Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Fachbereich Informatik
Department of Computer Science

Master Project

Computer Science

LitQEval: Measuring the Effectiveness of Literature Search Queries

Mohammad Sakinini

Supervisor	Philipp Baaden, Fraunhofer INT
Examiner 1	Prof. Dr. Jörn Hees, H-BRS
Examiner 2	Dr. Milos Jovanovic, Fraunhofer INT
Draft as of	2025-01-29 20:54:22+01:00
	(For submission: set `final` option in thesis.tex!)
To be submitted on	2025-01-29

Abstract

This project report is part of the Search Query Writer (SQW) initiative by Fraunhofer INT, designed to assist researchers in generating comprehensive literature search queries, particularly in unfamiliar topic areas. A critical limitation in current literature search workflows is the absence of an objective evaluation framework for query performance, which has so far relied on subjective assessments. This study addresses this gap by introducing a curated dataset containing core publications, which are relevant publications identified through systematic literature reviews (SLRs) or bibliometric analysis works, across multiple research fields, alongside novel metrics to evaluate the query results. Recall and precision are traditionally used for literature search query evaluation, with recall being the primary metric to ensure the retrieval of relevant publications. However, literature searches often yield excessively large result sets, complicating the identification of core publications. To address this, we introduce *semantic precision*, a metric that uses embedding space to identify possible true positives based on their semantic similarity to core publications. Our evaluation reveals that while SQW-generated queries outperform the baseline in recall, they often suffer in precision, particularly when query expansion introduces irrelevant keywords. Primarily, the main goal of this framework is to provide a systematic and automated approach to assess literature search queries. It lays the groundwork for improving tools like SQW, enabling more effective literature query generation across diverse research domains.

Contents

List of Figures

1. Introduction

The Fraunhofer Institute for Technological Trend Analysis (INT)[1] specializes in conducting technology foresight, tackling tasks and research questions across a diverse array of fields. These challenges often necessitate systematic and scientifically sound approaches, even when prior knowledge in the domain is sparse. To address this recurring need, a tool that assists researchers by generating effective search queries as entry points into unfamiliar subject areas becomes essential. For instance, when faced with a specific technological research question, the process typically begins with a thorough literature search using databases like Dimensions[2], Web of Science[3], and Scopus[4]. This step involves crafting a precise search query to locate relevant studies, enabling researchers to deliver foresight grounded in scientific evidence.

To address this, several internal tools such as Topic Modeling and Grants Analytics have been developed to analyze large volumes of scientific data from sources like Dimensions.ai and Web of Science. The rise of Large Language Models (LLMs) has further enhanced the appeal and accessibility of automation across numerous domains, including scientific research, spanning from idea generation and experimental iteration to paper composition [9].

In the realm of search queries, the main focus has been on text-to-SQL [3], where an LLM is prompted via natural language to generate a precise and valid SQL query. However, to our knowledge, there has been limited effort dedicated to the development of text-to-literature search queries. Thus this work introduces an evaluation pipeline and curates a dataset designed to help address this gap, with a particular focus on enhancing the evaluation the quality of literature search queries using a novel approach called *Semantic Precision*.

The evaluation of literature search queries is inherently complex due to several factors. One major challenge is the tendency to retrieve an overwhelming number of publications. In the end, only a small subset is considered relevant. Another challenge stems from the different objectives of the queries constructed. For example, Systematic Literature Reviews (SLRs) aim to identify every potentially relevant publication through exhaustive search strategies. In contrast, Bibliometric Analyses (BAs) focus on defining a large, relevant set of publications to be quantitatively evaluated. A common problem in both approaches is the initial identification of relevant publications within a large retrieved dataset.

[1] https://www.int.fraunhofer.de/
[2] https://www.dimensions.ai/
[3] https://clarivate.com/
[4] https://www.elsevier.com/

To address this issue, we introduce Semantic Precision: a method for assessing the relevance of publications based on their semantic similarity of the title and abstract to a defined set of core publications. This approach is the basis for the construction of an adjusted F_β metric, which includes the recall, the semantic precision, and an additional decay factor. The decay factor allows researchers to tailor the evaluation according to the specific intent of the literature review, whether it aligns with the comprehensive goals of SLRs or the quantitative focus of BAs. By accounting for these elements, our method provides a highly refined and focused framework for evaluating the effectiveness of search queries.

1.1. Motivation

The SQW tool is currently under development by the Fraunhofer INT and has generated interest among researchers internally. However, a primary challenge researchers face after testing earlier versions is evaluating the quality of the generated queries. Initially, we considered gathering human feedback from users by requesting them to rate the generated query on a scale of 0 to 5. While this approach could be useful for fine-tuning the underlying model, the quantity of feedback has so far been limited and remains subjective. This is especially problematic because the tool's purpose is to generate queries for researchers who are new to a given topic. Consequently, if the query quality is poor, the researcher may not immediately recognize this.

Identifying suitable evaluation metrics and datasets to assess the quality of the generated queries is a complex task, which forms the basis of this master's project. The project's objective is to find a robust solution for assessing the quality of literature search queries, enabling the further development of the SQW tool to provide more accurate results and improve productivity through the integration of LLMs.

1.2. Research Question

Our work is driven by a central research question that guides both the curation of the dataset and the formulation of metrics for evaluating the effectiveness of the generated queries. The root of this question is the following hypothesis: Given that we know the most important publications in a given field, referred to as Core Publications (CP), we can design metrics to evaluate the performance of search queries based on their ability to balance relevance and specificity. This leads to the following research questions: **Which metric can effectively penalize the generation of excessively large queries that achieve high recall at the cost of precision?** By addressing this question, we aim to develop an evaluation framework that discourages the trivial exploitation of large query sizes and instead rewards meaningful query design that aligns with the intent and context of the literature search.

1.3. Structure of this Work

The remainder of this work is structured as follows:

After this introduction, we will first focus on the foundations in Chapter 2, where the SQW tool will be briefly explained, primarily focusing on the format of the input and the stages that the SQW consists of. Subsequently, we will explore related works in Section 2.2 and review the currently available datasets, explaining why they are not suitable for our specific use case.

Next, we introduce our framework, which consists of two main components: the curated dataset in Section 3.1 and the evaluation metrics in Section 3.2. In the dataset section, we explain how the data was collected and perform a dataset analysis to gain deeper insights into its characteristics. In the evaluation metrics section, we present the metrics developed to assess the performance of literature search queries.

Following this, we present an evaluation of the framework and showcase the results in Chapter 4. This chapter begins with a description of the conducted experiments, where two types of queries are used: those written for systematic literature reviews (SLR) and those generated by the SQW, which are then used for evaluation.

Finally, we conclude this work with a summary of the main contributions and provide an outlook on future directions in Chapter 5.

2. Foundations

In this chapter, we introduce the SQW, a tool designed to generate literature search queries for evaluation in later stages of the project. We begin by outlining the required inputs for the tool, clarifying how users interact with it. Then, we provide an overview of the two main stages of the SQW, explaining the methodology and rationale behind its design.

We also review related work in the field, focusing on approaches that evaluate literature search queries. We assess the strengths and limitations of these methods and discuss how the queries they generate are evaluated using existing datasets. This review provides essential context for understanding the current landscape of literature query evaluation research, which, to our knowledge, has gaps in the area of automatic evaluation. While existing work offers valuable insights into specific aspects of query evaluation, it often overlooks the variety of factors influencing query performance across diverse datasets and domains. This gap highlights the need for further exploration into more robust and automated

evaluation frameworks.

2.1. Search Query Writer

The SQW is a tool based on an LLM, specifically using GPT-4o, to systematically generate literature search queries. The only required input for this tool, which is the main focus of this work, is the **Topic**. Users are required to provide a topic for generating a search query, irrespective of the scientific field—for example, *Synthetic Biology*.

Several optional inputs are available to enhance the quality of the generated query, including:

- **Negative Keywords:** Terms that should be excluded to avoid unwanted results.
- **Description:** A description that serves as an alignment mechanism to clarify the task's intent.
- **Modes:** Three selectable modes (Strict, Moderate, Creative) that control the temperature of the LLM to manage the level of randomness in responses.
- **Depth:** A parameter that specifies how comprehensively the topic should be analyzed.
- **Supporting Documents:** Users can upload a PDF, ideally a survey or overview document on the topic, which helps the tool acquire knowledge about the scientific field and better align with the research intent.

These additional inputs are intended to refine and tailor the search query to more closely match the user's research goals, but will not be extensively tested in this work.

To generate a literature search query, we designed the SQW to take a human-like approach, divided into two main steps: **Knowledge Enrichment** and **Iterative Scientific Fine-Tuning**.

The objective of the **Knowledge Enrichment** step is to provide the LLM with contextual information about the research topic. This is achieved by first retrieving information from Wikipedia based on the given topic. Specifically, the first 4,000 characters from the top-k pages are collected and summarized before being passed into the LLM's memory. ArXiv is queried in a similar manner to gather relevant research content. Additionally, we perform an online search using DuckDuckGo[1], aggregating results to offer a broader understanding of the topic.

To reduce Recency bias [2], which refers to the tendency of the attention mechanism to favor more recent information, each of the steps is conducted in a

[1] https://duckduckgo.com/

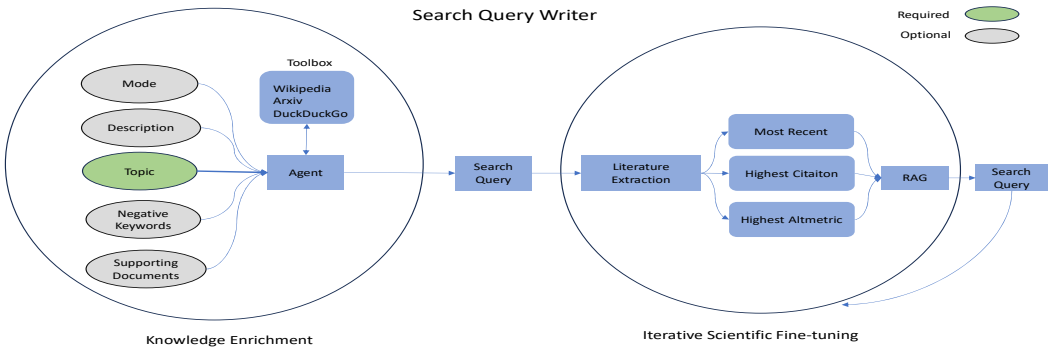

Figure 2.1.: A simplified overview of the SQW. The process begins with the Knowledge Enrichment stage, where the model receives input data and sends it to an LLM agent equipped with a suite of tools to gain insights into the topic. Based on this understanding, the model generates a well-structured search query, formatted and executed across multiple dimensions to retrieve a relevant selection of literature. Within this literature set, RAG is applied to identify the most pertinent keywords, which are compiled into an optimized search query. This query can be iteratively refined to enhance overall search quality.

separate API session, with results stored independently for future use. This approach ensures a that the LLM is biased towards the recently fetched results from prior steps.

The output of this first stage is a list of keywords, typically presented in a transfer-list format, as shown in Figure 2.2. This transfer-list contains two categories: specific and general keywords, along with additional information such as the number of publications found per keyword. The primary goal of this step is to enable the user to assess the relevance of each keyword. If a keyword is deemed too broad, it should be moved to the general list; if it accurately targets the specified topic, it should remain in the specific list. Additionally, the user is required to provide an overarching topic that limits the scope of the general keywords to align more closely with the research intent. The output of this stage will be the queries used for the final evaluation.

The **iterative scientific fine-tuning** on the other hand approaches more scientific sources, namely dimensions.ai, which is a literature database that offers quick access to publications across a wide range of journals. The query generated in the earlier stage is then used to prompt dimensions three times, once to retrieve

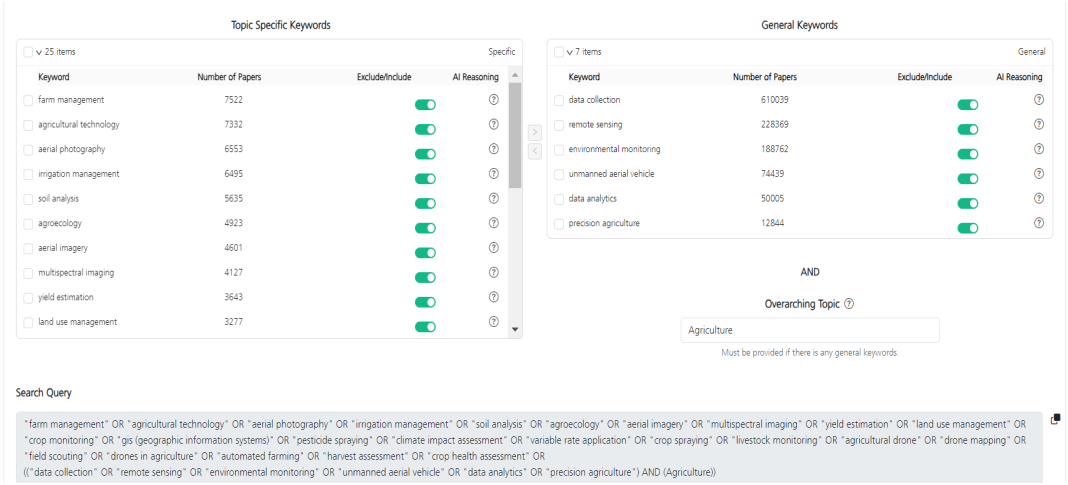

Figure 2.2.: A screenshot of the SQW UI after completing the Knowledge Enrichment stage. On the left, a list of keywords is displayed alongside the number of publications associated with each keyword when used as a search term. The keywords on the right-hand side were manually categorized as general and can be roughly assessed by the number of associated publications. To narrow the scope of general keywords, we selected "agriculture" as the overarching topic. The final generated query is displayed and updated interactively as values in the transfer lists are adjusted.

the most cited 1k literature, a second time to retrieve the newest 1k literature, and one last 1k to retrieve the most relevant literature based on their altmetric rating. The results are then combined, and duplicates are removed, yielding a set of publications. For this set, the titles and abstracts are extracted and processed using OpenAI's embedding model. Finally, a simple RAG pipeline is applied to retrieve publications which are then used to generate relevant keywords based on their content.

2.2. Related Work

SLRs are widely used across various fields, allowing researchers to conduct a comprehensive manual review of scientific topics and identify publications that answer a set of research questions. However, one significant challenge with this approach has been the exponential growth in the number of publications, which makes conducting unbiased reviews increasingly difficult. In the age of technological advancements, we can now use these technologies such as LLMs, Topic Modeling, Semantic Embeddings and much more to assist in investigating topics without the need to manually sift through extensive lists of potentially irrelevant publications. To address this issue, a series of works have been proposed within the Conference and Labs of the Evaluation Forum (CLEF) [5–7]. These works focus on the evaluation of empirical medical research, utilizing a dataset of medical

literature. They introduce two primary tasks: Task 1, which involves identifying relevant studies from the PubMed medical database, and Task 2, which assesses the ranking of studies following title and abstract screening. Notably, the evaluation pipeline, along with the dataset and descriptions of these tasks, are publicly accessible on GitHub[2].

LLMs have had significant impact on modern technology, including in scientific research, where they have provided remarkable improvements in efficiency. While the processing efficiency of LLMs is unprecedented, the quality of their output in various domains is still being explored. The work by Wang [15] investigated the performance of ChatGPT in generating Boolean search queries for literature reviews. Specifically, it evaluated the effectiveness of ChatGPT in constructing queries for SLRs using different prompting techniques, including zero-shot, few-shot and iterative guided approaches. The evaluation used the CLEF datasets [5–7] and an additional medical dataset containing a collection of seeds [14]. Although the results highlight the limitations of ChatGPT's performance, this work underscores the potential of LLMs to aid literature review, especially when supported by examples or more advanced, structured pipelines.

A broader and more diverse evaluation of the quality of automatically generated literature search queries for SLRs was conducted by Badami [1]. In this work, they introduced a pipeline that generates literature search queries based on a given research question and abstracts from previously identified relevant publications. The evaluation was performed against a dataset they constructed, which contains the results of 10 SLRs, including candidate papers, queries used, and relevant papers identified in each review. For example, in the review SLR_1, a total of 7,002 candidate papers were retrieved using search query S, from which a subset of 59 relevant papers RP was identified. To assess their proposed approach, they compared the generated queries in various settings using recall and precision metrics, benchmarking them against the original search query S. The dataset is publicly available on Zenodo[3].

3. LitQEval

Despite ongoing research on automatic literature query generation and related evaluations with medical datasets, such as CLEF [5–7] and the Collection of Seeds [14], the insights gained from these evaluation metrics are not particularly compelling for our use case. This limitation arises from two main factors.

First, the CLEF and Collection of Seeds datasets are exclusively focused on medical data. Although Badami's work [1] offers a more diverse dataset, it lacks a suitable evaluation metric. Their evaluation primarily aims to maximize recall,

[2] https://github.com/CLEF-TAR/tar/tree/master
[3] https://tinyurl.com/496zuar3

with minimal consideration for precision, as literature search queries often yield far more results than necessary, making precision a less effective measure in this context. A second limitation arises when recall is prioritized exclusively. For example, if we aim to train a model to generate queries that maximize recall, there is no penalty for generating overly broad queries, such as those that exploit wildcards, which could lead to an excessive number of irrelevant results. To address these issues, we introduce a dataset structured similarly to that of Badami [1] but designed to be more comprehensive and covering a wider range of research fields. Alongside this dataset, we propose new evaluation metrics that account for the inherently broad nature of literature search queries while penalizing excessively large queries. These metrics also emphasize the importance of accurately identifying core publications that are considered highly relevant within the field.

3.1. Dataset

The dataset we aim to create has three primary goals: First, it should encompass a wide range of randomly selected scientific research fields. Second, for each selected field, it should contain a set of highly relevant publications to serve as the CPs for evaluating additional publications found in these areas. Lastly, the data should consider different research intents, meaning a publication that is considered relevant by a bibliometric analysis (BA) might not be relevant for a Systematic Literature Review (SLR) work.

Selecting a new research fields is straightforward; however, to avoid biases from ongoing research interests, we used ChatGPT to generate a list of scientific fields that are recent and not overly broad. For instance, a field like *Artificial Intelligence* is vast, making it challenging to accurately and comprehensively identify core publications. Instead, we chose more specific, problem-focused fields such as *Drones in Agriculture*. To search for the corresponding BAs and SLRs we used the following query: *<FIELD> AND ("Bibliometric" OR "Scientometric" OR "Systematic literature" OR "Most Influential" OR "Most Cited" OR "Scientific Landscape" OR "Literature Landscape" OR "Core Literature")*

After identifying a sufficient number of diverse fields, 14 in our case, we sought to collect core publications for each field. Due to the difficulty of gathering core publications across a broad array of fields, we use some results from the bibliometrics community. Specifically, we searched for bibliometric studies that identify the most relevant publications within each research area. For example, a bibliometric analysis of *Drones in Agriculture* [11] lists the most cited publications from 1990 to 2021. In this case, 40 core publications were listed, which we manually allocated on Dimensions.ai and added to our dataset, omitting 15 that publications not found in Dimensions, resulting in a total of 25 core publications.

This process was repeated across all selected research fields, resulting in a dataset comprising 14 fields, each containing 25–50 core publications, as shown in Figure 3.1. For the SLR data, we used previously collected data [1] in the field

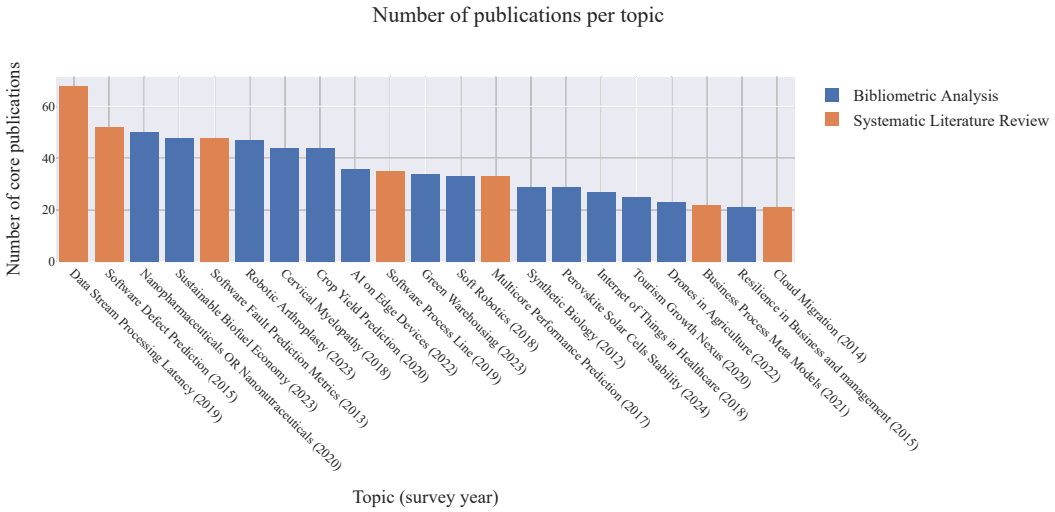

Figure 3.1.: An overview of the dataset and the selected 21 research fields with respective core publications identified through bibliometric analyses or systematic literature review. The number in brackets following the field name on the x-axis represents the year of survey publication.

of Software Engineering. Notably, the SLR data used here were replicated by executing the original query in Dimensions. However, only 7 out of the 10 original datasets were included, with SLR 2, 5, and 6 omitted due to extreme variations between the original datasets and the results retrieved from Dimensions. For instance, SLR 2 originally contained 8,911 candidate papers, but when executed in Dimensions, it yielded approximately 200,000. This discrepancy is attributed to the original SLR queries including additional constraints along side the query string, such as limiting the time span to between 2000 and 2010 or filtering by specific journals. These filters were intentionally excluded in this work, as the aim was to evaluate the performance of the query strings alone without any domain-specific adjustments. In total, the final dataset encompasses 21 fields.

Dataset Analysis

We recognize that potential biases may exist in our dataset due to its complete reliance on the bibliometric community for identifying core publications. This often implies that publications with higher citation counts are considered more relevant. To assess this, we analyzed the citation distribution per field, as provided by Dimensions, shown in Figure A.1. Additionally, we examined the distribution of publication years per field, illustrating the time span considered in the bibliometric analyses, as shown in Figure A.2. If we compare the distribution of publication years for the medical research field *Cervical Myelopathy* with that of *IoT in Healthcare*, both of which were published in 2018, we can observe distinct

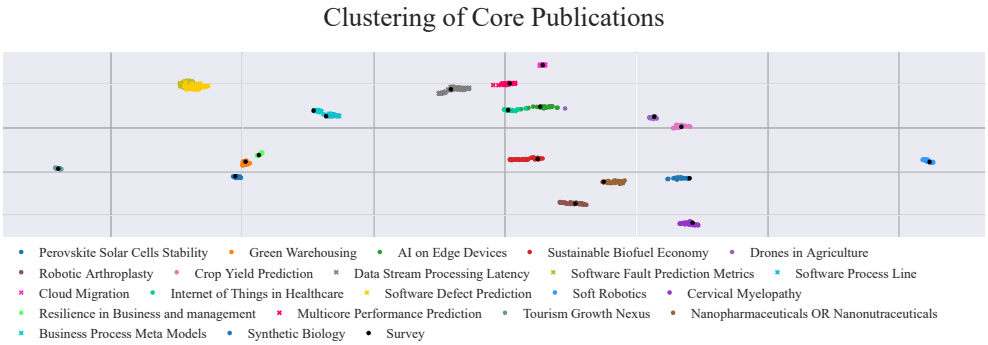

Figure 3.2.: This figure shows clusters of publication embeddings based on titles and abstracts, the ones marked with circles are from the BAs and crosses from SLRs. Embeddings were generated using OpenAI's small model, which yields high dimensional vectors. These vectors were clustered using K-means with $k = 21$. To enhance interpretability, the clustered high-dimensional embeddings were projected into two dimensions using UMAP. The colors in the figure correspond to the true fields of the publications, with the objective of testing the hypothesis that embeddings can effectively group publications from the same field into cohesive clusters. We see that the resulting clusters contain points from the same color, which indicates that the generated embeddings are meaningful and capable of capturing the semantic relationships between publications.

differences in the year distributions of their core publications. These variations may be attributed to factors such as the recency of the field, changes in terminology over time, or the nature of the research area, where one field may prioritize more established works while the other focuses on recent advancements.

For the evaluation pipeline that we will introduce, the embeddings of the core documents are essential for effectively assessing the search query, as detailed in Section 3.2. To validate this approach, we examine the clustered embeddings of the titles and abstracts for each core publication, as well as the bibliometric analyses survey in which these documents were initially referenced. This enables us to assess whether core publications within each field exhibit semantic similarity while also demonstrating some degree of dissimilarity from publications in other fields. The resulting clusters, shown in Figure 3.2, were generated using k-means clustering, where k is set to 21, which is the number of research fields in the dataset. To generate the embeddings we use OpenAI's small model alongside UMAP [10] to reduce the dimensionality to 2D.

3.2. Evaluation metrics

The standard metrics for query evaluation are recall and precision. We argue that both are of high importance, given that recall indicates how many of the

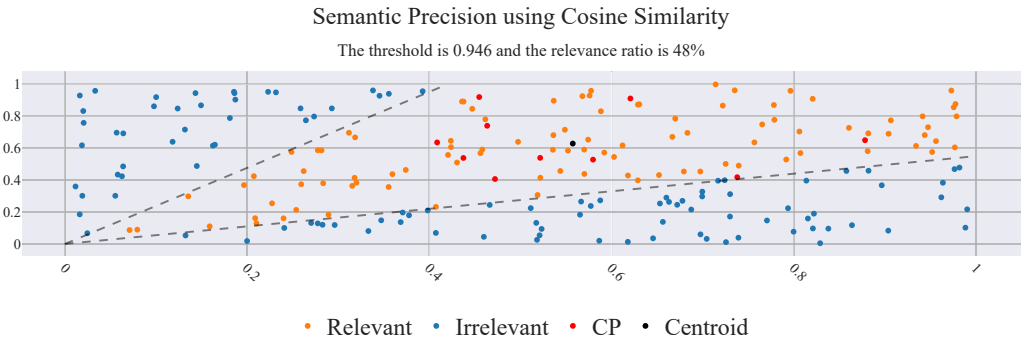

Figure 3.3.: This illustration demonstrates the effect of cosine similarity on randomly generated data in a 2D space. The core publications (CP) are shown in red, positioned between 0.4 and 0.9 on both the x- and y-axes. When we set the threshold to 0.946, based on the cosine similarity of the least similar core publication from the centroid, many retrieved publications on the opposite side of the spectrum are still assigned as relevant. This effect occurs because cosine similarity considers only the angle between vectors, ignoring their magnitude. In this case, this results in 48% of the retrieved publications being considered relevant.

relevant publications we have retrieved, while precision measures the size of relevant publications in comparison to the total number of retrieved results. An issue with precision is that anything other than the core publications is considered a true negative, which is not always true. To address this issue, we introduce the *Semantic Precision Heuristics*, which refines the standard precision by allowing non-core publications that are semantically similar to core publications to be considered a false negative.

The idea behind Semantic Precision is to evaluate the relevance of retrieved publications in comparison to the core publication set. If the retrieved publications are sufficiently similar to those in the core set, they are deemed to hold some relevance rather than being entirely unrelated. To achieve this, we assume that the core publications, encompass sufficient semantic breadth to gauge the quality of literature relevant to a specific field. We calculate Semantic Precision in three ways.

Semantic Cosine Precision

The first approach involves averaging the embeddings of the core publications. We then set an acceptance threshold based on the cosine similarity to the least similar core publication. This means that if the embedding of a retrieved publication is more similar to the center than the least similar core publication, we consider it a relevant publication, as shown in Figure 3.3. To do so we define:

- ECPs as the embeddings of the core publications.

- ERPs as the embeddings of the retrieved publications.
- $\cos(\mathbf{a}, \mathbf{b})$ as the cosine similarity between two vectors $\mathbf{a}$ and $\mathbf{b}$.

First, compute the centroid of the core publication embeddings:

$$\mathbf{c}_{\text{centroid}} = \frac{1}{|\,\text{ECPs}|} \sum_{\mathbf{c_i} \in \text{ECPs}} \mathbf{c_i}$$

Then, let the threshold similarity, θ, be the cosine similarity of the least similar core publication to the centroid:

$$\theta = \min_{\mathbf{c_i} \in CP} \cos(\mathbf{c}_{\text{centroid}}, \mathbf{c_i})$$

Finally, Semantic Precision using cosine similarity ($\text{SP}_{\cos}$) is defined as Equation 3.1, where $\mathbb{I}$ is an indicator function that equals 1 if the retrieved publication $\mathbf{r}$ meets the similarity criterion and 0 otherwise:

$$\text{SP}_{\cos} = \frac{\sum_{\mathbf{p} \in \text{ERPs}} \mathbb{I}\left(\cos(\mathbf{c}_{\text{centroid}}, \mathbf{p}) \geq \theta\right)}{|\,\text{ERPs}|} \tag{3.1}$$

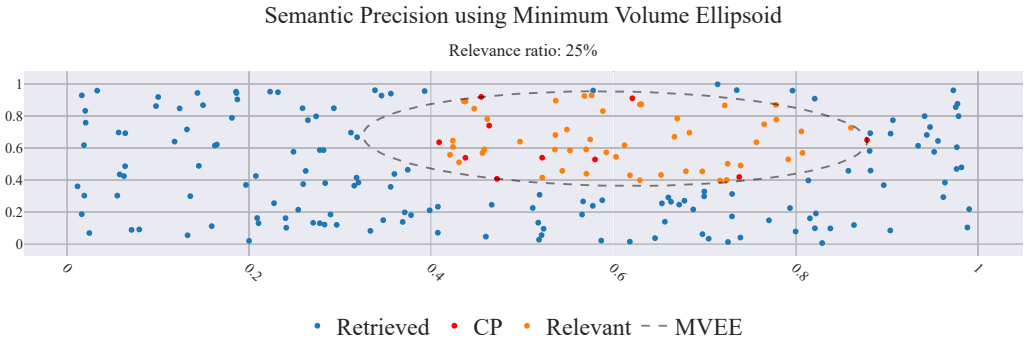

Figure 3.4.: This illustration demonstrates the effect of using the Minimum Volume Enclosing Ellipsoid (MVEE) on randomly generated data in a 2D space. The core publications (CP) are shown in red, positioned between 0.4 and 0.9 on both the x- and y-axes. An ellipsoid is generated using MVEE to define the scope of relevant publications, ensuring that only those within the maximal angles and magnitudes of the core publications are considered relevant. In this case, this approach results in only 25% of the retrieved publications being classified as relevant.

Semantic MVEE Precision

For the second approach, we omit the averaging of the embeddings and use a Minimum Volume Enclosing Ellipsoid (MVEE) [13], which creates the smallest

ellipsoid that includes our CP. We then use the ellipsoid to determine which of the retrieved publications are relevant by checking whether they are within MVEE or not, as illustrated in Figure 3.4. This approach allows us to take into account all the dimensions by not only considering the angle but also the magnitude

The MVEE for the core publication set CP is centered at δ with shape matrix A. To determine whether a retrieved publication p is relevant, we check if it lies within the ellipsoid by testing the following condition:

$$(\mathbf{p} - \delta)^T A(\mathbf{p} - \delta) \leq 1$$

Semantic Precision MVEE ($\mathrm{SP}_{\mathrm{MVEE}}$) for this approach is then:

$$\mathrm{SP}_{\mathrm{MVEE}} = \frac{\sum_{\mathbf{p} \in \mathrm{ERPs}} \mathbb{I}\left((\mathbf{p} - \delta)^T A(\mathbf{p} - \delta) \leq 1\right)}{|\,\mathrm{ERPs}|} \tag{3.2}$$

While MVEE generates a valid ellipsoid that captures the core publications and everything in between, it is not very robust against outliers. If a core publication is located far from the rest, MVEE may produce an overly large ellipsoid, thereby capturing a lot publications. To investigate this potential problem, we also evaluate using of the convex hull [8], which is the smallest convex set that encloses all the points by forming a polygon rather than an ellipsoid. A potential advantage of this approach is its greater robustness to outliers compared to MVEE.

Semantic Clustering Precision

For the final semantic precision approach, we apply a simple clustering algorithm, such as k-means, on the document embeddings. We iteratively adjusts the number of clusters K, starting with $K = 2$, and increasing its value until we get the smallest possible cluster that can allocate a certain percentage of the CPs. This is done as follows:
where:

- θ: Threshold parameter $\theta \in (0, 1]$ that determines the stopping condition.
- N: The maximum number of clusters.

The best case would be finding a cluster that has the size of C_{total}, whereby the worst case is when the threshold is only met if everything is clustered together, namely $k = 1$. All the above semantic precision metrics aim to identify potential true positives that were initially not considered as CPs. However, a key issue arises when the number of semantically relevant publications is large due to the broad scope of the initial query.

For instance, if a query retrieves 50,000 publications, with 30,000 deemed relevant, this still poses a challenge. Screening such a large volume of documents is infeasible, making the results problematic. To address this, we introduce a decay factor that can be multiplied with any of the introduced semantic precision metrics as a weighting factor. It is defined as follows:

Input: θ, N, ERPs
Output: Clustering Semantic Precision ($\mathrm{SP}_{\mathrm{clust}}$)
Initialize $K \leftarrow 2$;
$C_{\mathrm{total}} \leftarrow$ Total number of core publications;
while $K \leq N$ **do**
  Perform clustering on document embeddings into K clusters;
  $C_{\mathrm{ncores}} \leftarrow$ Number of CPs in the cluster that contains the most CPs;
  if $\frac{C_{ncores}}{C_{total}} \leq \theta$ **then**
    $C_{\mathrm{best}} \leftarrow$ Compute the clustering using $K-1$, and select the cluster with the most CPs;
    $\mathrm{SP}_{\mathrm{clust}} \leftarrow \frac{|C_{\mathrm{best}}|}{|\mathrm{ERPs}|}$;
    return $\mathrm{SP}_{\mathrm{clust}}$;
  end
  else
    Increase K by 1;
  end
end

Algorithm 1: Iterative Clustering

- p: Controls the initial slowness of the decay.
- q: Controls the acceleration of the decay near the end.
- α: The maximum threshold for the decay, representing the point at which the decay becomes negligible.

The decay function is expressed as:

$$\lambda = \left(1 - \left(\frac{n_{\mathrm{pubs}}}{\alpha}\right)^{p}\right)^{q}$$

The decay factor is designed as a hyperparameter that can be adjusted based on the specific task requirements. In the case of BAs, even though the number of CPs in a given topic may be limited, analysts often examine a broader set of publications to provide a more comprehensive overview of the field. An acceptable sample size for a BA is typically around 1,000 publications [12]. Beyond this point, the decay function ensures that the contribution of additional publications diminishes progressively, preventing an overwhelming volume from disproportionately influencing the semantic precision. This allows for a nuanced balance between breadth and relevance.

Now that we have metrics that can be used to penalize the model in case of generating a too broad of a query, we use can use it as a factor to calculate the F-Score, the goal of the standard F-score is to balance out between the recall and precision, but in our case we use F_β instead, whereby the β is the weighting factor of the recall, meaning the higher it is the more important the recall will

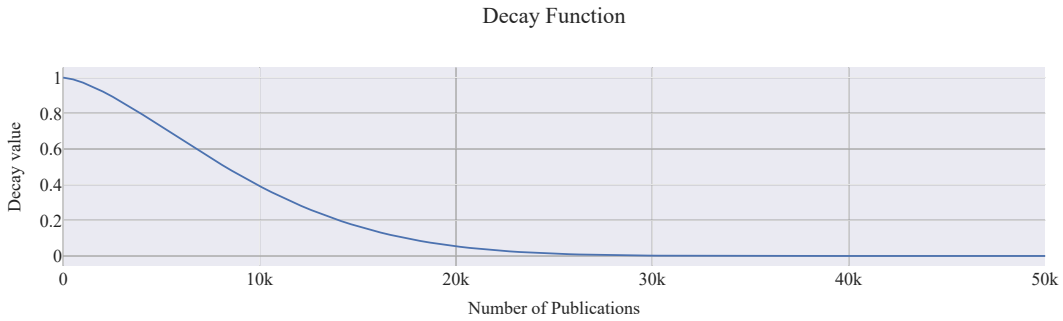

Figure 3.5.: This illustration demonstrates the effect of the decay factor, which ensures that the contribution of a large number of publications diminishes as the total count approaches the threshold. This prevents an overwhelming volume from biasing the semantic precision. For this example, we set the threshold (α) to 50k, $p = 1.5$, and $q = 10$.

be, in our case we set it to be 2, meaning that we want the recall is twice as important as the precision.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \tag{3.3}$$

For our specific case where $\beta = 2$, emphasizing the importance of recall, it is:

$$F_2 = 5 \cdot \frac{(\text{Precision} \cdot \lambda) \cdot \text{Recall}}{(4 \cdot \text{Precision} \cdot \lambda) + \text{Recall}}$$

Comparative Analysis

To further understand the metrics and their impact on evaluating the dataset, we conduct an in-depth analysis using a randomly selected topic, *Soft Robotics* and used its baseline query as a case study.

First, we visualize the embeddings of the baseline and predicted queries Figure 3.6. The baseline query is the exact topic name, *Soft Robotics*, while the predicted query is generated by the SQW. The embeddings are derived from the title and abstract of the retrieved publications and subsequently reduced to a 2D UMAP [10] space. It is important to note that a significant amount of information is likely lost due to the extreme dimensionality reduction from 1536 dimensions to 2.

Semantic Cosine Precision

At first, we test the Semantic Cosine Precision using the high-dimensional original embedding E_o, which was done as described in Equation 3.1. This resulted in 13,265 out of 17,573 publications being classified as relevant Figure 3.7. However, this high proportion of relevant publications appeared excessive, prompting

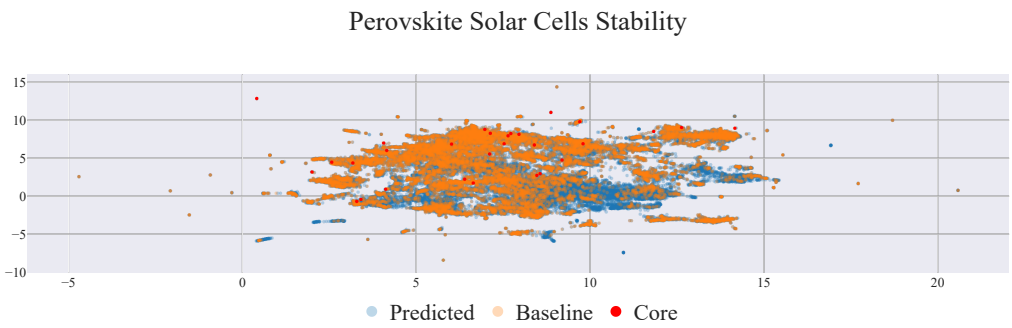

Figure 3.6.: This figure visualizes the distribution of publications retrieved by both the baseline and predicted queries in a 2D space. The baseline query retrieved 20 core publications, whereas the predicted query retrieved 26 core publications out of a total of 36.

further investigation into the threshold's effect on the number of semantically relevant documents.

Since we aim to use the F_2 score as our primary evaluation metric, we also factored it in the cost function which we want to maximize. The objective is to determine the optimal empirical threshold that balances retrieving core publications while controlling the number of false positives. The analysis presented in Figure 3.8 indicates that the optimal threshold is approximately 0.69. This means that publications scoring above this threshold are accepted. Interestingly, the results suggest that sacrificing a small number of core publications can be beneficial, as it helps reduce the overall number of semantically relevant but non-core publications, ultimately improving the precision of the retrieved set.

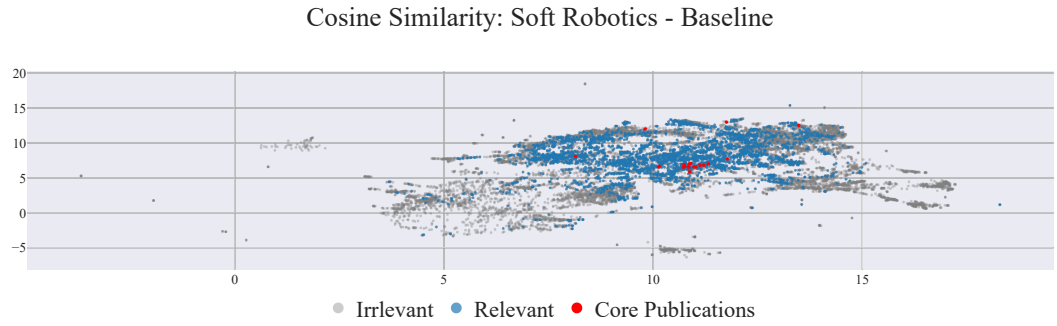

Figure 3.7.: This figure illustrates the publications identified as relevant using the cosine similarity measure with the threshold θ defined by Equation 3.1, which in this case was approximately 0.547. The query results included all core publications, as expected, but also classified 75% of the total retrieved publications as relevant.

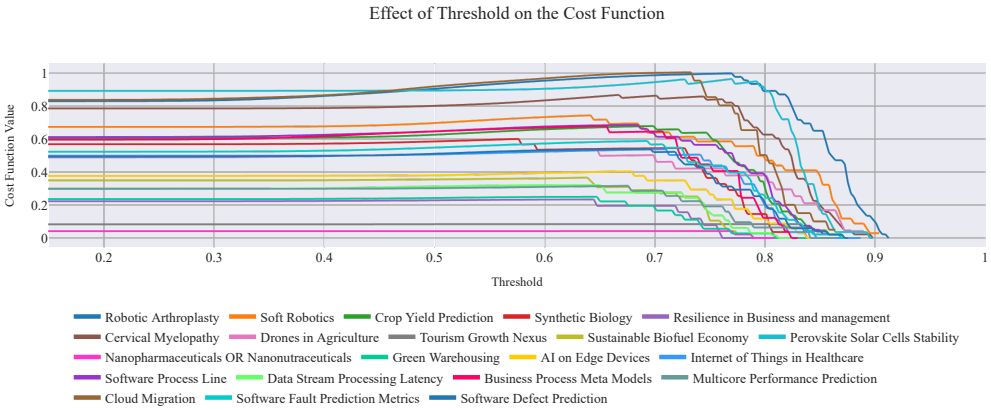

Figure 3.8.: This figure illustrates the effect of the threshold on the F_2 score. As the threshold increases, the number of semantically relevant publications and core publications identified decreases. However, in some cases, such as *Perovskite Solar Cells Stability*, the F_2 score continues to improve despite the loss of a core publication. This outcome is due to the F_2 score weighting recall twice as much as precision, allowing for stricter relevance criteria while sacrificing a single core publication.

After setting the threshold to the optimal empirical value, the Semantic Cosine Precision retrieves 19 out of the initially found 20 core publications while reducing the number of semantically relevant publications by a factor of 4. This adjustment results in only 3,424 publications being identified as relevant, compared to the initial 13,265. However, this refinement comes at the cost of missing only one core publication.

Semantic MVEE Precision

In contrast to Semantic Cosine Precision, we opt to use the 2D embeddings generated by UMAP (E_{umap}), rather than the original high-dimensional embeddings (E_o). This decision was made, because earlier evaluations of the dataset showed that the MVEE consistently classified at least 50% of the total retrieved documents as relevant, which we believe is related to the high-dimensional nature of the embedding vectors and the fact that their euclidean norm always is one, meaning that they have a constant magnitude.

We experiment with two enclosing shapes: has an ellipsoidal shape, and the Convex Hull, which forms a polygonal boundary. The primary difference is that the MVEE tends to be larger due to its ellipsoidal shape, whereas the convex hull strictly bounds the points. Using the MVEE approach, 9,595 publications were identified as relevant out of the total 17,573 publications, as shown in Figure 3.9. In contrast, the convex hull, being smaller as expected, identified 7,609 publications as relevant, as illustrated in Figure 3.10.

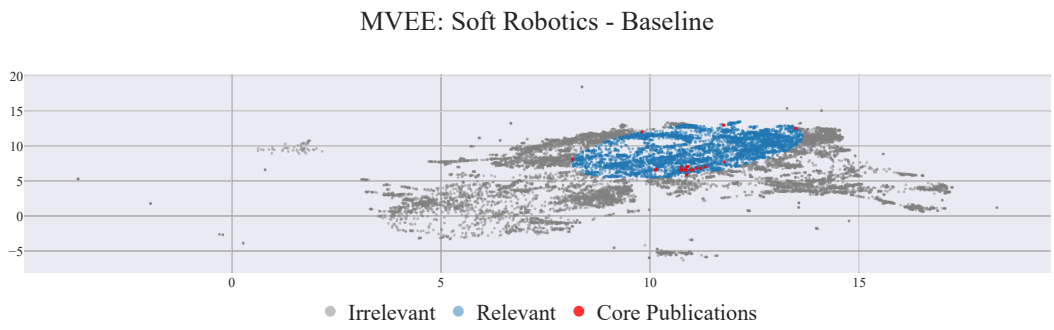

Figure 3.9.: This figure shows the relevant publications identified by the MVEE. An advantage of this approach is that we always expect that all the core publications to be included in the identified publications.

For further evaluation, understanding the quality of the UMAP embeddings (E_{umap}) is crucial, as they do not retain the same level of semantic meaning as the original higher-dimensional E_o. Unlike PCA, which is a linear transformation, calculating the exact semantic loss for UMAP embeddings is challenging due to its nonlinear nature. To approximate this loss, we utilized a Partial Least Squares Regression approach as outlined by Oskolkov[1]. Based on this method, we estimated that the explained variance of the two-dimensional UMAP embeddings is only 7.15%. This low explained variance underscores the reduction in information captured when transitioning from high-dimensional to two-dimensional space.

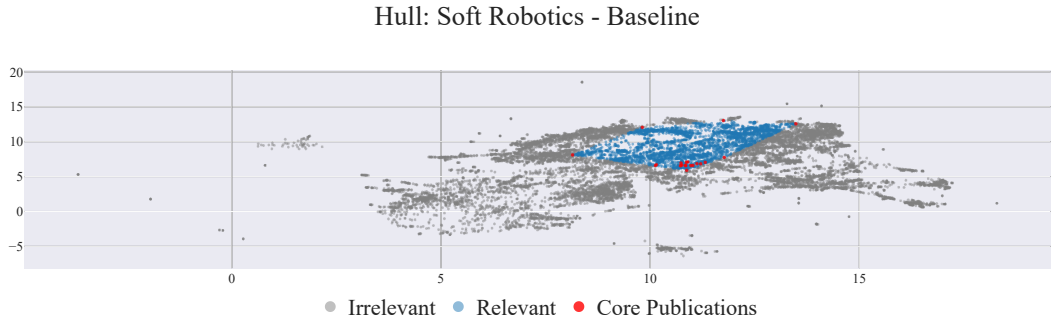

Figure 3.10.: This figure shows the relevant publications identified by the Convex Hull

Semantic Clustering Precision

To cluster the embeddings, we apply algorithm 1 using K-means with $N = 100$ clusters and a threshold of $\theta = 0.7$. The algorithm aims to identify the smallest

[1] https://towardsdatascience.com/umap-variance-explained-b0eacb5b0801

cluster that contains at least θ of the core publications. For this process, we use the high-dimensional embeddings, E_o, as input. Cosine similarity is used as the distance measure between points which requires input normalization, which is already handled by the output of OpenAI's text-embedding-3-small model.

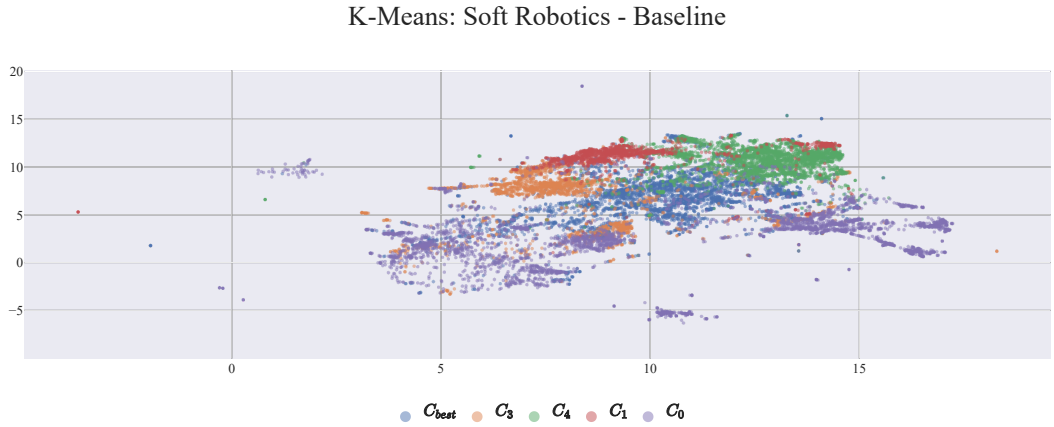

Figure 3.11.: This illustration shows the grouping of the embeddings E_o in higher-dimensional space using K-means, where $K = 5$ is identified as the optimal number of clusters (C) that contains C_{best}, which has at least 70% of the core publications. The spread-out nature of the clusters is uncommon in K-means but occurs here because clustering is performed in the higher-dimensional space before reducing the data to 2D using UMAP for visualization.

The clustering results Figure 3.11 indicate that the space can be divided into 5 groups. C_{best} has a size of 4,954 out of 17,573 publications and contains 15 out of the 20 core publications. We experimented with adjusting the threshold to match the quality of cosine similarity by increasing it to the next best solutions. At a threshold of approximately 0.75, the results included 6,861 publications with 17 core publications. At a threshold of approximately 0.85, all 17,573 publications were clustered together.

Additionally, we clustered the UMAP embeddings E_{umap}, using the same thresholds (0.7, 0.75, and 0.85). These thresholds consistently yielded similar results, with 11,644 out of 17,573 publications identified as relevant and 19 out of the 20 core publications included. This consistency can serve as an indicator of the information loss incurred when using UMAP.

As mentioned in Section 3.2, the F_β score is the metric we will use for evaluation, with $\beta = 2$, emphasizing recall by making it twice as important as precision. However, we extended the traditional F_β score by incorporating components such as semantic precision in place of typical precision and a decay factor to account for the number of semantically relevant retrieved publications. To better understand the influence of each component, we visualize their effects in Figure 3.12.

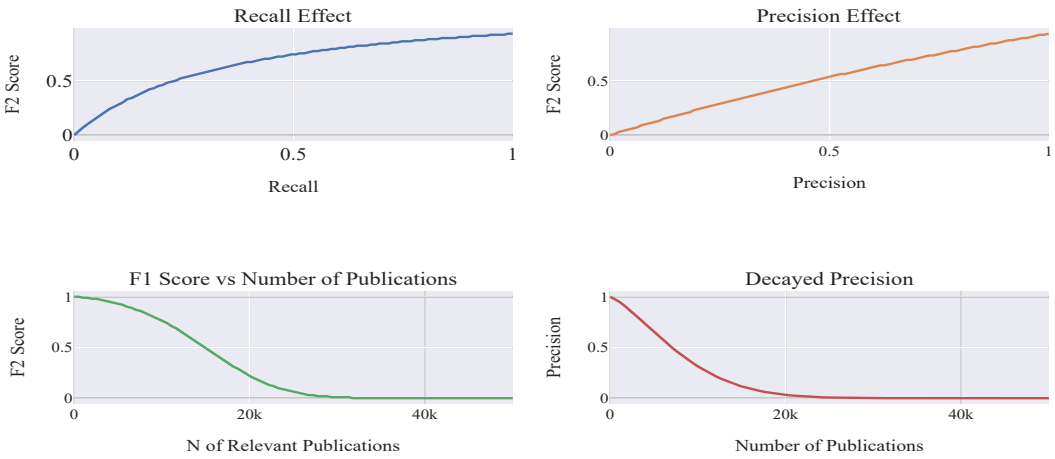

Figure 3.12.: This illustration demonstrates the effect of each component in the F_β score evaluation metric, where $\beta = 2$ and the decay hyperparameters are set to $p = 1.5$, $q = 10$, and $\alpha = 50000$. In the top left, we observe the impact of $\beta = 2$, which ensures better scaling even with lower precision if the recall is 1. The top right plot shows how the score scales almost linearly with the precision. The bottom left and right plots depict the dampening effect on the F_β score as the number of relevant publications increases, emphasizing the importance of controlling the decay to prevent inflated scores from overly large results.

4. Evaluation

In this section, we evaluate the performance of literature search queries based on the introduced metrics. This evaluation serves as a foundation for developing tools that can potentially generate automatic literature search queries in the future. It is crucial to note that the objective of this evaluation is not to only assess the SQW tool itself, but rather to evaluate any arbitrarily generated literature search query. Thus, the focus is solely on the quality of the query, independent of the method by which it was generated.

4.1. Experimental Setup

The curated dataset is constructed using two distinct methods to identify core publications: Bibliometric Analysis (14 topics) and Systematic Literature Review (7 topics), as illustrated in Figure 3.1. For the SLRs, the original queries used by the researchers are available. Consequently, we conduct two main experiments.

In both experiments we use Dimensions.ai to retrieve all required data. The retrieval process relies on their default relevance-based sorting method, which ranks publications based on the number of keyword matches between the title-abstract and the provided query.

The first experiment involves all 21 topics from both the SLRs and BAs, where we compare a baseline query against a query generated by the SQW. The baseline query consists of the exact topic name, passed into the search engine in a non-exact search fashion. For instance, the query *Soft Robotics* retrieves publications containing both words in their title or abstract, even if they do not appear consecutively.

The predicted query, however, is semi-automatically generated using the SQW tool. This process begins by providing the baseline query as input, which generates a list of keywords. These keywords are then manually sorted by the author into specific or general categories, as described in Figure 2.2. The overarching topic is derived from the topic itself; for example, in the case of *Soft Robotics*, the overarching keyword *Robot* is used. In some cases, the resulting queries produced excessively large results (>100k publications). To address this, keywords were filtered to limit the results to a maximum of 50k publications, balancing evaluation cost and processing speed. Importantly, the baseline query is always included in the predicted query. This ensures that recall is at least as high for the predicted query as for the baseline, making the primary goal of the evaluation to determine whether the expanded query retrieves more core publications than the baseline without becoming overly general by retrieving irrelevant publications.

The second experiment focuses exclusively on the 7 SLR topics. It uses the exact baseline queries and results from the first experiment but compares them to SLR queries manually crafted by experts in the field rather than those generated by the SQW. These SLR queries are designed with well-defined research questions aimed at retrieving the most relevant publications that help tackle these exact questions.

4.2. Results

Using the data from the first experiment, we computed all the metrics, namely: Cosine Precision, Clustering Precision, MVEE Precision, Hull Precision, Recall, and the F2 score for each precision metric, as shown in Figure Figure 4.1. When examining the precision metrics, the clustering precision distinctly stands out due to its high value in certain cases, which can be directly attributed to low recall. This recall issue is also evident in some instances for the MVEE and Hull metrics, such as the baseline for *Drones in Agriculture*, where they are set to 0 because fewer than three retrieved core publications are available, which is the minimum number required to define a plane.

A strong correlation is observed between cosine precision and the MVEE and Hull methods, despite relying solely on UMAP embeddings to define the enclosing

shapes. This highlights the robustness of these approaches in identifying semantically relevant publications. Additionally, we have two special case topics that had 0 recall, namely *Cloud Migration* and *Multicore Performance Prediction*. As expected, these resulted in a 0 across the board except for the cosine similarity, since it does not require any of the retrieved core publications to exist in order to compute. Instead, it only relies on the pre-computed average embeddings of the core publications.

Considering the F2 score, a notable example of the impact of overly large queries without any recall improvement is *Robotic Arthroplasty*. Both the baseline and predicted queries achieved a recall of 0.957, but the expanded predicted query from the SQW retrieved significantly more results overall. Specifically, the predicted query retrieved 22,892 publications, of which only 2,834 were relevant based on cosine similarity. In contrast, the baseline query retrieved 2,151 docu-

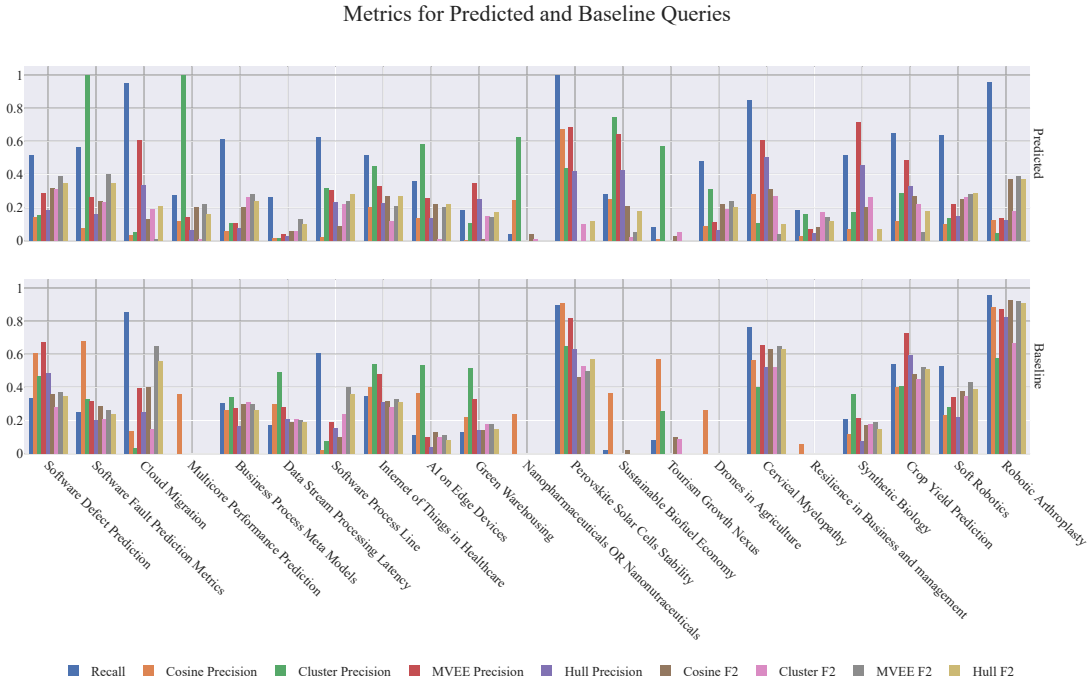

Figure 4.1.: This figure shows the results of the first experiment across all the datasets. Initially, the issue with the clustering, MVEE, and Hull precision metrics becomes apparent in cases such as *Cloud Migration* and *Multicore Performance Prediction*, where the value is 0. This occurs due to a recall that is <3 for MVEE and Hull and recall of 0 for the clustering. On the other-hand the impact of the crafted F2 score is particularly evident in cases like *Robotic Arthroplasty*, where the baseline score is very high. Conversely, for the predicted query, which retrieves more publications but maintains the same recall, the score is significantly lower.

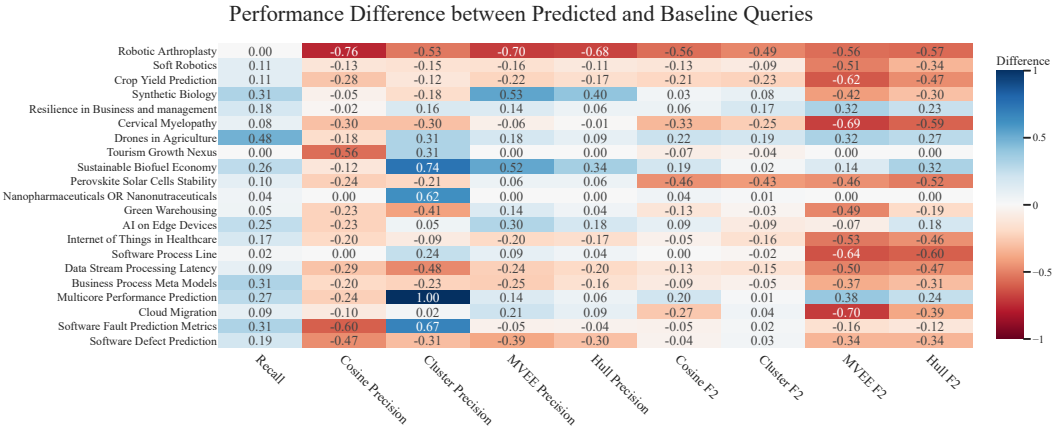

Figure 4.2.: In this figure we can see the difference in values between the predicted query from the SQW and the baseline, whereby a negative value means that the baseline is better. As anticipated we at least always achieve a similar recall, but in most cases, the SQW yields better recall. However, it severely suffers in precision. When looking at the F2 value, we can see that the tool only notably outperforms the baseline on the three topics *Drones in Agriculture*, *Sustainable Bio Fuel Economy*, and *Multicore Performance Prediction*, whereas it shows a clear disadvantage on the topics *Perovskite Solar Cells Stability*, *Robotic Arthroplasty*, and *Cervical Myelopathy*.

ments, with 1,904 classified as relevant. This demonstrates how an excessively large query can dilute the precision without improving recall or the number of relevant documents retrieved.

In Figure 4.2, we can better interpret the results of the first experiment by examining the differences between the scores of the predicted query and the baseline. Here, positive values indicate that the predicted query performs better, while negative values show that the baseline outperforms the predicted query.

As expected, the predicted query consistently achieves similar or better recall across all topics due to the inherent nature of the SQW. However, when evaluating precision, it is evident that the broader queries generated through query expansion often degrade the performance of the model. This effect is particularly visible in the F2 scores, where the increased number of irrelevant publications impacts the balance between recall and precision.

While the SQW demonstrates advantages in terms of recall, its over-expansion often leads to excessive noise in the results. This trade-off is especially clear for topics with a significant drop in precision or F2 scores due to the broader query scope.

The results of the second experiment, which compare the actual search queries used to identify the core publications in the SLRs, have yielded surprising yet explainable outcomes. It is important to re-emphasize that the queries initially used for the SLRs were adapted to fit dimension's query criteria and were only

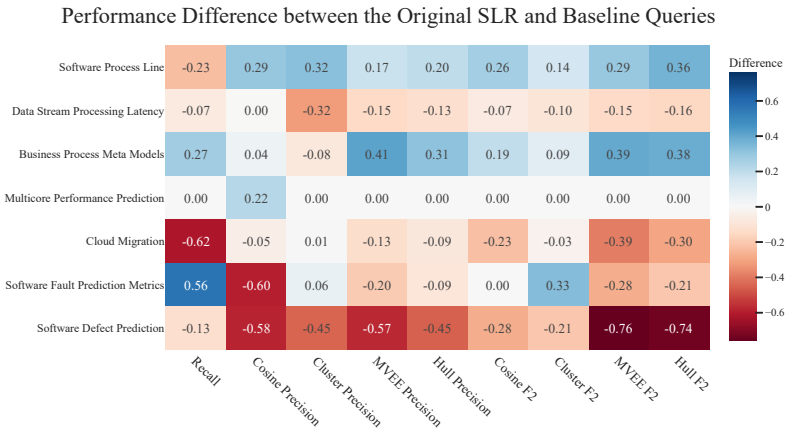

Figure 4.3.: This figure displays the metric values difference between the original SLR query and the baseline, where a negative value indicates that the baseline performs better.

applied to search the title and abstract. In contrast, the original queries were utilized across a variety of search indices, such as title, abstract, full text, and sometimes full data for specific fields, which is a form query fine-tuning that is search engine specific. Therefore, it is important to note that the results will always be search engine dependent, which, in this case, is dimensions.ai.

As shown in Figure A.3, the results of the SLR queries are not as anticipated, particularly since they were expected to yield high recall. This discrepancy arises from the queries' reconstruction and adaptation to fit dimensions' criteria. The performance difference between the SLR queries and the baseline can be observed in Figure 4.3. Overall, the baseline and the SLR queries performed on nearly equal footing, with two topics favoring the SLR and three favoring the baseline.

Interestingly, in case where the recall of the baseline significantly outperformed the SLR, namely *Software Fault Prediction Metrics*, the cosine precision was drastically lower. This results in an unwanted behavior which can be interpreted by the cosine-F2 score being 0, indicating that the recall gain was of no value due to the excessive number of irrelevant retrieved publications.

4.3. Discussion

The primary goal of this work was to determine the quality of literature search queries, emphasizing recall, which is widely regarded as an essential measure in the research community. However, precision has historically been less emphasized due to the intrinsic nature of literature search queries, which tend to favor comprehensiveness over specificity. By developing multiple metrics to evaluate the relevance of publications in a semantic space, we successfully integrated precision into the evaluation framework by introducing semantic precision.

To calculate semantic precision, we employed four metrics: semantic cosine, clustering, MVEE, and Hull precision. Each metric demonstrated distinct advantages and limitations. Initially, we hypothesized that the cosine similarity threshold should correspond to the least similar core publication. However, in certain cases, sacrificing a core publication to substantially reduce the number of irrelevant publications proved more beneficial in terms of the F2 score. Consequently, we adopted an empirical threshold estimated by maximizing the F2 score across all topics.

For Convex Hull and MVEE, we first tested their performance using the original embeddings in their high-dimensional space. However, these metrics consistently overestimated precision, often exceeding 50%. This discrepancy is likely due to the curse of dimensionality, which complicates the construction of accurate ellipsoids or hulls in high-dimensional spaces, possibly reflecting limitations in the embedding construction process. This observation led to the decision to switch to UMAP embeddings, which offer a reduced dimensionality and improved computational feasibility. However, while UMAP embeddings show potential, the exact amount of semantic value lost compared to the original high-dimensional embeddings remains unclear.

A common issue for clustering, MVEE, and Hull methods lies in their dependency on recall. Semantic clustering requires at least two core publications, and performance improves with more core publications as the clustering process narrows the focus on relevant publications. This limitation becomes problematic when the viable core publications count is less than two. Similarly, MVEE and Hull methods necessitate at least three core publications to construct a plane capable of enclosing other potentially relevant publications. In contrast, cosine similarity only requires the pre-computed mean embedding of the core publications and a similarity threshold. This independence from recall allows cosine similarity to determine relevance even when recall is minimal, making it a more robust metric in cases of sparse core publications.

5. Conclusion

5.1. Summary and Contributions

This work presents **LitQEval**, a novel framework addressing limitations in evaluating literature search query generation. Existing datasets like CLEF and Collection of Seeds focus on medical data, and while more diverse datasets like Badami's [1] exist, they lack robust evaluation metrics. The primary issues identified are the overemphasis on recall at the expense of precision and the problem of overly broad queries generating excessive irrelevant results.

LitQEval introduces a more comprehensive dataset covering 21 diverse topics. Core publications were collected using bibliometric analyses and SLRs to ensure relevance. The dataset is validated using techniques like clustering embeddings of publication titles and abstracts, confirming semantic similarity within fields while distinguishing between different topics.

New evaluation metrics are proposed to balance recall and precision. *Semantic Precision* evaluates the relevance of retrieved publications compared to core publications through four approaches: (1) cosine similarity, (2) Minimum Volume Enclosing Ellipsoid, (3) Convex Hull and (4) clustering. These metrics aim to determine relevant publications form an excessively broad queries via semantic similarity. A decay factor further adjusts precision to account for query breadth.

To balance recall and precision, the *F-β* score emphasizes recall $\beta = 2$ for evaluating queries. Comparative analyses, including a case study on "Soft Robotics," validate the metrics, revealing insights into the trade-offs between core publication retrieval and query specificity.

5.2. Outlook

This effort forms part of a broader initiative, the SQW, developed by Fraunhofer INT. The SQW aims to automate or semi-automate the creation of literature queries, expediting research on emerging topics and offering researchers a quick starting point. While this study concentrated on building a pipeline for evaluating search queries rather than testing SQW's performance, numerous opportunities remain to enhance the tool. These include using the second step, *Iterative Scientific Fine-Tuning*, to refine results further. Additionally, testing optional inputs such as detailed descriptions, uploads of relevant publications, or adjusting model parameters (e.g., temperature for exploration) holds significant potential for improvement.

Beyond the SQW, the framework for evaluating query quality introduces new opportunities. For instance, it facilitates the creation of scientific chatbots capable of answering complex questions by combining standard queries, research questions, or relevant publications as inputs. Such tools could leverage cosine similarity within a large semantic space to identify and retrieve additional relevant publications effectively.

This study highlights several promising directions for future research and development. One key avenue is the refinement and broader application of the semantic precision metrics introduced here. For instance, future work could explore how to adapt these metrics to dynamic and interdisciplinary research areas where core publications may be less clearly defined.

Lastly, the broader implications of this work in developing AI-driven research tools warrant continued exploration. From advanced literature search engines to domain-specific scientific assistants, the principles established here could inform the next generation of tools designed to augment and streamline academic

research workflows.

A. Appendix

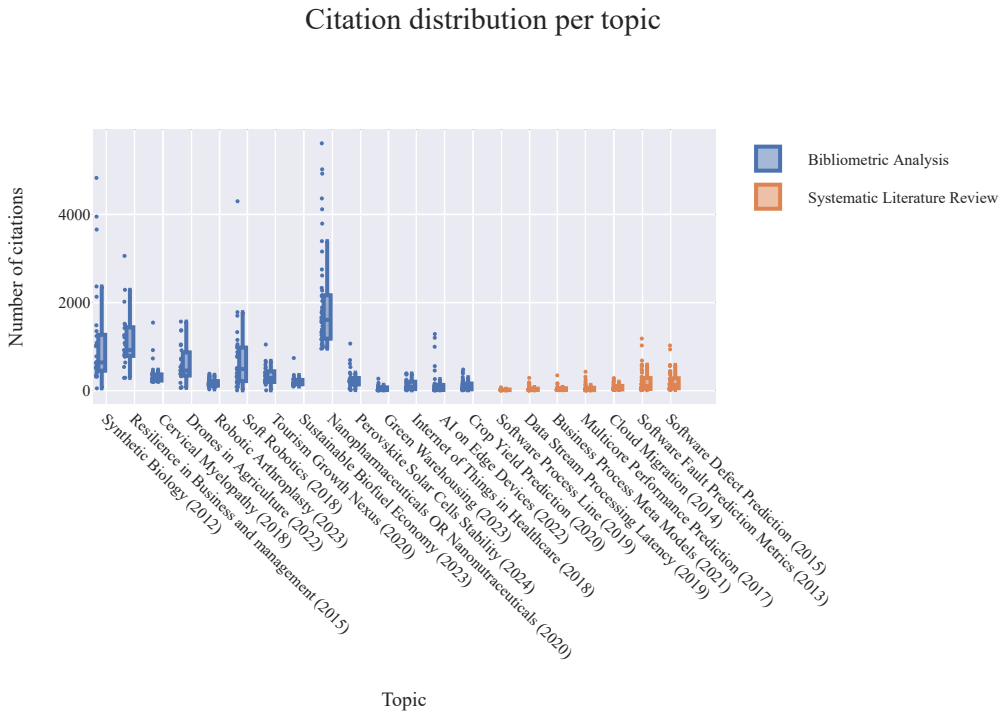

Figure A.1.: The citation ratio per topic, showing the relative citation counts of core publications compared to the average citation frequency within their respective research fields. This illustrates how the prominence of each publication compares to typical citation levels in its field.

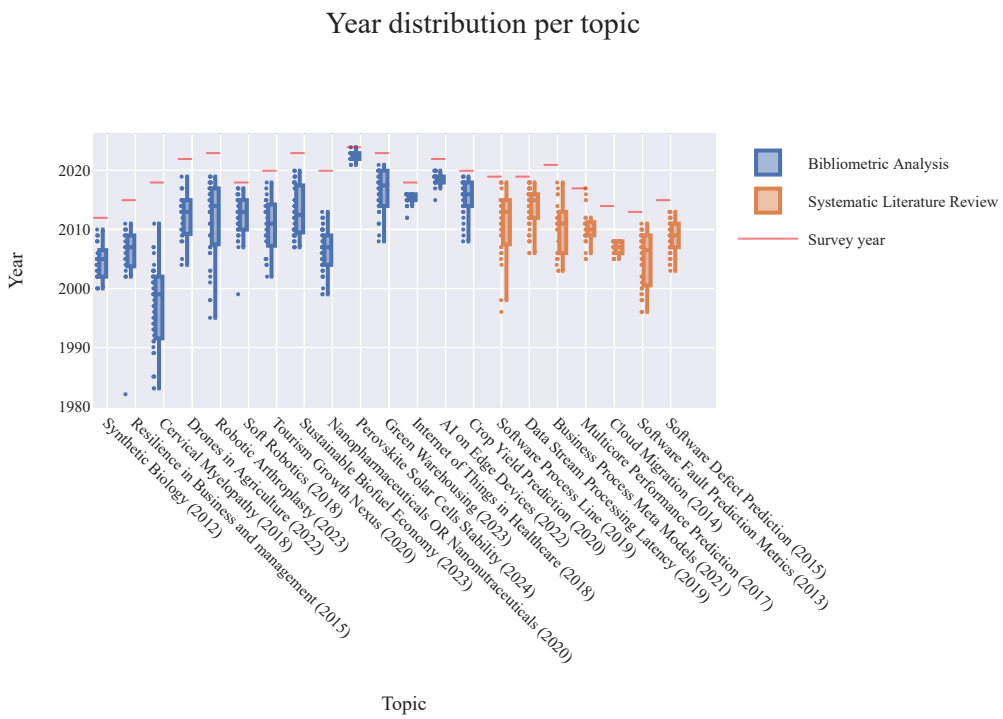

Figure A.2.: The distribution of publication years for core publications across various research topics, highlighting the historical range of studies considered in the bibliometric analyses for each field. Notably, for *Cervical Myelopathy*, the lower bound of publication years was set to 1980 for improved readability, although the actual range goes back to 1953.

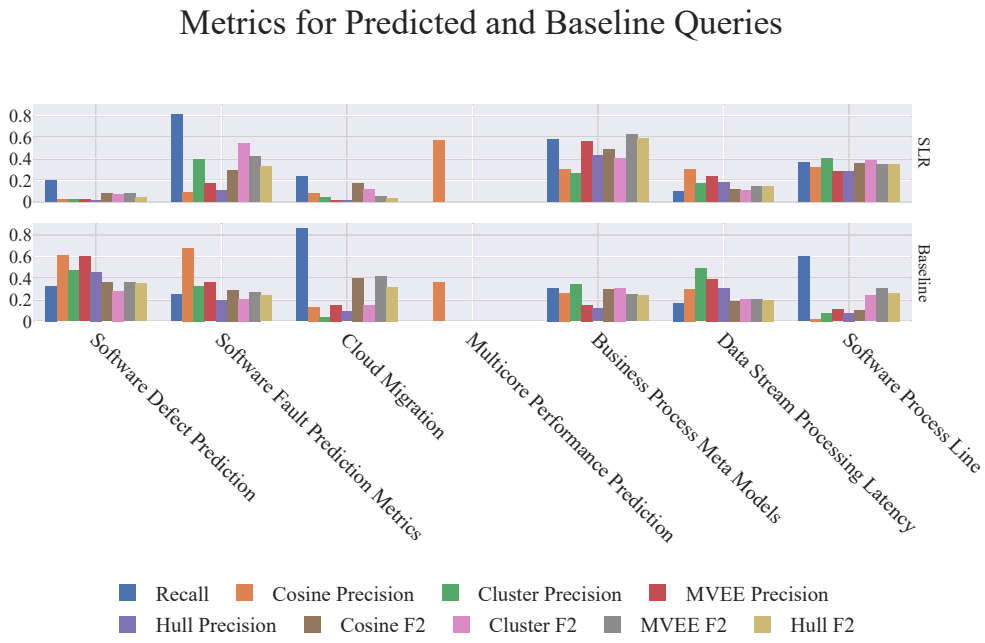

Figure A.3.: This figure shows the results of the second experiment across all the SLR datasets. Surprisingly, we can see that using the SLR query does not achieve outstanding results, which is attributed to its reconstruction and adaptation to fit dimension's search engine. Notably, issues similar to those from the first experiment due to the recall of 0 are also apparent in this case. Notably, the SLR query used for *Multicore Performance Prediction* is only partially available for public access [4], hence the very low scores.

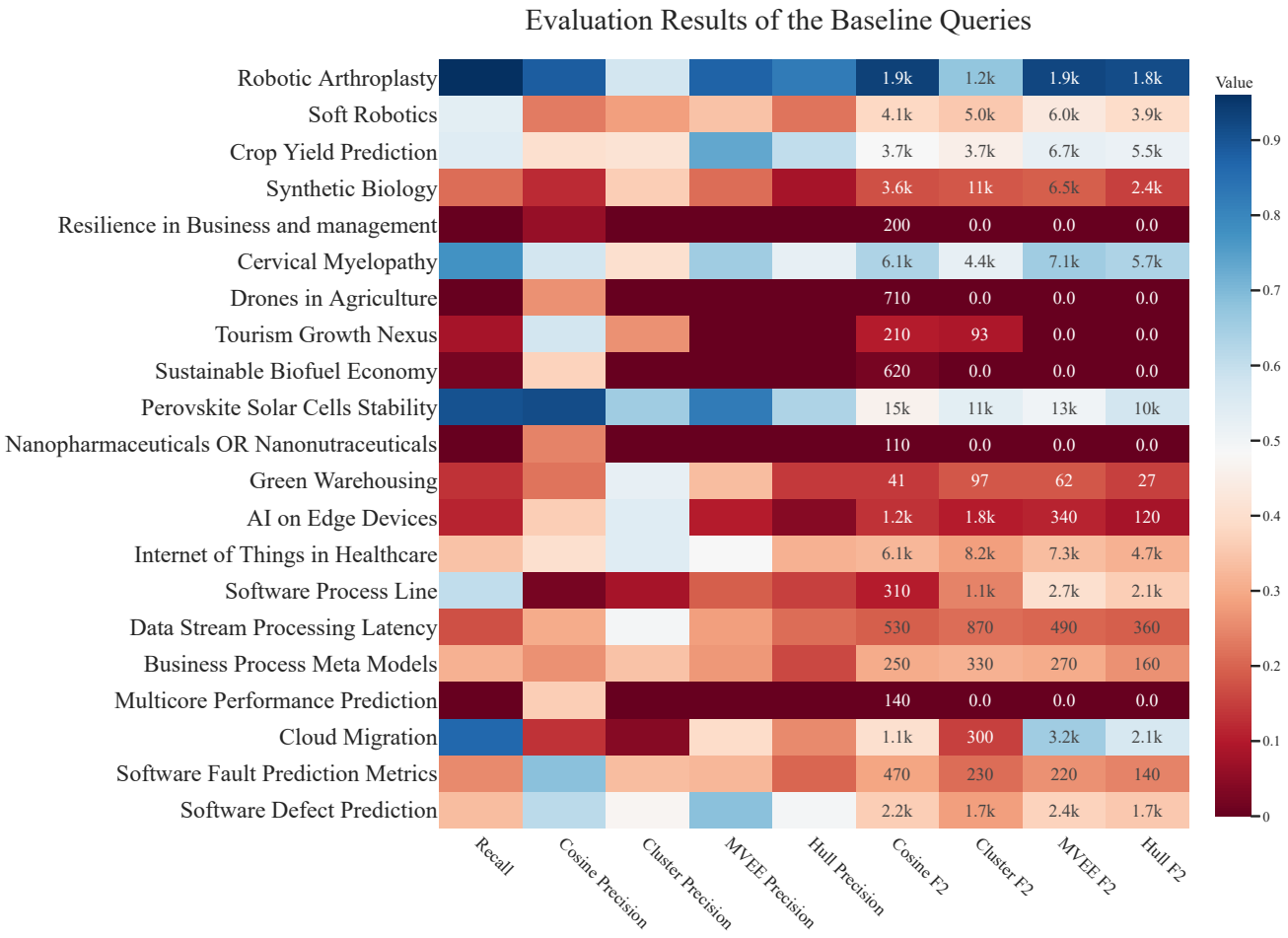

Figure A.4.: The evaluation results of the baseline queries. The color in the figure represents the value of the metric in their respective columns, where the text on F2 metrics indicates the total number of relevant publications used to compute the decay factor. The query for *Robotic Arthroplasty* demonstrates strong performance across precision and recall, and containing only 1.9k relevant publications, thus the high F2 score. In contrast, while *Perovskite Solar Cells Stability* achieves high recall and precision, its F2 score is only decent due to the large number of publications. For the rest of the topics, the F2 score is below average mostly due to the low recall.

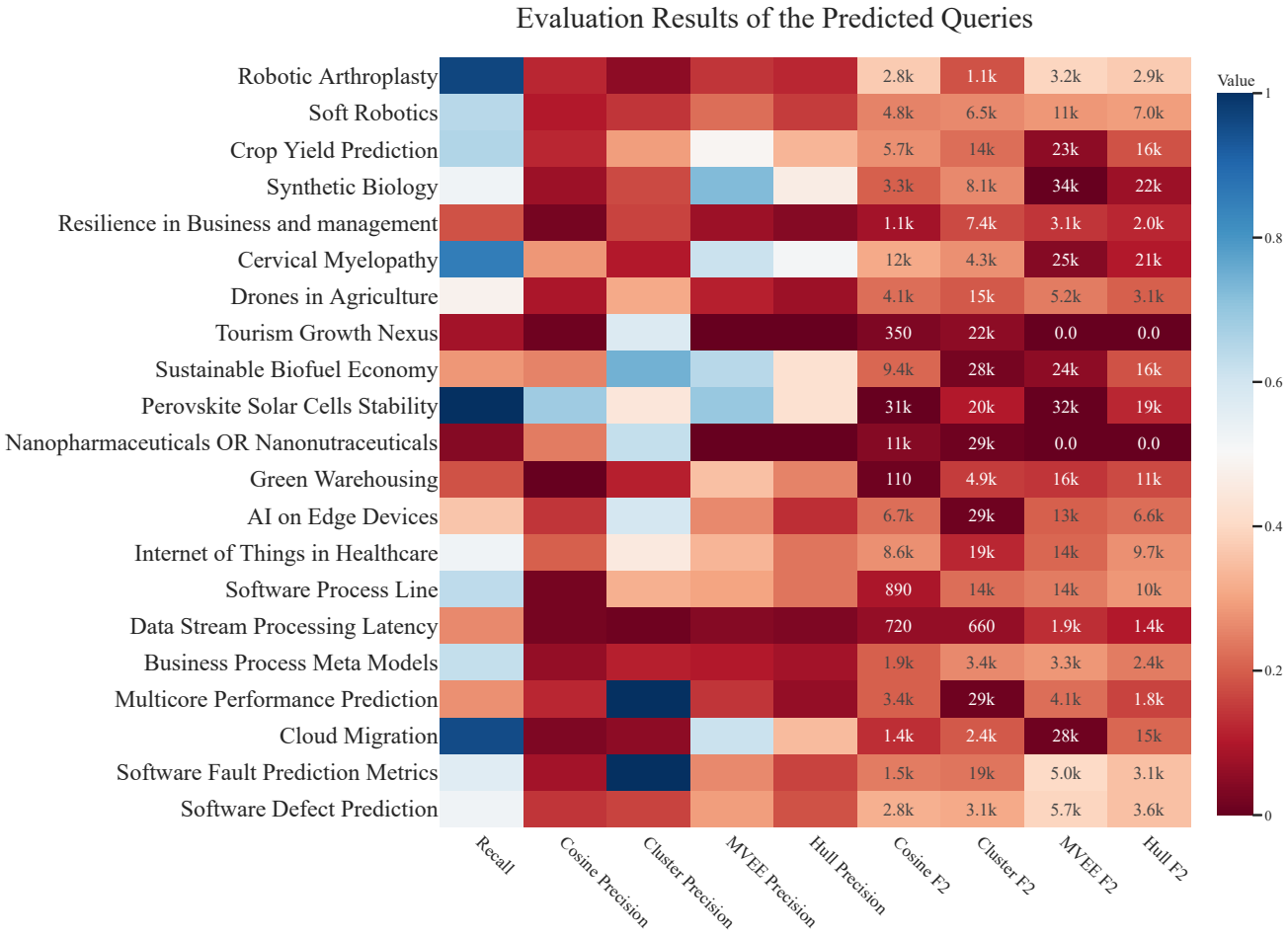

Figure A.5.: The evaluation results of the SLR queries. The color in the figure represents the value of the metric in their respective columns, where the text on F2 metrics indicates the total number of relevant publications used to compute the decay factor. The query for *Robotic Arthroplasty* demonstrates strong performance across precision and recall, and containing only 1.9k relevant publications, thus the high F2 score. In contrast, while *Perovskite Solar Cells Stability* achieves high recall and precision, its F2 score is only decent due to the large number of publications. For the rest of the topics, the F2 score is below average mostly due to the low recall.

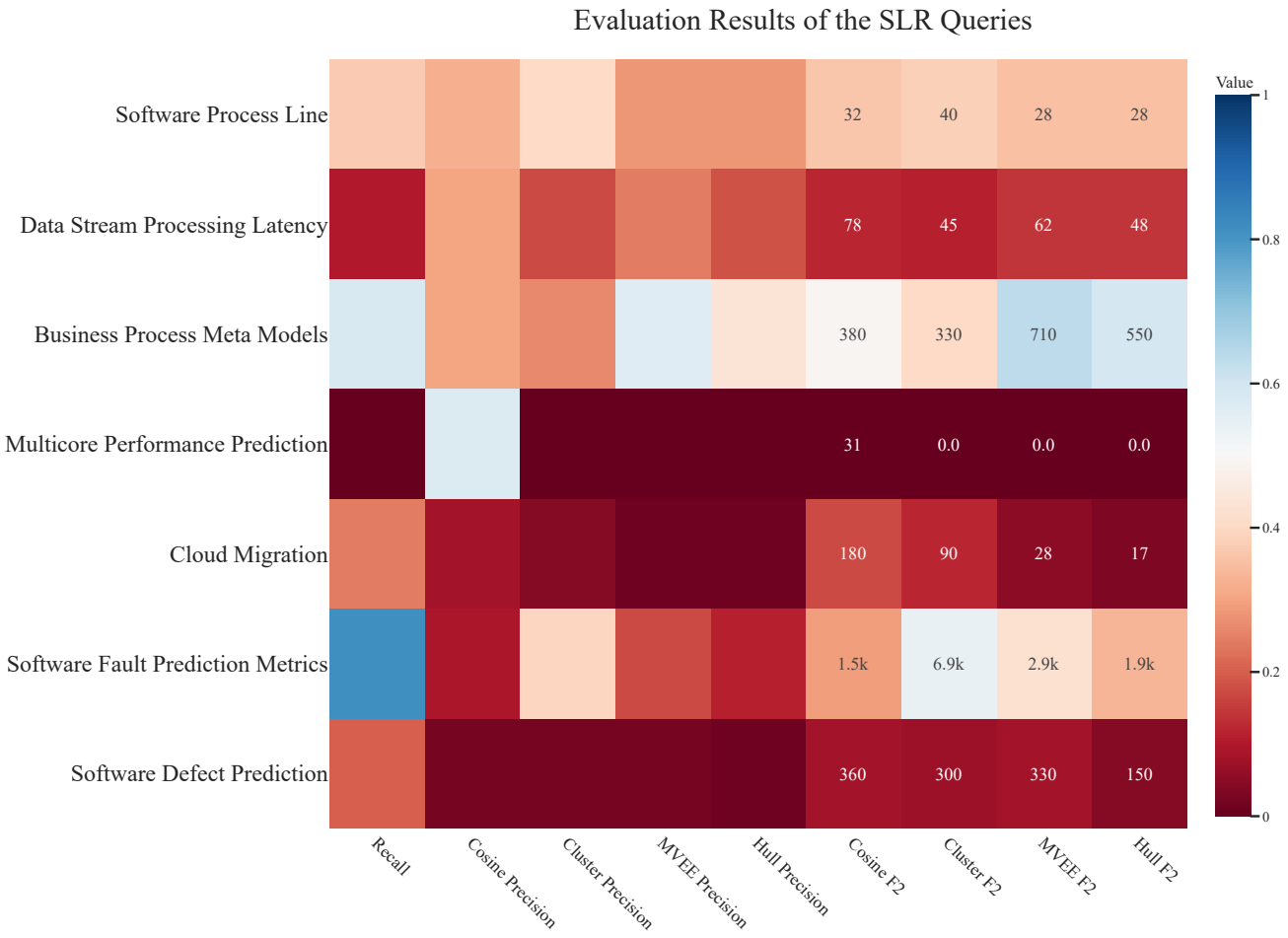

Figure A.6.: The evaluation results of the SLR queries. The color in the figure represents the value of the metric in their respective columns, where the text on F2 metrics indicates the total number of relevant publications used to compute the decay factor. Overall, the sample size of the handcrafted SLR queries is smaller compared to the baseline and predicted queries. This reduced count facilitates manual screening of results; however, it appears that precision is generally average to low across these queries. Notably, the SLR query used for *Multicore Performance Prediction* is only partially available for public access [4], hence the very low scores.

Bibliography

[1] M. Badami, B. Benatallah, and M. Baez. "Adaptive search query generation and refinement in systematic literature review". In: *Information Systems* 117 (2023), p. 102231 (cit. on pp. 7, 8, 25).

[2] Y. Deldjoo. "Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency". In: *ACM Transactions on Recommender Systems* (2024) (cit. on p. 4).

[3] X. Dong, C. Zhang, Y. Ge, Y. Mao, Y. Gao, J. Lin, D. Lou, et al. "C3: Zero-shot text-to-sql with chatgpt". In: *arXiv preprint arXiv:2307.07306* (2023) (cit. on p. 1).

[4] M. Frank, M. Hilbrich, S. Lehrig, and S. Becker. "Parallelization, Modeling, and Performance Prediction in the Multi-/Many Core Area: A Systematic Literature Review". In: *2017 IEEE 7th International Symposium on Cloud and Service Computing (SC2)*. 2017, pp. 48–55. DOI: 10.1109/sc2.2017.15. URL: https://app.dimensions.ai/details/publication/pub.1101547396 (cit. on pp. 30, 33).

[5] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. "CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview". In: Jan. 2017 (cit. on pp. 6, 7).

[6] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. "CLEF 2018 technologically assisted reviews in empirical medicine overview". In: *CEUR workshop proceedings*. Vol. 2125. 2018 (cit. on pp. 6, 7).

[7] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. "CLEF 2019 technology assisted reviews in empirical medicine overview". In: *CEUR workshop proceedings*. Vol. 2380. 2019, p. 250 (cit. on pp. 6, 7).

[8] D. G. Kirkpatrick and R. Seidel. "The ultimate planar convex hull algorithm?" In: *SIAM journal on computing* 15.1 (1986), pp. 287–299 (cit. on p. 13).

[9] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. 2024. arXiv: 2408.06292 [cs.AI]. URL: https://arxiv.org/abs/2408.06292 (cit. on p. 1).

[10] L. McInnes, J. Healy, and J. Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: (2020). arXiv: 1802.03426 [stat.ML]. URL: https://arxiv.org/abs/1802.03426 (cit. on pp. 10, 15).

[11] A. Rejeb, A. Abdollahi, K. Rejeb, and H. Treiblmaier. "Drones in agriculture: A review and bibliometric analysis". In: *Computers and Electronics in Agriculture* 198 (2022). https://doi.org/10.1016/j.compag.2022.107017, p. 107017. DOI: 10.1016/j.compag.2022.107017. URL: https://app.dimensions.ai/details/publication/pub.1147958699 (cit. on p. 8).

[12] G. Rogers, M. Szomszor, and J. Adams. "Sample size in bibliometric analysis". In: *Scientometrics* 125.1 (July 2020), pp. 777–794. ISSN: 1588-2861. DOI: 10.1007/s11192-020-03647-7 (cit. on p. 14).

[13] M. J. Todd and E. A. Yıldırım. "On Khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids". In: *Discrete Applied Mathematics* 155.13 (2007), pp. 1731–1744. ISSN: 0166-218X. DOI: https://doi.org/10.1016/j.dam.2007.02.013. URL: https://www.sciencedirect.com/science/article/pii/S0166218X07000716 (cit. on p. 12).

[14] S. Wang, H. Scells, J. Clark, B. Koopman, and G. Zuccon. "From Little Things Big Things Grow: A Collection with Seed Studies for Medical Systematic Review Literature Search". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. ACM, July 2022, pp. 3176–3186. DOI: 10.1145/3477495.3531748. URL: http://dx.doi.org/10.1145/3477495.3531748 (cit. on p. 7).

[15] S. Wang, H. Scells, B. Koopman, and G. Zuccon. "Can ChatGPT write a good boolean query for systematic review literature search?" In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2023, pp. 1426–1436 (cit. on p. 7).

Declaration

I declare that I have written this work by myself. I have identified as such all passages taken verbatim or in meaning from published or unpublished works by third parties. All sources and aids that I have used for the work are indicated.

(Example formulations follow, which you must adapt to your work for the sake of transparency. Of course, you should have discussed about the acceptability of such aids with your supervisor in advance.) In particular, the following AI systems were also used to create this work:

- ChatGPT in version ... was used for the initial text drafting based on bullet points given by me in the chapters ... / of the entire work.
- ChatGPT was consulted on the following topics: ... / was used to generate ideas regarding ... / for the structuring of ... / for the conception of the system ... .

  The wording of the dialogs and the version used were documented in the appendix of this work. Passages used are marked as such in the text.
- ChatGPT was used to create source code for ... . The wording of the dialogs and the version used were documented in the appendix of this work. The use is indicated in the header of the respective source file / class / method / parts.
- Copilot in version ... was used to create source code / auto-complete for ... . The use is documented in the header of the respective source file / class / method / parts.

I am aware that content generated by AI systems is no substitute for careful scientific work, which is why all such generated content has been critically reviewed and finalized by me.

This work has neither been submitted with the same content nor in essential parts to any other examination authority.

Your City, 2025-01-29

Mohammad Sakinini