Title:

Enhancing the readability of HS descriptions: Automatic Text Summarisation using LLMs

Abstract

This paper presents an automatic text summarisation method to generate a lexically simplified and shorter version of product descriptions in the Harmonised System (HS) using Large Language Models (LLMs). The results demonstrated that GPT-4 is effective in enhancing the readability of HS descriptions. On a scale of zero to five, the model was observed to reduce their complexity by 1.09 points when evaluated by machine-emulated high school student having an intermediate level of English and by 0.53 points when assessed by human trade professionals. Moreover, the summarisation has markedly enhanced the readability of highly complex descriptions, with a score of 5, which accounted for 14.8% of the dataset after the summarisation process instead of 36.7% initially. The use of semantic textual similarity as an automatic evaluation metric, coupled with human-perceived informativeness, has ascertained that the summaries are semantically and factually consistent with the original HS descriptions. Basic BERT and PEGASUS are limited in their ability to generate the desired lexically simplified and summarised version of the HS descriptions. On the one hand, BERT generates semantically inconsistent summaries even after increasing the candidate pools, improving complexity detection and relaxing filtering thresholds. For PEGASUS, paraphrasing gives the best results in terms of semantic preservation, but the structure of the output often does not match the original input. Consequently, both BERT and PEGASUS may require more advanced parameter adjustment and fine-tuning to increase their efficiency. The lexically simplified and summarised HS descriptions resulting from our ATS procedure, designated as the SIMPLEX, can be utilised in a multitude of trade compliance software and web-applications, such as a new generation of HS Search Engine, thereby expediting the reading and processing of the product descriptions or serving as a training dataset of a similarity search engine. Our ATS procedure could also be applied to reduce the complexity of other classification systems of goods and services used in other domains, such as the Nice Classification (NCL) for trademark registration or the International Standard Industrial Classification of all Economic Activities (ISIC).

Keywords (4-6 keywords)

Artificial Intelligence, Automatic text summarisation, Large Language Models (LLMs), Harmonized System, Product classification, International Trade.

Authors:

NGAVOZAFY Mampiandra Antonia (1,2)

MAMLOUK Youssef (3)

LATORRE Maria Conception (1,4)

Affiliations:

1. Universidad Complutense de Madrid
2. International Trade Centre (ITC)
3. Ecole Polytechnique Fédérale de Lausanne (EPFL)
4. Real Colegio Complutense at Harvard University

Enhancing the readability of HS descriptions: Automatic Text Summarisation using LLMs

1. Introduction

The Harmonized System (HS) is a structured product description and coding system edited by the World Customs Organization (WCO). The HS provides a uniform method of identification of all commodities exchanged internationally, forming the basis for tariffs and taxes levy in over 200 countries and customs territories and the collection of trade statistics of over 98% of merchandise. Additionally, the HS serves as a universal language in the negotiation and implementation of international treaties, trade agreements, and regulations. Furthermore, the HS is a crucial component of trade compliance, as the six-digit HS codes are a mandatory requirement for all global Cargo.

However, the HS classification process remains predominantly manual, necessitating substantial financial and human investments. Additionally, non-compliance with the HS has severe legal and pecuniary consequences, either directly in the form of fines or indirectly resulting from the disruption of the supply chain and the loss of consumers due to a deterioration in reputation. One significant challenge in HS classification is the inherent complexity of product descriptions. Indeed, Thomson Reuters and KPMG International (2016) revealed that 91% of traders encounter difficulties primarily because of unclear or ambiguous product descriptions. This complexity of product classification is primarily attributable to the length of the product descriptions and the use of jargon and non-commercial language in the nomenclature, making them challenging to interpret accurately. This situation highlights the necessity for solutions that facilitate clarity and accessibility in the product classification process.

The lexical simplification and summarisation of the HS descriptions is one potential solution to enhance the readability and the interpretation of the original nomenclature. Nevertheless, the 2022 edition of the HS comprises 5612 six-digit codes and associated descriptions, which are revised every five years. Their manual summarisation would be an unduly onerous task for a human expert. Automatic text summarisation (ATS) can alleviate this burden. In recent years, the field of ATS has made notable advancement thanks to the progression of Natural Language Processing (NLP) and the emergence of new Large Language Models (LLMs), which have demonstrated impressive capabilities in understanding and generating human-like texts. Notable examples of LLMs include the Generative Pre-trained Transformer (GPT) from OpenAI, as well as the Bidirectional Encoder Representations from Transformers (BERT) and the Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) by Google. They have become indispensable tools for solving practical problems in many industries, including in HS classification (Grainger, 2024). However, the direct application of LLMs to summarise HS descriptions presents significant challenges arising from their hierarchical structure, the diversity of commodities they describe, and the essential nature of every word they contain essential for the overall meaning.

This paper presents an approach to LLMs-powered abstractive summarisation of HS descriptions at the six-digit level, using the 2022 edition of the Harmonised System in English, published by the WCO, as an input dataset. Furthermore, it assesses LLMs' capabilities in improving the readability of complex HS descriptions through the utilisation of automatic and human evaluation metrics.

Our work contributes to the existing literature on the automatic text summarisation using artificial intelligence models in two ways. On the one hand, it broadens the scope of applications of ATS to encompass international trade and beyond the conventional domains of sentence, document, dialogue, speech, timeline, and code summarisation. This study is indeed the first to use this technique in the field of trade facilitation and product classification. On the other hand, it contributes to the future development of ATS evaluation metrics by introducing a new procedure for evaluation that is specific to HS classification. Furthermore, this paper makes a significant contribution to the existing literature on the use of Artificial Intelligence in international trade and its potential for facilitating trade compliance and enhancing market access. It proposes a concrete and efficient solution to ease the navigation through the often-complex international regulations. The resulting lexically simplified and summarised HS descriptions, designated as the SIMPLEX, can be utilised in a multitude of trade compliance software and web-applications, thereby expediting the reading and processing of the product descriptions.

The remainder of this paper is organised as follows. Chapter 2 lays out the methodology. Then, Chapter 3 presents the results. Chapter 4 is dedicated to the evaluation. Finally, Chapter 5 concludes.

2. Methods

2.1. Dataset

2.1.1. The Harmonised System

The Harmonised System (HS) is a widely used abbreviation for the Harmonised Commodity Description and Coding System, an international nomenclature edited by the World Customs Organisation (WCO) that enables the uniform identification of all internationally traded commodities. The HS consists of a series of four-digit headings, the majority of which are subdivided further into five- and six-digit subheadings (WCO, 2018). The HS also comprises legal notes to facilitate its interpretation.

As is common in many sectoral classification systems, the HS adheres to a hierarchical structure. The headings (HS4), identified with four-digit codes, constitute the initial legal level of classification, and all goods must be assigned to one. In the 2022 edition of the Harmonised System, over 95% of headings are further subdivided into five-digit and six-digit subheadings (HS6), which constitute the final legal level of classification harmonised internationally. The 2022 edition of the HS contains a total of 5612 HS6 codes. Figure 1 provides a visual representation of this hierarchical structure.

Figure 1: Hierarchical structure of the HS

Heading	H.S. Code	
84.70		**Calculating machines and pocket-size data recording, reproducing and displaying machines with calculating functions; accounting machines, postage-franking machines, ticket-issuing machines and similar machines, incorporating a calculating device; cash registers.**
	8470.10	- Electronic calculators capable of operation without an external source of electric power and pocket-size data recording, reproducing and displaying machines with calculating functions
		- Other electronic calculating machines :
	8470.21	-- Incorporating a printing device
	8470.29	-- Other
	8470.30	- Other calculating machines
	8470.50	- Cash registers
	8470.90	- Other
84.71		**Automatic data processing machines and units thereof; magnetic or optical readers, machines for transcribing data onto data media in coded form and machines for processing such data, not elsewhere specified or included.**
	8471.30	- Portable automatic data processing machines, weighing not more than 10 kg, consisting of at least a central processing unit, a keyboard and a display
		- Other automatic data processing machines :
	8471.41	-- Comprising in the same housing at least a central processing unit and an input and output unit, whether or not combined
	8471.49	-- Other, presented in the form of systems
	8471.50	- Processing units other than those of subheading 8471.41 or 8471.49, whether or not containing in the same housing one or two of the following types of unit : storage units, input units, output units
	8471.60	- Input or output units, whether or not containing storage units in the same housing
	8471.70	- Storage units
	8471.80	- Other units of automatic data processing machines
	8471.90	- Other
84.72		**Other office machines (for example, hectograph or stencil duplicating machines, addressing machines, automatic banknote dispensers, coin-sorting machines, coin-counting or wrapping machines, pencil-sharpening machines, perforating or stapling machines).**

Source: WCO, 2022

As illustrated in Figure 1, the denomination provided after each HS6 code is incomplete to adequately identify the product defined by the HS6 code. In fact, the text "*Comprising in the same housing at least a central processing unit and an input and output unit, whether or not combined*" following the HS6 code 8471.41 is merely a component of the description. To construct the full description, it is necessary to combine the latter with the denominations associated with the parent codes at the five- and four-digit levels. As a result, the complete description of 8471.41 is "*Automatic data-processing machines and units thereof; magnetic or optical readers, machines for transcribing data onto data media in coded form and machines for processing such data, not*

elsewhere specified or included : Other automatic data-processing machines : Comprising in the same housing at least a central processing unit and an input and output unit, whether or not combined". This is what is commonly referred to in practice as a hierarchical description. The hierarchical descriptions associated with the 5612 HS6 codes under the 2022 edition of the HS (WCO, 2022), as published by the WCO on 1 January 2022, is the primary input dataset of this study.

However, this hierarchical structure of the HS6 descriptions poses a significant challenge for the direct application of pre-trained LLMs to obtain concise and accurate summaries of the product descriptions. Therefore, the linearisation of hierarchical descriptions in order to generate their self-explanatory version that will be more straightforward to summarise efficiently is necessary. Table 1 provides examples of HS descriptions both before and after linearisation, for a selection of products.

Table 1: Hierarchical versus linearised HS descriptions

Example n°	HS6 code	Hierarchical Description	Linearised Description
1	0102.31	Live bovine animals : Buffalo : Pure-bred breeding animals	Pure-bred buffalo for breeding
2	0105.11	Live poultry, that is to say, fowls of the species Gallus domesticus, ducks, geese, turkeys and guinea fowls: Weighing not more than 185 g : Fowls of the species Gallus domesticus	Live fowls of the species Gallus domesticus, weighing <= 185 g (excl. turkeys and guinea fowls)
3	2517.20	Pebbles, gravel, broken or crushed stone, of a kind commonly used for concrete aggregates, for road metalling or for railway or other ballast, shingle and flint, whether or not heat-treated; macadam of slag, dross or similar industrial waste, whether or not incorporating the materials cited in the first part of the heading; tarred macadam; granules, chippings and powder, of stones of heading 2515 or 2516, whether or not heat-treated : Macadam of slag, dross or similar industrial waste, whether or not incorporating the materials cited in subheading 251710	Macadam of slag, dross or similar industrial waste, whether or not incorporating pebbles, gravel, shingle and flint for concrete aggregates, for road metalling or for railway or other ballast

The HS is concerned with the precise naming and description of commodities, employing legal terminology alongside complex industry-specific and scientific vocabulary. For instance, the

product description associated with the HS6 code 0105.11 is "*Live poultry, that is to say, fowls of the species Gallus domesticus, ducks, geese, turkeys and guinea fowls: Weighing not more than 185 g: Fowls of the species Gallus domesticus*", which is simply *domestic chicken* in the common English language. Consequently, the HS classification process necessitates comprehensive knowledge and expertise, making it challenging for laypeople. Besides, the use of jargon and non-commercial vocabulary raise some concerns about the consistency of their interpretation internationally (WCO, 2019).

Furthermore, the complexity of product classification is also attributable to the length of the product descriptions, making them challenging to interpret accurately. The challenge persists even when the product descriptions are linearised into their self-explanatory version, which may contain up to 111 words and 8111 characters. Consequently, the reader dedicates more time to read and fully understand the described product. In an experiment we conducted in 2023 with 20 international trade professionals who were already familiar with the HS nomenclature, fluent in English, and held at least a master's degree in economics, management, law, or statistics, we found that participants required an average of 22.4 minutes to read a random sample of 60 linearised product descriptions of 3190 words. This equates to an average reading speed of 142 words per minute (WPM), which falls in the mid-range of the 100-200 WPM average reading speed of complex texts for adults.

Therefore, our objective in this paper is to provide a lexically simplified version of the HS descriptions on one hand and to reduce their length on the other. The resulting simplified and summarised descriptions, which we name simplex descriptions, are not intended to replace the original HS nomenclature. The simplex descriptions are designed with the intention of expediting the comprehension of the HS description to facilitate the classification process and potentially serve as a training dataset for a semantic search engine of HS product codes.

2.1.2. The Flesch Reading Ease score: a traditional measure of text complexity

The complexity of the HS descriptions is the antithesis of their readability, which is defined as the ease at which they can be read and understood in terms of their linguistic features.

The Flesch Reading Ease score (Flesch, 1948) is a popular metric to evaluate text complexity. Alissa and Wald (2023) used this approach. The Flesch Reading Ease score ranges from 0 to 100, with a higher score indicating easier readability:

- 90-100: Very easy to read. Easily understood by an average 11-year-old student.
- 80-89: Easy to read. Understandable by a 13- to 15-year-old student.
- 70-79: Fairly easy to read. Understandable by a 12th to 15th-grade level student.
- 60-69: Plain English. Understandable by a 16th to 18th-grade level student.
- 50-59: Fairly difficult to read.
- 30-49: Difficult to read. Suitable for college students or experts in the field.
- 0-29: Very difficult to read. Suitable for college students or experts in the field

For a given text, the score is computed using the average number of words per sentence and the average number of syllables per sentence, as per the formula in Flesch (1948):

$$206.835 - 1.015\left(\frac{total\ \ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \quad (1)$$

The rationale behind this readability formula (1) is that the average number of words per sentence serves as a proxy for the syntactic complexity, whereas the average number of syllables per sentence should capture word sophistication.

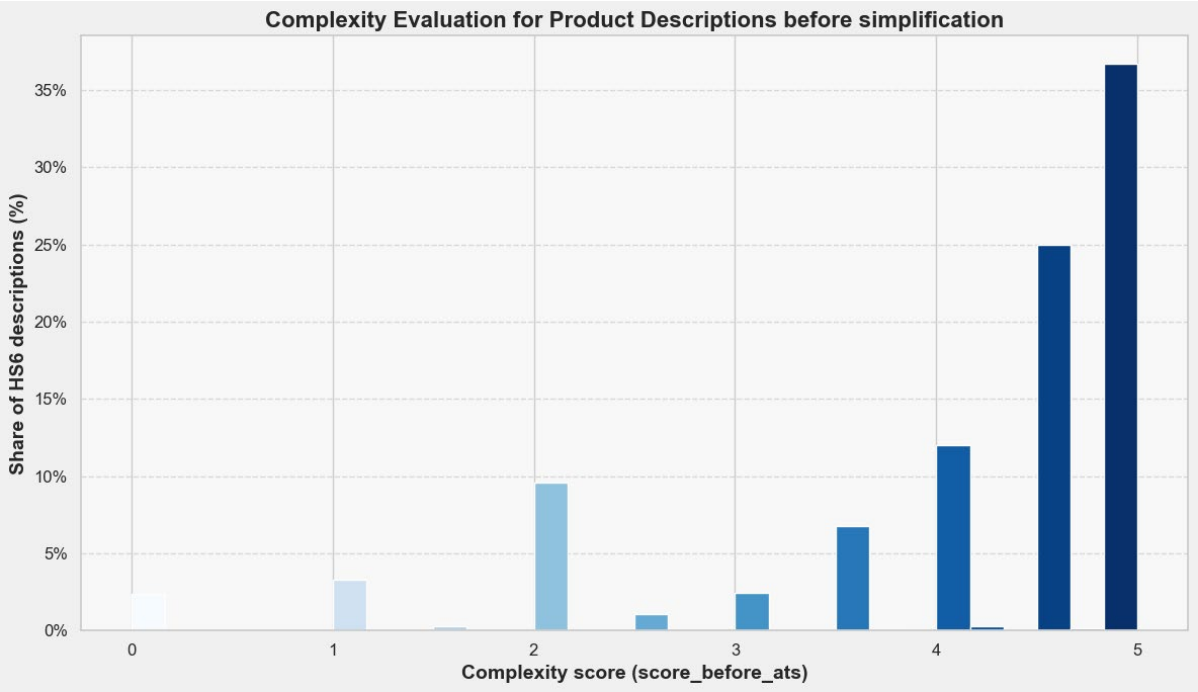
However, the Flesch Reading Ease presents several limitations. For example, it ignores the inherent relationship between elements within a text, the stylistic choices, the vocabulary and the grammar, as well as word familiarity and relevance. A complexity evaluation of the HS descriptions using this scoring system revealed numerous irregularities, including its inability to assign low scores to descriptions containing jargon and complex scientific terminology. Consequently, there is a necessity to investigate alternative approaches to complexity evaluation that are more applicable to HS descriptions.

2.1.3. Measuring the complexity of the HS descriptions

More recent studies have demonstrated that Natural Language Processing (NLP) are more reliable and effective approach for evaluating the readability of a text than traditional metrics. Crossley et al (2023) have conducted a comprehensive survey of these NLP-informed readability formulas. Considering the specific structure of the HS descriptions as exposed previously, our paper introduces a novel readability metric that is informed by the OpenAI's Generative Pre-Trained Transformer (GPT-4) and prompted through Python.

The GPT-4-based complexity metric assesses the entirety of the dataset comprising 5612 HS6 descriptions, on a continuous scale of zero to five. Through Python, we have prompted GPT-4 to rate the level of complexity of each description on a continuous scale of zero to five, with five representing the most complex and zero representing the simplest. To ensure that the evaluation captures the full range of linguistic features, we have established various sets of controls in the Python prompt, including the vocabulary sophistication, the lexical cohesion, the structure of HS descriptions, as well as the traditional words and syllables counts. For instance, words in Latin and scientific names of species such as Gallus domesticus, Xiphias gladus or Pleuronectidae were tagged as complex vocabulary. Such complexity would have been disregarded by the Flesch Reading Ease scoring metric. Furthermore, the evaluation was conducted from the perspective of an individual with an intermediate level of English proficiency (CEFR level B1) and who has completed up to a high-school degree. As a result, 84.4% of the linearised HS6 descriptions are evaluated as complex by GPT-4 under the established setting. 36.7% scores 5, the highest degree of complexity. Figure 2 depicts the scores from the GPT-4-informed complexity evaluation for the entire dataset of product descriptions.

Figure 2: Complexity of linearised HS6 descriptions, in a continuous scale of 0 to 5, where 0 = Very Simple, 1 = Simple, 2 = Rather Simple, 3 = Rather Complex, 4 = Complex, 5 = Very Complex.



Human evaluation of the results allows to confirm the reliability of the GPT-based complexity evaluation. Indeed, we have invited 17 international trade professionals to rate the complexity level of 30 linearised descriptions, randomly selected, using the same scale employed in the GPT-based evaluation. The human-rated sample contains 491 distinct product descriptions in total. The complexity levels under the two metrics appear to be consistent, although the ratings of human experts are generally less stringent than the GPT-based evaluation for an identical product description. The scores are aligned for simpler descriptions, whereas the discrepancies become more pronounced the more sophisticated are the descriptions. In fact, the GPT-4 complexity scores are 1.9 point higher than the human assessment in average, with 12% of exact score match. When the descriptions are evaluated as simple by GPT-4, i.e. scored up to 2.5, the gap with the human ratings is only 0.7 point in average. The gap widens to 2.2 point in average for complex descriptions which have received a GPT4 score higher than 2.5. This is explained by the fact that the participants to the survey were professionals who were already familiar with the HS nomenclature, fluent in English, and held at least a master's degree in economics, management, law, or statistics. Therefore, they have a higher ability to process more complex text and perceive them as simpler compared to individuals with lower qualification, such as a high-school degree only as in the configuration we set up in the GPT-4 complexity evaluation.

Furthermore, visual inspection of the GPT-4 results of complexity evaluation for the considered samples are in line with our expectation. Table 2 presents a few examples.

Table 2: Complexity scores by GPT4 vs by human Evaluators for a selection of linearised descriptions

HS6 code	Linearised Descriptions	Complexity Score by Human	Complexity Score by GPT-4
081060	Fresh durians	0	0
120910	Sugar beet seed, for sowing	0	0

010620	Live reptiles (e.g. snakes, turtles, alligators, caymans, iguanas, gavials and lizards)	0	1
820540	Hand-operated screwdrivers	1	1
030195	Live southern bluefin tunas (Thunnus maccoyii)	1	2
120760	Safflower (Carthamus tinctorius) seeds	2	2
722240	Angles, shapes and sections of stainless steel, n.e.s.	1	3
890323	Sailboats, with or without auxiliary motor, of a length > 24 m	2	3
843880	Machinery for the industrial preparation or manufacture of food or drink, n.e.s.	4	4
820340	Pipe-cutters, bolt croppers, perforating punches and similar hand tools, of base metal	3	4
251200	Siliceous fossil meals, e.g. kieselguhr, tripolite and diatomite, and similar siliceous earths, whether or not calcined, of an apparent specific gravity of <= 1	5	5
291419	Acyclic ketones, without other oxygen function (excl. acetone, butanone (methyl ethyl ketone) and 4-Methylpentan-2-one (Methyl isobutyl ketone))	5	5

2.2. Automatic Text Summarisation (ATS)

2.2.1. Principles

From a human perspective, text summarisation is a highly intellectual activity, as the process generally involves understanding the content of the text, i.e. interpretation, then identifying the most relevant information it contains, i.e. identification, and finally writing this information concisely, i.e. generation (Brandow et al., 1995). However, manual text summarisation can be daunting and inefficient, especially for large amounts of information, such as the 5612 HS descriptions that are the subject of this study. Automatic text summarisation techniques, which allows to produce abridged version of long texts or large documents, are a potential solution to the problem.

Automatic text summarisation (ATS) has regained interests in recent years. A growing number of researchers are attempting to develop ATS methods that ensure readability, fluency, informativeness and consistency. Significant advances in large language models (LLMs), which are capable of understanding, manipulating, and generating human-like text, have further increased the number of real-world applications of LLM-based ATS (Lu et al, 2023). Our study is the first to attempt LLM-based approach to summarising HS product descriptions at the very granular level of six-digit codes.

Considering the generated outputs, Nazari and Mahdavi (2019) mentioned two main methods for ATS: the extractive and the abstractive summarisation. The extractive summarisation consists of identifying, extracting, and then combining key elements without further modification of the words and the sentence structure to produce an abridged version of the original input. This method is a reasonable and simpler way of producing a summary. However, it may suffer from lack of fluency and incoherent repetition. Besides, the extractive approach is less suitable for summarising HS descriptions, since every word that forms the descriptions is relevant to definition of the commodity, and, on the other hand, our aim is also to simplify the sophisticated vocabulary used.

The abstractive method uses NLP techniques to understand the meaning of the text, then subsequently proceed to the summarisation task where key concepts are identified, interpreted, and transformed into another semantic form. In other words, the abstractive summarisation consists of rephrasing the original text into a shorter and lexically simpler version that is closer to human-generated summary ideal. Therefore, this approach is the most suitable for summarising HS descriptions. However, it may suffer from hallucination, as some models may generate factually incorrect summaries.

Our strategy to overcome this challenge involves testing multiple models and performing machine and human evaluations of the resulting summaries. The Generative Pre-Trained Transformer (GPT-4) by OpenAI as well as the Bidirectional Encoder Representations from Transformers (BERT) and the Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) by Google are among the powerful Large Language Models (LLMs) which have demonstrated near-human capabilities in text analysis and generation. BERT is a powerful model for NLP tasks, designed to pre-train deep bidirectional representations from unlabelled text by learning context from both left-to-right and right-to-left directions (Devlin et al., 2019). After pre-training, BERT can be fine-tuned to perform a diverse range of tasks, including text summarisation, with minimal task-specific architectural adaptations. GPT-4 is a large multimodal model with multiple capabilities, including the ability to understand and generate natural language text in more complex and nuanced scenarios (OpenAI, 2023). Text summarisation is one of GPT-4's numerous applications. PEGASUS is a state-of-the-art model to execute abstractive text summarisation tasks. When summarising a document, PEGASUS masks important sentences, then generates these sentences together as a unified output based on the remaining sentences which form the surrounding context (Zhang et al, 2020). PEGASUS demonstrated remarkable performance in summarising texts such as news, science, stories, instructions, emails, patents, and legislative bills.

We have tested the ability of BERT, GPT-4 and PEGASUS to summarise HS descriptions following the ATS strategies described below. The results are presented in the next chapter.

2.2.2. Processes

Regardless of the model used, our ATS strategies include three steps: data preprocessing, simplification, and evaluation.

Firstly, we perform tokenization and partitioning of the input data. This process consists of dividing text into smaller units, typically words, which are referred to as tokens. The objective is to encode HS descriptions in a manner that is comprehensible to machines without compromising their contextual integrity, thereby enabling them to discern patterns. Subsequently, we conducted the summarization per se using the pre-trained BERT, which has been adjusted to enhance the quality and coverage of the summarized outputs. This is achieved by expanding the candidate pool, enhancing the complexity detection, and relaxing the filtering thresholds. The expansion of the candidate pool enabled to consider a more diverse range of summary candidates from larger semantic sets. Additionally, we leveraged SUBTLEX_US word frequency metrics to further enhance the model's capacity to detect more complex words and phrases, thereby producing more concise and accurate summaries. The utilization of SUBTLEX_US word frequency data facilitates the identification of complex words that merit emphasis during the summarization process, which is achieved by establishing a complexity threshold based on the principle that words with lower frequency are more likely to be complex. In our ATS strategy, this frequency was set at 3.5. Finally, we lowered the filtering threshold values in order to increase the pool of candidates considered as relevant for the summary. While these candidates may not meet the criteria for simplicity or context, their inclusion has the potential to result in more detailed summaries.

3. Results

3.1. GPT-4-performed automatic text summarisation (ATS) and post-ATS complexity evaluation

This sub-chapter presents the results of the text summarisation tasks performed by GPT-4 for complex linearised descriptions. A product description is deemed to be complex if its GPT-4 complexity score, as evaluated in 2.1.2, is equal to or greater than 2.5. They concern 4738 HS6 codes.

Table 3 compares the original, linearised and simplex descriptions for a selection of HS6 codes. Qualitative observations of a representative sub-sample of 500 generated summaries indicate that GPT-4 is effective in reducing the length and the sophistication of the vocabulary used in HS descriptions. The simplex descriptions are marked less complex than the original texts yet remain coherent and retain the core information conveyed by the original descriptions.

To validate the results from the qualitative observations against the entire sample of complex descriptions, we have evaluated the complexity score of the summarised descriptions then performed a semantic similarity analysis.

We performed a GPT-4 complexity evaluation on the resulting HS description summaries, following a methodology similar to that developed in 2.1.2. Figure 3 depicts the results of the GPT-4-informed complexity evaluation for product descriptions identified as complex. The distribution of the complexity scores has flatten as a result of the summarisation process, indicating a general downward shift of the complexity level. Initially, complex descriptions, scoring at least 2.5, constitute 84.4% of the entire dataset. After the GPT-4-performed summarisation, this share has decreased to 64.1%. The count of HS6 descriptions identified as very complex, scoring 5 out of 5, has been more than halved. This category, which initially represents 36.7% of the entire dataset, is down to 14.8% after the summarisation process. Considering the average difference, GPT-4-performed summarisation allows to enhance the readability of complex HS descriptions by 1.09 points, with high variability (standard deviation of 1.07).

3.2. Automatic text summarisation with BERT and post-ATS complexity evaluation

3.3. Automatic text summarisation generated with PEGASUS and post-ATS complexity evaluation

Figure 3: Complexity scores before and after automatic text summarisation (ATS) using the three models. Note: The figures present the results for Complex descriptions only (original complexity > = 2.5), totalling 4738 HS6 products

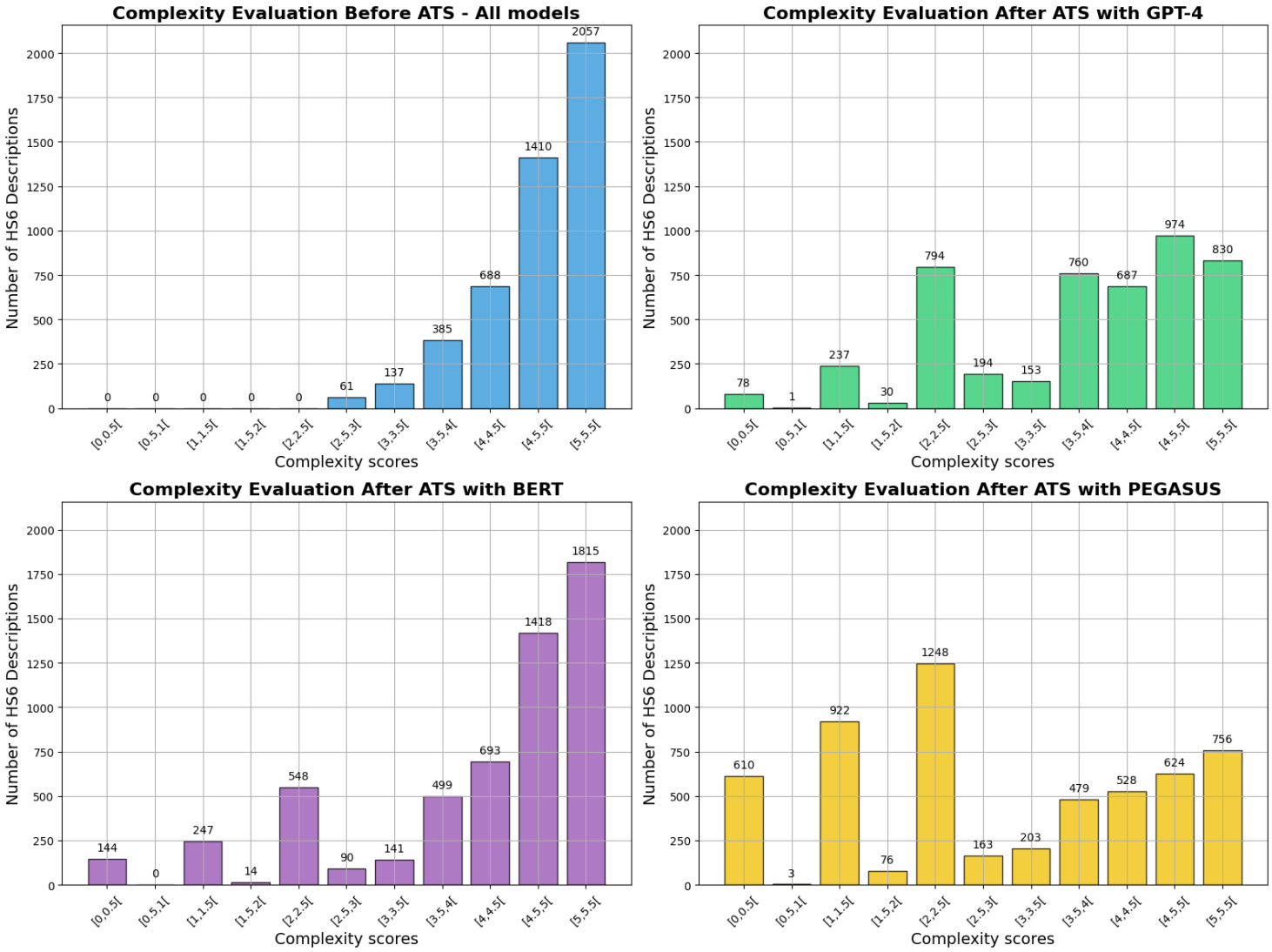


Table 3: Examples of GPT-4-generated summaries of HS descriptions, i.e. SIMPLEX Descriptions

HS6 code	Original HS Description	Linearised description	Complexity Score before Summarisation	GPT-4 Summary	Complexity Score after Summarisation
010511	Live poultry, that is to say, fowls of the species Gallus domesticus, ducks, geese, turkeys and guinea fowls : Weighing not more than 185 g : Fowls of the species Gallus domesticus	Live fowls of the species Gallus domesticus, weighing <= 185 g (excl. turkeys and guinea fowls)	5	Live domestic chickens weighing 185 grams or less, excluding turkeys and guinea fowls	2
030243	Fish, fresh or chilled, excluding fish fillets and other fish meat of heading 0304 : Herrings (Clupea harengus, Clupea pallasii), anchovies (Engraulis spp.), sardines (Sardina pilchardus, Sardinops spp.), sardinella (Sardinella spp.), brisling or sprats (Sprattus sprattus), mackerel (Scomber scombrus, Scomber australasicus, Scomber japonicus), Indian mackerels (Rastrelliger spp.), seerfishes (Scomberomorus spp.), jack and horse mackerel (Trachurus spp.), jacks, crevalles (Caranx spp.), cobia (Rachycentron canadum), silver pomfrets (Pampus spp.), Pacific saury (Cololabis saira), scads (Decapterus spp.), capelin (Mallotus villosus), swordfish (Xiphias gladius), Kawakawa (Euthynnus affinis), bonitos (Sarda spp.), marlins, sailfishes, spearfish (Istiophoridae), excluding edible fish offal of subheadings 030291 to 030299 : Sardines (Sardina pilchardus, Sardinops spp.), sardinella (Sardinella spp.), brisling or sprats (Sprattus sprattus)	Fresh or chilled sardines (Sardina pilchardus, Sardinops spp.), sardinella (Sardinella spp.), brisling or sprats (Sprattus sprattus)	5	Fresh or chilled sardines, sardinellas, brislings, or sprats	4
170211	Other sugars, including chemically pure lactose, maltose, glucose and fructose, in solid form; sugar syrups not containing added flavouring or colouring matter; artificial honey, whether or not mixed with natural honey; caramel : Lactose and lactose syrup : Containing by weight 99 % or more lactose, expressed as anhydrous lactose, calculated on the dry matter	Lactose in solid form and lactose syrup, not containing added flavouring or colouring matter, containing by weight >= 99% lactose, expressed as anhydrous lactose, calculated on the dry matter	5	99% pure lactose in solid form or syrup, without added flavors or colors	2
251720	Pebbles, gravel, broken or crushed stone, of a kind commonly used for concrete aggregates, for road metalling or for railway or other ballast, shingle and flint, whether or not heat-treated; macadam of slag, dross or similar industrial waste, whether or not incorporating the materials cited in the first part of the heading; tarred macadam; granules, chippings and powder, of stones of heading 2515 or 2516, whether or not heat-treated : Macadam of slag, dross or similar industrial waste, whether or not incorporating the materials cited in subheading 251710	Macadam of slag, dross or similar industrial waste, whether or not incorporating pebbles, gravel, shingle and flint for concrete aggregates, for road metalling or for railway or other ballast	5	Industrial waste-based aggregates for concrete, road or railway construction	3.5
870321	Motor cars and other motor vehicles principally designed for the transport of persons (other than those of heading 8702), including station wagons and racing cars : Other vehicles, with only spark-ignition internal combustion piston engine : Of a cylinder capacity not exceeding 1000 cm³	Motor cars and other motor vehicles principally designed for the transport of <10 persons, incl. station wagons and racing cars, with only spark-ignition internal combustion reciprocating piston engine of a cylinder capacity <= 1.000 cm³ (excl. vehicles for travelling on snow and other specially designed vehicles of subheading 8703.10)	5	Small-sized cars, including station wagons and racing cars, with spark-ignition engines of up to 1.000 cm³ capacity, excluding snow and other special vehicles	4
910111	Wristwatches, pocket-watches and other watches, including stopwatches, with case of precious metal or of metal clad with precious metal : Wristwatches, electrically operated, whether or not incorporating a stopwatch facility : With mechanical display only	Wrist-watches of precious metal or of metal clad with precious metal, whether or not incorporating a stop-watch facility, electrically operated, with mechanical display only (excl. with backs made of steel)	5	Electric wrist-watches made of precious metal or metal coated with precious metal, may include a stopwatch function, but exclude those with steel backs. They display time mechanically	3.5
020840	Other meat and edible meat offal, fresh, chilled or frozen : Of whales, dolphins and porpoises (mammals of the order Cetacea); of manatees and dugongs (mammals of the order Sirenia); of seals, sea lions and walruses (mammals of the suborder Pinnipedia)	Fresh, chilled or frozen meat and edible offal of whales, dolphins and porpoises (mammals of the order Cetacea), of manatees and dugongs (mammals of the order Sirenia) and of seals, sea lions and walruses (mammals of the suborder Pinnipedia)	4.5	Fresh, chilled or frozen meat and edible by-products of aquatic mammals like whales, dolphins, porpoises, manatees, dugongs, seals, sea lions, and walruses	2.5

4. Evaluation

4.1. Semantic Textual Similarity

The semantic textual similarity (STS) is the automatic evaluation metrics we employ. It aims to measure whether a pair of sentences are semantically equivalent (Agirre et al., 2015). In the context of our work, the STS calculates the likelihood in meaning between the original and the summarised descriptions for a given HS6 code. STS has a wide range of NLP applications, including machine translation, information retrieval, recommendations systems and question answering. As a metric, we employ STS to ascertain that the essence of the original descriptions is preserved after the summarisation, and that the ATS process has reduced the complexity without the loss of key information.

We select the STS as the automatic evaluation metric rather than the widely-used ROUGE developed by Lin et al. (2004) because of the limitations of the latter for this exercise, which includes ROUGE's inability to analyse syntactic and contextual information, therefore over-penalising abstractive summarisations compared to extractive ones (Kryscinski et al., 2019). Indeed, our objective is more to evaluate whether the ATS results and the source texts mean the same rather than they look the same.

We adopted the word embeddings approach using the *sentence_transformers* library of Python, then computed the cosine similarity between the two embeddings according to the following formula.

$$Similarity(A, B) = \frac{A \cdot B}{\| A \| \| B \|}$$

Where:

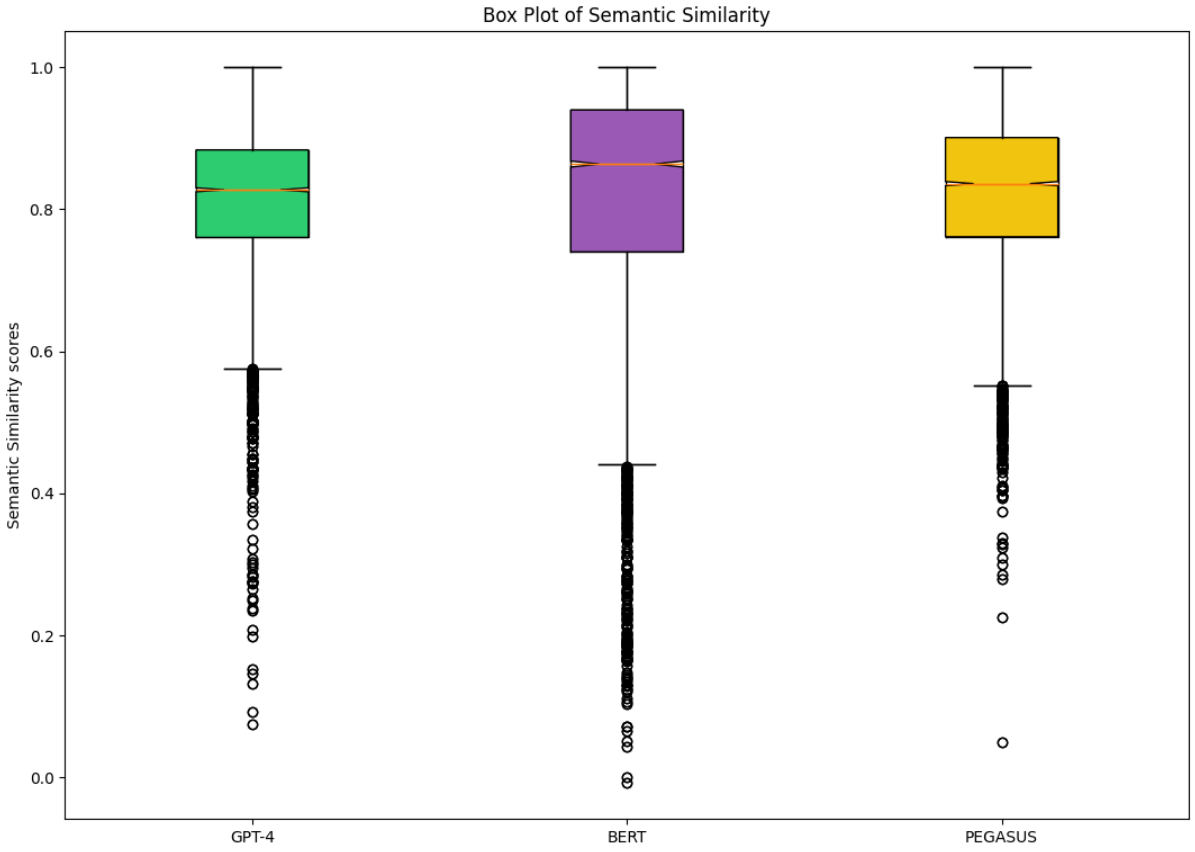
- A and B are the vectors being compared;
- $A \cdot B$ is the dot product of the vectors;
- $\| A \|$ and $\| B \|$ are the magnitudes of the vectors.

The values of the resulting cosine similarity range from -1 to 1. A cosine similarity of 1 indicates a high degree of similarity between the original and the simplex descriptions, as their representative vectors are perfectly aligned. For example, the cosine similarity scores are set to 1 for original descriptions considered as simple and which did not require further summarisation. Conversely, a cosine similarity of -1 indicates a high degree of dissimilarity between the original and the simplex descriptions, as their representative vectors are perfectly opposed. Finally, a cosine similarity of 0 indicates no similarity as the vectors are orthogonal.

The calculated cosine similarity scores between the original and simplex descriptions indicate that they have similar meanings and that the key information conveyed is preserved. We illustrate the results of the STS analysis in Figure 4. The cosine similarity scores are greater than zero for the entire dataset, indicating that the simplex descriptions are not completely dissimilar to the original descriptions. Only two HS6 codes, 1008.60 and 2522.10, are exception to this general pattern, as they scored close to zero, at 0.08 and 0.09. However, human evaluation of their original and simplex descriptions (*Triticale* versus *Hybrid wheat/rye grain* for 1008.60 and *Quicklime* versus *Calcined or heated limestone product* for 2522.10) indicates that they have similar meanings, so the low cosine similarity scores may be due to significant differences in wording. The average cosine similarity score is 0.84 (standard deviation of 0.12), which is very close to 1. Besides, 50% of the observation scored at least 0.85. Only 10% scored below 0.69, while more than 25% scored greater than 0.92, close to perfect similarity. Consequently, we can conclude that the summaries of HS descriptions generated by GPT-4 following our ATS strategy

are semantically and factually consistent with the original HS descriptions. This is an excellent result considering that 30% of the summaries generated by abstractive models contain inconsistencies (Cao et al., 2018; Goodrich et al., 2021; Falke et al., 2019). Therefore, the summaries of the HS descriptions that we have generated with this process can be used in practice to facilitate the understanding and interpretation of the HS descriptions.

Figure 4: Cosine Similarity of the original and summarised descriptions



4.2. Human evaluation

We perform a human evaluation centred around two indicators. The first one consists of the change of complexity scores following the ATS, which we developed to assess the readability enhancement after the summarisation and computed as the difference between the complexity scores before and after the summarisation process. The second indicator is an error-based metric as in Grusky et al. (2018), focusing on the informativeness to capture the semantic dimension.

We invited 17 international trade professionals to evaluate each a sample of 30 pairs of self-explanatory descriptions associated with their summaries, totalling 510 product descriptions and summaries at the HS6 level. They were given the prompts shown in Table 4 to rate the descriptions for each dimension. The distribution of the complexity scores, as perceived by human experts, before and after the simplification is illustrated in Figure 5.

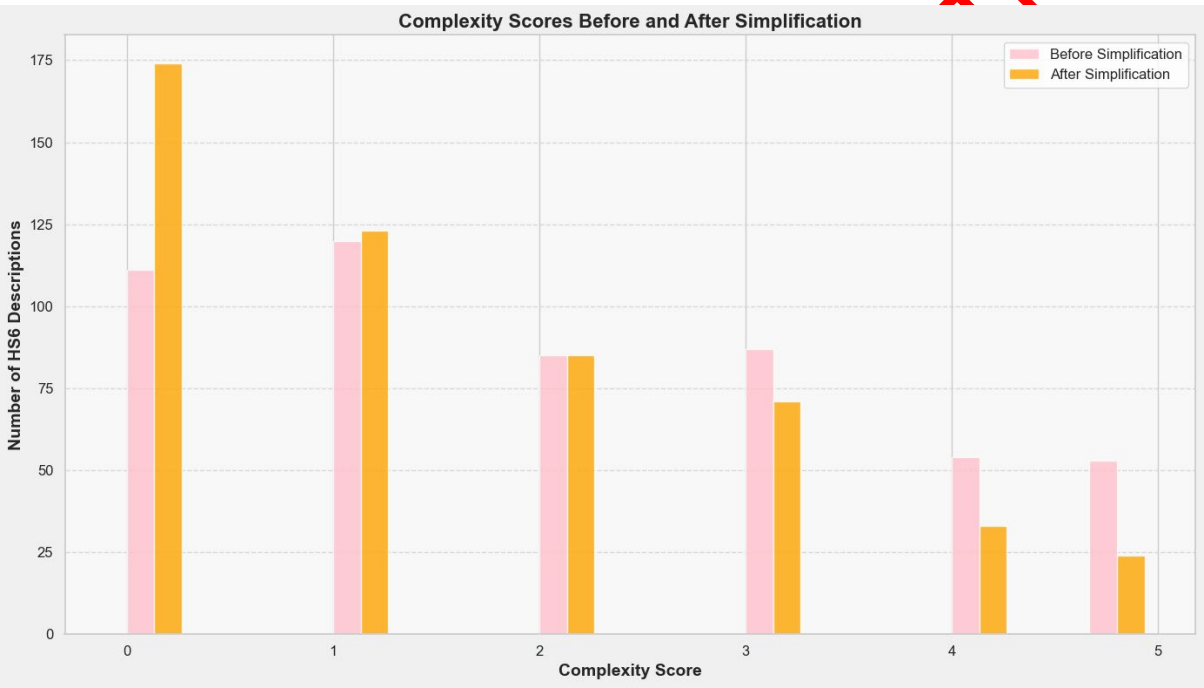
The complexity of the product descriptions, as perceived by human experts, has decreased for 89.5% of the sample and by 0.53 points in average (standard deviation of 1.14). The occurrence of very complex products, scoring 5, is more than halved. Regarding the semantic dimension, the

average informativeness score is 0.94, which means that the summaries mostly capture the essential information conveyed by the original descriptions, with a few exceptions.

Table 4: Prompts given to the Evaluators - Trade Experts

Dimension	Prompt
Readability	In a continuous scale of 0 to 5, 0 for Very Simple and 5 for Very Complex, please rate how difficult to understand are product descriptions selected in your sample? Diff_Complexity = Complexity_before – Complexity_after
Informativeness	Does the summary capture the key information of the original description? (Yes =1, No = 0)

Figure 5: Perceived Complexity Scores before and after summarisation



5. Conclusion

This study aimed to develop a methodology for generating a lexically simplified and shorter version of the product descriptions in the Harmonised System (HS) nomenclature, named simplex, using Large Language Models (LLMs), the final objective being the enhancement of their readability in order to facilitate the HS classification process.

Through automatic text summarisation (ATS) using Open AI's Generative Pre-Transformer (GPT-4), we were able to construct simplex descriptions that are significantly less complex than the original HS descriptions while maintaining the essential information conveyed. The average difference of pre- and post-ATS complexity scores is 1.09 points. The share of descriptions having a complexity score between 2.5 and 5, considered as complex in the perspective of a high school student having an intermediate level of English, has decreased from 84.4% to 64.1%, whereas the proportion of highly complex descriptions, scoring 5, has fallen from 36.7% to 14.8%. Furthermore, the average cosine similarity score is 0.84, with more than 25% of the observations scoring greater than 0.92, close to perfect similarity. Finally, human evaluation of the results centred around two indicators, the difference between the pre- and post-ATS complexity scores

as perceived by trade professionals and the informativeness index, has confirmed that our ATS strategy has enhanced the readability of HS descriptions without the loss of key information.

The simplex descriptions, which are lexically simplified and shorter version of the HS product descriptions, could find a practical application in the fields of trade compliance, particularly with regard to product classification, as they allow to expedite the reading and comprehension of product nomenclature. Nevertheless, 10% of the simplex descriptions, which have exhibited cosine similarity scores below 0.69, may necessitate additional verification by human trade experts and subsequent manual adjustments prior to their utilisation.

Finally, this study represents a pioneering attempt of LLMs-based Automatic Text Summarisation (ATS) in the field of international trade and trade compliance, extending ATS applications beyond the conventional domains. It further contributes to the development of ATS evaluation metrics by introducing novel indicators, including the GPT-4-generated complexity scores and the expert-assessed complexity scores. By investigating the potential utilisation of Artificial Intelligence in the context of international trade, where numerous avenues remain uncharted, our research provides a meaningful addition to the existing literature.

Reference

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Aguirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L. and Wiebe, J., 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, June. Association for Computational Linguistics.

Alissa, S. and Wald M., 2023. Text Simplification Using Transformer and BERT. Computers, Materials and Continua, Volume 75, Issue 2, 2023, Pages 3479-3495, ISSN 1546-2218. https://doi.org/10.32604/cmc.2023.033647.

Brandow, R., Mitze, K. and Rau, L.E., 1995. Automatic Condensation of Electronic Publications by Sentence Selection. Information Processing & Management, 31(5), pp. 675-685.

Cao, Z., Wei, F., Li, W. and Li, S., 2018. Faithful to the Original: Fact Aware Neural Abstractive Summarization. In Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). doi: 10.1609/aaai.v32i1.11912.

Crossley, S., Heintz, A., Choi, J.S., Batchelor, J., Karimi, M. and Malatinszky, A., 2023. A large-scaled Corpus for assessing text readability. Behavior Research Methods, 55, pp. 491-507. doi: 10.3758/s13428-022-01802-x.

Devlin, J., Chang, M.W., Lee, K., et al., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. doi: 10.48550/arXiv.1810.04805.

Falke, T., Ribeiro, L.F., Utama, P.A., Dagan, I. and Gurevych, I., 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 2214–2220.

Flesch, R., 1948. A new readability yardstick. Journal of Applied Psychology, 32(3), pp. 221-233.

Goodrich, B., Rao, V., Saleh, M. and Liu, P.J., 2021. Assessing the factual accuracy of generated text. The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4-8, 2019, Anchorage, AK, USA. doi: 10.48550/arXiv.1905.13322.

Grainger, A., 2024. Customs Tariff Classification and the Use of Assistive Technologies. World Customs Journal, 18(1), pp. 3-31. doi: 10.55596/001c.116525.

Grusky, M., Naaman, M. and Artzi, Y., 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 708-719, New Orleans, Louisiana. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065.

Kryściński, W., McCann, B., Xiong, C., et al., 2019. Evaluating the factual consistency of abstractive text summarization. arXiv preprint. arXiv:1910.12840.

Lin, C.Y., 2004. ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out, pp. 74-81.

Lu, L., Liu, Y., Xu, W., Li, H. and Sun, G., 2023. From task to evaluation: an automatic text summarization review. Artificial Intelligence Review, 56, pp. S2477-S2507. doi: 10.1007/s10462-023-10582-5.

Nazari, N. and Mahdavi, M.A., 2019. A survey on Automatic Text Summarization. Journal of AI and Data Mining. doi: 10.22044/jadm.2018.6139.1726.

OpenAI, 2023. GPT-4 technical report. arXiv preprint. arXiv:2303.08774.

Thomson Reuters and KPMG International, 2016. Global Trade Management Survey. Available at: https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2016/10/2016-global-trade-management-survey-from-thomson-reuters-and-kpmg-international.pdf.

World Customs Organisation (WCO), 2018. The Harmonized System: A universal language for international trade, 30 years on. Available at: https://www.wcoomd.org/-/media/wco/public/global/pdf/topics/nomenclature/activities-and-programmes/30-years-hs/hs-compendium.pdf

World Customs Organisation (WCO), 2019. Revitalizing the Harmonized System. Available at: https://www.wcoomd.org/-/media/wco/public/global/pdf/events/2019/hs-conference/hs-conference_guidance-document.pdf

World Customs Organisation (WCO), 2022. *HS Nomenclature 2022 edition*. Available at: https://www.wcoomd.org/en/topics/nomenclature/instrument-and-tools/hs-nomenclature-2022-edition/hs-nomenclature-2022-edition.aspx.

Zhang, J., Zhao, Y., Saleh, M., et al., 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. International conference on machine learning, PMLR, pp. 11328–11339. doi: 10.48550/arXiv.1912.08777.