Automatic Product Classification in International Trade: Machine Learning and Large Language Models*

Ignacio Marra de Artiñano
Université Libre de Bruxelles and ECARES
Franco Riottini Depetris
Inter-American Development Bank and UdeSA
Christian Volpe Martincus
Inter-American Development Bank and CESifo

This version: November 2023

Abstract

Accurately classifying products is essential in international trade. Virtually all countries categorize products into tariff lines using the Harmonized System (HS) nomenclature for both statistical and duty collection purposes. In this paper, we apply and assess several different algorithms to automatically classify products based on text descriptions. To do so, we use agricultural product descriptions from different public agencies, including customs authorities and the United States Department of Agriculture (USDA). We find that while traditional machine learning (ML) models tend to perform well within the dataset in which they were trained, their precision drops dramatically when implemented outside of it. In contrast, large language models (LLMs) such as GPT 3.5 and GPT 4 show a consistently good performance across all datasets, with accuracy rates ranging between 60% and 90% depending on HS aggregation levels. Our analysis highlights the valuable role that artificial intelligence (AI) can play in facilitating product classification at scale and, more generally, in enhancing the categorization of unstructured data.

Keywords: Product Classification, Machine Learning, Large Language Models, Trade
JEL Codes: F10, C55, C81, C88

*We would like to thank Peter Schott for valuable comments and suggestions. We are also grateful to Victoria Patience for her careful edition and María Lidia Viquez for her excellent assistance in the publication process. The views and interpretations in this paper are strictly those of the authors and should not be attributed to the Inter-American Development Bank, its executive directors, or its member countries.
Contact: Christian Volpe Martincus: christianv@iadb.org.

1 Introduction

Accurately classifying products is essential in international trade. Virtually all countries use the Harmonized System (HS) nomenclature to categorize products into tariff lines for both statistical and duty collection purposes. Misclassification, both intentional and unintentional, can be very costly. It can result in imprecise measurement of trade flows, inappropriate determination of origin, foregone duty collection, inadequate application of restrictions or prohibitions and significant delays to border monitoring and processing times. Furthermore, it can lead to the design and implementation of misguided trade policies, specially those related with trade remedies such as countervailing duties, antidumping, and safeguards.

Traditionally, the bulk of product categorization tasks has been carried out manually, frequently based on experts' judgments, and has accordingly been extremely time-consuming.[1] As a consequence, classification is challenging for governments, firms, and researchers, especially on a large scale. This has been magnified by the rise of cross-border e-commerce, which requires customs agencies to process several million small shipments per year. In many developing countries, this has generally resulted in most shipments being classified based on their value or size instead of the specific goods they consist of, thus limiting their ability to conduct risk assessments properly and that of their countries to accurately measure the composition of a growing portion of their international trade. Firms, in turn, particularly those that are small or have no previous experience in international trade, typically find it difficult to assign their products to HS codes and need to rely on costly specialized services to do so.[2] Last but certainly not least, various databases that could potentially provide inputs for novel, policy-relevant research report valuable product-level information through product names or text descriptions. This makes it hard for researchers to combine them, leading

[1] As highlighted by public customs' agencies resolutions, classification is often the object of firms' ex-ante consultations and is subject to ex-post adjustments.

[2] Furthermore, in an effort to reduce the wrong attribution of tariff lines, custom agencies often impose heavy misreporting fines, which can be burdensome for exporters.

to imperfect merges with standard trade databases based on the HS nomenclatures. In this paper, we look at an important example of one such attempt to classify products into the HS nomenclature using product descriptions.

The advent of machine learning (ML) is likely to reduce these classification efforts and increase their accuracy (see WCO, 2022a).[3] While there is an incipient literature that aims to assess the precision of ML for product classification, most existing models rely on tests on the same dataset used to train them. As a consequence, there is very limited evidence on how these models perform on real external datasets and hence on their general applicability. Further, such evidence is missing altogether in the case of large language models (LLMs), which are yet to be tested at scale for this purpose.

In this paper, we examine the performance of a variety of ML models along with LLMs (GPT-3.5 and GPT-4), at classifying products according to the HS nomenclature at different aggregation levels.[4] In doing so, we will go beyond the train-and-test dataset and thus explicitly assess the external validity of the models. For this, we will use three different datasets: (i) a dataset containing product descriptions from the Chilean customs agency to train and test ML algorithms, following earlier literature; (ii) a dataset containing product descriptions from the customs agency of a different country, Paraguay; and (iii) a database of product descriptions from the United States Department of Agriculture (USDA).[5] This third data source describes products for which firms obtain an organic certification. In all cases, our analysis will be limited to animal, vegetable, and food products, since these are the product categories for which firms can obtain organic certification (Marra de Artiñano et al., 2023).[6]

Our results reveal that, while standard ML algorithms performed very well within the

[3]The BACUDA project run by the World Customs Organization (WCO) is an example of ongoing work using these techniques for customs applications.

[4]We acknowledge that there are several other possible approaches to automatic product classification, including convoluted neural networks (CNN), recurrent neural networks (RNN) and other transformer-based ML models. In this paper, we restrict ourselves to some the most widely used and practitioner-friendly ML and LLMs.

[5]We use Paraguayan customs data because, like Chilean data, they are publicly available.

[6]This project was originally conceived with the objective to match product descriptions of organic certified firms with their corresponding HS Codes. In a future study, we will extend the analysis to all HS products.

test set, their accuracy dropped dramatically when these models were applied to datasets on which they were not explicitly trained. In contrast, GPT-3.5 and GPT-4 performed very evenly across all datasets.[7] Its accuracy was relatively high: it achieved percentages of approximately 60%—70% at the HS 6-digit level (highly granular product nomenclature), 70%—80% at the HS 4-digit level, and 80%—95% at the HS 2-digit level.[8] Perhaps surprisingly, due to its significantly larger parameter size and training set, we find that GPT-4 does not achieve higher accuracy in all three databases, but only in two of them.[9]

There are several important applications for this sort of scalable automatic product classification that uses product descriptions as inputs. First, it could help customs agencies identify patterns of intentional or fraudulent product miscategorization. Second, it would make it easier for both policymakers and researchers to categorize product descriptions from unstructured data sources (such as those obtained from e-commerce transactions or from historical sources) using established product nomenclatures. Finally, it could be used to develop chatbots that give HS code suggestions from simple text descriptions, which would greatly facilitate tariff line attribution for firms engaged in international trade and even consumers participating in cross-border e-commerce [10].

We make three main contributions to the existing literature. Overall, to the best of our knowledge, this study is the first to apply GPT to the WCO's HS product classification and, more generally, to a large multiclass classification problem in economics.[11].

[7] In addition, we use a "nested" version of GPT 3.5. In this case, we first prompted the model to assign a 2-digit HS code based on the product description. Subsequently, we asked the model to provide a 4-digit HS code among those that belong to the aggregate 2-digit HS code category that it previously identified. Finally, we requested the model to provide a 6-digit HS code belonging to the prior 4-digit HS code that it had identified. If different at all, this nested version of GPT 3.5 performed slightly worse than the 'unrestricted' counterpart, thus suggesting that LLMs might underperform in some product classification tasks when their answers are constrained. These results are available from the authors upon request.

[8] We also tested the performance of GPT 3.5 in mapping sector descriptions onto the North American Industry Classification System (NAICS). To do so, we used sectors reported by firms when registering with the online business platform *ConnectAmericas* The results indicate that the GPT-3.5 model achieves an efficiency of more than 60% at the HS 6-digit level. These results are available from the authors upon request.

[9] The parameters in LLMs are the number of variables that a model adjusts during training, typically the weights within neural network layers. A larger parameter space thus implies higher capacity to adjust to a variety of linguistic patterns and subtleties and hence to react to different prompts.

[10] The US Census has already developed similar chatbot, see https://uscensus.prod.3ceonline.com/.

[11] Kocon et al. (2023) carries out a simpler classification exercise focused on only a few categories.

A number of previous studies have proposed alternative approaches to automatically classify products into HS codes across a large number of tariff lines. Spichakova and Haav (2020) use ML methods to provide 6-digit HS code predictions and recommendations using a model trained with product descriptions from the United States Bill of Lading 2017 database. They show that the algorithm achieves a hit rate of 80% on the test dataset. Ruder (2020) uses a variety of ML and deep learning models to classify product descriptions from the US Bill of Lading and reaches accuracy levels of approximately 60%. Chen et al. (2021) apply unsupervised ML and an off-the-shelf embedding encoder to automatically assess whether reported HS codes in cross-border import declarations are correct. They achieve an overall success rate of 71% on an HS 6-digit dataset provided by Dutch customs. Turhan et al. (2015) adopt a different strategy whereby they use visual properties along with product labels and descriptions. The accuracy level they achieve is above 80% with 4-digit HS codes from a database of 4,494 binding tariffs published by the European Union in 2014. These papers use a single dataset, which is split into training and testing samples. Unfortunately, this approach does not allow the accuracy of the models on external datasets to be tested. This limitation is crucial because tariff databases often have significantly different product descriptions and text formats. One exception in this regard is He et al. (2021), who use data gathered directly from firms to train their models, along with a second dataset of product descriptions from a third firm that was not in the test dataset. However, they focus on very few HS products (12 6-digit potential product classifications) and their exercise is accordingly much simpler than product categorization across the universe of potential tariff lines.

We contribute to this literature on automatic product classification by assessing the accuracy of different ML algorithms on both the test-train-split dataset and two additional datasets for a large set of products. Our results indicate a very large decrease in the accuracy of standard ML algorithms outside the dataset on which the models are trained.

There is also a recent literature that aims to apply GPT and other LLM models to text-based data in the social sciences. Some recent papers that use GPT include Hansen et al. (2023), Lopez-Lira and Tang (2023), Hansen and Kazinnik (2023), Yang and Menczer (2023)

and Ko and Lee (2023).[12] Hansen et al. (2023) compare the performance of a predecessor of GPT-3 to their own model, WHAM, and find that WHAM outperforms GPT-3 in terms of the error rate at the task of classifying whether a job posting allowed the possibility of remote work at least one day per week. The authors also discuss the potential gains of adopting modern natural language processing (NLP) methods for text classification in economic environments. They suggest that other prediction problems using text in economics might similarly benefit from a large training sample combined with sequence embedding models, such as GPT-3.

Lopez-Lira and Tang (2023) examine the potential of ChatGPT (GPT 3.5) in predicting stock market returns by using analysis and the classification of news with potential impact for firms. Their analysis suggests that, even though ChatGPT (GPT 3.5) is not specifically trained for this task, it produces superior results in terms of predicting stock market returns than other traditional sentiment analysis methods commonly used in finance due to the comprehensiveness of the model. In a similar vein, Ko and Lee (2023) show that ChatGPT effectively helps improve portfolio management by selecting asset classes that statistically outperform random choices in diversification and returns.

Hansen and Kazinnik (2023) use GPT-3 and GPT-4 to decipher Fedspeak, the language used by the Federal Reserve to communicate monetary policy decisions. Their results suggest that these models obtain the lowest numerical errors, the highest accuracy rates, and the highest measure of agreement relative to human classification when compared to other pretrained linguistic models and dictionary-based approaches. Finally, Yang and Menczer (2023) use ChatGPT to study the credibility of news and conclude that they are able to correctly evaluate news sources by rating them.

We add to these papers by showing the usefulness of LLMs for product classification in international trade. We find that while GPT-3.5 and GPT-4 perform slightly worse than traditional ML algorithms on the test-train-split dataset, it significantly outperforms these models

[12] An exhaustive analysis of the recent literature using GPT (and its adjacent models) is beyond the scope of this paper. Nevertheless, it is worth mentioning papers such as Noy and Zhang (2023) on the effects on productivity, Biswas (2023) on its potential role in health, and Kasneci et al. (2023) on its potential impact on education.

on external databases. The reason is that LLMs are able to go beyond the specific context of the training dataset and thus have much higher external validity. Unlike traditional ML algorithms, they also require no additional data-cleaning or preprocessing, making them much simpler to use.

The rest of this paper is structured as follows. Section 2 describes the different data sources used in our analysis. Section 3 explains the methodological approach. Section 4 discusses the results of the classification process for the different databases. Finally, Section 5 concludes with a brief discussion of our results.

2 Data

In this paper, we used three different datasets: a database of product descriptions from Chilean customs, a database of product descriptions from Paraguayan customs, and a database of organic product descriptions from USDA. The first database (Chilean customs) was used to train and test the ML algorithms. The second database (Paraguayan customs) was employed to test the external validity of our models. Finally, the third database (USDA) was used to further test the models outside the context of customs product descriptions.

2.1 Train-Test-Split Dataset: Trade Transactions from the Chilean Customs

To generate and train the ML models that attempt to predict the HS nomenclator for a set of target products, we used the universe of Chilean export and import transactions between 2009 and 2021 as our train-and-test dataset. This comprehensive dataset contains more than 104 million observations, with granular information on trade transactions, including granular HS codes and detailed product descriptions. As is usual in the literature, we split this dataset into separate training and testing subsets. The training data set was used to develop and refine our models, whereas the test dataset was used to assess their performance and accuracy.

We focused our analysis on HS Chapters 1–22, which encompass agricultural, animal,

and food products. As mentioned in Section 1, our ultimate objective in this work was to accurately classify organic product descriptions into HS product nomenclatures, and thus we exclusively trained and tested in the categories these products are found in. To keep the computational load manageable, we randomly selected 1 million product descriptions in these HS chapters from the Chilean customs dataset. Following the standard practice in the ML literature, we used 70% of this sample for training purposes and the remaining 30% for testing purposes.

2.2 External Dataset 1: Trade Transactions from the Paraguayan Customs

To test our algorithms against a dataset outside the training set, we used a random sample of product descriptions from trade transactions recorded by Paraguayan customs. As before, we restricted the sample to agricultural, animal, and food products (HS chapters 1–22). Importantly, for this dataset, we not only had the product descriptions but also the HS codes assigned by firms, which enabled us to directly observe the accuracy of the HS codes provided by the different ML algorithms and by GPT models.

2.3 External Dataset 2: USDA Organic Product Descriptions

Finally, we used information on products for which the USDA has issued organic certifications to Latin American firms (see Marra de Artiñano et al., 2023). The original dataset comprises more than 26,000 product descriptions. These texts vary substantially in terms of how specific and clean they are (that is, whether they use clear, easy-to-understand wording that is narrow enough to accurately categorize the product). Thus, these descriptions may be significantly shorter than those usually found in customs databases (e.g., "maize" or "mangoes"), and may be highly specific or scant (e.g., "concentrate soursop pulp" or "ungurahui"). Table A1 in the Appendix shows selected descriptions for illustrative purposes.

3 Methodology

Classification algorithms play a vital role in a wide range of ML applications (Sarker, 2021).[13] Multiclass classification, a particularly challenging task, is one of the most widespread uses for classification algorithms. In this case, the objective is to categorize the data into three or more different and mutually exclusive categories (Aly, 2005), in such a way that what is sought is to train one or several models that can correctly assign a set of uncategorized data to the correct categories. Formally, given a training dataset of the form (x_i, y_i) where x_i is the ith input and y_i is the ith class label that belongs to the set $\{3, \ldots, N\}$ we want to find a model H such that $H(x_i) = y_i$ for new, uncategorized data.

The process of automatic product classification using ML models consists of several steps. First, the train-and-test data (in our case, the product descriptions in trade transactions from Chilean customs) needs to be preprocessed, which involves preliminary cleaning of the data, splitting it, tokenizing it, and extracting features. Second, the data must be divided into the training and testing sets. Third, a series of different ML algorithms are applied to the training set. After performing these steps, we also tested the estimated models on two alternative external databases (product descriptions in trade transactions from Paraguayan customs and the USDA organic product database).

In addition, we use OpenAI's GPT API to classify the different products through direct prompts and benchmark its performance against that of the ML models.

Our analysis was entirely conducted using Jupyter notebooks and Python open-source libraries such as NLTK, scikit-learn, spaCy, AST, and other commonly used libraries, along with the OpenAI library to conduct the GPT prompt requests.

[13] They have been used extensively in areas such as NLP (Otter et al., 2020), image recognition (Fujiyoshi et al., 2019; Lai, 2019), and sentiment analysis Mitra (2020), among others domains. In recent years, breakthroughs in NLP and text mining have propelled the adoption of these algorithms in applications as diverse as fraud detection, asset classification in finance and early detection of health problems (Kowsari et al., 2019).

3.1 Data Processing

As mentioned above, the Chilean customs dataset covers 2009–2021, contains more than 104 million observations, and lists 12,934 different products at the HS 8-digit level. We processed this dataset by first restricting the product descriptions to those in chapters 1–22 of the HS schedule, which correspond to animal, vegetable, and food manufacturing products. This first filter reduced the total number of observations to approximately 12 million and the total number of unique 8-digit HS codes to 2,866.[14] We then proceeded to randomly select 1 million product descriptions in an effort to reduce the computational burden of the exercise.

To clean and preprocess the product descriptions, we performed a series of tasks that are summarized in Table 1:

[14] In addition, we filter out 469,435 observations that do not correspond to any known product according to the standard HS nomenclature (e.g., 16.00.00).

Table 1: Preprocessing of Product Descriptions

Step	Description
Text preparation	We imported the Natural Language Toolkit (NLTK) library and apply the "word tokenize" function to break the text into individual words (tokens). This was crucial, as it made post-processing of text and feature extraction easier.
Lowercase	We converted all words to lowercase using a lowercase function. This helped to ensure that words are treated consistently in subsequent steps and to reduce data complexity.
Removal of non-ASCII characters	We applied a function to remove non-ASCII characters, except for the letter "ñ". This allows us to standardize and simplify the text, thus facilitating subsequent analysis.
Converting numbers written in words to digits	We used a function from the NLTK package to convert numbers written in words to digits. This helped reduce the complexity of the text and made it easier to extract relevant features.
Stop-word removal	We used a function to remove stop-words that do not provide relevant information for analysis, such as prepositions and conjunctions. This helped reduce the complexity of the text and allowed us to work on the most significant words.
Lemmatization	The lemmatize functions were used to transform words into their base or lemma form. This helped reduce the complexity of the text by grouping similar words together and made it easier to identify patterns in the data.[15]
Removing words that are not in English or Spanish	We applied a function to remove words that are not in English or Spanish. This helped focus the analysis on the relevant languages and reduced noise in the data.
English and Spanish noise removal	We applied some functions to remove irrelevant words in English and Spanish. This helped reduce noise in the data and allowed the most relevant words to be used for analysis.

Source: Authors' own elaboration.

By cleaning and preprocessing the text in the product descriptions as described in these steps, we got the data ready to be used with ML models and ensured that the models were

accurate and efficient at estimating HS codes. Table A2 in the Appendix illustrates the application of this procedure to a selected product description and shows the results thereof. This example provides a clear idea of the complexity of dealing with certain descriptions and demonstrates the importance of simplification if they are to be uses as inputs for traditional ML algorithms.

3.2 Traditional ML Algorithms

We used several different ML models for our multiclass classification problem. While offering an extensive explanation of such models is beyond the scope of this paper, this section contains a brief review of some of their characteristics, based primarily on Kowsari et al. (2019) and Aggarwal and Zhai (2012):

1. **Support Vector Machine (SVM)**: SVM is a supervised learning algorithm that identifies the optimal hyperplane that separates data points into their respective classes and maximizes the margin between the classes. The key in this classifier is to "determine the optimal boundaries between the different classes and use them for the purposes of classification" (Aggarwal and Zhai, 2012). It is one of the most efficient ML algorithms since its introduction in the 1990s.

2. **Rocchio**: It is a traditional and efficient method for text categorization. The algorithm represents documents as vectors in a high-dimensional space and calculates the centroid for each category. To classify a new product description, the algorithm measures the similarity of each to the centroids and assigns it to the closest category.

3. **Logistic Regression**: It is a linear model for binary classification, which can be extended to multiclass classification problems like categorizing product descriptions. Using a logistic function, the model estimates the probability of a product description belonging to a specific class. The class with the highest probability is then assigned to the product description.

4. **k-Nearest Neighbors (k-NN)**: It searches for the k most similar or closest items to the new object we want to classify, and then decides which category it belongs to, based on the most common category among its nearest neighbors.

5. **Random Forest**: It is an ensemble learning method that constructs multiple decision trees during training and combines their predictions to improve classification accuracy. This method can handle large datasets and effectively classify product descriptions into various HS chapters, but it is relatively slow to create predictions once trained.

6. **Naive Bayes**: It is a probabilistic classifier based on Bayes' theorem, which assumes independence between features. Although this assumption is often not valid in real-world applications, Naive Bayes classifiers still perform well in many cases.

7. **Decision Tree**: It is a flowchart-like structure that can be used for classification tasks. The tree is built by recursively splitting the dataset based on the feature that provides the best separation into classes.

3.3 LLMs: GPT-3.5 and GPT-4

GPT-3.5 and GPT-4 are advanced large-scale language, deep learning model.[16] They use transformer architecture to understand and generate human-like text. With billions of parameters and the ability to learn from vast amounts of text data, it has been fine-tuned to excel in a wide range of NLP tasks.

Some of the notable properties of GPT models include their autoregressive nature, which allows them to generate contextually relevant and coherent text by predicting the next word in a sequence given the previous words. The models are trained using unsupervised learning with a vast dataset that includes websites, books, and articles. The knowledge cut-off for both GPT-3.5 and GPT-4 is September 2021. [17].

[16] They were both developed by OpenAI. In our analysis, we use (1) the GPT-3.5 version -internally called "gpt-3.5-turbo"-, which powers the publicly available version of the ChatGPT chatbot, and (2) a more recent model, GPT-4 -internally called "gpt-4"-, which has an estimated parameter size x10 that of GPT 3.5.

[17] This is the knowledge cut-off as of the 7th of November of 2023, when the last exercise in this manuscript was

We applied the models through the OpenAI API, asking them to assign an HS category based on the product description we provide. For that purpose, we give a system command to act as a wizard that assigns 6-digit HS codes and then asked them to execute such function for a given product description. In this regard, it is worth mentioning that we asked it not only to assign each product an HS code but also to provide its best estimate if the product description was not clear enough, thereby "forcing" it to make a guess.

Preparing datasets for use with the model (that is, the data processing described in section 3.1) was not essential. When working with LLMs, which are trained on a diverse range of text typologies, preprocessing data may not be needed and may even be disadvantageous as it might obscure valuable contextual information. We therefore merely input orders one at a time, thus allowing GPT models to categorize products individually. The specific prompt used and the completion request associated with it are presented in the appendix (section A3).

4 Results

4.1 Results on the Train-Test-Split Dataset: Trade Transactions from the Chilean Customs

Figure 1 shows the accuracy of the different models on the Chilean customs data. It is important to stress that this is the dataset on which the ML algorithms are trained. Note that neither GPT-3.5 nor GPT-4 are "trained" using any of the datasets, since the outcomes are obtained from direct prompts to the model through the API.

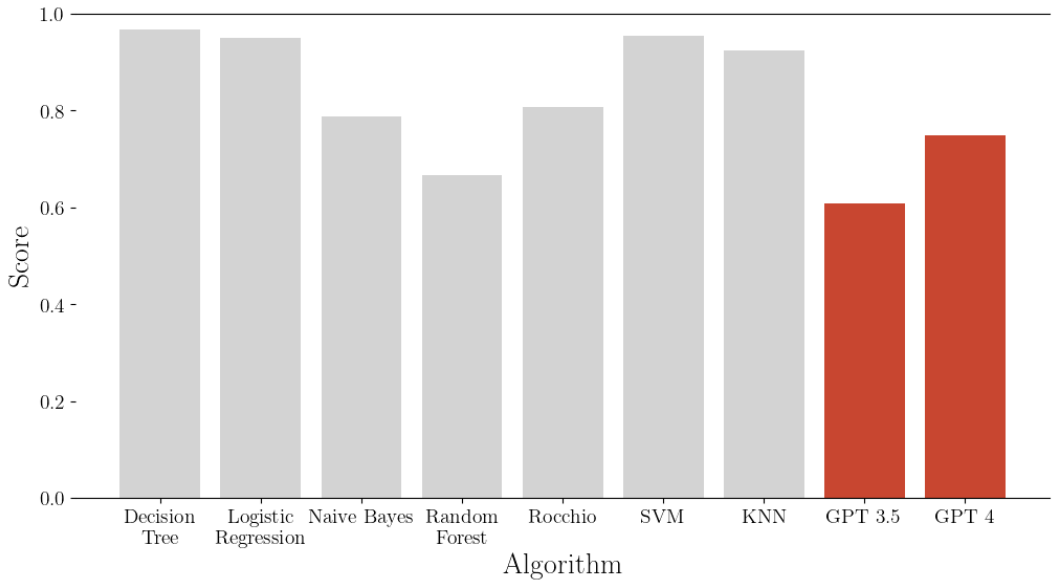
The trained algorithms had very high accuracy levels on the test dataset, especially in the case of the Decision Tree, Logistic Regression, and SVM algorithms. The results of this test are typically used to assess the predictive capability of an algorithm.

As expected, the accuracy levels were higher when less granular product categories were

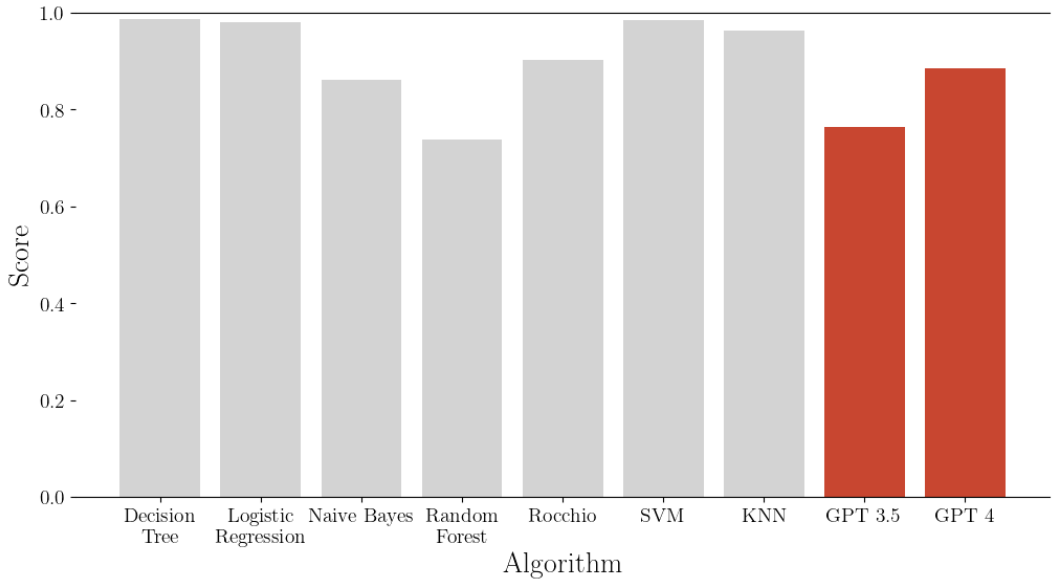
carried out. The models may be updated in the future. See https://platform.openai.com/docs/models/ for the latest knowledge cut-offs.

used (see Figures 1b and 1c). However, this increase in success rates is uneven. For example, the hit rate of the GPT-4 model increased by 13 percentage points (from 75% to 88%) when moving from HS 6-digit codes to HS 4-digit codes and by an additional 7 percentage points (from 88% to 95%) when HS 2-digit codes were used. Overall, these findings indicate that GPT-4 has a very high accuracy in the prediction of broad products' category.

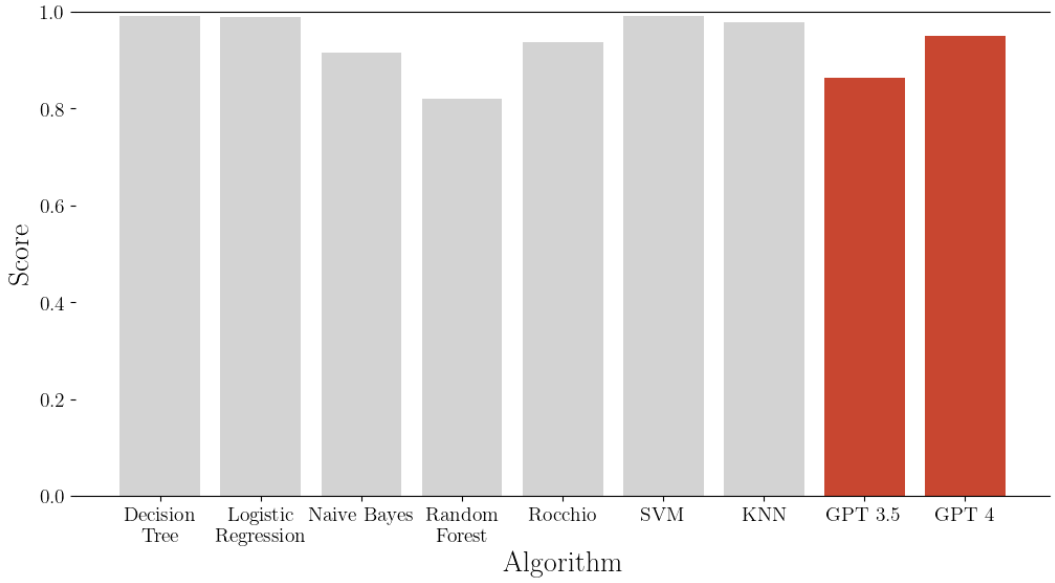
Figure 1: Algorithm's Accuracy in the Test-Train-Split Dataset: Chilean Customs.



(a) HS-6 digit level



(b) HS-4 digit level



(c) HS-2 digit level

Source: Authors' calculations based on Chilean customs data.

4.2 Results on the External Dataset 1: Trade Transactions from the Paraguayan Customs

In this subsection and in the following one, we tested the ML algorithms outside the dataset on which they were trained. This is key since the usefulness of such algorithms in real-world applications depends on their external validity. Real data imposes a clear challenge in this regard. It features a variety of product descriptions, including different formats. Hence, a model performing well on the training dataset may not be indicative of how well it will accomplish other classification tasks.

To explore this, we compared the models using data that was not part of the test dataset. Specifically, we selected a random sample of 10,000 product descriptions from Paraguayan customs records. This allows for a fairer comparison of ML models and GPT-3.5 and GPT-4, since it confronts both models with data on which neither was explicitly trained. The results are presented in Figure 2.

Traditional ML algorithms did not perform well when tested using real-world data on which they were not trained. Their accuracy rates dropped below 30% for 6-digit HS codes. In contrast, the GPT models performed much better, correctly assigning around 75% in the case of GPT-4 and 60% of the product codes in the case of GPT 3.5. These results are similar to those obtained on the Chilean dataset. This points to the consistency of GPT models in automatic product classification across customs datasets.[18]

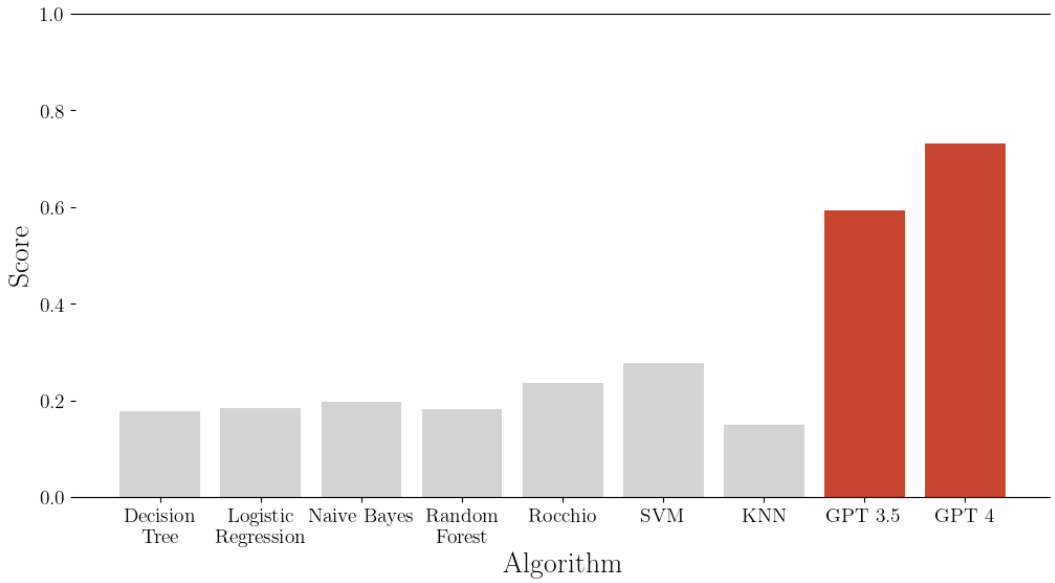
Next, we proceeded to check how the different algorithms performed at more aggregate levels. This allows us to better understand how these algorithms work and where the highest rates of success/failure occur. Figure 2b shows the accuracy of the different algorithms when using 4-digit HS codes. Once again, conventional ML algorithms achieved a relatively low accuracy level, with a maximum of 37%. GPT models, however, performanced at high accuracy rates. GPT-4 reached 88% (and GPT 3.5 of 77%), a 13-percentage-point increase (17 percentage-point increase for GPT 3.5) in its hit rate compared with HS 6-digit codes.

Figure 2c reports the results for the more aggregated 2-digit classification. In this case,

[18] In the appendix, we show GPT's accuracy at the HS 6-digit level for each broad HS 2-digit category. We failed to find any pattern across datasets, further suggesting a high level of consistency in its average performance

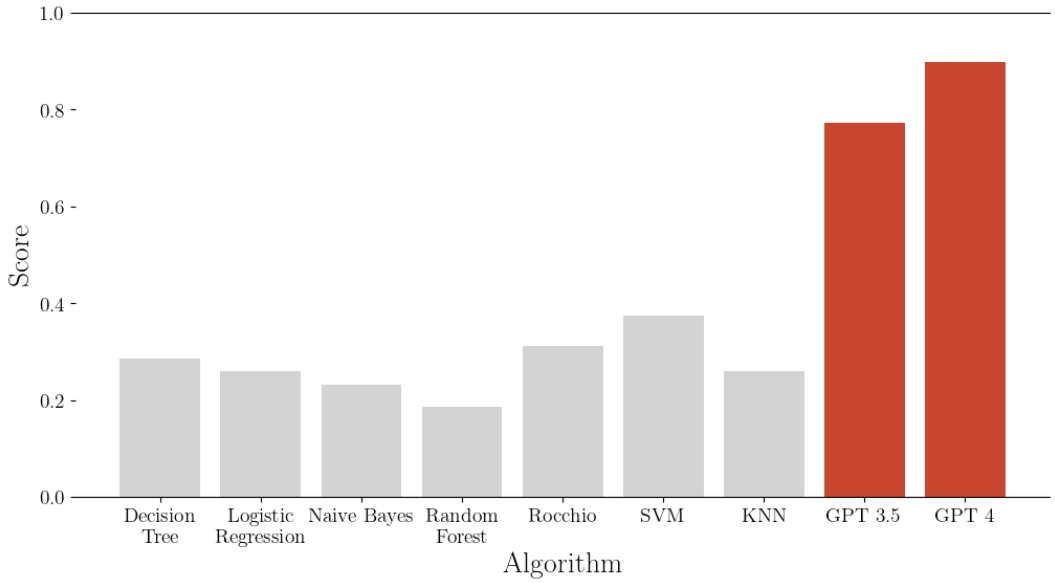
the GPT algorithms achieved more than 90% accuracy. However, it should be noted that the performance of the conventional ML algorithms also improved significantly, with the Decision Tree reaching 73%. This indicates that, even out of their training dataset, all algorithms can predict a product's HS relatively well, but LLMs perform much better in more granular classification levels. In the Appendix, we show that GPT models have high precision across all HS chapters included in our analysis (section A4). [19]

Figure 2: Algorithm's Accuracy in the First External Dataset: Paraguayan Customs.

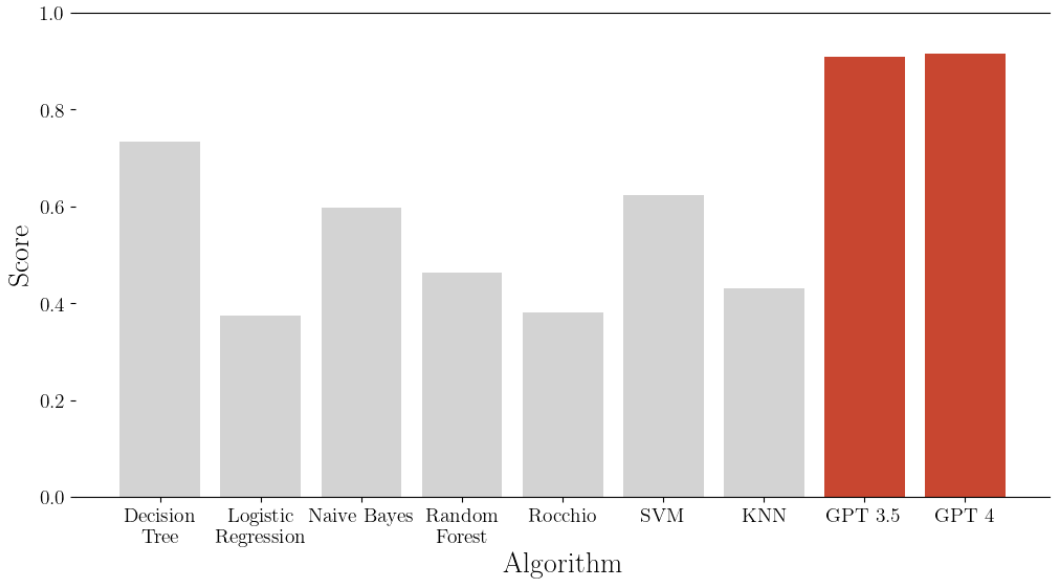


(a) HS-6 digit level

[19] In addition, a comparison of Figures 2a, 2b and 2c reveals differences in terms of the best-performing conventional ML algorithm. While at the HS 6-digit level SVM has the highest accuracy rate, Decision Trees seem to outperform other methodologies at a less disaggregated level.



(b) HS-4 digit level



(c) HS-2 digit level

Source: Authors' calculations based on Paraguayan customs data.

4.3 Results on the External Dataset 2: USDA Organic Product Descriptions

Finally, we assessed the ability of conventional ML algorithms and LLMs to accurately predict HS codes using text formats that differ from those traditionally used in customs. To do this, we used a set of descriptions of products for which Latin American firms are certified as organic producers and sellers by the USDA. As mentioned above, these product descriptions have different formats and vary significantly in terms of depth and specificity, which makes them potentially harder to categorize than the average customs product description. Furthermore, although this type of text contains descriptions of products, it does not specify the respective HS codes for each. Consequently, it cannot be used to train ML models to predict these. Similar cases can be found in many other data sources, such as cross-border e-commerce shipments, bank transactions, historical trade data and survey-based descriptions.

To conduct this exercise, we selected a random sample of 1,000 descriptions of USDA certified organic products and classified these by hand into 6-digit HS tariff lines. The results are fully in line with those based on the Paraguayan customs external dataset: the standard ML algorithms performed significantly worse than GPT-3.5 and GPT-4.[20]

Figure 3a shows the accuracy at the HS 6-digit level. The GPT-3.5 model achieved a success rate of 74.1%, while the traditional ML models scored 15% at most (Rocchio model). The differences were similar when HS 4-digit codes were used: the accuracy of GPT-3.5 was over 80%, an improvement of 6 percentage points on the HS 6-digit level. Among the traditional ML algorithms, the maximum hit rate increased to 26% (again, the Rocchio model). Finally, at the HS 2-digit level, GPT-3.5 classified almost 88% of the product chapters correctly (i.e., a 7-percentage-point improvement on the HS 4-digit classification).

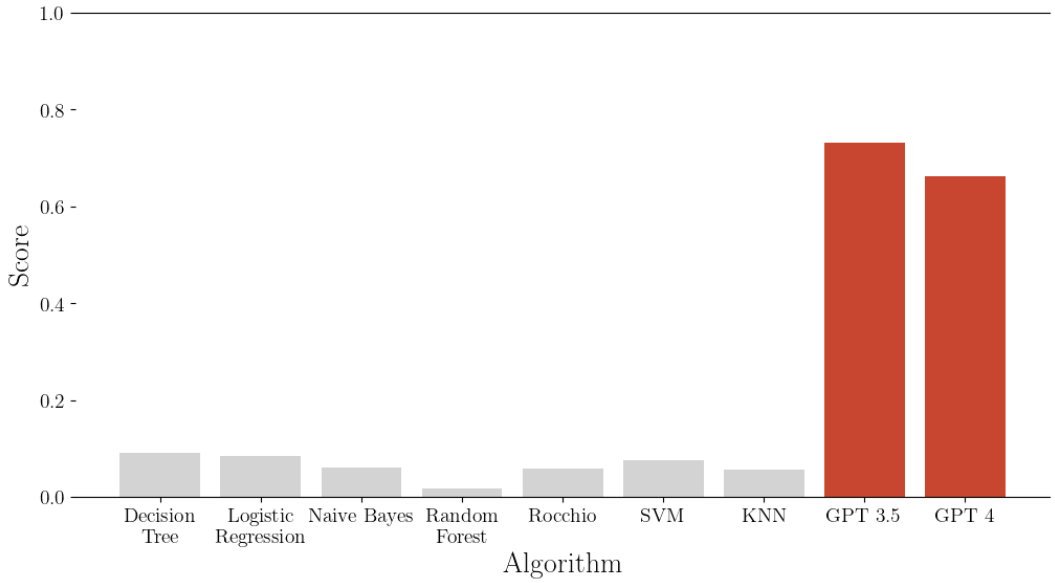
Also interestingly, in this case, GPT-4 performed slightly worse than the smaller GPT-

[20] To test the difference in performance from an increase of one order of magnitude in the number of products classified, we conducted a sensitivity analysis, in which we randomly divided the sample into 10 groups of 100 product descriptions and examined their accuracy. We found that GPT performed very similarly across the 10 groups, with a standard deviation of just 0.0136 for GPT 3-5 and 0.0149 for GPT 4). We also did this for the other datasets (see Appendix, Section A5).

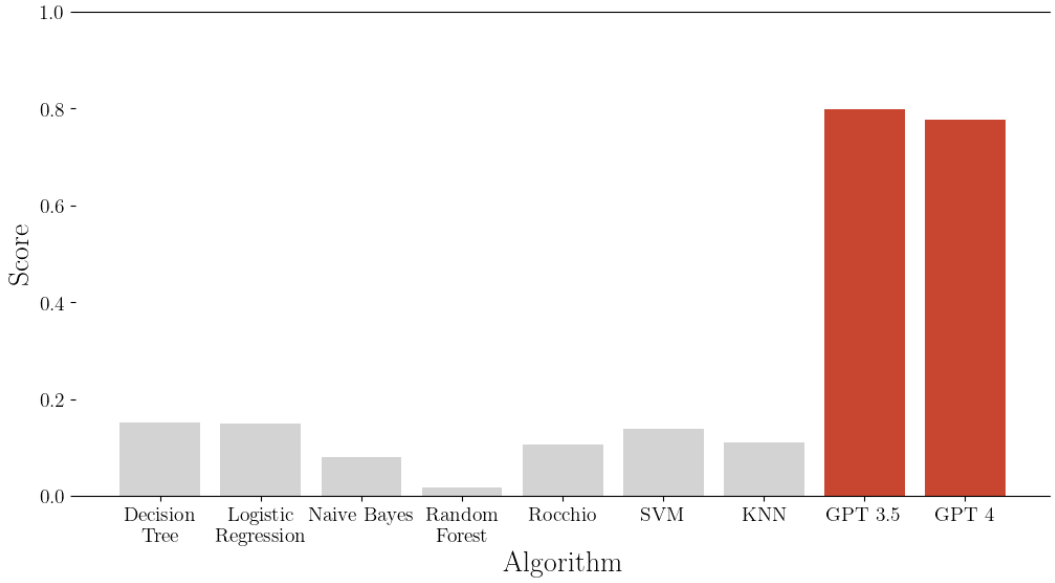
3.5 (7 percentage point less at the HS 6-digit level, 2 percentage point less at HS 4-digit level, but 1 percentage point more at HS 2-digit level). Given that GPT-4 is one-order of magnitude larger in terms of parameter space, this suggests a nonlinear relationship between performance in automatic product classification and model size. Above a certain size, an overall larger training set may not lead to sizeable gains in automatic product classification tasks.

Despite the latter, it is worth noting that the performance gap beween GPT models and traditional ML models increased when using this highly unstructured product descripton data relative that observed with traditional customs data.

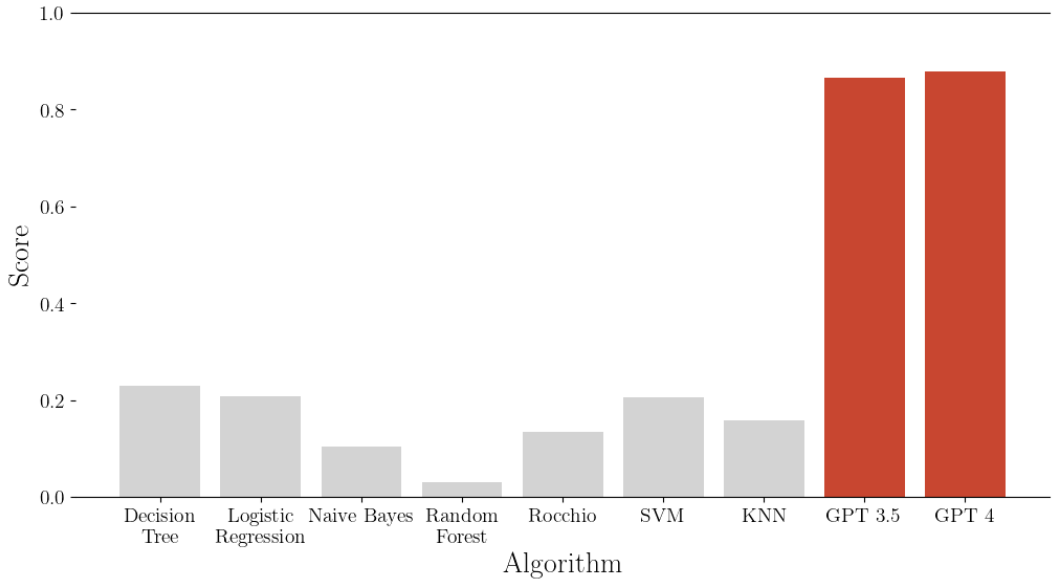
Figure 3: Algorithm's Accuracy in the Second External Dataset: USDA Organic Classification.



(a) HS-6 level



(b) HS-4 level



(c) HS-2 level

Source: Authors' calculations based on USDA data.

5 Discussion and Conclusions

LLM models showed high accuracy rates when classifying products according to the HS nomenclature. Traditional ML algorithms performed very well on their training dataset but their performance dropped dramatically when they were tested on external data. In such external validity tests, GPT-3.5 and GPT 4 significantly outperformed these algorithms. Importantly, this was the case even when the ML models were trained with 1 million observations of high-quality customs product descriptions and then subsequently tested on high-quality descriptions from a different customs agency from the same region.

Another major advantage of GPT models is their ability to work with product descriptions in different languages. Throughout our analysis, we used data in English and Spanish, but GPT-3.5 and GPT-4 are likely to perform very well across many other languages in which large amounts of data are publicly available (e.g., Chinese, French, German, etc.). Importantly, it is also able to successfully handle regional variants of the same language.

One interesting example from our study was *Physalis peruviana*, a fruit typically known as "goldenberry" in English. Our Chilean training data refers to them as "uchuva," but the fruit goes by other names in different countries: "uvilla" in Ecuador, "aguaymanto" in Peru, and "fisalis" in Spain. ML algorithms trained on the Chilean data failed to identify these regional variations and thus misclassified the product, whereas GPT-3.5 and GPT-4, trained on a much wider set of texts, recognized the fruit and classified it properly. This is an example of how the wide training dataset of LLMs allows them to outperform standard ML algorithms. LLMs can thus be very useful in comprehensive unilateral, regional, and multilateral trade policy initiatives involving product classifications over time and across countries (e.g., trade facilitation).

LLMs with chat interfaces are also significantly simpler since they do not require data-cleaning and preprocessing routines. Performing these tasks with traditional ML algorithms can be rather time-consuming and resource-intensive, especially those related to feature extraction.[21] While the API is necessary to work with both GPT-3.5 and GPT-4 at scale, the standard interface enables the classification functionality to be integrated easily into existing systems or applications. In our analysis, we worked with the base models, without making further adjustments, but GPT models could also be adapted for use with specific data through its fine-tuning mechanism. Fine-tuning LLMs with trade data may help them be more effective at product classification at scale.[22]

In terms of costs, the models we used (GPT-3.5 and GPT-4) are relatively inexpensive,

[21] We also assessed the model against a manual classification carried out by a research assistant (RA) using a sample of 100 observations. The results indicate that, while the RA's accuracy was slightly above that of GPT-3.5 at the HS 6-digit level, the difference fades when more aggregate classification levels are considered. At the 2-digit HS level, GPT-3.5 performed slightly better than the RA. It is worth stressing that while the RA needed four hours to accomplish the task, GPT 3.5 completed it in just one minute. This suggests that there is potentially a tradeoff between accuracy and time for highly disaggregated classifications in small samples. The terms of this tradeoff are highly likely to change as the number of observations increases, with GPT-3.5 clearly emerging as the better approach for large samples, especially given that human working time increases at a nonlinear rate due to marginal decreasing returns.

[22] Our results suggest that GPT-4 is only more effective than GPT-3.5 in some datasets and thus that there is a nonlinear relationship between model parameter size and accuracy rates in automatic product classification. Fine-tuning LLM models may be helpful in improving performance and avoiding this diminishing returns to scale. Note, however, that extensive fine-tuning of LLM models remains very costly.

except for very large tasks.[23]

Importantly, open-source LLMs are also becoming increasingly competitive and can be expected to be able to perform very well in large-scale product classification tasks in the short term.[24] Benchmarking automatic product classification at a larger scale across a wide range of LLMs and fine-tuning methods therefore remains and will be an important avenue for future research.

[23] In our work, we used both GPT-3.5 ("gpt-3.5-turbo") and GPT-4 ("gpt-4"). Without going into the billing system works in detail, our estimate as of the 1st of November of 2023 is that the total cost of classifying a dataset of 10,000 standard customs product descriptions is approximately $2.5 for GPT-3.5 and $40 for GPT-4. (see the pricing).

[24] In Appendix Section A6, we carry out a simple benchmark of other LLMs, including open-source models (LLaMa, SOLAR 70B) and other models developed by Google (Bard, PaLM) and Anthropic (Claude Instant) for a small random sample of 100 observations. We find that several other models perform well in automatic product classification -particularly Bard and PalM-, albeit none reaches the accuracy level of GPT-4.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining Text Data*, 163-222.

Aly, M. (2005). Survey on multiclass classification methods. *Neural Networks, 19*, 1-9.

Biswas, S. S. (2023). Role of ChatGPT in Public Health. *Annals of Biomedical Engineering, 51*(5), 868-869.

Chen, H., Van Rijnsoever, B., Molenhuis, M., van Dijk, D., Tan, Y. H., & Rukanova, B. (2021). The use of machine learning to identify the correctness of HS Code for the customs import declarations. *In 2021 Institute of Electrical and Electronics Engineers 8th International Conference on Data Science and Advanced Analytics*, 1-8.

Fujiyoshi, H., Hirakawa, T., & Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. *International Association of Traffic and Safety Sciences, 43*(4), 244-252.

Hansen, A. L., & Kazinnik, S. (2023). Can chatgpt decipher fedspeak? *Social Science Research Network*.

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., & Taska, B. (2023). Remote work across jobs, companies, and space. *National Bureau of Economic Research* (Working Paper No. 31007).

He, M., Wang, X., Zou, C., Dai, B., & Jin, L. (2021). A commodity classification framework based on machine learning for analysis of trade declaration. *Symmetry, 13*(6), 964.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*.

Ko, H., & Lee, J. (2023). Can chatgpt improve investment decision? from a portfolio management perspective. *Social Science Research Network.*

Kocon, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Kazienko, P. (2023). Chatgpt: Jack of all trades, master of none. *Information Fusion, 101861*.

Korinek, A. (2023). Language models and cognitive automation for economic research. *National Bureau of Economic Research* (Technical Report No. 30957).

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150.

Lai, Y. (2019). A comparison of traditional machine learning and deep learning in image recognition. *Journal of Physics: Conference Series*, 1314.

Lee, J. K., Choi, K., & Kim, G. (2021). Development of a natural language processing based deep learning model for automated HS code classification of the imported goods. *Journal of Digital Contents Society, 22*(3), 501-508.

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *ArXiv Preprint.*

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News, 40*(3), 26-29.

Marra de Artiñano, I., Scattolo, G., Volpe Martincus, C., & Zavala, L. (2023). The value of organic certifications. *Inter-American Development Bank Working Paper*. (Forthcoming).

Mitra, A. (2020). Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset). *Journal of Ubiquitous Computing and Communication Technologies, 2*(3), 145-152.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Social Science Research Network*.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *Institute of Electrical and Electronics Engineers Transactions on Neural Networks and Learning Systems, 32*(2), 604-624.

Ruder, D. (2020). Application of Machine Learning for Automated HS-6 Code Assignment. *Master's thesis, University of Tartu, Institute of Computer Science.*

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160.

Spichakova, M., & Haav, H. M. (2020). Application of Machine Learning for Assessment of

HS Code Correctness. *Baltic Journal of Modern Computing, 8*(4), 698-718.

Turhan, B., Akar, G. B., Turhan, C., & Yukse, C. (2015). Visual and textual feature fusion for automatic customs tariff classification. *2015 Institute of Electrical and Electronics Engineers International Conference on Information Reuse and Integration*, 76-81.

Xu, C.-J., & Li, X.-F. (2019). Research on the Classification Method of HS Code Products Based on Deep Learning. *Modern computer, 1*, 13-21.

Yang, K., Ji, S., Zhang, T., Xie, Q., & Ananiadou, S. (2023). On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *ArXiv Preprint.*

Yang, K.-C., & Menczer, F. (2023). Large Language Models Can Rate News Outlet Credibility. *ArXiv Preprint.*

Online Appendix

A1 Organic Descriptions

Table A1: Sample of 10 Randomly Chosen Organic Product Descriptions

Original Product
Ungurahui (Oenocarpus Bataua)
soy beans
Plátanos/Bananos - 1 Traboar_Finca Genoveva (F)
Banana puree acidulated deep frozen
Organic aseptic concentrate soursop pulp
Organic white corn powder
Banana puree without seeds
Organic coco
Safflawer
Maca flour - pre cooked

Source: Authors' elaboration based on USDA.

A2 Preparation Steps of Descriptions

Table A2: Preparation Steps of a Random Selected Description

Step	Result
Initial description	FROZEN DOUGHS EUROPASTRY-F CODE-81299 BERLIDOTS BOMBOM FOOD PREPARATION BASED ON WHEAT FLOUR AND WATER IN BOXES OF 36 UNITS FOR HUMAN CONSUMPTION
Text preparation	['FROZEN', 'DOUGHS', 'EUROPASTRY-F', 'CODE-81299', 'BERLIDOTS', 'BOMBOM', 'FOOD', 'PREPARATION', 'BASED', 'ON', 'WHEAT', 'FLOUR', 'AND', 'WATER', 'IN', 'BOXES', 'OF', '36', 'UNITS', 'FOR', 'HUMAN', 'CONSUMPTION']
Lowercase	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'on', 'wheat', 'flour', 'and', 'water', 'in', 'boxes', 'of', '36', 'units', 'for', 'human', 'consumption']
Removal of non-ASCII characters	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'on', 'wheat', 'flour', 'and', 'water', 'in', 'boxes', 'of', '36', 'units', 'for', 'human', 'consumption']
Converting numbers written in words to digits	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'on', 'wheat', 'flour', 'and', 'water', 'in', 'boxes', 'of', '36', 'units', 'for', 'human', 'consumption']
Stop-word removal	['frozen', 'doughs', 'europastery-f', 'code-81299', 'berlidots', 'bombom', 'food', 'preparation', 'based', 'wheat', 'flour', 'water', 'boxes', '36', 'units', 'human', 'consumption']
Lemmatization	['frozen', 'dough', 'europastery-f', 'code-81299', 'berlidot', 'bombom', 'food', 'preparation', 'base', 'wheat', 'flour', 'water', 'box', '36', 'unit', 'human', 'consumption']
Removing words that are not in English or Spanish	['frozen', 'dough', 'code', 'berlidot', 'bombom', 'food', 'preparation', 'base', 'wheat', 'flour', 'water', 'box', '36', 'unit', 'human', 'consumption']
English and Spanish noise removal	['frozen', 'dough', 'berlidot', 'bombom', 'food', 'preparation', 'base', 'wheat', 'flour', 'water', 'box', '36', 'unit', 'human', 'consumption']

Source: Authors' calculations based on Chilean customs data.

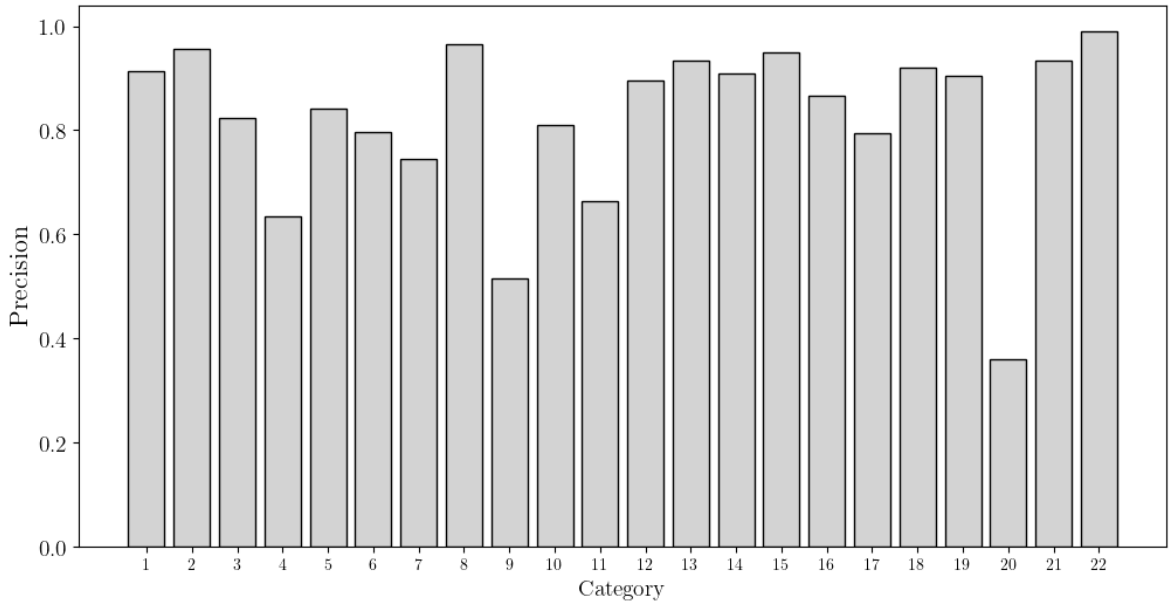
A3 GPT-3.5 Prompt

```python
@backoff.on_exception(backoff.expo, openai.error.RateLimitError, max_time=60)
def assign_code_forced(row, column):
    text = row[column]
    modelo = "gpt-3.5-turbo"
    response = openai.ChatCompletion.create(model=modelo,
    messages=[
        {"role": "system", "content": "You are a helpful assistant that assigns
            product codes in the HS6 product nomenclature categorization."},
        {"role": "user", "content": f'Please assign the harmonized system code
            number in the HS6 for the following description:"{texto}". Return "
            Code: number here". If you are unsure of the classification, provide
            your best possible option'}],
    temperature=0.1)
    assigned_code = response['choices'][0]['message']['content']
    return assigned_code
```

A4 Accuracy in Specific HS Chapters

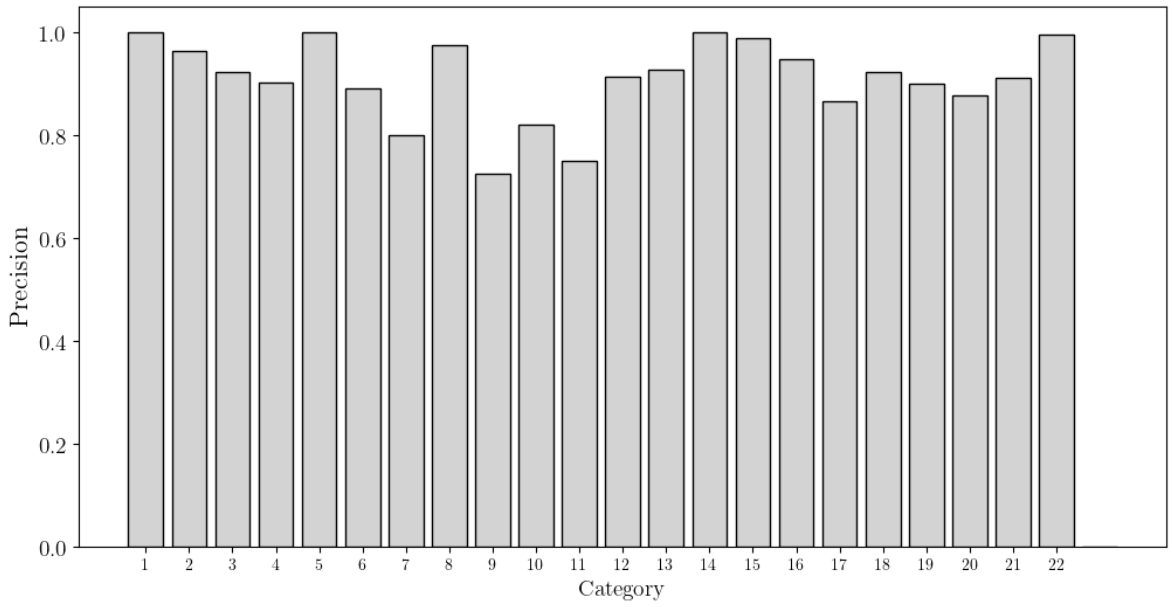
In this section, we show the accuracy of GPT-3.5 and GPT-4's HS 6-digit level classification across different relevant product categories, i.e., HS Chapters 1–22, which include all agricultural, animal, and food products –the focus of our study as explained in the main body of the paper– (see Figures A4a and A4b based on the Chilean data we used in the training set for the algorithms).

Figure A4a: Algorithm's Accuracy in Different HS Chapters for GPT-3.5, Chilean Customs Dataset



Source: Authors' calculations based on Chilean customs data.

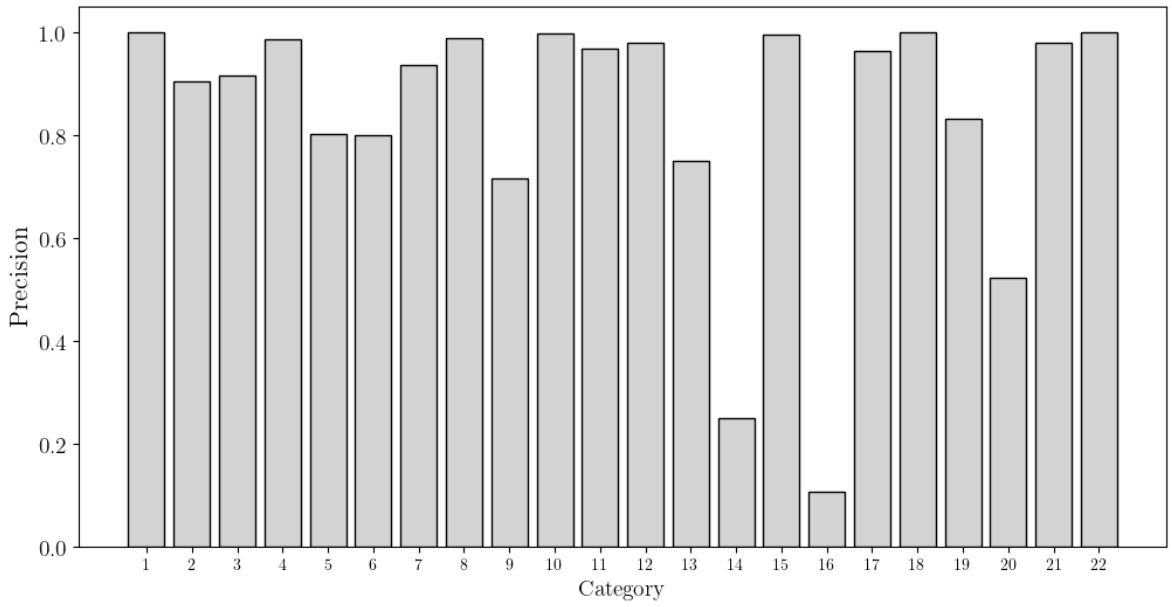
Figure A4b: Algorithm's Accuracy in Different HS Chapters for GPT-4, Chilean Customs Dataset



Source: Authors' calculations based on Chilean customs data.

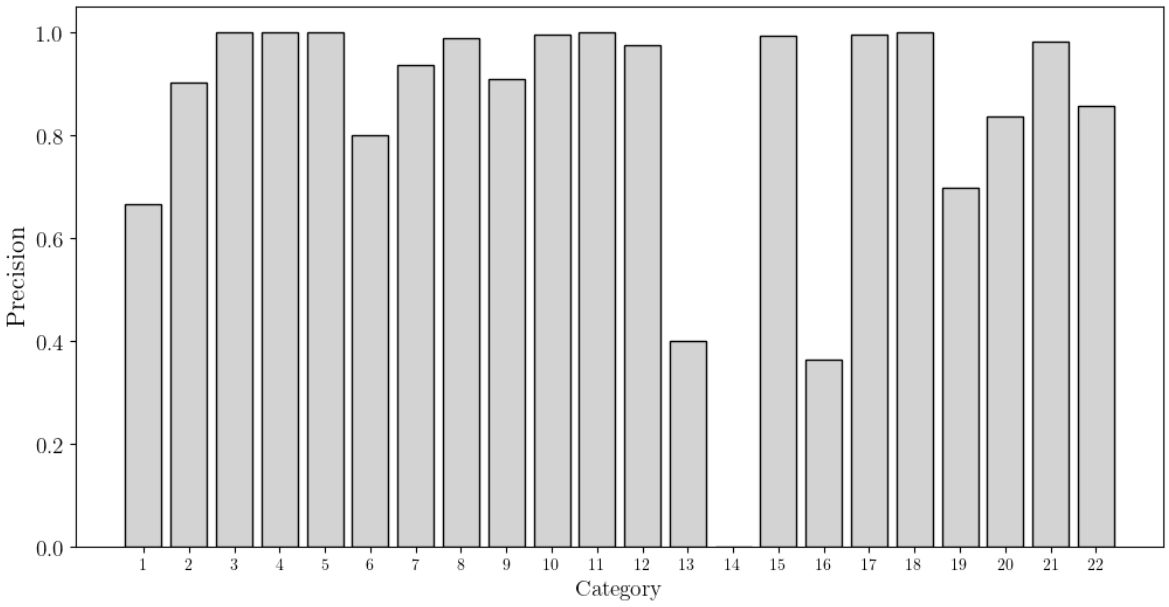
The chapters with the lowest hit levels are 4, 9, 11, and 20, which refer to "Dairy ; birds' eggs; natural honey; edible products of animal origin", "Products of the milling industry; malt; starches; inulin; wheat gluten", "Coffee, tea, mate and spices" and "Preparations of vegetables, fruit, nuts or other parts of plants," respectively. Despite being the lowest hit levels, they still have fairly high accuracy scores (0.63, 0.51, 0.66 and 0.35 with GPT 3.5 and 0.90, 0.72, 0.75 and 0.87 with GPT 4).

Figure A4c: Algorithm's Accuracy in Different HS Chapters for GPT-3.5, Paraguayan Customs Dataset



Source: Authors' calculations based on Paraguayan customs data.

Figure A4d: Algorithm's Accuracy in Different HS Chapters for GPT-4, Paraguayan Customs Dataset



Source: Authors' calculations based on Paraguayan customs data.

Using the Paraguayan data, we find low accuracy levels in two categories: "Vegetable plaiting materials; vegetable products not elsewhere specified or included" (HS chapter 14), with 25% GPT-3.5 accuracy, and "Preparation of meat, of fish or of crustaceans, molluscs or other aquatic invertebrates" (HS chapter 16), with 10% GPT-3.5 accuracy. Importantly, there are very few observations in these categories, which may be affecting the results.

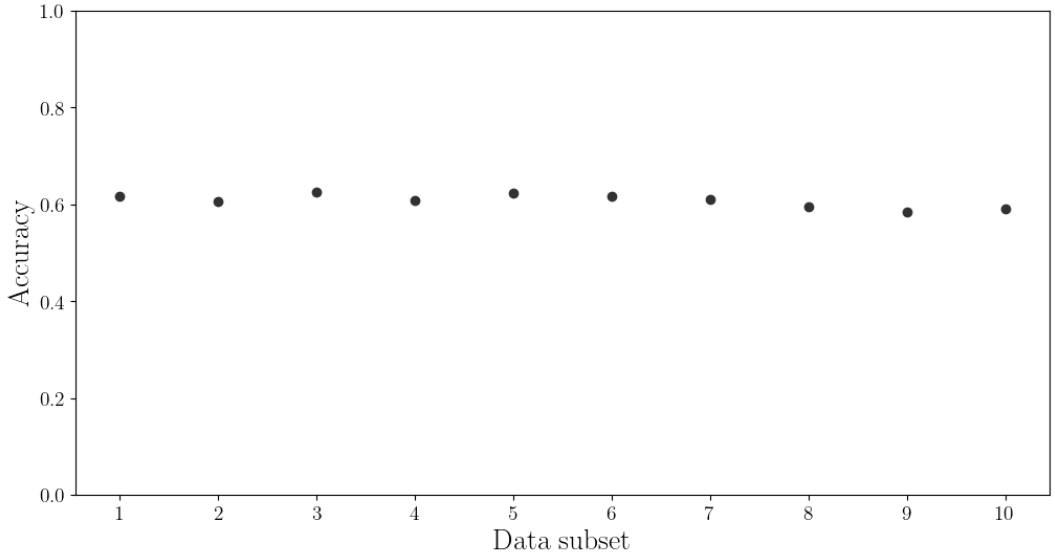
Overall, with the exception of a few HS Chapters in one of our two customs datasets, we find very high accuracy rates across all chapters for both GPT-3.5 and GPT-4.

A5 Subsample Accuracy

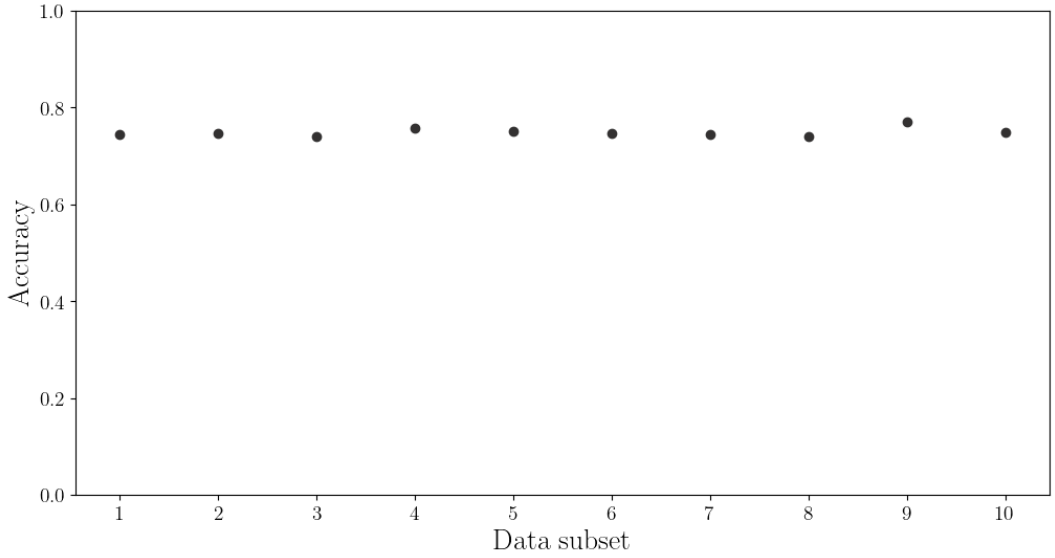
In this section, we report results of robustness checks of our findings to a reduction in the sample size by one order of magnitude. In particular, we randomly divide a 1000-observations sample into 10 subsamples of 100 observations and test the accuracy of GPT-3.5 and GPT-4 in each subsample for each dataset.

Figure A5a shows the point estimates after dividing the sample of observations from Chile into 10 groups. In this case, the point estimates across the 10 subsamples have a very small standard deviation of 0.0048 for GPT-3.5 and 0.0043 for GPT-4. Figure A5b does the same for the Paraguayan dataset. The standard deviation in this case is very similar (0.0049 and 0.0044 for GPT-3.5 and GPT-4, respectively). Finally, figure A5c does the same for the 1,000 classified observations from the USDA organic product database, for which the standard deviation is slightly larger, at 0.0136 for GPT-3.5 and 0.0149 for GPT-4. Overall, this exercise shows that the point coefficients are relatively stable when moving from a 1000-observations sample to 100-observations random subsamples.

Figure A5a: Algorithm's Accuracy in Chilean Subsets for GPT-3.5 and GPT-4



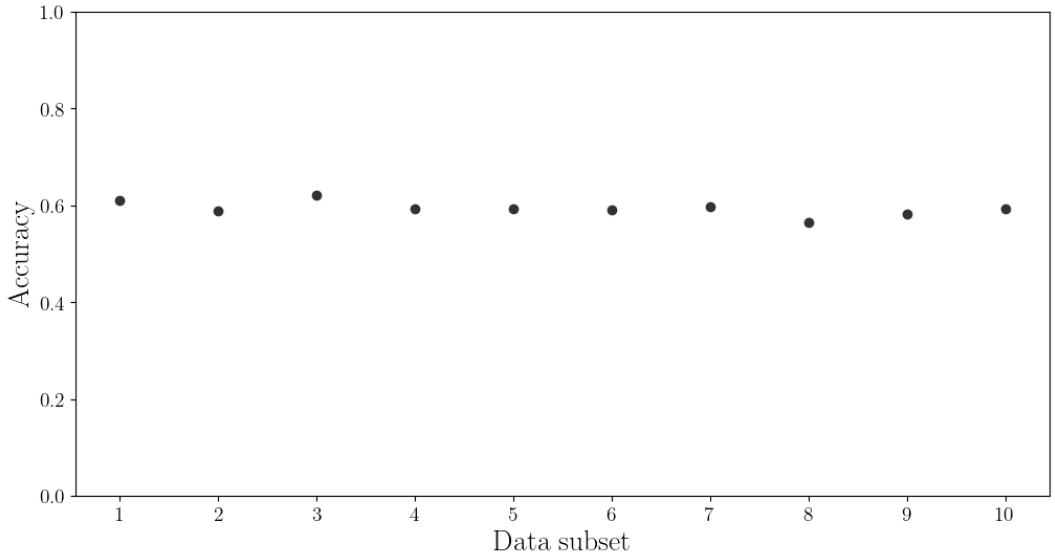
(a) GPT-3.5



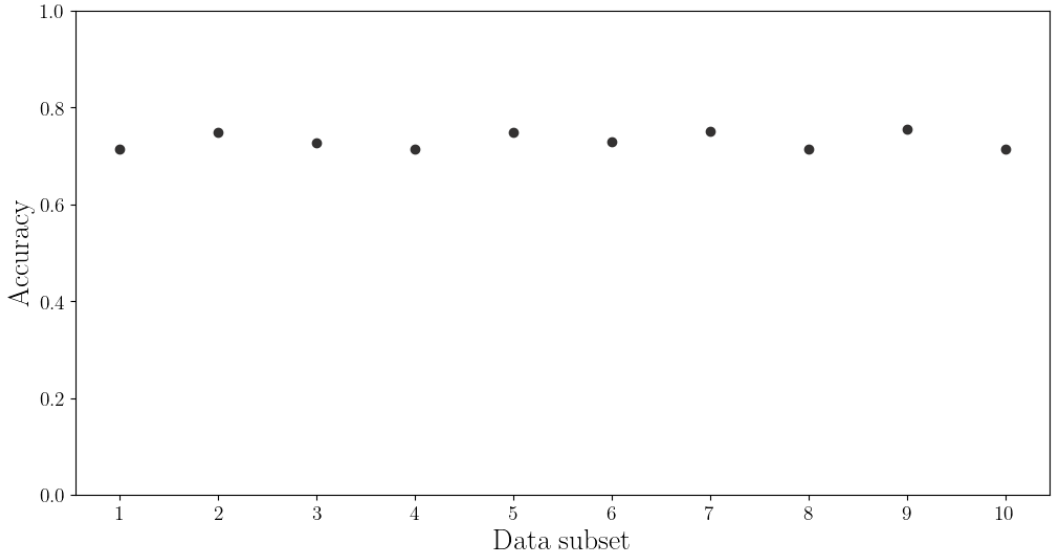
(b) GPT-4

Source: Authors' calculations based on data from Chilean Customs.

Figure A5b: Algorithm's Accuracy in Paraguayan Subsets for GPT-3.5 and GPT-4



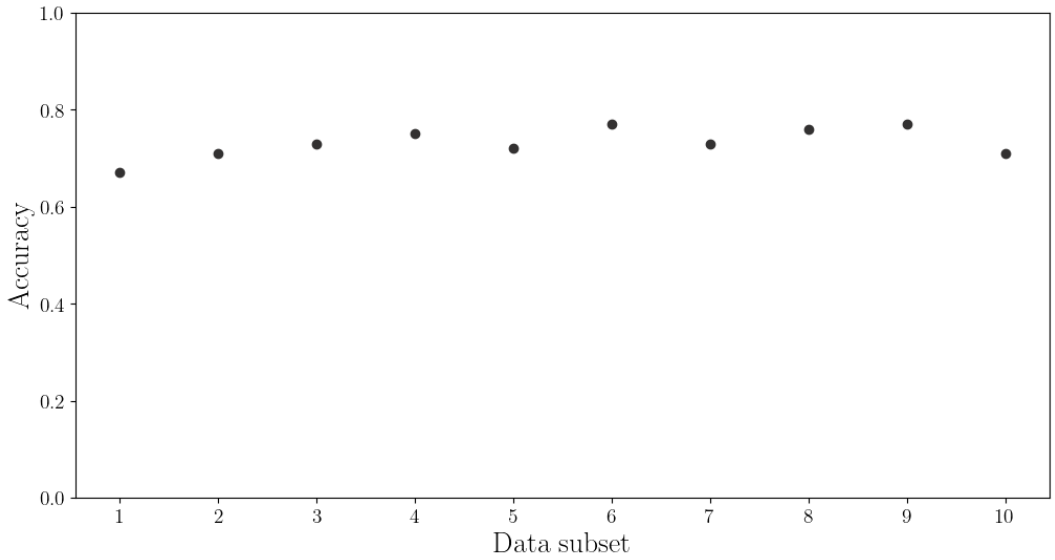
(a) GPT-3.5



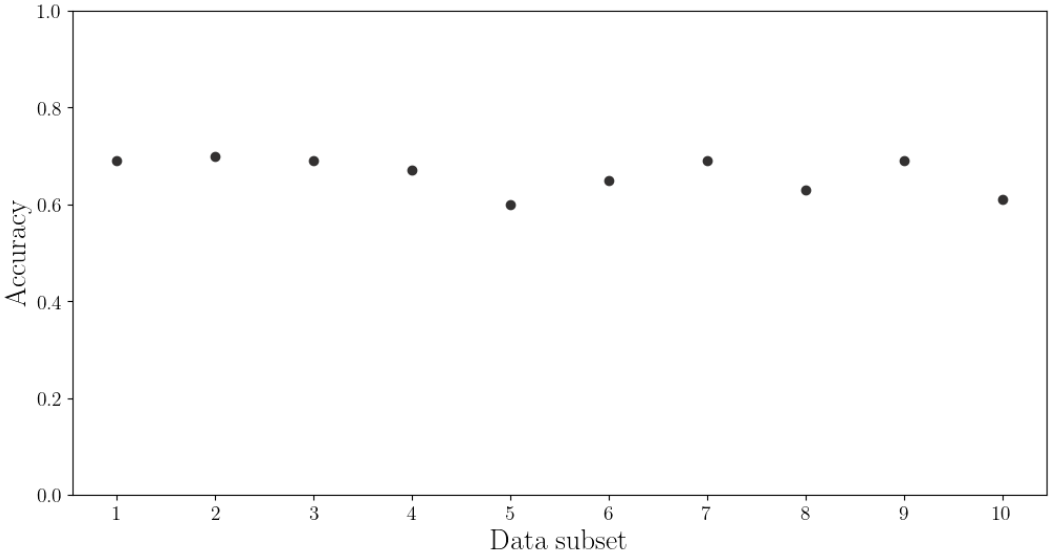
(b) GPT-4

Source: Authors' calculations based on data from Paraguayan Customs.

Figure A5c: Algorithm's Accuracy in the USDA Organic Subsets for GPT-3.5 and GPT-4



(a) GPT-3.5



(b) GPT-4

Source: Authors' calculations based on data from USDA.

A6 Exploring Other LLMs

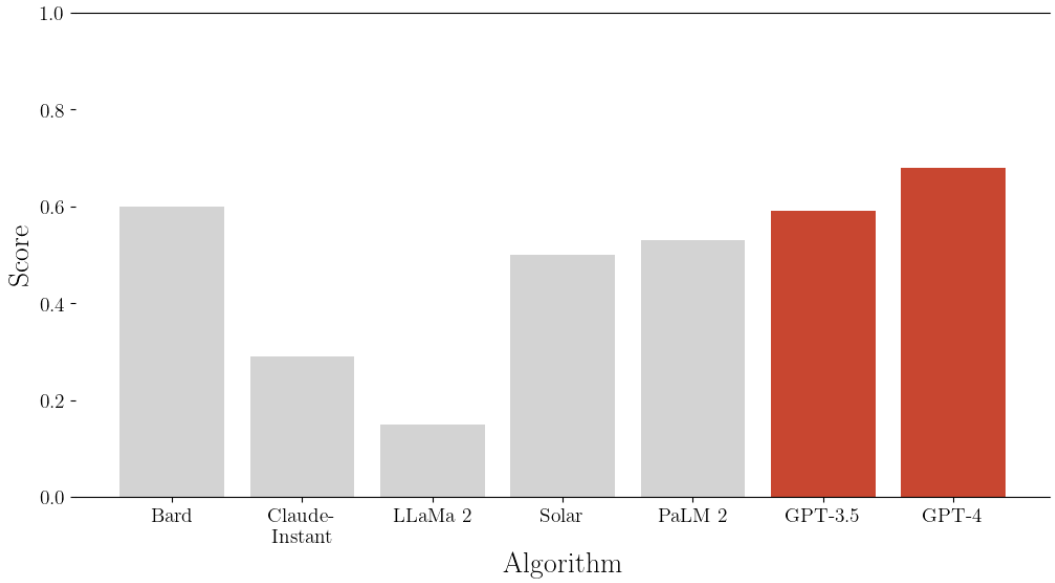
In this section, we present a simple benchmark of various LLMs tasked with classifying products from text descriptions. For this task, we selected 100 random observations from our test dataset with data from the Chilean customs (eee Section 2.1 for more details). This exercise is carried out through Poe, an open platform to explore LLMs. As of November 2023, this service is not accessible at scale through an API, preventing us from carrying out a more extensive analysis of these alternative LLMs. However, note that, while a random sample of 100 observations is relatively small, in Appendix Section A5, we show that our point estimates for GPT 3.5 and GPT-4 are very stable when moving from 100 to 1000 observations.

Figures A6a illustrates the accuracy rates of five distinct LLMs at HS 6-digit classifications: Bard[25], Claude-Instant[26], LLaMa 2[27], Solar[28], PaLM 2, GPT-3.5, and GPT-4. Each bar represents the algorithm's product classification efficacy. Both Bard, PaLM 2 and Solar exhibit relatively high accuracy rates, with Bard edging ahead slightly (60%). PaLM 2 demonstrates a robust performance (53%), aligning closely with Solar's (50%). GPT-4 emerges as the leading algorithm, with a score substantially higher than the others.

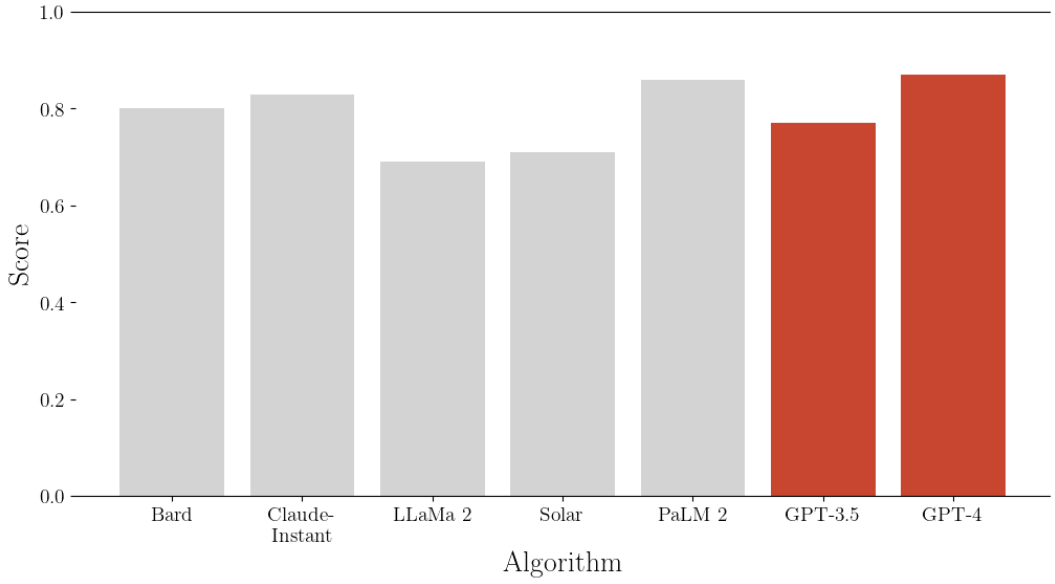
When we move to lower levels of disaggregation in Figures A6b (HS 4-digits) and A6c (HS 2-digits), the differences relative to GPT-4 become smaller. While GPT-4 remains the leading algorithm, Palm 2 (86%) is only 1 percentage point behind at the HS 4-digit level and show equal precision at HS 2-digit level (92%). All algorithms perform well at the HS 2-digit level, indicating a remarkable increase in accuracy in more aggregate categorizations. LlaMa 2, for example, assigns the correct HS 6 digit-code only in 15% of the cases, but its accuracy surges to 69% at HS 4-digit level and to 83% at the HS 2- digit level.

[25]Bard AI is a conversational AI chatbot developed by Google AI. It is powered by PaLM 2, a 540-billion parameter model, created by Google Research and trained with the Pathways system.
[26]Claude is an LLM developed by Anthropic with roughly 175 billion parameters.
[27]LlaMa 2 is a family of almost open-source LLMs (excluding commercial use). Here we used the 70-bilion parameter version.
[28]Solar-0-70b-16bit is a fine-tuned version of LlaMa 2 and a top-ranked model on the HuggingFace Open LLM leaderboard.

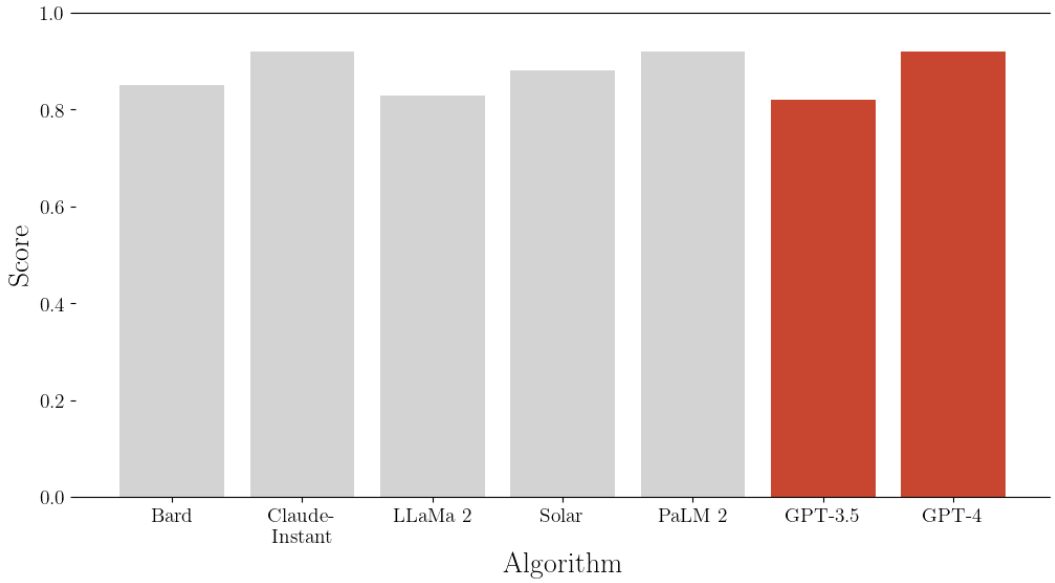
Figure A6: Comparative Performance of Other LLMs in HS6 Product Classification



(a) HS-6 level



(b) HS-4 level



(c) HS-2 level

Source: Authors' calculations based on Chilean customs data.