Using LLMs for Automatic Discipline Annotation in a Diachronic Corpus of English Scientific Papers

Sergei Bagdasarov, Diego Alves, Stefan Fischer, Elke Teich
Saarland University
Saarbrücken, Germany
sergeiba@lst.uni-saarland.de, diego.alves@uni-saarland.de
stefan.fischer@uni-saarland.de, e.teich@mx.uni-saarland.de

Abstract

This study investigates the potential of generative large language models (LLMs) to automatically identify the disciplines of scientific papers in the Royal Society Corpus (RSC) – an extensive collection of English scientific publications spanning more than three centuries. We evaluated eight open-source, state-of-the-art LLMs from four model families on a manually annotated subset and further validated the three best-performing models on a corpus of modern scientific texts. These models were subsequently used for large-scale annotation of the RSC. The models exhibited robust and consistent performance, with at least two LLMs agreeing on the same label for 98.3% of the documents. We then conducted an error analysis of papers assigned divergent labels and a diachronic case study of disciplinary trends within the corpus. The error analysis revealed that most discrepancies occurred in twentieth-century texts, reflecting the growing interdisciplinarity of research. The diachronic analysis showed a gradual decline in disciplinary diversity over time as well as fluctuations corresponding to major paradigm shifts such as the Chemical Revolution and key twentieth-century developments in Physics. The discipline labels generated by the three models will be made publicly available.

Keywords: LLMs, discipline annotation, scientific corpus

1. Introduction

Understanding the structure of modern science depends on how research is categorised. Major databases like Scopus[1] and Web of Science[2] classify papers into defined disciplinary areas using established annotation schemes, providing an essential foundation for studies that examine patterns and interactions within specific scientific domains.

However, this is not the case for other databases that contain historical scientific texts. The Royal Society Corpus (RSC) (Fischer et al., 2020) is a valuable resource for the analysis of the history of science and other topics in the digital humanities, comprising texts from the seventeenth to the twentieth century. Although it includes extensive metadata, the RSC lacks explicit disciplinary classifications for individual texts. Consequently, for a better understanding of phenomena such as knowledge transfer and the emergence of specific scientific trends within particular fields, a disciplinary classification is fundamental.

Motivated by this need, the present study investigates whether and how large language models (LLMs) can be used to classify scientific texts into disciplinary categories, with the goal of enriching the metadata of the RSC. Key challenges include selecting the most suitable model for historical texts, addressing the less explicit definition of disciplinary boundaries in earlier periods of the RSC, and evaluating whether the results obtained with LLMs are consistent with those from other corpora that use different discipline taxonomies.

To this end, we first tested eight LLMs (representing four different model families) on a sample of manually annotated RSC texts spanning all periods. We then evaluated the three best-performing LLMs on another dataset of scientific publications to assess the robustness of the discipline classification task across different types of texts and discipline labels. Finally, we applied the three selected models to annotate the entire RSC corpus[3] and conducted two case studies: (a) a diachronic analysis of disciplines in the RSC, and (b) a topic modeling experiment using the labels provided by the top three LLMs to examine papers where all three models disagreed. Our contributions are, therefore, the following:

- We evaluate eight state-of-the-art LLMs for automatic discipline identification on both diachronic and contemporary datasets.
- We enrich RSC metadata by providing consistent discipline labels.

[1] Scopus, Elsevier. Available at: https://www.scopus.com (accessed 10 October 2025).

[2] Web of Science Core Collection, Clarivate. Available at: https://www.webofscience.com (accessed 10 October 2025).

[3] Discipline labels generated by the three best-performing models as well as paper metadata are available on GitHub: https://github.com/s-bagdasarov/RSC-Disciplines

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2376–2386
11–16 May 2026.

- We show the disciplinary evolution of contributions sent to the Royal Society of London across +300 years.

2. Related Work

Large collections of digitised texts constitute a valuable resource for both corpus linguists and scholars in natural language processing (NLP). However, the usefulness of such data depends heavily on the quality of the accompanying metadata, which often has to be generated through (semi-)automatic procedures. Although NLP techniques can be employed to produce a wide range of metadata ranging from text target audience (Säily et al., 2025) and keywords (Kähler et al., 2025) to citation information (Choi et al., 2023) and table descriptions (Singh et al., 2025), text domain is arguably one of the most important types of information, as it can prove useful across various research contexts.

A considerable number of studies used Latent Dirichlet Allocation (LDA) topic modeling, an unsupervised probability-based technique that generates clusters of words related to different topics, for subject identification for web texts (Altarturi et al., 2023; Lyding et al., 2014), press releases (Glowacka-Musial, 2022) or news articles (Kuzman et al., 2023). (Fankhauser et al., 2016) used LDA on the RSC, tracing topic development diachronically and Menzel et al. (2021) used the resulting top six topic labels as document metadata for an enhanced release of the corpus.

The advent and rapid development of language models led to a wide use of this NLP technology for subject identification as well. Repo et al. (2024) fine-tuned a BERT model to classify historical English texts into different registers. Similarly, Roy and Ghosh (2023) used a BERT model to automatically identify subjects of scientific publications in the field of information studies. George and Sumathy (2023) proposed an interesting approach combining a BERT model with LDA for better topic modeling results.

In scenarios where no or very limited data is available for fine-tuning, generative LLMs offer an alternative and easy-to-implement solution. Kuzman et al. (2023) used GPT-3.5 for genre identification, achieving promising results. In turn, Säily et al. (2025) tested a variety of newer GPT models on the task of generating genre labels for novels from the Corpus of Historical American English. Focusing on contemporary English and German scientific texts, the second LLMs4Subjects shared task (D'souza et al., 2025) featured a subtask on automatic domain identification, in which participating systems were required to assign one or more labels from 28 possible categories (Ho, 2025; Shirali et al., 2025). While this task provided valuable insights and potential solutions to the problem of discipline annotation, its results are unfortunately not directly applicable to our study, as it did not account for diachronic variation and employed a classification scheme that differs considerably from that used by the Royal Society.

As shown above, applications on metadata generation using different NLP approaches abound, including for collections of modern scientific publications. However, historical scientific records seem to have received little attention in this regard. We will address this limitation in this paper by testing eight LLMs for scientific domain identification of historical papers and providing discipline labels for the entirety of RSC.

3. Data

3.1. Royal Society Corpus

The main focus of our work is on the Royal Society Corpus[4] (RSC) (Fischer et al., 2020) – a diachronic collection of more than 40 000 contributions to the Philosophical Transactions and Proceedings of the Royal Society of London. The texts span from the mid 17th century, when this periodical originated, to the end of the 20th century, covering almost 350 years of the development of scientific knowledge. Table 1 summarises some basic statistics about the corpus.

Years	# Texts	# Tokens
1665–1699	1,325	2,582,856
1700–1749	1,686	3,414,795
1750–1799	1,819	6,342,489
1800–1849	2,774	9,112,274
1850–1899	6,754	36,993,412
1900–1949	10,011	65,431,384
1950–1996	23,468	172,018,539

Table 1: Size of the RSC across time periods.

Founded in 1665, the Philosophical Transactions of the Royal Society of London, later joined by the Proceedings, was among the first journals devoted to sharing scientific knowledge. Its early issues consisted of letters and observations on diverse topics, from astronomy and medicine to travel accounts and observations of curious natural phenomena, reflecting the loosely defined disciplinary boundaries of early science.

Although its editorial structure evolved over the centuries, the Transactions remained broadly interdisciplinary until the 19th century, when increasing specialization and the rise of field-specific journals made its generalist format untenable. In response,

[4] https://fedora.clarin-d.uni-saarland.de/rsc_v6/

it was divided into Section A (mathematical and physical sciences) and Section B (biological sciences).

Despite major changes in science throughout the 20th and 21st centuries, the journal's two-section structure persists. The Royal Society now distinguishes between broad "Disciplines" and more specific "Subjects" within each section, and this disciplinary framework forms the basis of our labeling system for the LLM's classification of RSC texts.

3.2. SciTex Corpora

For additional validation of the selected models and prompting approaches, we also used contemporary scientific corpora. Specifically, the SciTex corpora[5] (Degaetano-Ortlieb et al., 2013) provide discipline labels as metadata. The SciTex corpora contain journal papers from the 1970s to the early 2000s that are annotated with discipline labels. The corpora have a three-way composition (see Figure 1) and were built to investigate linguistic usage in interdisciplinary fields (B) combining computer science (A) with another discipline (C).

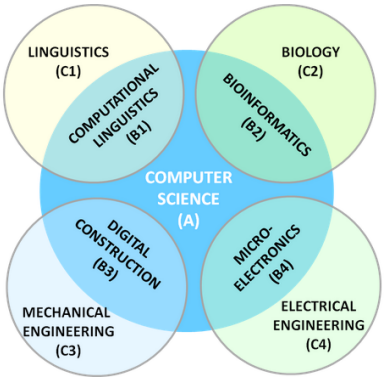

Figure 1: Disciplines in the SciTex corpora.

4. Model Selection

4.1. Model Testing on RSC

Since model performance may vary greatly depending on the input data, it was crucial for us to use a small fraction of the RSC for model testing. To this end, we randomly selected 763 texts, evenly distributed across the entire RSC time span, and had them labeled by two human annotators according to the following classification: *astronomy*, *biology*, *chemistry*, *computer science*, *earth sciences*, *engineering and technology*, *humanities*, *mathematics*, *medicine*, *miscellaneous*, and *physics*. This taxonomy is based on the one used by the Royal Society for more recent publications. Compared to the RSC discipline list, we added *humanities*, *medicine*, and *miscellaneous*[6], as older RSC articles can relate to these fields. Additionally, we removed the *cross-disciplinary studies* discipline, since our aim is to assign a single, most relevant discipline to each paper. As illustrated in Figure 2, the annotators achieved substantial agreement across all RSC periods (overall Cohen's κ = 0.7). Both annotators agreed on the label in 556 out of 763 cases (72.87%).

A qualitative analysis of the errors showed that most disagreements occurred between *biology* and *earth sciences*, particularly in papers about fossils describing the anatomy of prehistoric animals. Additional confusions were observed among *chemistry*, *physics*, and *engineering*, depending on the degree of technological focus or for papers from the early twentieth century involving experiments in atomic theory. Finally, some disagreements arose between *miscellaneous* and *humanities*, especially for texts describing historical events.

We used this annotated subset of texts to test eight LLMs from four families: Gemma (GemmaTeam, 2025), GPT-OSS (OpenAI, 2025), Llama (Grattafiori et al., 2024), and Qwen (Yang et al., 2025; QwenTeam, 2025). Table 2 offers an overview of the models' technical characteristics.

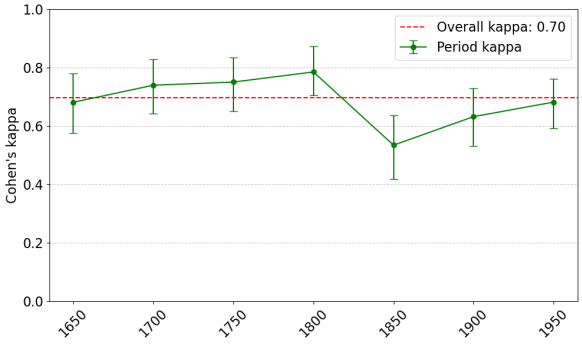

Figure 2: Inter-annotator agreement per period (e.g. label *1800* indicates the time period from 1800 to 1849). Dashed line shows the overall agreement.

We used three prompt versions, referred to as *minimal*, *extended*, and *detailed*. The *minimal* prompt included a brief description of the task, 11 discipline labels, a text snippet of 200 words, and output requirements. The *extended* prompt added the paper title, while the rest remained identical to the minimal prompt. Finally, the *detailed* prompt

[5] https://fedora.clarin-d.uni-saarland.de/scitex/

[6] The miscellaneous class is intended for non-scientific texts in the corpus, such as obituaries, indexes, etc.

Model	Parameters	Layers	Context
gpt-oss-20b	21B	24	128K
gemma-3-4b-it	4B	34	128K
gemma-3-12b-it	12B	48	128K
Llama-3.1-8B-Instruct	8B	32	128K
Qwen2.5-7B-Instruct	7.61B	28	131K
Qwen2.5-32B-Instruct-AWQ	32.5B	64	131K
Qwen3-8B	8.2B	36	32K
Qwen3-32B-AWQ	32.8B	64	32K

Table 2: Specifications of the tested LLMs.

included role assignment, data description, and more precise task instructions. Similar to the *extended* prompt, models were provided with both paper title and text snippet. The full prompts are provided in Appendix A.

The prompts described above can be classified as zero-shot. Additionally, we evaluated three few-shot settings, which differed in the number of examples per discipline (one, two, or three). For instance, the few-shot setting with one paper per label resulted in a prompt containing 11 examples. Only papers for which both human annotators assigned the same discipline label were used as examples. Each example, enclosed within `<EXAMPLE></EXAMPLE>` tags, was included in the user message and preceded by the sentence: "Below you can find examples of texts annotated by experts:". Depending on the prompt type, examples contained either text snippet alone or both text snippet and paper title.

All models were prompted using the Hugging Face `transformers` Python library.[7] Reasoning was enabled for models with reasoning capabilities.[8] The temperature was set to 0.2 to obtain more deterministic predictions and reduce endless repetitions in reasoning models. The maximum output length was set to 512 tokens for non-reasoning models and 4096 tokens for reasoning models.

A model's prediction was considered correct if it matched at least one of the labels assigned by human annotators. Model performance was quantified using accuracy, defined as the proportion of correct predictions relative to the total number of texts.

Three models achieved the highest performance, reaching 82% accuracy in both extended and detailed prompting settings: `gpt-oss-20b` and the two `Qwen3` models (see Table 3). These models demonstrated robust performance, showing minimal sensitivity to variations in prompt wording or the number of examples provided. Interestingly, the smaller `Qwen3` model with 8 billion parameters performed on par with the larger `GPT` and `Qwen3` models, likely due to the quantization applied to the latter two.

Prompt Type	gpt-oss-20b	gemma-3-4b-it	gemma-3-12b-it	Llama-3.1-8B-Instruct	Qwen2.5-7B-Instruct	Qwen2.5-32B-Instruct-AWQ	Qwen3-8B	Qwen3-32B-AWQ
minimal_0	.81	.68	.76	.49	.70	.75	.80	.80
minimal_1	.80	.74	.78	.71	.74	.79	.80	.80
minimal_2	.81	.73	.79	.72	.74	.79	.78	.81
minimal_3	.80	.75	.77	.71	.72	.78	.80	.81
extended_0	**.82**	.69	.77	.52	.72	.76	.80	.81
extended_1	**.82**	.75	.79	.75	.75	.80	.80	**.82**
extended_2	**.82**	.74	.79	.74	.76	.78	**.82**	**.82**
extended_3	.81	.75	.78	.74	.77	.78	.81	.81
detailed_0	.81	.71	.79	.68	.74	.77	.81	.80
detailed_1	.80	.75	.79	.76	.77	.79	.80	.80
detailed_2	.81	.76	.79	.74	.77	.78	**.82**	.81
detailed_3	.81	.76	.79	.75	.76	.79	.81	**.82**
average	.81	.73	.78	.69	.75	.78	.80	.81

Table 3: Model performance across prompt types. Digits after the prompt type indicate the number of examples per discipline. For instance, *minimal_1* stands for the minimal prompt complemented with one example per discipline. The best accuracy scores are presented in **bold**.

In contrast, the performance of smaller non-reasoning models fluctuated considerably depending on prompt design. The most pronounced change was observed in the `Llama` model: while it achieved only 49% accuracy in the minimal zero-shot setting, its accuracy improved to 76% when given a more explicit prompt and one example per discipline. This result is comparable to the performance of larger non-reasoning models and only 6% lower than that of the best-performing reasoning models. Contrary to our expectations, several models exhibited slightly lower accuracy when provided with three examples per discipline compared to two or even one.

Although smaller models combined with more advanced prompt-engineering techniques demonstrated promising results, we will proceed with the top three models for our annotation task to max-

[7] https://pypi.org/project/transformers/

[8] For the `gpt-oss-20b` model, the default medium reasoning effort was used.

imise overall accuracy. Among several prompting settings that achieved the highest accuracy, we select the simplest configuration for each model. For example, `Qwen3-32B-AQW` achieved 82% accuracy using both the extended prompt (with one or two examples per discipline) and the detailed prompt (with three examples per discipline). For this model, we will therefore adopt the extended prompt with one example per discipline.

4.2. Model Testing on SciTex

Before proceeding to annotate the entire RSC, we conducted an additional performance check using the SciTex corpora, considering both the full dataset with all nine disciplines and the core dataset with only five disciplines (*computer science*, *linguistics*, *mechanical engineering*, *electrical engineering*, *biology*). Figure 3 visualises the overall accuracy scores, while Tables 4 and 5 report the classification accuracy for each discipline in the core and full datasets, respectively.

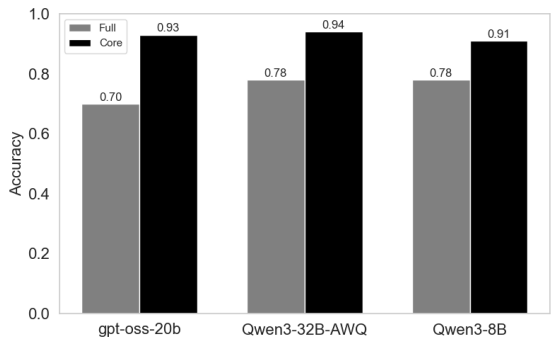

Figure 3: Overall accuracy on SciTex data (full and core).

Discipline	GPT	Q3-8B	Q3-32B
Biology	1.00	1.00	1.00
Computer science	0.97	0.98	0.99
Electrical engineering	0.86	0.71	0.79
Linguistics	1.00	1.00	1.00
Mechanical engineering	0.85	0.92	0.95

Table 4: Accuracy on core SciTex data for `gpt-oss-20b`, `Qwen3-8B`, and `Qwen3-32B-AWQ`.

All three models demonstrate excellent performance, achieving accuracy above 90% on the core discipline set. As expected, accuracy decreases on the full dataset, which includes highly interdisciplinary labels. The discipline with the lowest accuracy is *digital construction*, situated at the intersection of *computer science* and *mechanical engineering*. Both `Qwen3` models perform well on other interdisciplinary labels, whereas the `GPT` model struggles with *bioinformatics* and *microelectronics*.

Discipline	GPT	Q3-8B	Q3-32B
Biology	.96	.98	.96
Computer science	.95	.97	.96
Electrical engineering	.79	.72	.71
Linguistics	.99	.99	.98
Mechanical engineering	.83	.94	.87
Bioinformatics	.47	.80	.86
Computational linguistics	.89	.87	.90
Digital construction	.03	.18	.15
Microelectronics	.53	.66	.76

Table 5: Accuracy on full SciTex data for `gpt-oss-20b`, `Qwen3-8B`, and `Qwen3-32B-AWQ`.

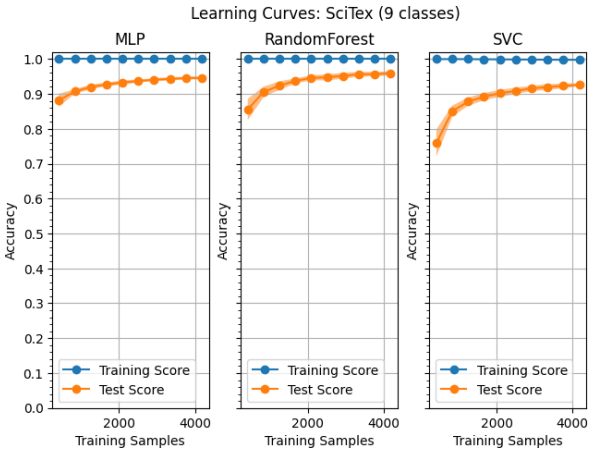

Figure 4: Learning curves of three different classifiers on SciTex.

For comparison, we additionally trained three traditional machine learning classifiers using the full SciTex dataset with nine categories (see Figure 4). While LLMs do not outperform traditional classifiers, they do represent a robust classification solution in scenarios with no or limited labeled data, performing well both on the RSC sample and the full SciTex dataset.

5. Annotation

After the model selection procedure described in Section 4, we proceeded to annotate the entire RSC using the following models and prompting configurations:

- **gpt-oss-20b**: extended zero-shot prompt;
- **Qwen3-8B**: extended prompt with two examples per discipline;
- **Qwen3-32B-AWQ**: extended prompt with one example per discipline.

All three models followed the instructions regarding output format in the vast majority of cases. Deviations were observed in 23 answers by `gpt-oss-20b`, 116 answers by `Qwen3-8B` and 49 answers by `Qwen3-32B-AWQ`.

In three cases, the `GPT` model endlessly repeated the same word sequence without outputting any label. For three papers, the model generated labels other than those required in the prompt: *mystery*, *psychophysical* and *science*. In the remaining 17 cases, the model did use valid labels but included some additional text or symbols.

For both `Qwen3-8B` and `Qwen3-32B-AWQ`, the wrong output format was mainly due to endless repetitions (113 and 44 cases respectively). `Qwen3-8B` generated two invalid labels: *psycology* (for two papers) and *materials science* (for one paper). `Qwen3-32B-AWQ` output two invalid labels as well: *psychology* and *zoology*, both occurring only once. Additionally, `Qwen3-32B-AWQ` used valid labels with wrong formatting in three further cases.

In general, the models show an excellent pairwise agreement with the Cohen's κ around 0.8: $\kappa_{\text{Qwen3-8B, Qwen3-32B}} = 0.799$, $\kappa_{\text{Qwen3-8B, GPT}} = 0.807$, and $\kappa_{\text{Qwen3-32B, GPT}} = 0.816$. Interestingly, the two `Qwen3` models achieved a slightly lower agreement with each other than with the `GPT` model. At least two models assigned the same label for 98.3% of the papers, and all three models agreed on the label for 76.8% of the papers. As shown in Figure 5, after an increase in the 18th century, the model agreement drops slightly in late 19th century, stabilising in the 20th century. In general, this pattern is similar to the one we observed for human annotators, although the fluctuations are much less pronounced.

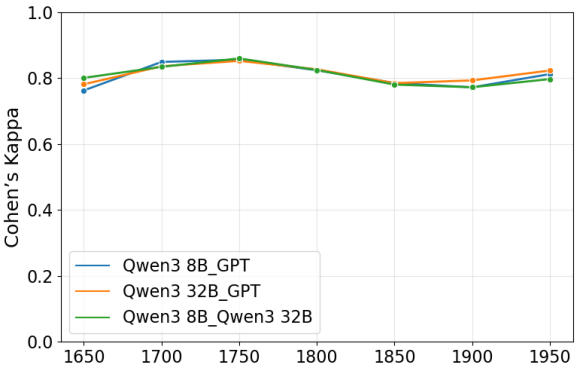

Figure 5: Model agreement per period (e.g. label *1800* indicates the time period from 1800 to 1849). Each line shows the agreement between one model pair.

As shown in Figure 6, the amount of papers with three diverging labels remains relatively stable from 17th century to the first half of the 19th century, starting to increase towards the end of the 19th century. More than half of such problematic papers are concentrated in the second half of 20th century when highly interdisciplinary studies begin to emerge.

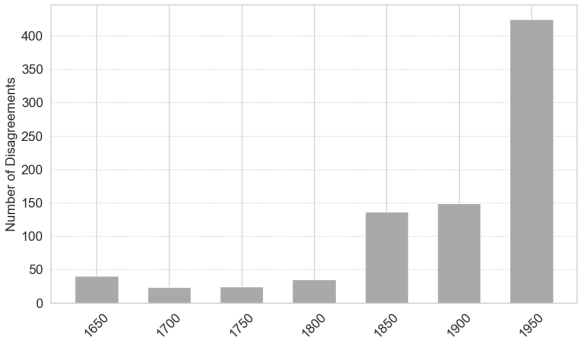

Figure 6: Distribution of papers where all models generated different labels by 50-year periods (e.g. label *1800* indicates the time period from 1800 to 1849).

Section	# Papers
Series A	244
Series B	110
Biographical Memoirs	172
Notes and Records	36
Other	12

Table 6: Distribution of papers with divergent labels across different journals in the 20th century. *Other* includes abstracts from Philosophical Proceedings that appeared before the journal was split into Series A and B in 1905.

Focusing on the 20th century, we observed that most of the problematic papers belonged to the Series A of the Royal Society, covering topics from mathematical, physical and engineering sciences (see Table 6). It is followed by Biographical Memoirs featuring tributes to prominent members of the Royal Society.

6. Results

6.1. Trends in Scientific Fields Over Time

The aim of this analysis is to trace how the relative prominence of scientific fields has shifted over time within the Royal Society Corpus. After filtering for texts where at least two of three large language models agreed on the disciplinary label, 46 988 texts were retained for the analysis. Figure 7 shows the distribution of disciplines across decades, normalized per decade so that each bar

represents the percentage of papers in each discipline.

It is possible to notice that the diversity of disciplines in RSC publications was greater in the earlier periods (up to the early 19th century). *Earth sciences*, as well as *medicine*, were relatively prominent during this time but declined in later decades. The impact of the Chemical Revolution in the late 18th century is evident in the rise of chemistry-related publications, while physics publications notably increase in the early 20th century, corresponding to major advances such as quantum theory and Einstein's theory of relativity. The observed trends indicate that the LLM-based annotations provide coherent and consistent discipline assignments to documents.

6.2. Topic Analysis: Disagreements Among LLMs

The aim of this analysis was to identify the main topics of the 849 articles for which all three LLMs disagreed on discipline assignment. Texts[9] were preprocessed with spaCy (Montani et al., 2023) to retain lemmatized nouns while removing stopwords. Using BERTopic (Grootendorst, 2022) with a CountVectorizer, we extracted ten latent topics and linked them to paper IDs and LLM-assigned disciplines. For each topic, the top three most frequent disciplines were then identified. Table 7 shows the topic number, the number of documents per topic, the representative keywords, and the most associated disciplines.

Broad or ambiguous content, as captured by Topic -1 (representing outlier documents that did not fit any coherent topic) with keywords like "time, result, paper, experiment," contains the largest number of documents, highlighting the challenges LLMs face in classifying general or heterogeneous content. Topic 0, the second biggest cluster, represents papers focused on biographical content. The special class, *miscellaneous*, was added to the discipline set specifically to capture this type of non-scientific text. However, for some articles within this topic, the LLMs struggled to assign a consistent discipline, particularly between *humanities* and *engineering and technology*. Similarly, Topic 3 contains articles with very generic keywords, suggesting that the provided text snippets may not contain enough information for a clear discipline assignment. This likely contributed to LLM disagreement within this topic.

On the other hand, the remaining topics display keywords that may be representative of multiple disciplines. In these cases, the limitation might again stem from the amount of text provided in the prompt, but it could also indicate the interdisciplinary nature of some articles.

[9]For the topic modeling, we used the same snippet of text as provided in the prompts.

7. Conclusion and Future Work

In this study, we leveraged the potential of generative LLMs to enrich metadata in the RSC, – an extensive corpus of English scientific publications spanning more than 300 years. Specifically, we produced discipline labels for the articles in the corpus following the taxonomy available on the website of the Royal Society. We tested eight open-source state-of-the-art LLMs from four different model families on a subset of the RSC manually labeled by two annotators, validating the performance of the top three models (`gpt-oss-20b`, `Qwen3-8B` and `Qwen3-32B-AWQ`) on a dataset of modern scientific publications.

Subsequently, we annotated the RSC articles using these three models. The models showed consistent performance and robust agreement, with at least two LLMs producing the same label for 98.3% of the documents. Our error analysis showed that most disagreements occurred in the 20th century, coinciding with the emergence of highly interdisciplinary studies. At the same time, LLMs struggle with classifying documents with excessively generic vocabulary such as biographical records.

Our diachronic analysis revealed a general decrease in discipline diversity over time. At the same time, fluctuations in the proportion of disciplines reflect important changes in the scientific paradigm like the Chemical Revolution or groundbreaking advancements in Physics in the 20th century.

In future work, we intend to improve our prompting settings by incorporating RAG-like approaches for dynamic example selection and/or by fine-tuning smaller models with resource-effective techniques like LoRA. Additionally, in the spirit of reproducible science, future work will incorporate recently published models, which rank high in the Open Source AI Index, e.g. `OLMo` and `Apertus`.

8. Limitations

This study is subject to several limitations that should be acknowledged. First, a fixed set of randomly selected examples was used for few-shot prompting, so in some cases, the retrieved examples may have not been relevant and may have lead to erroneous labeling. A more dynamic procedure for example selection should improve the robustness of our approach and will be implemented in future work. Second, models were prompted to return one single label for each text. This approach, while providing a clear discipline mapping,

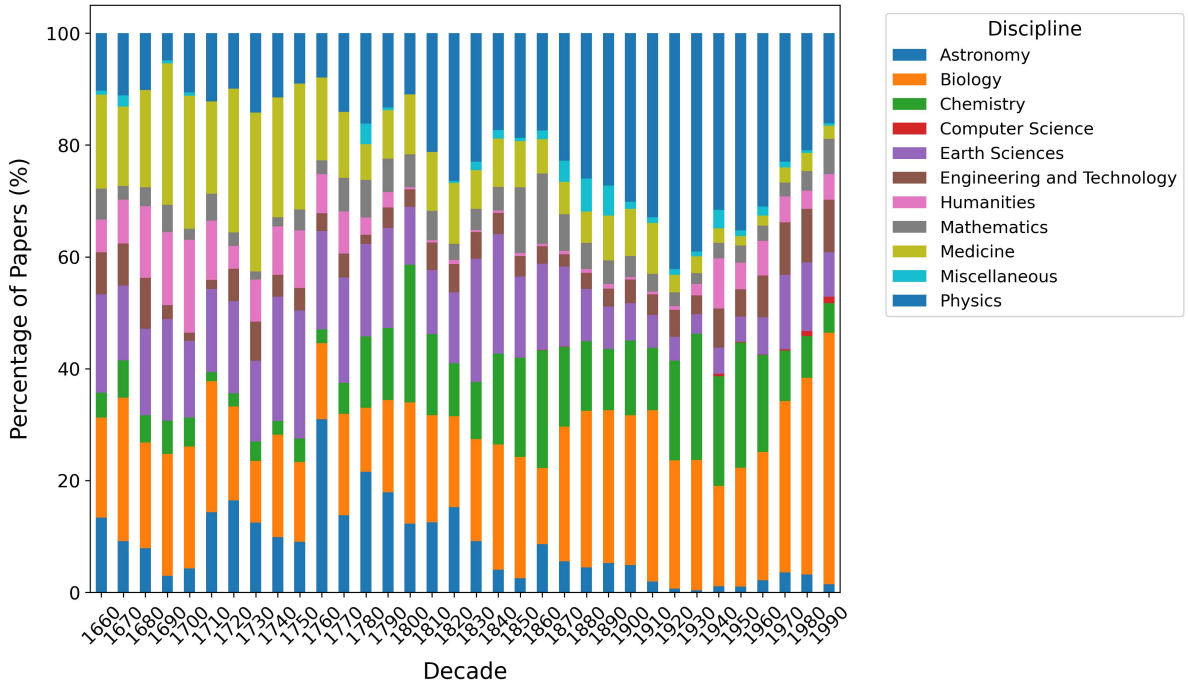

Figure 7: Normalized distribution of scientific disciplines per decade in the RSC, including only papers where at least two of three LLMs agreed on the label. Each bar represents the percentage of papers in each discipline.

Topic	Count	Key Words	Top 3 Disciplines
-1	269	time, result, paper, experiment	Physics, Miscellaneous, Chemistry
0	191	year, father, family, school	Miscellaneous, Humanities, Engineering and Technology
1	87	flame, temperature, solution, equation	Physics, Chemistry, Mathematics
2	65	cell, blood, action, muscle	Medicine, Chemistry, Biology
3	58	word, memory, letter, account	Miscellaneous, Humanities, Physics
4	49	satellite, orbit, atmosphere, observation	Physics, No Answer, Computer Science
5	41	telescope, light, glass, lens	Physics, Miscellaneous, Engineering and Technology
6	40	food, risk, use, problem	Miscellaneous, Engineering and Technology, Chemistry
7	37	meteorite, line, element, rock	Chemistry, Physics, No Answer
8	12	water, ice, pressure, air	Engineering and Technology, Physics, Chemistry

Table 7: Topics extracted from disagreed articles with top 3 assigned disciplines.

may be too restrictive in cases of extremely high interdisciplinarity, where multiple labels can apply. And finally, model evaluation on the RSC was performed only on a small fraction of the corpus. Increasing the number of manually annotated articles may lead to more accurate results. The additional evaluation on SciTex, while providing a general understanding of model performance, cannot replace the evaluation on diachronic datasets of comparable size.

9. Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

10. Bibliographical References

Hassan H. M. Altarturi, Mohammed Saadoon, and Nor Badrul Anuar. 2023. Web Content Topic

Modeling Using LDA and HTML Tags. *PeerJ Computer Science*, 9:e1459.

Won Choi, Hye-Min Yoon, Min-Ho Hyun, Hye-Jin Lee, Ju-Won Seol, Kyung-Doo Lee, et al. 2023. Building an Annotated Corpus for Automatic Metadata Extraction from Multilingual Journal Article References. *PLoS ONE*, 18(1):e0280637.

Stefania Degaetano-Ortlieb, Hans Kermes, Ekaterina Lapshinova-Koltunski, and Elke Teich. 2013. SciTex – A Diachronic Corpus for Analyzing the Development of Scientific Registers. In Paul Bennett, Martin Durrell, Sandra Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, volume 3, pages 93–104. Narr.

Jennifer D'souza, Sameer Sadruddin, Holger Israel, Mathias Begoin, and Diana Slawig. 2025. The Germeval 2025 2nd LLMs4Subjects Shared Task Dataset. Online dataset. Data managers and curators: Jennifer D'Souza (Data manager), Sameer Sadruddin (Data curator).

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2016. Topical Diversification over Time in the Royal Society Corpus. In *Proceedings of Digital Humanities (DH'16)*, Krakow, Poland.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802, Marseille, France. European Language Resources Association.

GemmaTeam. 2025. Gemma 3 Technical Report.

Lijimol George and P. Sumathy. 2023. An Integrated Clustering and BERT Framework for Improved Topic Modeling. *International Journal of Information Technology*, 15:2187 – 2195.

Monika Glowacka-Musial. 2022. Applying Topic Modeling for Automated Creation of Descriptive Metadata for Digital Collections. *Information Technology and Libraries*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan et al. 2024. The Llama 3 Herd of Models.

Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794*.

Clara Wan Ching Ho. 2025. UBFFM at the GermEval-2025 LLMs4Subjects Task: What if we take "You are an expert in subject indexing" seriously? In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 471–478, Hannover, Germany. HsH Applied Academics.

Maximilian Kähler, Lisa Kluge, and Katja Konermann. 2025. DNB-AI-project at the GermEval-2025 LLMs4Subjects Task: KIFSPrompt - Knowledge-Injected Few-Shot Prompting. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 455–464, Hannover, Germany. HsH Applied Academics.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden. Association for Computational Linguistics.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating Linguistically Relevant Metadata for the Royal Society Corpus. *Research in Corpus Linguistics, Challenges in combining structured and unstructured data in corpus development (special issue)*, 9(1):1–18.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. spaCy: Industrial-strength Natural Language Processing in Python.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card.

QwenTeam. 2025. Qwen3 Technical Report.

Lari Repo, Bryan Hashimoto, Antti Liimatta, Lotta Saario, Tanja Säily, Iiro L. I. Tiihonen, Mikko Tolonen, and Veronika Laippala. 2024. Towards Automatic Register Classification in Unrestricted Databases of Historical English. In Simon Coats and Veronika Laippala, editors, *Linguistics across Disciplinary Borders: The March of Data*, pages 97–126. Bloomsbury Publishing, London.

Aditi Roy and Saptarshi Ghosh. 2023. Automated Subject Identification using the Universal Decimal Classification: The ANN Approach. *SRELS Journal of Information Management*.

Parisa Shirali, Zahra Sarlak, and Ebrahim Ansari. 2025. Last Minute at the GermEval-2025 LLMs4Subjects Task: Few-Shot Contrastive Learning for Multilingual Multi-Label Classification. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 465–470, Hannover, Germany. HsH Applied Academics.

Mayank Singh, Abhijeet Kumar, Sasidhar Donaparthi, and Gayatri Karambelkar. 2025. Leveraging Retrieval Augmented Generative LLMs For Automated Metadata Description Generation to Enhance Data Catalogs.

Tanja Säily, Jukka Suomela, Florent Perek, Jimena Jiménez Real, and Turo Vartiainen. 2025. Using Large Language Models to Enrich Corpus Metadata: The Case of Novels in COHA. In *46th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 46)*, Vilnius, Lithuania.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report.

A. Prompts

A.1. Minimal Prompt

System message

Your task will be to classify scientific texts into disciplines.
The disciplines are defined as follows:
- Humanities
- Astronomy
- Biology
- Medicine
- Computer Science
- Chemistry
- Physics
- Mathematics
- Earth Sciences
- Engineering and Technology
- Miscellaneous

For each text, you must use only one discipline from the list above. Return the discipline only without any additional text like "The discipline of the text is:" or "Here is the discipline:".

User message

Identify the discipline of the following scientific paper:
Text: [200 words text snippet]

A.2. Extended Prompt

System message

Your task will be to classify scientific texts into disciplines.
The disciplines are defined as follows:
- Humanities
- Astronomy
- Biology
- Medicine
- Computer Science
- Chemistry
- Physics
- Mathematics
- Earth Sciences
- Engineering and Technology
- Miscellaneous

For each text, you must use only one discipline from the list above. Return the discipline only without any additional text like "The discipline of the text is:" or "Here is the discipline:".

User message

Identify the discipline of the following scientific paper:
Title: [Paper Title]
Text: [200 words text snippet]

A.3. Detailed Prompt

System message

ROLE:
You are an editor of a scientific journal. You are in charge of classifying papers into disciplines.
DATA:
You will be dealing with papers from different time periods ranging from mid 17th century to late 20th century.
TASK:
I will provide you with a title and a text snippet of a scientific paper. Read the title and the snippet carefully to understand the content of the paper and decide as accurately as possible which discipline the paper belongs to.
The disciplines are defined as follows:
- Humanities
- Astronomy
- Biology
- Medicine
- Computer Science
- Chemistry

- Physics
- Mathematics
- Earth Sciences
- Engineering and Technology
- Miscellaneous

For each text, you must use only one discipline from the list above. Return the discipline only without any additional text like ”The discipline of the text is:” or ”Here is the discipline:”.

User message

Identify the discipline of the following scientific paper:

Title: [Paper title]

Text: [200 words text snippet]