let me tell you what I did word for word and then instruct how to draw it so others understand fully

\*\*Data Collection\*\*

first, 60 participants data collected - big five personality test from that we extracted true levels of personality traits for all 60, and also asked for 2 short texts designed to mimic social media posts from each participant

Three psychology experts were asked to rank all participants' big five personality traits ranging from Low, Moderate and High based on the two posts collected earlier and to explain the rationale behind their selection

\*\*Generating Rules\*\*

I created an AI agent with api of openai to be an expert assistant supporting researchers in developing accurate, generalizable classification rules to detect personality traits in social media posts. for example a rule: "IF a person uses vivid and energetic language to describe a fun event with multiple friends and expresses excitement about the experience, THEN extraversion is likely High because it shows they derive joy and energy from social interactions."

\*\*Classifing text to personality traits levels\*\*

I asked an 3 AI agents to classify the posts. 1. zero shot classification, 2. 5 randomly selected shot classification, 3. incorporating experts-derived rules in the system prompt (my new method) trying to teach a machine to classify, the difference in the third agent is that I also instructed to return the rules that matched to the posts and the quotes from the posts that prove the connection.

for 15 replications, I divided the 60 participants into 2 randomly shuffled folds. in its turn, the 180 psychologists' comments were used for building set of rules. The remaining 30 particpants were formed as the test set. for each participant belonged to the test set, we extracted the trait level the AI agent predicted to be the correct one. As noted before, we were interested in understanding what rules match to each instance from test set.

We found out that some rules match to 90% of subjects but are not predicting correctly. we suggested to prune those rules and set minimum matches threshold of 15 matches and maximum correct ratio threshold of 0.67 between correct matches and all matches for each rule. Only if the two conditions meet that it is okay to prune out the rule and start over classifing the fold again without the rule pruned. I set the maximum iterations to be 5 iterations. We keep track on accuracy and Mean Absolute Error (MAE) because the classes are ordinaly sorted and choose the iteration with the highest accuracy and lowest MAE (currently not sure in what order is best)

In the end I want to analyze the rules that were responsible for the increase in accuracy and the decrease in MAE in comparison to zero shot and few shot.