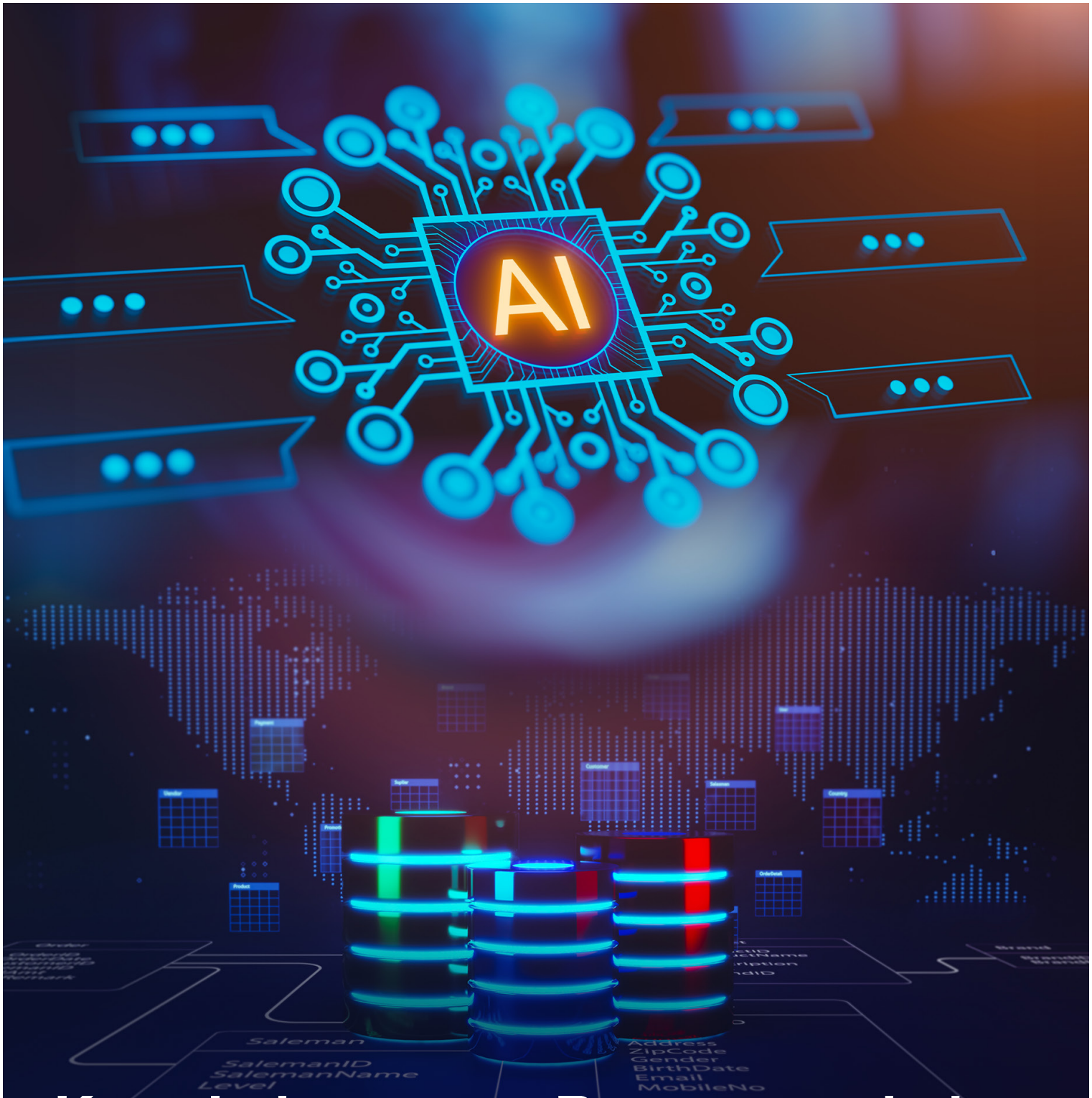


Knowledge-aware Recommendation Using Language Models:

A case of harmonized system code prediction

Whitepaper by

Chirag Tubakad, Data Scientist – Implementation of Solution
Sai Barath Sundar, Manager Data Science – Solution Development
Mohd. Juned Khan, (IIT-Kanpur) – Work done as part of internship at Mphasis NEXT Labs
Dr. Udayaadithya Avadhanam, Principal – Advisor
Bert Hooyman, Distinguished Fellow – Advisor
Biju Mathews, Partner – Advisor
Dr. Archisman Majumdar, Principal – Advisor

Mphasis
The Next Applied

Contents

1. Introduction

Categorization is the cognitive task of assigning items into groups based on shared characteristics. The task requires abstracting the characteristics. This comes from experience as well as reasoning through explicit rules. To facilitate organization, knowledge bases across various domains and processes employ taxonomies and coding systems. These systems enable easier search and retrieval of information, streamlining the categorization process.

A few examples of coding systems are:

- In healthcare, ICD (International Classification of Diseases) codes are used to categorize and classify diseases, symptoms and procedures. The coding system is used generally for analyzing mortality or disease patterns. However, its use also extends to processing claims and reimbursements.
- SIC (Standard Industrial Classification) is a system used to classify and categorize businesses and industries by their primary economic activity. SIC codes can be used to compare and analyze competitors in the same industry. They are used by the SEC (US Securities and Exchange Commission) to identify review responsibilities for a company's filings.
- In this paper, we are taking the case of HS (Harmonized System) codes. Harmonized System codes, also known as HS codes, are a standardized system of numerical codes applied to classify goods traded that cross borders. Customs officials, logistical companies, importers, exporters and border security authorities use HS codes to identify and track goods transiting across borders. The World Customs Organization (WCO) developed the HS code system, and it is employed in both domestic and international trade.

Knowledge-based (or expert) systems vs. data-driven systems

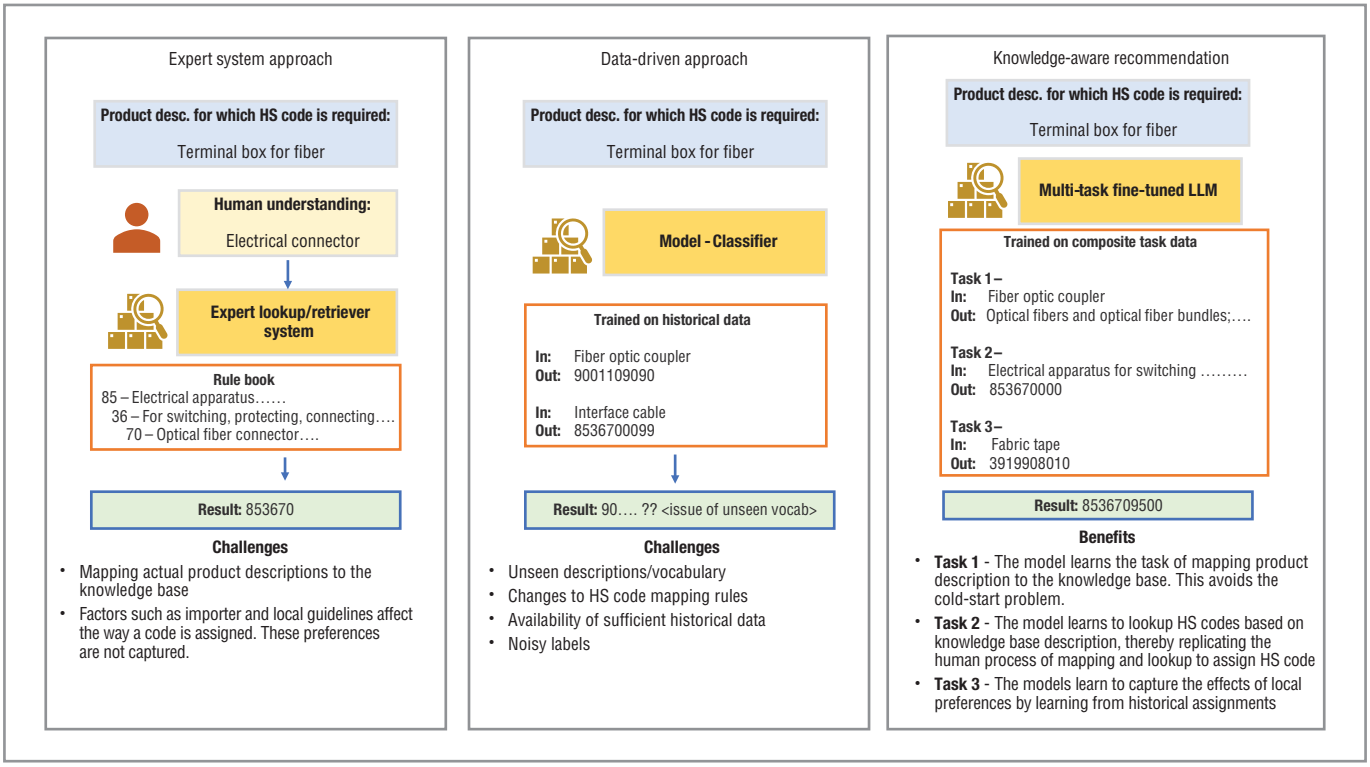
Traditionally, knowledge-based (or expert) systems were built using rules and ontologies that are codified by getting inputs from Subject Matter Experts (SMEs). While these systems ease the search and retrieval process, they are limited to looking up on the knowledge base alone – this meant one needed to know 'how to query' and map the inputs to the rules in the systems.

With the adoption of ML/AI (Machine Learning/Artificial Intelligence), large transactional data could be used to infer the rules and mimic the querying process. This purely data-driven approach has a few drawbacks – it cannot capture changes to the underlying coding schemes/rules, it assumes that all the required patterns are present in the training data and doesn't adapt to noisiness in the training data.

A hybrid "knowledge-aware recommendation" approach

To address these challenges, we present a hybrid approach that combines knowledge-driven and data-driven approaches to classification. This paper details the development and application of a novel approach for the prediction of HS codes from the description of products in commercial invoices by employing Large Language Models (LLMs). The solution is designed around a composite task model to address two key issues: (1) knowledge base mapping (incorporating rules engine like expert systems) and (2) historical best practice mapping (patterns in historical data). Large Language Models (LLMs) have shown capabilities of performing multiple cognitive tasks and cross-learning between multiple tasks. Being auto-regressive in nature, LLMs suit well for coding systems that are hierarchical in nature. The composite task model is trained on three distinct yet interrelated tasks: mapping product descriptions to the appropriate category within the rule book, predicting the HS code from the rule book description and predicting the HS code from the product description.

The figure below shows the differences between the three types of systems:



The proposed solution thus achieves the following: (1) enhances the precision in HS code assignment, (2) encapsulates historical best practices, (3) ensures best learnings and best practices are captured from different individuals performing the task historically, (4) adapts to changes in the rule book, ensuring the assignment of updated HS codes, and (5) maintains data privacy due to local development.

2. Background

There are two primary motivations for using HS code in trade. First, to standardize the classification of goods traded across countries globally and reduce the ambiguity in referencing a particular product. Second, to determine the tariffs and taxes applicable for the goods imported or exported.

HS codes are assigned to goods based on nature, composition and intended use of the goods. The HS codes are organized into sections, chapters and subheadings, with each code representing a specific type of product. For example, a laptop may be classified under HS code 8471.30, whereas a bicycle is classified under HS code 8712.00. The current HS codes manual consists of over 5000 articles and product types with a unique six-digit code assigned to a product type and the remaining 4 digits are specific to the importing country. Refer to Figure 1 for another example of the decomposition of an HS code.

The WCO monitors regulates and periodically updates HS codes, to reflect changes in technology, trade and addition or omission of products in a timely manner. In addition, the WCO provides training and support to customs officials and other stakeholders to ensure the accurate and consistent assignment of HS codes in international trade.

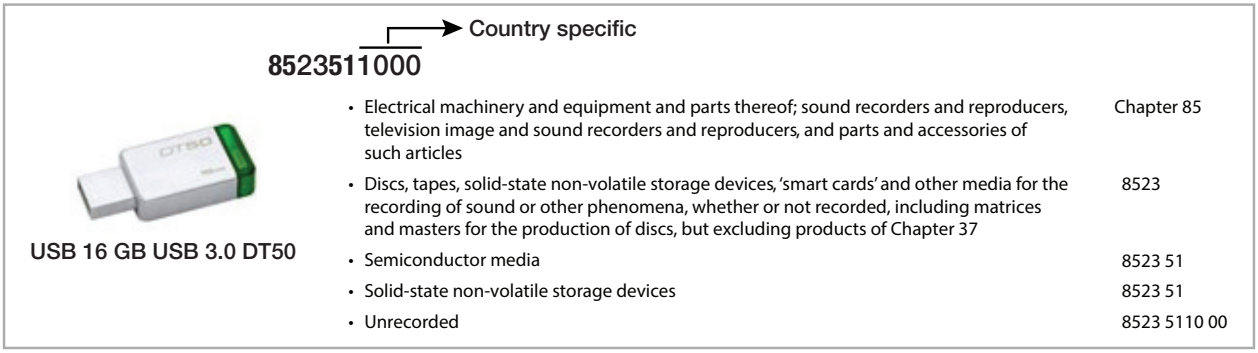


Figure 1: HS code breakdown

3. Business Process Challenges

Traditional approaches to identify the correct HS code for a product come with two business process challenges, and impact four key performance indicators discussed below:

Knowledge base mapping:

The process of HS code knowledge-base mapping to products is affected due to two reasons. First, the WCO periodically updates the HS codes of products depending on the product application. The business process must ensure that the latest HS codes from the WCO are assigned to products. Second, due to global variations in business processes, linguistic and cultural practices, both importers as well as exporters have different ways of describing a particular item. Such practices, introduce linguistic ambiguities in the description of items or products on a commercial invoice. Therefore, differences in the interpretation of product description and HS code lead to different HS code assignments resulting in rework and delays.

Historical best practice mapping:

The rules for the last four digits of the HS code are country-dependent. Therefore, human intervention is currently in practice to interpret the HS code when the last four digits are appended with the first six digits of the HS code. Furthermore, the parties involved in the transit of goods – importers, exporters and government – employ highly-trained professionals and invest in infrastructure with instructions to assign the HS codes for goods. When the trained professionals retire or quit, the years of accumulated knowledge leave with them. Training a new person is an expense in resources and time. Therefore, human dependency on trained professionals may be reduced to improve the process efficiency and minimize the expenses for all parties involved in this process.

KPIs impacted:

The two business process issues negatively impact the following business KPIs:

- Processing time: An increase in global trade across geographies and product categories, increases the labor intensity and manual interventions to assign HS codes for all products under exports and imports, and causes delays in processing the products at the borders of importing and exporting countries.
- Operational cost: Ambiguity in interpretation and incorrect HS code leads to the seizing of products at the entry/exit points of borders and holding for further clarifications. A holding fee must be paid for such shipments until clearance, which is an additional unforeseen expenditure for the suppliers.
- Regulatory constraints: For certain types of products such as pharmaceutical drugs and hazardous substances, importers/exporters must provide additional documentation for transit. Identifying such products early in the process empowers the suppliers to speed up the process of moving goods quickly from ports of entry/exit.
- Training cost/time for employees: Companies, partners and government agencies involved in HS codes for products need to retrain existing employees when HS code regulations change, and train new employees when the volumes of products increase for imports/exports. Either of the cases involves expenses on training and the risk of losing trained professionals due to attrition impacts.

Businesses and government agencies involved in the transportation of goods must address the challenges associated with the effective implementation of the HS code to facilitate seamless import and export processes within a country.

4. Problem Statement

To accurately identify HS code based on the natural language description of a product.

Technical challenges

Two technical challenges to solve the problem are: (1) systems challenge – the current IT systems do not have the ability to adapt to the complexities in the HS code of products, and (2) data challenges. The two technical challenges are elaborated below:

Current system challenges:

- The traditional IT search systems are not designed to effectively address the search for products. Reliance on keyword searches to find HS code from historical data or the WCO tariff rule book does not always result in a perfect match for product descriptions.
- Predictive Machine Learning solutions built on historical data struggle to comprehend the nature of a product and only match product text to HS code number combinations and pose a challenge while dealing with new product names and descriptions
- Machine Learning solutions cannot efficiently classify across thousands of categories (as the codes run to 10 digits)

Data challenges:

- The data challenges are due to two reasons. First, multiple products with single HS codes. For example, wristwatches are categorized in chapter 91 under the section: 'Clocks and watches, & parts thereof'. In case a vendor imports raw materials to assemble wristwatches in the destination country, all the raw materials fall under the same HS code as wristwatches. Therefore, multiple products are mapped to a common HS code based on the purpose of consumption of the products. Second, a single form of product with multiple HS codes. For example, aluminum is utilized in manufacturing aircrafts, cars and different industrial materials, and each of them has a different HS code for aluminum.

Capturing the nuances and processes in assigning HS code to a variety of products is hard for existing IT search systems and traditional Machine Learning approaches. Therefore, the large language model capabilities are explored to solve the HS code problem.

Large language models for sequence prediction

LLMs are known for generic text responses to conversational prompts. Solving HS code-related challenges is a niche problem for LLMs. The following six questions are explored while using LLMs to solve HS code problems.

- Utilize the generic pipelines to fine-tune pre-trained LLMs to identify HS code accurately
- Narrow down the generic sentence understanding capabilities of LLMs in smaller fine-tuned models for HS code-specific problem statements
- Capture the language/linguistics of customs officers, and map certain phrases or descriptions to a common HS code
- The training strategy to solve the HS code problem given the language ambiguity in the data
- The autoregressive decoder-only LLMs have the ability to understand tasks performed in identifying HS code from reviewing the chapter, heading, title and so on in a document. Explore the ability of autoregressive decoder-only LLMs to predict the next token depending on the preceding token.
- Review the performance of autoregressive decoder-only LLMs at the task of predicting HS codes as a sequence of tokens rather than approaching the problem statement as a classification problem

5. Solution

Data sources

Synthetically generated historical data mimicking real data from a logistics firm and the HS code document from the UK government as the rules book for the HS code problem.

LLM tasks

Instead of directly predicting the HS code of a product given a description, we broke the process into three composite tasks. Therefore, we had a dataset consisting of a mixture of tasks. The tasks were as follows:

- Historical to rules book: Given the historical description of a product, the task was to correctly identify the closest rules book product category description for the product
- Rules book to HS code: Given the rules book description for a product category, the goal was to sequentially predict the HS code token by token
- Historical to HS code: Given the description of a product, directly predict the HS code of a product without looking into the rules book for reference

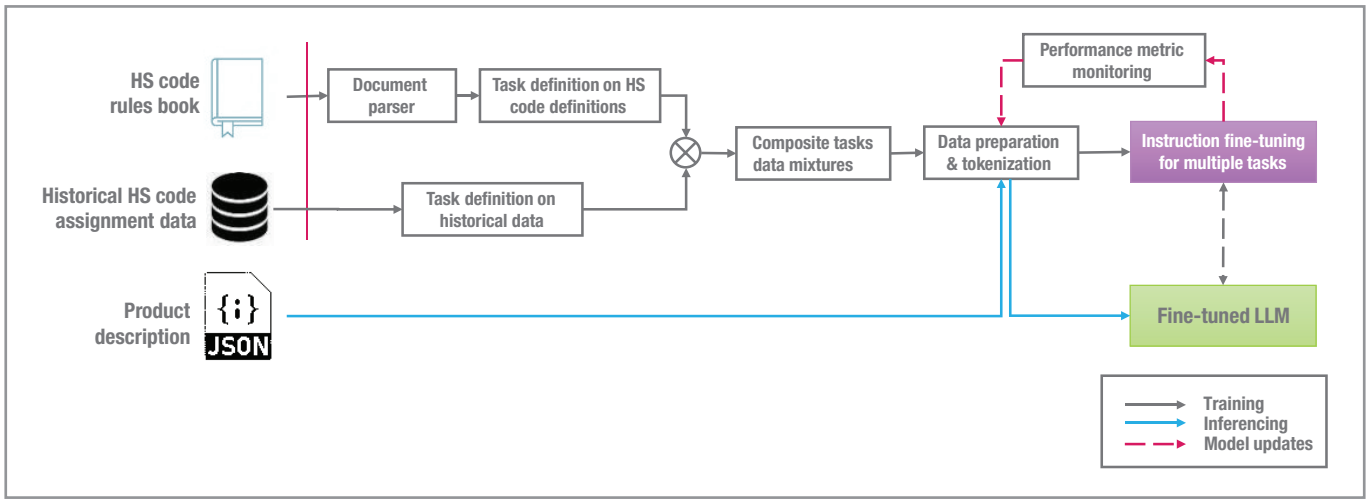


Figure 2: Solution pipeline for HS code model fine-tuning

The solution pipeline is built on Flan-T5, a small and robust model for autoregressive token prediction. Various training strategies and hyper-parameter tuning techniques are employed to find the perfect fit for the model. During the training phase, two versions of the model are created, one for two composite tasks and the other for all three composite tasks. The training pipeline of this process is given in Fig 2.

There are several ways to infer using the trained model depending on the prompt that is appended to the product description. Some of the inferencing strategies are listed below:

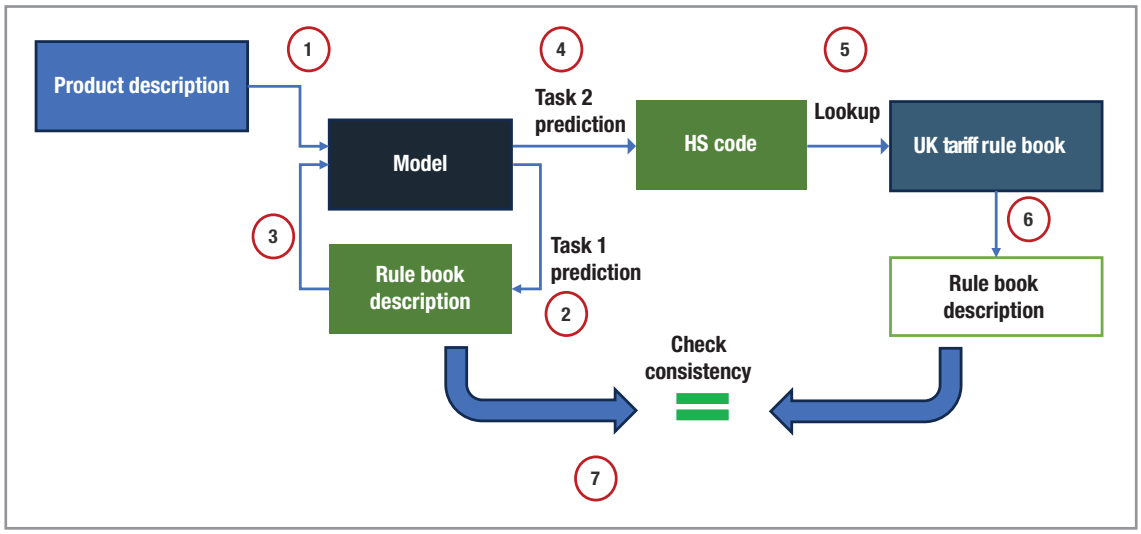
- Model with 2 composite tasks: Given the description of a product, first generate the accurate rules book description using beam search and then map that description to the right HS code using greedy search
- Model with 3 composite tasks: Given the description of a product, identify if there are any prior goods processed with the same description, then map that to the correct rules book description using beam search. This is then followed by identifying the right HS code for the product using greedy search. It consists of three steps in total.
- Model with 3 composite tasks: Given the description of a product, directly map it to the right HS code

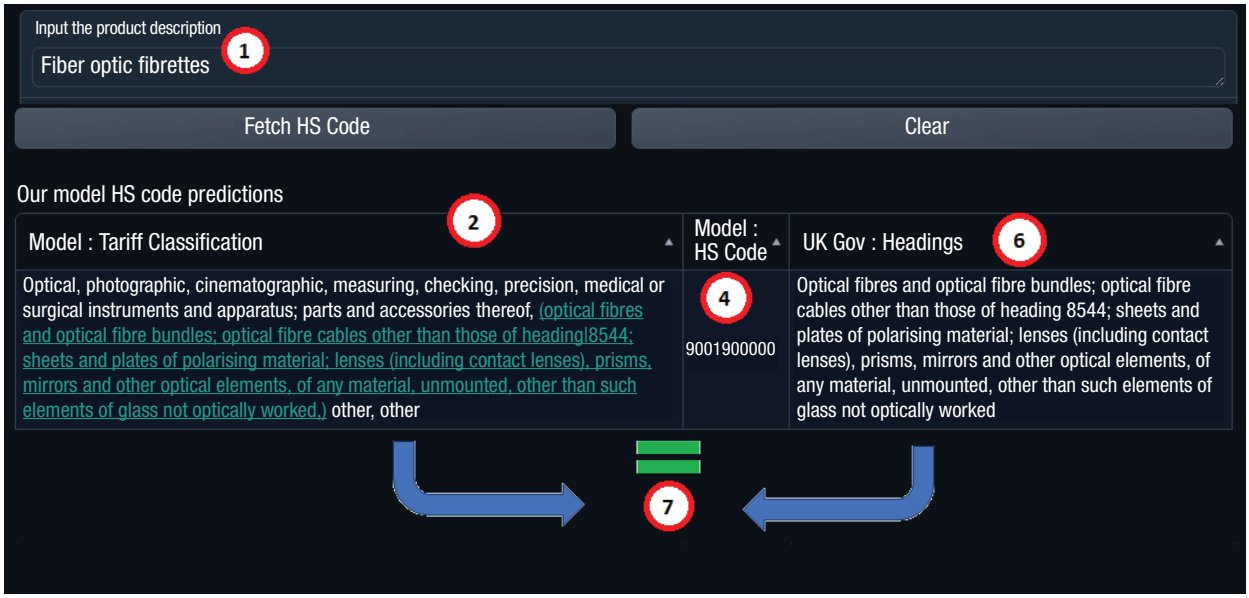
By performing a beam search, the inference becomes a recommendation of top 'n' possible HS codes, thus a recommendation engine.

Self-consistency from cross-learning

The rationale behind composite tasks is to induce cross-learning over three tasks. First, learn the relationship between a product description and its category (ontology). Second, learn the relationship between ontology and the chapter and sub-chapter numbers, and third, cross-learning the direct relationship between production description and chapter and sub-chapter numbers.

To evaluate whether the model has learned the relationships, we can test by checking if the model is self-consistent:

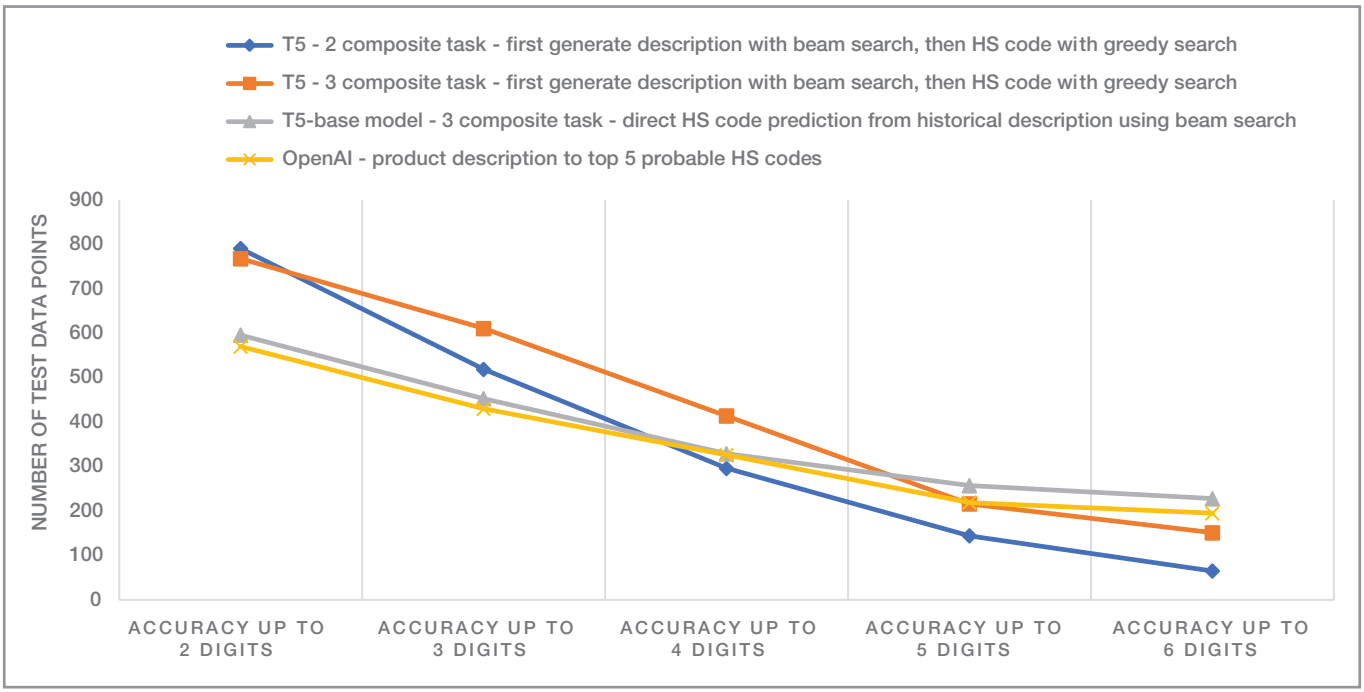




The results for the developed framework show a high degree of consistency, implying language understanding and domain understanding by cross-learning.

6. Results

The results across all the different model training procedures and inferencing strategies are shown below:



Evaluation method: Accuracy up to 'n' digits is computed by taking the top 5 recommendations from the model and comparing each of them with the ground truth (using the longest sequence match). The recommendation having the longest sequence match is considered for counting under the categories shown in the chart. The test dataset has 1000 data points. *It should be noted that a data point at a higher accuracy bucket would be double-counted at a lower accuracy bucket. For example: in the first strategy with two composite tasks, the model could predict accurately the first 2 digits for ~790/1000. At three digits, ~520/1000 were predicted accurately. Hence, the 520 is also part of 790, i.e., 790 is correct at 2 or more digits.*

The results indicate that the fine-tuned models can outperform OpenAI off-the-shelf inferencing. Flan-T5 is open source, small, requires less computational resources and performs well in local deployment without data leaving the enterprise system.

Let us look at the performance of the model in a few interesting cases:

Case 1: Ambiguous product description having multiple possible HS codes:

Product description	Top 3 - HS codes recommended by our model	Different ground-truths	Description from tariff rule book (at the level of matching digits)
Aircraft parts	7318158190	7318290090	Screws, bolts, nuts, coach screws, screw hooks, rivets, cotters, cotter pins, washers (including spring washers) and similar articles, of iron or steel
	8411222000	8411910090	Turbojets, turbo-propellers and other gas turbines
	9401910000	9401901090	Seats (other than those of heading 9402), whether or not convertible into beds, and parts thereof

Aircraft parts could refer to a range of different components being shipped for an aircraft. The model's top few recommendations capture these possibilities from historical data understanding.

Case 2: Product description contains unique variant/style of a product type:

Product description	Top 3 - HS codes recommended by our model	Description from tariff rule book	Ground-truth	Description from tariff rule book
Bohemian beachcomber	6110909900	Jerseys, pullovers, cardigans, waistcoats and similar articles, knitted or crocheted	6104520000	Women's or girls' suits, ensembles, jackets, blazers, dresses, skirts, divided skirts, trousers, bibs and brace overalls, breeches and shorts (other than swimwear), of cotton
	6204399090	Women's or girls' suits, ensembles, jackets, blazers, dresses, skirts, divided skirts, trousers, bibs and brace overalls, breeches and shorts (other than swimwear)		
	6105209000	Men's or boys' shirts, knitted or crocheted		

A Bohemian Beachcomber - a unique type of fashion apparel referring to a certain lifestyle and occasion (other examples – Bohemian tapestry, rugs, etc.). The language model being pre-trained on a vast corpus of web-crawled data (like C4) can make the connection to the fashion and clothing category of the tariff rule book.

Case 3: Product descriptions that can be worded similar but are different in nature:

Product description	Top 3 - HS codes recommended by our model	Ground-truth	Description from tariff rule book (at the level of matching digits)
Fiber optic fibrettes	9001900000	9001109090	Optical fibers and optical fiber bundles; optical fiber cables other than those of heading 8544; sheets and plates of polarizing material; lenses (including contact lenses), prisms, mirrors and other optical elements, of any material, unmounted, other than such elements of glass not optically worked
Fiber optic terminal box	8536100000	8536909599	Electrical apparatus for switching or protecting electrical circuits, or for making connections to or in electrical circuits (for example, switches, relays, fuses, surge suppressors, plugs, sockets, lamp holders and other connectors, junction boxes), for a voltage not exceeding 1000 V; connectors for optical fibers, optical fiber bundles or cables

A fiber optic cable and a terminal box that connects these optic cables fall under different categories. The results show that the model can differentiate the two products even though they share the words “Fiber optic”.

Case 4: Product descriptions with out-of-training vocabulary:

Product description	Top 3 - HS codes recommended by our model	Description from tariff rule book
James Clavell	4901908000	Printed books, brochures, leaflets and similar printed matter, whether or not in single sheets

In this test case, we provided the model with the name of a well-known British author – James Clavell. The words used aren’t part of our training corpus. The model is able to relate the author’s name to the commodity – books. This example shows that using language models can help avoid cold-start problems.

Case 5: Product descriptions that result in a high degree of accuracy:

Product description	Top 3 - HS codes recommended by our model	Ground-truth	Description from tariff rule book (at the level of matching digits)
Auto parts - injector	8708999099	8708999300	Parts and accessories of the motor vehicles of headings 8701 to 8705
Steel nuts and bolts	7318158190	7318156810	Screws, bolts, nuts, coach screws, screw hooks, rivets, cotters, cotter pins, washers (including spring washers) and similar articles, of iron or steel

In these examples, the model’s accuracy is six digits and above. (i.e., more than three levels of hierarchy from the tariff rule book). There are two reasons: (1) the product descriptions are worded closer to the tariff rule book descriptions and (2) historical training data has multiple occurrences of similar products. These examples thus give confidence that the model can be improved over time with appropriate training data preparation.

7. Conclusion

In this paper, we presented a novel approach for predicting Harmonized System (HS) codes from product descriptions using Large Language Models (LLMs). The proposed solution addresses the challenges faced by traditional knowledge-based and data-driven systems by combining both approaches in a hybrid "knowledge-aware recommendation" model.

The composite task model, trained on three distinct yet interrelated tasks, enables the LLM to learn the relationships between product descriptions, ontology and HS codes. This cross-learning approach ensures that the model captures the nuances and processes involved in assigning HS codes to a wide variety of products.

The results demonstrate that the fine-tuned models outperform off-the-shelf inferencing while being open-source, small and requiring less computational resources. The model's ability to handle ambiguous product descriptions, unique product variants and out-of-vocabulary terms highlights its robustness and adaptability.

8. Next Steps

In future courses, the solution will be extended into personalized recommendations and solve similar problems when the knowledge base is updated – i.e., adapting to changes to the rule book. Additionally, the proposed approach could be evaluated on other domains and coding systems, demonstrating its versatility.

9. References

World Customs Organization (wcoomd.org)

Trade Tariff: look up commodity codes, duty and VAT rates - GOV.UK (www.gov.uk)

[2306.12925] AudioPaLM: A Large Language Model That Can Speak and Listen (arxiv.org)

[2210.11416] Scaling Instruction-Finetuned Language Models (arxiv.org)

About Mphasis

Mphasis' purpose is to be the *"Driver in the Driverless Car"* for Global Enterprises by applying next-generation design, architecture and engineering services, to deliver scalable and sustainable software and technology solutions. Customer centricity is foundational to Mphasis, and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization, combined with an integrated sustainability and purpose-led approach across its operations and solutions are key to building strong relationships with marquee clients. Click here to know more. (BSE: 526299; NSE: MPHASIS)

For more information, contact: marketinginfo.m@mphasis.com

USA
Mphasis Corporation
41 Madison Avenue
35th Floor, New York
New York 10010, USA
Tel: +1 (212) 686 6655

UK
Mphasis UK Limited
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Mphasis Limited
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village, Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

IAS 13/06/24 US LETTER BASIL 9677

www.mphasis.com