Feasting on Knowledge: Crafting a Food-Centric Question Answering Model

Jiaxin (Violet) Li, Nicholas Luong

April 14, 2024

Abstract

This study explores the effectiveness of utilizing a Multiple Choice BERT (Bidirectional Encoder Representations from Transformers) model to provide dietary guidance by determining whether a given food item aligns with an individual's dietary requirements. The model will be trained on a set of dietary questions, including common and rare diets, with the food data provided by the US Department of Agriculture and questions generated by us. By training and fine-tuning a BERT model on a dataset of food and dietary information, the model is capable of comprehending nuanced dietary constraints and preferences through a multiple choice question format. The paper will also explore using Information Retrieval as a pre-processing tool to see its impact on the performance of the model. We ultimately find the fine-tuned model without information retrieval performed the highest, with an accuracy of 83.7%. We conclude that by facilitating informed dietary decision-making, this model and approach holds promise for improving dietary adherence and overall health outcomes.

1 Introduction

Individuals with complex dietary restrictions often face significant challenges in identifying suitable meal options. This project aims to develop a multiple-choice question-answering (QA) model that can reliably determine whether a given dish complies with specific dietary requirements, such as allergies, religious preferences, special diets, or general food preferences.

Originally, the team planned to create a generative QA model to provide tailored recommendations. However, to enhance the reliability and accuracy of the outcomes, we have shifted our focus toward a multiple-choice format. This adjustment allows for definitive answers, which can be systematically evaluated . The model will evaluate if a particular dish is suitable for a user based on the specified dietary constraint.

Through this project, we anticipate gaining a deeper understanding of model performance in specialized domains, which could be invaluable in other contexts where accuracy is critical to prevent the potentially harmful effects of misinformation. By focusing on the precision of multiple-choice responses, we hope to minimize risks and provide a dependable tool for individuals navigating complex dietary needs. The insights gained and the data collected from this model may support the development of future generative models, which can suggest compatible dishes to users.

2 Background

While the necessity for evaluating dietary restrictions persistently exists, research in this domain remains scarce. Clairet (2017) focuses on the detection of dietary conflicts from dish titles using a lexical-semantic network [Clairet, 2017]. However, her model is limited to classifying only 11 common diets, which is insufficient given the vast array of dietary needs. On the other hand, task-oriented dialogue systems like PolyResponse demonstrate the effective use of retrieval-based models in conversational settings, which could be adapted for dietary QA applications [Henderson et al., 2019]. Despite its innovative approach, PolyResponse lacks evaluation of answer accuracy, casting doubts on its reliability. Other research efforts diverge significantly on aspects like dish pricing [Chahuneau et al., 2012], which offer little in terms of dietary restrictions.

Given these challenges, we need to develop a model that not only covers a comprehensive range of dietary preferences and requirements, but also incorporates a robust evaluation framework to ensure reliability. To this end, we propose a multiple-choice QA model that is versatile and directly assessable.

3 Methods

3.1 Data

To develop our multiple-choice QA model, we utilized comprehensive data from the U.S. Department of Agriculture's FoodData Central [U.S. Department of Agriculture, Agricultural Research Service, 2023]. This data is publicly available and can be downloaded in CSV format. Specifically, we employed the branded_food.csv file, which includes nutrient values derived from the food labels provided by brand owners.

This dataset is rich in details such as serving sizes, importing country etc.. For our purposes, we only used two columns: ingredients list and food category. The list of ingredients is as stated on the product label; the food category classifies the food according to standards set by the Global Data Synchronization Network (GDSN) or Label Insight. These two fields were instrumental in developing our model.

3.2 Prompt Generation

In our question-answering model, each prompt paired a combination of ingredients with a specific dietary restriction. The response is naturally a binary choice between "yes" and "no".

First, we randomly selected 1000 kinds of food from the dataset, and randomly paied them with one of the general and common dietary restrictions, including: gluten-free, vegan, vegetarian, pescatarian, lactose intolerant, allergic to soy, nuts, shellfish, or animal byproducts.

In addition to the 1000 common diets, we also crafted 500 samples in a similar way of random combination, addressing more specific or rare dietary needs including:

- Medical conditions like Alpha-gal Syndrome, Histamine Intolerance, Favism, Phenylketonuria, and Nickel sensitivity
- Personal Dislikes such as not feeling like mushroom, don't like spinach.
- Religious diets including Buddhist, Muslim, Hindu.
- Contemporary diets like keto and paleo.

With these 1500 samples as our dataset, we then generated prompts for each of the sample in the structure as follows:

> *I am {diet type}, can I have this dish or not? The dish contains ingredients: {ingredients list}. The food category is {food category}*

This format prioritizes the diet label at the beginning to ensure it remains within the prompt in case of truncation due to the 512-token limit. The dual approach in prompt generation—catering to both common and rare diets—enables comprehensive training of our QA model. This strategy not only tests the model's ability to handle typical dietary inquiries but also its proficiency in dealing with specialized or less common dietary restrictions, which are critical for a segment of the population.

3.3 Label Generation

For the 1000 common diet prompts, we implemented a keyword search within the ingredient lists to identify specific dietary components like meat, gluten, shellfish, dairy, eggs, and animal by-products. We come up with a very comprehensive keywords list which you may see in data_wrangling.ipynb. This keyword search helped us label each food item with terms such as 'contains_meat', 'contains_gluten', and other relevant tags.

Based on the initial tags from the keyword search, we generated further classifications into dietary categories such as vegan, vegetarian, and pescatarian. This process involved a manual review and correction of certain labels. For instance, we ensured 'almond milk' was correctly identified as non-dairy and vegan.

For the 500 special diet prompts, we utilized GPT-4.0 to automatically label these samples. Given the complexity and rarity of some diet labels, we also conducted a careful manual review and adjustment of all 500 labels. As a side note, the GPT-4.0 model's accuracy on special diets achieved 83.6%. As we do not want to go pass the free tier of GPT and we could not train a comparably large model, we decided not to use it as the baseline, but only a reference.

This methodical preparation of data and prompt generation plays a foundational role in training our QA model to function effectively within the specified operational parameters, thereby supporting users with precise dietary needs efficiently.

3.4 Baseline

We used the pre-trained bert-base-uncased model from Google BERT [Devlin et al., 2018], which is available through Hugging Face's Transformers library [Wolf et al., 2020], as the baseline model. This model was set up in its default configuration as a multiple-choice classifier (model class TFAutoModelForMultipleChoice) without any fine-tuning.

We selected BERT as the baseline because it is the most widely used models in the field of NLP, making it suitable as a robust benchmark. Its extensive pre-training on Wikipedia articles may also help the model adapt to specific tasks like dietary restriction QA.

The baseline model achieved a training accuracy of 69.2%, validation accuracy of 77.0%, and test accuracy of 73.4%.

3.5 Modeling Architecture

We fine-tuned the baseline BERT model, and used a grid-search approach to explore a number of different parameters. [Devlin et al., 2019] claimed a range of possible values to "work well across all tasks", so we started with the suggested group of parameters and added more options we want to test on. We explored the combinations of below parameters, trying to find the model with best performance:

- Batch size: 2, 4, 16, 32
- Epoch size: 2, 4, 6, 8
- Learning rate: 5e-5, 3e-5, 2e-5, 1e-4(default)

3.6 Instructor Embedding for Information Retrieval

Our baseline model demonstrated satisfactory performance on common diets but underperformed when dealing with specialized diets, particularly religious and rare medical ones. We hypothesized that this discrepancy came from the model's insufficient understanding of these special diets. Additionally, the limitation in prompt length constrained our ability to incorporate detailed descriptions directly from extensive dietary articles into the model prompts. To address these issues, we utilized Instructor Embedding for information retrieval.

We sourced from the Wikipedia list of dietary restrictions and medical posts from Mayo Clinic, which detail (but not too much detail) necessary dietary practices and restrictions. Due to computational limits within the Google Colab environment, we conducted a manual selection of relevant articles, as fully integrating a larger database or employing more robust models like Instructor-XL was not feasible.

We employed the instructor-large tokenizer, which is known for information retrieval tasks [Su et al., 2023]. We chose it because it provides high-quality embeddings that facilitate good sentence similarity comparisons without the need for fine-tuning. We constructed queries in the format: "What does it mean by {diet type} diet?", and calculated the cosine similarities between the embeddings of article lines and the diet query sentence.

To ensure the relevance and accuracy of the information retrieved, we set a high similarity threshold of 0.9. If no sentences scored above the threshold for a particular diet, the prompt remained unchanged to avoid the inclusion of potentially misleading information.

The full architecture of the modelling process looks like Figure 1

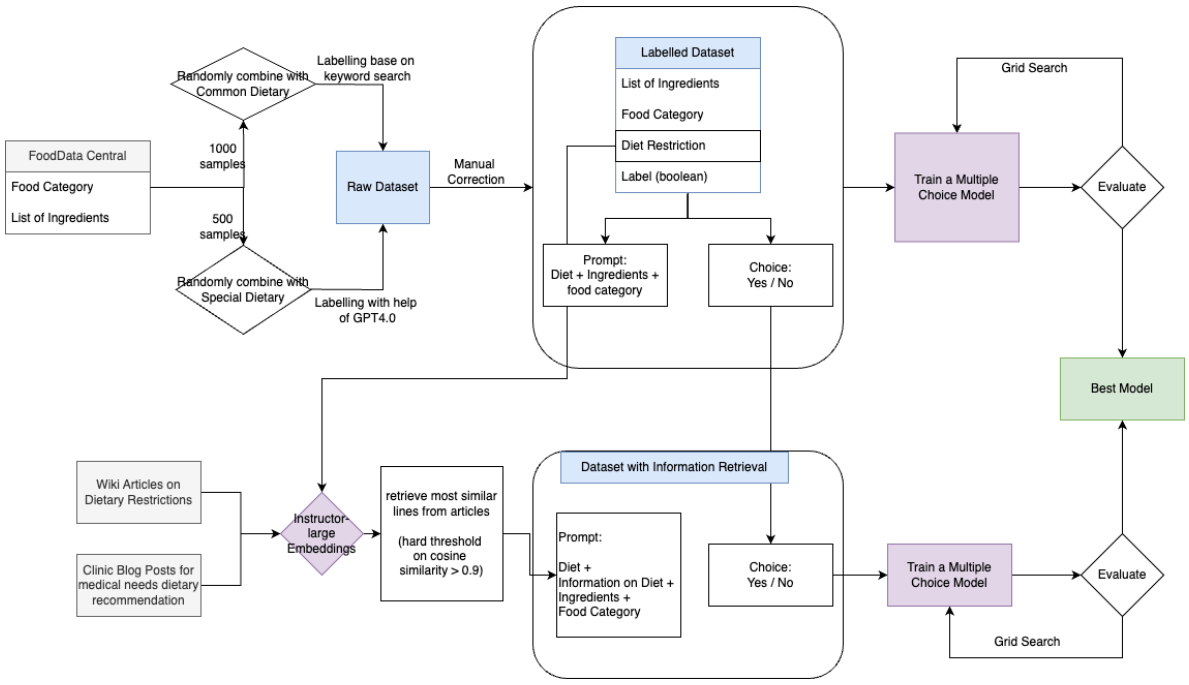

Figure 1: Modelling Architecture

4 Results

4.1 Accuracy

Model	Accuracy
Multiple-Choice Model	83.7%
Multiple-Choice Model with IR Pre-Processing	75.2%

Table 1: Model Performance

The baseline model initially performed at an accuracy of 73.4% but we implemented a GridSearch to find the optimal parameters for our model. In this case, we saw the highest performance of 83.7% with a batch size of 2 and an epoch of 8 on our test set. Notably, our model demonstrated differential performance across various dietary restrictions, with stronger performance observed for more prevalent diets compared to less common dietary restrictions. For instance, the model exhibited robust performance in identifying foods suitable for vegetarian diets, whereas its efficacy was comparatively lower for dietary restrictions such as Alpha Gal Syndrome. This suggests that the model's effectiveness may vary depending on the prevalence and specificity of the dietary restriction in question. In addition, it might also suggest that the model needs greater exposure to less common dietary restrictions in training to provide the knowledge needed to evaluate these diets.

To address this limitation, we propose incorporating relevant articles about these diets and leveraging information retrieval models to better summarize dietary behaviors. By incorporating this additional contextual information into our model's training data, we hoped to enhance its ability to accurately classify foods according to these more intricate dietary restrictions. Preliminary evaluations utilizing these augmented prompts yielded an improved accuracy of 75.2%, indicating the potential efficacy of this approach in bolstering the model's performance for more challenging dietary categories. While it did not outperform the original model, we still believe that incorporating more data for these dietary categories with additional contextual information will provide a better performance.

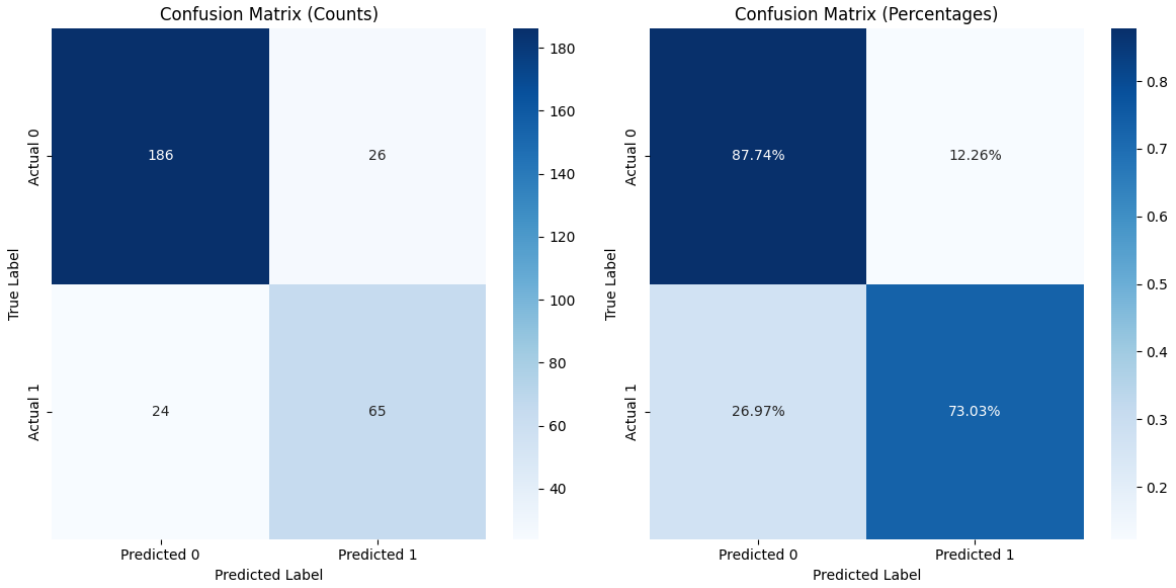

Figure 2: Confusion Matrix of the Finetuned Model

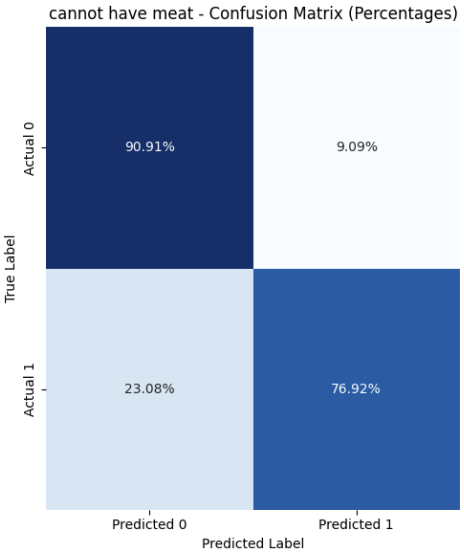

Figure 3: Confusion Matrix for a Meat-Free Diet

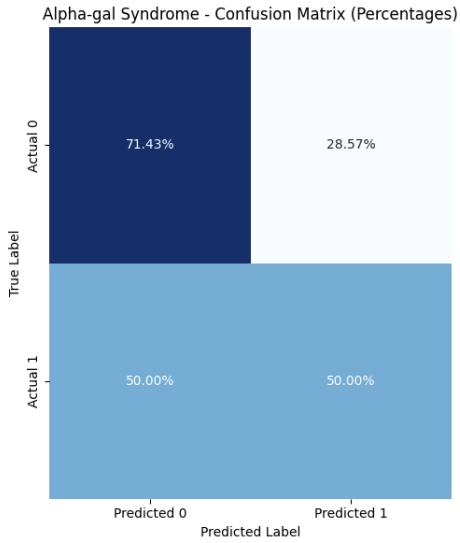

Figure 4: Confusion Matrix for Alpha-gal Syndrome

4.2 Discussion of Errors

Despite achieving notable accuracy, several factors have contributed to areas for improvement for future work.

One aspect mentioned in the previous section is how we chose to generate the questions for our training data. We used a structured template of the ingredients, food category, and the prompt to train the model on multiple choice tasks. Our training sets may not have been exposed to a sufficiently broad distribution of different dietary restrictions and variations in prompts, potentially impacting the model's ability to generalize effectively. The limited availability of diverse datasets may have influenced the performance of our model and underscores the importance of further efforts to curate more comprehensive datasets. This was the motivation for us to use an information retrieval model for pre-processing, in hopes that it would mitigate these issues.

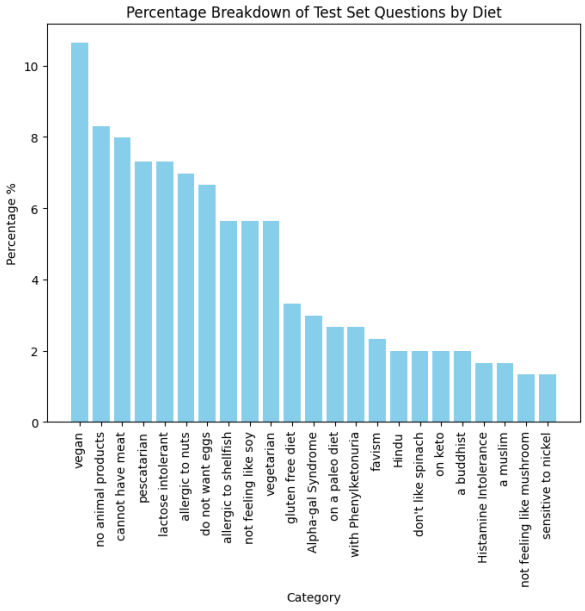

Figure 5: Distribution of Test Questions

Another discussion is the choice of using BERT over other model architectures. As previously stated, we chose BERT due to its ubiquity in NLP models and the model did perform well from our experience. However, picking a different architecture could be an area for further research to see if the performance improves. In addition, We chose to use a multiple choice task for ease of systematic evaluation, which is important in a domain like food restrictions, but there might be a need for other tasks incorporated such as asking questions about the food (Question-Answering) or summary about the food (Summarization).

5 Conclusion

Our evaluation of the application of a multiple choice model on distinguishing whether a meal fits a dietary restriction highlights the complexity of this topic. Our models performed well overall with a high accuracy rate but the disparity in the performance across diets demonstrates the challenges in working in this space as well. For certain people with less common allergies or restrictions, the accuracy of determining whether a food fits their needs is consequential to their health. Resources online lacks the clarity to help people navigate these complex diets and what food is acceptable and there are many obscure ingredients in packaged food where people won't know if the meal is acceptable to eat. Further work can be done to improve the model and, ultimately, provide a tool for people to help navigate their meals.

References

[Chahuneau et al., 2012] Chahuneau, V., Gimpel, K., Routledge, B. R., Scherlis, L., and Smith, N. A. (2012). Word salad: Relating food prices and descriptions. In Tsujii, J., Henderson, J., and Paşca, M., editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1357–1367, Jeju Island, Korea. Association for Computational Linguistics.

[Clairet, 2017] Clairet, N. (2017). Dish classification using knowledge based dietary conflict detection. In Kovatchev, V., Temnikova, I., Gencheva, P., Kiprov, Y., and Nikolova, I., editors, *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 1–9, Varna. INCOMA Ltd.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Henderson et al., 2019] Henderson, M., Vulić, I., Casanueva, I., Budzianowski, P., Gerz, D., Coope, S., Spithourakis, G., Wen, T.-H., Mrkšić, N., and Su, P.-H. (2019). PolyResponse: A rank-based approach to task-oriented dialogue with application in restaurant search and booking. In Padó, S. and Huang, R., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 181–186, Hong Kong, China. Association for Computational Linguistics.

[Su et al., 2023] Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., tau Yih, W., Smith, N. A., Zettlemoyer, L., and Yu, T. (2023). One embedder, any task: Instruction-finetuned text embeddings.

[U.S. Department of Agriculture, Agricultural Research Service, 2023] U.S. Department of Agriculture, Agricultural Research Service (2023). Fooddata central. https://fdc.nal.usda.gov. Accessed: 2023-04-10.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.