Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI
A. Salatino et al. (Eds.)

doi:10.3233/SSW240006

Zero-Shot Topic Classification of Column Headers: Leveraging LLMs for Metadata Enrichment

Margherita MARTORANA [1], Tobias KUHN, Lise STORK, Jacco VAN OSSENBRUGGEN
Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, Amsterdam, The Netherlands
ORCiD ID: Margherita Martorana https://orcid.org/0000-0001-8004-0464, Tobias Kuhn https://orcid.org/0000-0002-1267-0234, Lise Stork https://orcid.org/0000-0002-2146-4803, Jacco van Ossenbruggen https://orcid.org/0000-0002-7748-4715

Abstract. Traditional dataset retrieval systems rely on metadata for indexing, rather than on the underlying data values. However, high-quality metadata creation and enrichment often require manual annotations, which is a labour-intensive and challenging process to automate. In this study, we propose a method to support metadata enrichment using topic annotations generated by three Large Language Models (LLMs): ChatGPT-3.5, GoogleBard, and GoogleGemini. Our analysis focuses on classifying column headers based on domain-specific topics from the Consortium of European Social Science Data Archives (CESSDA), a Linked Data controlled vocabulary. Our approach operates in a zero-shot setting, integrating the controlled topic vocabulary directly within the input prompt. This integration serves as a Large Context Windows approach, with the aim of improving the results of the topic classification task.

We evaluated the performance of the LLMs in terms of internal consistency, inter-machine alignment, and agreement with human classification. Additionally, we investigate the impact of contextual information (i.e., dataset description) on the classification outcomes. Our findings suggest that ChatGPT and GoogleGemini outperform GoogleBard in terms of internal consistency as well as LLM-human-agreement. Interestingly, we found that contextual information had no significant impact on LLM performance.

This work proposes a novel approach that leverages LLMs for topic classification of column headers using a controlled vocabulary, presenting a practical application of LLMs and Large Context Windows within the Semantic Web domain. This approach has the potential to facilitate automated metadata enrichment, thereby enhancing dataset retrieval and the Findability, Accessibility, Interoperability, and Reusability (FAIR) of research data on the Web.

Keywords. Large Language Models, Metadata Enrichment, FAIR Guiding Principles, Retrieval Augmented Generation, Linked Data

[1]Corresponding Author: Margherita Martorana, m.martorana@vu.nl

1. Introduction

Traditional dataset retrieval systems index on metadata information rather than on the underlying data values. Despite the critical role of high-quality metadata for data retrieval, many datasets still lack informative metrics and annotations to facilitate their discovery [1]. Creating and enriching metadata with high-quality information and annotations is a labour intensive and challenging process to automate, and it often relies on knowledge from domain experts. Enriching metadata with column-level information, such as with the topic described by each column, poses particular difficulty due to sparse contextual information and reliance on domain-specific codebooks, often not available in digital structured format. Moreover, column-level information becomes even more critical in the context of restricted access datasets, where users cannot directly investigate the underlying data due to confidentiality issues. In such cases, the availability of high-quality metadata with detailed column-level information becomes even more a primary need to assess the relevance and suitability of the datasets retrieved.

The FAIR Guiding Principles [2] emphasise the importance of high-quality metadata to facilitate the Findability, Accessibility, Interoperability, and Reusability (FAIR) of data on the Web. Several studies have found that by applying the FAIR Principles, we can not only improve data management and stewardship [3,4], but also facilitate data transparency, reproducibility, discovery and reuse [5] and resource citation [6]. Recently, there has also been a concrete effort to incorporate column-level information into metadata schemas, recognising its essential role in facilitating the discovery and reuse of datasets [7]. However, the sparsity and fragmentation of information that can occur in the context of restricted access data leads to challenges in applying traditional topic classification techniques.

The rise of advanced Large Language Models (LLMs) has presented several opportunities and challenges in automating data annotation and metadata creation [8]. Studies have shown some preliminary results regarding the advantages and disadvantages of various LLMs and overall performance variations [9,10]. However, it is still not clear how different LLMs perform in topic classification tasks, particularly when dealing with short texts such as column headers, and in the context of restricted access data.

1.1. Use Case

To illustrate the motivation behind this research, consider the following scenario. A socioeconomic researcher is investigating the relationship between income inequalities and proximity to higher education institutions, and may need to use multiple datasets or fragments of datasets. However, given that socioeconomic data are likely to be confidential, direct examination may not be possible. In this context, the availability of high-quality metadata with rich column-level information is crucial to discover and explore common attributes across multiple datasets. For example, column metadata could be enriched with the CESSDA topic controlled vocabulary, which include terms related to both the topics of *'income inequality'* and *'education'*. Without rich metadata, the researcher would face significant challenges in identifying relevant datasets.

1.2. Research questions and Contributions

In this work we address the challenges of automated metadata enrichment in the context of restricted access data, by investigating how we can leverage Large Language Models in a zero-shot setting and by also following a Large Context Window approach. Specifically, we explore how LLMs can perform the column header topic classification task by using a controlled vocabulary of topics. In this approach, no fine-tuning of the models is necessary, as the controlled vocabulary is provided directly as part of the input. The controlled vocabulary of topics is used for the column header classification task, leveraging a Large Context Window approach and a zero-shot setting. It is important to note that because our research focuses on the domain of restricted access data, we will only use the column headers during our analysis and not the underlying data. Additionally, we will investigate whether incorporating contextual information about the datasets - i.e. the dataset descriptions provided by the publisher - results in any differences in the classification task. To guide our investigation, we formulate the following research questions:

1. What is the consistency of the LLMs in the topic classification task of column headers from a controlled vocabulary?
2. What are the difference in the topic classification task of column headers between LLMs and humans?
3. Do hierarchical and contextual information have any effect in the classification task of column headers?

Our work contributes to the current knowledge on the applications of LLMs by assessing the performance of three LLMs (GPT-3.5, GoogleBard, and GoogleGemini) in the topic classification task of column headers with a controlled vocabulary, and comparing it with human-made classifications. To the best of our knowledge, this is the first work to investigate the performance of various LLMs in this specific task and under these settings.

2. Related Work

2.1. Semantic Metadata Enrichment

Semantic metadata enrichment refers to the process of enhancing metadata with additional meanings and contexts to improve both human- and machine-readability. This generally involves the incorporation of semantic annotations derived from ontologies [11,12], thesauri [13,14], or specialised controlled vocabularies [15,16,17], which enrich the content and connections with external resources. Unlike simpler annotations, semantic enrichment provides a deeper layer of context and structure. Previous research shows that the application of semantic metadata enrichment and the use of controlled vocabularies helps in the FAIRification of data on the web and fosters cross-disciplinary cooperation between research entities and institutions [1,18]. Furthermore, it has previously been shown how the semantic annotation of metadata, such as column-level metadata [19,7,20], can improve the Findability, Accessibility, Interoperability, and Reusability of research data [21], a crucial aspect for reusing data with restricted access (i.e., medical records and microdata [22]), which typically lacks detailed information due to its

sensitive and confidential nature. Semantic annotations at the column level can facilitate the discovery of such data by adding more context while adhering to privacy preservation standards [19].

2.2. Topic Classification with LLMs

Large Language Models (LLMs) have revolutionised the field of Natural Language Processing (NLP), with advances in a variety of applications such as content creation, text classification, and question answering (QA). LLMs are trained on a very large amount of text data (and more recently multimodal data), allowing the models not only to recognise patterns and relationships between words and concepts, but also to handle a wide range of tasks, even those for which the models have not been specifically trained and without any explicit supervision [23,24]. LLMs are, in general, more powerful than traditional NLP methods, but they are often considered 'black boxes', as the mechanisms behind an LLM decision making are challenging to understand, which makes debugging and bias detection more difficult to investigate. Furthermore, a recent study has tested the performance of ChatGPT in a variety of NLP tasks, and it has been found that the more complex and pragmatic the task (e.g. emotion recognition), the more LLM loses performance [25]. Further, another work found significant biases and inconsistent performance between ChatGPT and GoogleGemini in the detection of sentiment analysis [26]. However, there are still open possibilities and challenges related to the use of LLMs for automated annotation of data and metadata [8]. Studies have shown evidence that ChatGPT have outperformed crowd-workers in the text classification of tweets [27], and other conventional baselines [28]. Another work suggests that text classification tasks could be improved with the addition of semantic technologies and knowledge graphs [29]. Further, a recent work has shown that GPT models have outperformed the SOTAB open model in Column Property Annotation tasks [30]. Based on these findings, our research investigates the performance of ChatGPT (GPT-3.5), GoogleBard, and GoogleGemini in the topic classification task of column headers with a controlled vocabulary of topics. In addition, we compare the classification results between the LLMs and human participants, as well as the effect of topic hierarchy and contextual information.

2.3. Large Context Window and Retrieval-Augmented Generation

Large language models often contain outdated or incomplete information, as training data lack real-time updates [31] and domain-specific expertise [32,33]. Furthermore, they are prone to generating irrelevant or factually incorrect content, a phenomenon commonly referred to as *'hallucinations'* [34,35,36,37].

Retrieval-Augmented Generation (RAG) systems [38] are considered to be a promising solution to these challenges [39,40,41,42], which combine internal information from LLM with external, and preferably precise, information (e.g. textbook) to improve the accuracy and reliability of information retrieval. RAGs use retrieval systems to index, for example, a textbook stored in a vector database. They have been shown to significantly improve the performance of LLM in a variety of tasks, such as code generation [43], and question-answering (QA) in both an open domain [38,44,45,46] and a domain specific setting [47]. Although current research shows promising results, there is a lack of understanding of the underlying mechanisms of RAG systems, and recent work has shown

limitations in terms of noise and counterfactual robustness, negative rejection, and information integration [48]. Moreover, RAGs require a well-functioning retrieval systems which can be more complex to set up.

In comparison, Large Context Window refers to the ability of a language model to process large textual information as input before generating a response. This allows the model to consider the input as context, without external retrieval systems. Following a Large Context Window approach, in this work we leverage knowledge from a controlled vocabulary - the CESSDA topic classification vocabulary - to optimise the topic classification task of column headers. In the following sections, we describe the experimental settings and evaluation metrics used in our investigation.

3. Experimental Design and Evaluation

In this section, we describe the data collection process, experimental design for human and machine topic classification tasks, and evaluation methods. Our experimental design aims to assess the consistency in topic classifications of column headers of three LLMs (ChatGPT using GPT-3.5, GoogleBard and GoogleGemini), comparing them with human-made classifications and investigating the impact of contextual information (e.g., dataset description). All experiments were carried out in February 2024. Initially, the analysis was supposed to include only ChatGPT and Bard. However, Bard was subsequently updated with Gemini, allowing us to also include the latter in the study. We selected OpenAI's ChatGPT and Gemini/Bard from Google because they can be considered the current state-of-the-art (SOTA) LLM, and previous research has also used them for comparison purposes [9,10,26]. A key aspect of our methodology is to provide the same prompt for both human and LLM tasks, allowing evaluation from a neutral point of view. Furthermore, our analysis also considers differences in the classifications based on the hierarchical structure of the topics in the CESSDA controlled vocabulary, which distinguishes between 'general' and 'specific' topics. For instance, 'Education' represents a general topic, while 'Higher and Further Education' falls under a more specific subset within the 'Education' topic. We aim to investigate how LLMs interpret and classify column headers with respect to both general topics and more specific subtopics, providing insights into the model's comprehension and granularity in the topic classification task.

3.1. Data Collection

This work explores how LLMs can be leveraged for the topic classification of column headers using a Controlled Vocabulary (CV). Our analysis uses the Topic Classification CV provided by the Consortium of European Social Science Data (CESSDA)[2]. The input column headers were sourced from the CBS Open Data Portal[3]. We opt for a random dataset selection approach while ensuring diversity between various topics. A total of 10 datasets were selected and we report some of the summary statistics and information below in Table 1. The chosen datasets varied in the number of columns (ranging from 3 to 68) and in the number of rows (ranging from 340 to 347,130), and were classified under different CBS themes. While we did not include any row-level information in our

[2] https://www.cessda.eu
[3] https://opendata.cbs.nl/statline/portal

Table 1. The table contains relevant information about the input datasets for the topic classification task.

Title	CBS Identifier	CBS Theme	N. of Columns	N. of Rows
Education expenditure and indicators	80393eng	Education	68	280
Health expectancy; since 1981	71950eng	Health and Welfare	14	4536
Listed monuments; region 2023	85538eng	Leisure and Culture	4	347130
Livestock	84952eng	Agriculture	3	708
Milk supply and dairy production	7425eng	Agriculture	11	379
Mobility per person, travel modes, travel purpose	84710eng	Traffic and Transport	12	52800
Plant protection products; sales	83566eng	Nature and Environment	4	494
Population dynamics; month and year	83474eng	Population	9	380
Social security; key figures	37789eng	Labour and Social Security	19	340
Trade and industry; finance, SIC 2008	81156eng	Trade, Hotels and Restaurants	43	4480

experiment, we include these statistics here to highlight the diversity among the selected datasets.

It is important to note that our research is conducted in a zero-shot setting, meaning there is no fine-tuning and pre-training of the models for the topic classification task. Therefore, selecting a vocabulary that aligns with the domain of the datasets becomes crucial to ensure that the topic classification task is effective. In our case, the chosen vocabulary (CESSDA) is relevant to the datasets (from CBS), because both of them are from the social science domain. All input data, CV, code, and results discussed in this work are available on GitHub [4].

3.2. LLMs Topic Classification

The prompt used for the LLMs task initiation includes the task specification, input data (i.e. column headers), the CESSDA CV and some formatting constraints. The same prompt was used to query all LLMs, and the same task was executed 10 times for each LLM. For each execution, we refreshed the page and initiated a new chat session. This approach aimed to prevent the risk that prior interactions could influence the execution of the current task and affect the results.

Additionally, to assess the impact of contextual information, we repeated the process with the inclusion of `*Dataset Description:.` in the prompt inputs. Here, it is important to note that the task with context (i.e., dataset description) could not be performed with GoogleBard due to the size limitation in the allowed prompt. The analysis of the effect of contextual information is therefore performed only with ChatGPT and GoogleGemini. A summary of the prompt is provided below.

We then labelled the classifications of each column header as follows: *'Specific'* for the CESSDA sub-topics, *'General'* for general topics, *'Other'* distinctively for classification of the CESSDA topic 'other', *'Unassigned'* when the classification was not executed, and *'Hallucination'* when the topic classified was not included in CESSDA. It is important here to highlight our deliberate focus on the 'Other' topic category. This choice was made because of its potential to indicate that the LLM recognises the absence of a CESSDA topic related to the column header. This behaviour could highlight a nuances understanding of the column header by the LLM, which opts to classify it under the topic of 'Other' rather than assigning an unrelated topic. Also, the label 'Unassigned' might also indicate that the LLM is not classifying the topic rather than assigning

[4] https://github.com/ritamargherita/LLMs-topic-classification

a wrong or random one. However, the former behaviour is preferable, as it suggests an understanding of the topic related to the header is beyond the scope of CESSDA. Also, when the topic is not classified by the LLM, we cannot be sure whether this behaviour is due to the fact that the LLM does not recognise the topic related to the column header within CESSDA, or if it is just an erroneous behaviour. As an evaluation, we performed a Tukey's Honest Significant Difference (HSD) test to assess significant differences between classifications.

```
Task: Column Header Classification with Controlled Vocabulary

You are provided with two inputs, below: 1) the column headers of a dataset
(in a list format), and 2) a controlled vocabulary of topics (in a CSV
format). Your goal is to classify each column header with a relevant topic.
The controlled vocabulary has two columns: the 'Topic Label' and 'Topic
Description'. For each column header, assign a topic from the controlled
vocabulary based on semantic relevance and the definition provided for each
topic. The result should be structured in JSON format, where each column
header is paired with its assigned topic's label.

**Constraints:
Use only topics provided in the controlled vocabulary, do not add any topics
that are not included.
Do not change the text of the column headers or topic's label.
Only return the output in a JSON format, and no additional text.

**Inputs:

*Column Headers (List):
[h(1), h(2), ....., h(n)]

*Controlled Vocabulary (CSV Format):
Topic Label,Topic Description
```

In addition, we measured the **Internal Consistency** of each LLM, implementing the Needleman-Wunsch (NW) algorithm [49]. The NW algorithm is conventionally used to align genomic sequences, but was adapted here to assess the uniformity of topic classification outcomes in the 10 repeated task executions for each LLM. The pairs of classifications for each LLM and the set of column headers were aligned and scored using the algorithm. We analysed these scores using a multi-way analysis of variance (ANOVA) and post hoc Tukey's HSD tests to investigate differences in outcomes within each LLM.

Similarly, we measured the similarities in classifications across LLMs - **Inter-LLMs alignment** - employing again the NW algorithm, to execute and score pairs of classifications across pairs of LLMs. The alignment scores for each pair of LLM were summed and averaged to obtain the *Inter-LLM Alignment* score. ANOVA and Tukey's HSD tests were performed to investigate differences in alignment scores for each pair of LLMs.

3.3. Human Topic Classification

The human topic classifications were performed by three participants: M.M. and T.K., both authors of this paper, and a social scientist who specialises in CBS data. Each participant performed the classification task twice: first without contextual information and then with the dataset descriptions included. Given the minimal difference (less than 5%)

between the topic classifications with and without contextual information, we decided to keep only the classification resulting from the task with context for simplicity.

To measure the agreement between humans and LLM classifications - **Human-Computer Agreement** - we compute the *joint probability distribution*, which measures the probability of two events happening at the same time. In our case, the events are: 1) each topic classification for a given column header by one LLM, and 2) all human-made classifications for that same column header. In addition, we sum the joint probabilities to measure the probability that the classifications are in agreement between each LLM and the human participants.

Specifically, equation 1 represents the joint probability P of a human classification c_H being the topic t and a machine classification c_m also being the topic t, given that the human classification belongs to the set of all human classifications C_H and the machine classification belongs to the set of all machine classifications C_m. Subsequently, equation 2, states that the probability P of human classification c_H and machine classification c_m being the same topic t is equal to the sum of the joint probabilities of both being t. In the following, we report the equations and their variables in Table 2.

$$P(c_H = t, c_m = t | c_H \in C_H, c_m \in C_m) = P(c_H = t | c_H \in C_H) \cdot P(c_m = t | c_m \in C_m) \quad (1)$$

$$P(c_H = c_m | c_H \in C_H | c_m \in C_m) = \sum_{t \in T} P(c_H = t, c_H \in C_H) \cdot P(c_m = t | c_m \in C_m) \quad (2)$$

Table 2. Variable descriptions of the joint probability scoring to measure the agreement between LLMs and humans topic classifications.

Variable	Definition	
Topic	$t \in \{1,..,95\} = T$	A topic t belongs to the set of all topics T
Classification	$c \in C$	A classification c belongs to the set of all classifications C
Human	$h \in \{1,2,3\} = H$	A human h belongs to the set of all humans H
Humans classification	$c_H \in T$	A human classification c_H belongs to the set of all topics T
Machine	$m \in \{1,2,3\} = M$	A machine m belongs to the set of all machines M
Machine classification	$c_m \in T$	A machine classification belongs to the set of all topics T

4. Results

In this study, we investigated the performance of ChatGPT, GoogleBard, and GoogleGemini in the task of topic classification of column headers, using the CESSDA controlled vocabulary of topics. We also compared human and LLM topic classifications and explored the effect of hierarchical and contextual information on classification outcomes. In the following pages, we report our findings.

4.0.1. LLMs Topic Classification Summary Statistics

Firstly, we have analysed the summary statistics of the LLM topic classifications in both the settings without context and with the context (that is, the description of the dataset) added to the prompt. We used a Tukey HSD test to investigate any significant differences in classified topics, based on the labels introduced in 3.2. To reiterate, the labels are 1) *'Specific'* for the CESSDA sub-topics, 2) *'General'* for general topics, 3) *'Other'* for the classification of the exact CESSDA topic 'Other', 4) *'Unassigned'* when the classification was not executed, and 5) *'Hallucination'* when the topic classified was not included in CESSDA.

For the task **without context** - i.e. without adding the description of the data set in the prompt - Tukey's HSD test revealed significant differences between the *ChatGPT-GoogleBard* pair in all topic classifications except for *'Specific'* topics ($p(2) = 0.02, p(3) = 0.01, p(4) = 0.03, p(5) = 0.03$). The *ChatGPT-GoogleGemini* pair, instead, showed significant differences only in the *'Other'* topic classifications ($p(3) = 0.01$). Lastly, the *GoogleGemini-GoogleBard* pair had significant differences for *'Unassigned'* ($p(4) = 0.02$) and *'Hallucinated'* ($p(5) = 0.01$) topic classifications. For the task **with context** - with the dataset description added to the prompt -, we found a weak significant difference in the classification of topics between the ChatGPT-GoogleGemini pair for the *'General'* topics ($p(2) = 0.0469$), and a stronger difference for the *'Other'* topics ($p(3) = 0.01$).

We present two boxplots below, where Figure 1 reports the distribution of classified topics in the setting without context, and Figure 2 when the task was performed with the addition of context. We can see that in 1 GoogleBard showed fewer instances of the classification of 'Specific' and 'General' topics. It also assigns the 'Other' topic more frequently, particularly compared to ChatGPT. Moreover, instances of 'Hallucinated' topics were more prevalent compared to the two other LLMs. In 2, instead, we report the distribution of classified topics based on the labels only for ChatGPT and GoogleGemini, because we were unable to perform this task with GoogleBard, as previously mentioned. This boxplot indicates that ChatGPT and GoogleGemini had, in general, a similar distribution of classified topics.

In summary, our findings and the Tukey's results suggest that ChatGPT and GoogleGemini are more similar in performances compared to GoogleBard in the classification task. They also show that GoogleBard is more likely to assign the topic 'Other', indicating a tendency to abstain from making a definite classification. In addition, the results suggest that contextual information does not have a strong impact on the types of classified topics.

4.0.2. LLM Internal Consistency

With this measure, we assessed the internal consistency of the LLM in the classification task of each dataset in the 10 task executions. Using the NW algorithm, we measured the internal consistency, and we evaluated it using Multi-Way ANOVA and Tukey's HSD tests. For the task **without context**, the overall consistency scores for each LLM were: $ChatGPT = 0.52$, $GoogleBard = 0.11$ and $GoogleGemini = 0.81$, where 1 is absolute consistency for all several executions for each dataset. Multi-way ANOVA showed a significant effect on consistency scores based on the dataset ($p = 1.87^{-7}$). Tukey's test confirmed significant differences in consistency between: the ChatGPT-GoogleBard

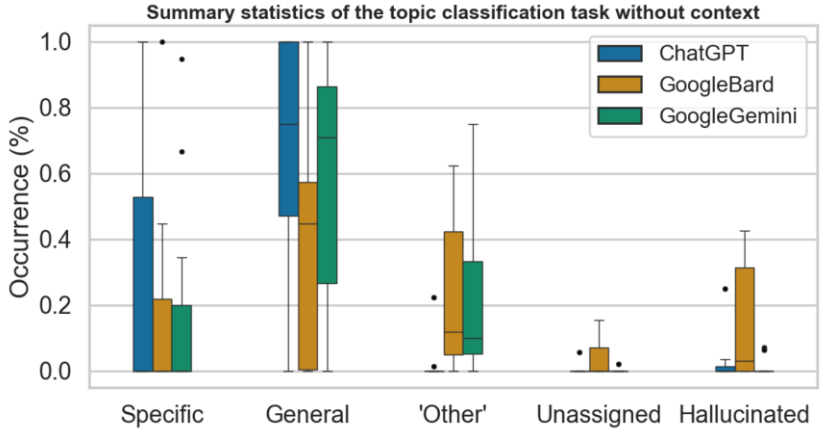

Figure 1. Summary of the topic classification task by the three LLMs, in the setting with no contextual information added to the prompt. We show the distribution of the topics classified based on 5 labels: 'Specific' topics, 'General' topics, the 'Other' topic, 'Unassigned' topics and 'Hallucinated' topics, i.e. outside of the controlled vocabulary.

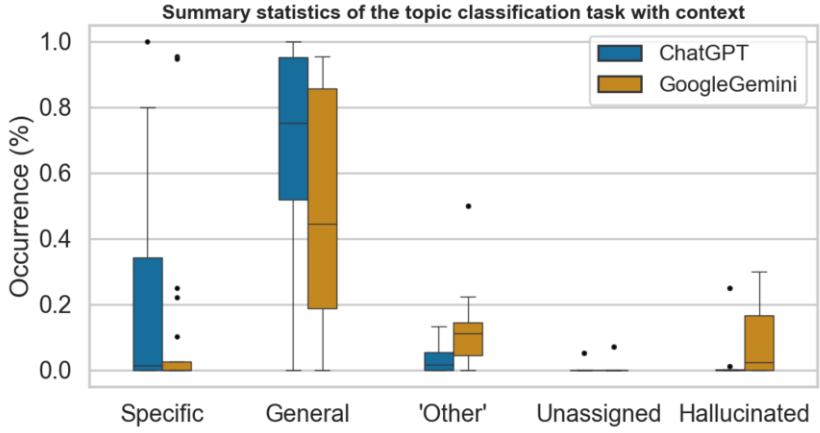

Figure 2. Summary of the topic classification task by the three LLMs, in the setting with contextual information added to the prompt. We show the distribution of the topics classified based on 5 labels: 'Specific' topics, 'General' topics, the 'Other' topic, 'Unassigned' topics and 'Hallucinated' topics, i.e. outside of the controlled vocabulary.

pair($p = 0.007$), and GoogleGemini-GoogleBard pair ($p = 0.0001$). No differences were found between the ChatGPT-GoogleGemini pair ($p = 0.3$), suggesting similar consistency scores across all task executions and datasets. For the task **with context**, we found the overall internal consistency scores of $ChatGPT = 0.46$ and $GoogleGemini = 0.51$. ANOVA and Tukey's test did not find significant differences in consistency scores between these LLMs, supporting the above findings.

These results indicate that, in general, GoogleGemini appears as the LLM that is

more consistent in the classification of topics across repeated task executions. GoogleBard, instead, shows much lower scores for the internal consistency measure. The ANOVA test also suggests that the dataset in which the column headers are classified can have an impact on the consistency score. This result needs further investigation, to evaluate whether there is a correlation between different aspects of the datasets (e.g. number of columns, domain, expressivity of column headers) and the internal consistency score. Furthermore, it appears that GoogleGemini and ChatGPT had similar internal consistency scores across all datasets, and no significant differences were found between these two LLMs even when the task was performed in context.

4.0.3. Inter-LLMs Alignment

To measure the agreement of the topic classifications between LLMs, we calculated the Inter-LLMs Alignment score using the NW algorithm and performed ANOVA and Tukey's tests. We computed the alignment scores for each LLM pair in both tasks with and without context. In all cases, the alignment scores were approximately 0, suggesting different classified topics for each LLM. For the task **without context**, ANOVA revealed that the datasets to which the column headers belonged had significant effects for the ChatGPT-GoogleBard and GoogleGemini-GoogleBard pairs ($p = 1.05^{-8}$ and $p = 2.75^{-9}$ respectively). No significant effects from the dataset were found for the GoogleGemini-ChatGPT pair. Furthermore, for the task **with context**, ANOVA found no significant effect of the dataset between the ChatGPT-GoogleGemini pair, supporting previous findings. Similarly to the results from the Internal Consistency scoring, the effect that the dataset might have on LLM performance needs further investigation. However, no significant effect was found for the ChatGPT-GoogleGemini pair for both the task with and without context, indicating that these two LLMs might have comparable underlying processes and performances.

4.0.4. Human-Computer Agreement

We calculated the Human-Computer Agreement (HCA) scores based on the joint probability metrics introduced in 3.3. The scores are reported in Table 3, where an agreement score of 1 indicates agreement between LLM classification and at least one human classification. The table reports the scores for the tasks with and without context, as well as based on the hierarchy of topics in CESSDA. In the table, the 'Exact Match' score represents the agreement between the human and machine classification when the topics are exactly the same. The 'Close Match' score, instead, involves mapping the topics to their general topic in the CESSDA controlled vocabulary. In other words, while 'Exact Match' requires exact agreement between topics (e.g. both the human and machine classifications are the CESSDA topic of *Education*), 'Close Match' allows for slight variations, as long as both machine and human classified topics belong to the same overarching CESSDA term (e.g. *Education* and *Higher and Further Education*). We find that ChatGPT classifications aligns most closely with the human-made classifications in the 'Close Match' setting for both tasks with and without context, and it also shows a slightly higher HCA for the task with context compared to the one without. Interestingly, GoogleGemini shows lower HCAs for the task with context compared to the one without context. Although we lack sufficient data for statistical significance between HCAs for tasks with and without context, these initial findings again support that contextual information may not have a significant effect on the classification task.

Table 3. The table shows for each LLM and settings (context and no-context) the agreement between machine and human classifications, where 1 is complete agreement and 0 is no agreement at all.

	No Context		With Context	
	Exact Match	Close Match	Exact Match	Close Match
ChatGPT	0.29	0.5	0.33	0.53
GoogleGemini	0.28	0.46	0.15	0.37
GoogleBard	0.24	0.31	X	X

5. Conclusion and Future Work

In this work we propose a novel approach that leverages LLMs for text classification of column headers with a topic controlled vocabulary. Our experimental design focuses on exploring the impact of contextual and hierarchical information in the topic classification task. We have evaluated the performance of three LLMs (ChatGPT, GoogleBard, and GoogleGemini) through various metrics: 1) we investigated the nature of classified topics, including the hierarchical structure of the controlled vocabulary; 2) we measured the internal consistency of each LLM in the classification task; 3) we evaluated the alignment of classified topics across LLMs; and 4) we measured the classification agreement between the LLMs and human participants. Our findings suggest that, in general, ChatGPT and GoogleGemini outperform GoogleBard in the column header topic classification task. Interestingly, contextual information appears to have no significant effect on the consistency and agreement of the classification tasks for the LLMs. We also did not find strong evidence suggesting that hierarchical information affects the classification task. Moving forward, our goal is to perform this investigation with a larger corpus of input data to better support statistical analysis and explore whether LLMs can capture semantic similarities based on the relationships between columns. We also intend to further explore the RAG approach by using available tools (e.g. the OpenAI RAG plug-in) in order to incorporate larger controlled vocabularies, as well as engage with additional domain experts to better validate and refine the results of the classification task. Although this study represents an initial exploration, it serves as a starting for the topic classification task of column headers, and it provides a groundwork approach for enhancing metadata with column-level information and advance the Findability, Accessibility, Interoperability, and Reusability of datasets on the Web.

Acknowledgement

The authors thank Emma Beauxis-Aussalet for her help with the mathematical notation of the Human-Machine Agreement scoring function. In addition, we acknowledge that ChatGPT was used to generate and debug part of the Python and LaTeXcode used in this work. This work is funded by the Netherlands Organisation of Scientific Research (NWO), ODISSEI Roadmap project: 184.035.014.

References

[1] Vlachidis A, Antoniou A, Bikakis A, Terras M. Semantic metadata enrichment and data augmentation of small museum collections following the FAIR principles. In: Information and Knowledge Organisation in Digital Humanities. Routledge; 2021. p. 106-29.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018.
[3] Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? European journal of human genetics. 2018;26(7):931-6.
[4] Mons B. Data stewardship for open science: Implementing FAIR principles. CRC Press; 2018.
[5] Wilkinson MD, Verborgh R, da Silva Santos LOB, Clark T, Swertz MA, Kelpin FD, et al. Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Computer Science. 2017;3:e110.
[6] Lamprecht AL, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. Data Science. 2020;3(1):37-59.
[7] Martorana M, Kuhn T, Siebes R, Van Ossenbruggen J. Advancing data sharing and reusability for restricted access data on the Web: introducing the DataSet-Variable Ontology. In: Proceedings of the 12th Knowledge Capture Conference 2023; 2023. p. 83-91.
[8] Tan Z, Beigi A, Wang S, Guo R, Bhattacharjee A, Jiang B, et al. Large Language Models for Data Annotation: A Survey. arXiv preprint arXiv:240213446. 2024.
[9] Singh SK, Kumar S, Mehra PS. Chat GPT & Google Bard AI: A Review. In: 2023 International Conference on IoT, Communication and Automation Technology (ICICAT). IEEE; 2023. p. 1-6.
[10] Rane N, Choudhary S, Rane J. Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation. Performance, Architecture, Capabilities, and Implementation (February 13, 2024). 2024.
[11] Lombardo V, Pizzo A. Ontologies for the metadata annotation of stories. In: 2013 Digital Heritage International Congress (DigitalHeritage). vol. 2. IEEE; 2013. p. 153-60.
[12] Bernasconi A, Canakoglu A, Colombo A, Ceri S. Ontology-Driven Metadata Enrichment for Genomic Datasets; 2018. Available from: https://swat4hcls.figshare.com/articles/journal_contribution/Ontology-Driven_Metadata_Enrichment_for_Genomic_Datasets/7343465.
[13] Koutsomitropoulos DA, Solomou GD. A learning object ontology repository to support annotation and discovery of educational resources using semantic thesauri. IFLA journal. 2018;44(1):4-22.
[14] Koutsomitropoulos DA. Semantic annotation and harvesting of federated scholarly data using ontologies. Digital Library Perspectives. 2019;35(3/4):157-71.
[15] Lisena P, Todorov K, Cecconi C, Leresche F, Canno I, Puyrenier F, et al. Controlled vocabularies for music metadata. In: ISMIR: International Society for Music Information Retrieval; 2018. .
[16] Gil Y, Garijo D, Ratnakar V, Khider D, Emile-Geay J, McKay N. A controlled crowdsourcing approach for practical ontology extensions and metadata annotations. In: The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16. Springer; 2017. p. 231-46.
[17] Afzal H, Stevens R, Nenadic G. Towards semantic annotation of bioinformatics services: building a controlled vocabulary. In: Proc. of the Third International Symposium on Semantic Mining in Biomedicine; 2008. p. 5-12.
[18] Sasse J, Darms J, Fluck J. Semantic metadata annotation services in the biomedical domain—a literature review. Applied Sciences. 2022;12(2):796.
[19] Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC medical research methodology. 2016;16:1-9.
[20] Magagna B, Rosati I, Stoica M, Schindler S, Moncoiffe G, Devaraju A, et al. The I-ADOPT Interoperability Framework for FAIRer data descriptions of biodiversity. In: JOWO 2021-Proceedings of the Joint Ontology Workshops 2021 Episode VII: The Bolzano Summer of Knowledge; 2021. .
[21] Jonquet C, Dutta B, da Silva Santos LOB, Pergl R, Le Franc Y. Common Minimum Metadata for FAIR Semantic Artefacts. In: Onto4FAIR 2023-2nd Workshop on Ontologies for FAIR and FAIR Ontologies; 2023. .
[22] Razick S, Močnik R, Thomas LF, Ryeng E, Drabløs F, Sætrom P. The eGenVar data management system—cataloguing and sharing sensitive data and metadata for the life sciences. Database. 2014;2014:bau027.
[23] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019;1(8):9.
[24] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. Advances in neural information processing systems. 2022;35:22199-213.

[25] Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. Information Fusion. 2023:101861.
[26] Buscemi A, Proverbio D. ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis. arXiv preprint arXiv:240201715. 2024.
[27] Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences. 2023;120(30):e2305016120.
[28] Chae Y, Davidson T. Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation. 2023.
[29] Shi Y, Ma H, Zhong W, Tan Q, Mai G, Li X, et al. Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. In: 2023 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE; 2023. p. 515-20.
[30] Korini K, Bizer C. Column Property Annotation using Large Language Models. In: Proceedings of the ESWC Conference; 2024. .
[31] He H, Zhang H, Roth D. Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv:230100303. 2022.
[32] Li X, Chan S, Zhu X, Pei Y, Ma Z, Liu X, et al. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. In: Wang M, Zitouni I, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Singapore: Association for Computational Linguistics; 2023. p. 408-22. Available from: https://aclanthology.org/2023.emnlp-industry.39.
[33] Shen X, Chen Z, Backes M, Zhang Y. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. arXiv preprint arXiv:230408979. 2023.
[34] Marcus G. The next decade in AI: four steps towards robust artificial intelligence. arXiv preprint arXiv:200206177. 2020.
[35] Cao M, Dong Y, Wu J, Cheung JCK. Factual Error Correction for Abstractive Summarization Models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020. p. 6251-8.
[36] Raunak V, Menezes A, Junczys-Dowmunt M. The Curious Case of Hallucinations in Neural Machine Translation. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tur D, Beltagy I, Bethard S, et al., editors. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics; 2021. p. 1172-83. Available from: https://aclanthology.org/2021.naacl-main.92.
[37] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Computing Surveys. 2023;55(12):1-38.
[38] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.
[39] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:231210997. 2023.
[40] Jiang Z, Xu FF, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, et al. Active Retrieval Augmented Generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 7969-92.
[41] Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. In: International conference on machine learning. PMLR; 2020. p. 3929-38.
[42] Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, et al. Improving language models by retrieving from trillions of tokens. In: International conference on machine learning. PMLR; 2022. p. 2206-40.
[43] Zhou S, Alon U, Xu FF, Jiang Z, Neubig G. DocPrompting: Generating Code by Retrieving the Docs. In: The Eleventh International Conference on Learning Representations;. .
[44] Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. arXiv preprint arXiv:230212813. 2023.
[45] Izacard G, Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In: Merlo P, Tiedemann J, Tsarfaty R, editors. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics; 2021. p. 874-80. Available from: https://aclanthology.org/

2021.eacl-main.74.

[46] Li D, Rawat AS, Zaheer M, Wang X, Lukasik M, Veit A, et al. Large Language Models with Controllable Working Memory. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023. p. 1774-93.

[47] Cui J, Li Z, Yan Y, Chen B, Yuan L. Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:230616092. 2023.

[48] Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38; 2024. p. 17754-62.

[49] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology. 1970;48(3):443-53.