Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Impact Classification within and beyond Academia: Domain-Robust Annotation and the Capacity of Large Language Models

Maria Becker

maria.becker@gs.uni-heidelberg.de

Leibniz Institute for the German Language

Kanyao Han

University of Illinois Urbana-Champaign

Rezvaneh Rezapour

Drexel University

Jana Diesner

Technical University of Munich

Andreas Witt

Leibniz Institute for the German Language

Research Article

Keywords: Impact classification, Sentence-level annotation, Social impact, Project report, Cross-domain, Large language model

Posted Date: January 23rd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5543205/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Impact Classification within and beyond Academia: Domain-Robust Annotation and the Capacity of Large Language Models

Maria Becker[1,3*†], Kanyao Han[2†], Rezvaneh Rezapour[4], Jana Diesner[2,5], Andreas Witt[1]

[1]Department of Digital Linguistics, Leibniz Institute for the German Language, Mannheim, Germany.
[2]School of Information Sciences, University of Illinois Urbana-Champaign, USA.
[3]Department of German Linguistics, Heidelberg University, Germany.
[4]Department of Information Science, Drexel University, Philadelphia, USA.
[5]School of Social Science and Technology, Technical University of Munich, Germany.

*Corresponding author(s). E-mail(s): maria.becker@gs.uni-heidelberg.de;
Contributing authors: kanyaoh2@illinois.edu; shadi.rezapour@drexel.edu; jana.diesner@tum.de; witt@ids-mannheim.de;
†These authors contributed equally to this work.

Abstract

Prior analyses and assessments of the impact of scientific research has mainly relied on analyzing its scope within academia and its influence within scholarly circles. However, by not considering the broader societal, economic, and policy implications of research projects, these studies overlook the ways in which scientific discoveries contribute to technological innovation, public health improvements, environmental sustainability, and other areas of real-world application. We expand upon this prior work by developing and validating a conceptual and computational solution to automatically identify and categorize the impact of scientific research within and especially beyond academia based on text data. We first empirically develop and evaluate an annotation schema to capture and classify the impact of research projects based on research reports from different

scientific domains. We then annotate a large dataset of more than 45k sentences extracted from research reports for the developed impact categories. We examine the annotated dataset for patterns in the distribution of impact categories across different scientific domains, co-occurrences of impact categories, and signal words of impact. Using the annotated texts and the novel classification schema, we investigate the performance of large language models (LLMs) for automated impact classification. Our results show that fine-tuning the models on our annotated datasets statistically significantly outperforms zero- and few-shot prompting approaches. This indicates that state-of-the-art LLMs without fine-tuning may not work well for novel classification schemas such as our impact classification schema, and in turn highlights the importance of diligent manual annotations as empirical basis in the field of computational social science.

Keywords: Impact classification, Sentence-level annotation, Social impact, Project report, Cross-domain, Large language model

1 Introduction

The outcomes of scientific research can have a multitude of impacts on our society, such as influencing people's values, knowledge, and opinions [1–5]. Research results enable and advance new industries, technologies, and other products and services that impact essential needs including communication, education, transportation, healthcare, and culture. Scientific research also provides evidence-based information relevant to policy makers and society at large [6, 7].

The object of previous analyses and assessments of the impact of scientific research was mainly academia. A large body of work has examined research publications and their dissemination using metrics and methods from bibliometric, infometric, or scientometric, such as calculating citation counts and measuring trends in the allocation of awards and grants to scholars [8–10].

However, these methods are limited in capturing impact beyond academia, such as societal, political, legal, and economic impact [11, 12]. More specifically, the societal impact of research may not be reflected by the number of citations, but rather by the tangible outcomes that some research has produced, such as policies and practices that concern the wider community.

To fill this gap, Witt et al. (2018) evaluated the impact of research projects beyond academia. Based on the analysis of scientific project reports, they empirically developed an alternative classification schema for categorizing impact, which distinguishes between monetary and non-monetary impact and includes the subcategories economic, technical, socio-cultural, political-legal, environmental impact and income for research institutions. Building on prior research, Rezapour et al. (2020) proposed a mixed-method framework that combines top-down and bottom-up strategies for extracting impact categories. This approach leverages domain-specific knowledge alongside impact categories derived from project reports to facilitate computational analysis of scientific impact. After manually annotating a corpus of funding reports, followed by employing supervised machine learning to infer impact categories, the

results show that research funding reports often aim to achieve impacts that extend beyond the academic domain, encompassing economic, socio-cultural, and political outcomes.

Building on these observations and insights, this paper aims to develop and evaluate a classification schema designed to effectively capture and categorize the impact of scientific research based on research reports. There are two main challenges in analyzing impact based on project reports. First, there is still a lack of a comprehensive classification schema to categorize the impact of scientific research. The PESTEL schema, a framework used by organizations and businesses to analyze external macro-environmental factors, has the potential to serve as an impact classification schema for scientific research, as it is comprised of a seven societally-relevant categories, i.e., **P**olitical, **E**conomic, **S**ocial, **T**echnological, **E**nvironmental, and **L**egal. However, while these categories are insightful for our purpose, their subcategories are more related to macro-environmental factors in business rather than science, such as "Free Trade Disputes" under the main category "Political". Given the lack of suitable category systems for our task, we needed to develop our own schema. Second, the absence of a standardized structure in project reports across programs and domains hinders the extraction, coding, and analysis of impact-relevant information from texts from different domains. As project reports can be lengthy, most of the content is unrelated to impact indicators, and impact descriptions are scattered throughout the texts, efficient solutions are needed to handle this genre.

In this paper, we propose and evaluate a schema for impact classification of scientific research. The resulting schema includes main- and subcategories for capturing the impact of research projects on academia, the economy, society, culture, politics, law, technology, and/ or the environment. Moreover, we show how our schema is applicable to different scientific domains, including the natural science, engineering, and the humanities. To tackle this challenge, we annotated a large dataset of over 45k sentences extracted from research reports based on the classification schema we developed. These research reports, which include a variety of research tasks, methodologies, and findings, are an important resource for measuring the the impact of scientific research using text data as a source. However, they have rarely been used or shared outside of academia. We further conduct an in-depth examination of our annotated dataset and report the findings and insights, including label distributions, correlations, keyword and example analysis, in order to answer the following research questions: What are the most frequent impact categories in general and for each domain? Which domain includes the largest amount of impact relevant sentences? And which words and language patterns (per domain and in general) indicate impact? Our investigations leads to a deeper understanding of domain-dependent and domain-independent patterns and features of scientific impact. Finally, using the annotated texts and the novel classification schema, we investigate and compare the performance of different large language models (i.e., model fine-tuning and few-shot prompting) for automated impact classification.

In summary, this paper makes the following contributions: (i) Developing an impact classification schema adequate for capturing and categorizing the academic and non-academic impact of scientific research from different domains. (ii) Providing a large

dataset (>45k sentences extracted from scientific reports across domains)[1] that has been annotated with main- and sub-categories of impact based on our classification schema. This dataset can be used, for example, to identify and analyze cross-domain, text-based evidence of the impact of scientific research within or beyond academia, and training machine learning models for automatic impact detection and classification. (iii) Presenting an in-depth, corpus-statistical, and linguistic analysis of text-based evidence of impact and discuss how our insights can inform and facilitate future work on the automatic analysis of impact using machine learning techniques. (IV) Experimenting with state-of-the-art large language models and natural language processing techniques to illustrate the opportunities and challenges for automated impact classification.

The remainder of this paper is structured as follows: Section 2 gives an overview of work related to (scientific) impact assessment and potential applications of LLMs for impact classification. Section 3 describes our dataset, followed by the presentation of our annotation schema (Section 4) and a description of the annotation process (Section 5). In section 6, we present the findings from our corpus-statistical and linguistic analysis of our annotated dataset. In section 7, we present the results of our experiments for automated impact classification using LLMs. Section 8 summarizes our results and discusses limitations and future work, especially related to using our annotations for automatic impact analysis.

2 Related Work

2.1 Impact Assessment

The private and public sectors have made significant investments to find answers to the following questions: how can we measure the influence of research funding not only on academia, but also on people and society, and what is that impact [15]. This paper focuses on the first question. Understanding and capturing this broader impact requires a nuanced approach that combines both qualitative and quantitative methods [4, 16]. This necessity also stems from the diverse nature of potential outcomes of funded research, which can vary from policy changes to shifts in public attitudes and direct impacts on health and wellness. As a result, there has been a strong focus on evaluating the societal impact of research, which is becoming increasingly crucial for stakeholders including funding agencies and governmental organizations who seek to justify the societal benefits of their investments [17–20].

To assess the impact of research projects in academia, different approaches have been proposed and widely used. Science of science, for instance, uses data-driven methods to study how scientific research is conducted, examining factors like research problem selection, collaboration patterns, and career development in science [21]. Bibliometric and altmetric approaches, on the other hand, offer robust, quantifiable measures of research influence [22]. Solutions from the field of bibliometrics focus on the statistical analysis of research publications and pertinent meta-data, using metrics such as citation counts to infer the significance of papers, authors, institutions, etc.

[1] Our annotated dataset and annotation manual are available here: https://github.com/khan1792/texttransfer.

[23–25]. This approach has been complemented by the altmetrics movement, which explores the broader digital footprint of research by capturing, for example, attention to research via online venues such as social media mentions, blog discussions, and policy document citations [26]. Karanatsiou et al. [27] discussed the evolution from bibliometrics to altmetrics, highlighting each approach's strengths and weaknesses, including susceptibility to manipulation, and emphasized the importance of a more holistic approach that combines both for a more accurate measurement of research impact. Erdt et al. [28] highlighted the value of altmetrics by analyzing 172 articles, demonstrating its multidimensional approach to evaluating research impact and emphasizing the importance of integrating both citation-based and alternative measures in an increasingly digital world. Despite their advancements, bibliometrics and altmetrics still face challenges in fully capturing the multifaceted impacts of research on society, policy, and public discourse [29].

The academic community is increasingly focused on the broader impact of research and grants beyond academia, with many scholars advocating for evaluations that go beyond citation-based methods, recognizing that such metrics may not fully capture the real-world influence of scholarly work [11, 12, 30–32] Penfield et al. [33] argues for a more nuanced approach that integrates both short-term and long-term societal outcomes, addressing the challenges of quantifying impacts that extend beyond immediate academic outputs. New methodologies are being developed to bridge these gaps, including frameworks that map the pathways from research activities to societal impacts. For example, the Kellogg Foundation's Logic Model [34] and the Payback Framework [35, 36] model impact as a journey involving stages such as research activities and long-term societal changes. These models emphasize outputs, outcomes, and broader impacts, suggesting a more comprehensive approach to impact evaluation [20, 37].

To build on prior research and offer a more comprehensive, long-term, and multidimensional view of impact, researchers have also proposed incorporating self-reported data from funding reports to better assess the societal impacts of research [13, 14, 38]. These studies have focused on initiatives funded by agencies, such as the UK and US National Science Foundation and Germany's research councils, aiming to directly connect research funding with societal outcomes, including technology adoption, health advancements, and economic growth [39]. Langfeldt and Scordato [40] examined how research funding agencies in Norway and other countries increasingly prioritize aligning research with societal impacts. They highlighted mechanisms like performance-based funding, which allocates resources based on measurable outcomes like citations or patents, and targeted programs that direct funding to specific areas with high potential for societal benefits, such as health, technology, or economic growth. While these approaches aim to ensure research directly contributes to societal goals, the authors also noted challenges in implementing them, including the risk of bureaucratic processes and balancing academic excellence with societal impact. Furthermore, Witt et al. [13] introduced a new classification system to assess the broader implications of scientific research, including its societal, political, legal, and economic effects. Their system is designed to extend the evaluation of research impact beyond the academic sphere to influence on finance, technology, and the environment using funding reports

as the input data. Similarly, Rezapour et al. [14] examined a set of project reports and used supervised machine learning methods to identify impacts such as economic, technical, socio-cultural, and political from the text. Bornmann [41] explored various frameworks and methodologies, including public value mapping [42], to evaluate how the societal impact of research can be effectively measured. The author also highlighted the inherent challenges of measuring societal impact, pointing out difficulties similar to those raised by Langfeldt and Scordato [40], including the tension between maintaining academic rigor and achieving societal relevance. While these studies have significantly advanced our understanding of the societal impacts of research funding, there remains a gap in evaluating the long-term and multi-dimensional impacts across different disciplines.

Our research builds upon previous work in this field. While most prior work has concentrated on assessing and extracting various types of scientific impact from a single research domain, our study broadens the scope by introducing a cross-domain impact framework that transcends the limitations of previous methodologies.

2.2 Automated Text Classification

Classifying texts to enhance the information available for analysis is a common approach in academic research and impact assessment [4, 14]. Over the past decade, automated classification models have become increasingly popular across various fields [43, 44] to reduce the labor and time required for categorizing large volumes of text data.

In recent years, the transformer architecture has become the most popular choice for developing text classification models, largely due to the transformers' superior performance over traditional machine learning or deep learning methods like support vector machines, gradient boosting, and recurrent neural networks [45]. Researchers have trained many large-scale transformer-based models using vast amounts of data, with the most notable models being the Bidirectional Encoder Representations from Transformers (BERT) [46] and Generative Pre-trained Transformer (GPT) families [47].

Significant efforts have also been made to adapt these general pre-trained models to scientific texts, involving fine-tuning existing pre-trained language models on domain- or task-specific data to leverage both the pre-existing knowledge in the model and the new knowledge from the specialized dataset [48–50]. For instance, SciBERT [49], a language model tailored for scientific text through fine-tuning BERT on a large corpus of scientific publications, has become one of the most widely used models for scientific document categorization [51, 52]. Additionally, benchmark datasets such as SDPRA 2021 [53] and arXiv [54] have been developed or adapted to aid in model training, fine-tuning, and performance evaluation.

Despite the success of transformer-based models in scientific text classification, challenges remain: Pre-trained models and benchmark datasets are often concentrated on a narrow range of domains, primarily Computer and Biomedical Sciences, leaving other fields, particularly the humanities and social sciences, under-resourced [55]. As a result, directly applying pre-trained models may lead to significantly lower classification performance in these areas. Fine-tuning on a substantial amount of

domain-specific data becomes essential when addressing classification tasks in these low-resourced domains [55, 56].

With the advent of LLMs like ChatGPT since 2023, zero-shot and few-shot learning have gained significant popularity among researchers [57, 58]. LLMs have been shown to outperform traditional models, mainly for two reasons. First, they are trained on vastly larger datasets and therefore have a significantly greater number of parameters. Second, they can process instructions and examples in natural language, guiding the model to perform specific tasks more effectively. However, recent studies have shown that these LLMs still disproportionately rely on high-resource fields, as knowledge and texts from these fields appear more frequently in training data. This limits their performance in low-resource fields [59], even though they generally achieve higher performance compared to previous pre-trained models [60].

In the field of impact classification, previous studies have used traditional machine learning methods, such as random forests and support vector machines, to identify impact categories from scientific project reports [14]. However, the potential of LLMs for impact classification remains largely unexplored. We therefore conduct a series of experiments to evaluate the impact of fine-tuning, as well as zero-shot and few-shot learning, on the prediction accuracy of state-of-the-art models like BERT, Llama2, ChatGPT, and GPT-4 on our annotated data and classification schema.

3 Data

3.1 Corpus Creation

Upon completion of a research project or grant, the results are usually summarized into a digitally archived project report (typically ranging from 50 to 150 pages). Many funding agencies require such a report. In addition to research tasks, methodologies, findings, and conclusions, these reports may also explain the research's necessity and potential applications. These reports build the empirical basis of this work. We consider final research reports of publicly funded research projects from the fields of artificial intelligence, mobility, linguistics, and musicology, accessible through the Leibniz Information Centre for Science and Technology (TIB)[2]. All reports are in German. The selection of these four domains allows an assessment of potential impact across different disciplines; from the natural sciences to engineering and the humanities. We collected a corpus of reports for this study based on the following criteria: 1) completed between 2005 and 2015, 2) publicly funded, 3) with two to ten collaborators, and 4) with at least one academic project partner. Each project can consist of more than one report, such as a final joint report and individual reports from different project partners. To avoid duplication and redundancy, we only consider final joint reports for annotation. Our corpus consists of 1,160 German research reports, with 497 reports from AI, 515 from mobility, 89 from linguistics, and 59 from musicology.

[2]

3.2 Data Preprocessing

Originally in PDF format, the texts were first converted into text files using a Python package *textract*. All non-textual data such as pictures or complex mathematical typesetting were removed. Subsequently, the texts were divided into passages using special tokens such as line breaks and section headings that typically indicate transitions between paragraphs.

To facilitate the analysis of lengthy reports and text labeling, we used the pipeline provided by [61] for identifying those sections in project reports that indicate potential impact. The framework first removes passages irrelevant for impact analysis, such as tables of contents (ToCs), bibliographies, headings, captions, and lists, using heuristic such as captions and the length of passages. A model that had been trained for for impact-relevant passage detection in a supervised manner is then used to automatically select those passages that indicate potential impact. This model relies on text passages from research reports that had been manually labeled in a binary way (impact-relevant vs. not impact-relevant), and these data were then used to train a random forest model to automatically predict if a given passage is impact-relevant or not. Further details on this methodology can be found in Becker et al. (2024).

Finally, the passages classified as potentially impact-relevant were split into sentences. Each sentence was manually annotated with impact categories as described next.

4 Bottom-Up Development of the Annotation Schema

To develop an annotation schema suitable for capturing and categorizing types of impact from project reports across and beyond domains, we first analyzed the structure and linguistic patterns of mentions, representations, and indicators of impact in our corpus. In these reports, impact-indicating content can be embedded in the description of methods and processes employed in a project or can be explicitly articulated by authors in the sections on outcome or an impact statement. We developed an annotation schema empirically, i.e., in a data-driven, bottom-up way: We initially provided human annotators, who were students with a background in (computational) linguistics, with guidelines that introduced the goals of our annotation project and impact definition (see Section 1). At this coding stage, we did not give any pre-defined categories to the annotators. Instead, we provided all annotators with a few identical passages from 12 reports (3 reports from each domain) and asked them to closely read the passages, mark up sentences that indicate any type of impact, and suggest labels for the types of impact they identified; allowing for main and sub-categories. We then systematically compared and discussed the annotations with our team members and consolidated the suggested categories. One emerging focal point of our discussion was whether each impact category was applicable across domains or if we needed different annotation schemes for different domains. Another focus lied on comparing our detected categories to existing impact schemes, in particular the PESTEL schema introduced in Section 1.

Impact Categories. Based on the proposed categories and comparison with the PESTEL schema, we converged on seven main categories with various subcategories of impact. These categories are applicable across domains. We iteratively tested and refined this schema in various data sessions, where project leaders and annotators jointly discussed initial annotations and cases of doubt. The resulting schema along with a short explanation of each category is shown in the following and examples for each domain in Table 1.

Societal Impact: Effects of a project on societal events, groups, or institutions)

- **Culture/Events** (Organization of (cultural) events, workshops, performances, etc.)
- **Education** (Education or training, e.g., new learning methods for schools)
- **Physical Health** (Alleviation/reduction of diseases, vaccination campaigns, etc.)
- **Life Quality** (Mental health, e.g., reduced depression, work-life balance)
- **Safety** (Traffic safety, hazard prevention, etc.
- **City Facility** (Public buildings, properties, parks, open spaces, vehicles, etc.)
- **Mobility** (freedom of movement and mobility)

Political & Legal Impact: Usage of project results in political/legislative contexts

- **Regulations** (Political regulations and deregulations)
- **Laws** (Development/contributions/amendments to laws/legal regulations)

Ethical Impact: Ethical effects of a project

- **Awareness** (Establishment/increase of awareness, perception, or attitudes)
- **Justice** (Establishment/improvement of justice and equality)
- **Data Policy** (Data privacy, Open Access)

Economic Impact: Utilization of project results for economic/financial developments

- **Business Models** (Development of business models or other economic strategies)
- **Income** (Generation/increase of income, new positions (outside university))
- **Employee Satisfaction** (Staff Contentment, service quality)
- **Optimizing Processes** (Enhancing processes in the economic/business sector)

Environmental Impact: Effects of a project on ecological or environmental aspects

- **Climate Protection** (Environmental/climate/conservation protection)
- **Sustainability** (Sustainability of products, methods, etc.)

Technical impact: Effects of technologies/models/data developed in the project

- **Protoytype** (Development of software prototypes)
- **Model Development** (Development of models/algorithms/methods)
- **IT Security** (Establishment/improvement of IT security)
- **Documentation** (Technical documentation)
- **Data Collection/Release** (collection, release or publication of datasets)

Other: Types of non-academic impact that do not fit into the other main categories

- **Product Development** (Development or improvement of a product)

- **PR/Visibility** (Promotion/other methods to increase the visibility of the project)
- **Knowledge Acquisition** (Obtaining knowledge through the project)
- **Knowledge Transfer** (Transfer of knowledge gained in the project to other areas)
- **Collaborations** (Establishment/improvement of collaborations beyond academia)

Academic Impact: Effects within the academic domain

- **Income Academia** (Income of research institutions, funding for team members)
- **Research Methods** (Development/application of new research methods)
- **Learning and Teaching** (New/improved learning/teaching methods in academia)
- **Publications** (Publication of research results e.g. in journals, books, conferences)
- **Academic Events** (Organization of academic events, e.g., workshops, conferences)
- **Collaborations** (Formation/improvement of collaborations/networks in academia)
- **Future Research** (Opening up new perspectives for future research projects)
- **Knowledge Acquisition** (Acquisition of academic knowledge through the project)

Our schema distinguishes between academic impact and impact beyond academia. The latter comprises the main categories of Societal Impact, Political and Legislative Impact, Ethical Impact, Economic Impact, Environmental/Ecological Impact, and Technical Impact. We also introduce the main category *Other*, which comprises subcategories beyond any of the main categories.

5 Top-Down Annotations of Impact Categories

Our annotators then labeled the text passages classified as impact relevant in the first step of our approach (cf. Section 3.2) according to the novel set of impact categories that we developed in a top-down approach. Considering the cost and time to comprehensively read and annotate lengthy reports, we opted for sentence-level annotation, while providing the annotators with the context of each sentence. Prior studies have proven that sentence-level annotation can reduce workload and tediousness. For example, a previous study in political science [62] demonstrated that employing non-expert annotators to annotate sentences from texts in the domain of politics could efficiently and cost-effectively yield accurately annotated information.

We build a codebook that lists and explains the main and subcategories of impact with examples and detailed instructions for the annotation process.[3]

For the annotation process, we used the annotation tool INCEpTION[4] which supports server-based, collaborative manual annotations of texts with customized annotation schemes (tagsets). Through this process, we created a custom tagset containing all the annotation categories defined in the codebook. The annotations are at the sentence level, and the annotators were instructed to consider each sentence independently, meaning that concrete indicators of impact must be present within the sentence itself. During annotation, the annotators were asked to select the most appropriate main and subcategory of impact per sentence. If more than one category were equally suitable, they could choose more than one label (both from the

[3] The codebook together with the annotated dataset can be found here: https://github.com/khan1792/texttransfer.

[4] https://inception-project.github.io/

Table 1 Examples for the impact categories identified from our corpus.

Sentence	Domain	Main Category	Sub Category
To conclude the workshop, the participant gave a concert at the Old Town Festival, where children could sign up for future courses.	Musicology	Societal	Culture/ Events
The insights gained from the analyses were presented to the Federal Ministry of Education and Research.	AI	Political/ Legal	Regulations
The developed battery lasts an average of 5 years longer than conventional batteries, contributing to reduced electronic waste in this way.	Mobility	Environmental/ Ecological	Sustainability
We demonstrate the impact of migration on the development of individual and social values as central aspects of youth identity.	Linguistics	Ethical	Justice
We can offer an effective and robust algorithm for extracting profession-specific information from English and German texts.	Linguistics	Technical	Model Development
Optimization of the module led to greater satisfaction among test users.	AI	Economic	Employee Satisfaction
One of the most interesting and important results is undoubtedly the collaboration with teachers from high schools, which differs from division-of-labor forms of cooperation and enables equal participation.	Musicology	Other	Collaborations
We propose a new, semi-automated research method for a contrastive comparison of German and English grammars.	Linguistics	Academic	Research Methods

main and subcategories). For sentences where annotators identified a main category but no matching subcategory, the subcategory *Other* must be selected for each main category.[5] If a sentence does not convey any impact-related content, the annotators had to assign the label *No Impact*. For each sentence to which an impact category was assigned, the annotators were asked to assess how significant the impact was; distinguishing between high, moderate, or low impact.

We trained a team of four student assistants with a background in linguistics in multiple sessions for the annotation process. None of them participated in the bottom-up coding process that was described in Section 4. These sessions included an overview of our project, the impact in academic texts, the codebook, and the annotation tool, training in how to do annotation, and discussions of initial annotations. We conducted the annotations in two phases:

The **first phase** primarily served for training the annotators, validating and fine-tuning the annotation manual, enhancing annotator agreement, and improving the quality and reliability of the annotations. In this phase, a selected subset of 3,000 sentences from all four domains was annotated independently by two annotators per sentence. We divided the data into three packages (with 1,000 sentences each) and discussed the annotations in several sessions after the completion of annotating each package. If necessary, we refined the guidelines by adding more concrete instructions

[5] E.g. for a sentence which belongs to the overarching category of *Societal Impact*, but not falling under any of the specified subcategories, the main category *Societal Impact* and the subcategory *Other* has to be labeled.

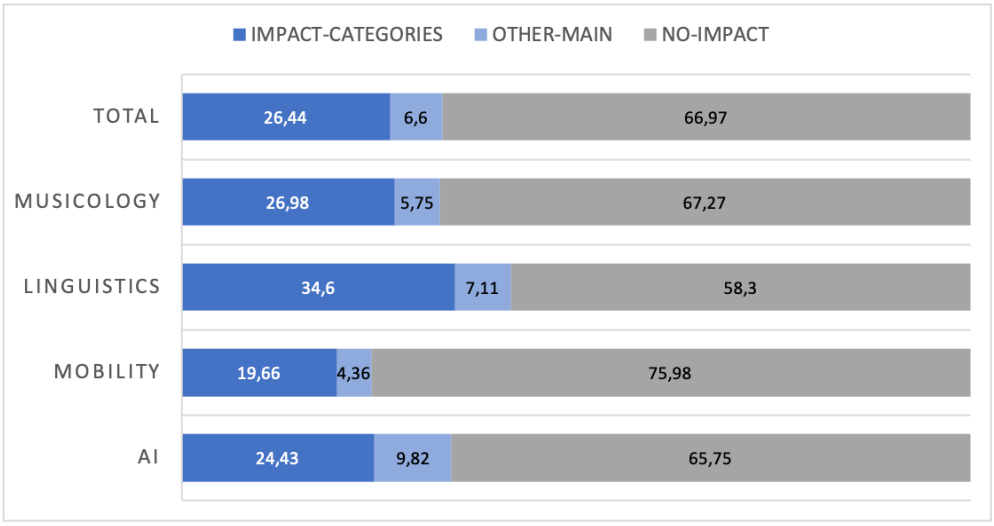


Fig. 1 Ratio of impact indicating sentences in our dataset, numbers in %.

or examples. We continuously measured the agreement of the annotations after completing each package and observed an increase in agreement among annotators for label assignments. The Cohen's Kappa score for the first text package was 34.68, which increased to 52.83 for the second package. For the last package, we finally observed a Kappa score of 74.48, which we interpret as a high level of agreement, given the complexity of the texts and our annotation schema. This suggests that our annotation schema adequately and distinctly captures and categorizes impact in project reports from different domains. It also indicates that the task is described and illustrated sufficiently in our codebook. Sentences with disagreement from this first annotation phase were finally adjudicated by an expert annotator (one of the authors).

After achieving a high level of agreement, we turned to the **second annotation phase**, where the remaining documents were annotated by a single annotator each. The combination of these two annotation stages allowed us to ensure the quality and reliability of the annotations, as well as an efficient way of manually coding huge amounts of texts with impact categories.

6 Data Statistics and Analysis

The outcome of our annotation process is a dataset tagged with different impact categories and subcategories, which builds the basis for the corpus-statistics and linguistic analysis of text-based evidence of impact, as presented in this section. Since our original dataset is very large and unbalanced in the number of research reports per domain (56,056 sentences in sum, where 54% of the sentences are from the domain Mobility, 37% percentage from AI, 5% from Linguistics and 4% from Musicology), we annotate a subset of 45,763 sentences, where 12,206 sentences were sampled from the domain mobility, 9,648 from AI, 11,126 from linguistics, and 12,783 from musicology.

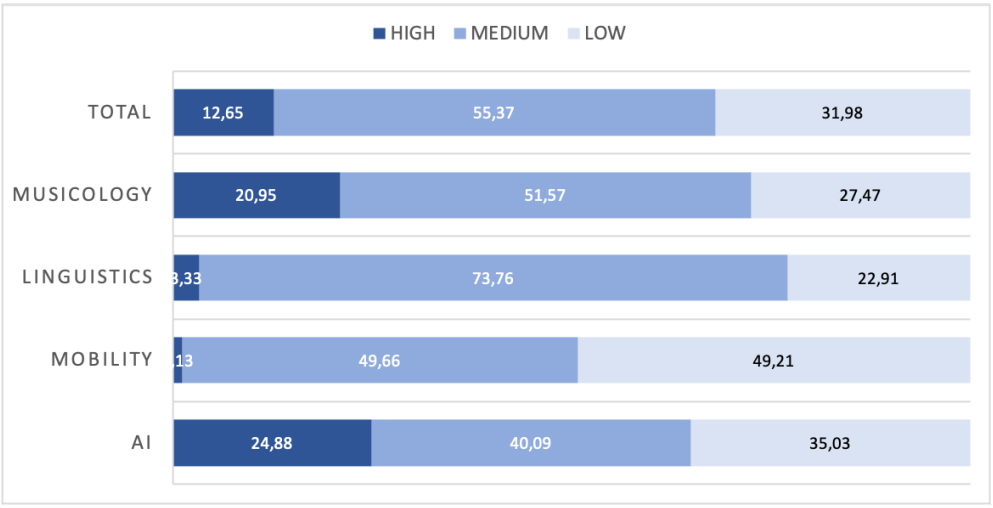


Fig. 2 Intensity of impact (for all sentences labeled as impact-relevant), in %

6.1 Ratio and Intensity of Impact Indicating Sentences

Figure 1 shows the **number of sentences** that were annotated as any one of the seven main impact categories, as *Other*, or as *No impact*. The Figure shows the distributions at the corpus and domain level, respectively. Overall, about a third (34.03%) of the sentences represent evidence of impact (*impact-categories* + *other-main*), while the majority of sentences (62.82%) were labeled as not being impact-relevant. [6] Linguistics is the domain with the highest amount of impact indicating evidence (measured by the number of sentences), while mobility and AI have the lowest amount of text-based evidence of impact.

Table 2 Intensity of impact: Examples from our dataset.

Sentence	Intensity	Domain
Specifically, the performance of the component has been significantly improved.	high	AI
The new structure facilitates an overview of the currently available technical documentation.	medium	Linguistics
The example of GAT makes it clear that trust and experience are guiding principles for the employees.	low	AI

Figure 2 shows the **intensity of the impact**. Overall, the majority of sentences were rated as indicating moderate impact (55%), 13% of the sentences were considered as expressing high impact, and 32% of the sentences as indicating low impact. Table

[6] This can be traced back to the fact that when choosing the model for extracting the passages for annotation (cf. Section 2.3), we chose a model with a wide scope of coverage (high recall) instead of a model with high precision. This is because we wanted to keep as many instances as possible for text-based evidence of impact in our dataset, and during manual coding, annotators can still filter out sentences that is not relevant to impact.

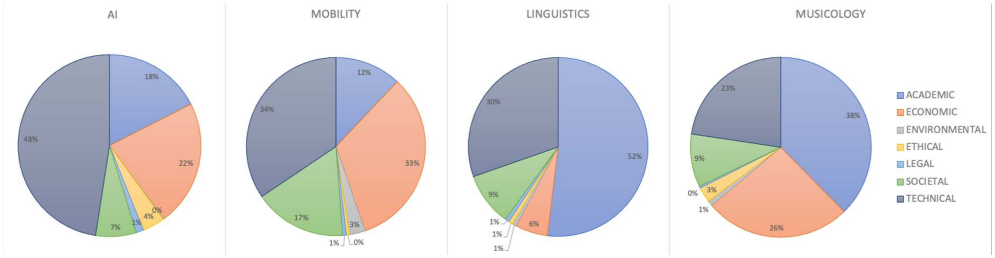


Fig. 3 Distribution of impact categories among domains, in %.

2 gives examples for these three categories. Note that all examples are originally in German and were translated into English by the authors. We find large differences between the domains: Out of all impact indicating sentences, sentences indicating strong impact have an above-average proportion in musicology (21%) and AI (25%) and account for 3% in linguistics and 1% in mobility.

6.2 Distribution and Co-Occurrences of Impact Categories

Figure 3 shows the **distribution** of the main categories (excluding the categories *Other* and *No Impact*) across the considered domains. We find the distributions of impact categories ti be domain-specific: in AI and mobility, a substantial number of sentences represent technical and economic impact, whereas the majority of sentences in the domains of linguistics and musicology mention academic impact. Ethical, legal, and environmental impact do not play an important role in any of these domains. Societal impact is most prevalent in mobility reports. Interestingly, economic impact is an important category in musicology while it occurs rather rarely in linguistics.[7]

As described in Section 5, our annotators could choose more than one impact category per sentence. While the majority of the sentences in our sample (11,584 sentences) were annotated with one impact category, out of all sentences labeled as impact indicating, 1,846 sentences (14%) are annotated with 2 labels, 175 sentences (1%) with 3 labels, and 14 sentences (0.1%) with 4 labels.

Figure 4 shows the **co-occurrences** of labels in our dataset, and Table 3 provides examples of sentences annotated with frequently co-occurring labels. Technical impact frequently co-occurs with academic as well as economic impact. This applies in particular to sentences from AI and mobility, where *Technical Impact* and *Economic Impact* are the most prevalent labels (cf. Figure 3). Closer examination of corresponding instances and respective subcategories of the combination of economic and academic impact (123 co-occurrences) and economic and societal impact (132 co-occurrences) showed that economic impact related to employee satisfaction often co-occurs with societal impact (second example in Table 3). When co-occurring with academic impact, economic impact often refers to optimizing processes (cf. example 3

[7] An example from our dataset for a sentence expressing economic impact in musicology is the following: *The result is internationally competitive production in order to open up new interest groups and sales markets.*

in Table 3); strengthening the assertion that optimizing processes is important in the the private sector and academia in Germany.

These results underline that even though our impact category schema is designed to be as exclusive and exhaustive as possible, there are still some cases where impact categories overlap. This, in turn, can be attributed to the broad spectrum and variability of the impact of scientific research.

6.3 Signal Words

Figure 5 gives an overview of the **most frequent words** for each main category. We lemmatized the texts and removed stopwords using *spaCy*. We find that all categories contain words that indicate impact in general (marked in yellow), such as *Ziel* (goal), *Ergebnis* (result), and *Beitrag* (contribution). In each domain, we also find words that can be interpreted as indicators for a specific category of impact (marked in green), such as *Wissenschaft* (science) for academic impact, *nachhaltig* (sustainable) for environmental impact, and *implementieren* (implement) for technical impact.

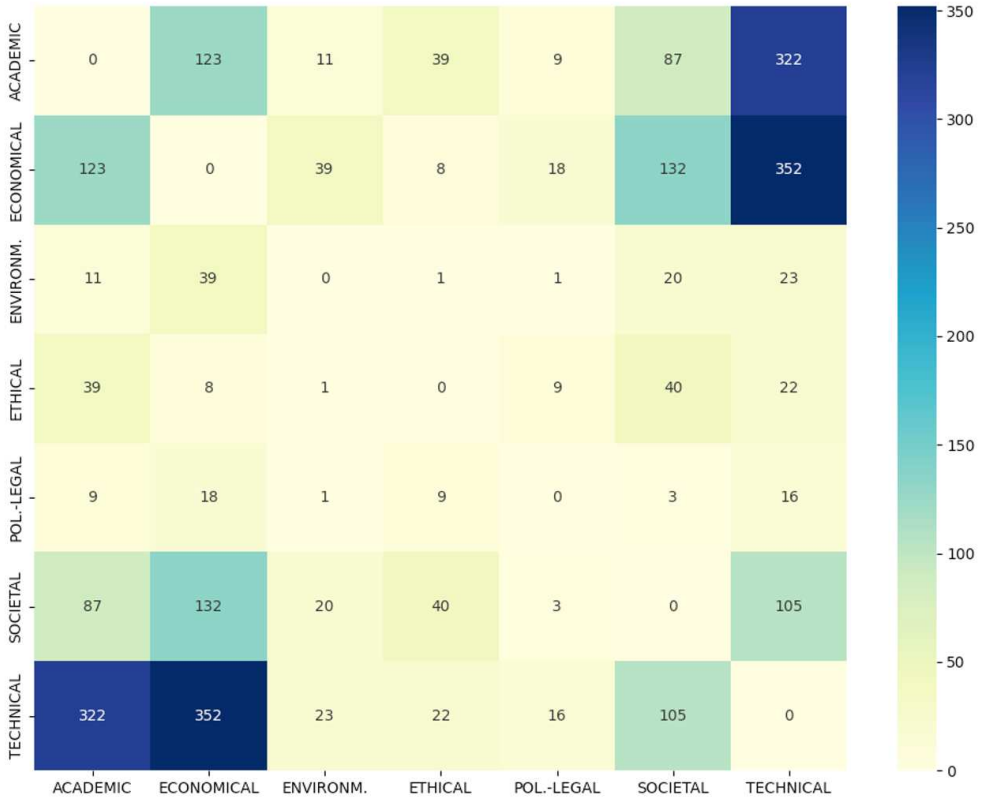


Fig. 4 Symmetrical heatmap of co-occurring impact labels, absolute numbers.

Table 3 Examples of co-occurring impact labels from our dataset.

Sentence	Main + Sub Category	Domain
The material selection and process control have been optimized to the extent that hollow cylinders with a density of approximately 95% and a length of approximately 20 mm can now be reliably and reproducibly manufactured.	ECONOMIC: Optimizing Processes; TECHNICAL: Prototype	AI
As part of the research project DigiLernPro, digital tools have been developed to assist users in creating learning scenarios and enabling intelligent playback of scenarios.	ECONOMIC: Employee Satisfaction; SOCIETAL: Education; TECHNICAL: Prototype	AI
Participation in the network supported by QPL has enabled the development of more specific and, consequently, more suitable methods for music conservatories.	ACADEMIC: Collaborations; ECONOMIC: Optimizing Processes	Musicology

academic		economic		environmental		ethic		legal		societal		technical	
projekt	583	möglich	217	nachhaltig	52	ziel	29	rechtlich	15	ziel	142	entwickeln	475
ergebnis	455	neu	212	neu	16	thema	24	ziel	13	projekt	108	projekt	319
weit	326	unternehmen	169	ziel	16	entwicklung	18	rahmenbedingung	9	möglich	89	entwicklung	291
rahmen	326	ziel	161	optisch	12	sozial	18	betrieb	9	schüler	87	ergebnis	271
können	325	projekt	155	konzept	11	möglich	16	möglich	8	können	79	rahmen	255
wissenschaft	287	wirtschaft	153	möglich	10	hochschule	16	technisch	8	entwicklung	78	neu	234
bereich	243	entwicklung	133	technisch	10	fh	16	rahmen	8	ergebnis	74	ziel	231
entwicklung	231	ergebnis	129	hersteller	9	umgang	14	entwicklung	8	rahmen	68	können	229
ziel	208	weit	117	rahmen	9	aspekt	13	neu	8	neu	66	weit	227
digital	207	können	116	betrieb	9	technisch	13	projekt	7	teilnehmen	64	möglich	207
entwickeln	206	entwickeln	113	nutzung	9	stark	12	daten	7	weit	63	daten	203
neu	199	rahmen	107	arbeit	9	massnahme	12	organisator	7	angebot	59	verfahren	181
möglich	196	nutzen	106	entwicklung	8	projekt	12	ergebnis	6	entwickeln	57	erstellen	172
hochschule	172	prozess	91	ergebnis	8	daten	12	verein	6	nutzen	53	basis	159
kooperation	170	region	88	vorhaben	7	apps	12	nachhaltig	6	öffentlich	52	technisch	151
universität	169	bereich	87	mobilität	7	studieren	12	grundlage	6	elektromobilit	51	gross	139
arbeit	162	geschäftsmodel	82	langfristig	7	können	11	jurist	6	bereich	46	implementieren	131
durchführen	161	gross	81	beitrag	7	weit	11	können	5	durchführen	44	analyse	131
verschieden	157	international	80	deutlich	7	datenschutz	11	ermöglichen	5	kind	40	automatisch	130
gemeinsam	157	kunde	79	reduzieren	6	nutzer	11	prüfen	5	zukunft	38	digital	126

Fig. 5 Most frequent words per impact category (absolute numbers).

Building on signal word analysis and as a first step towards **probing the feasibility of automatically classifying impact** in research reports, we then applied a topic model to our data. We used BERTopic [63], which leverages BERT embeddings to create dense clusters for a given document or several documents in order to find groups of semantically similar topics. Once clusters are formed, a class-based TF-IDF approach is applied to extract the most relevant keywords for each topic.

Figure 6 shows the selected terms for each detected topic. While some topics (e.g., topic 0 or 1) include mostly scientific vocabulary (e.g., *Projekt* (project)) or general impact indicating words (e.g., *Ergebnisse* (results)), we find topics that can be mapped to impact categories from our schema, e.g., topic 2 includes terms that hint at societal/cultural impact, while the terms included in topic 6 indicate economic impact.

These results show that even a simple, unsupervised topic modeling approach based on a large language model such as BERT and without any hyperparameter tuning or other adaptations is able to detect and distinguish certain impact categories of our schema from texts, which can be interpreted as support for our proposed classification schema.

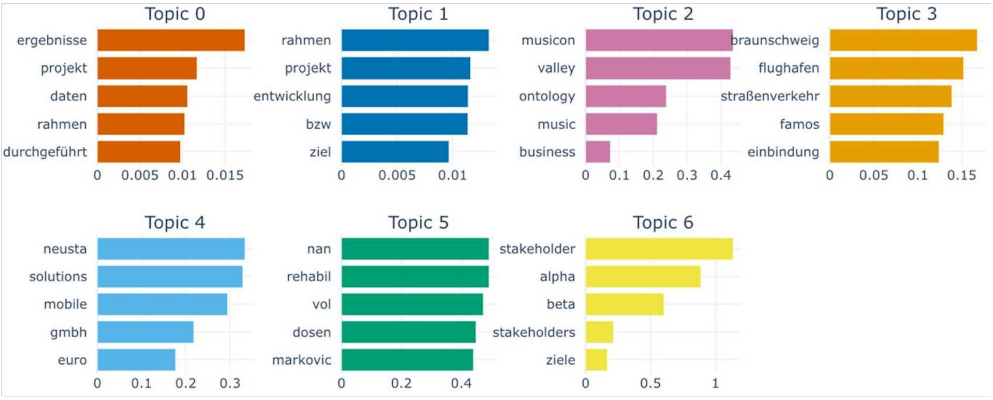


Fig. 6 Clusters from all sentences annotated as impact relevant from the project reports. The clusters together with the most relevant keywords for each are created with BERTopic (without using the annotations).

7 Automated Classification Experiments

We investigate the performance of large language models (LLMs) for the task of automated impact classification. This task can be described as a multiclass classification task on the sentence level. We used two widely adopted families of open-source models (BERT and Llama) as well as the most widely used commercial models (ChatGPT and GPT4). Specifically, German BERT [64] is a well-established tool for German classification tasks, and Llama2-chat models [65], ranging in scale from 7 billion to 70 billion parameters and serving as an open-sourced counterpart to commercial model ChatGPT and GPT4, are widely used state-of-the-art LLMs for various tasks and languages.[8]

7.1 Approaches and Models

To apply these models to our downstream classification task, we considered two main approaches: 1) fine-tuning with sentences and 2) few-shot or zero-shot learning. Traditionally, prior to the development of LLMs such as ChatGPT, GPT-4, and Llama2, fine-tuning was the preferred method for models like BERT when ample annotated data and computing resources were available because fine-tuning typically achieved significantly better performance than few-shot or zero-shot learning [66]. However, the advent of pretrained LLMs with billions of parameters, trained on extensive datasets and capable of handling various tasks and domains, has shifted the paradigm towards zero-shot and few-shot learning [66, 67]. Therefore, in our experiments, we used zero-shot and few-shot learning on Llama2, ChatGPT, and GPT4, and fine-tuning on both BERT and Llama2 models.

[8]The code for fine-tuning the models together with the prompts can be found here: https://github.com/khan1792/texttransfer.

- **Zero-shot and few-shot learning**: In recent studies and applications of LLMs, prompt-based learning, also known as zero-shot and few-shot prompting, has emerged as the most widely used technique, which directly models the probability of previous text to predict next words [67]. We also embrace this approach in our experiments. We conduct our experiments based on Llama2-chat-70b through an API provided by Replicate[9] as well as gpt-3.5-turbo and gpt-4 through an API provided by OpenAI. For zero-shot prompting, we provided the model with a task instruction, a classification schema, and a sentence to generate the prediction:

  (1) You will be provided with a text that possibly outlines the impact of a research project.
  (2) We define impact as an effect of scientific activities within academia or beyond the academic field, e.g. on the scientific field, economy, society, culture, politics, law, technology or the environment. In a report, impact may be represented by describing methods and routines implemented for a project, or by the impact that authors anticipated when writing their final reports. This means that estimated impact is also considered as impact. Impact could also be the maintenance or the avoidance of change.
  (3) Your task is to assign one impact category to the text from the following options:
  (4) NO-IMPACT: the sentence does not express any impact.
  TECHNICAL IMPACT: refers to technologies that are used outside of the original project, e.g. software prototype development, Improving IT security, or data release
  ECONOMIC IMPACT: refers to the use of research results for economic developments, e.g. development of business models, service quality, or economic strategies
  ACADEMIC IMPACT: refers to impact within academia - within or beyond the own field/institution, e.g.improved learning and teaching, new research methods, or publications
  SOCIETAL IMPACT: occurs when a project influences societal groups or institutions like schools, local authorities, foundations, or clubs, refugees/migration, religious persecution, etc.
  POLITICAL-LEGISLATIVE IMPACT: refers to using the project results in political or legislative contexts, e.g. contributions to laws or political regulations
  ETHICAL IMPACT: refers to ethical impact, e.g. equality, awareness, or charity
  ENVIRONMENTAL IMPACT: refers to changes of ecological or environmental aspects, e.g. climate protection, protection of species, or sustainability of products
  OTHER: any sentences that express impact, which are not captured by the other categories.
  (5) Simply return one category without any explanation.
  (6) Sentence: <Sentence>. Answer:

  We also tested shorter prompts, including removing definitions of impact categories in (4) and removing the definition of the term "impact" (2).
  Few-shot prompting follows a similar structure, and the prompt includes one example for each category:

[9] https://replicate.com/meta/llama-2-70b-chat/api

(1) - (6)
(7) For example:
Sentence: <NO-IMPACT sentence> Answer: No-IMPACT
Sentence: <TECHNICAL sentence> Answer: TECHNICAL
...
Sentence: <OTHER sentence> Answer: OTHER
Sentence: <Sentence> Answer:

- **Fine-tuning**: We fine-tune the Huggingface version of BERT[10] and Llama2[11] models using a training set with pairs of task prompts and answers. For BERT, considering that all classification tasks (main category, sub-category, and intensity) involve multiclass classification, where each sentence is assigned only one class, the output from the encoder layer can be forwarded to the Linear and Softmax functions to determine the most probable label. Additionally, the structure of the training set for BERT models is straightforward: The prompt is a sentence, and the answer is an annotated category. Fine-tuning Llama2 models is particularly challenging due to their substantial number of parameters as the smallest Llama2 model has 7 billion parameters and since many researchers, including us, have limited computing capacity available to them. To address this limitation, we used a newly developed Quantized Low-Rank Adaptation (QLoRA) method for fine-tuning. This method introduces 4-bit NormalFloat (NF4) quantization and double quantization techniques to reduce memory usage and enhance parameter efficiency during fine-tuning [68]. We used QLoRA to fine-tune the Llama2-chat-7b model. The structure of the training set follows zero-shot prompt is described in **Zero-shot and few-shot learning**.

7.2 Data Splits

Testing, validation, and training sets for model evaluation and fine-tuning are constructed as follows to address the imbalanced category distribution described in Section 6:

- **Testing set:** We randomly selected 5,391 sentences from our annotated corpus as the testing set. These sentences were used to investigate the performance of fine-tuned models as well as zero-shot and few-shot promoting for LLMs.
- **Validation set:** We randomly selected 7,257 annotated sentences from the corpus and then under-sampled sentences with the NO-IMPACT category to match the number of sentences in the second most frequent category (i.e., TECHNICAL).[12] This resulted in 2,327 annotated sentences in the validation set.
- **Training set:** We used both under-sampling and over-sampling methods on the corpus excluding the testing and validation sets as classes in our data are highly imbalanced (see Section 6.2). Specifically, we first under-sampled impact-irrelevant

[10] https://huggingface.co/google-bert/bert-base-german-cased
[11] https://huggingface.co/docs/transformers/en/model_doc/llama2
[12] We experimented with this under-sampling approach as well as using the validation set without sampling and did not observe significant differences in the performance of fine-tuned models.

sentences (labeled as NO-IMPACT) by randomly selecting 10,000 instances, and then over-sampled the minority impact categories using the Synthetic Minority Oversampling Technique (SMOTE), which duplicates examples in each minority category until the number of examples matches that of the majority impact category (i.e., ACADEMIC).[13] This resulted in a total of 35,384 sentences in the training set, with 12,527 unique impact-relevant sentences and 22,857 duplicates of minority categories. Among the 12,527 unique impact-relevant sentences, around 59% of sentences were relevant to academic or technical impact.

In addition to the original annotated data, we also conducted experiments after removing all sentences with NO-IMPACT category since people may have clean data in which every sentence indicates an impact. NO-IMPACT was removed from the prompts accordingly.

7.3 Results and Observations

Table 4 shows the performance of various models in terms of accuracy and F1 scores for classifying main categories. In the following, we summarize the results and draw conclusions from our observations.

(I) Original LLMs may not work well for low-resourced fields and novel classification schemas: Llama2 and ChatGPT exhibited low performance (accuracy and F1 below 0.17) in both zero-short and few-short prompting when the dataset contained impact-irrelevant sentences (labeled as NO-IMPACT) and the prompt lacked definitions for individual impact categories. While adding a definition for "impact" to the prompts slightly increased classification performance, incorporating detailed definitions of these categories in the prompt significantly improved their accuracy and F1 scores to the range of 0.35 to 0.4. GPT-4 performed better than Llama2 and ChatGPT, achieving higher accuracy and F1 by approximately 0.1-0.15. These models frequently misclassified "NO-IMPACT" instances as "TECHNICAL" or "ACADEMIC". Sentences from research projects often contain technical or academic content that does not reference impact. This indicates a limitation in the model's ability to grasp the concept of "impact", leading it to classification based at least partially on broader topic categories rather than only impact categories.

Removing impact-irrelevant sentences improved classification performance. However, even without impact-irrelevant sentences, the model's accuracy and F1 scores only reached around 0.49-0.0.57, with GPT-4 having the best performance. Additionally, few-shot prompting with concrete examples outperformed zero-shot prompting in most scenarios, even though the improvement is often minimal and not always consistent. This finding is aligned with prior research: Given that the training of LLMs heavily relies on word distributions of training data in high-resource fields, they may struggle to perform well in tasks with rare or out-of-vocabulary words including in low-resource fields [59].

(II) Without fine-tuning, free and open-source LLMs achieved slightly lower performance than commercial LLMs in most tested cases: Even though

[13] We experimented with various under-sampling and over-sampling strategies as well as using the original data. This strategy yielded optimal performance.

Table 4 Model Performance of Prompting and Fine-tuning. Subscript i and ii means adding definitions of the concept "impact" and "individual impact" in prompts, respectively. Subscript *sub* means this is a classification model for sub-categories. "With No-Impact" and "W/o No-Impact" in the column Data indicate whether the data contains impact-irrelevant sentences or only impact-relevant sentences, respectively.

Method	**Model**	**Accuracy (F1)**	
		w/ no-impact	*w/o no-impact*
Zero-shot	Llama2-chat-70b	0.13 (0.08)	0.49 (0.42)
	Llama2-chat-70b$_i$	0.15 (0.13)	0.49 (0.44)
	Llama2-chat-70b$_{ii}$	0.16 (0.14)	0.49 (0.45)
	Llama2-chat-70b$_{sub}$	0.08 (0.08)	0.25 (0.23)
	ChatGPT	0.17 (0.11)	0.53 (0.47)
	ChatGPT$_i$	0.18 (0.13)	0.54 (0.48)
	ChatGPT$_{ii}$	0.36 (0.40)	0.54 (0.48)
	ChatGPT$_{sub}$	0.09 (0.08)	0.29 (0.28)
	GPT4	0.29 (0.31)	0.54 (0.48)
	GPT4$_i$	0.30 (0.32)	0.55 (0.49)
	GPT4$_{ii}$	0.49 (0.50)	0.56 (0.53)
	GPT4$_{sub}$	0.13 (0.13)	0.34 (0.32)
Few-shot	Llama2-chat-70b	0.15 (0.10)	0.53 (0.49)
	Llama2-chat-70b$_i$	0.14 (0.09)	0.53 (0.49)
	Llama2-chat-70b$_{ii}$	0.19 (0.17)	0.54 (0.50)
	Llama2-chat-70b$_{sub}$	0.11 (0.09)	0.31 (0.28)
	ChatGPT	0.15 (0.14)	0.54 (0.49)
	ChatGPT$_i$	0.15 (0.14)	0.53 (0.48)
	ChatGPT$_{ii}$	0.23 (0.21)	0.55 (0.50)
	ChatGPT$_{sub}$	0.11 (0.10)	0.31 (0.29)
	GPT4	0.30 (0.31)	0.56 (0.50)
	GPT4$_i$	0.30 (0.31)	0.57 (0.52)
	GPT4$_{ii}$	0.46 (0.50)	0.57 (0.53)
	GPT4$_{sub}$	0.15 (0.14)	0.35 (0.34)
Fine-Tuning	Llama2-chat-7b	0.62 (0.65)	0.61 (0.62)
	Llama2-chat-7b$_{sub}$	0.49 (0.46)	0.43 (0.41)
	BERT	0.72 (0.74)	0.65 (0.64)
	BERT$_{sub}$	0.58 (0.64)	0.46 (0.43)

ChatGPT is currently the most widely used LLM, free and open-source models may achieve similar performance. In our experiments, while Llama2-chat-70b performs slightly worse than ChatGPT (by up to 0.05) in all zero-shot scenarios (except for task with the "no-impact" category), their accuracy and F1 scores in most few-shot scenarios were very similar. GPT-4 had a better performance than ChatGPT, but was more expensive to use [60]. Given the high cost of commercial models, free and open-source LLMs with few-shot learning prompting can be a viable option for specific tasks when compute power is limited.

(III) Fine-tuning is an preferred approach over zero- and few-shot prompting: Fine-tuning is thus the preferred approach to automated classification

for low-resouce fields and/or novel classification schema as the model can learn knowledge from our annotated data. Our results show that the fine-tuned Llama2 and BERT models have significantly higher accuracy and F1 scores than zero- and few-shot prompting using Llama2-chat-70b, ChatGPT, and GPT-4 by at least 0.08 in all cases. This highlights the importance of diligent manual annotations as the empirical basis for the impact classification task, which is one of the most important findings of our experiments.

(IV) Fine-tuning smaller models may outperform fine-tuning larger models when the training data is not large enough: Fine-tuned LLMs, such as Llama2, may not necessarily outperform smaller LMs like BERT with around 100 million parameters, at least in the given classification task. Our experiments revealed that while Llama2 and BERT have similar performance for impact-relevant sentences, and Llama2 achieved lower accuracy and F1 by around 0.1 when predicting the label of No-IMPACT. This finding aligns with the zero-shot and few-shot results, where the original Llama2 model failed to accurately predict the label of NO-IMPACT. This suggests that although LLMs encode extensive knowledge, this knowledge might not adequately cover low-resourced fields or specific downstream tasks. In other words, fine-tuning can only adjust certain aspects of this pre-existing knowledge unless a substantial amount of data from low-resource fields or tasks is added for fine-tuning. Clearly, the data we used for fine-tuning is not large enough to calibrate Llama2 effectively for impact classification. Therefore, fine-tuning smaller transformer models like BERT could be a more effective choice in low-resource fields because this approach would be less skewed towards pre-existing knowledge that is not helpful for impact classification.

(V) Different performance among different main categories: For the main categories, we observed that nearly all models exhibit superior performance in identifying technical, economic, and academic impacts, while they struggled with detecting "OTHER" impact. This outcome is unsurprising, given that technical, academic, and economic impacts are inherently more discernible while "OTHER" is inherently vague.

(VI) Sub-category classification is more challenging than main-category: Considering that sub-categories are more nuanced than main categories and that we have only few data points for many of the subcategories, classifying them poses a greater challenge. Specifically, the performance of fine-tuned models and original LLMs indicates that accuracy and F1 scores for sub-category classification are lower compared to main category classification.

(VII) All models perform best in the domain of linguistics and worst in the domain of mobility: Table 5 shows the classification accuracy (for main categories) across different methods, models, domains, and intensity. All models consistently achieves best performance in linguistics and worst performance in mobility. Accuracy scores in music and AI are generally close. This finding implies widely used pre-trained have a similar range of knowledge in terms of domains even though they are trained based on different methods.

(VIII) Higher impact may not suggests higher classification accuracy: All models exhibited higher accuracy on sentences of medium than of high and low impact intensity. Given that the domain of linguistics, where all models achieve the

Table 5 Comparison of the performance of the classification models on the main categories across different methods, domains, and intensities (without impact-irrelevant sentences)

		Domain				Intensity		
Method	Model	Music	Linguistics	Mobility	AI	High	Medium	Low
Zero-shot	Llama2	0.46	**0.58**	0.36	0.49	0.45	**0.53**	0.43
	ChatGPT	0.54	**0.62**	0.39	0.49	0.49	**0.59**	0.45
Few-shot	Llama2	0.54	**0.58**	0.44	0.51	0.45	**0.57**	0.50
	ChatGPT	0.55	**0.62**	0.41	0.53	0.51	**0.60**	0.46
Fine-tuning	Llama2	0.62	**0.65**	0.53	0.63	0.61	**0.65**	0.55
	BERT	0.66	**0.69**	0.61	0.61	0.61	**0.68**	0.61

Table 6 Accuracy and F1 of few-shot prompting in ChatGPT across different domains and intensities. The scores of high-intensity impact in the domain of mobility is excluded due to an insufficient number of instances for comparison.

	Intensity					
	High		Medium		Low	
Domain	Accuracy	F1	Accuracy	F1	Accuracy	F1
Music	0.52	0.43	**0.63**	**0.60**	0.43	0.33
Linguistics	0.50	0.49	**0.64**	**0.60**	0.57	0.53
Mobility	/	/	**0.47**	**0.43**	0.35	0.28
AI	0.48	0.41	**0.57**	**0.49**	0.52	0.45

highest performance, has the highest proportion of medium impact intensity (see Figure 3 in Section 6), the high classification accuracy for medium intensity might be attributed to the domain factor rather than the intensity factor itself. To investigate further, we examined the contingency table of domain and intensity. Table 6 shows the accuracy and F1 scores of few-shot prompting in ChatGPT across different domains and intensities, still revealing the highest performance with medium impact intensity. Similar results were observed when using other models, although we did not list all of them due to space constraints. This finding suggests that there might be a nonlinear relationship between the intensity of impact and classification accuracy.

8 Discussion and Conclusion

In this paper, we empirically developed and evaluated a novel annotation schema designed to effectively capture and categorize the impact of scientific research across domains within academia and beyond, based on research reports. Our schema classifies the impact of research on academia, economy, society, culture, politics, law, technology, and the environment. We showed that our schema is applicable to various scientific domains from natural science, engineering, and the humanities.

We annotated a large dataset consisting of over 45k sentences extracted from research reports based on the classification schema we developed. High Kappa scores indicate a strong agreement between our annotators, which suggests that our annotation schema adequately and distinctly captures and categorizes the impact in

project reports from different domains. We conducted an in-depth examination of our annotated dataset and found patterns regarding the distribution of impact categories among different domains, the co-occurrences of impact categories, and impact-indicating as well as category-specific keywords. These findings allow a deeper understanding of domain-dependent and domain-independent features of scientific impact and inspire future work on automatic impact detection and analysis. We finally probed the feasibility of automatic impact classification with a topic model, which was able to detect and distinguish certain impact categories from text.

We then applied transformer-based models to the annotated dataset to explore their potential for impact classification. Our results indicate that LLMs like Llama2, ChatGPT, and GPT-4 do not perform well in this task, likely because they lack sufficient knowledge about "impact" and tend to focus on general scientific topics. As a result, fine-tuning remains the most effective method for impact classification. While fine-tuning open-source LLMs like Llama2 may be promising when there is a substantial amount of data to adjust the model's parameters, fine-tuning smaller models from the BERT family is still preferred when the training dataset is relatively small, as it offers better classification accuracy along with lower costs and time requirements.

In summary, our results show that fine-tuning LLMs for the task of impact classification on annotated datasets significantly outperforms zero- and few-shot prompting approaches, which in turn indicates that state-of-the-art LLMs without fine-tuning may not work well for novel classification schemas such as our impact classification schema. This highlights the importance of diligent manual annotations as empirical basis in the field of computational social science.

References

[1] Bornmann, L., Daniel, H.-D.: Does the h-index for ranking of scientists really work? Scientometrics **65**(3), 391–392 (2005)

[2] Becker, H.A.: Social impact assessment. European Journal of Operational Research **128**(2), 311–321 (2001)

[3] Vanclay, F.: Principles for social impact assessment: A critical comparison between the international and us documents. Environmental Impact Assessment Review **26**(1), 3–14 (2006)

[4] Rezapour, R., Diesner, J.: Classification and detection of micro-level impact of issue-focused documentary films based on reviews. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1419–1431 (2017). ACM

[5] Latané, B.: The psychology of social impact. American psychologist **36**(4), 343 (1981)

[6] Council, N.R., *et al.*: Public Participation in Environmental Assessment and Decision Making. National Academies Press, ??? (2008)

[7] Esteves, A.M., Franks, D., Vanclay, F.: Social impact assessment: the state of the art. Impact assessment & project appraisal **30**(1), 34–42 (2012)

[8] Vinkler, P.: The Evaluation of Research by Scientometric Indicators. Elsevier, ??? (2010)

[9] Wildgaard, L., Schneider, J.W., Larsen, B.: A review of the characteristics of 108 author-level bibliometric indicators. Scientometrics **101**, 125–158 (2014)

[10] Wouters, P., Costas, R.: Users, narcissism and control: tracking the impact of scholarly publications in the 21st century (2012)

[11] Neylon, C., Willmers, M., King, T.: Impact beyond citation: an introduction to altmetrics; an overview of new altmetrics tools. Policy brief/Scholarly Communication in Africa Programme; 2, January 2014 (2014)

[12] Konkiel, S., Piwowar, H., Priem, J.: The imperative for open altmetrics. Journal of Electronic Publishing **17**(3) (2014)

[13] Witt, A., Diesner, J., Steffen, D., Rezapour, R., Bopp, J., Fiedler, N., Köller, C., Raster, M., Wockenfuß, J.: Impact of scientific research beyond academia: an alternative classification schema. Proceedings of the LREC 2018 Workshop on Computational Impact Detection from Text Data, 34–39 (2018)

[14] Rezapour, R., Bopp, J., Fiedler, N., Steffen, D., Witt, A., Diesner, J.: Beyond citations: Corpus-based methods for detecting the impact of research outcomes on society. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6777–6785. European Language Resources Association, Marseille, France (2020). https://aclanthology.org/2020.lrec-1.837

[15] Waldron, A., Mooers, A.O., Miller, D.C., Nibbelink, N., Redding, D., Kuhn, T.S., Roberts, J.T., Gittleman, J.L.: Targeting global conservation funding to limit immediate biodiversity declines. Proceedings of the National Academy of Sciences **110**(29), 12144–12148 (2013)

[16] Bornmann, L.: Measuring the societal impact of research: research is less and less assessed on scientific impact alone—we should aim to quantify the increasingly important contributions of science to society. EMBO reports **13**(8), 673–676 (2012)

[17] Williams, K., Grant, J.: A comparative review of how the policy and procedures to assess research impact evolved in australia and the uk. Research Evaluation **27**(2), 93–105 (2018)

[18] Heyeres, M., Tsey, K., Yang, Y., Yan, L., Jiang, H.: The characteristics and reporting quality of research impact case studies: A systematic review. Evaluation and program planning **73**, 10–23 (2019)

[19] Gomes, D., Stavropoulou, C.: The impact generated by publicly and charity-funded research in the united kingdom: a systematic literature review. Health research policy and systems **17**(1), 22 (2019)

[20] Ma, L., Agnew, R.: Deconstructing impact: A framework for impact evaluation in grant applications. Science and Public Policy **49**(2), 289–301 (2022)

[21] Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., *et al.*: Science of science. Science **359**(6379), 0185 (2018)

[22] Rogers, P., et al.: Theory of change: methodological briefs-impact evaluation no. 2. Technical report (2014)

[23] Ellegaard, O.: The application of bibliometric analysis: disciplinary and user aspects. Scientometrics **116** (2018) https://doi.org/10.1007/s11192-018-2765-z

[24] Park, M., Leahey, E., Funk, R.J.: Papers and patents are becoming less disruptive over time. Nature **613**(7942), 138–144 (2023)

[25] Kim, J., Diesner, J.: A network-based approach to coauthorship credit allocation. Scientometrics **101**, 587–602 (2014)

[26] Hourneaux Junior, F., Sandes-Guimarães, L.V.d.: Research impact–how to evaluate it? Editorial impact series part 2. SciELO Brasil (2021)

[27] Karanatsiou, D., Misirlis, N., Vlachopoulou, M.: Bibliometrics and altmetrics literature review: Performance indicators and comparison analysis. Performance Measurement and Metrics **18**(1), 16–27 (2017)

[28] Erdt, M., Nagarajan, A., Sin, S.-C.J., Theng, Y.-L.: Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. Scientometrics **109**, 1117–1166 (2016)

[29] Bornmann, L.: Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. Scientometrics **103**(3), 1123–1144 (2015)

[30] Donovan, C.: The qualitative future of research evaluation. Science and Public Policy **34**(8), 585–597 (2007)

[31] Bornmann, L.: Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. Higher Education **73**, 775–787 (2017)

[32] Foundation, N.S.: Broader Impacts. Accessed: 2024-10-18 (n.d.). https://new.nsf.gov/funding/learn/broader-impacts

[33] Penfield, T., Baker, M.J., Scoble, R., Wykes, M.C.: Assessment, evaluations, and

definitions of research impact: A review. Research evaluation **23**(1), 21–32 (2014)

[34] Foundation, W.K.: WK Kellogg Foundation Logic Model Development Guide. WK Kellogg Foundation, ??? (2004)

[35] Donovan, C., Hanney, S.: The 'payback framework'explained. Research Evaluation **20**(3), 181–183 (2011)

[36] Buxton, M., Hanney, S.: Evaluating the nhs research and development programme: will the programme give value for money? Journal of the Royal Society of Medicine **91**(35_suppl), 2–6 (1998)

[37] Woolston, C.: Measuring societal impact: how to go beyond standard publication metrics. Nature Publishing Group (2023)

[38] Nyboer, E.A., Nguyen, V.M., Young, N., Rytwinski, T., Taylor, J.J., Lane, J.F., Bennett, J.R., Harron, N., Aitken, S.M., Auld, G., *et al.*: Supporting actionable science for environmental policy: Advice for funding agencies from decision makers. Frontiers in Conservation Science **2**, 693129 (2021)

[39] Drummond Otten, C., Fischhoff, B.: Assessing broader impacts of funded research: the us national science foundation v. lamar smith. Science and Public Policy **49**(2), 313–323 (2022)

[40] Langfeldt, L., Scordato, L.: The role of research funding agencies in linking research to societal impacts (2015:8) (2015)

[41] Bornmann, L.: What is societal impact of research and how can it be assessed? a literature survey. Journal of the American Society for Information Science and Technology **64**(2), 217–233 (2013)

[42] Bozeman, B., Sarewitz, D.: Public value mapping and science policy evaluation. Minerva **49**, 1–23 (2011)

[43] Borg, D., Sestito, G.S., Silva, M.M.: Machine-learning classification of environmental conditions inside a tank by analyzing radar curves in industrial level measurements. Flow Measurement and Instrumentation **79**, 101940 (2021)

[44] Zhang, H., Pan, J.: Casm: A deep-learning approach for identifying collective action events with text and image data from social media. Sociological Methodology **49**(1), 1–57 (2019)

[45] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P.S., He, L.: A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology (TIST) **13**(2), 1–41 (2022)

[46] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. (2018)

[47] Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training. (2018). https://api.semanticscholar.org/CorpusID:49313245

[48] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 86 (2021)

[49] Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620 (2019)

[50] Wang, P., Fang, J., Reinspach, J.: Cs-bert: a pretrained model for customer service dialogues. In: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pp. 130–142 (2021)

[51] Maheshwari, H., Singh, B., Varma, V.: Scibert sentence representation for citation context classification. In: Proceedings of the Second Workshop on Scholarly Document Processing, pp. 130–133 (2021)

[52] Cai, X., Liu, S., Yang, L., Lu, Y., Zhao, J., Shen, D., Liu, T.: Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. Journal of Biomedical Informatics **127**, 103999 (2022)

[53] Reddy, S.M., Saini, N.: Overview and insights from scope detection of the peer review articles shared tasks 2021. In: Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25, pp. 73–78 (2021). Springer

[54] Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised hierarchical text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6826–6833 (2019)

[55] Shen, S., Liu, J., Lin, L., Huang, Y., Zhang, L., Liu, C., Feng, Y., Wang, D.: Sscibert: A pre-trained language model for social science texts. Scientometrics **128**(2), 1241–1263 (2023)

[56] Jeong, Y., Kim, E.: Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks. IEEE Access **10**, 60805–60813 (2022)

[57] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)

[58] Reynolds, L., McDonell, K.: Prompt programming for large language models: Beyond the few-shot paradigm. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–7 (2021)

[59] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al.: Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology, 100017 (2023)

[60] Loukas, L., Stogiannidis, I., Diamantopoulos, O., Malakasiotis, P., Vassos, S.: Making llms worth every penny: Resource-limited text classification in banking. In: Proceedings of the Fourth ACM International Conference on AI in Finance, pp. 392–400 (2023)

[61] Becker, M., Han, K., Werthmann, A., Rezapour, R., Lee, H., Diesner, J.: Detecting impact relevant sections in scientific research. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 4744–4749 (2024)

[62] Benoit, K., Conway, D., Lauderdale, B.E., Laver, M., Mikhaylov, S.: Crowd-sourced text analysis: Reproducible and agile production of political data. American Political Science Review **110**(2), 278–295 (2016)

[63] Grootendorst, M.R.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. ArXiv **abs/2203.05794** (2022)

[64] Chan, B., Schweter, S., Möller, T.: German's next language model. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6788–6796 (202)

[65] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[66] Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Computing Surveys **55**(13s), 1–40 (2023)

[67] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)

[68] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314 (2023)