LLM-Based forecasting of scientific papers impact based on abstract

Jaime Rueda Diví

Abstract—The dynamics of the scientific publishing industry, particularly under the current Open Access model, have incited substantial debate regarding the financial barriers faced by authors. Despite the availability of new publishing initiatives like Diamond Open Access journals, author engagement and public perception remains low due to the absence of traditional editorial oversight, which provides a benchmark for research quality and impact. This study explores the potential of leveraging Generative AI systems such as GPT-4o, specifically using advanced prompting techniques as an alternative mechanism to predict the impact of scientific articles, based on their abstract as it is the access-free part of scientific papers. By designing a diverse array of prompts based on the latest techniques of prompt engineering, the study evaluates the LLM's ability to categorize research articles into predetermined relevance quartiles, quartiles based on established academic rankings from Scimago. Initial trials demonstrate limited success, suggesting that reliance on abstracts alone may be insufficient for accurate impact forecasting, thus inviting further exploration.

Index Terms— Prompt Engineering, Large Language Models, Scientific papers prediction, SCImago Journal Rank Abstract

1 Introduction

Scientific publishing is a multimillion industry.[1] Under the current Open Access model, scientific articles are free to read by the general public, but authors must pay substantial fees (Article Processing Charges) to publish the results of their research. Several non-profit organizations have estimated that the costs associated with publishing a research article and maintaining it permanently available hover around €250, but publishing companies typically charge APCs more than 10 times higher. Several initiatives for free open access publishing (Diamond Open Access) have arisen in recent years, creating new scientific journals, but author buy-in is still relatively low. A main reason for the reticence of authors to submit research papers to these new Diamond OA journals is the fact that they lack editors vetting submissions to the journal. This creates a problem for authors. A journal prestige is established by its impact factor, which stems from the number of citations articles in the journal receive over time. While an author may justify the impact of their research based on citations, these take a long time to accumulate. Editors in traditional scientific journals decide whether an article is admissible for publication in a journal with an already established impact factor, thereby providing an immediate indicator of quality and impact for a scientific paper.

In the context of a rapidly evolving scientific inquiry and technological progress. The traditional model is increasingly regarded as inadequate in addressing the needs of a fast-paced collaborative research environment. One of the most cited limitations is the considerable time delay between the manuscript submission and final publication. This lag, often ranging from several months to over a year, hampers timely communication of scientific discoveries and could result in the same discovery being published elsewhere.

In addition to delays, concerns have been raised regarding the consistency and transparency of editorial and peer-review practices. Peer review, while foundational to the quality control of academic literature, varies considerably in rigor and objectivity across journals and disciplines. The lack of standardized review protocols can result in subjective evaluations or inconsistent recommendations. Moreover, the traditional system perpetuates structural inefficiencies. Reviewers typically work in isolation and their assessment is not shared across journal, leading to redundant efforts when manuscripts are declined by a publisher and resubmitted elsewhere.

Therefore, this study contemplates the option of a new tool set to be a helpful assistant to scientists predict potential impact of their studies without going through the traditional process. This project investigates whether generative AI systems, specifically large language models, can approximate editorial decision-making by predicting the perceived impact of scientific articles based solely on their abstract, as abstracts are the public part of papers.

- *E-mail de contacte: jaimeruedadivi@gmail.com*
- *Treball tutoritzat per: Ivan Erill Sagales (Area of Computer Science and Artificial Intelligence)*
- *Curs 2024/25*

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.

In fast-moving fields, being able to estimate the likely impact of a paper at the time of its release could inform funding priorities, collaborative decisions, and media outreach, enhancing the responsiveness of the scientific community. If successful, this approach could be extended to assist in peer review triage, preprint assessment, or even in recommending research directions based on anticipated scholarly attention.

1.1 Objectives

- Provide a prediction for the tier of publisher a set of papers of a specific topic should be published in.
- Investigate several prompting engineering techniques and strategies.
- Evaluate the results and prove if AI can be a helpful assistant in this type of predictions.

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.

2 State of art

The Journal Impact Factor (JIF)[2], introduced in the 1960s, has long served as a dominant metric for evaluating scientific research by averaging citations received over a two-year period. While widely adopted to gauge journal prestige and guide publication decisions, it has been increasingly criticized for misrepresenting scholarly value. JIF assumes that citation count equates to quality, yet citations are influenced by factors such as article type, disciplinary norms, and visibility rather than intrinsic merit. Strategic behaviors like excessive self-citation and citation rings further distort results, while journal-level aggregation obscures article-level impact. Moreover, the metric suffers from a significant time-lag, making it ill-suited for early-stage evaluations or fast-moving research. Citation spikes from novelty or controversy may later be invalidated, questioning its reliability. This system also entrenches elite journals, reinforcing structural inequities and limiting diversity in academic discourse. As a result, there is growing recognition of the need for alternative, more equitable and timely metrics to assess scientific influence.

In response to these limitations, alternative metrics have been proposed, among which the SCImago Journal Rank[3] (SJR) has gained notable traction. Unlike the raw citation-based impact factor, SJR considers both the number and the quality of citations received, providing a field-normalized and quartile-ranked indicator of journal prestige. Research comparing bibliometric indicators supports the view that SJR offers a more balanced and interpretable evaluation of journal influence, particularly in interdisciplinary or emerging fields where citation behavior may diverge from traditional norms.

Simultaneously, the advancement of Generative AI models, particularly large language models (LLMs), has introduced novel opportunities for automating editorial functions and predicting research impact. Earlier work has explored the use of machine learning for citation prediction and journal recommendation, but these approaches often rely on structured metadata and citation networks. Recent developments, such as GPT-4o, enable models to interpret and generate rich textual information based on abstract-level inputs. However, despite their capabilities in reasoning and text comprehension, LLMs still exhibit weaknesses in numerical processing, precision estimation, and arithmetic reasoning—limitations that must be accounted for when designing systems intended to replace or support editorial judgment.

Existing literature on prompt engineering[4] has shown that model output can vary significantly depending on the formulation and structure of the input prompt. Research in this area explores zero-shot, few-shot, and chain-of-thought prompting, each with varying levels of success across domains. While some studies have begun to evaluate AI in editorial and review processes, there remains a lack of comprehensive work assessing whether generative models can reliably predict journal-level impact metrics using only textual information available at the time of publication, particularly in a future-dated context.

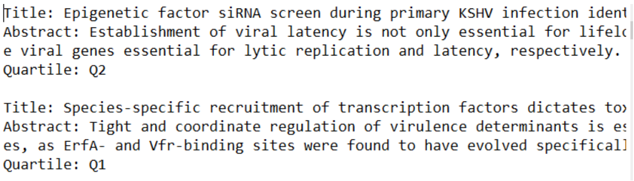
3 METHODOLOGY

3.1 Dataset

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[5]. 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.

SCImago Journal Rank includes almost every publisher in the scientific paper publishing ecosystem, every year they release an updated version of their ranking. It is disposed in an Excel file where every Journal has a SJR value and a Quartile assigned. The ranking holds the issue, in this case, of imbalanced classes as most publishers in this specific field of study fell in the Q1 and Q2 quartiles. To avoid this, a new classification of quartiles was processed based on the SJR value, increasing the strictness of the higher ranks.

Each article record consisted of the title, abstract, and the SCImago Journal Rank (SJR) quartile of the publishing journal as shown in Fig 1. The quartile was assigned accordingly to 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 in a Txt file, which fell within the knowledge cutoff of the LLM used (GPT-4o, October 2023).



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

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

Fig. 2 Format of Training set

Articles in the training dataset were carefully balanced in regard to which Quartile their publishers were categorized in as shown in Fig 2. A majority of Q2 and Q3 publishers were inevitable due to being the same journals publishing papers about this same topic.

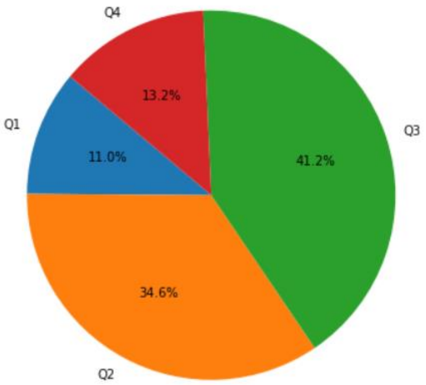


Fig. 1 Pie Chart of distribution of quartiles in the training set

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.

For both the Test set and the Validation set the same 38 articles were used, these were articles taken from PubMed about the same topics as the training and reviews, curated from the same query. Includes all 38 papers released in 2024 about the same specific field of study, transcriptional regulation in bacteria. Being released in 2024 was fundamental as GPT-4o cutoff's date is October 2023, therefore nullifying the possibility of the LLM having read them or knowing which publisher had released them. The only difference between both sets was the inclusion of the SJR Quartile as it was the goal of the query to predict which Quartile each paper belonged to.

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.[6]

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.

When prompting through the OpenAI 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. Also what kind of answer and what format was expected to be received.

The temperature parameter controls the randomness/creativity of the model's output. As this is a classification task, the model is required to be as deterministic and focused as possible in its outputs and the ability of reproducing the same prompt with similar results. Therefore, it's value is strictly set as 0.

- **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")

Due to GPT-4o Token Per Minute limit, 30k TPM, in 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 would not fit in with 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. Additionally, they were carefully looked at so that the combination of prompt+answer would not surpass the maximum context window allowed by the LLM, 128k tokens in this case.

3.3 Prompt Engineering Strategies

A critical component of this study involved the design and evaluation of prompt engineering strategies[7]. Drawing inspiration from best practices outlined in *The Prompt Report*, which emphasizes iterative refinement, multi-step reasoning, and contextual enhancement, several strategies were explored and tested during preliminary phases. From this broader set, six were selected for final evaluation based on clarity, diversity of reasoning structures, being representativeness of different prompting techniques are working adequately with text-based classifying predictions.

- **Chain-of-Thought** (CoT): based on the principle of prompting the model to reason step-by-step before delivering a prediction. This method is particularly effective in classification tasks derived from text-based inputs, as it encourages the model to parse content logically and break down abstract meaning into smaller interpretive units.

- **Thread-of-Thought** (ThoT): aims to sustain a more continuous reasoning flow, guiding the model through a reflective narrative rather than discrete steps—potentially capturing subtler contextual signals over longer text sequences.

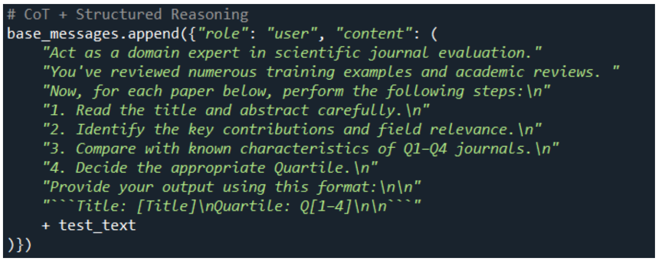
- **3-Panel Expert**: simulates a decision process involving three hypothetical domain experts. Each expert provides an individual judgment before a final consensus is formed. This technique is adapted from ensemble reasoning models and aims to mitigate bias by synthesizing multiple perspectives within a single prompt.

- **4-Expert Queue**: introduces a hierarchical panel of experts, each assigned to a specific quartile (Q1 through Q4). The simulated experts evaluate the paper sequentially, with the first one to "accept" the paper determining the final assigned quartile. This mirrors editorial workflows where papers are passed down tiers of prestige until accepted, embedding realistic decision logic into the prompt.

- **CoT with Structured Reasoning**: this strategy further expands the basic chain-of-thought method by introducing explicit reasoning categories—such as novelty, methodological soundness, and field relevance—into the prompt. This provides the model with a clearer schema for how to evaluate the abstract and make informed decisions

- **Self-Reflection**: encourages the model to critically assess its own reasoning process before issuing a final decision, leveraging metacognitive capabilities that can help reduce superficial errors and improve consistency.

Prompts were written in the following format as shown in Fig.3:



````
# CoT + Structured Reasoning
base_messages.append({"role": "user", "content": (
    "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"
    "Provide your output using this format:\n\n"
    "```Title: [Title]\nQuartile: Q[1-4]\n\n```"
    + test_text
)})
````

Fig. 3 Content prompt using Chain-of-Thought with Structured Reasoning technique

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 as "this paper is 80% similar to Q2 papers" instead of long texts. It was named Similarity_Score_RAG.

3.4 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 to evaluate each technique 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 behavior of each model, a confusion matrix was constructed to highlight how often each true class was predicted and how often it was confused. In addition, precision, recall and the harmonic mean (f1-score) were calculated for each quartile of each class, these metrics are specially relevant because the dataset distribution of each quartile is not perfectly balanced, having a slight majority of Q2 and Q3 papers.

For comparing the performance of different models it was done comparing accuracy, macro F1 and weighted F1.

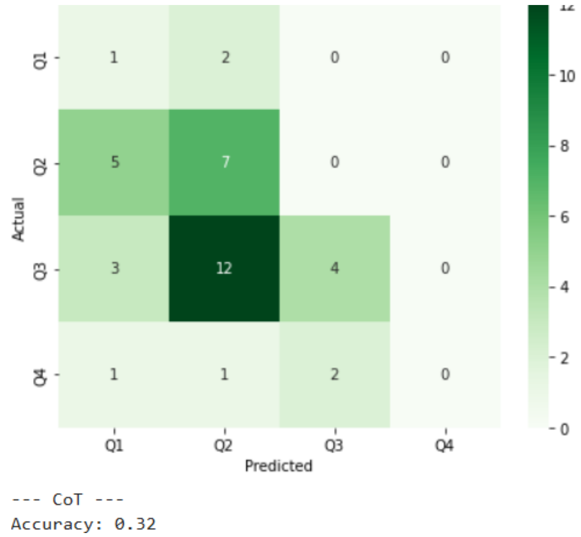
It was also performed for RAG's strategies.

4 Results & Interpretation

4.1 Result metrics

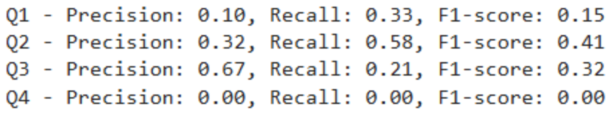
Individually, each one of the selected prompt engineering strategies was evaluated in the format shown in the following figures. Both a confusion Matrix indicating the predicted quartile vs the real quartile taken from the ground-truth file and the metrics measured by quartile in each different prompt.

It is shown in the results of the simpler Chain-of-Thought prompt and the more elaborate Chain-of-Thought with Structured Reasoning so the difference between them is clearer. The rest of the confusion matrixes and per-class metrics of each prompting technique can be viewed in the Appendix.



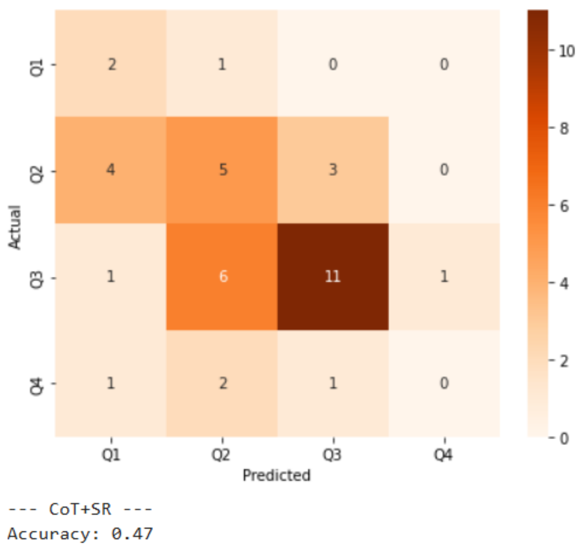
```
--- CoT ---
Accuracy: 0.32
```

Fig. 4 Confusion Matrix result of using Chain-of-Thought prompting



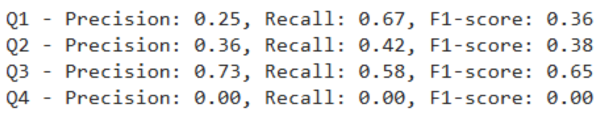
```
Q1 - Precision: 0.10, Recall: 0.33, F1-score: 0.15
Q2 - Precision: 0.32, Recall: 0.58, F1-score: 0.41
Q3 - Precision: 0.67, Recall: 0.21, F1-score: 0.32
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

Fig. 5 Per-Class metrics result of using Chan-Of-Thought prompting



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

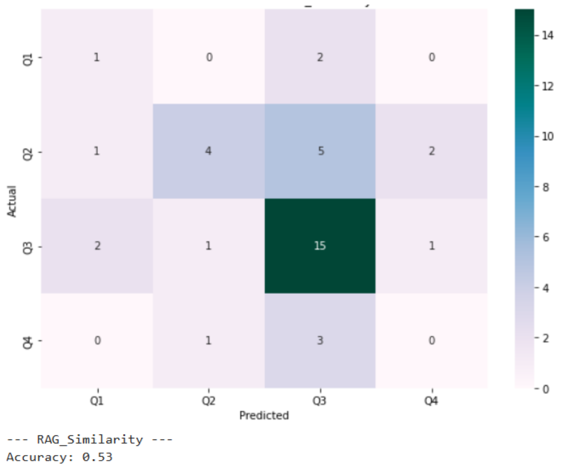
Fig. 7 Confusion Matrix result of using Chain-Of-Thought combined with Structured Reasoning prompting



```
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
```

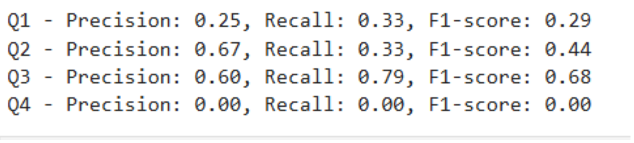
Fig. 6 Per-Class metrics result of using Chain-Of-Thought combined with Structured Reasoning prompting

Results for Retrieval-Augmented Generation showed better performance across the board while keeping biases towards middle quartiles.



```
--- RAG_Similarity ---
Accuracy: 0.53
```

Fig. 9 Confusion Matrix results for Retrieval-Augmented Generation using similarity scores prompting



```
Q1 - Precision: 0.25, Recall: 0.33, F1-score: 0.29
Q2 - Precision: 0.67, Recall: 0.33, F1-score: 0.44
Q3 - Precision: 0.60, Recall: 0.79, F1-score: 0.68
Q4 - Precision: 0.00, Recall: 0.00, F1-score: 0.00
```

Fig. 8 Per-Class metrics for results for Retrieval-Augmented Generation using similarity scores prompting

Comparing the metrics looked into the different methods between them produced the results shown in Table 1.

TABLE 1

Results of metrics for the selected prompting techniques

Strategy	Accuracy	Macro F1	Weighted F1
CoT	0.315789	0.221403	0.302177
ThoT	0.289474	0.206294	0.280824
Expert Panel	0.342105	0.262500	0.346053
Expert Queue	0.315789	0.226250	0.328299
CoT + SR	0.473684	0.348828	0.473695
Self Reflection	0.473684	0.323737	0.467943
RAG	0.411765	0.307353	0.354498
RAG – Similarity Score	0.526316	0.352994	0.503816

The improvement in performance by model is shown clearly in the following Fig 10.:

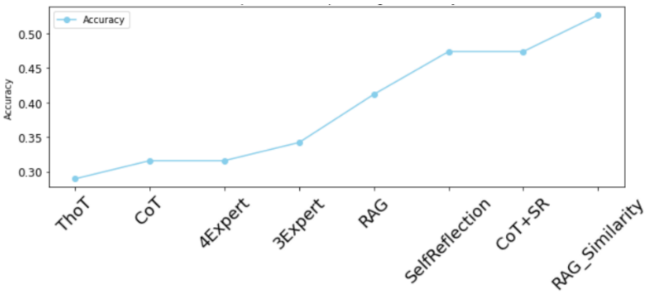


Fig. 10 Line Graph showing Accurary values vs prompting strategy

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 the distribution of quartiles from the training dataset, in which Q2 and Q3 were the most predominant quartiles adding up to a 75,6% of the papers. This distribution has been proved to skew the results.

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 closer 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 materials, 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.

5 Conclusions

This study explored the potential of large language models, specifically GPT-4o, to predict the potential tier of publisher deserved by 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.

5.1 Future workflows

This study represents an initial exploration of the use of LLM's for predicting scientific papers' potential impact. Future lines of work should focus on expanding input context available, as abstracts provide only a partial representation of a paper's contribution, issues for that are the privacy of the content of most papers.

Looking forward, the work serves as a starting point for further research into how LLMs can be integrated into editorial pipelines, peer-review support tools, or academic discovery platforms. The techniques applied here could also be extended to evaluate other impact indicators, such as citation trajectories, relevance to emerging trends, or even ethical considerations in scientific publishing. As language models continue to evolve, their application in scholarly contexts will likely expand, but careful design, validation, and interpretability must remain at the core of such efforts.

The involvement of domain-adapted language models trained on specific data. This would make the predicting program more focused on just specific areas where it is trained but could deliver better results.

Finally, expanding the scope of research beyond a single domain—such as transcriptional regulation in bacteria—would allow the generalizability of findings to be tested across disciplines with different publishing norms and impact dynamics. Comparing performance across fields could yield insights into how well LLMs generalize abstract-level impact assessments and where domain-specific calibration is necessary.

These continuations would not only improve the technical performance of such predictive systems but also contribute to ongoing discussions about the role of AI in academic publishing, peer review, and research evaluation frameworks.

6 Acknowledgments

I would like to express my sincere gratitude to my tutor for their invaluable guidance and support throughout the development of this project. I also wish to thank my family for their constant encouragement and for always believing in me. Lastly, a heartfelt thank you to my girlfriend for her patience, motivation, and unwavering support during every stage of this work.

7 References

[1] «For-science or For-profit?», ECS. Disponible en: https://www.electrochem.org/for-science-or-for-profit

[2] M. Sharma, A. Sarin, P. Gupta, S. Sachdeva, y A. V. Desai, «Journal Impact Factor: Its Use, Significance and Limitations», *World J. Nucl. Med.*, vol. 13, n.º 2, p. 146, 2014, doi: 10.4103/1450-1147.139151.

[3] «Scimago Journal & Country Rank». Accedido: 29 de junio de 2025. [En línea]. Disponible en: https://www.scimagojr.com/

[4] S. Schulhoff *et al.*, «The Prompt Report: A Systematic Survey of Prompt Engineering Techniques», 26 de febrero de 2025, *arXiv*: arXiv:2406.06608. doi: 10.48550/arXiv.2406.06608.

[5] «PubMed», PubMed. Disponible en: https://pubmed.ncbi.nlm.nih.gov/

[6] «Best practices for prompt engineering with the OpenAI API | OpenAI Help Center». Disponible en: https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api

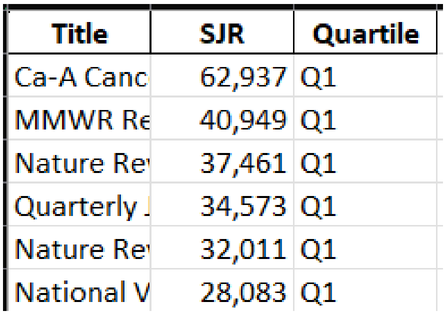
[7] «Prompt Engineering Guide». Disponible en: https://www.promptingguide.ai/es

APPENDIX

A1: Dataset Structure

- SJR Rankings:

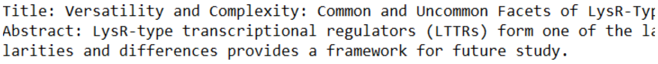
Excel file for each year where papers have been studied. Processed from the SCImago yearly Journal Ranks taking the 3 relevant columns. Title of publisher, SJR value, Quartile.



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

Fig. 11 SJR Rankings, processed to keep only the relevant values

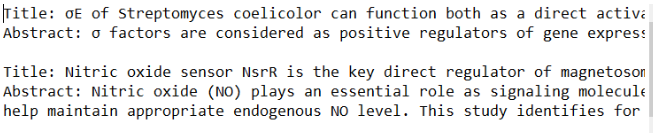
- Reviews set:



```
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.
```

Fig. 12 Reviews set format

- Test set:

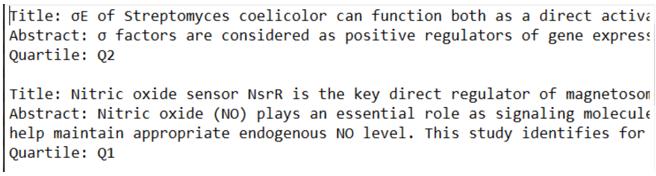


```
Title: σE of Streptomyces coelicolor can function both as a direct activa
Abstract: σ factors are considered as positive regulators of gene express

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
```

Fig. 13 Test set format

- Validation Set:



```
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
```

Fig. 14 Validation set format

A2: Prompt Engineering Queries

- **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}"

- Chain of Thought:

"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:

"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 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 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."

- Self-Reflection:

"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"

- RAG:

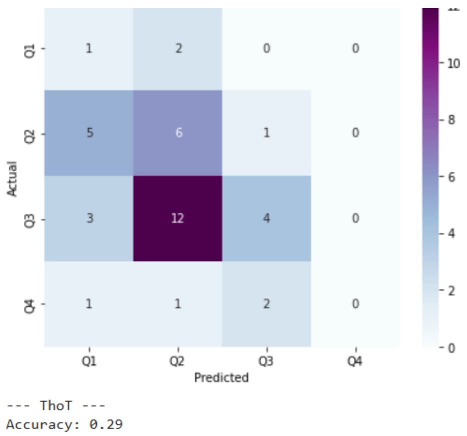
"You are an expert in evaluating scientific articles for Scopus Quartile classification (Q1–Q4).\n\n"
"Based on the following context extracted from previous reviews:\n\n"
f"{contexto_reviews}\n\n"
"And based on similar examples from training papers:\n\n"
f"{contexto_train}\n\n"
"Now evaluate the following 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 Similiraty_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]"

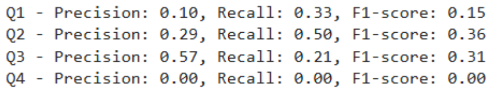
A3: Prompt Results

- Thread of thought



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

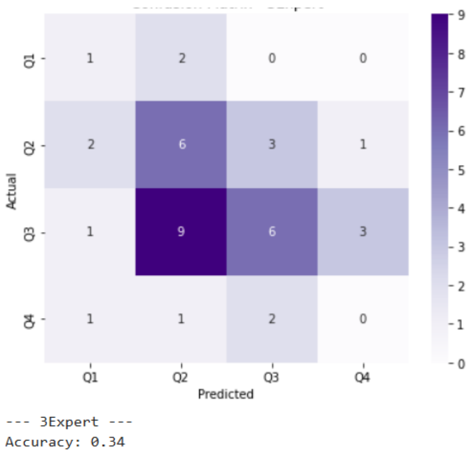
Fig. 15 Confusion Matrix for Thot



```
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
```

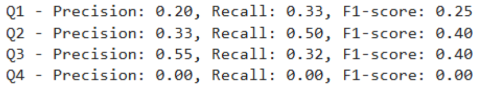
Fig. 16 Per-Class metrics for Thot

- Expert Panel:



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

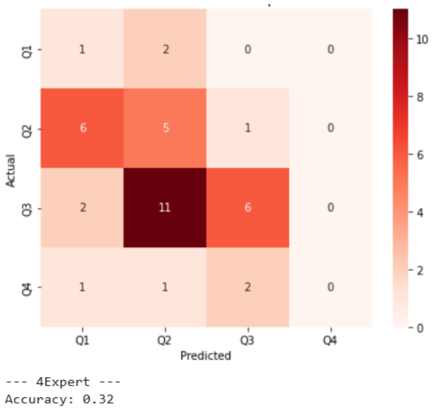
Fig. 17 Confusion Matrix for Expert Panel



```
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
```

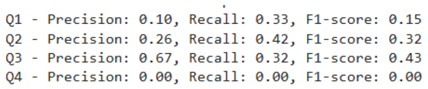
Fig. 18 Per-Class Metrics for Expert Panel

- Expert Queue:



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

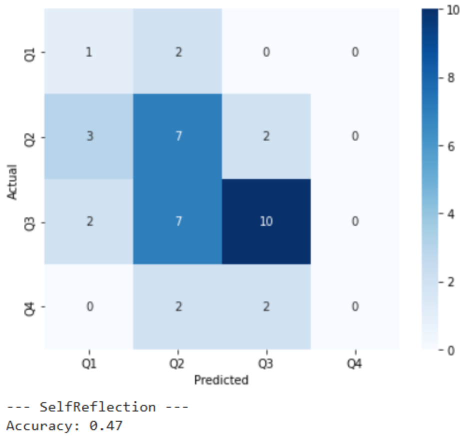
Fig. 19 Confusion Matrix for Expert Queue



```
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
```

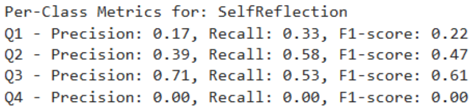
Fig. 20 Per-Class Metrics for Expert Queue

- Self-Reflection:



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

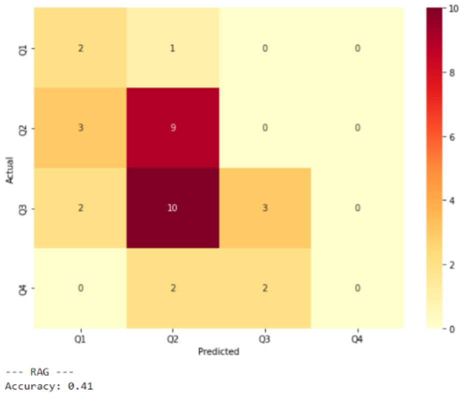
Fig. 21 Confusion Matrix For Self-Reflection



```
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
```

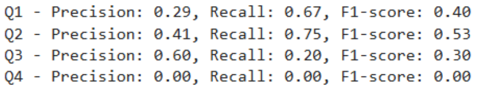
Fig. 22 Per Class Metrics for Self-Reflection

- RAG:



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

Fig. 23 Confusion Matrix for RAG



```
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
```

Fig. 24 Per-Class metrics for RAG