

A Systematic Literature Review on LLM-Based Information Retrieval: The Issue of Contents Classification

Diogo Cosme¹^a, António Galvão²^b and Fernando Brito E Abreu¹^c

¹ISTAR-IUL, Instituto Universitário de Lisboa (Iscte-IUL), Av. das Forças Armadas, 40, 1649-026 Lisboa, Portugal

²CENSE, School of Science and Technology, NOVA University Lisbon, 2829-516 Caparica, Portugal

Keywords: Systematic Literature Review, Large Language Model, Information Retrieval, Contents Classification.

Abstract: This paper conducts a systematic literature review on applying Large Language Models (LLMs) in information retrieval, specifically focusing on content classification. The review explores how LLMs, particularly those based on transformer architectures, have addressed long-standing challenges in text classification by leveraging their advanced context understanding and generative capabilities. Despite the rapid advancements, the review identifies gaps in current research, such as the need for improved transparency, reduced computational costs, and the handling of model hallucinations. The paper concludes with recommendations for future research directions to optimize the use of LLMs in content classification, ensuring their effective deployment across various domains.

1 MOTIVATION

Generative AI (GenAI), particularly LLMs, which were designed for Natural Language Processing (NLP) tasks, has changed the paradigm of Information Retrieval (IR). An interesting list of IR topics and themes based on LLMs is presented in (Liu et al., 2024). Notably, automatic content classification has improved thanks to LLMs. Before their rise, achieving accurate and efficient content classification, mainly of textual content, was challenging. LLMs have successfully overcome these limitations.

Besides being trained on vast amounts of data, most LLMs follow the transformer architecture (Vaswani et al., 2017). 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 tasks, text generation, and context understanding. The key concepts of the transformer models are:


- **Model Architecture:** It can be encoder-only, designed to understand the meaning and context of each word in relation to others, making it suitable for classifying texts, answering questions, and other


comprehension-based applications. It can also be decoder-only and used to generate a new sequence of words, making it suitable for various generative tasks such as text generation, language modeling, and conversational agents. Lastly, combining both is also possible, resulting in encoder-decoder models. The foundational models¹ that stand out in each architecture are, respectively: BERT (Devlin et al., 2019), GPT (Radford et al., 2018), and BART (Lewis et al., 2020).


- **Adapting a LLM:** There are two main ways to specialize a LLM for specific tasks. One method is fine-tuning the model, which consists of adjusting the model's weights based on the new data. The larger the model, the greater the computing resources required. A more resource-efficient alternative, though potentially less effective, is In-Context Learning (ICL). It involves giving the model examples of the task during inference² without additional training, allowing it to learn from these examples. It can receive zero examples (zero-shot), i.e., the hy-

¹A 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.

²Inference in the context of LLMs refers to generating a response or prediction based on a given input.

^a <https://orcid.org/0009-0001-1245-286X>

^b <https://orcid.org/0000-0002-6566-9114>

^c <https://orcid.org/0000-0002-9086-4122>

pothesis that the model is already capable is tested, or it can receive some examples (few-shot).

Due to the immense potential and inherent complexities of LLMs, it is essential to evaluate or conduct literature reviews to support the field of 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 corresponded to literature reviews (Mahadevkar et al., 2024; Yu et al., 2023). However, none of these were about LLM-based content classification. On (Yu et al., 2023), a literature review addressed the critical need for guidelines for incorporating LLMs and GenAI into healthcare and medical practice. In contrast, a systematic literature review on (Mahadevkar et al., 2024) 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 write this paper. It is organized as follows: Section 2 describes the review methodology used to identify and conduct the study; Section 3 analyzes the studies obtained; and Section 4 provides a summary of the existing research and identifies the threats to this literature review.

2 METHODOLOGICAL APPROACH

A systematic literature review (SLR), in contrast to an unstructured review approach, reduces bias by following a strict and methodical sequence of stages for conducting literature searches (Wohlin, 2014; Kitchenham and Brereton, 2013). The ability of an SLR to methodically search, extract, analyze, and document findings in stages depends on carefully designed and evaluated review protocols. The technique for these efforts is described in this section.

2.1 Planning the Review

2.1.1 Research Questions

The following research questions were formulated:

- **RQ1:** What type of empirical studies have been conducted in LLM-based content classification?
- **RQ2:** How extensive is the research in this area?
- **RQ3:** What were the relevant contributions of the existing studies?
- **RQ4:** Can LLMs be used to assess the quality of studies?

2.1.2 Review Protocol

Based on the research conducted by (Stahlschmidt and Stephen, 2020), Scopus offers more extensive subject coverage than Web of Science and Dimensions, encompassing the majority of articles found in these two databases. As a result, we chose to use the Scopus database exclusively for our formal literature search.

2.1.3 Search String

Keywords were derived from the research questions and used to search the primary study source. The search string included the most important terms related to the research questions, including synonyms, related terms, and alternative spellings.

To carry out the intended research, the following search string was drawn up:
("Large Language Model" OR "Foundational Model") AND ("Contents Classification" OR "Topic Classification")

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.

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". This indicates that it is a robust and trustworthy AI solution for summarizing, finding, and extracting details from scientific articles.

Elicit allows us to extract several details from scientific articles, but we have only selected these: research question; summary of introduction; dataset; limitations; research gaps; software used; algorithms; methodology; main findings; Study Objectives; study design; intervention effects; hypotheses tested; experimental techniques.

All the information extracted with *Elicit* is available online here (Cosme et al., 2024).

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 scientific research involves using LLMs, we decided to carry out a performance comparison test to evaluate the quality of articles between a manual assessment and an LLM-based one.

The information extracted from *Elicit* was then used as a basis for the manual and LLM-based quality assessment. For the LLM-based evaluation, we carried it out using 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.

The prompt template used, which is outlined below (**Prompt 1**), is organized in the following manner: 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. 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.

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.

Microsoft Copilot was the LLM used. For Q2, 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** (see the corresponding box).

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?"

2.2 Conducting the Review

2.2.1 Execute Search

Applying the specified search string resulted in the retrieval of nineteen scientific articles. Seven studies were rejected, and twelve articles were accepted.

One of the accepted studies, (Russo et al., 2023), is an overview 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. So, articles of that challenge relevant to the research topic that did not appear in the search string results and fulfill the inclusion criteria have been added. This resulted in a total of thirteen accepted articles.

2.2.2 Apply Quality Assessment

Figure 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 method is also helpful in understanding the performance (Figure 2). Both graphs show that Q7 has the most significant disparity, with the highest mean absolute score difference between the two methods and the largest gap between the mean scores ($|2.77 - 1.08| = 1.69$). 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, with $|3.08 - 3.00| = 0.08$, 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 ab-

solute 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.

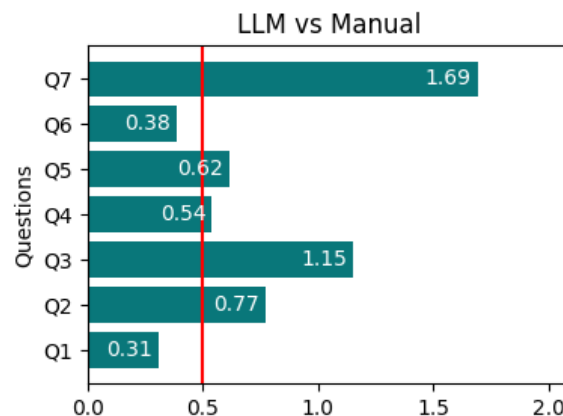


Figure 1: Mean Absolute Score Difference Between Methods Per Question.

The data obtained in the comparison between manual (M) and LLM (L) analysis is available online here (Cosme et al., 2024).

Mean Score by Question and Method

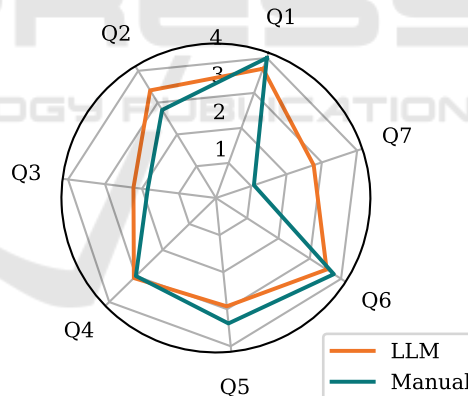


Figure 2: Radar Chart Displaying the Average Scores Given to the Studies by M and L.

Although the results indicate that using ICL zero-shot is not yet reliable, we conclude that assessing 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.

Due to the few studies, this task did not remove any studies and was only useful for assessing their overall quality.

3 DOCUMENT THE REVIEW

3.1 Demographics

Figure 3 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 5), 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 one article involving cooperation between teams from Indonesia and Turkey. Regarding authors' affiliation countries, while no single country dominates, Europe emerges as the most active continent (Figure 4).

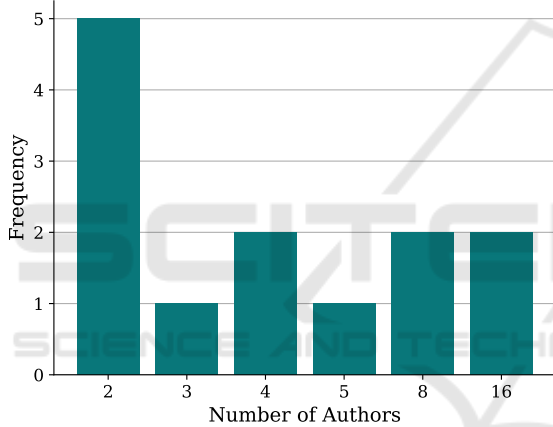


Figure 3: Publication Frequency by Authors Count.

Figure 6 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.

3.2 Analysis and Findings

A methodology was proposed in (Rodríguez-Cantelar et al., 2023) 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 (Li et al., 2017)

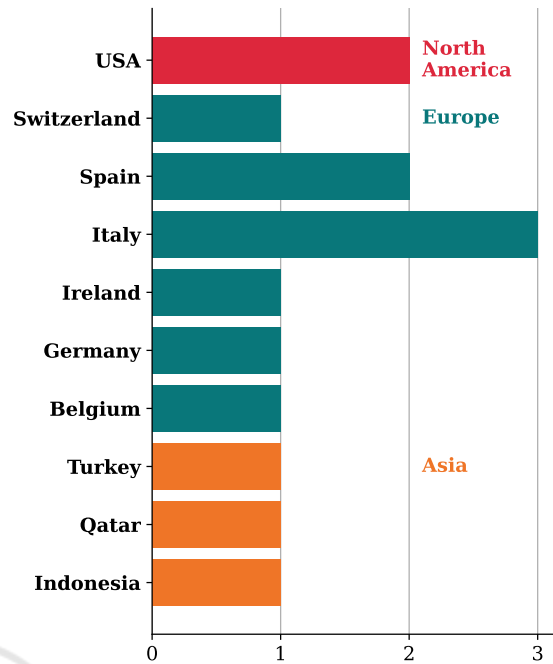


Figure 4: Publication Frequency by Author Affiliations' Country.

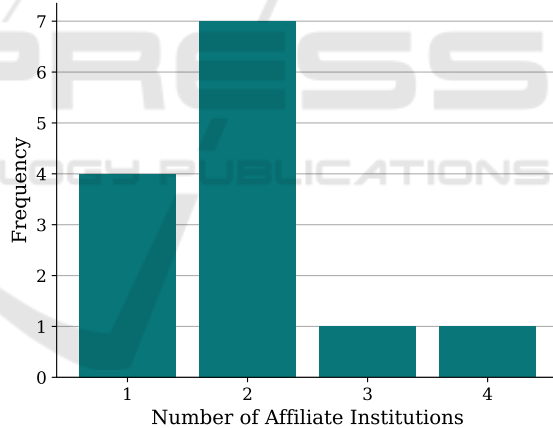


Figure 5: Publication Frequency Affiliates Count.

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

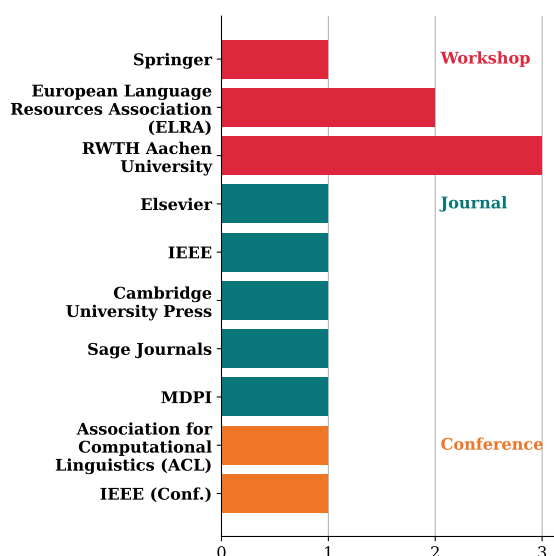


Figure 6: Publication Frequency by Publisher.

(ACTI)" is presented in (Russo et al., 2023). 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 (Cignoni and Bucci, 2023) compared the performance between two fine-tuned encoder-only transformer models (bert-base-italian-xxl-cased and XLM-RoBERTa (Conneau

et al., 2020)) and a non fine-tuned decoder-only transformer model (LLaMA 7B (Touvron et al., 2023)). 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.

- In (Hromei et al., 2023), the authors 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)³(Hu et al., 2022), 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.

For Subtask A, the approach in (Cignoni and Bucci, 2023) achieved the sixth rank, while the one in (Hromei et al., 2023) secured the second position. For Subtask B, their rankings were fourth and fifth. The winning team in both subtasks employed an approach that leveraged data augmentation through LLMs.

In (Trust and Minghim, 2023), query-focused sub-modular mutual information functions are proposed to select diverse and representative demonstration examples for ICL in prompting. In addition, an interactive tool is presented to explore the impact of

³LoRA fine-tuning significantly reduces the computational and storage costs of training large language models by only adjusting a subset of low-rank parameters.

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) (Socher et al., 2013), and a topic classification task with the AG News Classification Dataset (Zhang et al., 2015). 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 (Reimers and Gurevych, 2019) 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 (Radford et al., 2019), OPT (Zhang et al., 2023), and BLOOM (Le Scao et al., 2022).

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 (Leong et al., 2024). 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 (Liu et al., 2019). 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.

The paper (Borazio et al., 2024) introduces 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, and to provide explicit support to analysts through the generation of 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 (Lewis et al., 2020) and previously trained on the Multi-Genre Natural Language Inference corpus (Williams et al., 2018). 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 (Abdelali et al., 2024) 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 Ope-

nAI Studio Chat playground and PromptSource (Bach et al., 2022), 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.

In (Peña et al., 2023), the potential of LLMs to enhance the classification of public affairs documents is studied. 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 (Gutiérrez-Fandiño et al., 2022): RoBERTa-base, RoBERTa-large, RoBERTa-Talex, 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 (Balkus and Yan, 2023). 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 per-

formance, 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 (Nasution and Onan, 2024) 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 (Devlin et al., 2019), BERTurk (a fine-tuned Turkish version of BERT), RoBERTa (Liu et al., 2019) (fine-tuned on specific datasets), and T5 (Mastropaolo et al., 2021). 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 (Gehweiler and Lobachev, 2024). 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 (Schabus et al., 2017). 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

Table 1: Articles summary information.

Article	Method	Evaluation Metrics	Description
(Rodríguez-Cantelar et al., 2023)	ICL	Weighted F1	Topic: 0.34; Subtopic: 0.78 (DailyDialog)
		Accuracy	Topic: 81%; Subtopic: 62% (SGC5)
(Cignoni and Bucci, 2023)	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.
(Hromei et al., 2023)	Fine-tuning	F1	Subtask A: 0.82 (extremIT5); 0.86 (extremITLLaMA).
			Subtask B: 0.81 (extremIT5); 0.86 (extremITLLaMA)
(Trust and Minghim, 2023)	ICL	F1	Sentiment Classification: 88.35%.
			Topic Classification: 90.56%.
(Leong et al., 2024)	Fine-tuning	Macro-avg F1	0.80 compared to the best ML model with 0.76
		Weighted F1	0.90 compared to the best ML model with 0.88
(Borazio et al., 2024)	ICL	Accuracy	Single Class: 67%; Top two system suggestions: 90.56%.
(Abdelali et al., 2024)	ICL	Macro-avg F1	Few-shot (10-shot): 0.721; Zero-shot: 0.656.
(Peña et al., 2023)	Fine-tuning	Accuracy	Accuracies higher than 85%.
(Balkus and Yan, 2023)	ICL	Accuracy	LLM: 76%; Estimated Human: 85%.
(Nasution and Onan, 2024)	Fine-tuning; ICL	Avg F1	Human: 0.883; GPT-4: 0.865.
(Gehweiler and Lobachev, 2024)	Fine-tuning	F1	Identifying user intent: 0.755.
(Van Nooten et al., 2024)	Fine-tuning; ICL	F1 score	Zero-shot experiments lag behind fine-tuned models.

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 (Van Nooten et al., 2024) 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 (Sanh et al., 2019), BERT (Devlin et al., 2019), RoBERTa, and RoBERTa-large (Liu et al., 2019). For the themes dataset, the authors used fine-tuned versions of Multi-Lingual DistilBERT, XLM-RoBERTa, and XLM-RoBERTa-large (Conneau et al., 2020). Their zero-shot experiments with GPT models show they still lag behind fine-tuned models in multi-label

classification.

Table 1 shows the training methods used, the evaluation metrics, and the main results of this evaluation.

4 CONCLUSIONS

4.1 Recap of Research Questions

RQ1: What Type of Empirical Studies Have Been Conducted in LLM-Based Content Classification?

Although the number of studies is limited, 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 research areas explored:

- Hierarchical topic/subtopic detection in inconsistent chatbot responses (Rodríguez-Cantelar et al., 2023)
- Socio-political phenomena during health crises (Borazio et al., 2024);

- Public affairs documents (Peña et al., 2023);
- Customer feedback (Leong et al., 2024);
- Corporate Social Responsibility themes and topics (Van Nooten et al., 2024);
- Conspiracy Content (Cignoni and Bucci, 2023; Hromei et al., 2023)
- Sentiment (Trust and Minghim, 2023; Nasution and Onan, 2024)
- Emotion (Nasution and Onan, 2024)
- Benchmarking of NLP and speech processing tasks (Arabic) (Abdelali et al., 2024)
- Short questions (Balkus and Yan, 2023)
- User intent in online discussions (Gehweiler and Lobachev, 2024)
- Comparison of generated annotations (Nasution and Onan, 2024)

RQ2: How Extensive Is the Research in this Area?

Although there are currently only a few approaches to topic/content classification using LLMs, this field is emerging. We believe it will grow and improve significantly in the future.

RQ3: What Were the Relevant Contributions of the Existing Studies?

Based on the available studies, fine-tuned LLMs outperform LLMs prompted with ICL techniques (Balkus and Yan, 2023; Van Nooten et al., 2024). 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. 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 (Abdelali et al., 2024), revealing that few-shot outperformed zero-shot. Notably, the selection of few-shot examples plays a crucial role (Trust and Minghim, 2023), and there are limitations related to the reasoning abilities of LLMs. Researchers (Abdelali et al., 2024; Borazio et al., 2024) reported challenges arising from model hallucinations.

RQ4: Can LLMs Be Used to Assess the 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 either through more extensive

research with a fine-tuned model or by using ICL few-shot examples.

4.2 Threats to Validity

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

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.

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.

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.

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.

4.3 Future Work

The use of LLMs in information retrieval is promising, as shown by recent studies and their years of publication. Future research should optimize LLMs for different domains, focusing on domain-specific fine-

tuning and possibly hybrid models to maintain broad knowledge while adapting to specialized domains.

Improving the interpretability of LLM-based classifiers is critical because they often operate as black boxes, limiting trust in sensitive areas such as healthcare and finance. Creating explainability frameworks within LLM architectures can increase transparency and trust by clarifying classification decisions.

Ethical considerations are also critical. Research should focus on mitigating biases in LLM training data and outputs to ensure fair content classification.

Efficiency, scalability, and dynamic adaptation of LLMs are growing challenges. Future studies should improve computational efficiency through model compression or streamlined architectures, and explore continuous or reinforcement learning to help keep LLMs up to date with evolving content such as social media and news.

Lastly, enhancing cross-domain transfer learning can improve LLM adaptability across different applications. By refining these techniques, LLMs could become more versatile and excel at content classification across various industries.

ACKNOWLEDGEMENTS

This work was partially funded by the Portuguese Foundation for Science and Technology (FCT), under ISTAR-Iscte project UIDB/04466/2020 and CENSE NOVA-FCT project UIDB/04085/2020.

REFERENCES

- Abdelali, A., Mubarak, H., Chowdhury, S. A., Hasanain, M., Mousi, B., Boughorbel, S., Abdaljalil, S., Kheir, Y. E., Izham, D., Dalvi, F., Hawasly, M., Nazar, N., Elshahawy, Y., Ali, A., Durrani, N., Milic-Frayling, N., and Alam, F. (2024). LARA-Bench: Benchmarking Arabic AI with Large Language Models. In Y., G., M., P., and M., P., editors, *Proc. of the 18th EACL Conf.*, volume 1, pages 487–520. ACL.
- Bach, S., Sanh, V., Yong, Z. X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Fevry, T., Alyafeai, Z., Dey, M., Santilli, A., Sun, Z., Bendavid, S., Xu, C., Chhablani, G., Wang, H., Fries, J., Al-shaibani, M., Sharma, S., Thakker, U., Almubarak, K., Tang, X., Radev, D., Jiang, M. T.-j., and Rush, A. (2022). PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In Basile, V., Kozareva, Z., and Stajner, S., editors, *Proc. of the 60th Annual Meeting of the ACL: System Demonstrations*, pages 93–104. ACL.
- Balkus, S. V. and Yan, D. (2023). Improving short text classification with augmented data using GPT-3. *Natural Language Engineering*.
- Borazio, F., Croce, D., Gambosi, G., Basili, R., Margiotta, D., Scaielli, A., Del Manso, M., Petrone, D., Cannone, A., Urdiales, A. M., Sacco, C., Pezzotti, P., Riccardo, F., Mipatrini, D., Ferraro, F., and Pilati, S. (2024). Semi-Automatic Topic Discovery and Classification for Epidemic Intelligence via Large Language Models. In *Proc. of PoliticalNLP@LREC-COLING Workshop*, pages 68–84.
- Cignoni, G. and Bucci, A. (2023). Cicognini at ACTI: Analysis of techniques for conspiracies individuation in Italian. In *CEUR Workshop Proceedings*, volume 3473.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics*, pages 8440–8451. ACL.
- Cosme, D., Galvão, A., and Brito e Abreu, F. (2024). Supplementary Data for "A Systematic Literature Review on LLM-Based Information Retrieval: The Issue of Contents Classification". *Zenodo*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository (CoRR)*.
- Gehweiler, C. and Lobachev, O. (2024). Classification of intent in moderating online discussions: An empirical evaluation. *Decision Analytics Journal*, 10.
- Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Armentano-Oller, C., Rodríguez-Penagos, C., González-Agirre, A., and Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, page 39–60.
- Hromei, C. D., Croce, D., Basile, V., and Basili, R. (2023). ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme. In *CEUR Workshop Proceedings*, volume 3473.
- Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. of ICLR Conf.*
- Kitchenham, B. and Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(12):2049–2075.
- Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *Computing Research Repository (CoRR)*.
- Leong, M., Abdelhalim, A., Ha, J., Patterson, D., Pincus, G. L., Harris, A. B., Eichler, M., and Zhao, J. (2024). MetRoBERTa: Leveraging Traditional Customer Relationship Management Data to Develop a Transit-Topic-Aware Language Model. *Transportation Research Record*.

- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). 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*, pages 7871–7880. ACL.
- Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *Computing Research Repository (CoRR)*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository (CoRR)*, abs/1907.11692.
- Liu, Z., Zhou, Y., Zhu, Y., Lian, J., Li, C., Dou, Z., Lian, D., and Nie, J.-Y. (2024). Information Retrieval Meets Large Language Models. In *Proc. of the ACM Web Conf. (WWW Companion)*, pages 1586–1589.
- Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., and Choudhury, T. (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, 11(1).
- Mastroianni, A., Scalabrino, S., Cooper, N., Nader Palacio, D., Poshvanyk, D., Oliveto, R., and Bavota, G. (2021). Studying the usage of text-to-text transfer transformer to support code-related tasks. In *Proc. of ICSE Conf.*, pages 336–347.
- Nasution, A. H. and Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*, 12:71876–71900.
- Peña, A., Morales, A., Fierrez, J., Serna, I., Ortega-García, J., Puente, I., Córdova, J., and Córdova, G. (2023). Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs. *Lecture Notes in Computer Science*, 14193 LNCS:20–33.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Computing Research Repository (CoRR)*.
- Rodríguez-Cantelar, M., Estecha-Garitaigotia, M., D’Haro, L. F., Matía, F., and Córdoba, R. (2023). Automatic Detection of Inconsistencies and Hierarchical Topic Classification for Open-Domain Chatbots. *Applied Sciences (Switzerland)*, 13(16).
- Russo, G., Stoehr, N., and Ribeiro, M. H. (2023). ACTI at EVALITA 2023: Automatic Conspiracy Theory Identification Task Overview. In *CEUR Workshop Proc.*, volume 3473. CEUR-WS.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Computing Research Repository (CoRR)*.
- Schabus, D., Skowron, M., and Trapp, M. (2017). One Million Posts: A Data Set of German Online Discussions. In *Proc. of the 40th SIGIR Conf.*, page 1241–1244. ACM.
- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*, pages 1631–1642.
- Stahlschmidt, S. and Stephen, D. (2020). Comparison of Web of Science, Scopus and Dimensions databases. Technical report, KB forschungspoolprojekt, DZHW Hannover, Germany.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *Computing Research Repository (CoRR)*.
- Trust, P. and Minghim, R. (2023). Query-Focused Submodule Demonstration Selection for In-Context Learning in Large Language Models. In *Proc. of the 31st Irish AICS Conf.*
- Van Nooten, J., Kosar, A., De Pauw, G., and Daelmans, W. (2024). Advancing CSR Theme and Topic Classification: LLMs and Training Enhancement Insights. In *Proc. of FinNLP-KDF-ECONLP@LREC-COLING*, pages 292–305.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-December, pages 5999–6009.
- Williams, A., Nangia, N., and Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Walker, M., Ji, H., and Stent, A., editors, *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies*, volume 1, pages 1112–1122. ACL.
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proc. of the 18th EASE Conf.* ACM.
- Yu, P., Xu, H., Hu, X., and Deng, C. (2023). Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare (Switzerland)*, 11(20).
- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2023). Opt: Open pre-trained transformer language models, 2022. *Computing Research Repository (CoRR)*, 3:19–0.
- Zhang, X., Zhao, J., and Lecun, Y. (2015). Character-level convolutional networks for text classification. *Computing Research Repository (CoRR)*.