Overview and Methodology

Background: Company XYZ wants to explore building an NER model which can correctly annotate and classify components of recipes. XYZ is confident in its team's annotation skills but wants to explore where to allocate its resources in obtaining the best training data for the model. Brandon was tasked to develop a proof of concept and deliver insights and potential next steps.

Process: Brandon obtained two datasets and also created his own evaluation sample to test potential training data. The evaluation dataset was kept constant across all test cases in order to compare all model performances consistently. The datasets consist of:

1. **Self-trained (evaluation):** 100-manually annotated sample of unlabeled data. Benefit: internal with documented annotation guidelines
2. **Third-party trained (external):** 1250-annotation sample. Annotations written by spectators at a 2023 data workshop in NYC. Benefit: large sample size. Consequence: potential quality issue due to no annotation guidelines
3. **GPT-generated (AI):** 500-annotation sample. Benefit: automated sample size augmentation. Consequence: No human intervention

Initial model accuracy resulted in the external model at 63% and the AI model at 48%.
Hypothesis 1 (merging): Brandon hypothesized that it is possible that there could be valid annotation techniques that are unique to each initial model, and it is worth checking if merging the external and AI datasets would increase accuracy. After doing so, there was no change in the model's best performance (63%) and it was determined that he would omit AI data from the testing due to its low original accuracy and no impact to model performance. The external dataset would be used as a baseline, and as such a new model pipeline "model-pydata" was generated for keeping track of the best results for exclusively the external data.
Hypothesis 2 (additional data): Although not available, Brandon suspected that adding additional external data would be beneficial to increasing model accuracy. After running a train-curve analysis, the accuracy continued to rise between the 75% and 100% thresholds which supports the hypothesis (Appendix 1.1).
Hypothesis 3 (correcting external data): Brandon attempted a third approach and corrected 100 annotations from the external dataset to see the impact of internal knowledge on external data. After merging these corrected annotations back to the original pydata-nyc data, we witnessed a slight increase in accuracy to 64%. It is possible that the 100 corrected annotations held little weight in the large training sample size.

Outcome and Next Steps:
Based on hypothesis testing and results, company XYZ should invest more resources in acquiring external data and having its employees apply consistent guidelines to correct the annotations accordingly. Acquiring a larger sample of data showed a directional improvement in accuracy and can be beneficial. According to test results, it is recommended to provide more detailed guidelines for the "DISH" labels as the accuracy scores skewed overall model performance downward with its relatively poor results (Appendix 2.1).

Annotation Guidelines

Task: Label comments from reddit to classify words into three categories: Dish, Ingredient, Equipment to be able to understand the contents of recipes and in the context of food.

- General
    - **Dish**
        - Treated similarly to a response variable (outcome) of an equation; the end result/final consumable
    - Ingredients
        - Treated similarly to independent variables of an equation; items necessary in created a dish
    - *Equipment*
        - Physical objects needed to handle ingredients and facilitate a part of or the entire process of making a dish
- Edge Case Examples
    - Ingredient-based dish names
        - “Ever had **beer brats**? Boil bratwurst in beer and then come and talk to me. You'll thank me.”
            - The dish “beer brats” is the final outcome of combining ingredients “bratwurst” and “beer”. Mark the outcome as “dish” and its components as “ingredients”
    - Dishes being used as ingredients
        - “I put jelly or jam on toast, then put the egg on that. Get a nice, gooey, sweet and savory delight”
            - Although having jelly/jam on toast in itself is a dish, by adding an egg to it makes all prior components ingredients. There is no reference of what this process is making, so this example would only contain ingredients
    - Brands used to specify/describe
        - “I have the Kuhn Rikon *peeler*. It is excellent!”
            - “Kuhn Rikon” is likely a brand of peeler, so we would treat the brand like an adjective an annotate “peeler” as equipment

Appendix

1.1 Train Curve

%	Score	ner
0%	0.02	0.02
25%	0.59 ▲	0.59 ▲
50%	0.59	0.59
75%	0.60 ▲	0.60 ▲
100%	0.63 ▲	0.63 ▲

2.1 Label Stats in Best Model

	P	R	F
DISH	53.66	43.14	47.83
INGREDIENT	62.99	65.04	64.00
EQUIPMENT	73.68	80.77	77.06