TFG

LLM-Based forecasting of scientific papers impact based on abstract

Directed by Ivan Erill Sagales (Area of Computer Science and Artificial Intelligence)

Data Engineering

Universitat Autònoma de Barcelona

Barcelona, June 2025

Jaime Rueda Diví

1566320

Content

1. Introduction – Problem & Context

1.1 Scientific publishing landscape & impact factor relevance

Scientific publishing is a multimillion-dollar industry. Under the current Open Access model, articles are freely available to readers, but authors are often required to pay high Article Processing Charges (APCs) to publish their work. While several Diamond Open Access initiatives have emerged to eliminate author fees, their adoption remains limited. One key barrier is the absence of established editorial structures and journal prestige — factors that authors often rely on to signal quality and impact.

Journal prestige is frequently tied to citation-based metrics, especially the Impact Factor, which reflects how often articles in a journal are cited over time. However, citations accumulate slowly, and authors publishing in newer or less-recognized journals may struggle to demonstrate early impact. In contrast, traditional journals offer immediate visibility and institutional recognition due to their established reputation and editorial processes.

Despite their influence, conventional publishing and review systems have been criticized for slow turnaround times, inconsistent peer review, and structural inefficiencies that delay the communication of scientific findings. These issues are particularly problematic in fast-paced research environments and raise the question of whether alternative, AI-assisted models could help address some of these challenges.

1.2) Motivation for prediction

In the current publishing ecosystem, there is a growing need for tools that can estimate the potential influence of scientific articles without waiting for citations to accumulate. Traditional impact assessment is slow and relies on post-publication citations while this tool could give a prediction even before being published. The ability to forecast a paper's impact at the time of publication could help in many different ways. It could accelerate editorial decisions, decide funding allocations, enhance research research visibility, etc. Moreover, it would give Diamond Open Access journals the editorial prestige of the already well-stablished ones.

One of the most fundamental perks of this tool would be an even playing field for everyone trying to publish, abandoning traditional editor decisions whose standards aren't public and could be different depending on the author/authors of the study.

Finally, it is a new way of exploring the powerful tools that are LLMs, especially for text-based prediction tasks which they are known to be strongly reliable.

1.3) Statement of the problem

This project addresses the problem of early-stage impact estimation for scientific publications. Specifically, it explores whether large language models, when prompted effectively, can predict the relative impact of a paper—measured through SCImago Journal Rank quartiles—based solely on abstract-level information. The feasibility and reliability of such predictions remain largely unexplored, particularly in the context of journals and articles not seen during the model's training.

1.4) Objectives

- Provide a prediction for the impact a set of papers of a specific topic will have.
- Investigate several prompting engineering techniques and strategies.
- Evaluate the results and prove if AI can be either an assistant or a substitute for traditional editors.

1.5) Scope & Limitations

This study focuses on evaluating the capacity of GPT-4o to predict the SCImago Journal Rank (SJR) quartile of journals publishing scientific articles related to bacteriology, using only the abstract text as input. All tests were conducted on a curated subset of publications extracted from PubMed using a domain-specific query, limited to articles published between 2020 and 2024. The primary task is framed as a classification problem, where the model is asked to predict the SJR quartile (Q1–Q4) of the journal in which each article appeared. A secondary goal is to assess how different prompt engineering strategies affect prediction accuracy within this specialized scientific field.

The scope of this work is intentionally limited to bacteriology to ensure topic consistency and reduce variability in language patterns and citation behaviors. This narrow focus enables more precise evaluation of the model's reasoning capabilities within a well-defined domain, but it also constrains the generalizability of the findings. The results may not be transferred to other scientific disciplines, particularly those with different publishing norms, terminology, or citation dynamics.

Additional limitations arise from the use of GPT-4o, whose knowledge cutoff is October 2023. Articles published in 2024 were included in the test set to ensure the model could not have "seen" them during training, but this also means the model lacks awareness of journal developments or reclassifications that occurred after its cutoff date. Moreover, the model was used in its general-purpose form, without fine-tuning or domain-specific retraining. Finally, while SCImago's quartile ranking provides a more robust metric than raw citation counts, it remains a journal-level indicator and may not fully reflect the impact of individual articles.

2. Background / State of art

2.1) Traditional Impact Metrics and their limitations

The impact of scientific research has historically been evaluated through citation-based metrics, with the Journal Impact Factor (JIF) being the most prominent. Introduced by Eugene Garfield in the 1960s, the JIF calculates the average number of citations received by articles published in a journal over a two-year period. This metric has become widely adopted as a proxy for journal prestige, influencing where researchers choose to publish and how research output is evaluated by institutions, funders, and hiring committees.

Despite its widespread use, the JIF has been subject to extensive criticism due to both methodological and conceptual shortcomings. Chief among these is the assumption that citation counts are a direct reflection of a paper's scientific quality or influence. However, citation behavior is shaped by numerous external factors that often have little to do with the intrinsic value of a publication. For instance, review articles and methodological papers generally attract more citations due to their broad applicability, while original research in emerging or specialized fields may be cited less frequently, regardless of significance.

Further distortions arise from strategic citation practices, including excessive self-citation, citation rings, and the deliberate promotion of certain articles to boost journal metrics. These behaviors can artificially inflate a journal's impact factor, misrepresenting its true scholarly value. Moreover, the JIF aggregates data at the journal level, meaning that a small number of highly cited articles can disproportionately raise a journal's average, masking substantial variability in article-level impact.

Another critical limitation of citation-based metrics is their inherent time-lag. Citations accrue slowly, often taking years to reflect the actual influence of a work. This delay makes such metrics unsuitable for early-stage evaluation, particularly in contexts where timely assessment is essential—such as in funding decisions, preprint screening, or fast-moving disciplines. Papers may also experience citation spikes due to controversy or novelty, only to be later discredited, highlighting the volatility of early citation signals.

Lastly, reliance on the JIF reinforces structural inequalities in the publishing ecosystem. High-impact journals attract more submissions, citations, and attention, thereby maintaining their elite status in a self-reinforcing cycle. This dynamic can marginalize newer or regional journals, limit the diversity of scholarly discourse, and constrain innovation by incentivizing publication in a narrow set of prestigious venues. These systemic flaws have led to growing calls for alternative metrics that offer more equitable, reliable, and timely measures of research impact.

2.2) Scimago Journal Rank

The SCImago Journal Rank (SJR) is a journal-level metric developed by the SCImago Research Group, designed as an alternative to the traditional Impact Factor for assessing the influence and prestige of academic journals. Based on data from the Scopus database, SJR calculates not only the number of citations a journal receives but also considers the quality and reputation of the citing journals. It uses a PageRank-like algorithm to weight citations, such that a citation from a highly ranked journal contributes more to a journal's score than one from a lesser-known source. This approach seeks to correct the inflationary biases seen in raw citation counts and aims to provide a more nuanced and field-normalized view of scholarly influence.

One of the primary advantages of SJR is its ability to mitigate citation gaming and field-based disparities. By accounting for the prestige of the source of citations and normalizing across disciplines, SJR allows for more meaningful comparisons between journals in different scientific areas. It also provides quartile rankings (Q1 to Q4), making it easier to interpret journal standing at a glance, and can support more informed publication decisions by authors and institutions alike. However, SJR is not without limitations. Like other bibliometric indicators, it still aggregates impact at the journal level rather than the article level, which may obscure the quality or influence of individual papers. Moreover, its reliance on the Scopus database means it inherits any coverage biases present in that corpus, potentially underrepresenting certain regions, languages, or newer journals.

Despite these limitations, SJR remains one of the more robust and interpretable alternatives to the Impact Factor. It balances citation volume with citation quality and provides a transparent, multi-dimensional measure of journal influence that is especially useful in comparative and evaluative contexts. When used alongside other indicators and qualitative assessments, it contributes meaningfully to a more equitable and informed understanding of scholarly impact.

2.3) Generative AI models in prediction tasks

While there have been researches in impact prediction using Machine Learning models and neural networks, relatively little research has been done making use of Generative AI systems.

Generative AI systems are trained on a massive corpus drawn from the Internet at a given date. These systems can handle complex prompts and generate statements about the putative relevance and impact of a journal article, based on its abstract, at the time of its publication. This project seeks to design advanced prompts and systematically evaluate the performance of generative AI systems at predicting the impact factor of the journal in which scientific articles published after the known training date of the generative AI system were published, thereby providing an AI-based alternative to the role of editors in traditional scientific journals.

Large language models (LLMs) have increasingly been used in prediction tasks beyond text generation, including domains such as medical triage, legal risk assessment, and academic peer review support. These models, trained on massive corpora of unstructured data, can infer complex patterns and make probabilistic predictions based on input prompts. Unlike traditional models that require structured features, LLMs can evaluate free-text inputs—such as scientific abstracts—to estimate characteristics like novelty, relevance, or potential impact, even without explicit numerical training.

However, their predictive power has clear limitations. Generative models rely on statistical associations rather than causal reasoning, making them prone to errors in tasks requiring arithmetic accuracy or real-world judgment. They may also reproduce biases present in their training data. Despite this, when carefully designed and prompted, generative AI offers a scalable way to forecast qualitative outcomes—such as the perceived importance of a research article—making them a valuable complement to human editorial decision-making in high-volume or time-sensitive evaluation settings.

2.4) Prompt Engineering

Prompt engineering refers to the deliberate design and structuring of inputs (prompts) to guide the behavior of large language models (LLMs) toward desired outcomes. According to The Prompt Report, this encompasses a taxonomy of 58 distinct techniques, including zero-shot, few-shot, chain-of-thought, decomposition, ensembling, and self-criticism, which fall into six major categories. Effective prompt engineering involves not only selecting appropriate strategies but also refining the wording, format, and context to improve model performance through iterative experimentation .

Although powerful, prompt engineering remains largely empirical rather than principled, with outcomes varying significantly across models and tasks. The Prompt Report emphasizes the need for standardized terminology and structured methodologies, noting that subtle changes in phrasing or template design can produce dramatically different results. In high-stakes applications—such as scientific forecasting or editorial assessments—systematic evaluation of prompting strategies is critical to ensure reliability, reduce ambiguity, and improve reproducibility.

2.5) Positioning in the field of study

This project marks an initial step into the emerging field of using large language models (LLMs) for scientific impact prediction based solely on abstract content. Rather than offering a definitive methodology, it serves as an exploratory proof of concept to test the feasibility of prompt-based inference using GPT-4o. Given the model's limitations and the complexity of academic publishing, the aim is not to replace editorial judgment or established metrics, but to contribute to early discussions on how generative AI might support or accelerate aspects of research evaluation. Future work is needed to refine the approach, validate it across disciplines, and assess its broader implications.

3. Design / Methodology

3.1) Dataset description:

To evaluate the ability of a large language model (LLM) to predict the perceived impact of scientific articles, a domain-specific dataset was curated from PubMed. The dataset focused on a narrow but scientifically rich topic area: transcriptional regulation in bacteria. Articles were selected using the query: "bacteria transcription, regulator, repressor, activator, promoter", which ensured that the texts centered on a coherent subject, enabling the model to develop contextual understanding across the training corpus.

Each article record consisted of the title, abstract, and the SCImago Journal Rank (SJR) quartile of the publishing journal, as assigned for the article's year of publication. Quartile labels were determined using the SCImago Journal Rankings corresponding to the publication year (e.g., 2020 SJR for 2020 papers), offering a time-sensitive and field-normalized measure of journal prestige. The training set included 187 articles published between 2020 and 2023, which fell within the knowledge cutoff of the LLM used (GPT-4o, October 2023). To provide broader contextual grounding, a set of 7 review articles on the same topic—containing only title and abstract—was added to the training set. These reviews were intended to help guide the model's understanding of what constituted relevant or significant research within the domain during those years.

The test set consisted of 38 articles published in 2024, sharing the same structure (title and abstract) but without an assigned SJR quartile. This allowed for evaluation of the LLM's performance on unseen, post-cutoff data. The validation set used the same 38 articles but included their actual 2024 SJR quartile labels, which had been retrieved from the SCImago database. This setup enabled a fair assessment of the model's predictions relative to current journal rankings.

To address class imbalance—common in SCImago quartiles due to the overrepresentation of Q1 and Q2 journals in the domain—a balancing strategy was implemented. This involved reclassifying journals based on SJR score thresholds, ensuring a more even distribution across quartiles and making the classification task more challenging and meaningful. Without this adjustment, the model might have learned to default to high-quartile predictions, reducing its discriminatory power.

In the early stages of the project, alternative impact measures were considered. The original plan involved using citation counts and the "Highly Influential Citations" metric from Semantic Scholar as predictive targets. Early versions of the dataset contained these values instead of SJR quartiles. However, due to their temporal delay, volatility, and the limited availability of citations for recent publications, the project shifted to using SJR quartiles as a more stable and interpretable classification target.

3.2) Model Setup

This study employed GPT-4o, a state-of-the-art large language model developed by OpenAI, as the predictive engine for estimating the impact of scientific papers based on their abstracts. The model was accessed via its API interface and operated in a prompt-based inference setting. The model's knowledge cutoff was October 2023, ensuring that it had no access to the 2024 test set data during training, thus preserving the integrity of the evaluation.

The prediction task was formulated as a multiclass classification problem, in which the model was asked to assign each input abstract to one of four SCImago Journal Rank (SJR) quartiles: Q1, Q2, Q3, or Q4. These quartiles represent relative journal prestige within their respective fields, as determined by SCImago's field-normalized rankings. The labels were drawn from historical SJR data corresponding to the year of publication, while the predictions were based solely on the abstract and title content.

Prompt Engineering Strategies

To guide the model toward accurate classification, a series of prompt engineering strategies were tested. These involved varying the structure, and instructions presented in the input prompt. They consisted of large prompts that were sent in chunks due to GPT-4o max limit in Tokens Per Minute (30k TPM) but were carefully designed so that the prompt and answer would not go past the maximum limit of window size of GPT-4o (128k tokens).

When prompting through the Open AI API, first a “system” message is required. It gives context and is used to define the role and what behavior is expected from the model before any input is given. For this project it was defined to be a Scientific Paper Editor whose role were to be to classify scientific papers based on their abstract and what type of answer and it what format was expected to be received. Temperature is set to 0 so the model is as deterministic as possible.

- **System Prompt:**

"role": "system", "content": (

"You are a Scientific Paper editor. Predict the SJR Quartile of given papers based on training for examples and using reviews as context of what is relevant."

"Answer only with the Title and predicted Quartile (Q1 to Q4). Format: Title: xxx \n Quartile: xxx"

"Q1 for the most relevant and Q4 for the least")

- **Simple Prompt:**

  f"Context: I will give you a list of articles in format: Title: xxx , Abstract: xxx , SJR Quartile: xxx \n. Read the content in: \n\n{train}. \nThis are to help your predictions."

  f"Extra context: I will give a list of Reviews in format: Title: xxx, Abstract: xxx. They explain relevant matters on subjects related to the papers. Read the content in \n\n{reviews}. Also a test set for you to do the predictions in {test}"

For more complex strategies the user prompts were structured differently. The training set consisted of 76174 tokens, the review 1333 tokens and the test set 15150 tokens. As they can’t fit in a single prompt they were reduced to 20k tokens chunks that were sent every minute. After having sent both training and reviews set, the user prompt was included in the last chunk with the test set.

- **Chain-of-thought:**
  Instead of directly requesting a label, the prompt was structured to elicit step-by-step

reasoning, encouraging the model to first reflect on the article's content before issuing a prediction.

"You have seen training and reviews. Now predict the Quartile for each paper below.:\n"

"For each one, reason step by step using title and abstract, compare with context, and then assign a Quartile."

+ test_text)})

- **Thread-of-thought:**
  An improved thought inducer for CoT reasoning. Instead of "Let's think step by step," it uses "Walk me through this context in manageable parts step by step, summarizing and analyzing as we go."

"Based on your prior exposure to relevant training examples and review articles, you will now evaluate a sequence of papers. As you progress, maintain a consistent thread of reasoning across all predictions."

"For each paper, consider how its title and abstract align with the standards and patterns identified in the context. Use step-by-step reasoning to analyze novelty, relevance, and scientific contribution. Reflect on how each compare with previous papers you have seen."

"Document your reasoning clearly and consistently, then assign a final SJR Quartile (Q1–Q4) based on that thread of thought."

- **Expert-Panel:**
  Simulate an expert panel with 3 different ways of evaluating and choose a result. This makes the LLM reason from 3 different perspectives and makes a thoughtful decision.

  "Simulate a panel of three expert reviewers (R1, R2, R3) discussing each paper based on its title and abstract. Each reviewer gives a short analysis, and a final Quartile is chosen based on consensus.\n"

- **Expert Queue:**
  Simulate 4 different experts, each one of a specified Quartile, every paper goes through the queue of experts. First the Q1 expert rates it, if it meets his criteria Q1 will be assigned, if not it will go to the next expert until the last one.

  "Simulate a panel of 4 expert reviewers, each expert is an expert of a specific Quartile."
  "For each test paper, first go through the expert of Q1, then Q2, then Q3 and lastly Q4."
  "Each expert will have a criterion based on the papers of their knowledge."
  "Assign the Quartile value of the first expert that met his standards."
  "If no one accepted it, assign it to Q4."

- **Structured Reasoning + Chain-of-Thought:**

A more complex CoT where the thinking process follows a set structure with specific rules to follow. This gives the model a detailed guide of how to process the test papers when predicting its score

"Act as a domain expert in scientific journal evaluation. You've reviewed numerous training examples and academic reviews."
"Now, for each paper below, perform the following steps:\n"
"1. Read the title and abstract carefully.\n"
"2. Identify the key contributions and field relevance.\n"
"3. Compare with known characteristics of Q1–Q4 journals.\n"
"4. Decide the appropriate Quartile.\n"

- **Self Reflection:**
  Makes the model write down a reason for each prediction, read it and justify the prediction based on it. This iterative process makes the model produce a more reasonable prediction as it has been justified by itself.

  "You are an experienced reviewer for Scopus-indexed journals. Given the examples and reviews you've seen, analyze the following articles. "
  "For each article:\n"
  "- Summarize the focus based on the title and abstract.\n"
  "- Reflect on its potential impact and relevance.\n"
  "- Justify your decision briefly.\n"
  "- Assign a Quartile from Q1 (most relevant) to Q4 (least relevant).\n"

Alternative prompts:

Multiple other prompts techniques and styles were tested in this project but it was decided to only show the mentioned above. these techniques did not yield improved performance or consistency in the context of this task, which required domain-specific classification based on subtle textual cues in scientific abstracts. In particular, few-shot examples sometimes introduced unintended biases or overwhelmed the prompt window without contributing meaningful guidance, while more elaborate reasoning formats often led to verbose or unfocused outputs. As a result, the selected prompting strategies prioritized clarity, task-specific framing, and conciseness to maintain precision in quartile prediction.

These strategies were implemented following a progressive logic: as the task requires abstract judgement based on limited input, every new strategy introduced a new or different layer of reasoning to assist the model in predicting better results. The underlying hypothesis was that increasing the content in the prompts with useful guidance would enhance the model's ability to produce reliable classifications.

To further explore more options apart from just prompt engineering, two different Retrieval-Augmented Generation strategies were developed. The first one worked as usual RAG strategies do, where it found the most relevant texts (in training papers and review) for each test paper and include snippets of those texts in the prompt to help in the decision making process of the LLM. It gives real examples and background directly from similar papers.

A second and more interesting type of RAG was developed where instead of working like classic RAGs, it calculated how similar the test paper was to each of the quartile groups, and gave those numbers as part of the prompt to GPT-4o. It would use numeric clues such as “this paper is 80% similar to Q2 papers” instead of long texts. It was named Similarity_Score_RAG.

- RAG:

```
"You are an expert in evaluating scientific articles for Scopus Quartile classification (Q1–Q4).\n\n"
    "Based on the following context extracted from previous reviews:\n\n"
    f"{contexto_reviews}\n\n"
    "And based on similar examples from training papers:\n\n"
    f"{contexto_train}\n\n"
    "Now evaluate the following paper:\n"
    f"Title: {title}\nAbstract: {abstract}\n\n"
    "First, reason step-by-step how its content compares to both the training and reviewed works.\n"
    "Then, assign a Quartile (Q1 = highest, Q4 = lowest).\n"
```

- RAG- Similarity-Score:

```
f"You are a scientific journal expert. A paper has the following title and abstract:\n"
    f"Title: {title}\n"
    f"Abstract: {abstract}\n\n"
    f"It has the following average cosine similarities to past articles grouped by quartile:\n"
    + "\n".join([f"- {q}: {score:.3f}" for q, score in sim_scores.items()]) + "\n\n"
    "Based on this similarity profile, assign the most likely Quartile (Q1–Q4).\n"
    "Respond strictly in the format:\nTitle: [title]\nQuartile: Q[1–4]"
```

3.2) Evaluation framework

To assess the performance of the language model in predicting the SCImago Journal Rank (SJR) quartile of scientific articles, a classification-based evaluation framework was employed. The model's outputs were compared against the ground truth quartile labels derived from the 2024 SCImago rankings for each paper in the test set. As the task involved assigning one of four possible classes (Q1–Q4) based solely on the title and abstract of each paper, standard multiclass classification metrics were used to evaluate the model's effectiveness.

The primary metric reported was accuracy, defined as the proportion of correctly predicted quartiles across all examples. While accuracy provides a straightforward measure of overall correctness, it does not capture the nuances of performance across individual quartile classes, especially in the presence of any residual class imbalance.

To obtain a more detailed view of the model's behavior, a confusion matrix was constructed. This matrix highlights how often each true class (actual quartile) was confused with each predicted class, offering insight into systematic biases, for example, whether the model tends to overpredict high-prestige quartiles such as Q2 or Q3.

In addition to accuracy and confusion matrix analysis, precision, recall, and F1-score were calculated for each quartile class. These metrics offer class-specific perspectives:

- Precision indicates how often a predicted quartile was correct.
- Recall measures how well the model identified all papers belonging to a given quartile.
- F1-score, the harmonic mean of precision and recall, balances these two aspects into a single measure.

Finally, a macro-averaged F1-score was reported to reflect the model's performance across all classes equally, regardless of how many examples fall into each quartile. This is particularly important in this study, as the dataset was deliberately balanced to offset the natural overrepresentation of Q1 and Q2 journals in the domain.

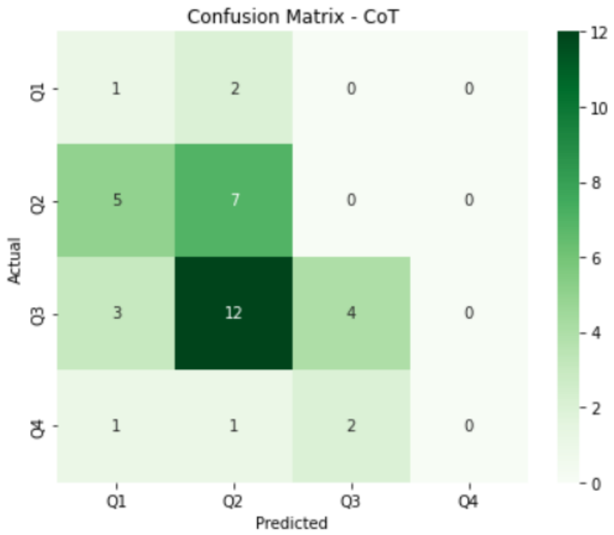
Together, these metrics provide a comprehensive view of the model's classification performance, enabling a nuanced evaluation of its potential for automating early-stage impact estimation in scientific publishing.

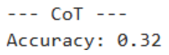
4. Results & Interpretation

4.1) Results

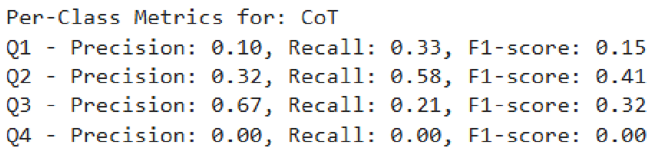
This section presents the evaluation of various prompt engineering strategies applied to the task of predicting the SCImago Journal Rank (SJR) quartile of scientific papers using a large language model. Each strategy was assessed using a common validation set, and performance was measured using standard classification metrics, including accuracy, macro-averaged and weighted F1-scores, as well as precision, recall, and F1-score per quartile class. Additionally, confusion matrices were generated to visualize the distribution of predictions and identify any systematic biases. The comparative analysis highlights the relative strengths and limitations of each prompting approach in capturing domain-relevant signals from paper abstracts.

- **Chain of Thought**



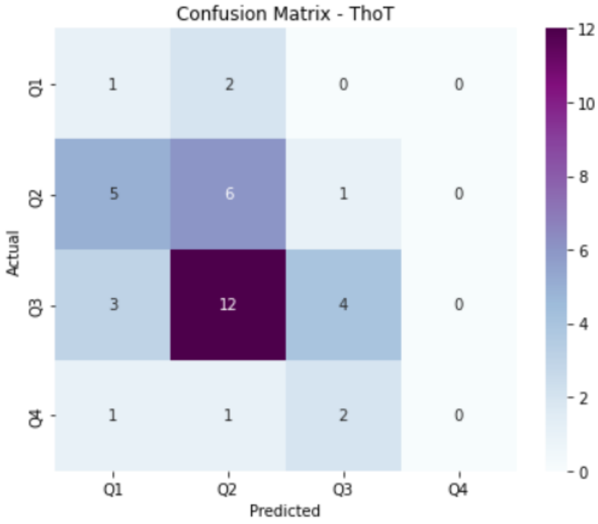


```
--- CoT ---
Accuracy: 0.32
```

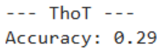


```
Per-Class Metrics for: CoT
Q1 - Precision: 0.10, Recall: 0.33, F1-score: 0.15
Q2 - Precision: 0.32, Recall: 0.58, F1-score: 0.41
Q3 - Precision: 0.67, Recall: 0.21, F1-score: 0.32
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

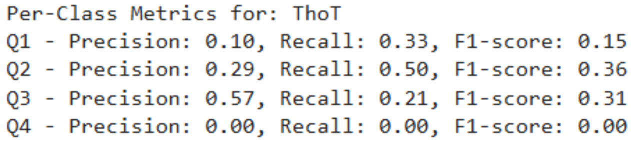
- ### Tree of Thought



	Q1	Q2	Q3	Q4
Q1	1	2	0	0
Q2	5	6	1	0
Q3	3	12	4	0
Q4	1	1	2	0

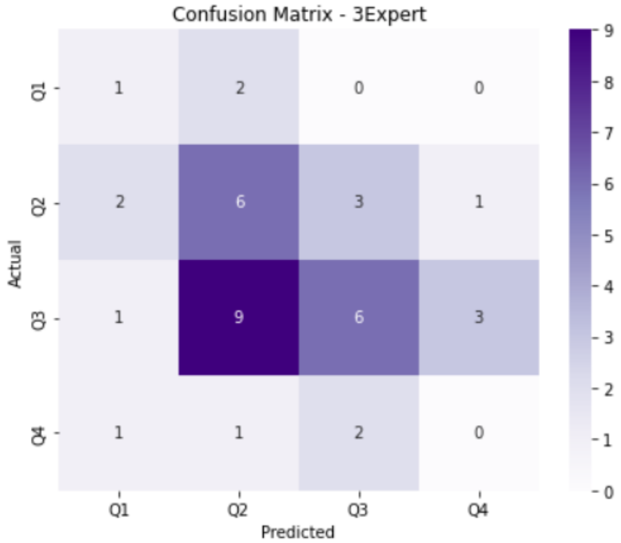


```
--- ThoT ---
Accuracy: 0.29
```

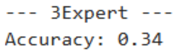


```
Per-Class Metrics for: ThoT
Q1 - Precision: 0.10, Recall: 0.33, F1-score: 0.15
Q2 - Precision: 0.29, Recall: 0.50, F1-score: 0.36
Q3 - Precision: 0.57, Recall: 0.21, F1-score: 0.31
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

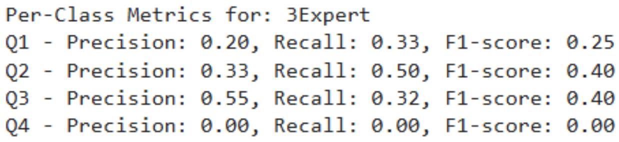
- ### Expert Panel



	Q1	Q2	Q3	Q4
Q1	1	2	0	0
Q2	2	6	3	1
Q3	1	9	6	3
Q4	1	1	2	0

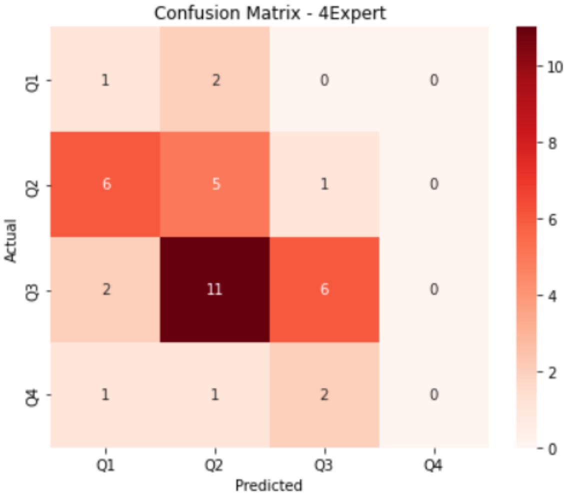


```
--- 3Expert ---
Accuracy: 0.34
```



```
Per-Class Metrics for: 3Expert
Q1 - Precision: 0.20, Recall: 0.33, F1-score: 0.25
Q2 - Precision: 0.33, Recall: 0.50, F1-score: 0.40
Q3 - Precision: 0.55, Recall: 0.32, F1-score: 0.40
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

- **Expert Queue**

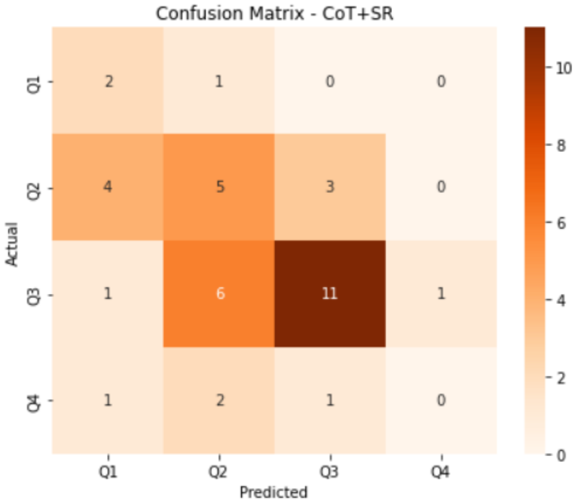


	Q1	Q2	Q3	Q4
Q1	1	2	0	0
Q2	6	5	1	0
Q3	2	11	6	0
Q4	1	1	2	0

```
--- 4Expert ---
Accuracy: 0.32
```

```
Per-Class Metrics for: 4Expert
Q1 - Precision: 0.10, Recall: 0.33, F1-score: 0.15
Q2 - Precision: 0.26, Recall: 0.42, F1-score: 0.32
Q3 - Precision: 0.67, Recall: 0.32, F1-score: 0.43
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

- **CoT + Structured Reasoning**

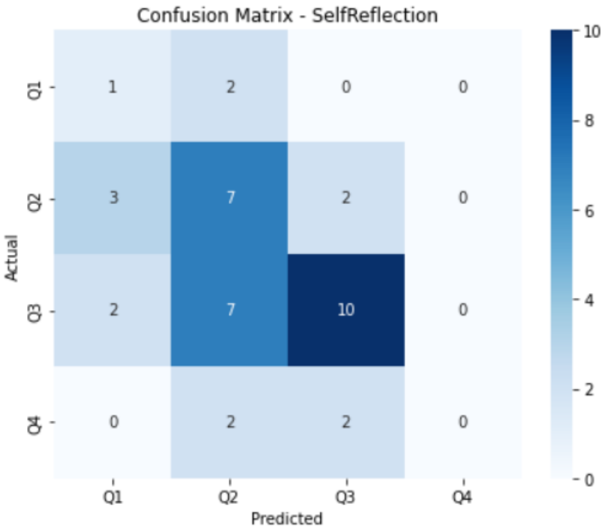


	Q1	Q2	Q3	Q4
Q1	2	1	0	0
Q2	4	5	3	0
Q3	1	6	11	1
Q4	1	2	1	0

```
--- CoT+SR ---
Accuracy: 0.47
```

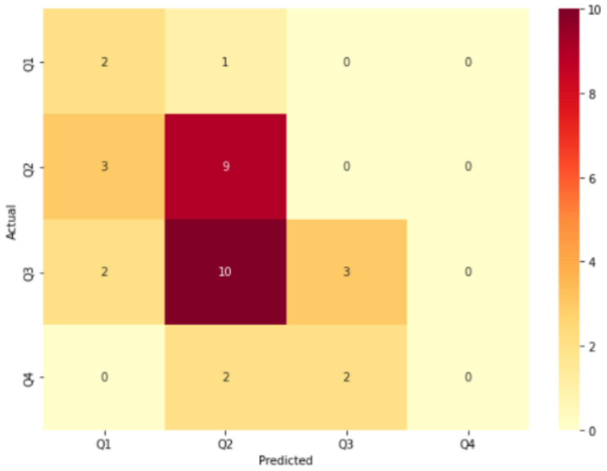
```
Per-Class Metrics for: CoT+SR
Q1 - Precision: 0.25, Recall: 0.67, F1-score: 0.36
Q2 - Precision: 0.36, Recall: 0.42, F1-score: 0.38
Q3 - Precision: 0.73, Recall: 0.58, F1-score: 0.65
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

- **Self-Reflection**



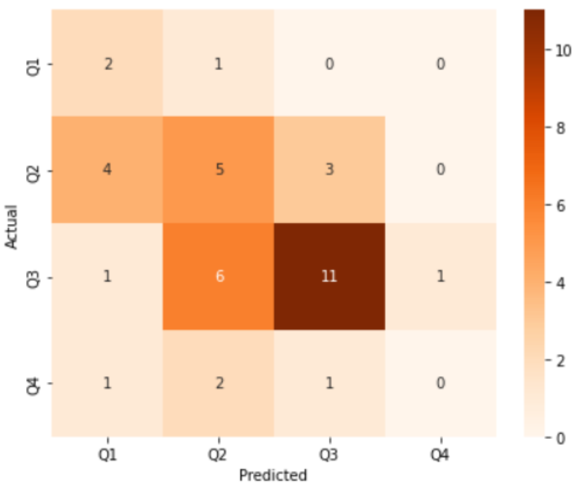
```
--- SelfReflection ---
Accuracy: 0.47

 Per-Class Metrics for: SelfReflection
 Q1 - Precision: 0.17, Recall: 0.33, F1-score: 0.22
 Q2 - Precision: 0.39, Recall: 0.58, F1-score: 0.47
 Q3 - Precision: 0.71, Recall: 0.53, F1-score: 0.61
 Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

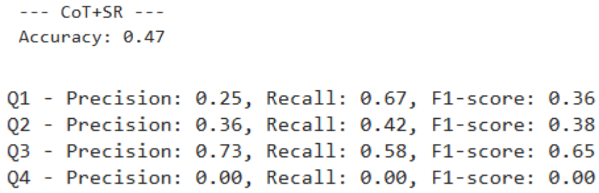
- **RAG**



```
--- RAG ---
Accuracy: 0.41


Q1 - Precision: 0.29, Recall: 0.67, F1-score: 0.40
Q2 - Precision: 0.41, Recall: 0.75, F1-score: 0.53
Q3 - Precision: 0.60, Recall: 0.20, F1-score: 0.30
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

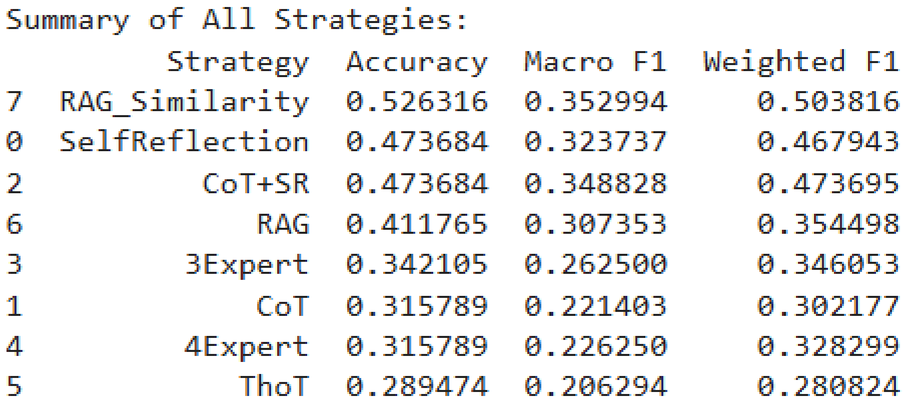
- **RAG – Similarity Context**





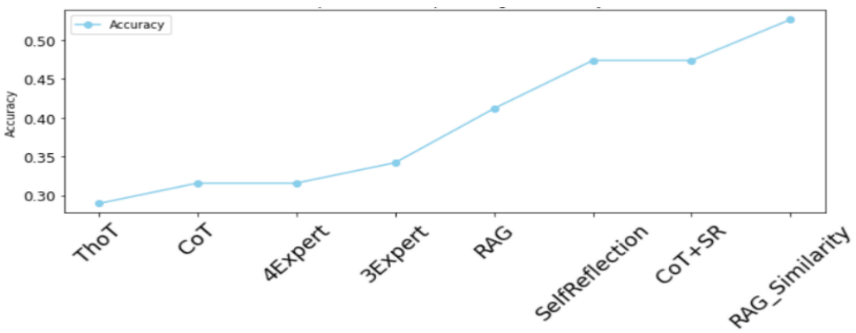
```
--- CoT+SR ---
Accuracy: 0.47

Q1 - Precision: 0.25, Recall: 0.67, F1-score: 0.36
Q2 - Precision: 0.36, Recall: 0.42, F1-score: 0.38
Q3 - Precision: 0.73, Recall: 0.58, F1-score: 0.65
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

- Summary of all strategies:



```
Summary of All Strategies:
        Strategy  Accuracy  Macro F1  Weighted F1
7 RAG_Similarity  0.526316  0.352994     0.503816
0 SelfReflection  0.473684  0.323737     0.467943
2         CoT+SR  0.473684  0.348828     0.473695
6            RAG  0.411765  0.307353     0.354498
3        3Expert  0.342105  0.262500     0.346053
1            CoT  0.315789  0.221403     0.302177
4        4Expert  0.315789  0.226250     0.328299
5           ThoT  0.289474  0.206294     0.280824
```

- **Comparison between all methods:**



The line chart compares the performance of eight different prompt engineering strategies in terms of overall accuracy. It is evident that strategies incorporating more sophisticated reasoning mechanisms, such as RAG or Self-Reflection, achieve comparatively better results. These approaches outperform simpler strategies like expert ensemble prompting (3Expert, 4Expert) or Thread-of-Thought (ThoT) or Chain-of-Thought, suggesting that prompts encouraging step-by-step justification may enhance the model's ability to assess abstract content relative to impact. Nevertheless, the overall results remain modest, with none of the strategies surpassing the 60% accuracy threshold. This supports the underlying assumption that predicting scientific impact from abstracts alone is inherently limited, as critical context and evaluative criteria influencing journal rankings are often absent from title and abstract-level information

4.2) Interpretation

The results obtained across all evaluated prompting strategies demonstrate that predicting a paper's SJR quartile solely from its abstract presents considerable challenges. Overall accuracy levels ranged from modest to moderate, with the best-performing strategy achieving less than 50% accuracy. Precision, recall, and F1-scores varied significantly across quartile classes, with lower performance consistently observed for Q1 and Q4 categories. This imbalance suggests that the model struggled to differentiate between extremes of perceived impact, potentially due to limited discriminative cues in the input data.

These outcomes align with the initial expectation that abstracts alone would offer an incomplete basis for robust impact estimation. While abstracts do summarize the scope, methods, and findings of a paper, they often lack critical contextual factors that influence citation dynamics and editorial decisions—such as author reputation, institutional affiliation, methodological rigor, or timeliness relative to trends in the field. Furthermore, some impactful studies may be presented in understated terms, while less influential works may use assertive language, making linguistic features an unreliable proxy for scholarly impact.

The low precision and recall for Q4 in particular suggest a strong tendency of the model to overpredict middle quartiles, possibly due to implicit biases in training data or the dominance of Q2/Q3 papers in the domain. The confusion matrices support this, showing frequent misclassification between adjacent quartiles. This misalignment reflects the model's difficulty in distinguishing fine-grained differences in perceived journal prestige, which even human experts often debate.

The integration of Retrieval-Augmented Generation (RAG) strategies yielded a modest but consistent improvement in the predictive performance of the language model across several evaluation metrics. Compared to purely prompt-engineered approaches, RAG-based methods, especially the version retrieving relevant abstracts and training examples—demonstrated slightly higher accuracy and F1 scores. This enhancement is attributed to the fact that retrieved content enriches the context provided to the model, offering domain-specific knowledge that is temporally aligned and semantically similar to the test inputs. Unlike general prompts that rely solely on the abstract of the paper, RAG allows the model to ground its reasoning in previously seen mate-rials, simulating a more informed editorial review process.

The improvement observed can be explained by the fact that language models, while capable of generalizing from training data, benefit significantly from targeted conditioning when faced with nuanced classification tasks such as journal impact estimation. By surfacing thematically similar papers or summaries of trends via RAG, the model receives direct cues about the type of research historically associated with specific quartiles. This reduces ambiguity and supports more consistent classification decisions, particularly in middle quartiles where most papers tend to cluster. While the gain is not dramatic, it highlights the value of combining structured retrieval with prompting to better align model outputs with human-like editorial judgments.

In summary, while certain prompting strategies performed better than others, none achieved a level of precision that would justify the use of LLMs as standalone tools for impact prediction. These results emphasize the limitations of abstract-only inference and point to the need for more comprehensive input features, for example full-text analysis, author metadata, or citation context, should future work aim to improve predictive reliability. Nonetheless, the findings provide a valuable baseline for understanding how generative models interpret academic content and suggest directions for refinement in prompt design and dataset construction.

5. Conclusions

This study explored the potential of large language models, specifically GPT-4o, to predict the impact of scientific articles as measured by the SCImago Journal Rank (SJR) quartile, using only the title and abstract of each paper. Through the evaluation of multiple prompt engineering strategies, it was shown that while the model can identify certain general patterns and make informed guesses, its predictive performance remains limited across most metrics.

In addition to assessing model output, the project also highlights the central role of prompt design in shaping the behavior of generative systems. Each strategy examined offers a different approach to guiding the model's attention, simulating human reasoning, or encouraging self-evaluation. As such, the comparative analysis of these strategies adds to the growing methodological toolkit available to practitioners aiming to direct LLMs toward specific evaluative goals in academic or technical domains.

The best-performing strategies achieved moderate accuracy and F1-scores yet often struggled with consistent classification across all quartiles, particularly at the extremes. These results reinforce the hypothesis that abstract-level information alone is insufficient for reliably estimating a paper’s future impact or journal placement. Important predictive signals such as methodological depth, novelty within the full context, and broader scholarly relevance are not always captured in abstracts and are beyond the model’s inference capabilities without additional input.

From a methodological perspective, this work contributes a replicable framework for testing LLM performance in classification tasks tied to real-world scientific metadata. The use of known journal quartiles as labels, the construction of a training and validation split based on publication year, and the inclusion of multiple reasoning strategies provide a solid foundation for future experiments. Additionally, the project underscores the value of domain specificity: by selecting a very narrow field, transcriptional regulation in bacteria, it was possible to reduce noise and focus the model’s attention, which may be crucial in settings where subtle textual cues drive significant differences in classification.

Despite these limitations, the project contributes to the emerging body of work investigating the role of generative AI in academic evaluation and editorial decision support. It demonstrates both the potential and the current boundaries of language models in scholarly impact prediction and suggests practical directions for improvement, including the use of richer input data and refined prompting techniques. Ultimately, this work should be seen as a foundational step toward integrating LLMs into more nuanced, data-informed workflows rather than as a definitive solution for editorial judgment.

6. Evolution of the project

This study has taken different routes during the time of research. Initially the plan was to navigate different LLM's online and perform the same prompt strategies in different LLM's to test which would give best results. The project changed its course for many reasons:

First and foremost, LLM's online free to use are, in most cases, old versions of the company's LLM, for whom you're allowed to use a *mini* version with much less capacity, or you have a limited usage of the tool. This was especially relevant as the size of the prompts I wanted to implement, revolving around 100k tokens per prompt, needed a large quantity of usage available. To access the more powerful LLM's with bigger window sizes and that have been proven to work well with text-based analysis, a monthly subscription or initial charging fee is required. Llama would be an alternative, but it lacks a reasoning model at the moment. As this project was thought to be just an introductory exploration of the usage of LLM's as a tool for predicting scientific papers' potential impact, it was considered that the better idea would be to specify the funding destined for it to just one subscription, which was decided to be the Open AI LLM, GPT-4o. First because it is widely considered to be the LLM with best results across the board when performing text-based analysis and reasoning and secondly because the context size is large enough to support the type of prompts that were intended in the project.

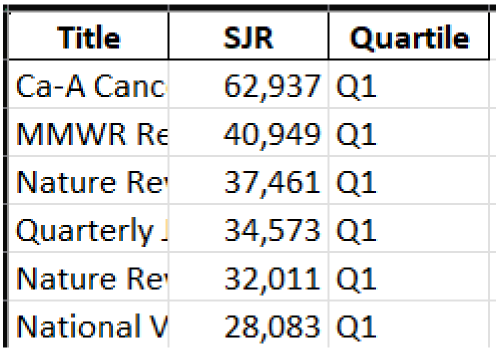
Other ideas held at the start of the project were measuring impact factor of papers based on possible citation counts. The issue of not all citations being of the same quality but counting the same was an obstacle that deserved to be investigated. It was found that in the paper repository of Semantic Scholar with each paper it was also given a number of "Highly influential citations" which defined a more relevant metric and would give a possible solution to the issue mentioned before. Anyway, it wasn't a good enough solution as another issue showed up when discovering how LLM's worked, and their poor performance with arithmetic and predicting numerical events. Therefore, the decision was made to change the scope of the study and use the SJR value to make predictions. While it still is a number and making the LLM predict its possible value would mean falling into the same trap, SJR rankings were easy to divide into 4 quartiles, changing the query asked from predicting a number to classifying in categories, an area where LLM's are much more stable and produce better results.

Many prompting strategies were thought and tested but just a selection of the most interesting/most reasonable were selected to be shown in the report, as the model struggled to show accurate results, it was decided that techniques that were relevant to the question and how it was prompted would be selected.

7. Appendices

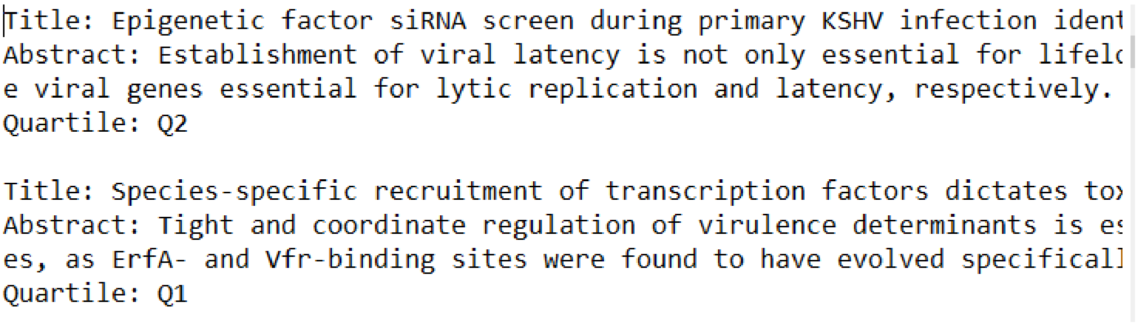
7.1) Dataset Structure:

- SJR Rankings:
  Excel file for each year where papers have been studied. Processed from the SCImago yearly Journal Ranks taking the 3 relevant columns. Title of publisher, SJR value, Quartile.



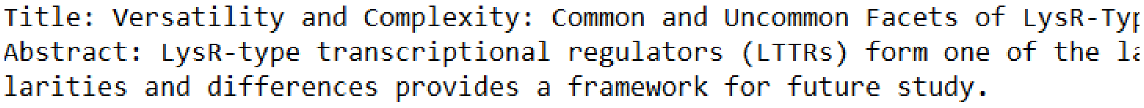
Title	SJR	Quartile
Ca-A Canc	62,937	Q1
MMWR Re	40,949	Q1
Nature Re	37,461	Q1
Quarterly J	34,573	Q1
Nature Re	32,011	Q1
National V	28,083	Q1

- Training Set:
  Consisted in a TXT file including 187 papers that followed the structure:
  - Title:
  - Abstract:
  - Quartile:



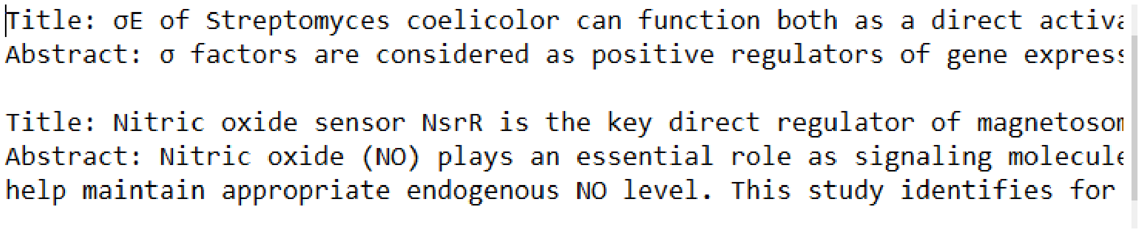
```
Title: Epigenetic factor siRNA screen during primary KSHV infection ident
Abstract: Establishment of viral latency is not only essential for lifelo
e viral genes essential for lytic replication and latency, respectively.
Quartile: Q2

Title: Species-specific recruitment of transcription factors dictates tox
Abstract: Tight and coordinate regulation of virulence determinants is es
es, as ErfA- and Vfr-binding sites were found to have evolved specificall
Quartile: Q1
```

- Review Set:
  TXT file including 7 reviews with formt: title and abstract.



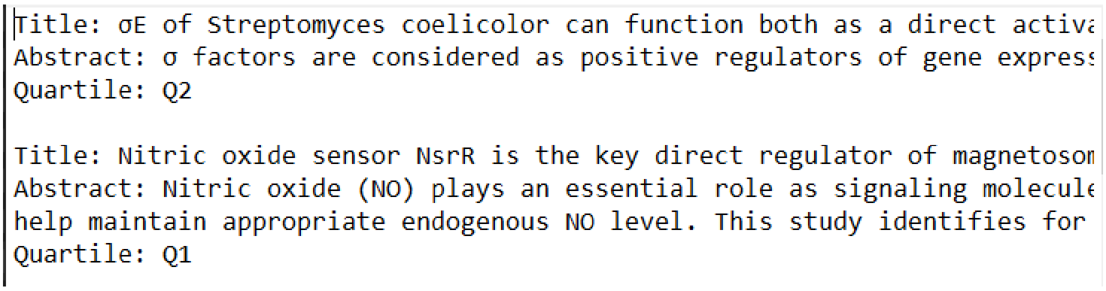
```
Title: Versatility and Complexity: Common and Uncommon Facets of LysR-Typ
Abstract: LysR-type transcriptional regulators (LTTRs) form one of the la
larities and differences provides a framework for future study.
```

- Test Set:
  TXT file including 38 papers for the LLM to predict.



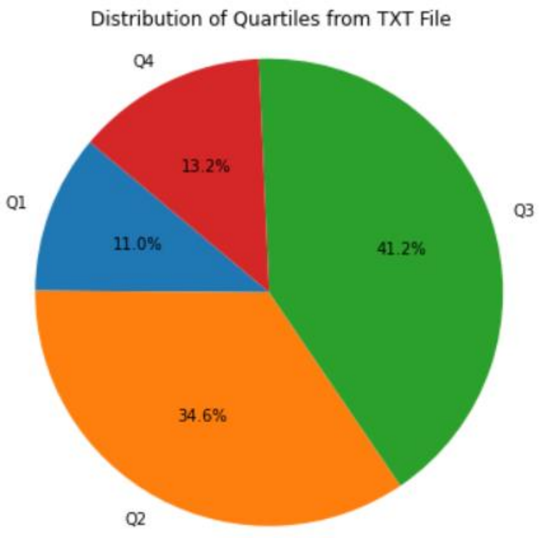
```
Title: σE of Streptomyces coelicolor can function both as a direct activa
Abstract: σ factors are considered as positive regulators of gene express

Title: Nitric oxide sensor NsrR is the key direct regulator of magnetosom
Abstract: Nitric oxide (NO) plays an essential role as signaling molecule
help maintain appropriate endogenous NO level. This study identifies for
```

- Validation Set:
  TXT file the same 38 papers from the test set including the true Quartile



```
Title: σE of Streptomyces coelicolor can function both as a direct activa
Abstract: σ factors are considered as positive regulators of gene express
Quartile: Q2

Title: Nitric oxide sensor NsrR is the key direct regulator of magnetosom
Abstract: Nitric oxide (NO) plays an essential role as signaling molecule
help maintain appropriate endogenous NO level. This study identifies for
Quartile: Q1
```

- Distribution of journals by Quartile in the training:


Distribution of Quartiles from TXT File
Q4
Q1
Q3
Q2