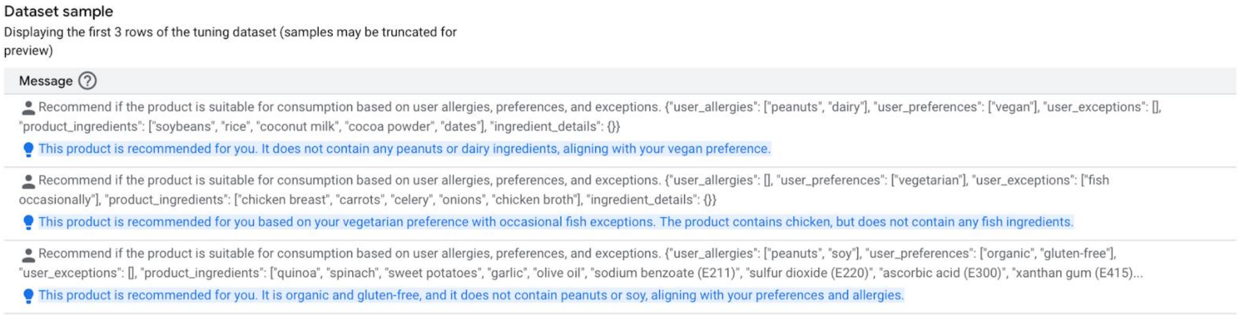
Dataset of Fine-tuning

The original dataset contains instruction, input and output. Instructions are all the same questions, "Recommend if the product is suitable for consumption based on user allergies, preferences, and exceptions.”. Inputs are users’ allergies, preferences, exceptions and product ingredients with details. Outputs are suggestions based on instructions and inputs to users whether nutrition profressions recommend to take foods with suggestions. We also transformed the data to the format of questions and answers to fit the input format, where questions combine intructions and inputs from original datasets and anwsers are outputs. Here is the examples and overview of the dataset.



Dataset sample

Displaying the first 3 rows of the tuning dataset (samples may be truncated for preview)

Message ?

Recommend if the product is suitable for consumption based on user allergies, preferences, and exceptions. {"user_allergies": ["peanuts", "dairy"], "user_preferences": ["vegan"], "user_exceptions": [], "product_ingredients": ["soybeans", "rice", "coconut milk", "cocoa powder", "dates"], "ingredient_details": {}}

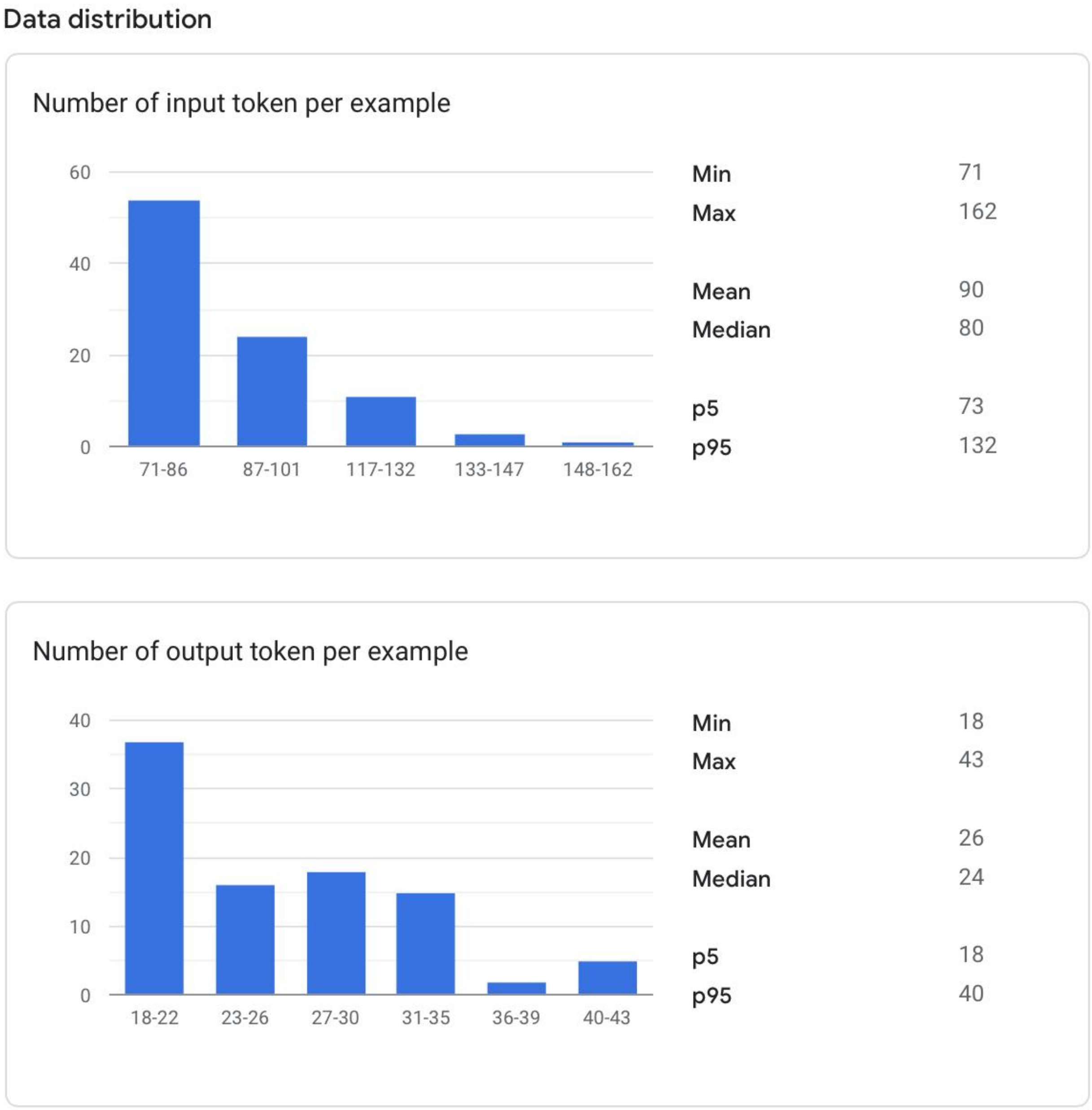
This product is recommended for you. It does not contain any peanuts or dairy ingredients, aligning with your vegan preference.

Recommend if the product is suitable for consumption based on user allergies, preferences, and exceptions. {"user_allergies": [], "user_preferences": ["vegetarian"], "user_exceptions": ["fish occasionally"], "product_ingredients": ["chicken breast", "carrots", "celery", "onions", "chicken broth"], "ingredient_details": {}}

This product is recommended for you based on your vegetarian preference with occasional fish exceptions. The product contains chicken, but does not contain any fish ingredients.

Recommend if the product is suitable for consumption based on user allergies, preferences, and exceptions. {"user_allergies": ["peanuts", "soy"], "user_preferences": ["organic", "gluten-free"], "user_exceptions": [], "product_ingredients": ["quinoa", "spinach", "sweet potatoes", "garlic", "olive oil", "sodium benzoate (E211)", "sulfur dioxide (E220)", "ascorbic acid (E300)", "xanthan gum (E415)...

This product is recommended for you. It is organic and gluten-free, and it does not contain peanuts or soy, aligning with your preferences and allergies.


Data distribution

Number of input token per example

Min	71
Max	162
Mean	90
Median	80
p5	73
p95	132

Number of output token per example

Min	18
Max	43
Mean	26
Median	24
p5	18
p95	40

```
Hyperparameters:
    "max_output_tokens": 3000,  # Maximum number of tokens for output
    "temperature": 0.75,  # Control randomness in output
    "top_p": 0.95,  # Use nucleus sampling
epochs=3, # change to 2-3
        adapter_size=4,
        learning_rate_multiplier=1.0
```

Model for fine-tuning
We aim to use Question Answering datasets to fine tune Gemini–1.5-pro-002 model. Later we may fine-tuning Llama 3.18 as suggested by our TA.