Large Language models for recipes
Project Proposal for NLP Course, Winter 2022

Maciej Chrabaszcz
Warsaw University of Technology
maciej.chrabaszcz.stud
Aleksander Kozłowski
Warsaw University of Technology
aleksander.kozlowski.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Large language models(LLM), such as those trained by OpenAI, can be used to classify and extract information from the text in a prompt-based manner. Given a prompt containing text data, these models can process the information and accurately identify relevant details such as dietary tags or specific ingredients. This approach allows for efficient classification and extraction of ingredients, as only a small amount of data (i.e. the prompt) needs to be processed at a time. This can be particularly useful in cases where annotating a large amount of data is infeasible or impractical and can be applied to various applications such as food labelling and recipe creation. Because of that, we will use pre-trained LLM in a prompt-based manner for the recipes domain. We aim to check how well those models can give dietary tags and extract product names without training.

1 Scientific goal

The scientific goals of using large language models (LLMs) in a prompt-based manner for the recipes domain include the ability to accurately classify and extract relevant information from text data, such as dietary tags and specific ingredients. This approach can be applied to various applications, such as food labeling and recipe creation, and is particularly useful in situations where annotating a large amount of data is infeasible or impractical. The aim is to determine the effectiveness of using pre-trained LLMs to provide dietary tags and extract product names without additional training.

2 Significance of the project

The significance of this project lies in its potential to streamline and improve various processes related to handling text data in the recipes domain. Using large language models in a prompt-based manner makes it possible to efficiently classify and extract relevant information from text data, such as dietary tags and specific ingredients. This can have several practical applications, such as improving the accuracy and speed of food labeling and recipe creation. Additionally, effectively extracting this information without additional training can save time and resources, making the process more efficient and cost-effective.

3 Concept and work plan

The concept of this project is to utilize large language models (LLMs) in a prompt-based manner to classify and extract relevant information from text data in the recipes domain. Specifically, we will use pre-trained LLMs to identify dietary tags and extract specific ingredients from prompts containing text data.

The work plan for this project will involve the following steps:

- Create a set of prompts templates which can be used for our tasks (classification and extraction).
- Use pre-trained large language models (LLMs) to process the prompts filled with text from the food.com dataset, classify them according to relevant dietary tags and extract specific ingredients.
- Evaluate the accuracy of the LLMs in classifying the prompts and extracting relevant information.
- Identify any areas where the LLMs are not performing as expected and explore potential

solutions to improve their accuracy.

- Determine the effectiveness of using pre-trained LLMs for classifying dietary tags and extracting ingredients in the recipes domain, as demonstrated by their performance on the dataset and prompt template.

4 Approach & research methodology

The approach for this project will involve using pre-trained large language models (LLMs) to classify and extract relevant information from text data in the recipes domain. This will be done in a prompt-based manner, with the LLMs processing individual prompts containing text data and identifying relevant details such as dietary tags or specific ingredients.

To research the effectiveness of this approach, we will follow the work plan outlined in the previous section. This will involve identifying a dataset containing text relevant to the recipes domain and using pre-trained LLMs to classify and extract information from the texts in the dataset. We will then evaluate the LLMs' accuracy and identify areas where they are not performing as expected.

To ensure the validity and reliability of our research, we will follow standard research methodology practices such as clearly defining our research question and objectives, selecting an appropriate sample and dataset, and using appropriate statistical analysis to evaluate the results. We will also ensure that the pre-trained LLMs are configured and used consistently throughout the research process.

In addition to evaluating the effectiveness of using pre-trained large language models (LLMs) for classification in the recipes domain, we will also compare our results to a previous project in which we trained LMs specifically for this task. This will allow us to determine the relative benefits and limitations of using pre-trained versus trained models for this task and to identify any potential improvements that could be made to the training process. By comparing our results to those from the previous project, we will gain a deeper understanding of the capabilities and limitations of LLMs for classification in the recipes domain.

References

[1] Conneau Alexis, Khandelwal Kartikay, Goyal Naman, Chaudhary Vishrav, Wenzek Guillaume, Guzmán Francisco, Grave Edouard, Ott Myle, Zettlemoyer Luke, and Stoyanov Veselin. Unsupervised Cross-lingual Representation Learning at Scale. *Cornell University - arXiv.*

[2] Wróblewska Ania, Kaliska Agnieszka, Pawłowski Maciej, Wiśniewski Dawid, Sosnowski Witold, and Ławrynowicz Agnieszka. Tasteset – Recipe Dataset and Food Entities Recognition Benchmark. *Cornell University - arXiv.*

[3] Tom B., Brown, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Daniel M., Ziegler, Wu Jeffrey, Winter Clemens, Hesse Christopher, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, McCandlish Sam, Radford Alec, Sutskever Ilya, and Amodei Dario. Language Models are Few-Shot Learners. *Cornell University - arXiv.*

[4] Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28, Dublin, Ireland, December 2020. Association for Computational Linguistics.

[5] Larissa Britto, Luciano Pacífico, Emilia Oliveira, and Teresa Ludermir. A Cooking Recipe Multi-Label Classification Approach for Food Restriction Identification. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2020).*

[6] Howard Jeremy and Ruder Sebastian. Universal Language Model Fine-tuning for Text Classification. *Cornell University - arXiv.*

[7] Kaggle. Food.com recipes with search terms and tags, 2021.

[8] Diwan Nirav, Batra Devansh, and Bagler Ganesh. A Named Entity Based Approach to Model Recipes. *Cornell University - arXiv.*