**Writing notes for Master’s thesis**

Outline:

1. Introduction
   1. Common Misinterpretations of p values
   2. Focus on the 'p > .05 = no effect' Misinterpretation
   3. Why This is a Human Factors Problem
   4. Existing (Automated) Human-Factors Solutions
2. Methods
   1. Project Components
      1. Statement Detection
         1. RegEx and Papercheck
      2. Statement Classification
         1. NLP, BERT
      3. Statement Correction
         1. Groq and LLMs
   2. Sample and Annotation Process
      1. Descriptives
   3. Human Evaluation and Validation Strategy
      1. Where what was evaluated by a human and how it will be compared with the automated approach
3. Results
   1. Detection Accuracy
      1. How many of the human identified statements with nonsig. p values does the automated approach get and how many false positives are there?
      2. Examples of false positives
   2. Classification Performance
      1. How well did the model predict correct labels of 0 or 1
      2. Examples of falsely classified statements
   3. Correction Evaluation
      1. How many of the (model predicted) incorrect statements that were corrected by the LLM were now actually correct, incorrect or otherwise problematic (as in it’s not useful as writing assistance), as assessed by a human expert?
      2. Examples of incorrect ‘corrections’
4. Discussion
   1. Summary of Findings
   2. Strengths of the Approach
   3. Limitations and Challenges
   4. Implication of the Tool
   5. Future Directions and Improvements
   6. Conclusion
5. Appendix
   1. Supplementary Material and Final ToDos
      1. Clean GitHub Repo
         1. Requirements file or so for R packages and environment?
      2. Check References (especially for non papers like packages, tools, and websites, …)
      3. Create streamlit app to test model, get prediction etc.?
      4. Check all codebooks