LEAF: Predicting the Environmental Impact of Food Products based on their Name

Bas Krahmer
Independent Researcher
baskrahmer@gmail.com

Abstract

Although food consumption represents a substantial global source of greenhouse gas emissions, assessing the environmental impact of off-the-shelf products remains challenging. Currently, this information is often unavailable, hindering informed consumer decisions when grocery shopping. The present work introduces a new set of models called **LEAF**, which stands for **L**inguistic **E**nvironmental **A**nalysis of **F**ood Products. LEAF models predict the life-cycle environmental impact of food products based on their name. It is shown that LEAF models can accurately predict the environmental impact based on just the product name in a multilingual setting, greatly outperforming zero-shot classification methods. Models of varying sizes and capabilities are released, along with the code and dataset to fully reproduce the study.

1 Introduction

Reducing global greenhouse gas emissions is a key objective for mitigating rapid climate change. Recent estimates based on life-cycle assessment (LCA) data say the global food system accounts for up to 30% of global greenhouse gas emissions (Li et al., 2022). Although one can not completely eliminate the emissions from food consumption, the environmental impact can be greatly reduced by avoiding foods with a high climate impact. For most food products however, this information is not readily available, which makes it more difficult for consumers to make informed decisions[1].

In this work, a new set of models is introduced which can predict the environmental impact of a product based on the product name. These models learn relationships from products with existing LCA data, which can subsequently be applied to any text.

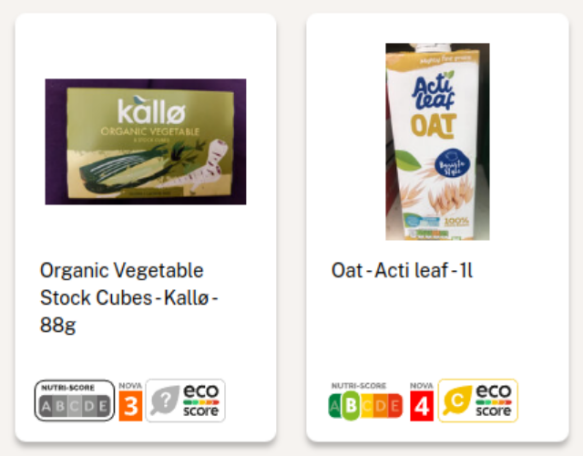

Figure 1: Two example products from the Open Food Facts platform. The left product has no Eco-Score, whereas the score for the product on the right is known.

1.1 Related Work

In recent years a handful of studies have described models that predict certain aspects of food products. Hu et al. (2023) use a BERT model for food classification of Canadian branded products in the context of nutrition. Balaji et al. (2023) do emissions estimation of general consumer products using zero-shot classification based on a sentence BERT model.

The Open Food Facts (OFF) data (Section 2.1) is an excellent resource for the current work. OFF has developed a computer vision model called Robotoff (2024(b)) which predicts missing data fields like category, weight and brand based on the product image. These predictions are subsequently verified in a crowd-sourcing environment called Hunger Games (2024(a)).

To the best of available knowledge, the current study is the first work exploring the usage of NLP methods specifically for the estimation of environmental impact of food products.

[1] On Open Food Facts, 73% of products have no Eco-Score

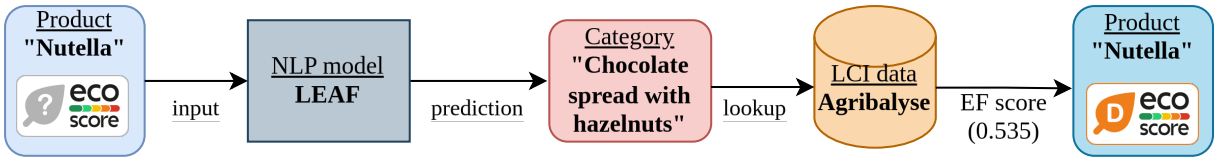

Figure 2: Conceptual diagram of LEAF. A product with an unknown Eco-Score can be processed by LEAF to make a category prediction. This category is linked to the Agribalyse database to fetch an EF score, which can be discretized into an Eco-Score.

2 Methods

2.1 Open Food Facts Dataset

The OFF database consists of open-access, crowdsourced products. With 40% of products, the dominant language of the database is French, followed by English (32%) and Spanish (10%). Although the OFF platform is mostly popular in France, the website has dedicated pages for a large variety of countries and territories, and contributions to the database are made by consumers and producers globally. The platform has quality control measures like community review and error detection when uploading a product. The product entries vary in completeness, where some products have a detailed list of ingredients and others just a picture and a name. A more elaborate data analysis can be found in Appendix A.

The dataset is filtered for products that have an associated Agribalyse class (Colomb et al., 2015). In total there are 2518 Agribalyse classes present in the dataset, each of which has an associated life-cycle assessment (LCA) that estimates the environmental impact measured as the environmental footprint (EF) score (Colomb et al., 2015; European Commission, 2021). The EF score is a weighted combination of 14 different factors, expressed in millipoint (mPt) per kilogram of product. EF score factors are related to the full life-cycle of a product, including manufacturing, packaging, transport, consumption and disposal. The biggest contributing factor in this score is the climate impact, measured in carbon dioxide equivalent or CO_2Eq (Brander and Davis, 2012). This unit is also used to compute the Eco-Score (Facts, 2023) as displayed on the OFF website (Figure 1). While the concept of a discrete A/B/C/D/E rating system for the Eco-Score is similar to that of the widely-adopted Nutri-Score (Chantal et al., 2017), the formulae and methodology behind the two label values differ.

The OFF database primarily gathers its data through crowdsourcing. The dataset is licensed under the Open Database License (ODBL) (Open Data Commons). The dataset used for this work was exported on March 31st 2024.

2.2 Task and Models

The current work introduces a set of models to predict the EF score of a product based on its name. A high-level task overview is given in Figure 2.

LEAF models consist of a pretrained sentence embedding base model and a readout head. The `distiluse-multilingual-base-v2` (DU) transformer model (135M parameters) (Reimers and Gurevych, 2019) and a larger `bge-m3` (M3) transformer model (561M parameters) (Chen et al., 2024) are used due to their cross-lingual capabilities where semantically similar texts across languages are nearby in vector space. The parameters of the base model are frozen and not fine-tuned; instead the static sentence embeddings serve as input for the learnable task-specific heads. Three different LEAF model configurations are introduced, based on their unique readout heads:

- **LEAF$_\mathbf{c}$**: Standard classification head comprising a dense layer with 2518 output nodes, followed by a softmax function and optimised by a cross-entropy loss function.
- **LEAF$_\mathbf{r}$**: Regression head that predicts a single continuous value using a dense layer, followed by a softplus function that maps the output to a non-zero positive value and optimised by a MSE loss function.
- **LEAF$_\mathbf{h}$**: Hybrid head that combines the classification head and the regression head in a sequential way, such that each logit can contribute individually to the resultant regression value. Both the logits and the regression output are simultaneously optimized, with a hyperparameter α (set to 0.5 in experiments) controlling the weight of the individual loss terms (details in Appendix B.1).

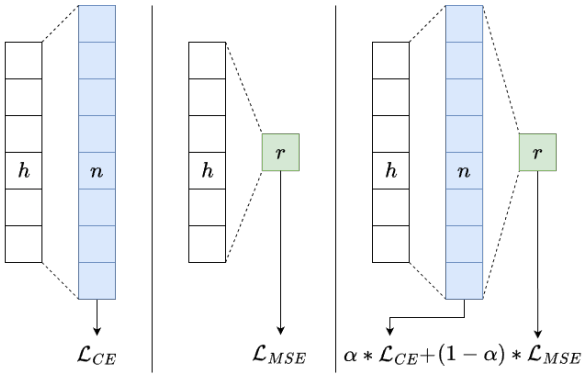

Figure 3: Readout heads of different model configurations. From left to right: LEAF_c, LEAF_r, LEAF_h. $\mathcal{L}_{CE}$ and $\mathcal{L}_{MSE}$ denote cross-entropy and mean-squared error loss respectively. Dimensionalities are denoted as h (hidden size, 768), n (number of classes, 2518) and r (regression, 1).

The different model configurations are illustrated in Figure 3.

2.3 Experimental Setup

Models are evaluated based on the predicted deviance from the ground truth EF score, as measured by the mean absolute error (MAE). For the classification model, the EF score is computed by mapping the predicted class to the corresponding value. For the other model heads, the predicted EF score is the actual regressand.

Models are tested on a holdout set comprising 20% of the original training data. The test set is randomly sampled using stratified sampling based on the product language, such that the language distribution of the test set reflects that of the training set. Metrics are micro-averaged across all data points. For each LEAF configuration, a grid search is performed to identify good values for the batch size and learning rate (parameters in Table 5). The grid searches use the smaller DU embedding model, and the best model configuration is also trained using the M3 embedding model.

2.4 Baseline Models

LEAF models are evaluated against two different types of baseline models. The first baseline model is a zero-shot autoregressive classifier using the OpenAI `gpt-3.5-turbo` model API (175B parameters) (Brown et al., 2020). For budgetary reasons, this model is chosen over the more powerful `gpt-4-turbo` variant and the test sample is deliberately smaller to reduce API costs. The model achieves an accuracy of 0.374 and a MAE of 0.110 on a random sample of 1000 products when ignoring any hallucinated class predictions. When considering hallucinations as random guesses, the accuracy is corrected to 0.302. Optimising and paraphrasing the textual prompt has negligible impact on performance. Details of the exact baseline methodology are supplied in Appendix B.2.

In addition, an untrained DU and M3 are evaluated on the entire test dataset. Predictions are obtained by constructing an embedding table by embedding each of the class names. Given an embedded product name, its predicted class is defined by taking the class with the lowest cosine similarity. Although the M3 model with CLS pooling achieves the best open-source baseline performance with an accuracy of 0.193 and a MAE of 0.300, it does not outperform the OpenAI baseline model. Interestingly, the mean-pooled DU embedding models seem to substantially outperform CLS-pooled ones.

3 Results

The accuracy and MAE values of different models are summarised in Table 1. There is a clear performance gap between classification and regression models, where classifying products in concrete classes seems to result in higher accuracies and lower MAE scores than predicting a continuous value. Among classification models, LEAF_c outperforms the OpenAI baseline despite having about a thousand times less parameters. Although the accuracy of the OpenAI model is substantially lower than that of LEAF_h, the MAE of OpenAI is actually better, which implies that misclassifications of the OpenAI model are typically within a smaller error bound than those of the more accurate LEAF_h model. A sample-based qualitative comparison of the OpenAI baseline and LEAF_c models supports this claim (details in Appendix C.1). Note that the MAE score of 0.071 is considerably smaller than the global dataset standard deviation of 0.448, and that a more granular class distribution like the Eco-Score is less sensitive to small numerical deviations.

Ablation studies are performed on the best-performing LEAF_c model configuration to gain deeper insights into which configurations contribute to it's performance. The results are summarized in Table 2. Firstly, it is noteworthy that the M3 model brings an increased performance versus the DU model in a classification setting. This can be explained partly by the larger parameter count

Model	Accuracy	MAE
LEAF_c	**0.731**	**0.071**
LEAF_r	N/A	0.233
LEAF_h	0.696	0.224
$\text{Cosine}_{DU,CLS}$	0.057	0.406
$\text{Cosine}_{DU,mean}$	0.109	0.356
$\text{Cosine}_{M3,CLS}$	0.193	0.300
$\text{Cosine}_{M3,mean}$	0.193	0.301
$\text{OpenAI}_{GPT-3.5}$	0.374	0.110

Table 1: Test set results for best-performing grid search configurations and baseline models. The LEAF_r accuracy is not available since the model only produces a single numeric output. Note that the OpenAI metrics are for a subset of 1000 valid test set predictions, ignoring any hallucinated class predictions.

Ablation	Accuracy	MAE
$\text{LEAF}_{c,M3}$	**0.772**	**0.057**
$\text{LEAF}_{c,CLS}$	0.720	0.075
$\text{LEAF}_{c,LLFT}$	0.364	0.196

Table 2: Test set results for other classification configuration models. CLS denotes using CLS-pooled embeddings; M3 denotes using the M3 base model instead of DU; LLFT denotes finetuning of the last layer (at 0.1 times the learning rate).

of the base model (561M parameters versus 135M). Secondly, using CLS pooling, we observe a slight drop in performance compared to the mean-pooled configuration, which is also as expected considering the baseline scores for the DU models. Lastly, finetuning the last attention layer of the DU model while training the classifier results in a sharp performance decrease. This can have several causes, but it seems likely that the model starts overfitting the attention mechanism to the task at hand, losing meaningful capabilities attained during pretraining (Ramasesh et al., 2021).

An analysis of multilingual performance shows that $\text{LEAF}_{\text{c, M3}}$ performs best for the top-5 languages in the dataset. The results can be found in Tables 9 and 10 of Appendix C.2.

4 Conclusion

Modern NLP methods can accurately predict the environmental impact of food products using only their names across various languages. Empirical evidence shows that classification is preferred over regression for estimating EF scores. Among different model configurations, LEAF_c models substantially outperform others in both accuracy and error, and they also surpass GPT-3.5 in a zero-shot classification setting. All in all, predictions based on the product name are a simple yet powerful approach, and LEAF models can be considered for tasks like Eco-Score prediction.

5 Limitations and Future Work

While LEAF offers a novel approach to predict EF scores for a variety of products, this work has certain limitations which are transparently outlined to raise awareness and encourage further research.

Limited Class Specificity: There are no individual differences within an Agribalyse class. For example, an apple belongs to the apple class, regardless of whether that apple is produced locally or overseas. Various factors influencing a product's environmental impact are abstracted and averaged out, although the difference can be significant. Future work can address this by e.g. using more fine-grained LCA data or by working on EF score explainability.

Fixed Consumption Location: Current models assume the product is consumed in France, as per Agribalyse assumptions. The effects of certain large emission factors, such as transportation, are location-specific and substantially contribute to greenhouse gas emissions (Li et al., 2022). Therefore, caution is needed when interpreting LEAF results for locations with significantly different food supply chains than France.

Additional Data Sources. The current work examines the relationship between product name and environmental impact. Other (potentially unlabelled) data sources, such as ingredient lists, country of production, country of consumption, transportation method, and packaging data, could provide additional insights for more accurate environmental impact predictions. A new model that combines different data sources under varying levels of uncertainty could be superior.

Processing of LCA Data. The current dataset has redundancy among certain classes. For example, there are three classes for almonds (peeled, unpeeled, and salted), all with the same LCA values. A compressed mapping for a new class distribution specific to EF score estimation could improve stability and performance by reducing the parameter count.

References

Bharathan Balaji, Venkata Sai Gargeya Vunnava, Geoffrey Guest, and Jared Kramer. 2023. Caml: Carbon footprinting of household products with zero-shot semantic text similarity. In *Proceedings of the ACM Web Conference 2023*, pages 4004–4014.

Matthew Brander and Gary Davis. 2012. Greenhouse gases, co2, co2e, and carbon: What do all these terms mean. *Econometrica, White Papers*.

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*, 33:1877–1901.

Julia Chantal, Serge Hercberg, World Health Organization, et al. 2017. Development of a new front-of-pack nutrition label in france: the five-colour nutri-score. *Public health panorama*, 3(04):712–725.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Vincent Colomb, Samy Ait-Amar, Claudine Basset-Mens, Armelle Gac, Gérard Gaillard, Peter Koch, Jerome Mousset, Thibault Salou, Aurélie Tailleur, and Hayo MG Van Der Werf. 2015. Agribalyse®, the french lci database for agricultural products: high quality data for producers and environmental labelling.

Directorate-General for Environment European Commission. 2021. Commission recommendation (eu) 2021/2279 of 15 december 2021 on the use of the environmental footprint methods to measure and communicate the life cycle environmental performance of products and organisations.

Open Food Facts. 2023. Eco-score: The environmental impact of food products. Open Food Facts.

Open Food Facts. 2024(a). Hunger games. Open Food Facts Wiki.

Open Food Facts. 2024(b). Robotoff: Machine learning for food label insights. GitHub repository.

Guanlan Hu, Mavra Ahmed, and Mary R L'Abbé. 2023. Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. *The American Journal of Clinical Nutrition*, 117(3):553–563.

Mengyu Li, Nanfei Jia, Manfred Lenzen, Arunima Malik, Liyuan Wei, Yutong Jin, and David Raubenheimer. 2022. Global food-miles account for nearly 20% of total food-systems emissions. *Nature Food*, 3(6):445–453.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what art? a call for multi-prompt llm evaluation. *arXiv preprint arXiv:2401.00595*.

Open Data Commons. Open database license (odbl).

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2021. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

SIL International. 2023. Ethnologue: Languages of the world.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Class name	Frequency	EF score
Camembert	33017	0.485
Biscuit (cookie)	25451	0.345
Honey	18638	0.175
Yogurt	18478	0.220
Tea	16140	0.013

Table 3: The five most frequent Agribalyse classes in the dataset. Full class names for shortened names are "Camembert cheese, from cow's milk", "Yogurt, fermented milk or dairy specialty, plain" and "Tea, brewed, without sugar".

A Open Food Facts Data Analysis

The OFF dataset has a skewed class distribution and a skewed language distribution with French being its most prevalent language.

In total, there are 800,589 products in the dataset. Products have an average EF score of 0.448 ($\sigma = 0.454$) and percentile values of 0.175, 0.310 and 0.588 for the 25th, 50th and 75th percentiles respectively. The 5 most-common classes are summarised in Table 3. The 5 products with the highest EF score all relate to lamb meat prepared in different ways, with a EF score of 5.813. The top-1 percentile of highest EF score products ($n = 25$) have an average EF score of 5.243 ($\sigma = 0.594$) and are composed of different types of meat ($n = 21$), different types of seafood ($n = 3$) and decaf instant coffee ($n = 1$).

B Supplemental Information on Methods

B.1 Hybrid Loss Function

In total, three different loss functions are implemented to train the LEAF models, of which one is a custom implementation. The cross-entropy loss and mean-squared error loss are common loss functions for training classification and regression models respectively. The hybrid loss function is defined as:

$$\mathcal{L}_h = \alpha * \mathcal{L}_{CE} + (1 - \alpha) * \mathcal{L}_{MSE} \qquad (1)$$

Where α is a non-negative value between 0 and 1. Although α could be a learnable parameter, this could lead the model to finding a local optimum by learning a value close to either 0 or 1 and thereby eliminating one of the loss terms. All experiments in the current work have a constant α value of 0.5.

B.2 OpenAI Baseline Methodology

The OpenAI model was evaluated on a limited set of 1236 samples due to financial constraints. Of these, the model was able to make a valid prediction for 1000 samples, indicating a hallucination rate of 0.191. For these samples, the model had an accuracy of 0.374. When considering the invalid predictions as random guesses on a balanced dataset, the model achieves an accuracy of 0.302. The context length of this model is 16k tokens, which is insufficient to encode a single sample in one prompt including the model instructions, product name and all possible classes. To accommodate for the context window limit, the total number of classes are split into two random partitions for each test sample. The model in total does 3 classifications per sample; one for each of the two partitions, and another one to choose between the two partition classifications. The categories and partition splits are randomly shuffled for each sample. The generation temperature is set to 0 and the seed is set to 42. API calls were made on April 28th 2024 using the OpenAI Python SDK.

The system prompt consists of the following text:

> You are a helpful assistant. Your task is to classify the text string given by the user. The string can be presented in any language. You must pick a class from the permitted categories you are provided, even if the correct class is not in the list.

B.2.1 Handling Hallucinations

A significant challenge of the baseline model is its tendency to hallucinate new class labels that would fit the sample. If the model hallucinates a non-existent category in one of the partitions, the other category is automatically picked as the predicted class. If both partition categories are hallucinated, the prediction is rendered invalid. Baseline metrics are provided for two scenarios: one where invalid predictions are considered random guesses and one where they are not considered. The probability that the model hallucinates is likely higher for more difficult samples, so it is important to interpret both numbers. Hallucinations can be prevented by limiting token generation to the possible class names, but to the best of available knowledge this is not currently supported in the OpenAI API.

System prompt	Accuracy	MAE	HR
Original	0.325	**0.120**	**0.160**
Paraphrased	**0.345**	0.122	0.197
Minimal	0.340	0.123	0.222
Linguist	0.303	0.131	0.251
Environmentalist	0.330	0.133	0.216

Table 4: Results for 200 valid predictions given different system prompts, where HR denotes hallucination rate.

B.2.2 System Prompt Sensitivity

Since it has been shown that prompting can have a significant impact on model performance (Mizrahi et al., 2023; White et al., 2023), an additional experiment is performed to establish the sensitivity of `gpt-3.5-turbo` to system prompt variability for the classification task. Apart from the forementioned original system prompt, the following system prompts were evaluated:

1. **Paraphrased**: You are an assistant dedicated to providing support. Your objective is to categorize the text provided by the user. This text may be in any language. You must choose a category from the allowed list of options, even if the most appropriate category isn't included.

2. **Minimal**: Your task is to classify the text string given by the user. The string can be presented in any language. You must pick the correct class from the list of permitted categories, even if the correct class is not in the list.

3. **Linguist**: You are an expert linguist and text classifier. Your task is to classify the text string given by the user. The string can be presented in any language. You must pick the correct class from the list of permitted categories, even if the correct class is not in the list.

4. **Environmentalist**: You are an expert in assessing the environmental impact of food products. Your task is to classify the text string given by the user. The string can be presented in any language. You must pick the correct class from the list of permitted categories, even if the correct class is not in the list.

The methodology mostly is unchanged except for a reduced sample size (200 valid predictions per prompt instead of 1000) and the experiment time (API calls are made on June 24$^{\text{th}}$ 2024 instead of April 28$^{\text{th}}$). The results are summarised in Table 4. Firstly, the original assistant prompt has a lower accuracy of 0.345 ($n = 200$) than in the earlier experiment where it had an accuracy of 0.375 ($n = 1000$). This can be a result of intermittent OpenAI model updates that made the model slightly worse on the task. It can also be due to difference in sample size. Furthermore, it can be observed that the paraphrased prompt has a slightly higher accuracy than other variants, and that the original prompt has a slightly lower MAE and hallucination rate. It seems sensible that hallucinations are more likely for ambiguous (and therefore harder) samples, making it hard to objectively assess the efficacy of individual prompts. Nevertheless, prompt variation empirically has a negligible effect on overall performance, where LEAF models yield a substantially better performance than zero-shot classification models based on GPT-3.5. The accuracy can be further improved by utilising a more powerful language model or formulating as a few-shot classification task.

Parameter	Values
Peak Learning Rate	{1e-3, 5e-3, 1e-2, 5e-2}
Batch Size	{64, 128, 256}
Sequence Length	32
Pooling Mode	Mean
Warm up Steps	10k
Training Steps	100k
Weight decay	0.01
Gradient Clipping	None
Precision	FP32
Learning Rate Decay	Linear
Adam β_1	0.9
Adam β_1	0.999
Adam ϵ	1e-8

Table 5: LEAF training hyperparameters, where curly brackets denote the set of values used for grid searches.

B.3 Training Hyperparameters

The grid search hyperparameters are summarised in Table 5. Only the readout head, peak learning rate and batch size are varied among models. The random seed is set to a constant value of 42. For LEAF_c, LEAF_r and LEAF_h, the best observed peak learning rate values are 256, 64 and 256 respectively. Similarly, the best observed batch sizes are 5e-2, 5e-3 and 5e-3 respectively.

For M3 base models, it is observed that a lower learning rate is required than for LEAF_c using a DU base model. Based on 3 training runs with learning rates of 5e-2, 8e-3 and 5e-3, it is observed that 5e-3 performs best. For the LLFT ablation run, a classifier learning rate of 5e-2 and a attention layer learning rate of 5e-3 were used.

B.4 Reproducibility

The codebase is available on GitHub[2], including scripts for dataset creation and model training. The trained LEAF_c with DU and M3 base models are available on the Hugging Face model hub under the aliases `baskra/leaf-base` and `baskra/leaf-large` respectively. The train and test datasets are available on the Hugging Face datasets hub under the alias `baskra/leaf`.

C Additional Results

C.1 Qualitative Analysis

A random sample of test set predictions is analysed to compare qualitative differences between the OpenAI baseline and LEAF_c with DU and M3 base models, across each of the major languages in the dataset.

Table 6 shows misclassifications from DU where GPT-3.5 predicts the correct class. It is observed that GPT-3.5 misclassifications are generally more precise than the misclassifications of DU, which are relatively coarse. For example, GPT-3.5 correctly classifies the product name "Kräuteressig" as *Vinegar*, whereas DU misclassifies it as *Camembert*. This is notably in line with the observed metrics: while GPT-3.5 has a worse overall accuracy, it achieves a better MAE compared to DU. Conversely, Table 7 shows misclassifications of GPT-3.5 where DU predicts the correct class. Here is seen that GPT-3.5 misclassifies the product name "Chipolata aux herbes au sel de l'ile de re" as *Sausage*, which, while correct, is not as specific as the ground-truth *Chipolata* class.

Lastly, a comparison between LEAF_c variants is made in Table 8, which shows misclassifications of DU where M3 predicts the correct class. Here one can see that both coarse and precise misclassifications of DU are not made by M3, indicating that the larger M3 model is overall more accurate than DU.

[2]URL: https://github.com/baskrahmer/LEAF

C.2 Multilingual Performance

Metrics for the most-frequent languages in the dataset are present in Table 9. In addition, metrics for the most-spoken languages (according to (SIL International, 2023)) are present in Table 10.

Sample product name	Language	$LEAF_{c,\ DU}$ prediction	Ground truth
Curaçao bleu	French	Camembert cheese, from cow's milk	Liqueur
Compotée de Cerises	French	Jam, cherry	Fruits compote, miscellaneous
the noir lapsang Souchong	French	Soy sauce, prepacked	Black tea, brewed, without sugar
Antiuxixona, milk chocolate	English	Milk, semi-skimmed, UHT	Milk chocolate bar
Natural Sharp Cheddar Cheese	English	Camembert cheese, from cow's milk	Cheddar cheese, from cow's milk
Graham teddy bears	English	Candies, all types	Biscuit (cookie)
Maíz palomitas	Spanish	Camembert cheese, from cow's milk	Pop-corn or oil popped maize, salted
8 Fettine di formaggio fuso	Italian	Camembert cheese, from cow's milk	Processed cheese with fresh cream cheese and walnuts
Kräuteressig	German	Camembert cheese, from cow's milk	Vinegar

Table 6: Sample of $LEAF_{c,\ DU}$ misclassifications where GPT-3.5 accurately predicts the correct category (ground truth column). Predictions are randomly sampled from the test set across major languages.

Sample product name	Language	GPT-3.5 prediction	Ground truth
Filet de maquereau fumé au poivre	French	Mackerel, smoked	Mackerel, canned in brine, drained
Chipolata aux herbes au sel de l'ile de re	French	Sausage meat, pork and beef, raw	Chipolata slim sausage, raw
Comte AOP	French	Tomme cheese, from cow's milk	Comté cheese, from cow's milk
Unsweetened applesauce	English	Apple, pulp, raw	Apple compote
Mint green tea with japanese matcha tea bags	English	Green tea, brewed, without sugar	Tea, brewed, without sugar
Golden vegetable rice	English	Rice, mix of species (white, wholegrain, wild, red,etc.), raw	Rice, parboiled, raw
Garbanzos	Spanish	Chick pea, cooked	Chick pea, canned, drained
Banane	Italian	Banana, pulp, raw	Plantain banana, raw
Makrelenfilets	German	Mackerel, fillet, in white wine, canned, drained	Mackerel, canned in brine, drained

Table 7: Sample of GPT-3.5 misclassifications where $LEAF_{c,\ DU}$ accurately predicts the correct category (ground truth column). Predictions are randomly sampled from the test set across major languages.

Sample product name	Language	$\text{LEAF}_{\text{c, DU}}$ prediction	Ground truth
Confiture extra de griottes	French	Jam, strawberry	Jam, cherry
Goûters Noisette	French	Hazelnut	Biscuit (cookie)
Saint Émilien GrandCru 2014	French	Wine, red	Wine, white, dry
Kreams gold orange	English	Marmalade, orange	Biscuit (cookie)
Bolachas de Água e Sal	English	Salt, white, for human consumption (sea, igneous or rock), no enrichment	Wafer biscuit, crunchy (thin or dry), plain or with sugar, prepacked
Adnams southwold dry hopped lager	English	Dry sausage	Beer, dark
Galleta espelta de arandanos y manzana	Spanish	Muesli, flakes (Bircher-style)	Biscuit (cookie)
Burrata di buffala	Italian	Turkey, meat and skin, raw	Camembert cheese, from cow's milk
Porridge mit Vollkornhafer Beerentrio	German	Beer, dark	Breakfast cereals, mix of puffed or extruded cereals, fortified with vitamins and chemical elements

Table 8: Sample of $\text{LEAF}_{\text{c, DU}}$ misclassifications where $\text{LEAF}_{\text{c, M3}}$ accurately predicts the correct category (ground truth column). Predictions are randomly sampled from the test set across major languages.

Model	Acc_{fr}	MAE_{fr}	Acc_{en}	MAE_{en}	Acc_{es}	MAE_{es}	Acc_{it}	MAE_{it}	Acc_{de}	MAE_{de}
$\text{LEAF}_{c,DU}$	0.764	0.065	0.760	0.058	0.716	0.076	0.730	0.068	0.630	0.098
$\text{LEAF}_{c,M3}$	0.799	0.050	0.781	0.050	0.769	0.060	0.772	0.059	0.705	0.076
LEAF_{r}	N/A	0.249	N/A	0.190	N/A	0.243	N/A	0.227	N/A	0.231
LEAF_{h}	0.724	0.236	0.741	0.184	0.676	0.240	0.697	0.222	0.579	0.231
$\text{N}_{samples}$	81568		35312		14428		9005		11602	

Table 9: Performance for 5 most frequent languages in the dataset (fr=French, en=English, es=Spanish, it=Italian, de=German)

Model	Acc_{zh}	MAE_{zh}	Acc_{ar}	MAE_{ar}	Acc_{hi}	MAE_{hi}	Acc_{bn}	MAE_{bn}	Acc_{pt}	MAE_{pt}
$\text{LEAF}_{c,DU}$	0.397	0.108	0.375	0.137	0.0	0.033	0.333	0.168	0.500	0.136
$\text{LEAF}_{c,M3}$	0.381	0.197	0.417	0.155	0.0	0.788	0.0	0.142	0.584	0.098
LEAF_{r}	N/A	0.272	N/A	0.240	N/A	0.228	N/A	0.195	N/A	0.235
LEAF_{h}	0.452	0.244	0.440	0.223	0.0	0.355	0.0	0.067	0.486	0.229
$\text{N}_{samples}$	42		59		1		3		933	

Table 10: Performance for 5 most spoken languages globally (zh=Chinese, ar=Arabic, hi=Hindi, bn=Bengali, pt=Portuguese)