analysis as follows.

• WPS (Words Per Sentence): Average number of words per sentence, measuring sentence length and complexity

• Dic (Dictionary words): Proportion of standard vocabulary recognized by the LIWC dictionary

• ppron (Personal pronouns): Overall frequency of personal pronoun usage

• i: Frequency of first-person singular pronouns ("I", "my") usage

• we: Frequency of first-person plural pronouns ("we", "our") usage

• you: Frequency of second-person pronouns ("you", "your") usage

• shehe: Frequency of third-person singular pronouns ("he", "she") usage

• they: Frequency of third-person plural pronouns ("they", "them") usage

• article: Frequency of article usage ("a", "the", "that")

• prep (Prepositions): Frequency of preposition usage ("in", "through", "about")

• adverb: Frequency of adverb usage ("very", "quickly", "obviously")

• negate: Frequency of negation word usage ("not", "no", "never")

• verb: Overall frequency of verb usage

• adj: Overall frequency of adjective usage

• affect: Overall frequency of all emotion-related vocabulary

• posemo: Frequency of positive emotion vocabulary ("love", "happy", "excellent")

• negemo: Frequency of negative emotion vocabulary ("hate", "terrible", "sad")

• anx: Frequency of anxiety-related vocabulary ("worry", "nervous", "fear")

• anger: Frequency of anger-related vocabulary ("hate", "angry", "furious")

• sad: Frequency of sadness-related vocabulary ("crying", "sad", "depressed")

• social: Overall frequency of social reference vocabulary

• family: Frequency of family-related vocabulary ("mom", "brother", "family")

• friend: Frequency of friend-related vocabulary ("friend", "neighbor", "roommate")

• female: Frequency of female-related vocabulary ("lady", "she", "mom")

• male: Frequency of male-related vocabulary ("gentleman", "he", "dad")

• cogproc (Cognitive processes): Overall frequency of cognitive process-related vocabulary

• insight: Frequency of insight-related vocabulary ("think", "know", "consider")

• cause: Frequency of causality-related vocabulary ("because", "effect", "cause")

• discrep: Frequency of discrepancy-related vocabulary ("should", "hope", "if")

• tentat (Tentative): Frequency of tentative/uncertainty vocabulary ("perhaps", "maybe", "guess")

• certain: Frequency of certainty vocabulary ("always", "absolute", "definitely")

• differ: Frequency of differentiation vocabulary ("but", "except", "however")

• focuspast: Frequency of past tense-related vocabulary ("went", "did", "was")

• focuspresent: Frequency of present tense-related vocabulary ("is", "now", "today")

• focusfuture: Frequency of future tense-related vocabulary ("will", "plan", "soon")

• informal: Overall frequency of informal language

• swear: Frequency of profanity/swear words

• netspeak: Frequency of internet slang ("btw", "lol", "thx")

• assent: Frequency of agreement vocabulary ("yes", "okay", "agree")

• nonflu (Nonfluencies): Frequency of non-fluent expressions ("um", "uh" and other fillers)

• filler: Frequency of filler words ("you know", "I mean")

LIWC-22 Word Category Selection Interface Screenshot:

:

图片包含 表格

AI 生成的内容可能不正确。

After this, based on the LIWC2015 Dictionary, we obtained the proportion of occurrences for these categories of words. The results have been uploaded to GitHub, in the Task5\_AI and Task5\_Human files within the Task5 folder.

Screenshot:

图形用户界面, 应用程序, 表格

AI 生成的内容可能不正确。

Figure 1\_AI

日历

AI 生成的内容可能不正确。

Figure 2\_Human

Since the proportions of 41 linguistic categories are relatively difficult to analyze directly, we integrated the data to obtain the mean, median, quartiles, mode, standard deviation, and variance for each category.

The results have been uploaded to GitHub in the Task5 folder as Task5\_Extended\_Statistics\_Comparison.xlsx.

Screenshot：

表格

AI 生成的内容可能不正确。

**Data analysis:**

图表, 条形图

AI 生成的内容可能不正确。

Human use more words when they answer the question. But we limited the word when AI answers the question, so these data no longer have any reference value.

图表, 条形图

AI 生成的内容可能不正确。

AI uses personal pronouns more than twice as frequently as humans. AI more like to use first and second pronouns and never use third pronouns and “We”. The use of personal pronouns in "human answer" is more diverse. However, no third-person pronouns were used.

图表, 条形图

AI 生成的内容可能不正确。

Human more like to use descriptive language than AI. This indicates that human is more vivid when answering questions.

图表, 条形图

AI 生成的内容可能不正确。

Human more like to use verb when answering questions. This indicates that humans tend to construct vivid scenarios when answering questions by using verbs, while AI places greater emphasis on the accuracy and logic of the information.

图表, 条形图

AI 生成的内容可能不正确。

humans use more emotion words than AI. It can be observed that when humans answer questions, they tend to incorporate more of their emotional factors, while AI tends to respond in a more rational and logical manner.

图表, 瀑布图

AI 生成的内容可能不正确。

In terms of language features, AI tends to answer questions from a perspective outside of the social members, or in a friendly manner. However, human responses are more likely to be from a certain social member's perspective.

图表, 条形图

AI 生成的内容可能不正确。

The use of cognitive vocabulary is similar between AI and humans. However, compared to humans, AI tends to use more causal and deterministic words, while compared to AI, humans tend to use more uncertain and tentative words. This indicates that AI pays more attention to causal logic and the answers it provides are mostly definite. In contrast, human responses may include some uncertain expressions.

图表, 条形图

AI 生成的内容可能不正确。

Humans tend to pay more attention to tense when answering questions, which indicates that humans pay more attention to the context when answering and have a more concrete understanding of time. In statements using tense, both AI and humans mainly use the present tense.

图表, 条形图

AI 生成的内容可能不正确。

Human responses tend to use more informal language compared to AI responses. Humans are more flexible when answering questions, which also highlights the limitations of current AI in terms of context perception and cultural adaptation.