Use of large language models in the classification of scientific texts into disciplines

Diego Alves, Sergei Bagdasarov, Badr Mohammed Abdullah, and Stefania Degaetano-Ortlieb

1. Introduction

The classification of scientific publications into disciplines is a fundamental task for organizing and understanding the architecture of knowledge. Yet, this task is complicated by the fact that disciplinary boundaries are not static; they are fluid, historically contingent, and shaped by institutional practices, funding priorities, and evolving research agendas (Abbott, 2001; Frickel and Gross, 2005). This fluidity is evident in the proliferation of interdisciplinary fields and the continuous emergence, differentiation, and reconfiguration of subdisciplines.

In practice, there is no universal consensus on what constitutes a discipline, a fact reflected in the divergent classification systems used across academia and research analytics. Frameworks such as the OECD's Fields of Research and Development (OECD, 2015), the QS Subject Areas[1], and Elsevier's All Science Journal Classification[2] (ASJC) employ varying levels of granularity and organizational logic (Alves et al., 2024). For instance, as summarized in Table 1, these systems differ significantly in the number of categories they define and their use of hierarchical grouping or "other" labels. This lack of a consistent framework underscores a fundamental challenge: any attempt to classify scientific texts must first define a suitable disciplinary taxonomy, as the categories themselves are moving targets.

In linguistic terms, the evolution of disciplines also manifests in lexical and semantic change, as specialized vocabularies develop in response to emerging knowledge domains (Halliday and Martin, 2003; Ure, 1982). These sublanguages or registers are closely tied to socio-cultural and cognitive developments in the respective scientific communities. Against this backdrop, computational tools offer promising avenues for tracing how such categories evolve over time. In particular, the rise of large language models (LLMs) has

1 https://support.qs.com/hc/en-gb/articles/4539968720924-QS-Subject-Categorisation

2 https://support.qs.com/hc/en-gb/articles/4406036892562-All-Science-Journal-Classifications-ASJC-Codes

introduced new possibilities for the automated classification of scientific texts and for mapping disciplinary development on a large scale.

Table 1: Overview of subject area counts in major classification frameworks

	Broad area	Subject
QS	5	56
ASCJ	-	388
FORD	6	42

This study investigates whether and how LLMs can be applied to classify scientific texts into disciplines, using the Royal Society Corpus (RSC; Fischer et al., 2020), a diachronic corpus spanning over 300 years of scientific publication. By employing a fine-tuned LLM (Hermes-2-Pro-Llama-3-8B), we assess the model's capacity to assign discipline labels to historical scientific articles from 1665 to 1996, based on a predefined taxonomy. Building on previous research on topic modeling in historical corpora (Bizzoni et al., 2020; Menzel et al., 2021) and guided by the evolving structure of the Royal Society's own disciplinary divisions, we aim to inspect diachronic patterns of specialization and interdisciplinarity in scientific discourse.

In doing so, the paper contributes to two lines of inquiry. First, it provides a computationally grounded method for the historiography of science, enabling the large-scale tracking of how the prevalence and relationships of disciplinary identities shift over time. Second, it evaluates the reliability and limitations of using a fine-tuned LLM for historical text classification, a task complicated by lexical and semantic change (Halliday and Martin, 2003). By clearly defining our taxonomic framework and applying it consistently, we aim to shed light on the dynamic architecture of scientific knowledge while also highlighting the critical methodological considerations required when using AI tools in historical research.

2. Disciplines in the Royal Society Philosophical Transactions and Proceedings

Brought to existence in 1665, the *Philosophical Transactions* of the Royal Society of London, joined two centuries later by the *Proceedings*, was the world's earliest periodical dedicated explicitly to the dissemination of high-quality scientific knowledge. In its early stages, the *Transactions* consisted of letter-like observations, travel accounts, experiment reports, and book reviews and read more like a patchwork of intellectual curiosities with no clear disciplinary boundaries than a modern scientific journal. Astronomy, medicine, biology, chemistry, physics, and an eclectic array of other subjects coexisted on its pages.

Throughout the following decades and centuries, the *Philosophical Transactions* experienced substantial changes in its organization, procedure for paper selection and paper format, yet these changes had little to no influence on the notion of discipline in the jour-

nal. Well into the first half of the 19th century, it remained a smorgasbord of topics, where works on optics or anatomy of the eye were directly followed by an "Account of a Monstrous Lamb" or an "Account of an Elephant's Tusk, in Which the Iron Head of a Spear was Found Imbedded".

However, the growing specialization of scientific knowledge, coupled with the rapid development of new techniques and methods throughout the 19th century, gradually rendered the journal's generalist approach untenable. Adding to this pressure was the emergence of newer, field-specific periodicals, which began to compete with the *Philosophical Transactions and Proceedings* both for audience and contributors. A natural, if somewhat delayed, response to this new reality in the scientific world was the eventual division of the journal into Sections A and B, devoted respectively to the mathematical & physical sciences and biological sciences.

The branching of scientific inquiry found further expression in the establishment of six Sectional Committees, each dedicated to a major domain: Mathematics, Botany, Zoology, Physiology, Chemistry & Physics, and Geology. These committees assumed the task of reviewing incoming submissions, embodying the growing need for disciplinary expertise. Today, the Royal Society recognizes eleven such Sectional Committees, spanning Computer Sciences, Mathematics, Astronomy and Physics, Chemistry, Engineering and Materials Science, Earth and Environmental Sciences, Molecules of Life, Cell Biology, Multicellular Organisms, Patterns in Populations, and Health and Human Sciences. While these committees no longer oversee peer review, they nonetheless offer a revealing insight into how the Royal Society continues to structure and conceptualize disciplinary boundaries—at least at their broadest contours.

Turning again to publishing practices, it is interesting how the continued evolution of science in the 20th and 21st centuries has left the overall structure of the *Philosophical Transactions and Proceedings* largely untouched. Their bisectional division has persisted into the present day, with new disciplines and emerging fields of study absorbed into one section or the other. On its website, the Royal Society draws a distinction between Disciplines and Subjects within the two sections: the former denote broad, overarching domains of knowledge, while the latter serve as more granular, often interdisciplinary tags used to describe individual papers.

The disciplines in both sections are summarized in Table 2. The inclusion of Biology and the creation of a separate "discipline" for cross-disciplinary research within Section A as well as the appearance of Mathematics in Section B are particularly striking, illustrating the difficulties inherent in assigning a single disciplinary label to research that might be actually equally rooted in multiple scientific domains and highlighting potential challenges related to automatic classification of disciplines.

However, providing information on document domain is crucial for large collections of data. An attempt to accomplish this task for Royal Society texts was made by Menzel et al. (2021) who trained a probabilistic topic model using the Royal Society Corpus (RSC; Fischer et al., 2020). The model used an unsupervised machine-learning approach that scanned the documents in the corpus and identified different topics based on words occurring together in similar contexts. Then, labels were assigned manually after the examination of the most distinctive words of each topic. This procedure resulted in a hierarchically organized system of topics illustrated in Figure 1.

Figure 1: Topics in the Royal Society Corpus as identified by an unsupervised topic-modelling approach

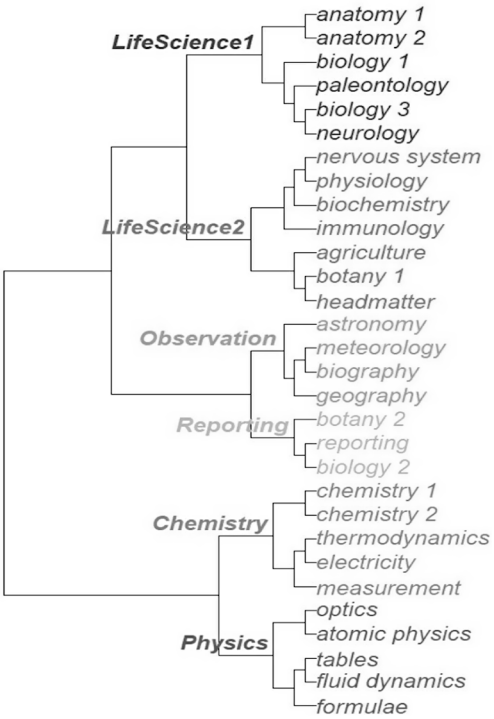


Figure taken from Menzel et al. (2021)

Table 2: Disciplines in Sections A and B of the Philosophical Transactions and Proceedings according to the Royal Society website.

Section	**Disciplines**	**Number of Papers (Transactions + Proceedings)**
Section A	Astronomy	503
	Biology	15
	Chemistry	946
	Computer Science	1,246
	Cross-Disciplinary Studies	13
	Earth Sciences	1,164

Section	Disciplines	Number of Papers (Transactions + Proceedings)
	Engineering and Technology	2,595
	Mathematics	3,153
	Physics	3,903
Section B	Biology	15,568
	Mathematics	1

The appearance and rapid development of LLMs in recent years offers researchers and the DH community in general new opportunities in many NLP tasks. Evidently, this applies to metadata generation and text classification as well. We explored the potential of LLMs in this field in a case study described in the next section.

3. Case study: using LLM to classify texts into scientific disciplines

In this case study, we aimed to analyze the ability of a large language model to classify scientific texts spanning from 1665 to 1996 according to a predefined list of disciplines. The task is challenging because the concept of discipline is fluid and undergoes constant change over time. However, this type of classification can still be useful for studies in the history of science and in other research areas, for example, in linguistics, to analyse differences within the scientific register.

Our aim is to provide an overview of the distribution of disciplines over time and to identify potential areas of interdisciplinarity.

3.1 Methodology

In our study, we use the Royal Society Corpus (RSC; Fischer et al., 2020), a diachronic corpus of scientific English spanning the period from 1665 to 1996. The corpus is based on the Philosophical Transactions and Proceedings of the Royal Society of London, and comprises 47,837 texts (295,895,749 tokens). The table in Figure 2 shows the overall distribution of texts across different periods of the RSC.

Regarding the LLM, we used Hermes-2-Pro-Llama-3-8B[3], a fine-tuned version of the Llama 3 language model family. Based on the 8B-parameter version of Llama 3, this model has been instruction-tuned and further optimized for producing structured outputs, such as JSON and YAML. It outperforms earlier models like Llama 2 and Mistral, particularly in tasks requiring clear, machine-readable responses. Its enhanced ability to

3 https://huggingface.co/NousResearch/Hermes-2-Pro-Llama-3-8B

follow instructions and generate well-formatted outputs makes it especially suitable for classification tasks.

Figure 2: Number of Texts and Tokens per Period in the RSC.

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

As is common in prompt-based workflows, we defined a system prompt to describe the task, while the user prompt contained the information and data from the texts to be classified. Although our main focus in this study is the classification of RSC texts by discipline, the prompt used included four tasks: (a) suggesting an alternative title; (b) providing a TL;DR[4] summary; (c) identifying the main topics of the paper; and (d) determining the primary discipline (from a predefined list) along with a suitable sub-discipline.

The system prompt is structured as follows:

1. Role: Act as a librarian
2. Objectives: Read, analyze, and organize
3. Input: Texts from the RSC
4. Tasks:
   a. Provide an alternative title
   b. Generate a TL;DR summary
   c. Identify main topics
   d. Determine primary discipline and sub-discipline
5. Examples: Sample input and output

The user prompt, on the other hand, follows an iterative process, providing information about the text to be analyzed:

1. Article ID
2. Author(s)
3. Journal
4. Year
5. Title
6. Text snippet (i.e., the first 1,000 words of the paper)

4 "Too Long; Didn't Read"

The detailed scripts, including prompt configurations, are available on GitHub[5] for reference.

Given the ever-evolving nature of scientific disciplines, using a predefined list helps avoid an overly fine-grained classification. Therefore, based on the information described in Section 2 about the RSC (i.e., its history and the topic-modeling classification by Menzel et al. (2021)), we defined the following list:

1. Physics
2. Chemistry
3. Environmental and Earth Sciences
4. Astronomy
5. Biology and Life Sciences
6. Medicine and Health Sciences
7. Mathematics
8. Engineering and Technology
9. Social Sciences, Art, and Humanities

Requesting the LLM to provide the main topics and a subdiscipline categorization for each text is also intended to enhance the classification, as some subdisciplines in the corpus may be highly recurrent and could help refine and expand the predefined list.

3.2 Evaluation

The evaluation of the LLM's output shows a high level of reliability, with 99.42% (47,558 out of 47,837) of the generated responses being valid YAML formats, indicating strong respect of the required structured output.

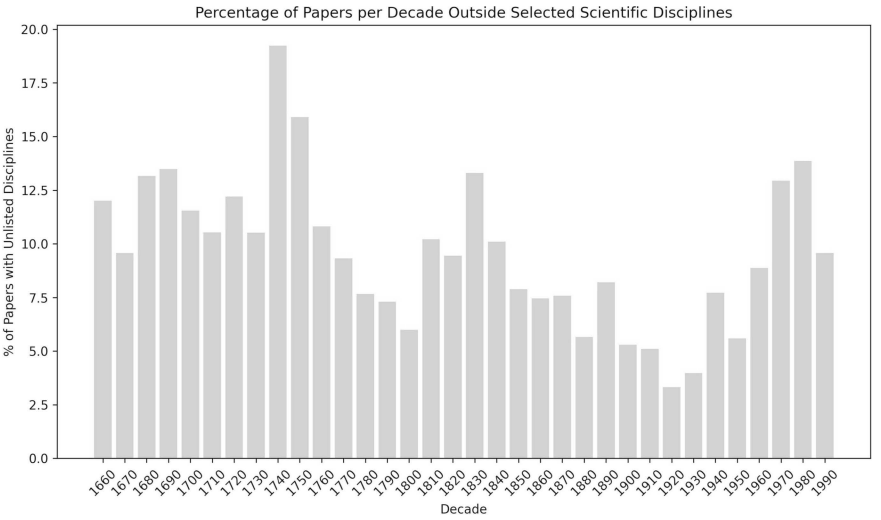
Regarding the classification accuracy, 90.95% (43,257 out of 47,558) of the predicted scientific disciplines matched those in the predefined list, demonstrating the model's effectiveness in assigning appropriate categories. However, some discrepancies were observed, such as labeling *Earth Sciences* instead of the broader *Environmental & Earth Sciences* and the occasional generation of novel categories, like *Music*. These findings suggest that while the model performs well overall, there is room for refining category definitions and handling outliers.

The *Music* category, for example, was used to classify seven articles: one in the subdiscipline *Acoustics* and the other six in *Theory*. Two papers are from the 18th century (one from the beginning and one from the end), while the rest are from the second half of the 19th century. The themes vary. One article presents a method for tuning musical instruments using precise mathematical proportions to achieve perfect harmonies. Another examines the conditions, extent, and realization of a perfect musical scale on instruments with fixed tones. There is also a discussion of the Hindoo division of the octave and its implications for music theory. Thus, while the LLM's reasoning was accurate, it considered the topics too specific to classify these papers within the broader category of *Social Sciences, Arts, and Humanities*.

5 sci_scope/llm_code at main · badrex/sci_scope · GitHub

Figure 3 shows the distribution by decade of RSC articles, indicating the percentage of articles published in each decade that were categorized under disciplines not included in our predefined list.

Figure 3: Proportion of articles published in each decade of the RSC classified under disciplines not covered by the predefined categories.



It is possible to observe better performance by the LLM in the second half of the 18th century, which may be attributed to the clearer delineation of disciplines following the changes brought about by the scientific revolutions. After a slight increase in hallucinations during the first half of the 19th century, performance improves again. However, in the second half of the 20th century, the number of papers classified under disciplines not included in the predefined list increases once more. These peaks may also reflect periods of greater interdisciplinarity, possibly associated with the emergence of new disciplines.

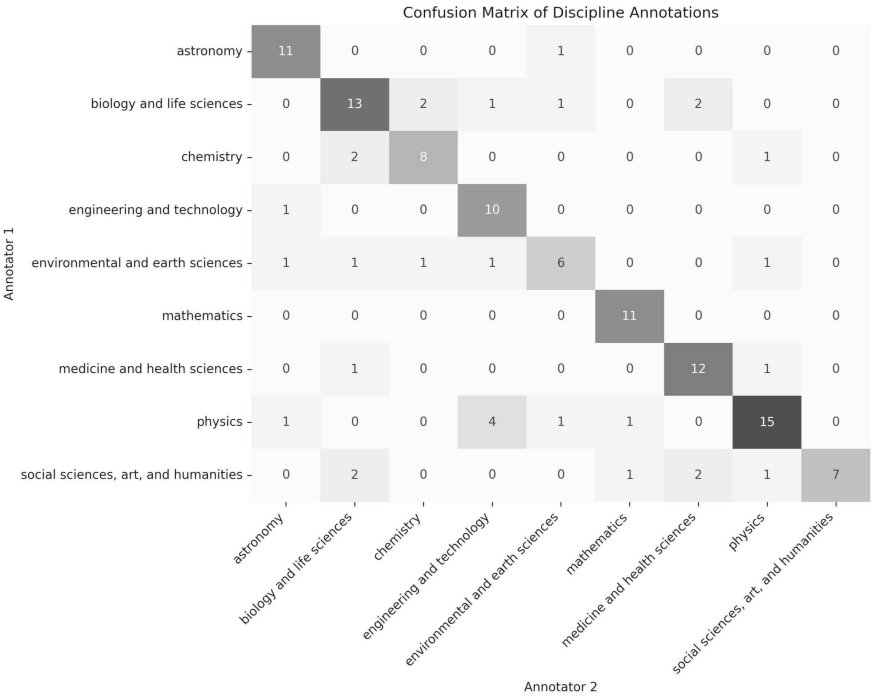
Regarding the texts classified into the nine predefined categories, we also conducted an accuracy evaluation of the LLM's classification. First, we randomly selected 15 texts from each discipline category identified by the LLM, ensuring that at least two texts from each 50-year period of the RSC were included. These texts were then labeled according to the list of disciplines by two German annotators, both university students in the humanities and fluent in English.

To evaluate the reliability of the manual annotations, we calculated inter-annotator agreement using Cohen's Kappa coefficient, a statistical measure that accounts for the degree of agreement between two annotators while correcting for chance agreement. The annotation campaign resulted in a Cohen's Kappa of 0.72, which indicates substantial agreement, indicating a high level of consistency between annotators. Figure 4 presents the confusion matrix indicating the disagreements between the 2 annotators.

Some expected discrepancies between annotators can be observed. For instance, four papers labeled as *Physics* by Annotator 1 were classified as *Engineering and Technology* by

another annotator. Differences in annotation are also noticeable in the fields of *Biology*, *Medical and Health Sciences*, and *Chemistry*.

Figure 4: Confusion matrix showing the agreements and disagreements between the two annotators across the assigned labels.

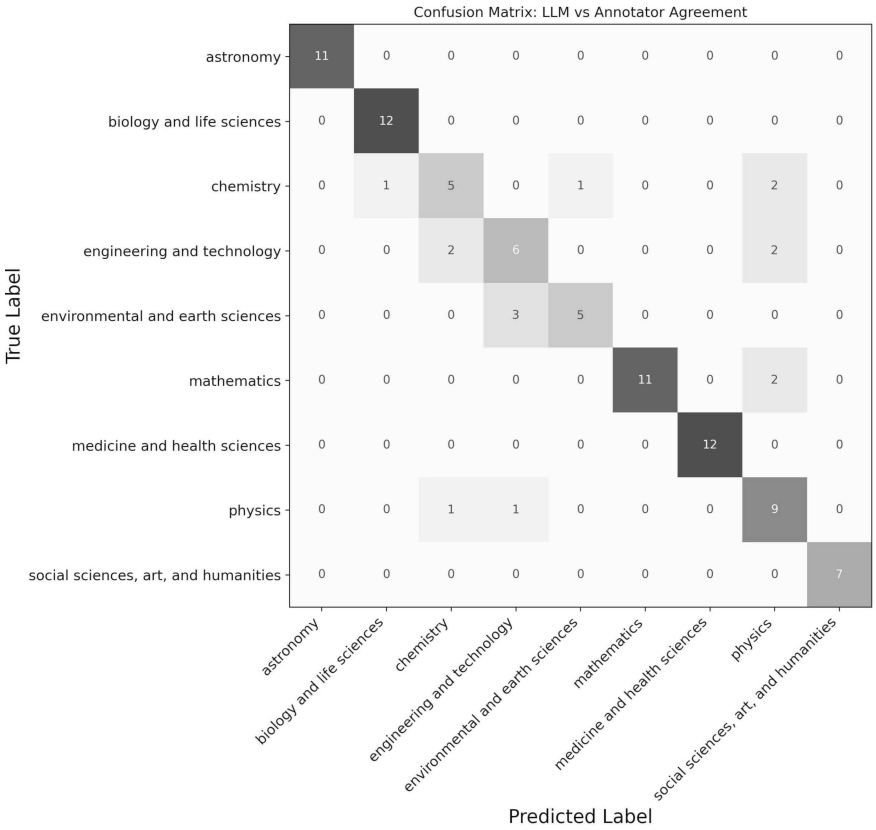


To evaluate the consistency of discipline classification by the LLM, we compared the output from our two annotators against labels provided automatically. First, we calculated the accuracy by considering a classification correct if the reference label matched the label assigned by at least one of the two annotators, resulting in an accuracy of 82.22%. Additionally, we computed accuracy only on the subset of cases where both annotators independently agreed on the discipline, resulting in a slightly higher accuracy of 83.87%. These results demonstrate the potential of using LLMs for discipline categorization of texts, even with a diachronic dataset spanning three centuries of scientific publications.

Figure 5 presents the confusion matrix comparing the LLM labels to the categories defined by the annotators, considering only cases where both agree.

The disciplines that appear most challenging for the LLM are Chemistry, Engineering and Technology, and Physics. This may be due to the application of some concepts from the two more theoretical disciplines within practical and technological contexts in some papers.

Figure 5: Confusion matrix showing the agreements and disagreements between the LLM (Predicted label) and the annotators (True label).

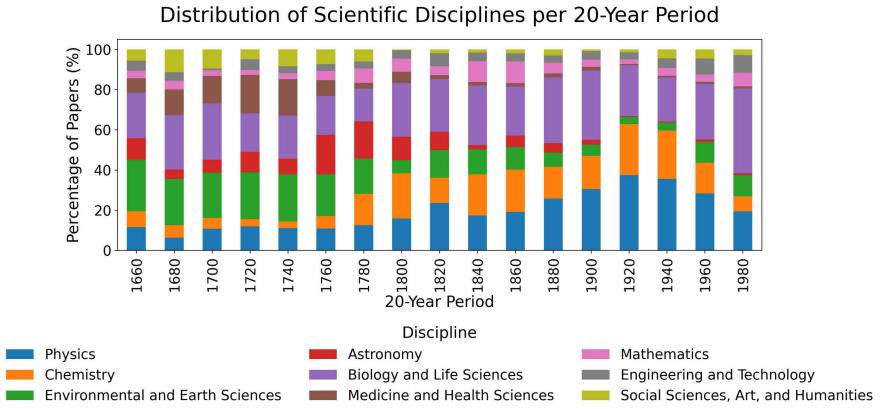


3.3 Analyzing interdisciplinarity

Figure 6 presents the distribution of the number of RSC papers per discipline, grouped by 20-year periods.

The graph in Figure 6 indicates a greater diversity of disciplines represented in publications from the earlier periods of the RSC (up to the beginning of the 19th century). *Environmental and Earth Sciences* and *Medicine and Health Sciences* were relatively well represented during this time, but their presence diminished over the years. However, *Environmental and Earth Sciences* shows a slight increase in the final two 20-year periods, possibly linked to growing scientific awareness of environmental issues and ecology-related research. The impact of the Chemical Revolution (late 18th century) is evident in the increased number of chemistry-related publications from that period onward. Moreover, an increase in physics publications is also noticeable in the early 20th century, reflecting the discipline's major transformations with the development of quantum theory and Einstein's theory of relativity.

Figure 6: Distribution of RSC Papers by Discipline (20-Year Periods)



To analyze interdisciplinarity, we used the TL;DR summaries generated by the LLM to create text vectors, using the FastEmbed library dictionary[6] with the BAAI/bge-small-en-v1.5 model, for all RSC papers classified under one of the nine predefined disciplines. These summaries were chosen to minimise the influence of text length and linguistic features specific to different historical periods. We then applied a t-SNE projection to plot the resulting vectors in the two-dimensional graph shown in Figure 7.

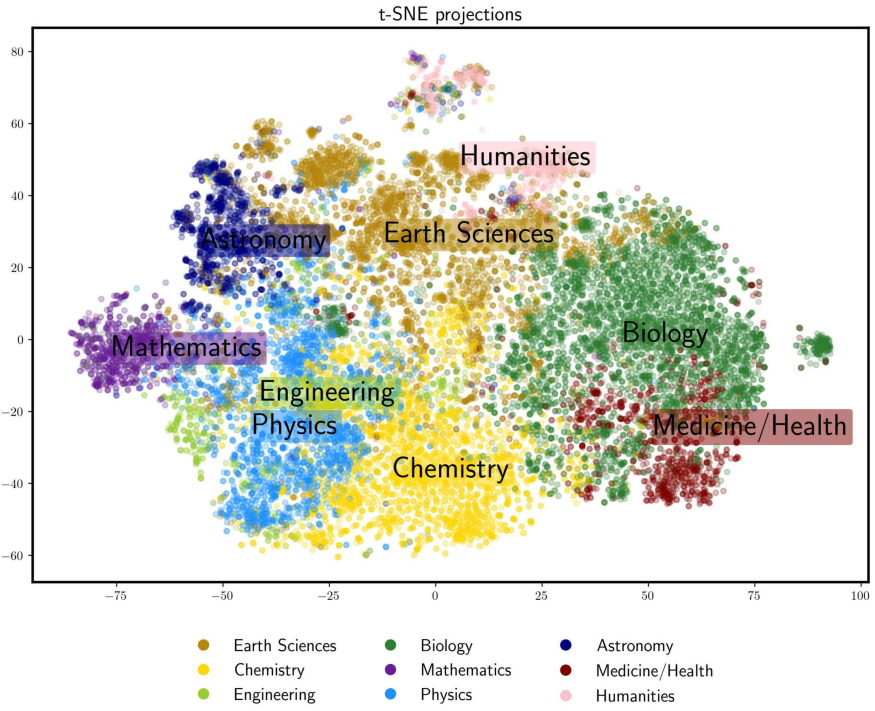
Although discipline clusters can be identified, their boundaries are not clearly defined. *Astronomy* and *Mathematics* appear more distinct, though signs of interdisciplinarity are visible, particularly between *Astronomy* and *Physics*, and between *Mathematics* and *Earth Sciences*. *Chemistry* is positioned between *Physics* and *Biology*, showing a broad area of overlap. As expected, *Biology* is also closely interconnected with *Medicine and Health Sciences*, and *Physics* with *Engineering*.

This large overlap between Chemistry and Biology suggests the potential need to add Biochemistry as a tenth discipline in the list. However, Biochemistry was not recognized as a distinct scientific discipline until the late 19th to early 20th century. Thus, depending on the research question, this could introduce a bias in the analysis if texts are taken from a limited period onward.

These results demonstrate that using LLMs to classify scientific publications into predefined disciplines can be a valuable tool for analyzing interdisciplinarity. This approach can be extended to a more detailed diachronic analysis, for example by comparing different periods, revealing how disciplines and their intersections evolve over time and providing a foundation for the emergence of new disciplines. Moreover, the method of Aguilar-Valdez et al. (2026), originally applied to the entire RSC, can also be used to trace conceptual shifts within individual disciplines.

6 https://github.com/qdrant/fastembed

Figure 7: t-SNE projection of RSC texts based on vectors generated from TL;DR summaries. Each color represents one of the nine predefined disciplines.



4. Conclusions

The classification of scientific disciplines is a complex and dynamic process, shaped by historical developments, institutional structures, and evolving research paradigms. This study shows that the boundaries between disciplines are fluid, with fields continuously fragmenting, merging, and redefining themselves in response to new knowledge and societal needs. The case study of the Philosophical Transactions and Proceedings of the Royal Society illustrates how scientific publishing has adapted to increasing specialization.

Using large language models (LLMs) to classify historical scientific texts presents both opportunities and challenges. While the model in this study achieved high accuracy in assigning predefined discipline labels, discrepancies arose due to the inherent ambiguity in disciplinary boundaries and the emergence of new fields over time. The analysis also revealed significant interdisciplinary overlaps, particularly between chemistry, biology, and physics, suggesting that rigid classifications may not always capture the true nature of scientific inquiry.

These findings highlight the importance of carefully defining disciplinary frameworks when conducting digital humanities research. Researchers must consider the historical context, the granularity of classification, and the potential biases introduced

by automated systems. Future work includes testing other LLMs, refining discipline lists (with different predefined lists across various time periods), incorporating subdisciplinary distinctions, and exploring diachronic trends in interdisciplinarity.

Ultimately, this study underscores the value of computational methods for tracing the evolution of scientific knowledge while emphasizing the need for critical, context-aware approaches to classification.

5. Limitations

Historical texts might be underrepresented in our model's pre-training data, which might have negatively impacted the classification results for earlier periods. Moreover, while the model was fine-tuned for structured output generation and instruction following, it was not fine-tuned for discipline classification. Although we prompted the model in a few-shot setting, providing relevant examples, an additional task-specific fine-tuning might improve the classification results.

Because our prompt required both classification and generation, we opted for an encoder-decoder model as a unified solution. However, for classification alone, a fine-tuned encoder-only model might offer better results.

We defined our list of disciplines based on the historical evolution of the Transactions and Proceedings of the Royal Society and on the previous topic modelling results. However, our labels fail sometimes to account for highly interdisciplinary texts, resulting in misclassifications and model hallucinations. Moreover, applying a modern disciplinary framework to historical texts can introduce bias, as many current fields were not conceived as independent disciplines in the past.[7]

References

Abbott A (2001) *Chaos of Disciplines*. Chicago: University of Chicago Press.

Aguilar-Valdez S, Phan-Tắt B, Speelman D, et al. (2026) Discursive parallels of the chemical revolution. Topic modelling and distributional analysis. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Alves D, Tadić M and Rehm G (2024) Which Domains, Tasks and languages are in the focus of NLP research on the languages of Europe? In: *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability @ LREC-COLING 2024* (eds F Gaspari, J Moorkens, I Albade et al.), Torino, Italia, 25 May 2024, pp. 18–32. Totina: ELRA and ICCL.

7 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Fischer S, Knappen J, Menzel K, et al. (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* (eds. N Calzolari, F Béchet, P Blache et al.), Marseille, France, 11–16 May 2020, pp. 794–802. Marseille: European Language Resources Association (ELRA).

Frickel S and Gross N (2005) A general theory of scientific/intellectual movements. *American Sociological Review* 70(2): 204–232.

Halliday MAK and Martin JR (2003) *Writing Science: Literacy and Discursive Power*. London: Routledge.

Menzel K, Knappen J and Teich E (2021) Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics* 9(1): 1–18.

OECD (2015) *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. Report, 8 October. Paris: OECD Publishing.

Ure J (1982) Introduction: approaches to the study of register range. *International Journal of the Sociology of Language* 1982(35): 5–24.