UNIVERSITY
INSTITUTE
OF LISBON

Smart ETL and LLM-based contents classification:
the European Smart Tourism Tools Observatory experience

Diogo Francisco Machado Cosme

Master in Computer Engineering

Supervisor:
PhD Fernando Brito e Abreu, Associate Professor,
Iscte-IUL

Co-supervisor:
MSc António Miguel Portugal Galvão, Researcher,
FCT-NOVA

September, 2024

[This page has been intentionally left blank]

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Smart ETL and LLM-based contents classification:
the European Smart Tourism Tools Observatory experience

Diogo Francisco Machado Cosme

Master in Computer Engineering

Supervisor:
PhD Fernando Brito e Abreu, Associate Professor,
Iscte-IUL

Co-supervisor:
MSc António Miguel Portugal Galvão, Researcher,
FCT-NOVA

September, 2024

[This page has been intentionally left blank]

Smart ETL and LLM-based contents classification:
the European Smart Tourism Tools Observatory experience

[This page has been intentionally left blank]

[This page has been intentionally left blank]

Acknowledgements

I want to express my gratitude to Fernando Brito e Abreu and António Galvão, my supervisor and co-supervisor, respectively, for their continuous guidance, dedication, commitment, and motivation and for all the knowledge they have shared throughout this dissertation.

My research was sponsored by a scholarship offered in the scope of the RESETTING project, funded by the European COSME Programme (EISMEA), under grant agreement COS-TOURINN 101038190. This experience allowed me to embark on a European project and gain an insider's perspective on how different institutions work together. It has also allowed me to get to know various people, from colleagues and professors to business professionals.

I would also like to thank ISTAR—Information Sciences and Technologies and Architecture Research Center, the research unit that provided support and conditions for this dissertation.

Finally, I would like to express my sincere gratitude to my family and friends for their unwavering support, patience, dedication, and assistance whenever needed.

Lisbon, September, 2024

Diogo Francisco Machado Cosme

[This page has been intentionally left blank]

Resumo

Esta dissertação trata da conceção, implementação e validação de um processo Smart ETL (Extract, Transform, and Load) para integração automática de conteúdos no Observatório Europeu de Smart Tourism Tools. Um dos seus aspectos chave é a classificação automática dessas ferramentas de acordo com uma taxonomia de Smart Tourism Tools (STTs), baseada na utilização de Large Language Models (LLMs).

Foi efectuada uma revisão sistemática da literatura sobre a aplicação de LLMs na recolha de informação, com especial incidência na classificação de conteúdos. Durante a etapa de avaliação da qualidade dos estudos primários, comparámos os resultados obtidos utilizando métodos manuais e baseados em LLMs.

Após a configuração da PaaS para o Observatório, foi concebida a metodologia para o processo Smart ETL. Para a fase de extração, catálogos de STTs em formato PDF são inicialmente utilizados como fontes de dados e o seu conteúdo é extraído. Na fase de transformação, cada STT extraída dos catálogos é classificada automaticamente utilizando LLMs. Finalmente, na fase de carregamento, o conteúdo é carregado automaticamente no observatório através da sua API.

Embora os resultados da tarefa de classificação não tenham correspondido totalmente às expectativas, os primeiros resultados confirmam a viabilidade desta abordagem e marcam um passo significativo no sentido de uma classificação eficiente baseada em conteúdos, não só no domínio do turismo inteligente, mas também adaptável a outros domínios. Foram identificadas direcções de trabalho futuras para melhorar estes resultados.

Palavras-chave: Smart ETL; Large Language Model; Classificação de conteúdos; Turismo Inteligente; Smart Tourism Tools; Observatório Online

[This page has been intentionally left blank]

Abstract

This dissertation is about the conception, implementation and validation of a Smart ETL (Extract, Transform, and Load) process for automatic content integration on the European Smart Tourism Tools Observatory. One of its key aspects is the automatic classification of those tools according to a taxonomy of Smart Tourism Tools (STT), based on the usage of Large Language Models (LLMs).

A systematic literature review was conducted on the application of LLMs in information retrieval, with a particular focus on content classification. During the quality assessment step of primary studies, we compared the results obtained using manual and LLM-based methods.

After configuring the PaaS for the Observatory, the methodology for the Smart ETL process was designed. For the extraction phase, STT catalogs in PDF format are initially used as data sources and their content is extracted. In the transformation phase, each STT extracted from the catalogs is classified automatically using LLMs. Finally, in the loading phase, the content is automatically loaded into the observatory via its API (Application Programming Interface).

Although the results in the classification task did not fully meet the expectations, the first results confirm the feasibility of this approach and mark a significant step towards efficient content-based classification, not only in the field of smart tourism, but also adaptable to other fields. Future work directions were identified to improve these results.

Keywords: Smart ETL; Large Language Model; Contents Classification; Smart Tourism; Smart Tourism Tools; Online Observatory

[This page has been intentionally left blank]

Contents

List of Figures

[This page has been intentionally left blank]

List of Tables

[This page has been intentionally left blank]

Listings

[This page has been intentionally left blank]

Acronyms

AI Artificial Intelligence.

API Application Programming Interface.

BART Bidirectional and Auto-Regressive Transformers.

BERT Bidirectional Encoder Representations from Transformers.

CRM Customer Relationship Management.

CSR Corporate Social Responsibility.

CTO Chief Technology Officer.

DCMES Dublin Core Metadata Element Set.

DCMI Dublin Core Metadata Initiative.

DCTERMS DCMI Metadata Terms.

DSRM Design Science Research Methodology.

ETL Extract, Transform, and Load.

EU European Union.

GenAI Generative AI.

GPT Generative Pre-trained Transformer.

ICL In-Context Learning.

LDA Latent Dirichlet Allocation.

LLM Large Language Model.

LoRA Low-Rank Adaptation.

ML Machine Learning.

NLP Natural Language Processing.

RDF Resource Description Framework.

RESETTING **R**elaunching **E**uropean **S**mart and sustainabl**E** **T**ourism models **T**hrough digitalization and **IN**novative technolo**G**ies.

RR Rapid Review.

SEO Search Engine Optimization.

SMEs Small and medium-sized enterprises.

SMI Submodular Mutual Information.

SOTA State-of-the-art.

STT Smart Tourism Tool.

T5 Text-to-Text Transfer Transformer.

URI Uniform Resource Identifier.

WMATA Washington Metropolitan Area Transit Authority.

Chapter 1

Introduction

Contents

This chapter presents the context, justifies its importance, presents the methodology, the scientific methodology, and organization of this dissertation.

[This page has been intentionally left blank]

Chapter 1
Introduction

1.1 Context and Motivation

1.1.1 RESETTING Project

The work underlying this dissertation was developed in the scope of the RESETTING (**R**elaunching **E**uropean **S**mart and sustainabl**E** **T**ourism models **T**hrough digitalization and **IN**novative technolo**G**ies) project, funded by the European Union's COSME Programme. Its main objective was to support the shift of European tourism companies towards more resilient, circular, and sustainable operating models by testing and implementing cutting-edge digitally-driven solutions that lessen needless burdens, enhance the quality of the travel experience, aid in the decarbonization of the tourism sector, and promote more equitable economic growth that benefits both SMEs (Small and medium-sized enterprises) and destination residents. RESETTING aimed to help SMEs overcome their challenges of not having the resources and skills to keep up with technological developments and, as a result, present innovative solutions. Participating business associations, universities, and public sector organizations throughout the project strived to stimulate business innovation in european tourism SMEs.

1.1.2 STT Taxonomy

Smart Tourism (ST) is a common designation for using ICT-based innovation in the tourism sector. The corresponding ICT tools in this context are dubbed Smart Tourism Tools (STTs). A literature review on this topic [16] concluded that *"Despite all the hype around ST, there is a lack of consensus on the definition of ST"* and proposed a definition of STTs, drawing from the insights of 330 worldwide tourism experts: *"seamlessly interconnected digital tools designed to benefit all stakeholders in the tourism industry, with a special focus on the tourist and the destination, that aim at sustainable development"*. A taxonomy was also proposed there to aid in organizing these STTs, based on products, services, and applications "made in Europe". The taxonomy is divided into three application domains: *(Part of) the Touristic Offer*, *Marketing*, and *Management & Operations*, each further subdivided into STT categories, and in some cases subcategories, as represented in Figure 1.1.

1.1.3 European Smart Tourism Tools Observatory

The tourism industry's shift towards digitalization and innovation requires STTs. However, there is a digital divide between large enterprises and SMEs in terms of their stake in the tourism industry [34], [42], and their ability to capitalize on the opportunity of a digital transformation of their core business due to financial and technical limitations, as well as an unawareness of existing STT. The RESETTING project proposed to address the unawareness limitation by developing the European STT Observatory. It aims to provide a comprehensive and up-to-date overview of the STT offer in Europe. However, the objective extends beyond creating and

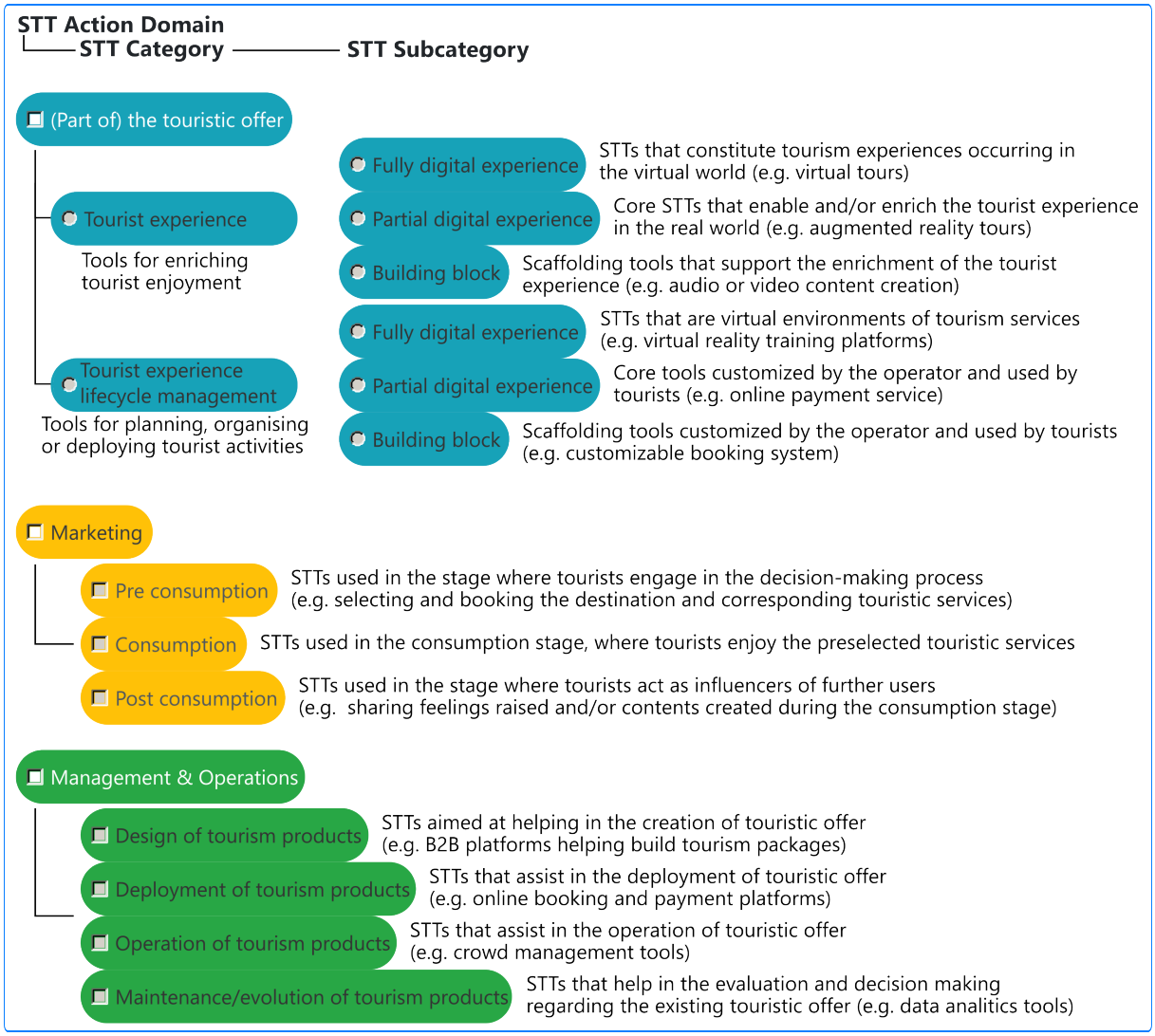

Figure 1.1: STT Taxonomy

populating the Observatory with STTs. The aim is also to automatically categorize them based on the STT taxonomy, making it easier for users to find exactly what they need.

1.1.4 Topic Overview

The main challenge in our research, scoped in the data integration subarea of the information retrieval knowledge area [32], aims to automate the process of incorporating and classifying STTs, according to their taxonomy, into the observatory.

Data integration means combining data from different sources into a unified view. It involves cleaning, transforming, and consolidating data from various databases, applications, systems, or services. Data integration aims to provide meaningful and valuable information that can be easily used for analytical, operational, or transactional purposes [14]. Besides moving data from one place to another, data integration should ensure its consistency, reliability, and quality. An interesting taxonomy for data integration features can be found online in [52].

ETL (Extract, Transform, and Load) focuses explicitly on the data integration process's extraction, transformation, and loading phases. Therefore, ETL can be considered a subset of

the data integration landscape. In data integration, ETL tools are crucial in collecting data from various sources, transforming it into a consistent and usable format, and loading it into a target database or data warehouse. These tools enable organizations to efficiently manage, consolidate, and analyze data from disparate sources, providing a unified view for reporting and decision-making. ETL can be particularly challenging when extracting multimedia contents (e.g., plain text, street addresses, URLs, logos, images, video) [31].

1.1.4.1 Smart ETL

Our research aims to extract data from unstructured PDF-format STT catalogs, organize them in a human-readable and orderly manner, classify the STTs, and upload them to the observatory. To emphasize the use of AI for automating the STT classification process, we propose the term *Smart ETL*. Our approach for the automation of STT classification differs from conventional Machine Learning (ML) classification methods on two main aspects:

1. The absence of a categorized dataset, usually with thousands of records, to be divided into training and testing sets.

2. The dataset does not consist of several features with categorical, ordinal, or numeric values. The classification of STTs relies solely on their textual descriptions. In addition, the model must be able to identify implicit concepts. For example, if a STT is described as having sensors to measure crowding, this implies that it can monitor the flow of people and that the collected data can be used to adjust a tourist service. However, this information is not explicitly stated in the description; it should be inferred.

1.1.4.2 Generative AI and Large Language Models

Generative AI (GenAI) is a currently popular subset of AI involving algorithms that generate new content based on their training data, including images, text, and audio. Large Language Models (LLMs) constitute a specific category of generative models explicitly designed to understand, generate, and manipulate human language. Among these, Transformer-based models [58] have gained prominence, which, according to NVIDIA, *"70 percent of arXiv papers on AI posted in the last two years mention transformers" (March 25, 2022)*. These models effectively capture context and dependencies using self-attention mechanisms, excelling in NLP (Natural Language Processing) tasks, text generation, and context understanding. The transformer architecture can be divided in [1]:

1. **Encoder:** takes in a sequence of words (like a sentence or paragraph) and processes it to understand the meaning and context of each word in relation to the others. It generates vectors of numerical representations (embeddings) that represent the input in a way the machine can understand. Among the foundational models [2] in the encoder-only architecture, BERT (Bidirectional Encoder Representations from Transformers) [13] stands out.

[1] Although transformer-based models are applicable beyond LLMs, the explanation of this architecture in this dissertation will focus exclusively on its use in LLMs.

[2] A foundational model refers to a large, pre-trained model that serves as a starting point or base for various specialized tasks and applications. These models are typically trained on vast amounts of data and are designed to capture general patterns and features that can be fine-tuned for specific use cases.

Learning from left and right contexts during pre-training enhances its understanding of natural language. It is well suited for text classification, question answering, and other understanding-based applications.

2. **Decoder:** takes the encoded information (the embeddings) and generates a new sequence of words. It produces the output one word at a time, considering the context provided by the encoded information. GPT (Generative Pre-trained Transformer) [38], a renowned foundational model of the decoder-only architecture, generates text by predicting the next word in a sequence, making it suitable for various generative tasks, such as text generation, language modeling, and conversational agents.

Combining both components results in encoder-decoder models that use both an encoder to process the input and a decoder to generate the output. The transformer model was the one that introduced this architecture. Besides that, there are other famous encoder-decoder models like T5 (Text-to-Text Transfer Transformer) [40], which treats every NLP problem as a text-to-text task, and BART (Bidirectional and Auto-Regressive Transformers) [26], which combines the bidirectional context of BERT with the autoregressive nature of GPT, making it versatile for both understanding and generation tasks.

Encoder-only and decoder-only models can achieve greater efficiency and effectiveness within their respective domains by specializing in specific task types. In contrast, encoder-decoder models offer versatility for tasks involving input processing and output generation.

There are two main ways of teaching or adapting a LLM to specific types of tasks:

- **Fine-Tuning:** involves taking a pre-trained model and further training it on a smaller, task-specific dataset. This process adjusts the model's weights based on the new data, allowing it to perform well on the specific task.

- **In-Context Learning (ICL):** involves giving the model examples of the task during inference[3], without additional training, allowing it to learn from these examples to complete the assignment. The model may receive zero examples (zero-shot) or a small number (few-shot) within the prompt. This approach is prominently used in decoder-only models.

Fine-tuning is ideal for achieving high performance on specific tasks when there is a sufficient amount of labeled data. However, it requires additional training, which can be resource-intensive, especially for larger models with many parameters and layers. In contrast, ICL excels in flexibility and quick adaptation to new tasks without further training. This approach is instrumental when labeled data is scarce or insufficient computational resources are available for fine-tuning.

In short, LLMs have brought significant advances to the field of NLP, from text understanding and generation to translation and summarization.

[3] Inference in this context refers to the process of generating a response or prediction based on a given input.

1.2 Expected Results

The first objective is to identify a platform to host the European STT Observatory that supports automatic upload of content (e.g., STTs) and the creation of accounts for each STT producer. This will allow producers to update their associated content. These two requirements are not the only criteria for selecting the PaaS to host the Observatory. A detailed explanation of the platform selection is given in Chapter 2. Our initial goal is to use a platform that creates a digital tradeshow with virtual booths/stands (one for each STT producer) within a 3D digital environment. This immersive experience will allow users to explore the tradeshow and visit each booth as if they were walking through a physical exhibition.

The second goal of this dissertation, which is intended to contribute to the scientific research community, is to develop a smart ETL approach to load STTs into the selected PaaS hosting the Observatory. This involves identifying and extracting STTs from different data sources (in this case, PDF catalogs), applying necessary transformations, and finally uploading them. Additionally, to make it easier for users of the Observatory to find what they are looking for, we plan to automatically classify the STTs based on the STT taxonomy. This classification will allow users to filter and search for STTs by categories, enhancing their overall search experience.

For the automatic classification, the AI model must excel in NLP since the only data available for classification are the textual descriptions of the STTs. Therefore, we believe that LLMs are the optimal solution for our classification approach, given the capabilities of these models outlined in the previous section.

1.3 Scientific Methodology

As a development methodology, this work follows the Design Science Research Methodology (DSRM) proposed in [36] and displayed in Figure 1.2. Here, the artifact will be the European STT Observatory loaded with STTs classified according to their taxonomy, which aims to facilitate the user's search process. Table 1.1 provides an overview of the steps in the methodology, with their description and the sections of this dissertation where they are covered.

Table 1.1: Descriptive steps and corresponding sections of the DSRM

Step	Description	Sections
Identify Problem & Motivate	Identify the specific research problem and explain the importance of finding a solution. This task requires an understanding of the current state of the problem and the importance of addressing it.	Section 1.1, Chapters 2 and 3
Define Objectives of a Solution	Derive the goals of the desired solution.	Section 1.2, Chapters 2, and 3
Design & Development	This involves defining the desired functionality and architecture of the artifact, followed by its creation.	Chapter 4
Demonstration	Demonstrate how the artifact addresses instances of the problem.	Section 5.1
Evaluation	Assess and evaluate the artifact's effectiveness in solving the problem.	Sections 4.4 and 5.2
Communication	Conveying the meaning of the problem, the objective, the usefulness and innovation of the artifact, the rigor of its design, and its effectiveness to researchers and other relevant audiences.	Section 5.3

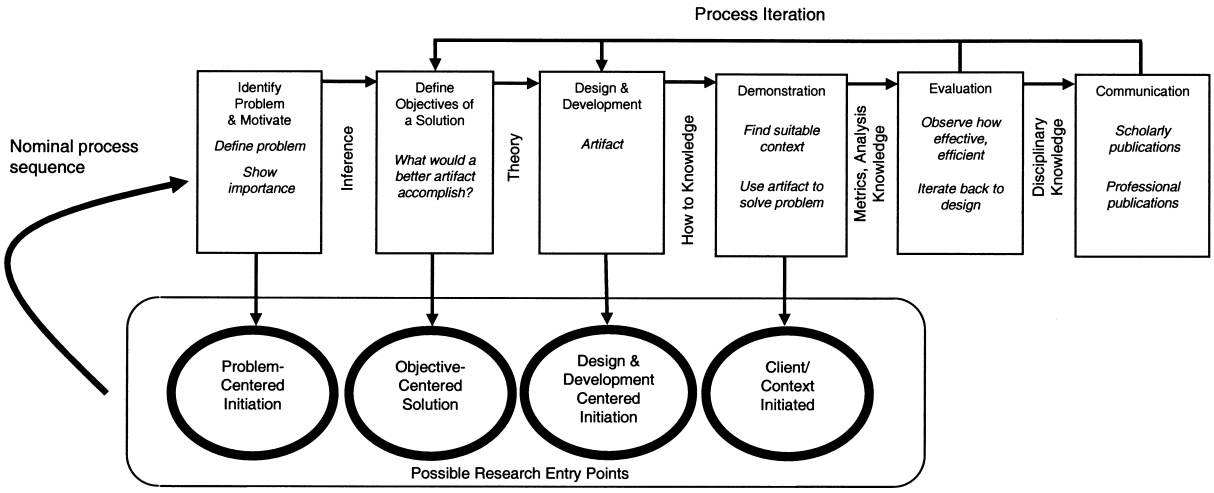

Figure 1.2: DSRM Process Model [36]

1.4 Dissertation Organization

This dissertation is structured as follows: Chapter 2 covers the necessary background context, Chapter 3 presents the literature review, Chapter 4 details the developed and implemented methodology and discusses the main results, Chapter 5 focuses on the verification and validation of the proposal, and Chapter 6 concludes with the findings and outlines future work.

Chapter 2.

Background

Contents

This chapter provides an overview of some fundamental concepts. We start with an introduction to the European STT Observatory, then discuss the catalogs used as data sources and other important details.

[This page has been intentionally left blank]

Chapter 2
Background

As mentioned previously, the initial goal was to create a 3D digital tradeshow with virtual booths/stands (one for each STT producer). This immersive experience will allow users to explore the tradeshow and visit each virtual booth/stand as if they were at a physical event. Furthermore, unlike events such as the Tourism Innovation Summit that last for a few days, this setup will allow us to offer a permanent, 24/7 event. To achieve this goal and to meet other prerequisites, the platform selected should meet the following requirements:

- Be developed by a company based in the EU since the observatory is part of an European project
- Be open to allow search engine optimization (SEO)
- Have an Application Programming Interface (API) to allow the creation of STT producers as virtual stands (booths) and add custom classification fields for each one
- Be stable to guarantee the online permanence required by an Observatory (365/24 and 24/7)
- The license cost should be fixed (i.e., not depending on the number of booths and users accessing it)
- Each STT must be presented in a booth with editing permissions only for the corresponding STT producer
- The layout of the fair (complete set of booths) should be configurable to allow the grouping of related STTs
- Visitors do not need to register

After analyzing the available online offers, we found that most products did not meet our criteria, mainly because most of them aimed at closed events and, therefore, did not provide access to search engines. Some companies offer turnkey solutions that do not allow for user customization/booth creation, and most offers are located in Asia or the US.

We met with representatives of two EU-based companies Swapcard, and Fairsnext, and despite our best efforts, we were never able to meet with meetyoo and Virtual Booth. Swapcard was discarded as it did not offer 3D solutions. Regarding Fairsnext, although it met most of our requirements and we developed an initial prototype in the platform, we ruled it out after discovering that their API does not allow the automatic creation of stands. We could only do it manually after experimenting without success with the virtual stand upload tool available on the platform.

Due to the previously identified limitations, we decided to use the Omeka Net (gold version) platform. This is a hosted version of the open-source Omeka platform, a web publishing platform for sharing digital collections and creating media-rich online exhibits. Instead of

offering a user experience as a virtual trade fair with 3D booths, it is similar to visiting a website. Omeka offers the stability of a mature project, online since 2007. Furthermore, it contains an API allowing us to insert STTs into the observatory automatically. It works through HTTP requests with the data field in JSON format. Tasks related to inserting, updating, and deleting content require an API Key that only observatory managers have. Omeka allows us to define items, collections, and exhibits. Items are the most essential element, and it is possible to upload files to them. In our case, an item will represent an STT. A collection is like a folder that groups related items together, and an exhibit is a way to combine items in a narrative text. While an item can be in multiple exhibits, it can only be in one collection.

2.1 STTs Catalogs

The current observatory content is extracted from two types of catalogs:

- **Catalogs of STTs:** This category includes four catalogs. The 2022 and 2023 versions of the catalog described in [49], produced by SEGITTUR - Sociedad Mercantil Estatal para la Gestión de la Innovación y las Tecnologías Turísticas in Spain. From now on, they will be identified as *SEGITTUR 2022* and *SEGITTUR 2023*, respectively. The other two catalogs are the first and second versions of the [7], produced by ADESTIC - Clúster de Empresas Innovadoras para el Turismo de la Comunitat Valenciana also in Spain, which will be referred to as *ADESTIC V1* and *ADESTIC V2*, respectively. These were used as data sources for the STT classification.

- **Catalogs of ST Pratices:** Rather than listing various STTs, these catalogs detail the services and initiatives implemented using STTs, categorized as ST practices. The versions used were 2022 and 2023 of the catalog described in [47], commissioned by the European Commission, which will be referred to as *EU 2022* and *EU 2023*, respectively.

The *SEGITTUR* and *ADESTIC* catalogs contain only STTs developed by Spanish companies. Additionally, the *ADESTIC* catalogs are in Spanish, so the extracted content must be translated into English. As for the European Commission's catalogs, they only list services and initiatives implemented in Europe. Next, we will examine the structure and types of content present in each catalog in detail.

The *SEGITTUR* and *ADESTIC* catalogs share a similar structure. Both include an iconographic glossary featuring various graphical icons, each associated with a specific label, representing the types of solutions designated by the respective catalog producers. For each STT, the associated icons are extracted and matched against those in the glossary to obtain the corresponding label. *SEGITTUR* also offers an additional iconographic glossary illustrating the types of destinations suitable for STTs (i.e., Culture and Urban, Nature and Sport, Niche, and Beach). Thus, a process similar to the one described previously is carried out. *ADESTIC* also identifies the types of destinations suitable for STTs, but uses a text label instead of an icon.

The relevant elements that the Spanish catalogs have in common are:

- **STT's Name**

- **STT's Description:** A text description of the STT.
- **Producer's Name:** The name of the company that produced the STT.
- **Producer's Description:** A text description of the producer.
- **Producer's Logo:** A graphical logo of the producer.
- **Producer's URL**
- **Type of Destination:** In *ADESTIC* catalogs, it appears as a text element, whereas in *SEGITTUR* catalogs, it is represented as an icon.
- **Type of Solution:** This element is an icon in both catalogs.

In addition to these elements, there are unique features. *SEGITTUR* catalogs identify the scope of the STT (e.g., governance, technology, sustainability, innovation, or accessibility) with a text label. *ADESTIC* catalogs include the producer's phone number, email address, and QR codes pointing to videos and additional files. The number indicator may also identify the producer's regional location when the phone number is a landline.

Concerning European catalogs, the relevant elements are:

- **STT Application's Title**
- **STT Application's Location:** The city where each service or initiative was implemented.
- **STT Application's Description:** A text description of the service or initiative implemented.
- **STT Application's URL**
- **STT Application's Image(s):** Associated image(s), if any.

2.2 Observatory Ontology

Each item published in the observatory must match the Dublin Core metadata structure to describe digital or physical resources. The Dublin Core, also known as the Dublin Core Metadata Element Set (DCMES), is a set of fifteen core metadata elements that have been standardized as ISO 15836-1, IETF RFC 5013, and ANSI/NISO Z39.85. Each element has a Uniform Resource Identifier (URI)[1], and they have been assigned to the *dc:* namespace, i.e., to ensure consistent identification and use across applications and systems, each URI has the prefix *dc:*. Table 2.1 provides descriptions of some of these elements, with the *dc:* namespace included in their URIs.

These core elements now belong to a wider set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI), named DCMI Metadata Terms (DCTERMS), and available on the *dcterms:* namespace. It was also standardized as ISO 15836-2. When mentioning Dublin Core, *elements* refers to the DCMES properties, while *terms* refers to the DCTERMS.

[1] A URI is a string of characters used to identify an abstract or physical resource. There are two types of URIs: Uniform Resource Locator (URL), to specify the location of a resource; and Uniform Resource Name (URN), to identify a resource by name within a particular namespace.

Table 2.1: Some descriptions of DCMES elements

contributor	
URI	http://purl.org/dc/elements/1.1/contributor
Label	Contributor
Definition	An entity responsible for contributing to the resource.
Comment	Examples of a contributor include a person, an organization, or a service. Typically, a contributor's name should be used to identify the entity.
creator	
URI	http://purl.org/dc/elements/1.1/creator
Label	Creator
Definition	An entity primarily responsible for making the resource.
Comment	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
date	
URI	http://purl.org/dc/elements/1.1/date
Label	Date
Definition	A point or period of time associated with an event in the lifecycle of the resource.
Comment	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].

Compared to the definition of *elements*, *terms* are more detailed and precise, including the type of term, formal ranges, and classes. To maintain compatibility with existing RDF (Resource Description Framework)[2] implementations of DCMES, 15 *terms* with the same names were created in DCTERMS and defined as sub-properties of the corresponding DCMES *elements*. This compatibility is essential for several reasons: it ensures interoperability, allowing different systems and applications to work together seamlessly; it ensures data consistency, allowing data described with DCMES to be understood and processed accurately; it maintains backward compatibility, preventing current implementations from becoming obsolete; and it supports a seamless transition for organizations and developers, allowing them to adopt the new terms without making major changes to their existing RDF-based systems. Table 2.2 contains the definitions of the *terms* for the corresponding *elements* shown in Table 2.1.

While implementations can use these fifteen properties in either the older format *dc:* or the newer format *dcterms:*, DCMI advises using the more recent one to adhere to Semantic Web's best practices. The observatory currently follows the DCTERMS structure.

[2]RDF is a standard model for data interchange on the Web

Table 2.2: Corresponding *terms* of the *elements* in the Table 2.1

contributor	
URI	http://purl.org/dc/terms/contributor
Label	Contributor
Definition	An entity responsible for making contributions to the resource.
Comment	The guidelines for using names of persons or organizations as creators apply to contributors.
Type of Term	Property
Range Includes	http://purl.org/dc/terms/Agent
Subproperty of	Contributor (http://purl.org/dc/elements/1.1/contributor)
creator	
URI	http://purl.org/dc/terms/creator
Label	Creator
Definition	An entity responsible for making the resource.
Comment	Recommended practice is to identify the creator with a URI. If this is not possible or feasible, a literal value that identifies the creator may be provided.
Type of Term	Property
Range Includes	http://purl.org/dc/terms/Agent
Subproperty of	Creator (http://purl.org/dc/elements/1.1/creator)
date	
URI	http://purl.org/dc/terms/date
Label	Date
Definition	A point or period of time associated with an event in the lifecycle of the resource.
Comment	Date may be used to express temporal information at any level of granularity. Recommended practice is to express the date, date/time, or period of time according to ISO 8601-1 [ISO 8601-1] or a published profile of the ISO standard, such as the W3C Note on Date and Time Formats [W3CDTF] or the Extended Date/Time Format Specification [EDTF]. If the full date is unknown, month and year (YYYY-MM) or just year (YYYY) may be used. Date ranges may be specified using ISO 8601 period of time specification in which start and end dates are separated by a '/' (slash) character. Either the start or end date may be missing.
Type of Term	Property
Has Range	http://www.w3.org/2000/01/rdf-schema#Literal
Subproperty of	Date (http://purl.org/dc/elements/1.1/date)

[This page has been intentionally left blank]

Chapter 3

State-of-the-Art

Contents

This chapter presents the literature review methodology

[This page has been intentionally left blank]

Chapter 3
State-of-the-Art

3.1 Related Literature Reviews

Due to LLMs' immense potential and inherent complexities, it is essential to review the existing literature on LLM-based content classification, especially for textual content. By understanding the current landscape and methodologies, researchers can realize LLMs' full potential and ensure their applications are innovative and effective in various fields. To check if the characterization of that landscape (aka state of the art) was already performed, we searched for literature reviews on this topic in the SCOPUS database using this search string:

"literature review"AND ("information retrieval"OR "contents classification"OR "topics classification") AND (LLM OR "large language model"OR "foundational model"OR GPT)

We obtained ten hits, but only two were literature reviews [30], [63], and none were about LLM-based content classification. A literature review, as described in [63], addressed the critical need for guidelines for incorporating LLMs and GenAI into healthcare and medical practice. In contrast, a systematic literature review, detailed in [30], identified potential research directions for information extraction from unstructured documents.

In summary, the importance of LLM-based content classification and the lack of previous literature reviews on this topic motivated us to develop the one presented hereinafter. Due to its novelty, we submitted it to a relevant conference, where it was accepted very recently [9] (supplemented by data available in [10]).

In addition to reviewing studies, we aim to group them in clusters based on their structural quality. To achieve this, we conduct a study quality assessment to ensure that the papers are well-written and include common sections such as limitations, research questions, related work, etc. The objective is to evaluate the structure of each study rather than its content and results. The primary author of this work, with the support of the co-supervisor, performed a manual assessment, and an LLM-based assessment was also conducted to explore the feasibility of automating this task.

This literature review is organized as follows: section 3.2 describes the methodology used for searching and screening the primary studies; section 3.3 analyzes the final set of selected studies; and section 3.4 provides a summary of the existing research and identifies the threats to the validity of this literature review.

3.2 Methodology

As defined in [55], a Rapid Review (RR) is *"a form of knowledge synthesis in which components of the systematic review process are simplified or omitted to produce information in a timely manner"*. In our study, we conducted a comprehensive, rapid literature review to ensure a thorough and meticulous examination of the issues at hand. This approach provided a solid foundation on the

subject matter, enabling us to gather pertinent evidence on existing technologies and identify gaps in current research.

3.2.1 Planning the Review

3.2.1.1 Research Questions

The following research questions were formulated:

- **RQ1:** What empirical studies have been conducted in LLM-based content classification?
- **RQ2:** What were the relevant contributions of the existing studies?
- **RQ3:** Can LLMs be used to assess the structural quality of studies?

3.2.1.2 Review Protocol

According to the results reported in [51], Scopus offers a more extensive subject coverage than Web of Science and Dimensions, encompassing the majority of articles found in these two other databases. As a result, we chose to use the Scopus database exclusively as a source of primary studies for our literature search.

3.2.1.3 Search String

The search string used to search the source included the most important terms related to the research questions, including synonyms, related terms, and alternative spellings, as follows:

("Large Language Model"OR "Foundational Model") AND ("Contents Classification"OR "Topic Classification")

3.2.1.4 Inclusion Criteria

A careful review of the abstracts and overall structure of the studies was conducted to determine their relevance to our research. The decision to include a study in our selection was based on the fulfillment of the following inclusion criteria:

- Be written in English
- Be a primary study
- Match at least one of the literature review objectives
- Be the most up-to-date and comprehensive version of the document
- Be available for download through the Portuguese b-on scheme[1]

[1] This is an initiative that provides access to a wide range of academic journals and publications for researchers and students in Portugal.

3.2.1.5 Data Extraction

The *Elicit* AI Research Assistant was used to extract details from papers into an organized table. According to its website, it has been used by more than 2 million researchers. Besides, it is claimed that *Elicit* uses various strategies to reduce the rate of hallucinations such as *"process supervision, prompt engineering, ensembling multiple models, double-checking our results with custom models and internal evaluations, and more to reduce the rate of hallucinations"*. In other words, its authors claim that it is a robust and trustworthy AI solution for summarizing, finding, and extracting details from scientific articles.

We have selected these excerpts with *Elicit* and made them available in [10]: research questions; summary of introduction; dataset; limitations; research gaps; software used; algorithms; methodology; main findings; objectives; study design; intervention effects; hypotheses tested; experimental techniques.

3.2.1.6 Quality Assessment

Despite the limited number of articles under review, the studies from the preceding phase were evaluated and analyzed to gauge their quality.

The quality assessment of the studies consists of 7 questions (see box with **Prompt 1** and box with **Prompt 2**), each to be answered with a score from an ordinal scale: 0—Strongly Disagree, 1—Disagree, 2—Neither Agree nor Disagree, 3—Agree, 4—Strongly Agree.

Since the main objective of our research is about using LLMs for classification, we decided to evaluate how an LLM classifies the quality of articles, taking manual (human) classification as the ground truth.

The information extracted with *Elicit* was then used as a basis for the manual and LLM-based quality assessment. For the LLM-based evaluation, we used prompting combined with the ICL Zero-shot technique, as this is the fastest and most cost-effective approach compared to fine-tuning and few-shot ICL techniques. Microsoft Copilot was chosen as the model for this study because it was one of the best available at the time for analyzing PDF documents, which was essential for the second question.

We then used **Prompt 1**, which is organized as follows: it begins with an introduction to the task, followed by the expected output that the LLM should produce: a JSON object where each key represents a question indicator, and the values are the assigned scores. Lastly, for every article, the term *"""ARTICLE"""* is substituted with the corresponding JSON object, in which each key signifies an *Elicit* field, and the values are the related information.

Prompt 1

Your task is to assess the quality of a study article based on the information provided. You'll receive two JSON objects:
1 - A JSON object with question indicators as keys and the corresponding questions as values.
2 - Another JSON object containing information about the article, where keys represent specific parameters.
Your goal is to assign to each question a score from 0 to 4 (0 - strongly disagree, 1 - disagree, 2 - neither agree nor disagree, 3 - agree, 4 - strongly agree).
Please provide your evaluation in the following JSON format: {"Q1": <score>, "Q2": <score>, . . . }.

Questions: {
"Q1": "Were the study's goals and research questions clearly defined?"
"Q3": "Was the research design clearly outlined?"
"Q4": "Were the study limitations evaluated and identified?"
"Q5": "Was the data used for validation described in sufficient detail and made available?"
"Q6": "Were answers to the research questions provided?"
"Q7": "Were negative or unexpected findings reported about the study?"
}
Article:
"""ARTICLE"""
Please provide the requested JSON.

An important note is that none of the available *Elicit* fields refer to related work, so it is impossible to answer Q2 the same way as the other questions. For the latter, the procedure was as follows: via the Copilot sidebar section in the Microsoft Edge browser, we can restrict the relevant information sources to the open page only, which in this case is a PDF opened in Microsoft Edge. We then provided **Prompt 2** as illustrated.

Prompt 2

Your task is to assign a score from 0 to 4 (0 - strongly disagree, 1 - disagree, 2 - neither agree nor disagree, 3 - agree, 4 - strongly agree) to a question from a study quality assessment about this article. Besides the score, you must provide a detailed justification and identify the sections or pages (if possible both) that contribute to your answer.
The question is: "Was previously published related work exposed and compared with the research results claimed in the study?"

3.2.2 Conducting the Review

3.2.2.1 Execute Search

Applying the specified search string resulted in the retrieval of nineteen primary studies. Given the previously defined inclusion criteria, only twelve articles were accepted.

One of the accepted studies, [44], is a review of a challenge in which several teams presented their approach to classifying the content of messages as conspiratorial or non-conspiratorial and their conspiratorial type. Therefore, articles submitted to this challenge and relevant to the research topic of this literature review were included if they did not appear in the search string results and met the inclusion criteria. This resulted in a total of thirteen accepted articles.

The initial search was conducted in June 2024. To determine whether new articles had been published prior to the submission of this dissertation, a follow-up search was conducted in September 2024, resulting in the inclusion of six additional articles that met the criteria.

3.2.2.2 Apply Quality Assessment

The articles obtained (thirteen from the first search and six from the second) were assessed for quality after each search.

- **1st Search**

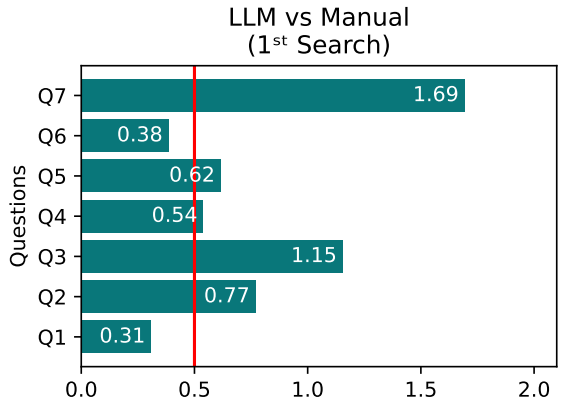

Figure 3.1: Mean Absolute Score Difference Between Methods Per Question (1st Search)

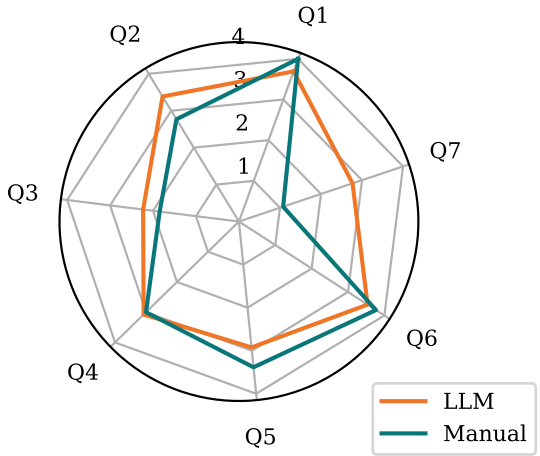

Figure 3.2: Radar Chart Displaying the Average Scores Given to the Studies (1st Search)

Figure 3.1 shows the mean absolute score difference between the two methods (LLM and manual) for each question, highlighting the response variability. A lower difference indicates that the responses, while not identical, are relatively similar. Inversely, a higher difference indicates significant variability in responses. A red line is drawn at a mean absolute difference of 0.5 to help visualize the variability. We consider an average difference of 0.5 or less across the 13 studies to be a strong indicator of agreement between the methods. For example, for questions Q1 and Q6, the number of questions without agreement was 4 for each.

Nevertheless, analyzing the mean scores assigned to each question by the method is also helpful in understanding the performance (Figure 3.2). Both graphs show that Q7 has the most significant disparity, with the highest mean absolute score difference between the two methods and the most significant gap between the mean scores. Given that Q7 relates to identifying negative or unexpected findings in the study, the higher scores assigned by the LLM-based method may indicate that LLMs have difficulty penalizing score assignments. Q4 shows a minimal difference in average scores but a mean absolute score difference of 0.54. This discrepancy occurs because one study had opposite responses (4 vs 0), significantly affecting the mean absolute score difference.

This suggests that the most effective way to evaluate performance on this test is to examine the mean absolute difference in scores. For example, if Study X scored 2 and 4 on the same question using the LLM and Manual methods, respectively, and Study Y scored 4 and 2, the difference between the mean scores would be 0: $3 - 3 = 0$. However, the mean absolute difference would be 2: $(\,|\,4 - 2\,|\,+\,|\,2 - 4\,|\,)\,/\,2 = 2$. In other words, focusing only on the difference between the average scores could misleadingly suggest that the LLMs gave the same answers as humans when, in fact, they did not.

The data obtained in the comparison between manual (***M***) and LLM (***L***) analysis is available online in [10].

- **2nd Search**

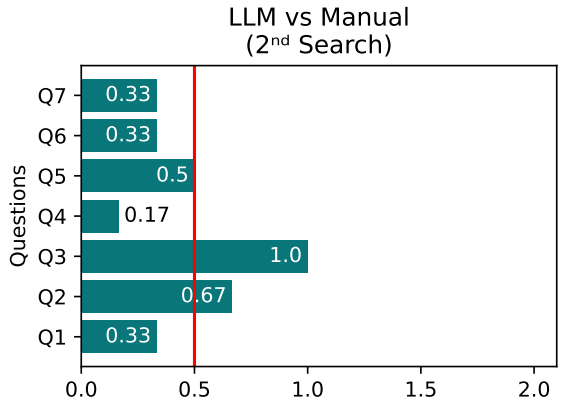

Figure 3.3: Mean Absolute Score Difference Between Methods Per Question (2^{nd} Search)

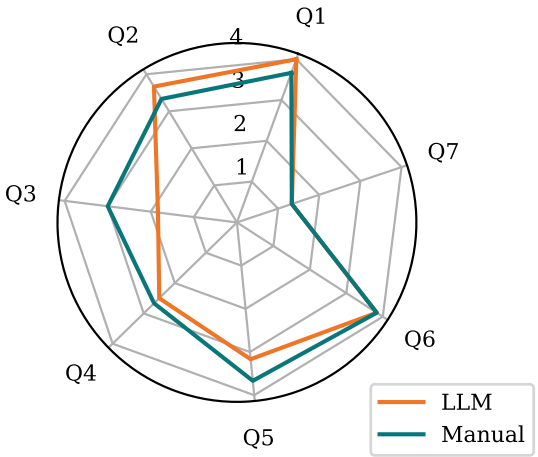

Figure 3.4: Radar Chart Displaying the Average Scores Given to the Studies (2^{nd} Search)

The articles of the second search were analyzed in the same way (Figures 3.3 and 3.4).

As the graphs of the previous and current searches show, the performance was similar except for Q4 and especially Q7. The difference in Q4 could be due to the outlier identified in the previous search, which negatively impacted the results. In contrast, Q7 now shows a significant positive performance difference. The mean absolute score difference between the LLM and manual assignments decreased from 1.69 to 0.33. This may indicate an improvement in the model's reasoning capabilities over the three months between the searches.

Given the significant improvement in Q7, we evaluated the studies from the first search to determine if the performance improvement extended to all studies.

- **Reevaluation of 1^{st} Search**

  Figure 3.5 illustrates the mean absolute score difference between the first LLM test and the manual test and between the second LLM test and the manual test. From that, it is possible to conclude that:

  1. Although not significant, a general improvement has been registered.
  2. Unexpectedly, the performance in Q7 differed from that observed in the six studies from the second search. One possible explanation is that we used an online model whose performance can be affected by the number of concurrent users. When many users are accessing the model at the same time, the model's reasoning ability may be reduced compared to times of lower usage. This is a limitation because Copilot does not provide information on the capacity of the model at the time of use, making it difficult to identify the reasons for the difference in this question.
  3. In the second LLM test, the Q1 scores are exactly the same as the manual test scores.

  A detailed comparison of the scores given to each study from the first search by each LLM test is displayed in Figure 3.7. The n value under each question label indicates the number of times the scores were different for the same study.

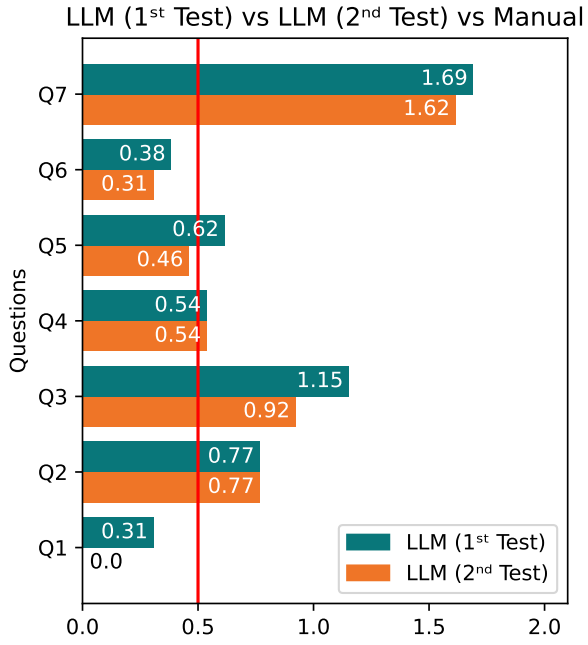

Figure 3.5: Mean absolute score difference between 1^{st} and 2^{nd} LLM tests and the manual test

To draw conclusions, we decided to use the results from the second LLM test for the first thirteen studies. Figure 3.6 shows the mean absolute score difference between the LLM assignments (the second LLM test for the thirteen studies and the only LLM test for the six studies) and the manual assignments. Figures 3.8 and 3.9 details the scores assigned to each question for all nineteen studies. Both graphs show that using Copilot to answer Q1, Q4, Q5, and Q6 is reliable, as shown by the results. Q2 could also be included, despite the 12 cases where the scores did not match, considering Copilot's slight tendency to be optimistic when answering this question. In 10 out of 12 cases, Copilot gave a score one unit higher than the manual test.

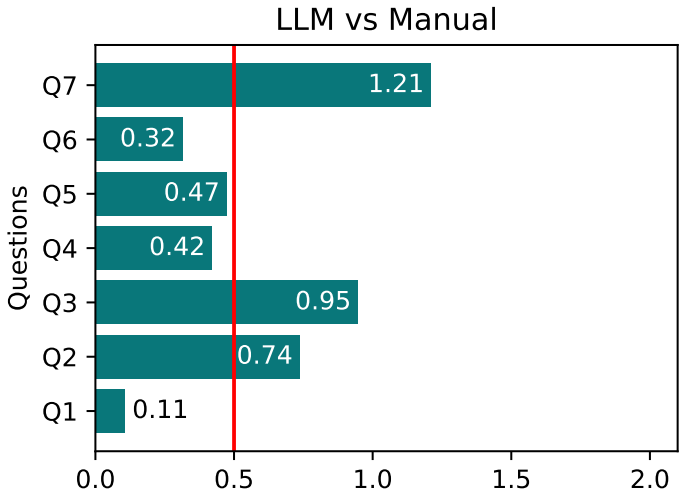

Figure 3.6: Mean absolute score difference between methods per question (2^{nd} LLM test with 2^{nd} search)

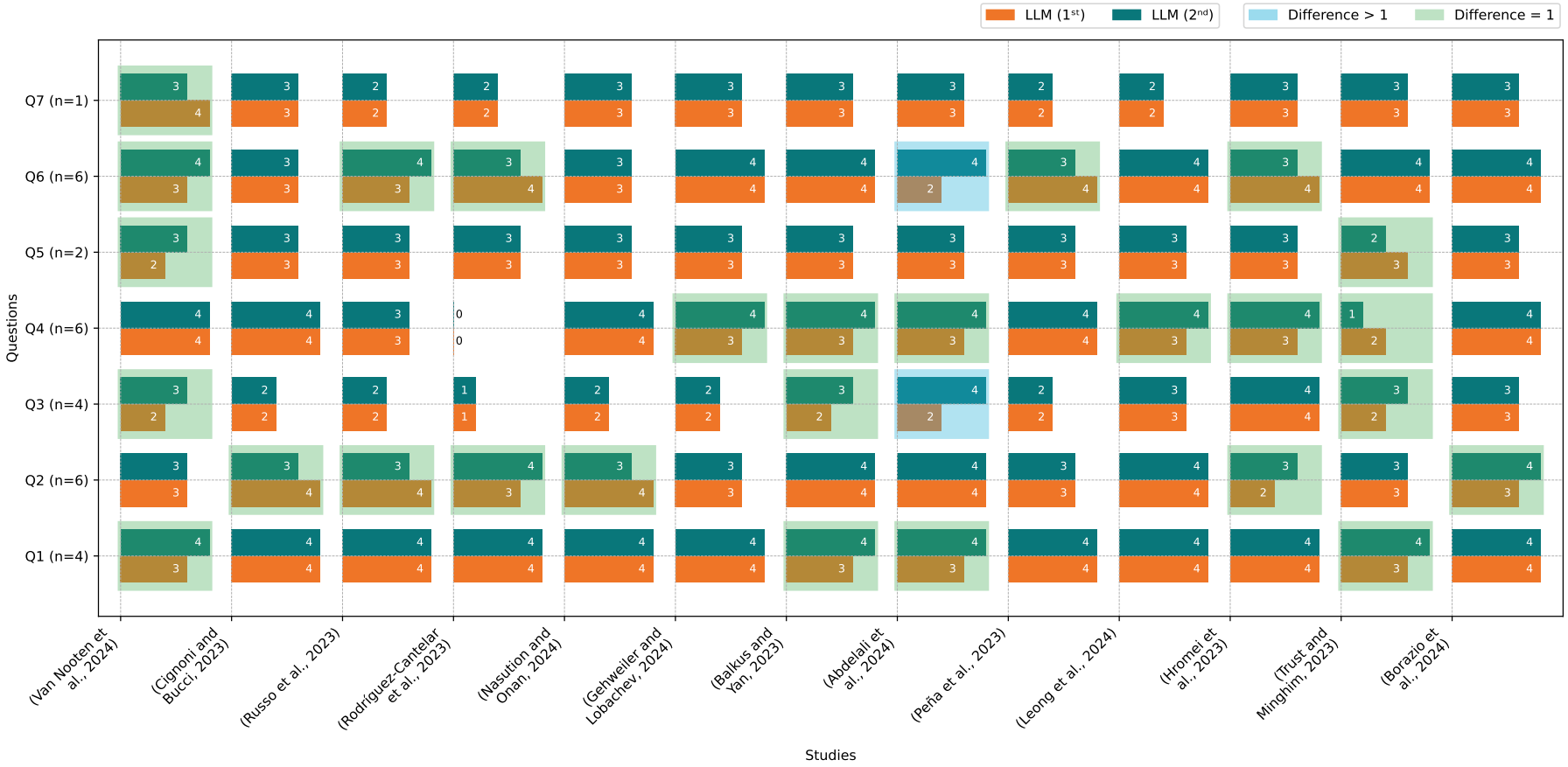

Figure 3.7: Scores Assigned to Each Study from the First Search by Each LLM Test

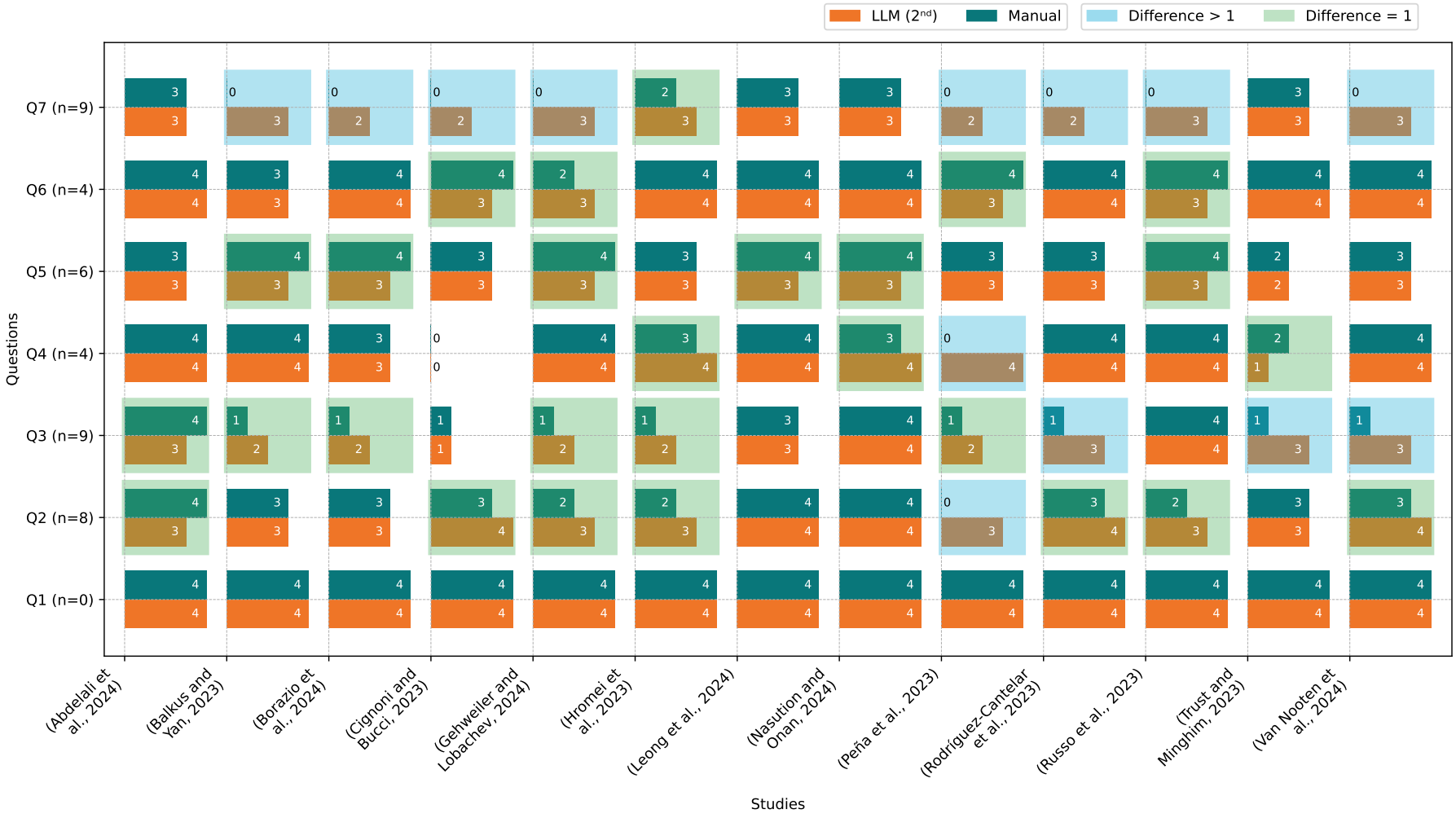

Figure 3.8: Scores Assigned to Each Study from the First Search by the 2^{nd} LLM test and the Manual test

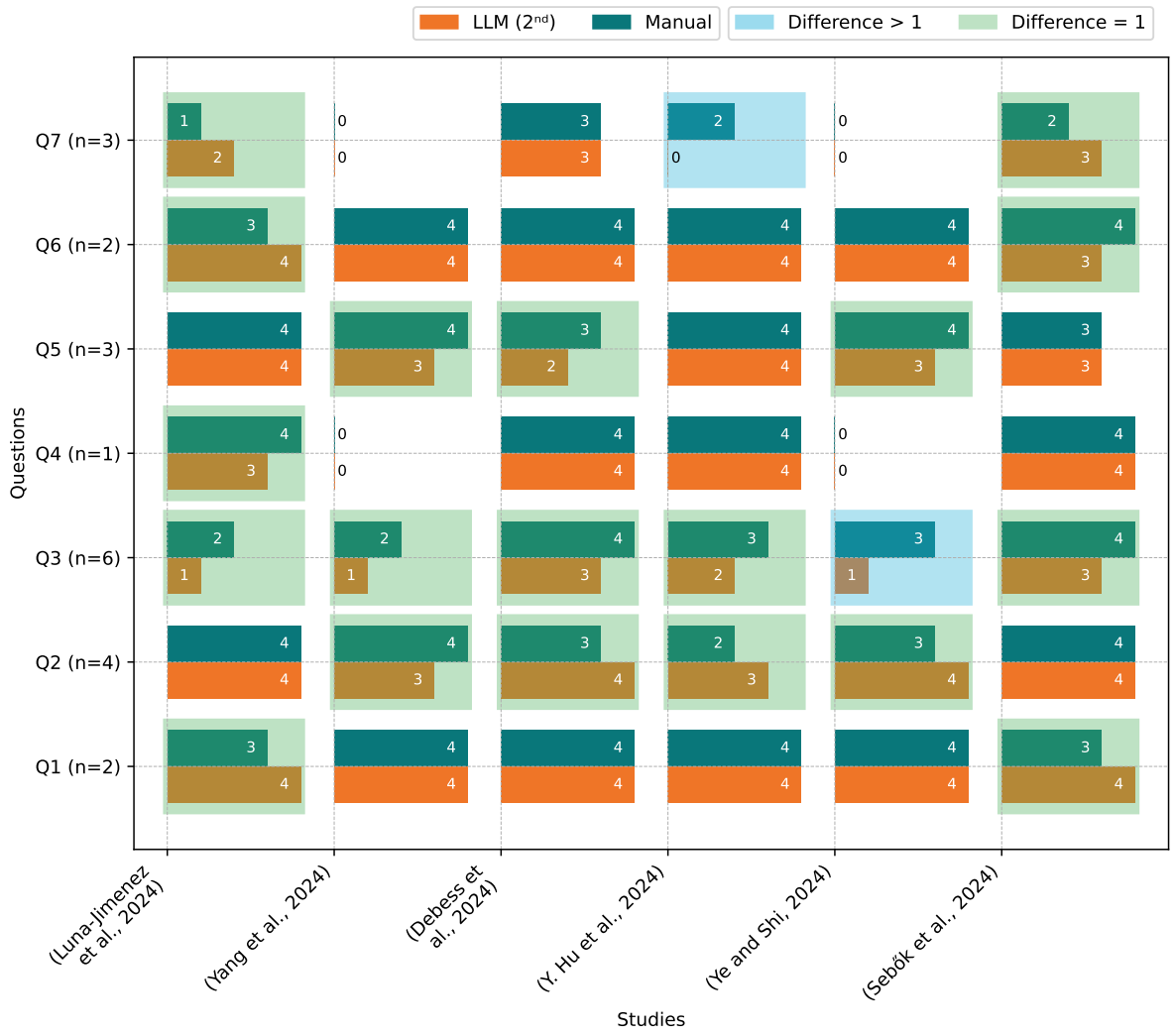

Figure 3.9: Scores Assigned to Each Study from the Second Search by the 2^{nd} LLM test and the Manual test

Although the results suggest that the use of ICL zero-shot is not yet reliable for assessing the full study quality, we conclude that it may be feasible to fully assess the quality of scientific articles with LLMs. This could be achieved through more extensive research with a fine-tuned model or by using ICL few-shot examples with a larger number of studies. Apart from the limitations of Copilot, there are also restrictions when using information extracted by the *Elicit* AI Research Assistant. As it is an AI system, there is a possibility that the extracted information could be inaccurate or, if presented without full context, could be misinterpreted. An example of this problem is the information extracted from [37] and labeled as *Limitations* by *Elicit*. In fact, the information in this field is about the authors' future work, not about limitations. This incorrect extraction may explain why Copilot gave the question about identifying and evaluating limitations (Q4) a score of 4, while the manual classification gave it a score of 0.

Given the limited number of studies, this task did not exclude any studies and was only useful for assessing their overall quality.

3.2.2.3 Overview of Quality Assessment

Figure 3.10 shows the total score of each study from the manual assignment, along with the corresponding cluster determined by the K-means algorithm. The color of each bar indicates a different cluster. To improve the readability of the graph, a red dashed line is drawn at the total score of 19.6, which is 70% of the maximum score, calculated as $(4 \times 7) \times 0.70 = 19.6$. Studies with a score above this threshold are considered to be of good quality. The two clusters containing studies with higher total scores are all above this threshold. Therefore, we categorize the clusters (from top to bottom) as follows: high quality, good quality, high moderate quality, and low moderate quality. The last cluster is still considered moderate, as the scores are around 50% of the maximum score. Some studies are highlighted in red to distinguish those in the second search (red) from those in the first search (black).

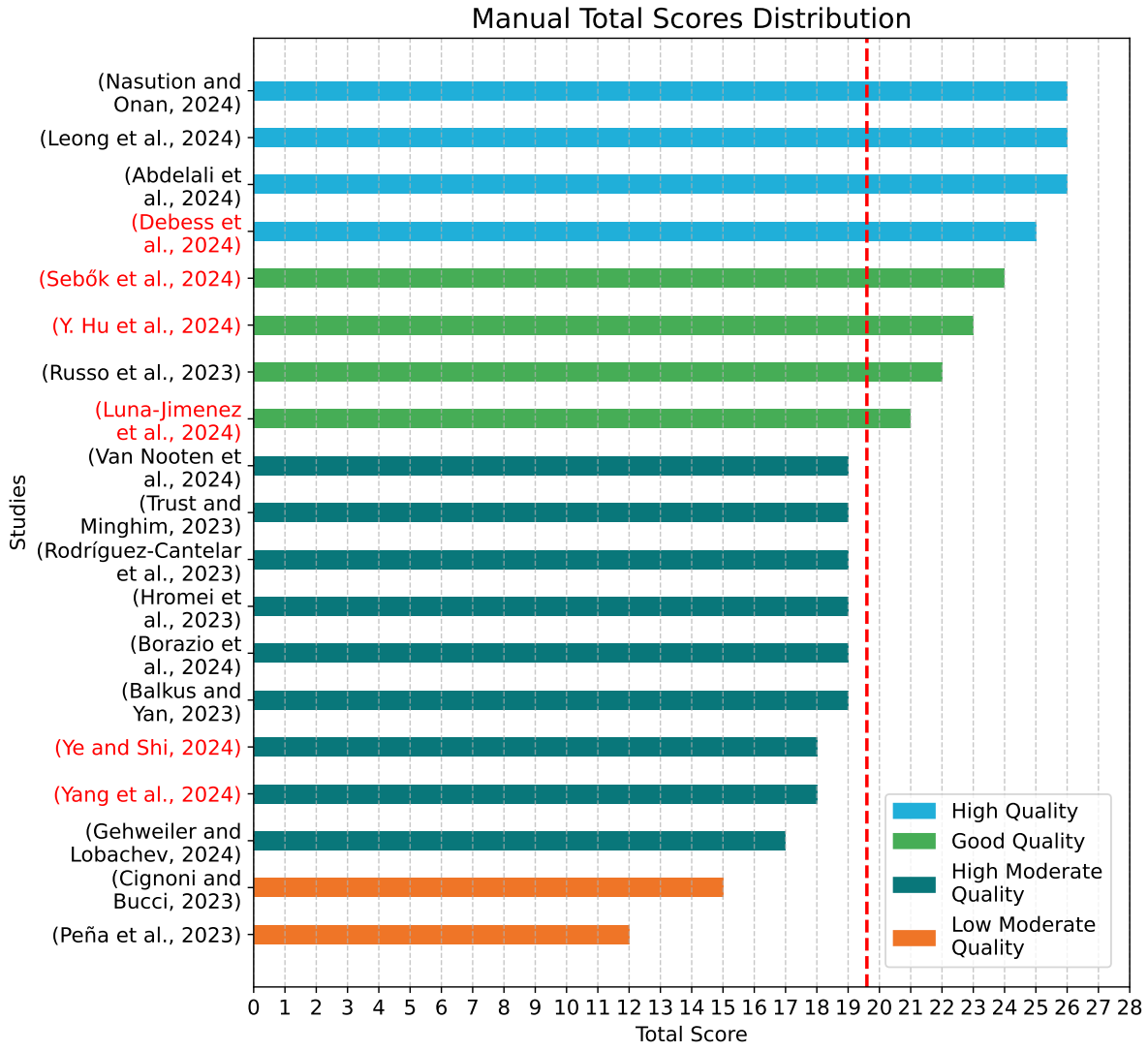

Figure 3.10: Study Quality Clustering of the Manual Assignment

3.3 Document the Review

3.3.1 Demographics

Figure 3.11 illustrates that all studies are collaborative efforts with multiple authors, with most having two authors. There are also two rare cases with many researchers (16). Regarding the authors' affiliation (Figure 3.12), the most common scenario involves one or two institutions. The relatively low number of institutions compared to the number of authors suggests a gap in inter-institutional collaboration that could improve research. This is further emphasized by the lack of international partnerships, with only two articles involving cooperation between teams from Indonesia and Turkey ([35]), and Faroe Islands and Iceland ([11]). Regarding authors' affiliation countries, while no single country dominates, Europe emerges as the most active continent (Figure 3.13). All the demographic information presented so far includes data from both the first and second searches. The most significant differences between these searches pertain to the countries of author affiliations and the venue types. In the second search, Chinese affiliations appeared with three studies, while Europe continued to contribute with four new studies. Except for one (Spain), all these European studies came from new countries. Furthermore, in the second survey, the published studies were either related to conferences or journals.

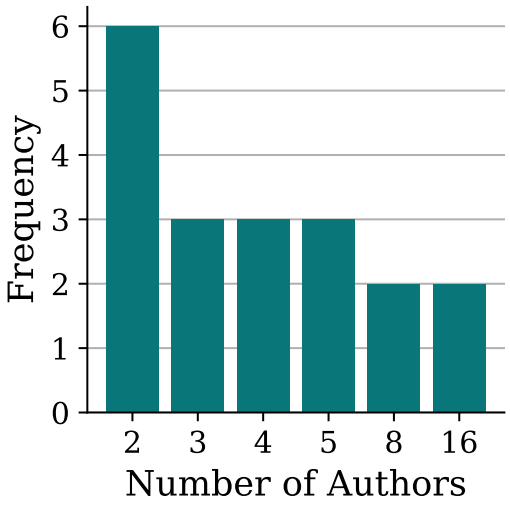

Figure 3.11: Publication Frequency by Authors Count

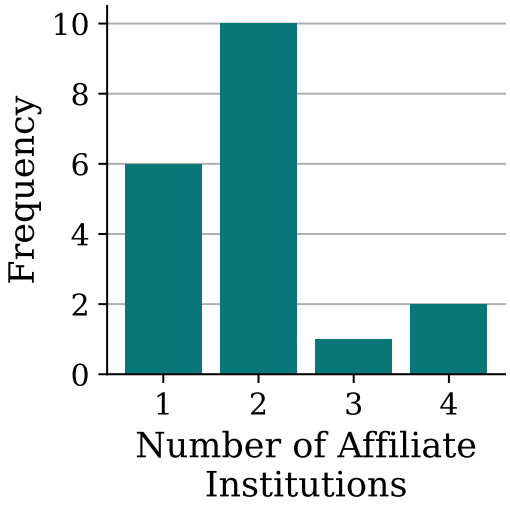

Figure 3.12: Publication Frequency Affiliates Count

Figure 3.14 clearly shows that most selected studies were published in workshops and journals. It should be remarked that three articles come from the same workshop (EVALITA 2023). This "high concentration" in a single workshop may indicate the topic is still niche, with limited venues for broader exposure. It can also be considered a sign that a community is emerging, with the possibility of broader interest in the future.

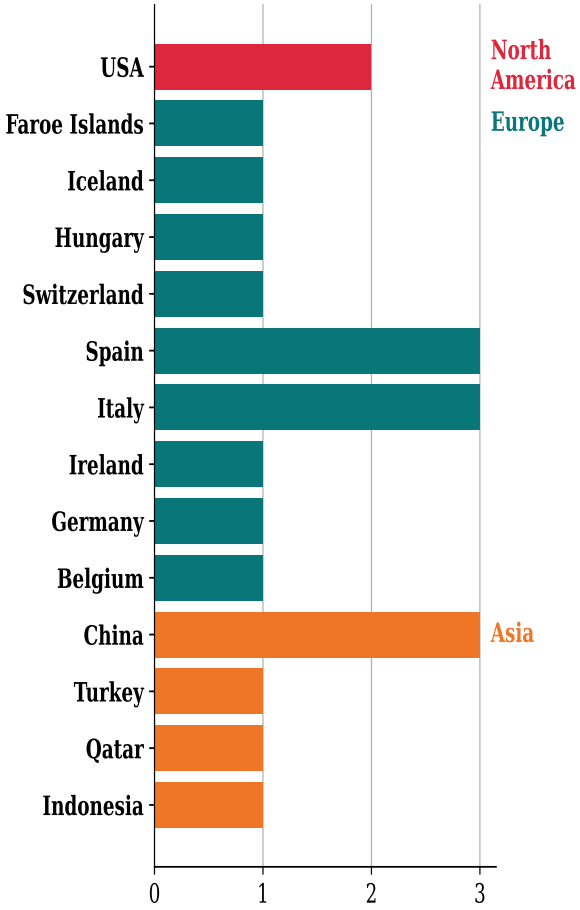

Figure 3.13: Publication Frequency by Author Affiliations' Country

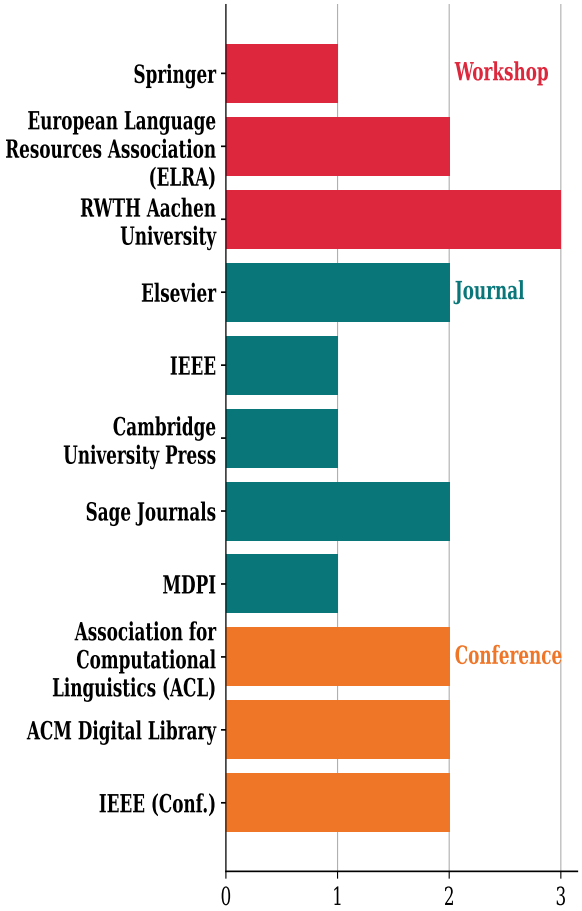

Figure 3.14: Publication Frequency by Publisher

3.3.2 Analysis and Findings

3.3.2.1 1^{st} Search (June, 2024)

A methodology, described in [43], was proposed to address the problem of inconsistent responses in chatbots. It consists of hierarchical topic/subtopic detection using zero-shot learning (through GPT-4), and detecting inconsistent answers using clustering techniques. The datasets used in the study were the DailyDialog corpus ([27]) and data collected by the authors' Thaurus bot during the Alexa Prize Socialbot Challenge (SGC5). Using the *DailyDialog* dataset, the authors achieved a weighted F1 score of 0.34 for topic detection and 0.78 for subtopic detection. The SGC5 dataset obtained an accuracy of 81% and 62% for topic and subtopic detection, respectively. Notably, there is room for improvement in the *DailyDialog* topic detection, as the authors recorded a lower weighted F1 score, indicating a significant number of false positives or false negatives.

An overview of the EVALITA 2023 challenge "Automatic Conspiracy Theory Identification (ACTI)" is presented in [44]. The challenge focuses on identifying whether an Italian message contains conspiratorial content (Subtask A) and, if so, classifying it into one of four possible conspiracy topics: "*Covid*", "*Qanon*", "*Flat Earth*", or "*Pro-Russia*"(Subtask B). A total of eight teams participated in Subtask A and seven teams in Subtask B. The provided dataset was

the same for each team and each task. It used a collection of Italian comments scraped from 5 Telegram channels known for hosting conspiratorial content, collected between January 1, 2020, and June 30, 2020. The comments were manually annotated by two human annotators to identify conspiratorial content (as "*Not Relevant*", "*Non-Conspiratorial*"or "*Conspiratorial*") and categorize it into specific conspiracy theories. The authors calculated inter-annotator agreement rates using Cohen's Kappa coefficient to evaluate the consistency among annotators. They achieved high agreement levels: a Cohen's Kappa of 0.93 for Subtask A and 0.86 for Subtask B. For data integrity reasons, comments that didn't receive the same classification were excluded, and "Not Relevant"comments were also discarded to focus solely on relevant conspiratorial content. The final datasets consist of 2,301 comments labeled with a binary label for Subtask A and 1,110 comments labeled with a value from 0 to 3, representing the specific conspiracy topic. The articles in this challenge that are relevant to the subject of this paper are:

- The authors of [6] compared the performance between two fine-tuned encoder-only transformer models (bert-base-italian-xxl-cased and XLM-RoBERTa, [8]) and a non fine-tuned decoder-only transformer model (LLaMA 7B, [53]). The BERT models achieved a higher test score than the LLaMa model in both subtasks. For Subtask A: 0.83, 0.82 and 0.80, respectively. For Subtask B: 0.83, 0.85 and 0.74, respectively. The article does not provide details regarding the study's limitations and how LLaMa was used.

- The authors of [20], took a distinct approach. Initially, they introduced a model to address all tasks in the EVALITA 2023 challenge, not just the ACTI task. Consequently, their dataset was significantly larger than the one provided for the ACTI task, comprising 134,018 examples from various tasks. For each task, the authors compared the performance of two models. One is an encoder-decoder model named *extremIT5*, based on IT5, consisting of approximately 110 million parameters. It was fine-tuned by concatenating task names and input texts to generate text solving the target tasks. The other model is a decoder-only model named *extremITLLaMA*, based on LLaMa 7B. It was first trained on Italian translations of Alpaca instruction data using LoRA (Low-Rank Adaptation)[2]([21]), to enable the model to comprehend instructions in Italian. Then, it is further fine-tuned using LoRA on instructions reflecting the EVALITA tasks. In their final results, the authors achieved an F1 score of 0.82 for Subtask A using *extremIT5* and 0.86 with *extremITLLaMA*. For Subtask B, the F1 scores were 0.81 and 0.86, respectively. The biggest limitations of this study are the computational cost and inference speed of the larger *extremITLLaMA* model and the limited exploration of architectures and hyperparameters due to time constraints. In conclusion, the authors suggest that exploring zero-shot or few-shot learning could benefit sustainability, as it reduces the need for large amounts of annotated data.

The approach described in [6] achieved the sixth rank in Subtask A, while the one detailed in [20] secured the second position. In Subtask B, their respective rankings were fourth and fifth. The winning team in both subtasks utilized a strategy that leveraged data augmentation through LLMs.

[2]LoRA fine-tuning significantly reduces the computational and storage costs of training large language models by only adjusting a subset of low-rank parameters.

As proposed in [56], query-focused submodular mutual information functions can be used to select diverse and representative demonstration examples for ICL in prompting. In addition, an interactive tool is presented to explore the impact of hyperparameters on model performance in ICL. For evaluation purposes, the authors have applied their method to the following tasks: two sentiment classification tasks with Stanford Sentiment Treebank datasets (SST-2 and SST-5) ([50]), and a topic classification task with the AG News Classification Dataset ([66]). Their methodology consists of the following two steps.

i. **Retrieval:** The goal here is to, based on the input test, select representative and diverse in-context demonstration examples from the training data. The input test and the training dataset undergo embedding via the sentence transformer ([41]) to achieve this. Subsequently, specialized selection occurs by leveraging Submodular Mutual Information (SMI) functions to choose examples from the training data. The selected examples are then incorporated into a prompt template alongside an optional task directive or as stand-alone demonstrations.

ii. **Inference:** The prompt template and input test are fed into a pre-trained language model to deduce the corresponding label. They used three open-source pre-trained models: GPT-2 ([39]), OPT ([65]), and BLOOM ([23]).

According to the authors, their approach can yield performance enhancements of up to 20% when compared to random selection or conventional prompting methods, and the size and type of the language model do not always guarantee better performance.

A transit-topic-aware language model that can classify open-ended text feedback into relevant transit-specific topics based on traditional transit Customer Relationship Management (CRM) feedback is proposed in [24]. The primary dataset includes around 180,000 anonymous customer feedback comments, manually labeled, from the Washington Metropolitan Area Transit Authority (WMATA) CRM database, covering January 2017 to December 2022. Given 61 distinct labels, the authors used Latent Dirichlet Allocation (LDA) to group customer feedback into broader topics. Due to the limitation of LDA in detecting significantly less represented topics, these topics were excluded from the CRM dataset before applying LDA and grouped according to their original topic (2 niche groups). LDA failed to identify a primary topic for approximately 62,000 complaints. As a result, the final dataset included around 120,000 complaints categorized into 11 topics (9 LDA-detected topics and two niche topics). They evaluated the performance of five ML models (Random Forest, Linear SGD, SVM, Naive Bayes, and Logistic Regression) against the proposed MetRoBERTA LLM. MetRoBERTA is a fine-tuned version, with the CRM dataset, of the RoBERTa LLM open-sourced by Meta Research ([28]). MetRoBERTA outperformed the traditional ML models with a macro average F1 score of 0.80 and a weighted average F1 score of 0.90, compared to the best ML model with 0.76 and 0.88, respectively. A significant limitation of this study is the exclusion of approximately 60,000 initial complaints, accounting for over one-third of the entire dataset.

A novel framework that uses LLMs to identify and categorize emergent socio-political phenomena during health crises, with a focus on the COVID-19 pandemic, is introduced in [5]. This framework also provides explicit support to analysts by generating actionable statements

for each topic. For this aim, they used a dataset of 2,254 news articles manually categorized by ISS (Istituto Superiore di Sanità) experts into five topics: "*Covid Variants*,""*Nursing Homes Outbreaks*,""*Hospital Outbreaks*,""*School Outbreaks*,"and "*Family/Friend Outbreaks*,"collected from February 2020 to September 2022. Then, their system generates linguistic triples to capture fine-grained concepts, which analysts can refine to correlate themes. For the following step, they have employed a model based on BART ([25]) and previously trained on the Multi-Genre Natural Language Inference corpus ([59]). The model uses zero-shot classification to associate news articles with the identified topics without fine-tuning. Preliminary results demonstrate accurate mapping of news articles to specific, detailed topics. The system achieved an accuracy of 67% when proposing a single class, which increased to 88% when considering the top two system suggestions. However, the authors acknowledge potential limitations, including hallucinations from integrating a decoder LLM (GPT-4) for prompting generation.

The benchmarking study LAraBench ([1]) addresses the gap in comparing LLMs against state-of-the-art (SOTA) models used already for Arabic natural language processing and speech processing tasks. 61 publicly available datasets were used to support 9 task groups: Word Segmentation, Syntax and Information Extraction; Machine Translation; Sentiment, Stylistic and Emotion Analysis; News Categorization; Demographic Attributes; Factuality, Disinformation and Harmful Content Detection; Semantics; Question Answering; Speech Processing. The models GPT-3.5-Turbo, GPT-4, BLOOMZ, and Jais-13b-chat were used for NLP tasks combined with zero and few-shot learning. Following the recommended format from Azure OpenAI Studio Chat playground and PromptSource ([2]), various prompts were explored, and the most reasonable one was selected. The study revealed that in specific multilabel tasks, like propaganda detection, the LLMs sometimes generated outputs that did not fit the predefined labels. Besides that, they mention that deploying LLMs seamlessly requires substantial effort in crafting precise prompts or post-processing to align outputs with reference labels. While GPT-4 has made significant strides by closing the gap with state-of-the-art models and outperforming them in high-level abstract tasks like news categorization, consistent SOTA performance in sequence tagging remains challenging. In addition, the authors registered an averaged macro-F1 improvement from 0.656 to 0.721 by using few-shot learning (10-shot) instead of zero-shot learning.

The potential of LLMs to enhance the classification of public affairs documents was studied in the work described in [37]. The researchers gathered raw data from the Spanish Parliament, spanning November 2019 to October 2022. They acquired approximately 450,000 records, with only around 92,500 of them labeled. They concentrated on the 30 most frequent topics out of 385 labels to mitigate the impact of significant class imbalances. As models, they have used four transformer models pre-trained from scratch in Spanish by the Barcelona Supercomputing Center in the context of the MarIA project ([19]): RoBERTa-base, RoBERTa-large, RoBERTalex, and GPT2-base. Their approach involves employing transformer models in conjunction with classifiers. They conducted experiments using four models combined with three classifiers (Neural Networks, Random Forests, and SVMs). The results demonstrate that utilizing an LLM backbone alongside SVM classifiers is an effective strategy for multi-label topic classification in public affairs, achieving accuracy exceeding 85%.

An improvement of the GPT-3 performance on a short text classification task, using data

augmentation, is explored in [3]. The authors pretend to classify whether a question is related to data science by comparing two approaches: augmenting the GPT-3 Classification Endpoint by increasing the training set size and boosting the GPT-3 Completion Endpoint by optimizing the prompt using a genetic algorithm. Both methods are accessible via the GPT-3 API, each with advantages and drawbacks. The Completion Endpoint relies on a text prompt followed by ICL (zero-shot or few-shot), but its performance is notably influenced by the specific examples included. In contrast, the Classification Endpoint utilizes text embeddings and offers more consistent performance, although it necessitates a substantial number of examples (hundreds or thousands) to achieve optimal results. The dataset used in the study consists of 72 short text questions collected from the University of Massachusetts Dartmouth Big Data Club's Discord server. In Classification Endpoint Augmentation, GPT-3 was employed to generate new questions. Among the approaches, the embedding-based GPT-3 Classification Endpoint achieved the highest accuracy, approximately 76%, although this falls short of the estimated human accuracy of 85%. On the other hand, the GPT-3 Completion Endpoint, optimized using a genetic algorithm for in-context examples, exhibited strong validation accuracy but lower test accuracy, suggesting potential overfitting.

The study in [35] presents a comparison on the quality of annotations generated by humans and LLMs for Turkish, Indonesian, and Minangkabau NLP tasks (Topic Classification, Tweet Sentiment Analysis, and Emotion Classification). In their study, the authors used three Turkish datasets, each designed for one of the NLP tasks. Additionally, they employed two Indonesian datasets: one customized for Tweet Sentiment Analysis and the other for Emotion Classification. Furthermore, they included two Minangkabau datasets translated from the Indonesian datasets. The study employed the following LLMs: ChatGPT-4, BERT ([13]), BERTurk (a fine-tuned Turkish version of BERT), RoBERTa ([28]) (fine-tuned on specific datasets), and T5 ([33]). Human annotations consistently outperformed LLMs across various evaluation metrics, serving as the benchmark for annotation quality. While ChatGPT-4 and BERTurk demonstrated competitive performance, they still fell short of human annotations in certain aspects. The trade-off between precision and recall was observed among the LLMs, highlighting the need for better balance in these two measures.

The use of LLMs for moderating online discussions is investigated in [17]. The focus is on identifying user intent in various types of content and exploring content classification methods. As data sources, the authors have used various datasets, such as the One Million Posts Corpus dataset by the Austrian Research Institute for Artificial Intelligence (OFAI) of German comments made on the Austrian newspaper website's [46]. Another dataset used was the New York Times Comments collection with over two million comments on over 9,000 articles. The LLMs they used were obtained from the Detoxify python library. Their research highlights effective LLM approaches for discerning authors' intentions in online discussions and that fine-tuned AI models, based on extensive data, show promise in automating this detection.

The authors of [57] report their results for classifying the Corporate Social Responsibility (CSR) Themes and Topics shared task, which encompasses cross-lingual multi-class and monolingual multi-label classification. The shared task involved two subtasks: cross-lingual, multi-class classification for recognizing CSR themes (using one dataset) and monolingual multi-label text classification of CSR topics related to Environment (ENV) and Labour and

Human Rights (LAB) themes (using two datasets). For text classification, the LLMs used were GPT-3.5 and GPT-4 (both zero-shot and without fine-tuning), as well as fine-tuned versions of DistilBERT ([45]), BERT ([13]), RoBERTa, and RoBERTa-large ([28]). For the themes dataset, the authors used fine-tuned versions of Multi-Lingual DistilBERT, XLM-RoBERTa, and XLM-RoBERTa-large ([8]). Their zero-shot experiments with GPT models show they still lag behind fine-tuned models in multi-label classification.

3.3.2.2 2^{nd} Search (September, 2024)

The use of embeddings in the mental health domain is investigated in [29]. Various feature selection techniques were applied to the embeddings derived from the Llama-2 ([54]) and MentaLlama ([60]) models to achieve detailed classification of mental health topics by removing redundant features. The authors used two datasets consisting of requests or responses along with the corresponding questions.

- *Counsel-Chat*: Sourced from the HuggingFace library, it contains 3,451 samples across 28 unbalanced topics. The distribution of samples per class is as follows: minimum = 148, maximum = 589, mean (μ) = 123.25, and standard deviation (σ) = 143.71.

- *7Cups*: It contains 142,230 samples across 39 unbalanced topics, with a minimum of 14 and a maximum of 24,891 samples per topic. The μ is 3,646.92 and the σ is 6,091.36. This dataset contains topics similar to those in Counsel-Chat, with some variations in names and additional topics.

Topics with less than three questions and duplicate samples were removed. Finally, the datasets were divided into training, validation, and test sets with proportions of 80%, 10%, and 10%, respectively. To provide more detail about the models, two versions of Llama-2 with 7 billion parameters were used: the original developed by Meta, and a variant trained on ten mental health datasets covering eight mental health tasks (MentaLlama). Using double NF4 quantization (loaded in 4 bits) and 16-bit floating-point format for parameter computation, the authors managed to fit each model on less than 4 GB of GPU. To extract the 4,096 dimensional embeddings from the LLMs, the text of each sample was fed into the Llama-2 model without any modifications. An average over the temporal dimension was then computed to create a compact representation for the feature selection stage. During this stage, ANOVA F-values were used to rank the best feature sets. Two additional stages were then performed:

1. By selecting the optimal combination of hyperparameters and various classifiers (such as SVM, logistic regression, k-NN, etc.) that performed best on each dataset with the full set of 4,096 dimensional features, the authors determined the weighted F1 score, which serves as a benchmark for the next phase.

2. In this phase, the authors selected the classifier that previously achieved the highest weighted F1 score and retrained it using subsets of reduced dimensions based on the ranking from the feature selection strategy. The number of dimensions was incrementally increased until the weighted F1 score reached or approached the reference value.

Their results show that it is possible to reduce the dimensions of the input embeddings by up to 75% (from 4,096 dimensions to about 1,000), thereby reducing the complexity of the models without significantly affecting their performance. Additionally, they discovered other interesting facts. By comparing the selected embeddings from Llama-2 and MentalLlama, they found that the percentage of repetitiveness in the optimal feature set is 81.80% for Counsel-Chat with 1,000 features and 77.60% for 7Cups. This raises the question of whether it is more effective to extract embeddings from a fine-tuned model tailored to a specific task, or to use an open-domain version. When comparing embeddings across different datasets, the overlap percentage for 1,000 features is 46.20% for Llama-2 and 46.40% for MentaLlama. These results indicate that the selected embeddings can vary across different datasets, even when the data and tasks appear similar.

An advanced methodology for classifying financial news topics is proposed in [61]. The dataset used consists of 21,107 financial tweets annotated with twenty different labels. The data is divided into 80% for training and 20% for testing and comes from a Kaggle competition. For LLM, the Chatglm3-6b model ([15], [64]) was used and fine-tuned using LoRA integrated with Noise Enhanced Fine-Tuning (NEFTune). This method aims to mitigate overfitting and improve model performance at no additional computational cost by incorporating uniform noise into the input embeddings. Their experimental settings were as follows: a learning rate of 5e-5, a batch size of 8, gradient accumulation set to 2, and a LoRA rank of 9. The results show that the Chatglm3-6b model outperforms other models, such as Bert-Base, Bert-Large, and Deberta variants, in accurately understanding and classifying financial news content.

Another study in the same research area as [61] that also uses the Chatglm3-6b LLM is [62]. While both studies use the same model, they also appear to use the same dataset. Although [62] states that the data is from the Twitter Financial News dataset, which is available in Kaggle, [61] only mentions that it is from Kaggle. However, the number of samples and features are identical, leading us to believe that they are the same dataset. Despite the similarities in LLM and dataset, the fine-tuning techniques differ. In this study, the authors employed the QLora technique ([12]). Their configuration included a batch size of 4, a learning rate of 5e-5, the Adam optimizer, a quantization level of 4, a LoRA rank of 8, and gradient accumulation of 4. They benchmarked their model against others such as *Roberta-Base*, *Roberta-Large*, *Deberta-V2-base*, and *Deberta-V2-large*. The proposed fine-tuned model slightly outperformed the competitors, achieving an accuracy of 0.8815 compared to the lowest accuracy of 0.8615.

The effectiveness of GPT-4 for sentiment analysis and topic classification in low-resource languages, specifically Faroese news texts, is investigated in [11]. The study used the Basic Language Resource Kit 1.0 for Faroese, an open source text corpus. This corpus contains short news articles from the Faroese online news sites Portalurin and Dimmalætting, comprising 44,042 words in 170 articles. For the dataset, one to three complete sentences were randomly selected from these articles for annotation and analysis, resulting in a total of 225 sentences. In the sentiment analysis, each article was scored at both the sentence level (using the randomly selected sentences) and at the full article level. Each sentence and article was classified into one of three classes: positive (1), neutral (0), or negative (-1). In the topic analysis, GPT was instructed to assign one or more topics from a predefined list to each article. Two linguists, native speakers of Faroese, served as human annotators and verified GPT's responses for both tasks. In order

to reproduce the annotation results as accurately as possible, the annotators developed a set of guidelines for annotating news items. GPT-4 was prompted with a temperature of 0 using its function API to extract structured information from news articles according to a specified JSON schema. The schema required GPT-4 to split the text into sentences, translate to English each sentence, assign a sentiment, assign one or more topics to the article, and determine an overall sentiment for the article. One variation of the study eliminated the translation requirement. It was found that GPT-4 occasionally failed to split sentences correctly (such errors were omitted), and sometimes suggested topics that were not in the predefined list, which were corrected by the annotators. Moreover, the authors also studied the use of few-shot instructions for sentiment analysis, describing what the task consists of and providing an example for each possible class. Their results showed a moderate inter-annotator agreement between human annotators and GPT-4, suggesting that GPT-4 can reliably classify sentiment in Faroese news texts. Regarding the different approaches used, they did not observe any improvement by translating the news into English. However, they did observe an increase in agreement when using the few-shot sentiment instruction.

The development of a threat intelligence knowledge graph that leverages LLMs to classify topics in collected reports, extract entities and relationships, and identify TTPs (The Tactics, Techniques, and Procedures) from attack descriptions is presented in [22]. The authors gathered the necessary data from various open source threat intelligence platforms, including security company content platforms, security news, and influential personal security blogs. These online resources provide detailed information about malware, vulnerabilities, threat actors, and attack activity, making them valuable sources of threat intelligence. To create a dataset for the tasks *Topic Classification* and *Entity and Relationship Extraction*, GPT-3.5-Turbo with few-shot learning was used to generate data for fine-tuning the Llama2-7B model. Different prompts and few-shot examples were used for each task. Due to the specialized knowledge required for network security and TTP, GPT could not be used for this purpose. Instead, two sources of manually labeled data, MITRE and CTID, were used for fine-tuning. The training dataset follows the format 'instruction+input+output'. The instruction provides a natural language description of the task, the input is the specific segment to be processed, and the output is the expected response. For the *Entity and Relationship Extraction* task, the instruction includes several output examples to ensure higher quality responses. A report only proceeds to the *Entity and Relationship Extraction* task if it successfully passes the *Topic Classification* task. This ensures that only relevant reports that pertain to malware, security vulnerabilities, or attack activities are analyzed further. For cost-effectiveness, the Llama2-7B model was fine-tuned using the LoRA technique on Python 3.9, utilizing 2 x 3090 GPUs. The maximum learning rate was set to 1×10^{-4}, with a maximum of 10 epochs. The maximum sequence length was configured to 1,024 tokens for the entity task and 512 tokens for the other tasks. In addition to the Llama2 model, non-fine-tuned GPT-3.5 and GPT-4 models were also used for comparison, in the *Entity Recognition* task. The Llama2-7B fine-tuned with 1,600 samples achieved the highest precision, GPT-3.5 had the best recall, and GPT-4 obtained the best F1 score. In *TTP Classification*, the Llama model achieved accuracy, precision, recall, and F1 scores above 96% for broad categories. For more specific categories, it achieved an F1 score of 87.50%.

The authors of [48] present *The Babel Machine*, a system for automatically classifying input

files based on the 21 major policy topics outlined in the codebook of the Comparative Agendas Project (CAP). The data used for this system come from publicly available datasets on the CAP website and data provided from the international CAP community. According to the authors, the datasets were labeled by human coders and met the quality standards of the CAP project. The training data includes 22 categories (21 major CAP policy topics and one "no policy content"category), spans nine languages, and covers ten domains (e.g., media, legislative, executive speech), totaling 2.66 million records. The data was split into training (80%) and test (20%) sets. The XLM-RoBERTa model ([8]), which supports 100 languages and has 270 million parameters, was used as the LLM. The model, which accepts up to 512 tokens as input, was accessed via the HuggingFace repository and tuned using the Transformers library. Fine-tuning parameters included a batch size of 8, a learning rate of 5e-6, and a dropout rate of 0.1 in the final classification layer to prevent overfitting. The hardware used was a single NVIDIA A100 GPU with 80 GB of RAM. In total, 61 XLM-RoBERTa models were tuned: 1 pooled model (all data), 10 domain models (one for each domain), 9 language models (one for each language), and 41 language-domain models (one for each language-domain pair). For the language-domain models, 24 achieved a weighted macro F1 score above 0.75, with 6 exceeding 0.90. In general, the language-specific models did not provide any improvement over the pooled model, which had already reached gold standard quality, with the exception of some languages, such as Hungarian. While the domain-specific models also showed improved performance, with the legislative domain achieving the highest average scores. Based on these results, the authors recommend using LLMs as a primary option, while using human coding for validation or active learning processes in more challenging domains or languages.

Table 3.1 shows the training methods used, the evaluation metrics, and the main results of this literature review.

Table 3.1: Articles summary information

Article	**Method**	**Evaluation Metrics**	**Description**
[43]	ICL	Weighted F1	Topic: 0.34; Subtopic: 0.78 (DailyDialog)
		Accuracy	Topic: 81%; Subtopic: 62% (SGC5)
[6]	Fine-tuning	Macro-avg F1	Subtask A: 0.83, 0.82 and 0.80, respectively.
			Subtask B: 0.83, 0.85 and 0.74, respectively.
[20]	Fine-tuning	F1	Subtask A: 0.82 (extremIT5) and 0.86 (extremITLLaMA).
			Subtask B: 0.81 (extremIT5) and 0.86 (extremITLLaMA)
[56]	ICL	F1	Sentiment Classification: 88.35%.
			Topic Classification: 90.56%.
[24]	Fine-tuning	Macro-avg F1	0.80 compared to the best ML model with 0.76

Continued on next page

Table 3.1: Articles summary information (continued)

Article	Method	Evaluation Metrics	Description
		Weighted avg F1	0.90 compared to the best ML model with 0.88
[5]	ICL	Accuracy	Single Class: 67%; Top two system suggestions: 90.56%.
[1]	ICL	Macro-avg F1	Few-shot (10-shot): 0.721; Zero-shot: 0.656.
[37]	Fine-tuning	Accuracy	Accuracies higher than 85%.
[3]	ICL	Accuracy	LLM: 76%; Estimated Human: 85%.
[35]	Fine-tuning; ICL	Avg F1	Human: 0.883; GPT-4: 0.865.
[17]	Fine-tuning	F1	Identifying user intent: 0.755.
[57]	Fine-tuning; ICL	F1	Zero-shot experiments lag behind fine-tuned models.
[29]	Fine-tuning	Weighted F1	In this case, the LLM was not used as a classifier, but rather to extract its embeddings.
[61]	Fine-tuning	Accuracy	88.56%
		Rouge metrics	Rouge-1 = 0.8979, Rouge-2 = 0.5975, Rouge-L = 0.8960
[62]	Fine-tuning	Accuracy	88.15%
[11]	ICL	Accuracy	91.2% in the Topic Annotation
		Cohen's Kappa	Between Humman Annotators and GPT-4: 0.70 at sentence-level; 0.57 at document-level
[22]	Fine-tuning; ICL	Accuracy	97.47% - TPP broad categories; 83.60% - TPP specific categories
		F1	98.21% - TPP broad categories; 87.50% - TPP specific categories; 86.33% - Named Entity Recognition
[48]	Fine-tuning	Weighted macro F1	Various tests performed

3.4 Conclusions

RQ1: What empirical studies have been conducted in LLM-based content classification?

Although there are currently only a few approaches to topic/content classification using LLMs, the field is emerging and is expected to grow and improve significantly in the future, as evidenced by the appearance of six new studies (50% increase) in just three months.

Despite the limited number of studies, their analysis reveals a wide variety of methodologies, including different approaches (e.g., ICL vs fine-tuning, prompting strategies) and model architectures (encoder-only, encoder-decoder, decoder-only), as well as diverse research areas being explored (Table 3.2).

Table 3.2: Research Areas Explored

Research Area	**Articles**
Hierarchical topic/subtopic detection in inconsistent chatbot responses	[43]
Socio-political phenomena during health crises	[5]
Public affairs documents	[37]
Customer feedback	[24]
Corporate Social Responsibility themes and topics	[57]
Conspiracy Content	[6], [20]
Sentiment Analysis	[11], [35], [56]
Emotion	[35]
Benchmarking of NLP and speech processing tasks (Arabic)	[1]
Short questions	[3]
User intent in online discussions	[17]
Comparison of annotations generated by Humman and LLM	[11], [35]
Mental-Health	[35]
News	[11], [56], [61], [62]
Threat Intelligence	[22]
Public Policy Analysis	[48]

RQ2: What were the relevant contributions of the existing studies?

Based on the available studies, fine-tuned LLMs outperform LLMs prompted with ICL techniques ([3], [57]). When fine-tuning models, it is essential to carefully consider the choice between an encoder-only model, a decoder-only model, or an encoder-decoder model. Each architecture has distinct characteristics and implications for the model's behavior and performance. Furthermore, an analysis of the studies by [62] and [61] shows that different fine-tuning strategies can lead to similar results. Both authors used the same LLM for the same problem, but used different fine-tuning techniques. However, achieving optimal performance requires substantial computational resources and a dataset containing hundreds or thousands of examples. LLMs can be prompted using zero-shot or few-shot techniques as a more cost-effective alternative. A comparison between these two methods for a specific case was conducted in [1], [11], revealing that few-shot outperformed zero-shot. Importantly, the choice of few examples is critical ([56]), and there are inherent limitations in the reasoning abilities of LLMs. Researchers [1], [5], [11], [22] have pointed out problems related to model hallucinations. For example, [11] found that GPT-4, when applied to a classification task, suggested topics that were not in a predefined list. Another important observation was made by [29], who found significant

redundancy in the embeddings derived from Llama models. In some cases, it may not be beneficial to extract embeddings from a fine-tuned version, which can help reduce computational complexity by using only a small subset of the embeddings or by not needing to fine-tune a model.

RQ3: Can LLMs be used to assess the structural quality of studies?

While the results suggest that using ICL zero-shot is not yet reliable, we conclude that evaluating the quality of scientific articles with LLMs may be feasible. This could be achieved through more extensive research with a fine-tuned model or by using ICL few-shot examples.

3.4.1 Summary

In short, this is a new field, as evidenced by the number of studies and the years of publication. Consequently, further research is needed to improve performance, efficiency, transparency, and reasoning while reducing model hallucination. Possible areas could include analyzing the influence of the architecture, the number of layers and parameters, and the quality of the training dataset to be used.

3.4.2 Threats to validity

The following types of validity issues were considered when interpreting the results from this review.

3.4.2.1 Construct Validity

A literature database of relevant books, conferences, and journals served as the source for the research found in the systematic review. Therefore, bias in selecting publications is a potential drawback of this strategy, especially considering that three of the thirteen articles were submitted to the same workshop. To address this, we used a research protocol that included the study objectives, research questions, search approach, and search terms. Inclusion and exclusion criteria for data extraction were established to reduce this bias further.

Our dataset only includes studies published in the last two years (2023 and 2024), making it challenging to identify trends due to the recent and limited sample size. Moreover, the studies on LLM-based content classification only used well-established taxonomies, such as news categorization and fake news topics. None of the studies used a taxonomy the model had not encountered during its training process.

3.4.2.2 Internal Validity

No studies were excluded during the quality assessment due to the low number of documents retrieved in the search, so there is no potential threat to internal validity. In other words, we did not exclude studies that could contribute significantly despite their lower quality.

3.4.2.3 External Validity

There may be other valid studies in digital libraries that we did not search. However, we attempted to mitigate this limitation using the most relevant literature repository. Additionally, studies not written in English were excluded, which may have omitted important papers that would otherwise have been included.

3.4.2.4 Conclusion Validity

There may be some bias during the data extraction phase. However, we have addressed this by defining a data extraction form to ensure consistent and accurate data collection to answer the research questions. While there is always a small chance of inaccuracies in the numbers, we mitigate this by publishing our final dataset, allowing for replication and further validation.

[This page has been intentionally left blank]

Chapter 4

Design and Implementation

Contents

This chapter describes the proposed Smart ETL approach.

[This page has been intentionally left blank]

Chapter 4
Design and Implementation

4.1 Smart Extraction Phase

In this section, we present our current extraction phase, where textual and graphical data, if available, are extracted from the catalogs.

4.1.1 Extraction of elements from Catalogs

Before explaining the scraping of the STT catalogs, Table 4.1 outlines the specific elements extracted from each catalog. It is important to note that not all elements were extracted from every catalog due to variations in the available information. Cells with the value 'AI' instead of 'X' indicate that the meaning of the extracted element was identified using AI models. Although the main focus of this research is the application of LLMs for the automatic classification of STTs, AI has also been used to extract content from catalogs.

Table 4.1: Identification of the elements extracted from each catalog

	Catalogs of SEGITTUR	Catalogs of ADESTIC	Catalogs of EU
STT's Name	X	X	
STT's Description	X	X	
Producer's Name	X	X	
Producer's Description	X	X	
Producer's Logo	X	X	
Producer's URL	X	X	
Scope of SD application	X		
Type of Destination	AI	X	
Type of Solution	AI	AI	
Producer's Phone number and Email address		X	
Producer's Address		X	
QR Code		X	
STT Application's Title			X
STT Application's Location			X
STT Application's Description			X
STT Application's URL			X
STT Application's Image(s)			X

The *Producer Logo* and *STT Application Image(s)* elements represent images, while the *Type of Destination* in SEGITTUR catalogs and the *Type of Solution* represent icons with corresponding labels. In contrast, the remaining elements consist of textual data.

Regardless of the element's data type, the PyMuPDF library was used to extract them.

In the ADESTIC catalogs, explicit addresses were not available. Instead, when a landline phone number was provided (mobile phone numbers were not feasible for this purpose), we

attempted to identify its associated region. To achieve this, we extracted all Spanish prefix numbers corresponding to different Spanish regions. Once a match was found, we used the GeoPy Python library to obtain the corresponding coordinates, so that we could insert them into the observatory's map area. Consequently, the *Address* element does not represent the precise producer address but rather indicates the region from which the producer originates. The Pyzbar library was used for the QR code extraction.

4.1.2 Leveraging AI for Graphical Element Extraction

Besides extracting graphical elements, it's crucial to obtain their labels from the initial glossary of each respective catalog. However, for the *Producer Logo* element, this is not necessary since it does not represent a category. Given that the corresponding label for each icon is exclusively shown in the glossary and not on the STT pages, we utilized the resnet-18 ML model for image classification. This model is accessible via the fastai Python library, and here is how we used it:

1. The icons from the graphic glossary were extracted and saved in different image formats and resolutions.
2. Data augmentation techniques were applied to increase our training data. It resulted in a dataset three times larger than the original.
3. resnet-18 was fine-tuned with the augmented training data.
4. Our objective is to classify images from the glossary (which served as our training dataset). Specifically, we aim to classify images the model has already found rather than dealing with unseen images. So, we intentionally induce overfitting, which is typically not the goal.

We followed this procedure for every catalog containing an iconographic glossary.

4.1.3 Extracted STTs Overview

Figure 4.1 presents the BPMN process model for our smart extraction phase.

The code used in this section is available on the RESETTING page on GitHub.

4.2 Smart Transformation Phase

Tables 4.2 and 4.3 present some extracted elements that were incorporated into DCTERMS properties. Also, a DCTERMS property may comprise more than one element.

Omeka allows us to incorporate the content of each item in HTML format. So, after extracting the elements, we transform them into HTML format, which gives us more flexibility in how the content is laid out. In addition, since ADESTIC's catalogs are in Spanish, the deep-translator library was used to translate them into English.

In the Transformation phase, LLMs were used for removing duplicates (Section 4.2.2) and classifying STTs (Section 4.2.3). Before diving into the tasks, the choice of the LLMs is detailed (Section 4.2.1).

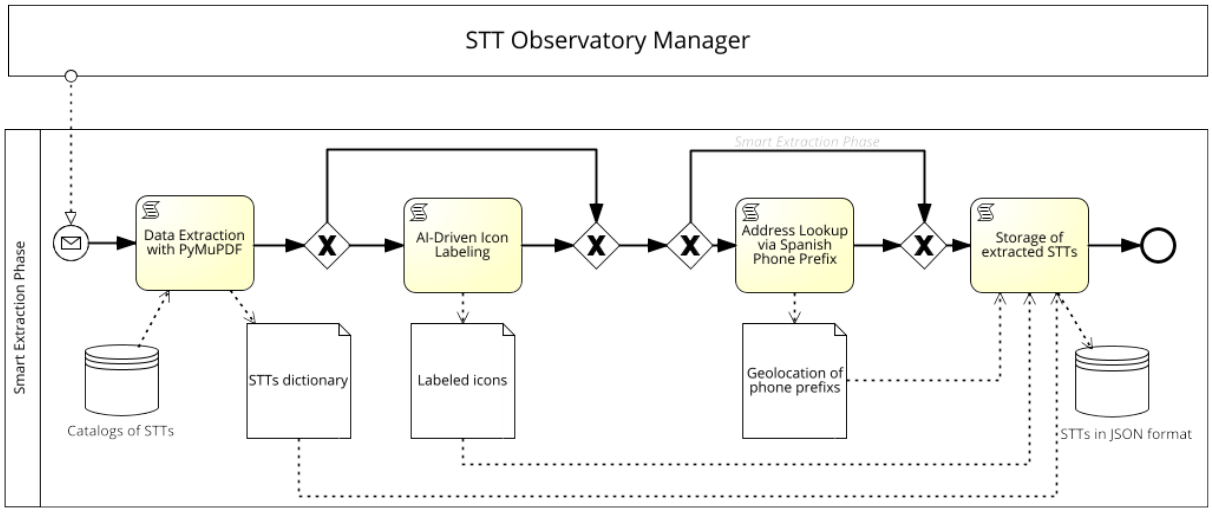

Figure 4.1: BPMN process model for the Smart Extraction Phase

Table 4.2: DCTERMS property associations from Spanish catalogs

DCTERMS property	Element(s) Extracted
Title	STT Name
Description	STT Description
Creator	Producer Name; Producer Description; Producer Logo; Phone Number
Local URL	Producer URL

Table 4.3: DCTERMS property associations from EU catalogs

DCTERMS property	Element(s) Extracted
Title	STT Application Title
Description	STT Application Description
Address	STT Application Location
Local URL	STT Application URL

4.2.1 Choice of LLMs

Due to resource constraints, fine-tuning a LLM with satisfactory performance was not feasible. Hosting an LLM solely for inference was also not a viable option. Consequently, model selection should exclusively consider decoder-only models, since they are the only type accessible online without the need for local execution.

Hugging Face is a French-American company and an open-source community that focuses on NLP models and tools. They are renowned for their Transformers library, an open-source platform that offers user-friendly interfaces to cutting-edge pre-trained NLP models. Besides that, they have also developed HuggingChat, e.g., a free AI-powered conversational agent with the latest NLP models. Hugchat is an unofficial Python API for HuggingChat, but notably, Hugging Face's Chief Technology Officer (CTO) has expressed appreciation for the project on their GitHub page.

For the duplicates removal task we used, through the hugchat, the *mistralai/Mixtral-8x7B-Instruct-v0.1*. At the time of execution, this model was hugchat's default LLM, and since we had difficulties trying to change the LLM (i.e., we tried to choose a different LLM but in the end the model used was the default one), we continued with it. In any case, it was one of the best LLMs available at the time.

Regarding the primary subject of this paper, the LLM-based STT classification, we employed different LLMs to assess their performance: *Microsoft Copilot*, *NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO*, *mistralai/Mixtral-8x7B-Instruct-v0.1*, and *meta-llama/Meta-Llama-3.1-70B-Instruct* (the last three from hugchat).

4.2.2 Duplicates Removal

Since both the ADESTIC and SEGITTUR catalogs focus on Spanish STTs, there are repetitions between them. However, when ADESTIC is translated into English, the STT names may not match those in the SEGITTUR catalogs. Furthermore, solution types in different catalogs may not share identical names, even if they represent the same concept, and a type in one catalog may cover several types in another catalog. As a result, manual associations have been established between solution types. However, some solution types remain unassociated because there is no possible association with the types in another catalog. The manual association is available here.

Therefore, the duplicate removal process works as follows:

1. The STT and producer names between two STTs are compared.

2. If no match is found, the STTs associated with similar solution types are retrieved. For instance, when searching in the ADESTIC V2 catalog for duplicates of a STT from the SEGITTUR 2022 catalog with the solution type *"Efficient Management: Energy"*, we will be looking for STTs having the solution type *"Efficient Management: water, air, energy or waste"*.

Example of Similar Solution Types

```
{
    "Segittur 2022": [

        "Efficient Management: Energy",
        "Efficient Management: Water",
        "Efficient Management: Air quality",
        "Efficient Management: Waste"

    ],
    "Segittur 2023": "Efficient resource management: water/energy/waste/air quality",
    "Adestic V1": "Efficient Management: water, air, energy or waste",
    "Adestic V2": "Efficient Management: water, air, energy or waste"
}
```

However, there is another limitation that has been prevented. It is explained using an example to facilitate interpretation. When searching within the SEGITTUR 2023 catalog

for duplicates of a STT from the ADESTIC V1 catalog with *"Accessibility"* as the solution type, and if the SEGITTUR 2023 catalog lacks that solution type (see below **Example of Solution Type Not Available in All Catalogs**), all solution types without an association between these two catalogs are retrieved (see below **Unassociated Solution Types**). This process helps identify solutions that are the same but were categorized differently by the catalog administrators.

Example of Solution Type Not Available in All Catalogs

```
{
   "Segittur 2022": "Accessibility",
   "Segittur 2023": "",
   "Adestic V1": "Accessibility",
   "Adestic V2": "Accessibility"
}
```

In a real application, the "{...}"in **Unassociated Solution Types** would be replaced by the remaining types.

Unassociated Solution Types

```
{
   "Segittur 2022": "Accessibility",
   "Segittur 2023": [

      {

        "Segittur 2022": "",
        "Segittur 2023": "Intelligent Signage/Totems/Tourism Signage",
        "Adestic V1": "",
        "Adestic V2": ""

      }, {...}],

   "Adestic V1": "Accessibility",
   "Adestic V2": "Accessibility"
}
```

3. Subsequently, each STT will undergo comparison using the LLM. The prompt provided in a zero-shot setup is available on the European STT Observatory GitHub page, accompanied by an application example.

Through the application of this process, unanticipated outcomes emerged. The LLM not only successfully identified identical STTs, but also revealed instances where an STT, originally comprising multiple functionalities within one catalog was fragmented into several distinct STTs in another catalog. We decided to keep those that were separate and eliminate those that included several STTs.

The number of duplicates identified were:

- **By STT and producer names comparison:** 156
- **By LLM evaluation:** 57

The code created for this task is available here.

4.2.3 STTs Classification

As noted above, the European Commission catalogs were not used as a data source for this task because they do not specifically present STTs.

While several LLMs were evaluated, only Microsoft Copilot in Precise Mode was employed to define the necessary prompts. This decision was influenced by the fact that, at that time, the GPT-4 Turbo powered Copilot Precise mode and was considered one of the top-performing models.

4.2.3.1 Needle in a Haystack Challenge

Although the context length of Copilot is known, we conducted a test to determine the model's ability to detect fine details within the provided context. This detection is essential for our task, so we can know the model's limit to understand the taxonomy and the few-shot examples provided. The test was a needle in a haystack challenge, where the haystack was the text of an extracted book and the needle was a short sentence out of the context. The needle was randomly placed in the book text, and the LLM was asked to find it.

- **Haystack:** The Sonnets by William Shakespeare
- **Needle:** *"Portugal's national team became European champions in 2016 against France. The final was played in Paris and the final score was 1-0 after extra time."*

We initiated the test with a prompt character limit of 10,000. If the LLM responded correctly, we increased the limit by 500 characters. However, if the model provided incorrect answers 6 times consecutively, the test was concluded. In addition, we calculated the quartile in which the needle was randomly inserted in the text of the book so that we could get a better idea of the influence on the model's response.

The prompt provided to the model was:

Prompt template for the Needle in the Haystack Challenge

"Let's do the needle in a haystack challenge. In the following text, you have to find the needle, which is a sentence out of context. Good luck!
The text is: "###TEXT TO REPLACE###"
You have to return this JSON object. The 'Needle' value is the sentence out of context. If you don't find the needle the value must be 'NOT FOUND': {"Needle": <SENTENCE>}."

We conducted a total of four tests. However, all tests were prematurely terminated due to reaching the Copilot character limit. As a result, we were unable to accurately determine the precise character limit at which the LLM can still successfully locate the needle. Despite this limitation, we were able to derive several valuable insights:

- Copilot may be capable of finding the needle in the haystack when the character limit is above 23,000.
- Figure 4.2 shows that Copilot performs better at detecting small details in the second half of the prompt, as evidenced by the lower percentage of errors in the third and fourth quartiles. Note again that the needle was placed randomly, so n is not the same for all quartiles.
- The requested JSON object was provided in 85% of the responses, for a total of 151 responses. This higher accuracy may confirm the previous point that the model is better at detecting small details in the second part of the prompt, since the JSON request is the final statement. However, for a comprehensive analysis, the JSON request should be placed in different parts of the prompt for comparison, which was not done in this research as it was not the primary focus.
- Figure 4.3 illustrates that the probability of errors increases as the maximum character limit increases. The 19,500-23,000 range had almost as many errors as the 10,000-19,000 range, with 22 errors compared to 21. It is important to note that the test was terminated after six errors, not after the first error. This approach allows us to estimate the likelihood of Copilot making errors at each character limit.

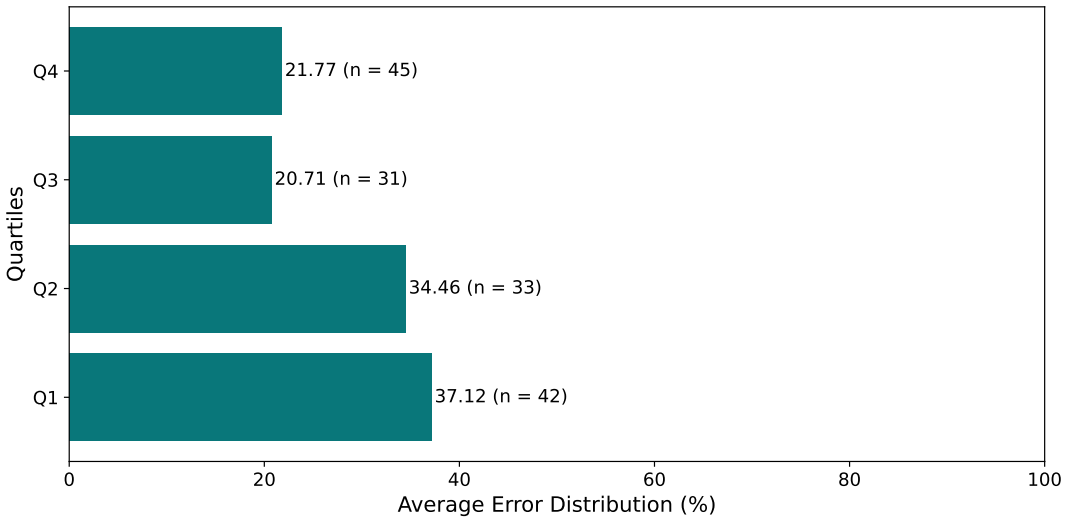

Figure 4.2: Quartile-based analysis of error distribution

4.2.3.2 Classification Methodology

Based on these results, we decide on a maximum character limit of 15,000 characters for the STT classification task. For that, we employ both training and testing datasets. However, due to the absence of a large categorized dataset, we manually categorized some STTs as follows:

- **STTs Selection:** We randomly selected STTs from each Spanish catalog, since only those catalogs contained STTs. We extracted an STT from one catalog and categorized it, then repeated the process for the other catalogs.

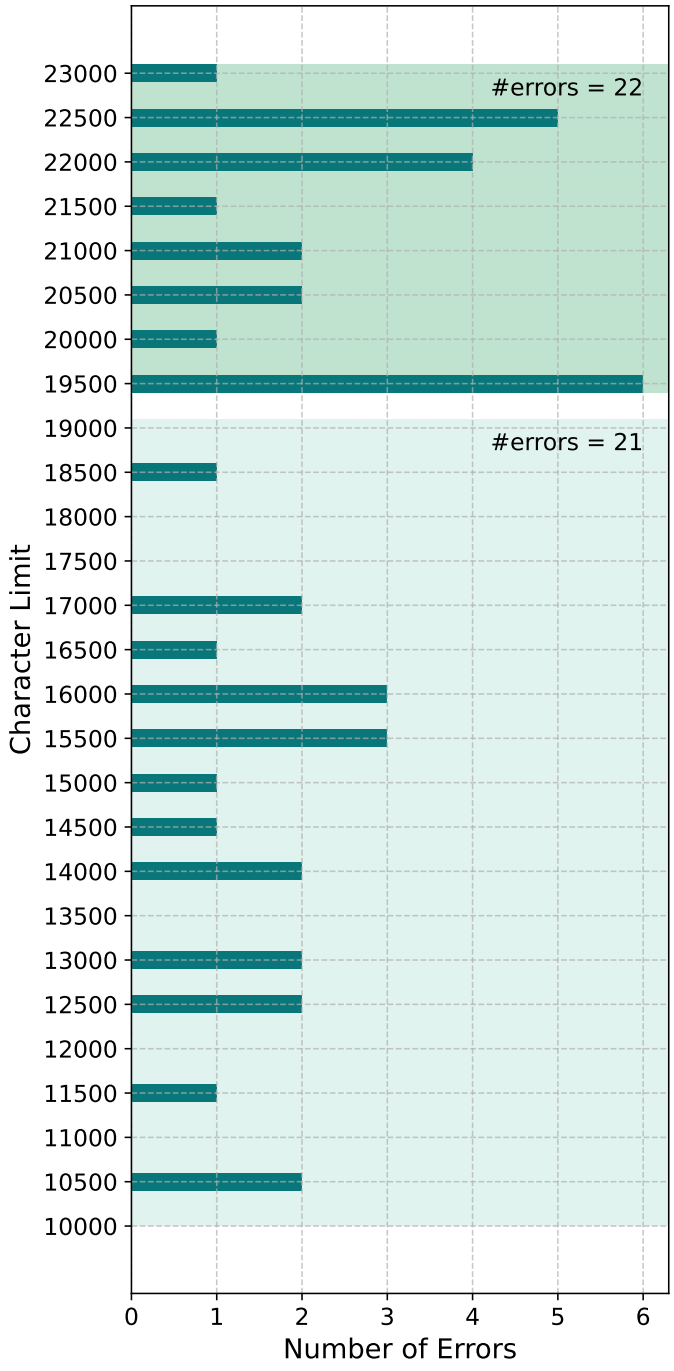

Figure 4.3: Character limit-based error analysis

- **Dataset Sizes:** After reviewing all four catalogs, we returned to the first one and repeated the process until we created two datasets, each containing STTs that covered all the categories in the taxonomy. In one test, one dataset serves as the training set and the other as the testing set. In the subsequent test, the roles are reversed, with the training dataset becoming the testing set and vice versa. Since an STT can cover multiple categories, the size of the datasets is smaller than the number of categories. Thus, Dataset A contains 8 STTs, and Dataset B also contains 9 STTs.

The prompts used for this task were created through an iterative approach, with changes or additions made in each iteration. Three types of prompts were utilized: one to describe the STT taxonomy, one for the few-shots, and another to introduce each STT to be classified. The way we verified the need to update the prompts was through the answer given by the LLM about the STTs given to be classified. From his answers, it is clear when a definition is not clear enough or when it confuses two or more categories. When it was necessary to shorten or clarify a sentence or paragraph in the prompts, we found it advantageous to use the LLM itself for this task, simply by asking the LLM to do so and providing the respective text to be updated.

- **Prompt of Taxonomy Definition:** the earlier versions of this proposal included an outdated taxonomy. We recognized the need for an update in the *"(Part of) the tourist offer"* domain because the LLM struggled to fully understand it, and we concluded that people might face the same difficulty. A report on the progress of this prompt is available online.

- **Few-shot Prompt:** we differentiated the few-shot examples from the STTs to be classified using initial identifiers in the prompts: *"###EXAMPLE###"* or *"###Classification###"*. A structure identical to the one below was used for each shot:

  Few-Shot Example Structure

  ```
  ###EXAMPLE###
  {
      "Solution Name": "...",
      "Description": "..."
  },
  {
      "Classification": [...],
      "Justification": "..."
  }
  This message is only used for context and does not require a response.
  ```

  While it does not always work (i.e., the LLM did not comply with the request), the last statement is essential in preventing, in most cases, the LLM from giving long answers to the few-shot examples. Besides that, for the justification of the classification, we created an expression for each STT category so that we could repeat it in the justification every time that category was selected, thereby achieving consistency and repetition between the examples of a few-shots to try to make the job of interpretation easier for the LLM. An example is:

  Tourist Experience: *"It is "Tourist Experience" because it requires the tourist's active participation (i.e., . . .) and it's not about the planning, organization, or execution of the activity or future activities."*

  This sentence is repeated in the justifications each time this category is selected, with "..."replaced by information that classifies the STT as "Tourist Experience". If the combination of two or more few-shots did not exceed the limit identified in the Needle in a Haystack test, they were sent in the same prompt.

- **Prompt to introduce the STT to be classified:** our initial idea was to use a prompting approach where the LLM would assume the roles of three different appropriate experts to solve the classification problem. At each step of reasoning, the experts would share their thoughts with the group, and each expert would score their peers' responses on a scale from 1 (highly unlikely) to 5 (highly likely). Finally, the LLM would analyze the three expert analyses and provide either a consensus definition or its best guess solution. However, this approach proved too complex for the LLM to complete the assigned task, so we adopted a different strategy. Instead of asking the LLM to identify and behave as three different experts suitable for the problem, we instructed the model to assume the roles of three experts, each specializing in a different application domain of the taxonomy. Each expert assigned a score from 1 (highly unlikely) to 5 (highly likely) and an explanation to their respective categories and subcategories, if applicable. The LLM then identified the STT categories with scores of 4 or higher, or those with the highest scores if none reached 4. At the end, the LLM should return a JSON object with a key for each domain and the corresponding identified categories as values (demonstrated below). This approach proved to be more viable and had better performance. In addition to the prompt, the STT to be classified is provided in a JSON object with "Solution Name"and "Description"as keys and their respective texts as values. A report on the progress of this prompt is available online.

STT Classification JSON

```
{

    "(Part of) the touristic offer": [...],
    "Marketing": [...],
    "Management & Operations": [...]

}
```

A maximum of ten prompts were sent per context length, i.e., the same conversation with an LLM. These prompts were divided into the taxonomy definition prompt, the few-shot prompts, and the STTs to be classified prompts. The number of prompts sent in the same conversation varies depending on the possible combinations of few-shots within a single prompt (see explanation of **few-shot prompt**) and the split of the test dataset. Since the total number of STTs to be classified, combined with the other prompts, exceeds ten prompts, the test data set was split into two parts. Additionally, dataset B has an odd number of entries, resulting in an additional prompt being sent in one of the conversations.

As mentioned above, the datasets were manually labeled. However, since this taxonomy is new, our experience in classifying STTs according to it is not fully mature. Therefore, during the prompt refinement process, when the LLM suggested a category for a STT that we had not initially assigned, we reviewed the justification provided. If it made sense, we added that category to the STT. The opposite also happened when we had a category that was not properly assigned.

The results obtained, the limitations, and the negative or unexpected findings of the STT classification are presented in Sections 4.4 and 4.5.

4.3 Load Phase

This was the only phase where AI was not used. One of the reasons for choosing Omeka.net is that it includes an API that allows us to automatically insert STTs into the observatory. It works via HTTP requests with the data field in JSON format, and for tasks related to inserting, updating, and deleting content, it requires an API key that only the observatory managers have. To facilitate the API processes, we have created a Python class called OmekaAPI. The class does not contain all the features that Omeka.net allows, but only the ones we need, and is available here.

To help sum up these last two sections, Figure 4.4 presents the BPMN process of the Smart Transformation Phase and the Load Phase applied in this research.

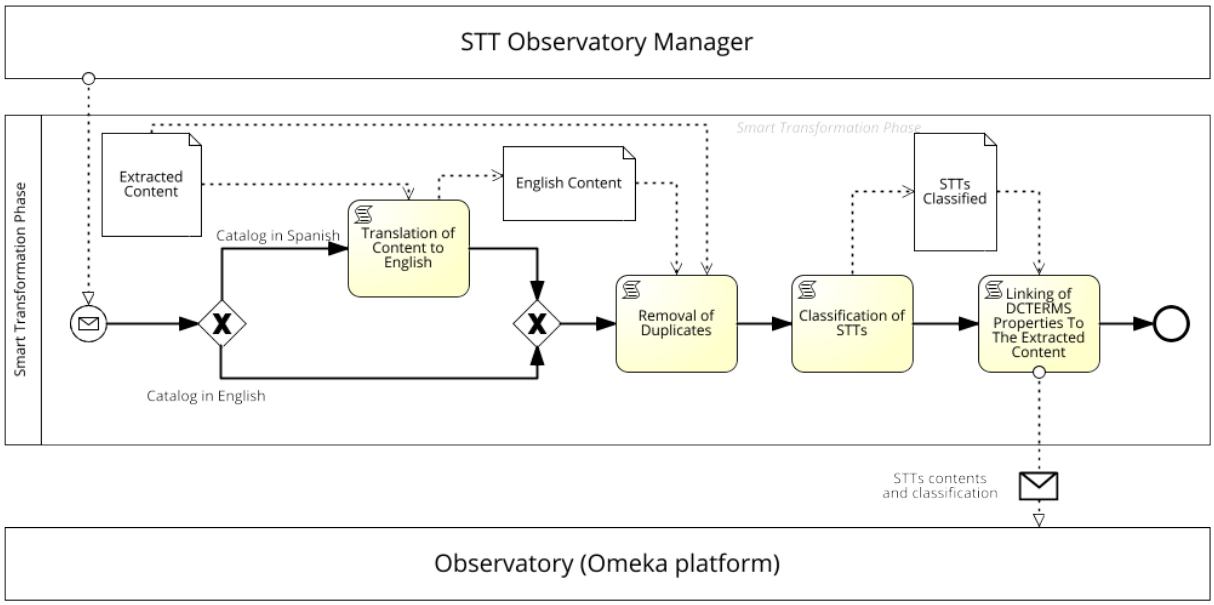

Figure 4.4: BPMN process model for the Smart Transformation and Load Phases

4.4 Results

Here, we will focus on presenting the task of the Smart ETL proposal that received the most attention since the results of the other smaller tasks have already been described. Each test corresponds to the classification of an STT, following a black-box approach, and its response is considered valid as long as the returned JSON contains the required keys and the STT categories identified as values.

For ease of interpretation, we discuss the results of the STTs classification for each LLM separately:

- ***NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO:*** The LLM failed the tests due to his difficulty in understanding the taxonomy and the classification task. This was evident after completing the first full set of tests with dataset A, as the results did not match the format of the expected results and showed a poor understanding of the taxonomy.

- ***mistralai/Mixtral-8x7B-Instruct-v0.1:*** We discarded the use of this LLM due to its recurrent hallucination, i.e., most of the classified STTs were assigned to non-taxonomy categories such as *"Tourist Information Systems"*, *"Centralized Database & Distribution System"*, *"Booking Systems"*, or *"Appointment Manager"*, among others. Running two full sets of tests, one with dataset A and the other with dataset B, was sufficient to conclude that the reasoning abilities of this model were inadequate for this task.

- ***meta-llama/Meta-Llama-3.1-70B-Instruct:*** This LLM "understood" the required task. However, the output for the *(Part of) the Touristic Offer* domain was sometimes inaccurate. Like in the previous model, but less recurrently, it hallucinated, categorizing an STT with non-taxonomy categories like *"Tourist Infrastructure"* or *"Parking Infrastructure"*. At other times, it only returned the first-level category, such as *"Tourist Experience"* or *"Tourist Experience Lifecycle Management"*. Another minor issue was splitting categories like *"Tourist Experience Lifecycle Management, Building Block"* into separate categories. About this minor error, we considered the responses valid for metrics because it is possible to conclude implicitly that they belong to the domain *(Part of) the Touristic Offer*. Table 4.4 shows the average precision, recall and F1 scores for each domain, together with the overall F1 score. Each dataset was tested twice on different days and times to assess performance variability.

Table 4.4: Performance Metrics of the *meta-llama/Meta-Llama-3.1-70B-Instruct* Model

#	Testing Dataset	Total F1	Domain	Precision	Recall	F1
1	A	0.50	(Part of) the Touristic Offer	0.13	0.19	0.15
			Marketing	0.88	0.77	0.81
			Management & Operations	0.75	0.44	0.54
2	A	**0.61**	(Part of) the Touristic Offer	0.19	0.31	0.23
			Marketing	1.0	0.85	0.90
			Management & Operations	0.81	0.69	0.71
1	B	0.47	(Part of) the Touristic Offer	0.17	0.22	0.19
			Marketing	0.67	0.54	0.57
			Management & Operations	0.74	0.63	0.64
2	B	0.37	(Part of) the Touristic Offer	0.0	0.0	0.0
			Marketing	0.61	0.54	0.54
			Management & Operations	0.70	0.57	0.59

- ***Microsoft Copilot:*** This model achieved the best results but also had the most hallucinations. The *(Part of) the Touristic Offer* domain continued to present difficulties. Common errors included:

- Treat the first-level categories (*"Tourist Experience"* or *"Tourist Experience Lifecycle Management"*) as domains. In the returned JSON, these categories were incorrectly placed as keys, which should represent the domains, with the categories as their values. Despite this, these answers were accepted because they are easily associated with the corresponding domain.
- In some tests, there was repetition in the answers. Specifically, the LLM assigned the same categorization to multiple STTs. As a result, these tests were considered invalid and excluded from the metric calculations. Since the LLM is available online, we chose not to include these invalid tests in the metrics, as they reflect instances where the LLM hallucinated or repeatedly gave the same answer, potentially indicating times when Copilot was more overloaded and performing sub-optimally.
- Occasionally, the model used information from previously classified STTs within the same context window to justify the classification of the current STT.
- Another error, which occurred only once but highlighted the hallucination issues, involved a test where instead of returning the requested JSON, the LLM returned a JSON with the categories as keys and the STT name as a value for each key.

Unfortunately, most tests were considered invalid for the reasons mentioned above. However, it was possible to carry out some tests where we would be able to measure performance. Table 4.5 shows the average precision, recall and F1 scores for each domain, together with the combined F1 score. Each dataset was tested twice at different times and on different days to assess performance variability.

Table 4.5: Performance Metrics of the *Microsoft Copilot* Model

#	Testing Dataset	Total F1	Domain	Precision	Recall	F1
1	A	0.76	(Part of) the Touristic Offer	0.88	0.81	0.83
			Marketing	0.94	0.92	0.90
			Management & Operations	0.52	0.63	0.56
2	A	0.75	(Part of) the Touristic Offer	0.75	0.69	0.71
			Marketing	0.81	0.83	0.81
			Management & Operations	0.65	0.88	0.73
1	B	0.53	(Part of) the Touristic Offer	0.33	0.22	0.26
			Marketing	0.67	0.78	0.70
			Management & Operations	0.56	0.85	0.63
2	B	0.68	(Part of) the Touristic Offer	0.78	0.67	0.70
			Marketing	0.67	0.57	0.61
			Management & Operations	0.76	0.76	0.76

4.5 Limitations

One of the limitations of this research is that the prompts were only refined using Microsoft Copilot and were not customized for each model used. As a result, the prompts may have been

over-optimized for Microsoft Copilot, potentially leading to sub-optimal performance in other models. In addition, because the models were used online rather than locally, we were affected by the variability of other users' interactions with the model. This left us uncertain about the model's capabilities during testing, making it unclear whether occasional negative results were due to potential model overload or our methodology.

Chapter 5

Verification and Validation

Contents

This chapter covers the demonstration, evaluation, and communication phases of the Design Science Research Methodology (DSRM) applied in this dissertation.

[This page has been intentionally left blank]

Chapter 5
Verification and Validation

5.1 Demonstration

This Design Science Research Methodology phase aims to demonstrate the use of the artifact created. The intended artifact is the European STTs Observatory, populated with classified STTs to support the search process of European SMEs in the tourism sector, the primary audience of the Observatory. Although the classification functionality has not been fully achieved, several demonstrations of the observatory have been carried out.

For **online demonstration**, we created a web site containing all the information related to the observatory. This site includes a user manual detailing the Observatory's functionalities, a video demo, and technical documentation summarizing the main technical aspects of the project, such as Smart ETL, Large Language Models, the methodology used, as well as the tools and a link to the GitHub repository with the implemented code.

Several **onsite demonstrations** also took place:

- At the inauguration of the new "Iscte - Knowledge and Innovation" building in Lisbon on November 20, 2023. This demonstration was attended by the then Prime Minister, the Minister for Science, Technology and Higher Education, the Minister for Cohesion, the Minister for Culture, two Secretaries of State, a large entourage representing the most diverse sectors, and the media.
- At the RESETTING Hackathon, held at Tecnoparc in Reus, Spain, on May 2024, which focused on "TourismTech solutions to address the challenges of European tourism". During this event, each tool developed in the RESETTING project, including the Observatory, was presented, accompanied by a poster, to the hackathon participants. All posters are available here.
- At the RESETTING Final Conference, held at Auditori Diputació in Tarragona, Spain, also on May 2024. During this event, the tools and posters were presented to the conference attendees, including some SME owners, our desired main target audience, who received funding from the RESETTING project.

5.2 Evaluation

The objective of this phase is to monitor and evaluate how effectively the artifact addresses the problem, specifically the lack of SMEs awareness of existing STT offerings.

5.2.1 Continuity

We sent a survey to different smart tourism stakeholders to draw conclusions on the usability and usefulness of the European STTs Observatory and get insights on possible missing features.

The survey asked about the types of professionals for whom the Observatory could be helpful, how they rated the demo, user manual, and technical documentation, how likely they would be to use the Observatory in the future, and what additional features they would like to see implemented.

The survey can be found at https://tinyurl.com/sttobservatory-survey, and at the time of writing, it had already received 40 responses, most of them complete (not all questions were mandatory). Preliminary results show that researchers, producers of Smart Tourism Tools (STTs), and managers of tourism-related businesses are likely to find the Observatory most useful (Figure 5.2). Tourists themselves and public service managers will also benefit. Figure 5.1 shows the grades received for the demo, user manual, and technical documentation. We can see that they all received high scores (7 or above) in most responses, reflecting good, though not perfect, quality. In addition, most of the respondents said they were likely to use the Observatory in the future (Figure 5.3). Among the suggested functionalities, providing information on events and research publications stands out as a feasible addition to the European STT Observatory. Among the remaining suggestions, the most interesting were:

- *"I recommend you include new technologies on your website, such digital twins, cybersecurity, metaverse, generative AI, and quantum computing in smart tourist destinations. These disruptive tools are being used in some STDs, and they will change the new paradigm of travel and tourism activities in tourism cities and STDs."*
- *"Easy way to submit STTs"*

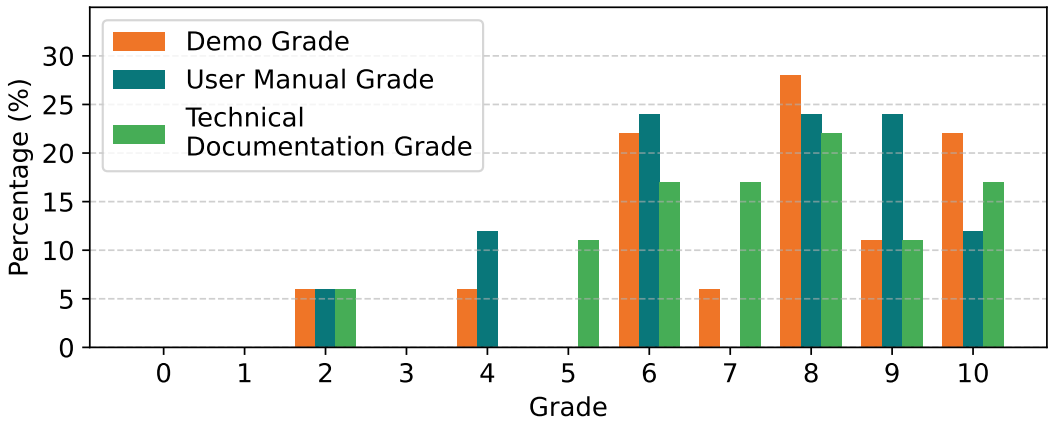

Figure 5.1: Grades of Demo, User Manual, and Technical Documentation

5.2.2 Monitoring

The Google Analytics platform is a crucial tool used to monitor access to the Observatory. The snapshot below (Figure 5.4), corresponds to the last eight months (from January to August 2024). This robust monitoring process ensures the Observatory's performance is meticulously tracked, and any necessary adjustments can be made to maintain its effectiveness.

Figure 5.2: Type of professionals most likely to use the Observatory

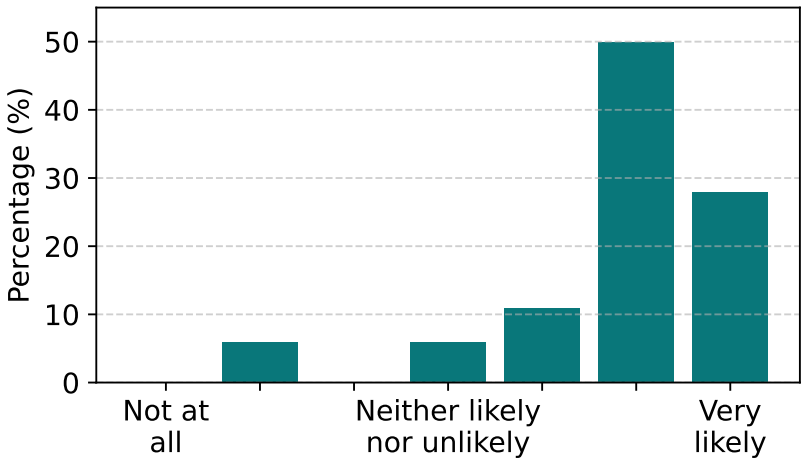

Figure 5.3: Likelihood of future use of the observatory

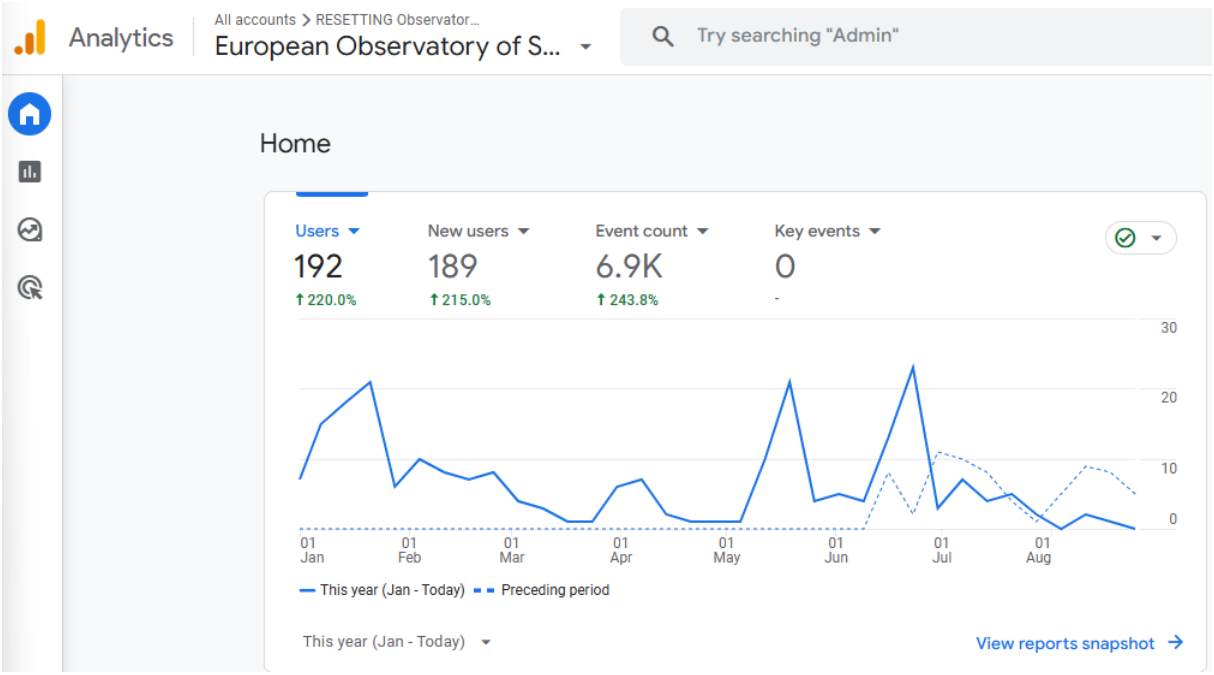

Figure 5.4: Monitoring of the Observatory with Google Analytics

5.3 Communication

To highlight the work done in this dissertation and its significance, a paper titled "A Systematic Literature Review on LLM-Based Information Retrieval: The Issue of Contents Classification," covering the literature review (chapter 3), was submitted and accepted in the 16th International Conference on Knowledge Discovery and Information Retrieval, scheduled for November 17-19, 2024, in Porto, Portugal ([10], supplemented by [10]).

Another paper covering essentially the material from chapters 4 and 5, titled "Smart ETL and LLM-based contents classification: the European Smart Tourism Tools Observatory experience", was submitted to Springer's journal Knowledge and Information Systems and is currently under review.

Chapter 6

Conclusion and Future Work

Contents

This chapter presents the conclusions and future work

[This page has been intentionally left blank]

Chapter 6
Conclusion and Future Work

6.1 Conclusion

A prerequisite for this dissertation was selecting an online platform to host the European Smart Tourism Tools Observatory. The original plan was to use a European virtual trade show platform provider that would allow the creation of customized virtual booths for each STT producer and the automatic integration of new content. However, no platform was found that met these and other requirements. So, we shifted the research target to less specific online platforms that could meet as many of our requirements as possible. As a result, Omeka was chosen as the observatory's hosting solution. Despite not being based in Europe, it provides an API for automatic insertion of STTs, has a fixed license cost, does not require visitor registration, has a customizable layout, and is robust, having been online for many years. It also supports standardized metadata and is open-source, so we are not strictly dependent on the Omeka managers.

The initial objective of this dissertation was to create a smart ETL process, where "smart"means integrating AI into a conventional ETL framework. The extraction phase was completed by extracting various types of content, including images, text, links, QR codes, phone numbers, email addresses, and more, from STT catalogs in PDF format. The transformation phase focused on developing an automatic classification system for STTs using AI, particularly LLMs. That classification (STT labeling) will facilitate tourism sector SMEs searching for innovative and sustainable business solutions.

It is essential to strive for high accuracy in classification, but achieving a perfect one (100%) is unnecessary because eliminating all errors is often unfeasible. The results demonstrated that classification is feasible but lacked the consistency required for large-scale implementation. This could be due to either the use of potentially inappropriate few-shot examples or the volatility of the LLM's reasoning capabilities, which, as an online model, can sometimes underperform due to overload. There is also the problem of hallucinations in LLMs. In this dissertation, the model occasionally classified an STT as if it understood the problem but then returned categories not part of the given STT taxonomy. In some cases, it even used information from previously classified STTs to justify the classification of the current STT. A recent study, [4], examined the issue of hallucinations in LLMs and found that while there are techniques to mitigate them, these hallucinations are unavoidable, and systems using LLMs must be prepared to deal with them. The loading phase was completed by integrating the content into the European STT Observatory using the API of its hosting platform.

An additional objective that emerged during the completion of this dissertation was the LLM-based quality assessment of literature review studies. Although not entirely reliable, our initial results demonstrated the feasibility of this approach. Once refined to produce sound results, this methodology can improve the efficiency of the literature review process, a critical step in any research endeavor.

6.2 Future work

Given the promising initial results for the LLM-based classification, there are several pathways for future work along the two contexts used in this dissertation: the article's quality and STT's labeling.

Regarding the classification of articles' quality, there are two specific improvement opportunities:

- Since the ground truth for LLM-based classification is the manual (human) assignment, a significant improvement would be to perform it with a larger group of participants, following predefined classification guidelines. Interrater agreement techniques [18] in the classification process would then minimize human subjectivity. This approach draws inspiration from the methodology used in [11], which similarly compared human-generated and LLM-generated labels. See Appendix 6.2 for the initial guidelines, which may be updated later.

- We should explore how the content of the primary studies is provided to the LLM. This can be done by providing structured information, as performed in this dissertation, for classifying an STT.

Regarding the specific context of STTs for the observatory, there are several pending features, as follows:

- Extract and classify content from the STTs' URLs found in catalogs. The main challenge here is that those pages do not have a standard layout as in the catalogs, so LLMs are also required for content extraction, similarly to what Elicit does for extracting content from primary studies, as described in chapter 4. If the content is then segmented such as in catalogs, then the technique proposed in this dissertation can be used for STT classification (labeling).

- Identify new STTs by web searches automatically. The main challenge here is finding reliable STT candidates. Analyzing the outputs of a search engine with an LLM for soundness checking seems to be a viable first step. Then, a validation step performed by a human actor would allow applying a reinforcement learning approach to fine-tune the selection process. Once a sound STT candidate is found, we will resort to the previous pending problem.

- Allow STT producers to independently update the information on their STTs made available in the observatory. This will require hosting the open-source version of Omeka in a dedicated server and implementing a new Omeka plugin for user access control, therefore extending Omeka's core functionality. By customizing the user roles and permission system, associating collections with specific users, and intercepting workflow processes, it is possible to create a flexible and secure access control system. That plugin will probably use design patterns like the Observer, Factory, or Strategy to enforce clean and maintainable access control logic.

For both contexts, other research threads may be explored.

- Experimenting with open LLMs such as the ones made available in the HuggingFace Hub, in dedicated hardware, to mitigate the effect of user load in limited access commercial LLMs, such as Microsoft Copilot used herein.
- Experimenting with fine-tuning an LLM using frameworks like PyTorch, TensorFlow, or Hugging Face's transformers library. These provide the tools to manage and customize training procedures, such as learning rate schedules, gradient accumulation, etc.
- Experimenting with the Retrieval Augmented Generation (RAG) concept. RAG involves retrieving relevant information from an external knowledge base before the LLM generates a response. This ensures that the LLM only accesses the necessary information not present in its training data, resulting in more accurate and contextually relevant answers.

Fine-tuning and RAG are computationally expensive, requiring powerful hardware such as high-memory GPUs or TPUs. The specific resources depend on the size of the LLM, the dataset, and the duration of the training. Fortunately, the first supervisor of this dissertation was recently awarded the required resources in the context of the call for Advanced Computing Projects (A1 Development Access) promoted by the Portuguese FCT. Those computing resources will be made available starting on October 1st, 2024, by the Portuguese Network for Advanced Computing (RNCA).

Bibliography

[1] A. Abdelali, H. Mubarak, S. A. Chowdhury, M. Hasanain, B. Mousi, S. Boughorbel, S. Abdaljalil, Y. E. Kheir, D. Izham, F. Dalvi, M. Hawasly, N. Nazar, Y. Elshahawy, A. Ali, N. Durrani, N. Milic-Frayling, and F. Alam, "LAraBench: Benchmarking Arabic AI with Large Language Models," in *Proc. of the 18th Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL'2024)*, G. Y., P. M., and P. M., Eds., vol. 1, ACL, 2024, pp. 487–520, ISBN: 979-889176088-2. DOI: 10.48550/arXiv.2305.14982.

[2] S. Bach, V. Sanh, Z. X. Yong, *et al.*, "PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts," in *Proc. of the 60th Annual Meeting of the Assoc. for Computational Linguistics: System Demonstrations*, V. Basile, Z. Kozareva, and S. Stajner, Eds., ACL, May 2022, pp. 93–104. DOI: 10.18653/v1/2022.acl-demo.9.

[3] S. V. Balkus and D. Yan, "Improving short text classification with augmented data using GPT-3," *Natural Language Engineering*, 2023. DOI: 10.1017/S1351324923000438.

[4] S. Banerjee, A. Agarwal, and S. Singla, "Llms will always hallucinate, and we need to live with this," *arXiv preprints*, 2024. DOI: 10.48550/arXiv.2409.05746.

[5] F. Borazio, D. Croce, G. Gambosi, R. Basili, D. Margiotta, A. Scaiella, M. Del Manso, D. Petrone, A. Cannone, A. M. Urdiales, C. Sacco, P. Pezzotti, F. Riccardo, D. Mipatrini, F. Ferraro, and S. Pilati, "Semi-Automatic Topic Discovery and Classification for Epidemic Intelligence via Large Language Models," in *Proc. of the 2nd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP@LREC-COLING)*, 2024, pp. 68–84. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195143331&partnerID=40&md5=fdbd936465c630ce6e4818b32757a00f.

[6] G. Cignoni and A. Bucci, "Cicognini at ACTI: Analysis of techniques for conspiracies individuation in Italian," in *CEUR Workshop Proc.*, vol. 3473, 2023. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85173564577&partnerID=40&md5=9284c028d4c4500a7d60667484346379.

[7] Clúster de Empresas Innovadoras para el Turismo de la Comunitat Valenciana, *Nuestras Soluciones Tecnológicas*, Accessed in Dec 2023, 2023. [Online]. Available: https://adestic.org/uploads/catalogo_soluciones_turisticas.pdf.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics*, ACL, 2020, pp. 8440–8451. DOI: 10.48550/arXiv.1911.02116.

[9] D. Cosme, A. Galvão, and F. Brito e Abreu, "A Systematic Literature Review on LLM-Based Information Retrieval: The Issue of Contents Classification," in *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR2024*, INSTICC, SciTePress, 2024, pp. 135–146, ISBN: 978-989-758-716-0. DOI: 10.5220/0013062300003838.

[10] ——, *Supplementary Data for 'A Systematic Literature Review on LLM-Based Information Retrieval: The Issue of Contents Classification'*, Sep. 2024. DOI: 10.5281/zenodo.13354076.

[11] I. N. Debess, A. Simonsen, and H. Einarsson, "Good or Bad News? Exploring GPT-4 for Sentiment Analysis for Faroese on a Public News Corpora," in *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, 2024, pp. 7814–7824. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195997116&partnerID=40&md5=17e8c604738f2655cb7494ee26843ae9.

[12] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in *Advances in Neural Information Processing Systems*, vol. 36, 2023. DOI: 10.48550/arXiv.2305.14314.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies conference (NAACL HLT)*, vol. 1, ACL, 2019, pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805.

[14] A. Doan, A. Halevy, and Z. Ives, *Principles of Data Integration*. San Francisco, CA: Elsevier, 2012, ISBN: 978-0124160446.

[15] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General language model pretraining with autoregressive blank infilling," *arXiv preprints*, 2021. DOI: 10.48550/arXiv.2103.10360.

[16] A. Galvão, F. Brito e Abreu, and J. Joanaz de Melo, "Towards a Consensual Definition for Smart Tourism and Smart Tourism Tools," in *Smart Life and Smart Life Engineering: Current State and Future Vision*, ser. Lecture Notes in Business Information Processing, Springer Nature, 2024, pp. 1–25.

[17] C. Gehweiler and O. Lobachev, "Classification of intent in moderating online discussions: An empirical evaluation," *Decision Analytics Journal*, vol. 10, 2024, ISSN: 27726622. DOI: 10.1016/j.dajour.2024.100418.

[18] N. Gisev, J. S. Bell, and T. F. Chen, "Interrater agreement and interrater reliability: key concepts, approaches, and applications," *Research in Social and Administrative Pharmacy*, vol. 9, no. 3, pp. 330–338, 2013.

[19] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas, "MarIA: Spanish Language Models," *Procesamiento del Lenguaje Natural*, pp. 39–60, 2022, ISSN: 1989-7553. DOI: 10.26342/2022-68-3.

[20] C. D. Hromei, D. Croce, V. Basile, and R. Basili, "ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme," in *CEUR Workshop Proc.*, vol. 3473, 2023. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85173568550&partnerID=40&md5=0d3882fff863f524a297de457d6abb20.

[21] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *ICLR 2022 - 10th Int. Conf. on Learning Representations*, 2022. DOI: 10.48550/arXiv.2106.09685.

[22] Y. Hu, F. Zou, J. Han, X. Sun, and Y. Wang, "LLM-TIKG: Threat intelligence knowledge graph construction utilizing large language model," *Computers and Security*, vol. 145, 2024. DOI: 10.1016/j.cose.2024.103999.

[23] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, *et al.*, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," *arXiv preprints*, arXiv:2211.05100, arXiv:2211.05100, Nov. 2022. DOI: 10.48550/arXiv.2211.05100.

[24] M. Leong, A. Abdelhalim, J. Ha, D. Patterson, G. L. Pincus, A. B. Harris, M. Eichler, and J. Zhao, "MetRoBERTa: Leveraging Traditional Customer Relationship Management Data to Develop a Transit-Topic-Aware Language Model," *Transportation Research Record*, 2024. DOI: 10.1177/03611981231225655.

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics*, ACL, 2020, pp. 7871–7880. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115443344&partnerID=40&md5=68efdf1732c50aade60d049337fb10e1.

[26] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *arXiv preprints*, 2019. DOI: 10.48550/arXiv.1910.13461.

[27] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," in *Proc. of the Eighth Int. Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*, 2017. DOI: 10.48550/arXiv.1710.03957. arXiv: 1710.03957 [cs.CL].

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*, vol. abs/1907.11692, 2019. DOI: 10.48550/arXiv.1907.11692.

[29] C. Luna-Jimenez, Z. Callejas, and D. Griol, "Mental-Health Topic Classification employing D-vectors of Large Language Models," English, in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 199–204, ISBN: 979-835038472-7. DOI: 10.1109/CBMS61543.2024.00041.

[30] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," *Journal of Big Data*, vol. 11, no. 1, 2024, ISSN: 21961115. DOI: 10.1186/s40537-024-00948-z.

[31] H. Mallek, F. Ghozzi, and F. Gargouri, "Systematic Literature Review," *Intelligent Systems Design and Applications: Industrial Applications, Volume 6*, vol. 6, p. 55, 2024.

[32] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2008, p. 482, ISBN: 978-0-521-86571-5.

[33] A. Mastropaolo, S. Scalabrino, N. Cooper, D. Nader Palacio, D. Poshyvanyk, R. Oliveto, and G. Bavota, "Studying the usage of text-to-text transfer transformer to support code-related tasks," in *Proc. of the Int. Conf. on Software Engineering (ICSE)*, 2021, pp. 336–347. DOI: 10.1109/ICSE43902.2021.00041.

[34] V. Minghetti and D. Buhalis, "Digital Divide in Tourism," *Journal of Travel Research*, vol. 49, no. 3, pp. 267–281, 2010. DOI: 10.1177/0047287509346843.

[35] A. H. Nasution and A. Onan, "ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks," *IEEE Access*, vol. 12, pp. 71 876–71 900, 2024. DOI: 10.1109/ACCESS.2024.3402809.

[36] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. DOI: 10.2753/MIS0742-1222240302.

[37] A. Peña, A. Morales, J. Fierrez, I. Serna, J. Ortega-Garcia, I. Puente, J. Córdova, and G. Córdova, "Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14193 LNCS, pp. 20–33, 2023. DOI: 10.1007/978-3-031-41498-5_2.

[38] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, *Improving language understanding by generative pre-training*, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[39] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. DOI: 10.48550/arXiv.1910.10683.

[41] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. of the Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. on Natural Language Processing conference (EMNLP-IJCNLP)*, 2019, pp. 3982–3992. DOI: 10.48550/arXiv.1908.10084.

[42] F. G. Reverte and P. D. Luque, "Digital Divide in E-Tourism," in *Handbook of e-Tourism*. Springer, 2020, pp. 1–21, ISBN: 978-3-030-05324-6. DOI: 10.1007/978-3-030-05324-6_109-1.

[43] M. Rodríguez-Cantelar, M. Estecha-Garitagoitia, L. F. D'Haro, F. Matía, and R. Córdoba, "Automatic Detection of Inconsistencies and Hierarchical Topic Classification for Open-Domain Chatbots," *Applied Sciences (Switzerland)*, vol. 13, no. 16, 2023. DOI: 10.3390/app13169055.

[44] G. Russo, N. Stoehr, and M. H. Ribeiro, "ACTI at EVALITA 2023: Automatic Conspiracy Theory Identification Task Overview," in *CEUR Workshop Proc.*, L. M., M. S., P. M., R. V., S. R., and V. G., Eds., vol. 3473, CEUR-WS, 2023. DOI: 10.48550/arXiv.2307.06954.

[45] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprints*, 2019. DOI: 10.48550/arXiv.1910.01108.

[46] D. Schabus, M. Skowron, and M. Trapp, "One Million Posts: A Data Set of German Online Discussions," in *Proc. of the 40th Int. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, ACM, 2017, pp. 1241–1244, ISBN: 9781450350228. DOI: 10.1145/3077136.3080711.

[47] Scholz & Friends Agenda Berlin GmbH, *Leading examples of Smart Tourism Practices in Europe*, Accessed in Dec 2023, Feb. 2023. [Online]. Available: https://smart-tourism-capital.ec.europa.eu/leading-examples-smart-tourism-practices-europe_en.

[48] M. Sebők, Á. Máté, O. Ring, V. Kovács, and R. Lehoczki, "Leveraging Open Large Language Models for Multilingual Policy Topic Classification: The Babel Machine Approach," *Social Science Computer Review*, 2024. DOI: 10.1177/08944393241259434.

[49] SEGITTUR, *Catalogue of Technological Solutions for Smart Tourism Destinations*, Accessed in Dec 2023, Jan. 2023. [Online]. Available: https://www.destinosinteligentes.es/en/formacion/catalogue-of-technological-solutions-for-smart-tourist-destinations-third-edition/.

[50] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. of the Empirical Methods in Natural Language Processing conference (EMNLP)*, 2013, pp. 1631–1642. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84926358845&partnerID=40&md5=aee25e7557c51d87ca49204c286b2813.

[51] S. Stahlschmidt and D. Stephen, "Comparison of Web of Science, Scopus and Dimensions databases," KB forschungspoolprojekt, DZHW Hannover, Germany, Tech. Rep., 2020, pp. 1–37.

[52] Y. Tatarnikova and E. Parsadanyan, *13 Data Integration Tools: Comparative Analysis*, Accessed in Dec 2023, Oct. 2023. [Online]. Available: https://blog.n8n.io/data-integration-tools/.

[53] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, *LLaMA: Open and Efficient Foundation Language Models*, 2023. DOI: 10.48550/arXiv.2302.13971. arXiv: 2302.13971 [cs.CL].

[54] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprints*, 2023. DOI: 10.48550/arXiv.2307.09288.

[55] A. C. Tricco, J. Antony, W. Zarin, L. Strifler, M. Ghassemi, J. Ivory, L. Perrier, B. Hutton, D. Moher, and S. E. Straus, "A scoping review of rapid review methods," *BMC Medicine*, 2015. DOI: 10.1186/s12916-015-0465-6.

[56] P. Trust and R. Minghim, "Query-Focused Submodular Demonstration Selection for In-Context Learning in Large Language Models," in *Proc. of the 31st Irish Conf. on Artificial Intelligence and Cognitive Science (AICS)*, 2023. DOI: 10.1109/AICS60730.2023.10470628.

[57] J. Van Nooten, A. Kosar, G. De Pauw, and W. Daelemans, "Advancing CSR Theme and Topic Classification: LLMs and Training Enhancement Insights," in *Proc. of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services and the 4th Economics and Natural Language Processing (FinNLP-KDF-ECONLP@LREC-COLING)*, 2024, pp. 292–305. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195191165&partnerID=40&md5=994484ba62f44f8f5c0352f56a983dbb.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 2017-December, 2017, pp. 5999–6009. DOI: 10.48550/arXiv.1706.03762.

[59] A. Williams, N. Nangia, and S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in *Proc. of the 2018 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds., ACL, Jun. 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101.

[60] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models," in *Proceedings of the ACM Web Conference 2024*, Association for Computing Machinery, 2024, pp. 4489–4500. DOI: 10.1145/3589334.3648137.

[61] L. Yang, Y. Huang, C. Tan, and S. Wang, "News Topic Classification Base on Fine-Tuning of ChatGLM3-6B using NEFTune and LORA," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2024, pp. 521–525. DOI: 10.1145/3675249.3675339.

[62] C. Ye and X. Shi, "Optimizing News Topic Classification with Instructional Fine-Tuning of Chatglm3," in *ACM International Conference Proceeding Series*, 2024, pp. 573–577. DOI: 10.1145/3672758.3672851.

[63] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration," *Healthcare (Switzerland)*, vol. 11, no. 20, 2023, ISSN: 22279032. DOI: 10.3390/healthcare11202776.

[64] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, *et al.*, "GLM-130b: An open bilingual pre-trained model," *arXiv preprints*, 2022. DOI: 10.48550/arXiv.2210.02414.

[65] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, "OPT: Open Pre-trained Transformer Language Models," *arXiv preprints*, pp. 1–19, 2022. DOI: 10.48550/arXiv.2205.01068.

[66] X. Zhang, J. Zhao, and Y. Lecun, "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems*, vol. 2015-January, 2015, pp. 649–657. DOI: 10.48550/arXiv.1509.01626.

[This page has been intentionally left blank]

Appendix 1

By detailing the study, its questions, and the output format, we led Microsoft Copilot in creating the classification guidelines for future use, described below (Quality Assessment Future Guidelines).

Quality Assessment Future Guidelines

Q1: Were the study's goals and research questions clearly defined?
Scale and its rationale:

- **0 (Strongly Disagree):** No clear goals or research questions are stated.
- **1 (Disagree):** Goals or research questions are vaguely mentioned.
- **2 (Neither Agree nor Disagree):** Goals and research questions are somewhat clear but lack detail.
- **3 (Agree):** Goals and research questions are clearly defined but could be more specific.
- **4 (Strongly Agree):** Goals and research questions are explicitly and clearly defined.

Guidelines for classification:

1. **Abstract and Introduction:** Look for a summary of the study's goals and research questions.
2. **Objectives Section:** Check if a dedicated section outlines the study's objectives.
3. **Research Questions:** Ensure the research questions are specific and directly related to the study's goals.

Q2: Was previously published related work exposed and compared with the research results claimed in the study?
Scale and its rationale:

- **0 (Strongly Disagree):** No related work is mentioned.
- **1 (Disagree):** Related work is mentioned but not compared with the study's results.
- **2 (Neither Agree nor Disagree):** Some related work is mentioned and partially compared.
- **3 (Agree):** Related work is adequately exposed and compared with the study's results.
- **4 (Strongly Agree):** A comprehensive comparison with related work is provided.

Guidelines for classification:

1. **Literature Review Section:** Look for a detailed review of related work.
2. **Discussion Section:** Check if the study's results are compared with previous findings.
3. **References:** Ensure the study cites relevant and recent related work.

Q3: Was the research design clearly outlined?
Scale and its rationale:

- **0 (Strongly Disagree):** No clear research design is provided.
- **1 (Disagree):** Research design is mentioned but lacks detail.
- **2 (Neither Agree nor Disagree):** Research design is somewhat clear but incomplete.

- **3 (Agree):** Research design is clearly outlined but could be more detailed.
- **4 (Strongly Agree):** Research design is explicitly and comprehensively outlined.

Guidelines for classification:

1. **Methods Section:** Look for a detailed description of the research design.
2. **Study Protocol:** Check if the study protocol is clearly explained.
3. **Diagrams and Flowcharts:** Ensure using visual aids to explain the research design.

Q4: Were the study limitations evaluated and identified?
Scale and its rationale:

- **0 (Strongly Disagree):** No limitations are mentioned.
- **1 (Disagree):** Limitations are mentioned but not evaluated.
- **2 (Neither Agree nor Disagree):** Some limitations are identified but not thoroughly evaluated.
- **3 (Agree):** Limitations are adequately identified and evaluated.
- **4 (Strongly Agree):** Comprehensive evaluation and identification of limitations are provided.

Guidelines for classification:

1. **Discussion Section:** Look for a dedicated subsection on study limitations.
2. **Critical Analysis:** Check if the limitations are critically analyzed.
3. **Future Work:** Ensure the study suggests ways to address the limitations in future research.

Q5: Was the data used for validation described in sufficient detail and made available?
Scale and its rationale:

- **0 (Strongly Disagree):** No data description or availability
- **1 (Disagree):** Data is mentioned but not described in detail or made available.
- **2 (Neither Agree nor Disagree):** Data is somewhat described but not fully available.
- **3 (Agree):** Data is adequately described and partially available.
- **4 (Strongly Agree):** Data is comprehensively described and fully available.

Guidelines for classification:

1. **Results Section:** Look for detailed descriptions of the data used.
2. **Supplementary Materials:** Check if additional data is provided in supplementary materials.
3. **Data Repositories:** Ensure the data is available in public repositories or upon request.

Q6: Were answers to the research questions provided?
Scale and its rationale:

- **0 (Strongly Disagree):** No answers to the research questions are provided.
- **1 (Disagree):** Answers are vaguely mentioned.
- **2 (Neither Agree nor Disagree):** Some answers are provided but lack detail.

- **3 (Agree):** Answers to the research questions are adequately provided.
- **4 (Strongly Agree):** Comprehensive and clear answers to all research questions are provided.

Guidelines for classification:

1. **Conclusion Section:** Look for a summary of the answers to the research questions.
2. **Discussion Section:** Check if the answers are discussed in detail.
3. **Results Section:** Ensure the results directly address the research questions.

Q7: Were negative or unexpected findings reported about the study?
Scale and its rationale:

- **0 (Strongly Disagree):** No negative or unexpected findings are reported.
- **1 (Disagree):** Negative or unexpected findings are mentioned but not detailed.
- **2 (Neither Agree nor Disagree):** Some negative or unexpected findings are reported but lack detail.
- **3 (Agree):** Negative or unexpected findings are adequately reported.
- **4 (Strongly Agree):** Comprehensive reporting of negative or unexpected findings.

Guidelines for classification:

1. **Results Section:** Look for any mention of negative or unexpected findings.
2. **Discussion Section:** Check if these findings are discussed and analyzed.
3. **Conclusion Section:** Ensure the study acknowledges and addresses negative or unexpected findings.

[This page has been intentionally left blank]

iscte UNIVERSITY INSTITUTE OF LISBON

Smart ETL and LLM-based contents classification: the European Smart Tourism Tools Observatory experience

Diogo Cosme