# DeepCoffee : Coffee Flavors Prediction Using Deep Learning

Manisha Sri Suresh
mssuresh@ucdavis.edu
University of California, Davis
Davis, California, USA

Zhuoheng Li*
pipli@ucdavis.edu
University of California, Davis
Davis, California, USA

## ABSTRACT

Over the years, the growing popularity for the coffee has led to the massive development of the coffee industry across the world. The world's leading coffee-producing countries are Brazil, Vietnam, Colombia, Indonesia,Ethiopia, Honduras, India, Peru, Uganda, and Mexico[25][32]. The coffee retailers of these countries use the most fascinating flavors and mix it to make it more special for the people however these flavor descriptions are not widely available due to rare number of certified professionals and tremendous market demand[15]. Currently, there are only few studies that is available to predict the overall score based on several determinants of the coffee including acidity,aroma,flavor,sweetness etc. In this paper, we proposed a novel idea to predict the flavor of the coffee with feature collection including country, region, processing methods, altitudes etc using GPT-3[14], a transformer based model along with the fine-tuning which we refer as a DeepCoffee model in this paper. We used OpenAI API [33] for GPT-3 model with fine-tuning [14] and built a web application which the prompts the user to enter the coffee features that includes coffee growing region, altitude, variety and the processing method and provides the coffee flavors as a result from the prediction. We were able to achieve a good of 83.5% using our DeepCoffee models while comparing over the other models available in openAI and emphasize that the fine-tuned GPT-3 [14] is a promising method for the coffee flavor prediction and can provide a new future objective in predicting the other complex flavors of food and beverages in the world[15].

## CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; *Neural networks*; *Model development and analysis*; *Natural language processing*.

## KEYWORDS

datasets, coffee flavor prediction, deep learning, fine-tuning, GPT-3, preprocessing, web application

## 1 INTRODUCTION

Coffee is one of the most widely consumed commodities worldwide and there are several distinct characteristics that makes it a unique and special beverage[15]. The worlds top coffee producing countries include Brazil, Vietnam, Colombia, Indonesia,Ethiopia, Honduras, India, Peru, Uganda, and Mexico[25][32].These countries drive the coffee manufactures with innovative solutions to the processing methods based on the country, region and altitude. There are several global brands across the countries competing with each other to provide the high quality and tasteful coffee comprising of roasted, instant, and ready-to-drink (RTD) coffee [35]. Coffee drinking has become one of the most important aspect in our individual daily routine while offering benefits of health and personal care along with the richness in texture, aroma and most importantly flavours. The flavor of a coffee is influenced by factors such as the geographical location of origin including country, region, altitude, variety, climatic factors, processing methods, roasting process, and preparation methods[19][37][26][31][18][11][7][23][22]. The differences in sensory characteristics also attribute towards coffee drinking[12]. Coffee drinking emphasise a significant role in the daily routines of many people as such it is beneficial for health and personal care providing the consumers with refreshment and convenience. Apart from these benefits, coffee is also famous for its texture,aroma and flavours. Especially, flavor is the most definite aspect for the users in a good coffee and continues to be a driving force for growth of Coffee in today's market[35]. There are hundreds of flavors available such as french vanilla, hazelnut, peppermint, pumpkin, caramel, mocha and further classifying the flavors into fruit flavors - stone fruit flavor such as peach, apricot, nectarine, citrus flavor including orange, lemon, grapefruit, nectarine; berries flavor such as raspberry, strawberry, blackberry, black currant, cherry; chocolate flavors such as milk chocolate, dark chocolate and cacao. Due to its growing demand, coffee industries tending to introduce new flavoring ideas by mixing multiple interesting flavors to attract the consumers. However, multiple flavors in a single coffee and limited availability of the coffee professionals leads to more complex problem for the prediction of descriptive coffee flavors[35].

Recently, deep learning[27] techniques are continuously drawing much attention from the industry, academy and other known fields. This is mainly because of its superior performance compared to the other previous machine learning techniques[29] and with the breakthrough of the transformer[39] architecture via the Attention mechanism. The existing works on pre-trained language representations in NLP systems have been directly fine-tuned removing the need for task-specific architectures completely. [34][28][21][14]. However, there is still a major limitation to this approach even while the architecture remains task-agnostic[14]. The NLP tasks requires a task-specific datasets and task-specific fine-tuning on a dataset of thousands to hundreds of thousands of examples specific to that task for achieving a strong performance[14]. Given some examples of the task as input, large language models (LMs) are allowed to perform a wide range of natural language processing (NLP) tasks[33]. However, these models often misalaigned and express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions [10] [13][24][41][38][20][33]. It is important to remove these biased and unintended behaviors especially for language models that are

* Q Arabica Grader.

deployed and used in hundreds of applications[33]. In order to encompass both explicit intentions such as following instructions and implicit intentions such as staying truthful, and not being biased, toxic, or otherwise harmful, the approach of fine-tuning can be used and such that it helps in aligning the language models to act in accordance with the user's intention[33][30]. Additionally the models hould be helpful (they should help the user solve their task), honest (they shouldn't fabricate information or mislead the user), and harmless (they should not cause physical, psychological, or social harm to people or the environment)[33]. The improved transformer based models like Generative Pre-Training [8] developed a novelty for Deep Learning [27] based NLP models. GPT architectures were trained on large datasets to create pre-trained models. Thereafter transfer learning was used to fine-tune these models for task-specific features resulting in significant performance on several NLP tasks[36]. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and close tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic[14].

We created the dataset by collecting the reliable and important coffee features data by performing the web scrapping from the real websites royalcoffeedb [4], thecaptainscoffee[5]. Followed by the data collection, we performed the data preprocessing to generate a more balanced dataset from the raw data to result in more accurate and efficient model performance results. We then performed the training process using our DeepCoffee model i.e fine-tuned GPT-3 using supervised learning with OpenAI [33]. Finally, we built a web application that is GUI based for providing a user with a good interface visualisation to display the coffee flavors result based on the input coffee features including coffee growing region, altitude, variety and the processing method. Overall in this work, we evaluated our data on our DeepCoffee, a GPT-3 fine-tuning model using open AI[33] and determined the model performance experimentally to show that our method and results are highly significant for the prediction of the coffee flavors while achieving a reliable and good model accuracy (83.5%).

The following sections of the paper is organised as follows: In section 2, we review our related work for our project followed by the detailed methodology in section 3 that contains the subsection of Data collection in section 3.1, Data Preprocessing in section 3.2, Data Modeling in section 3.3 that provides the training methods and process and a Web Application development in section 3.4. In section 4 , we present our results and effectiveness of the proposed methods followed by the Limitations and Future Work in section 5. Finally, we conclude our work in section 6 and our complete work is available on the GitHub link provided in section 7.

## 2 RELATED WORK

The flavor of the coffee and its correlation between combination of various factors including Altitude, Country/Region, Variety, and Processing Methods were understudied. Meanwhile, there are more limited research studying such subject through Machine Learning or Deep Learning methods. One recent study focused on investigating the feasibility of predicting complex coffee flavors with DCNN(ResNet) using NIR spectra readings as the inputs, have demonstrated the capabilities of ResNet in doing such task.[15] While (Chang Et.al, 2021) used NIR spectra readings as input to predict, such process could be both time and economical costly. The prior works [16] used several fine-tuning approaches to provide the text summarizing which provided an idea of using the fine-tuning approach for the coffee flavor in texual formats. Other works [42] included the work in the difference of training and testing data. The paper [33] provides us the knowledge of using the language models on natural language processing tasks improves the task performance on both few-shot and zero-shot. The paper [9] shows us on how the models are trained to follow the natural language instructions in a simulated environment. There are papers [10] that provided us the consequences of the possible risks in the real world associated with the behavior of the language models and tells us on how the model architcture can be modified mitigate the harms.

The thriving of Transformers[39] based models give the task another approach which to focus on its lexical meaning and giant amount of infomation existed on the internet which wrote sensory evaluation mainly done by coffee experts and Q Graders. Moreover, as Transformers were largely used Word Embeddings as their input and output, we could have a better measurement of the similarities of predicted flavors. Lastly,some recently proposed NLP model such as GPT-3 has showed an outstanding ability in few-shot learning.[14] We intrigued the ability of GPT-3 in flavor prediction and formulated a testing query to text-davinci-002 engine. In result, we found out that GPT-3 engine is capable to understand the context and generated flavors in the sentence level; however, the generated flavors seemed oversimplified, and it urged the need to propose a new model.

## 3 METHODOLOGY

In this section, we describe the detailed methodology of our proposed work.

### 3.1 Data Collection

Over the years, though coffee has become one of the worlds most consuming beverages, the predication of the coffee flavor still remained as an unpopular task in deep learning, and we have found only limited coffee dataset publicly accessible. One of the most popular coffee database called as coffee quality databse is gathered by coffee quality institute CQI.[17]. The CQI database contains the collection of the detailed coffee beans information. This CQI dataset is the largest coffee dataset that is available publicly with around 1300 entries by merging the two different popular coffee species (arabica and robusta) and contains the labels of coffee features that includes species, Owner, Country of Origin, Farm Name, Lot Number, Mill, ICO Number, Company, Altitude, Region, Producer, Number of Bags, Bag Weight, In Country Partner, Harvest Year, Grading Date, Owner , Variety, Processing Method, Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Cupper Points, Total Cup Points, Moisture, Category One Defects, Quakers, Color, Category Two Defects, Expiration, Certification Body, Certification Address, Certification Contact, unit_of_measurement, altitude_low_meters, altitude_high_meters, altitude_mean_meters. However, these coffee information labels that describes the sensory characteristics[12](Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean.Cup, Sweetness) of the coffee does not provide the user with the its textual description but rather it simply provides the predicted mathematical score out of 10 based on the training processing of the supervised labels. More importantly, There is no single dataset that is publicly available providing the textual description of the coffee flavors considering several coffee features as the labels provided in the coffee quality institute data set.

The flavors of the coffee is one of the most significant driving force for several coffee industries however the prediction of the coffee flavor remains as a more complex problem due to the limited availability of the coffee professionals. Even with the availability of the coffee experts, there is a possibility of the biased and time-consuming data resulting from the coffee professionals which provides an inefficient coffee information results. In. general, this remains as a common problem for any model dependent with the humans for predicting the result. Hence in this paper, we propose a novel idea of predicting the coffee flavor in a textual description using a novel deep learning technique. We understood the need to create a data set that can provide the result of the coffee flavors along with other essential factors to overcome the limitations of human dependencies and other existing bias problems. Initially, we performed an excessive search of the coffee websites search for extracting the coffee features and other information that are useful for creating the reliable coffee data set for predicting the coffee flavors. We identified the coffee websites that includes the royalcoffee [4], onyxcoffeelab[2], blackwhiteroasters[1], thecaptainscoffee[5], yce[6]. However, these coffee websites does not contain thousands of data that is generally required for the model training. The royalcoffee[4] contains 1073 coffee data entries, thecaptaincoffee[5] website contains 40 entries and the others having only around 5 to 10 coffee data entries. However to perform the analysis on the real existing data, we considered the data from the two major websites containing some sufficient data of reasonable coffee related information and those includes the royalcoffee[4] having 1073 coffee data entries, thecaptaincoffee[5] having 40 entries. Though the coffee data entries are limited, there are several coffee features that are available for each of these entries. The coffee data of the royalcoffee[4] webiste includes the features of Grower, Variety, Region, Harvest, Altitude, Soil, Process, Certifications and Flavors. The coffee data of the thecaptaincoffee[5] website includes Arrival Date, Acidity & Brightness, Balance & Finish, Body & Texture, Flavors, Grade, Processing, Grower, Region, Varietals, Roasting. We collected these coffee features by implementing a python script that can perform the web scrapping. The webscrapping is achieved with the help of the most suitable BeautifulSoup python library. We gathered all the coffee related information from each of the websites and stored the result in a JSON file that is fed to the next Data Preprocessing stage.

## 3.2 Data Preprocessing

The data that are unreliable and irrelevant can aim to produce misleading results and can highly affect the model accuracy and efficiency. In order to ensure better model performance, we performed the data preprocessing of the dataset generated in section 3.1. We identified that the coffee features are not uniform across all the websites as in general the site owners corresponding to each of the coffee webistes would have their own design style of web page creation. The coffee information or features available in some of the coffee websites is not available in the other coffee websites. This gave us an idea to decide the important features of the coffee beverage based on the the accessibility of the coffee features that are available across all the websites. We identified the features that are most important, useful are existing across the available websites and are consistent in the occurrence of the coffee features in the provided coffee websites. Finally, we concluded that the following coffee features that have to be considered for our data set those include coffee cultivated country region, coffee growing altitude above the sea level, variety of the coffee, processing method of the coffee and finally the flavors of the coffee. We decided that these coffee features are the most important features for the prediction of the flavors. We have eliminated the unreliable features that includes Grower, Harvest, Soil, Certifications from the royalcoffee[4] raw dataset and Arrival Date, Acidity & Brightness, Balance & Finish, Body & Texture, Grower, Roasting from the thecaptaincoffee[5] raw dataset. We also noticed the inconsistency in the naming and text similarity across the coffee features in the dataset that we generated in section 3.1. The coffee features of the thecaptaincoffee[5] website that includes Varietals, Processing, Grade contains the textual naming mismatch with the Variety, Process, Altitude coffee feature in royalcoffee[4] respectively. We implemented the python script to remove all these inconsistencies, missing, redundant and unreliable data and created the dataset that is uniform and consistent across all the coffee data entries to obtain accurate results from our proposed model. Removing these inconsistencies also prevents the over-fitting of our DeepCoffee model and serves to provide faster performance and more accurate results. Hence using the python script, we created the data set with the following labels for the coffee features: Region, Altitude, Variety, Processing method and Flavors.

## 3.3 Data Modeling

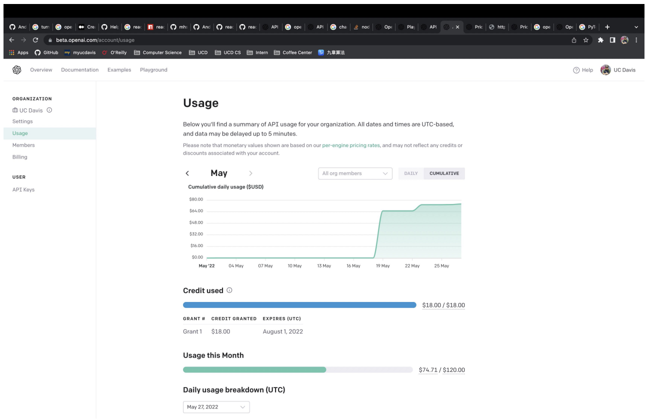
GPT-3[14], a transformer based model has demonstrated a profound performance in few-shot learning, and some recent studies[40] has demonstrated more satisfying result in zero-shot performance using instruction-tuning. Coffee Flavors prediction in another term requires the language model to understand a context involving geographic location, altitude, coffee species, processing methods, and generates vocabularies or sentences to perform sensory evaluation(describing flavor).

The OpenAI API[3] provides a wide set of training language models that are suitable for the intentions of the users while making them to stay more truthful, and not being biased, toxic, or otherwise harmful[33]. The models perform a wide variety of of natural language tasks that are suitable for different levels of tasks as well as the ability to fine-tune the models to improve the ability of the overall model performance. In general, the language models provided by openAI API[3] are significant in understanding and generating the text.

Prompts and completions - The openAI provides a simple interface to the models that are extremely powerful and flexible. The users are required to provide some text input as a "prompt" and the model will generate a text completion that matches with that input prompt provided by the user[3]. The text completion is the core of openAI API. The text completion can be moderately considered as an auto complete feature, content generation, expansion and summarization in which the model predicts the result based on the user input included in the prompt[3].

Tokens - It is very important for the models to understand the model input for processing and providing the results more accurately. The openAI provides the method for the models to understand and process the input text from the user by breaking the text into tokens[3]. The number of tokens for processing depends on both the length of the text inputs and outputs provided by the user. One token is considered to be approximately 4 characters or 0.75 words for English text[3]. The openAI imposed a limitation that on including both the text prompt and the generated model completion, the token size to exceed no more than the model's maximum context length[3]. The openAI contains most of the models provided with the limit in its maximum context length which is around 2048 tokens, or about 1500 words approximately[3].

However, there is a major drawback of openAI models. The models of the openAI includes the price points as shown in Figure 4. The price of the model training increases with the increase in the size of the dataset as the number of tokens increases and hence the openAI model is not affordable for the datasets with a large number of textual words. Currently, GPT-3 is not a open source model and also it requires hundreds of GPUs that is not possible to run on local computer and requires a superpower computer. Microsoft paid for most of the training costs (several million dollars) and optioned to exercise their contractual rights to exclusively license source code. Though we have the other transformed model GPT-2 which is a open-source model and can be run on a local machine, the performance of the model is poor. Hence we opted to use GPT-3 for our experiment to not compromise on the results for our work and that overall costed us around $90 pricing charge for using the model of openAI as shown in Figure 4 and Figure 2.



Usage

Below you'll find a summary of API usage for your organization. All dates and times are UTC-based, and data may be delayed up to 5 minutes.

Please note that monetary values shown are based on our per-engine pricing rates, and may not reflect any credits or discounts associated with your account.

Credit used

$18.00 / $18.00

| GRANT # | CREDIT GRANTED | EXPIRES (UTC) |
|---|---|---|
| Grant 1 | $18.00 | August 1, 2022 |

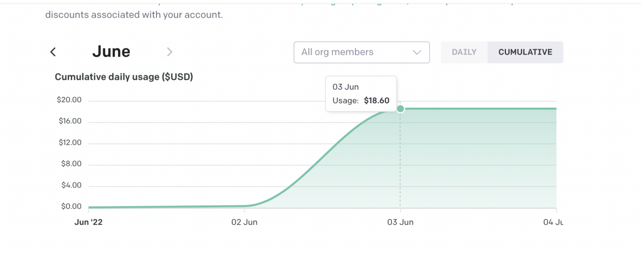
Usage this Month

$74.71 / $120.00

Daily usage breakdown (UTC)

May 27, 2022

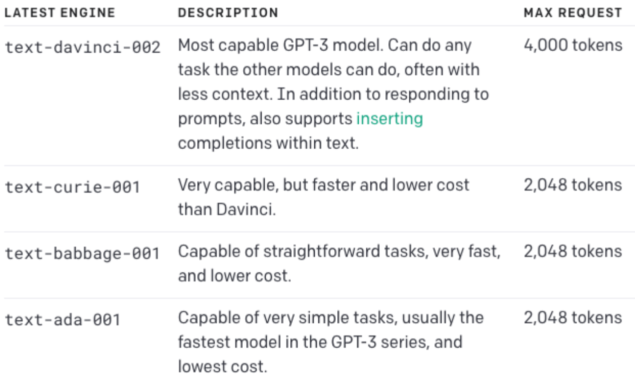
**Figure 1: OpenAI Fine-tuned GPT-3 Pricing**



discounts associated with your account.

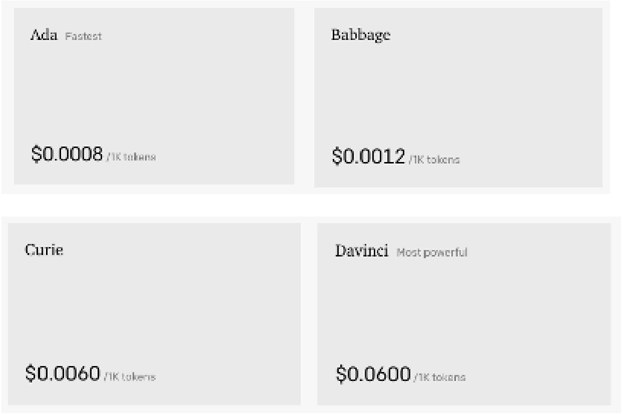
**Figure 2: OpenAI models comparison Pricing**

The openAI provides four different models that are suitable for different level of tasks as shown in Figure 3 along with the other base models davinci, curie, babbage, and ada shown in Figure 4. We used "Davinci" with fine-tuning that is highly suitable in performing the complex tasks and understanding the content that requires a lot of understanding and computing resources. Fine-tuning the model provides various advantages as follows: Increased capability of solving more complex problems with cause and effect Highly accurate and efficient results Ability to perform the few-shot learning on more inputs that can fit in a prompt Limits the token size with the short text prompt



| LATEST ENGINE | DESCRIPTION | MAX REQUEST |
|---|---|---|
| text-davinci-002 | Most capable GPT-3 model. Can do any task the other models can do, often with less context. In addition to responding to prompts, also supports inserting completions within text. | 4,000 tokens |
| text-curie-001 | Very capable, but faster and lower cost than Davinci. | 2,048 tokens |
| text-babbage-001 | Capable of straightforward tasks, very fast, and lower cost. | 2,048 tokens |
| text-ada-001 | Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost. | 2,048 tokens |

**Figure 3: Model description with token size**



Ada Fastest

$0.0008 /1K tokens

Babbage

$0.0012 /1K tokens

Curie

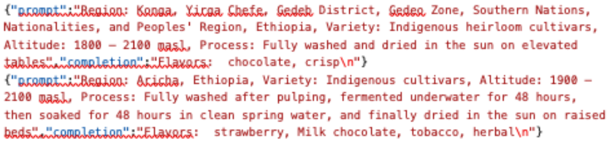
$0.0060 /1K tokens

Davinci Most powerful

$0.0600 /1K tokens

**Figure 4: Pricing of the base models**

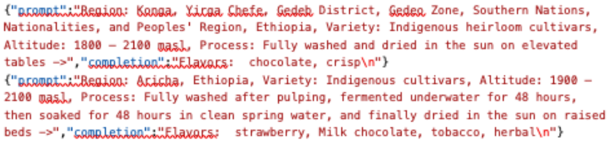
We installed the OpenAI command-line interface (CLI) for setting up the Davinci model using the Package Installer for Python (PIP)[3]. The openAI model provided the ability to fine-tune specified tasks that can be aimed to achieve best performance results on wide number of tasks[3]. We loaded our API key in the variable OPENAI_API_KEY for setting up the local environment for the model processing[3].The key variable is the environment variable provided by the openAI . Followed by the setup, we implemented the python script to customise the preprocessed dataset created in section 3.2 to the prompt-completion pair in JSON format as shown in Figure 5 for preparing the training data as shown in Figure 6.



```
{"prompt":"Region: Konga, Yirga Chefe, Gedeb District, Gedeo Zone, Southern Nations,
Nationalities, and Peoples' Region, Ethiopia, Variety: Indigenous heirloom cultivars,
Altitude: 1800 – 2100 masl, Process: Fully washed and dried in the sun on elevated
tables","completion":"Flavors:  chocolate, crisp\n"}
{"prompt":"Region: Aricha, Ethiopia, Variety: Indigenous cultivars, Altitude: 1900 –
2100 masl, Process: Fully washed after pulping, fermented underwater for 48 hours,
then soaked for 48 hours in clean spring water, and finally dried in the sun on raised
beds","completion":"Flavors:  strawberry, Milk chocolate, tobacco, herbal\n"}
```

**Figure 5: Preprocessed data in JSON format**

The openAI imposes the requirement of the training data for their provided models in JSONL format. In order to make the proposed requirement easier, the openAI developed a CLI data preparation tool that provides the users to directly reformat the dataset[3]. The dataset in different formats such as CSV, TSV, XLSX, JSON or JSONL file formats can be converted into a JSONL with the only requirement that the input file format should contain a prompt and a completion column/key[3] using openAI CLI data preparation tool. This allows the users to save time, effort and avoid human errors. We prepared the training data by converting our JSON file into JSONL file as shown in Figure 6. A string "->" has been added to the end of the prompt for indicating the model to start generating completions, rather than continuing with the prompt[3].



```
{"prompt":"Region: Konga, Yirga Chefe, Gedeb District, Gedeo Zone, Southern Nations,
Nationalities, and Peoples' Region, Ethiopia, Variety: Indigenous heirloom cultivars,
Altitude: 1800 – 2100 masl, Process: Fully washed and dried in the sun on elevated
tables ->","completion":"Flavors:  chocolate, crisp\n"}
{"prompt":"Region: Aricha, Ethiopia, Variety: Indigenous cultivars, Altitude: 1900 –
2100 masl, Process: Fully washed after pulping, fermented underwater for 48 hours,
then soaked for 48 hours in clean spring water, and finally dried in the sun on raised
beds ->","completion":"Flavors:  strawberry, Milk chocolate, tobacco, herbal\n"}
```

**Figure 6: Training Data in JSONL format**

We performed our model fine-tuning process using the highly capable Davinci model by creating two different datasets. The first dataset contains the real data of coffee related information that is obtained from the websites as described in section 3.1 with around 1300 coffee data entries. We have introduced the second dataset for the two main reasons. First, GPT-3 is a very large language model and can take thousands of entries for evaluating the model accuracy and so we created the second dataset with 10000 entries to study the behavior of GPT-3 model performance. Second, coffee flavor prediction is a variable data and so we decided to explore the model prediction of the coffee flavors by creating a large number of additional entries and to study on how our. DeepCoffee model behaves with the changes in the coffee features. For example, we would like to explore the flavors when a coffee grown altitude is changed in a different region or what if the the processing method of a region is changed. Does the coffee flavors change ? Whether our DeepCoffee model can provide any new innovations that can drive the coffee industries ?. Hence, we created 10000 entries including the real dataset entries by implementing a python script. The script is written in such a way that it creates the 10000 entries using only the coffee feature values from the real dataset entries to avoid non-existing, fake, toxic and unrelaiable in our dataset. We then evaluated the performance of the model on two different datasets. We split each of our datset into 80% training data and 20% testing data and achieved a good model accuracy of 83.5%.

## 3.4 Web Application

As GPT-3[14] has demonstrated a SOTA performance in many NLP tasks, such large language model have enormous potential in the real world application. DeepCoffee built upon those abilities, and tuned to understand professional specialty coffee context. We were surprised by the performance of the model in its knowledge of coffee producing regions though we did not specifically designed our dataset while testing, and then we decided to utilize its capacity in building an interactive Web Application.

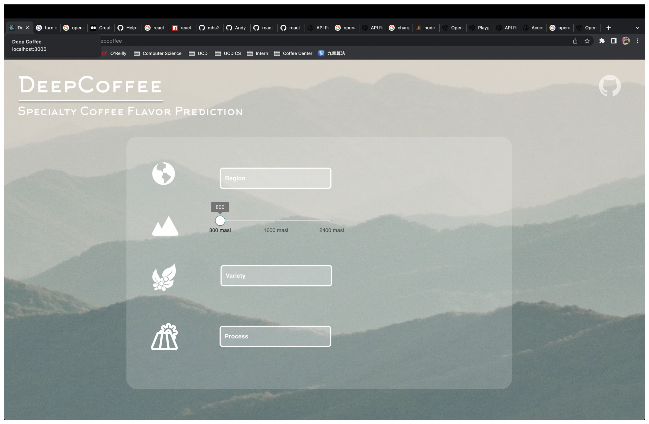
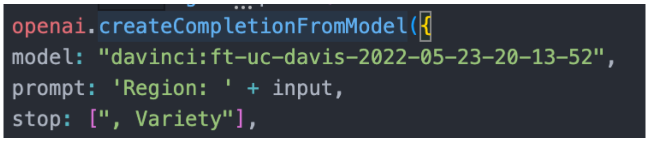


**Figure 7: Web application for coffee flavor prediction**

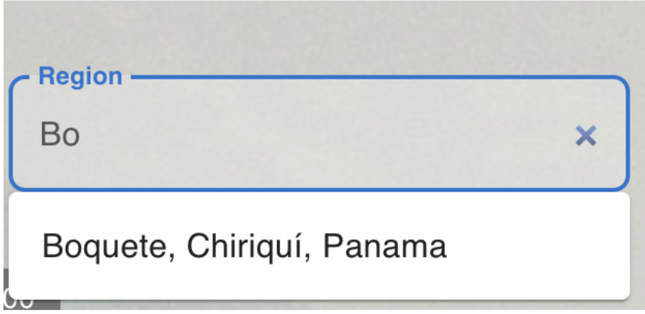
Empowering by existing OpenAI API on Node.js, we chose to use react.js as our front-end framework in building the website. Except for other design elements, the main input frame in which user input prompts to DeepCoffee contains four major categories: Region, Altitude, Variety, and Process. Those four major categories corresponded to the input query we trained our DeepCoffee model, and should find a great prediction by the model. For Altitude, we have regulated the scale to be 800 - 2400 masl which corresponded to the highest and lowest altitude we have seen when we collected our dataset, and which the optimal growing altitude for overall Arabica Coffee Trees. Moreover, for the Variety and Processing, we chose to add few representing examples(Heirloom, Typica, and etc. in Variety and Naturla, Washed, Honey, and etc in Process) in the autocomplete bar but also allow users to input text other than we the existing ones.

*3.4.1 Autocomplete.* The major novelty we contributed to user input frame, however, is at the region text field input experience. During our training process we found that DeepCoffee is able to predict accurate regions with user inputted only few characters without any spcific training on coffee regions dataset. The observation then became our stimuli to develop autocomplete functionality on region input. We have developed a structured query8 in order to let the model generate response on predicted input based on few characters.



```
openai.createCompletionFromModel({
model: "davinci:ft-uc-davis-2022-05-23-20-13-52",
prompt: 'Region: ' + input,
stop: [", Variety"],
```

**Figure 8: Structured Requesting Query**

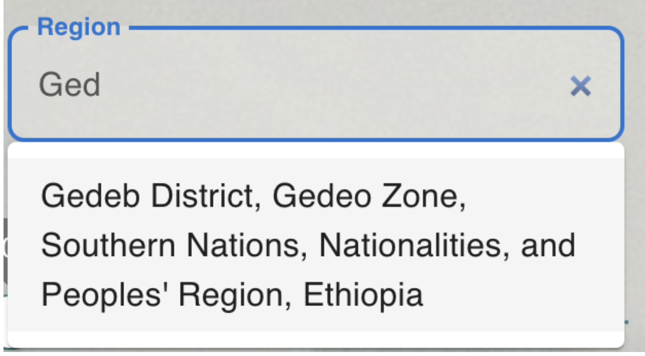


Region

Bo

Boquete, Chiriquí, Panama

**Figure 9: Autocomplete Region**



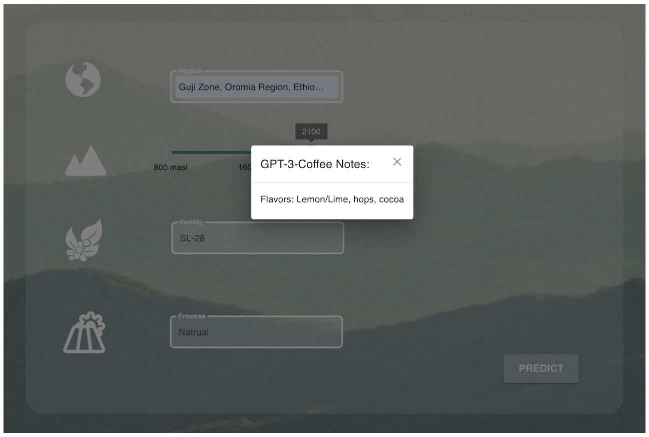
Region

Ged

Gedeb District, Gedeo Zone, Southern Nations, Nationalities, and Peoples' Region, Ethiopia

**Figure 10: Autocomplete Region**

*3.4.2 Flavor Prediction.* After user inputted all 4 text input, we then organized user inputs in the similar constuct as training dataset as prompts to feed DeepCoffee. We have tested some example we found on specialty coffee roasters to the model, and find a great demonstration of flavors that similar to what the roasters have wrote.11 However, we did not perform systematic evaluation of the accuracy of the model at this point.



Region: Guji Zone, Oromia Region, Ethio…

GPT-3-Coffee Notes:

Flavors: Lemon/Lime, hops, cocoa

Variety: SL-28

Process: Natrual

**Figure 11: Web Result of the Coffee flavors Prediction**

## 4 RESULTS

In this section, we provide our evaluation results as an evidence for our work.

We conducted the experiments of fine-tuning GPT-3 model on two different datasets as described in section 3.3. We compared the training accuracy and loss of the two datsets having the 1300 entries and 10000 entries to study our DeepCoffee(GPT-3 + fine-tuning) model behavior with scaling the size of the entries in the dataset. The training accuracy of our DeepCoffee model with the real 1300 entries dataset shows a highly significant result of 91.3% as can be seen in Figure 12 while achieving a low loss as shown in Figure 13.

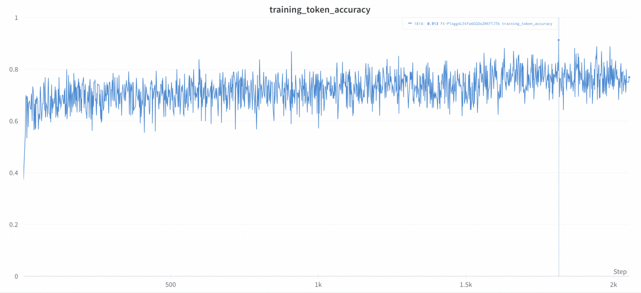


**Figure 12: Training accuracy of the real dataset with 1300 entries**

The training accuracy of our DeepCoffee model with the 10000 entries in the coffee dataset (that is created using the randomness from the real dataset containing 1300 entries) is evaluated to study the performance of the model with higher number of data entries in the dataset and that the model shows a result of 77.1% training accuracy as shown in Figure 14 and the training loss as shown in Figure 15.

We could observe the decrease in the training accuracy however this is not conclusive for determining the training accuracy of our DeepCoffee model due to the randomness introduced in in the generation of the dataset. Due to the price charge by openAI in the number of tokens used in the dataset. This costed us with around 74.71 dollars as shown in Figure 10.

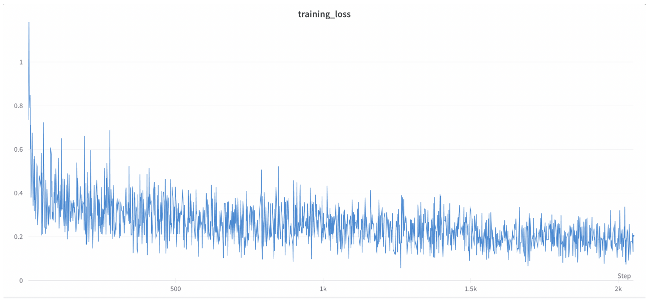


Figure 13: Training loss of the real dataset with 1300 entries

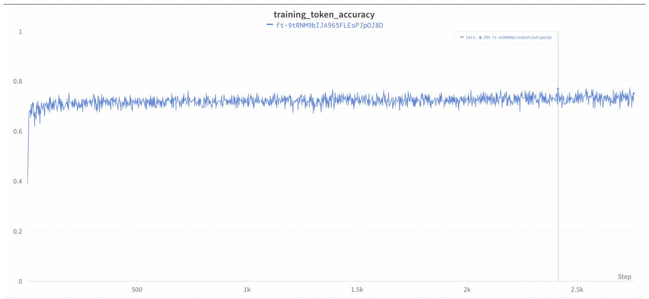


Figure 14: Accuracy with training large dataset

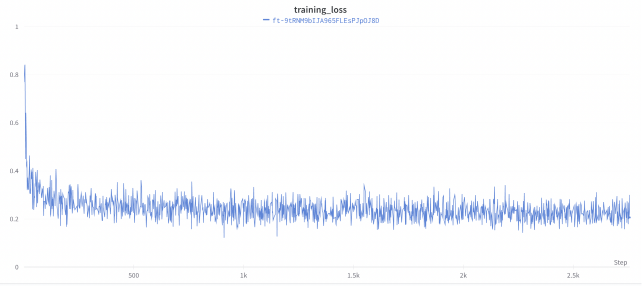


Figure 15: Training loss of the large dataset with 10000 entries

Figure 16 provides the comparison of the training accuracy of 1300 entries dataset and 10000 entries dataset.Figure 17 provides the comparison of the training loss of 1300 entries dataset and 10000 entries dataset.

We decided to study the performance of the model over the real dataset containing 1300 entries for providing the accurate results without any toxic and unreliable data that can impact the model behavior. We divided the data set with real 1300 coffee information data entries into 80% training data and 20% testing data i.e in the ratio of 4:1 such that the training data contains 1040 entries and testing data contains 260 entries. We compared our Deepcoffee (Davinci+ Fine-tuning) model with the other models that includes text-ada-001, text-babbage-001, text-curie-001, text-davinci-002 provided by the OpenAI. Our DeepCoffee model achieved a high accuracy of 83.5% over the other OpenAI models with text-davinci-002 resulting in 73.2% accuracy followed by text-ada-001 with 66.8% accuracy, text-curie-001 with 64.2% accuracy , text-babbage-001 63.2% accuracy as shown in Figure 18. However with the price points of the models in openAI, the comparison with the other charged us an additional cost of 18.60 dollars as shown in figure2

Overall, the experiments prove that our DeepCoffee model (Davinci with Finetuning) are the coffee experts that can provide the more accurate predication of the the coffee flavors with the given coffee features region, altitude, variety and finally the processing method.

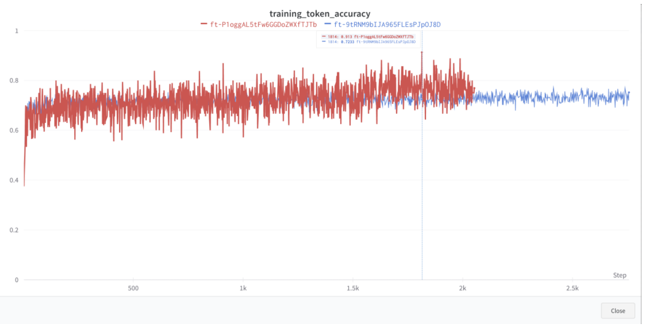


Figure 16: Accuracy comparison with training real and large dataset

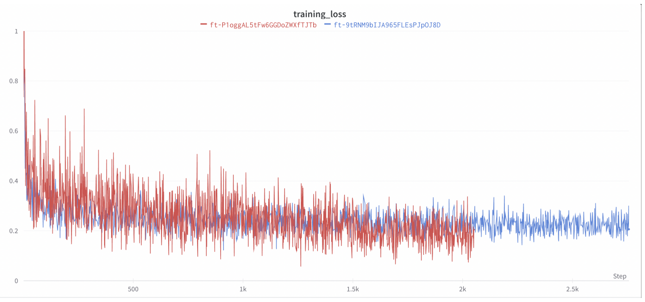


Figure 17: Training loss comparison of the real dataset with 1300 entries in red and 10000 entries in blue

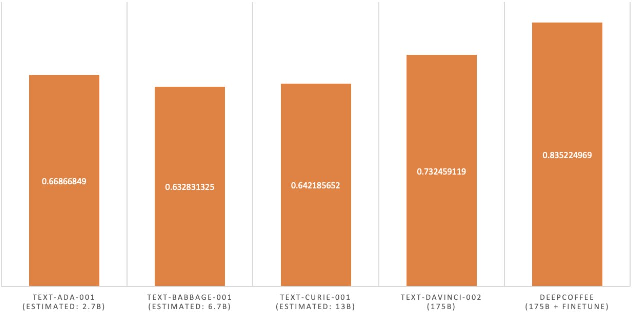


Figure 18: Testing accuracy between different model

## 5 LIMITATIONS AND FUTURE WORK

Due to the limited availability and consistency of the coffee related information across the websites, we considered only five important coffee features (that includes Region, Altitude, Variety, Processing method and Flavors) for the prediction of the coffee flavors and we would like to further explore our DeepCoffee model behavior by largely scaling the coffee features. One of the major limitation of our DeepCoffee is the pricing for the model usage being GPT-3 not an open source model. Overall, the work on fine-tiuning our DeepCoffee model costed us around $90. The model is still not fully reliable as they still can generate biased or toxic text, or simply not following user instructions[33]. We would like to further research the complications and other problems with GPT-3 not being open sourced apart from the computing resources. We currently used the GPT-3 and fine-tuning approach by the openAI and we would like to explore our future research on creating a new model design for predicting the coffee flavors while achieving better performance and high accuracy.

## 6 CONCLUSION

We have shown our model DeepCoffee, a GPT-3 based transformer model with the fine-tuning can better predict the flavors of the coffee. We presented our DeepCoffee model performance with two different dataset scaling the tokens. We provided our model comparison with the other GPT-3 based models that includes ada, curie and babbage and showed that our model provides the flavor prediction result with a higher accuracy. Finally, we implemented a web application which is GUI-based that allows the users to provide the input to the coffee features region, variety, altitude and processing method and obtain the coffee flavors as the output with an easy interfacing and a good visualisation.

## 7 CODE

https://github.com/Andy-LZH/GPT-3-Coffee

## ACKNOWLEDGMENTS

We would like to thank our Instructor Prof. Jiawei Zhang and teaching assistant Xiao Liu for providing valuable inputs, feedback and support on this work.

## REFERENCES

[1] [n. d.]. blackwhiteroasters database. https://www.blackwhiteroasters.com/collections/frontpage. Accessed: 2022-06-02.

[2] [n. d.]. onyxcoffeelab database. https://onyxcoffeelab.com/collections/coffee. Accessed: 2022-06-02.

[3] [n. d.]. openai website. https://beta.openai.com/docs/. Accessed: 2022-06-02.

[4] [n. d.]. royalcoffee database. https://royalcoffee.com/offerings/. Accessed: 2022-06-02.

[5] [n. d.]. thecaptainscoffee database. https://thecaptainscoffee.com/. Accessed: 2022-06-02.

[6] [n. d.]. yce database. http://yce.coffeeauction.org. Accessed: 2022-06-02.

[7] Masayuki Akiyama, Kazuya MURAKAMI, Michio IKEDA, Keiji IWATSUKI, Sadayuki KOKUBO, Akira WADA, Katsuya TOKUNO, Masanobu ONISHI, Hisakatsu IWABUCHI, and Kiyofumi TANAKA. 2005. Characterization of flavor compounds released during grinding of roasted robusta coffee beans. *Food Science and Technology Research* 11, 3 (2005), 298–307.

[8] Radford Alec, Narasimhan Karthik, Salimans Tim, and Sutskever Ilya. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI* (2018).

[9] Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, and Edward Grefenstette. 2018. Learning to understand goal specifications by modelling reward. *arXiv preprint arXiv:1806.01946* (2018).

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[11] Natnicha Bhumiratana, Koushik Adhikari, and Edgar Chambers IV. 2011. Evolution of sensory aroma attributes from coffee beans to brewed coffee. *LWT-Food Science and Technology* 44, 10 (2011), 2185–2192.

[12] Natnicha Bhumiratana, Mona Wolf, Edgar Chambers IV, and Koushik Adhikari. 2019. Coffee drinking and emotions: Are there key sensory drivers for emotions? *Beverages* 5, 2 (2019), 27.

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[15] Yu-Tang Chang, Meng-Chien Hsueh, Shu-Pin Hung, Juin-Ming Lu, Jia-Hung Peng, and Shih-Fang Chen. 2021. Prediction of specialty coffee flavors based on near-infrared spectra using machine-and deep-learning methods. *Journal of the Science of Food and Agriculture* 101, 11 (2021), 4705–4714.

[16] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[17] Diego Volpatto CQI. [n. d.]. Coffee Quality database from CQI. https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi. Accessed: 2022-04-27.

[18] Suzana Reis Evangelista, Maria Gabriela da Cruz Pedrozo Miguel, Cecília de Souza Cordeiro, Cristina Ferreira Silva, Ana Carla Marques Pinheiro, and Rosane Freitas Schwan. 2014. Inoculation of starter cultures in a semi-dry coffee (Coffea arabica) fermentation process. *Food Microbiology* 44 (2014), 87–95.

[19] Adriana Farah. 2012. Coffee constituents. *Coffee* (2012), 21–58.

[20] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[21] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[22] Thierry Joët, Andréina Laffargue, Frédéric Descroix, Sylvie Doulbeau, Benoît Bertrand, Stéphane Dussert, et al. 2010. Influence of environmental factors, wet processing and their interactions on the biochemical composition of green Arabica coffee beans. *Food chemistry* 118, 3 (2010), 693–701.

[23] Daneysa Lahis Kalschne, Marcelo Caldeira Viegas, Antonio José De Conti, Marinês Paula Corso, and Marta de Toledo Benassi. 2018. Steam pressure treatment of defective Coffea canephora beans improves the volatile profile and sensory acceptance of roasted coffee blends. *Food Research International* 105 (2018), 393–402.

[24] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659* (2021).

[25] Ji Yoon Kim. 2022. Coffee Beans Quality Prediction Using Machine Learning. *Available at SSRN 4024785* (2022).

[26] Kenji Kumazawa and Hideki Masuda. 2003. Investigation of the change in the flavor of a coffee drink during heat processing. *Journal of Agricultural and Food Chemistry* 51, 9 (2003), 2674–2678.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[28] J Devlin M Chang K Lee and K Toutanova. 2018. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[29] Ji-Yoon Lee and Young-Seob Jeong. 2022. Prediction of Defect Coffee Beans Using CNN. In *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 202–205.

[30] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871* (2018).

[31] Kazuya Murakami, Masayuki Akiyama, Masahiro Sumi, Michio Ikeda, Keiji Iwatsuki, Osamu Nishimura, and Kenji Kumazawa. 2010. Differences in flavor characteristics of coffee drinks originating from thermal sterilization process. *Food science and technology research* 16, 2 (2010), 99–110.

[32] International Coffee Organization. 2019. Total production by all exporting countries.

[33] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[34] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[35] Denis Richard Seninde and Edgar Chambers. 2020. Coffee flavor: A review. *Beverages* 6, 3 (2020), 44.

[36] Sushant Singh and Ausif Mahmood. 2021. The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access* 9 (2021), 68675–68702. https://doi.org/10.1109/ACCESS.2021.3077350

[37] Wenny B Sunarharum, David J Williams, and Heather E Smyth. 2014. Complexity of coffee flavor: A compositional and sensory perspective. *Food Research International* 62 (2014), 315–325.

[38] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503* (2021).

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[40] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

[41] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from Language Models. *arXiv preprint arXiv:2112.04359* (2021).

[42] Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015* (2019).