



Submitted by: Prompt Pirates

Team Members:

Kushagra Tiwari

Aditya Kumar Giri

Matru Prasad Mohanty

# **Abstract**

In response to the challenges posed by the increasing volume of research papers, we have developed an automated solution to streamline the evaluation and conference selection process for academic submissions. The first component of our solution leverages large language models (LLMs) and web scraping APIs to classify research papers as "Publishable" or "Non-Publishable." By analyzing critical aspects such as methodology, argument coherence, and evidence validation, the system identifies issues that may impact the paper's suitability for publication. In the second part, we apply a hybrid search methodology and a Retrieval-Augmented Generation (RAG) framework to recommend the most suitable academic conferences for papers deemed "Publishable." By utilizing the Pyncone database and comparing the paper's subject matter, methodology, and findings with the profiles of prestigious conferences like CVPR, NeurIPS, DAA, EMNLP, TMLR, and KDD, the system offers conference suggestions along with a detailed rationale. This approach ensures that each recommendation is well-aligned with the paper's content and the conference's focus areas. Our solution significantly enhances the efficiency and objectivity of the research evaluation and conference selection process, providing a scalable framework for future academic workflows.

**Introduction**

The evaluation and selection of research papers for conference submissions are critical steps in the academic process, shaping the trajectory of knowledge dissemination and driving innovation. However, these processes are traditionally labor-intensive, time-consuming, and highly dependent on the expertise and subjective judgment of reviewers. With the exponential growth in research outputs across disciplines, manual evaluation has become increasingly unsustainable, often leading to delays and inconsistencies in decision-making.

To address these challenges, this hackathon presents a unique opportunity to innovate by developing an AI-driven system using the Pathway Framework. The aim is to streamline the evaluation process, focusing on two key tasks: assessing the publishability of research papers and recommending the most suitable conferences for submission. By harnessing cutting-edge technologies such as advanced language models, comparative analysis techniques, and streaming data frameworks, participants are tasked with building a system that not only optimizes the evaluation process but also ensures objectivity, scalability, and alignment with the standards of academic excellence.

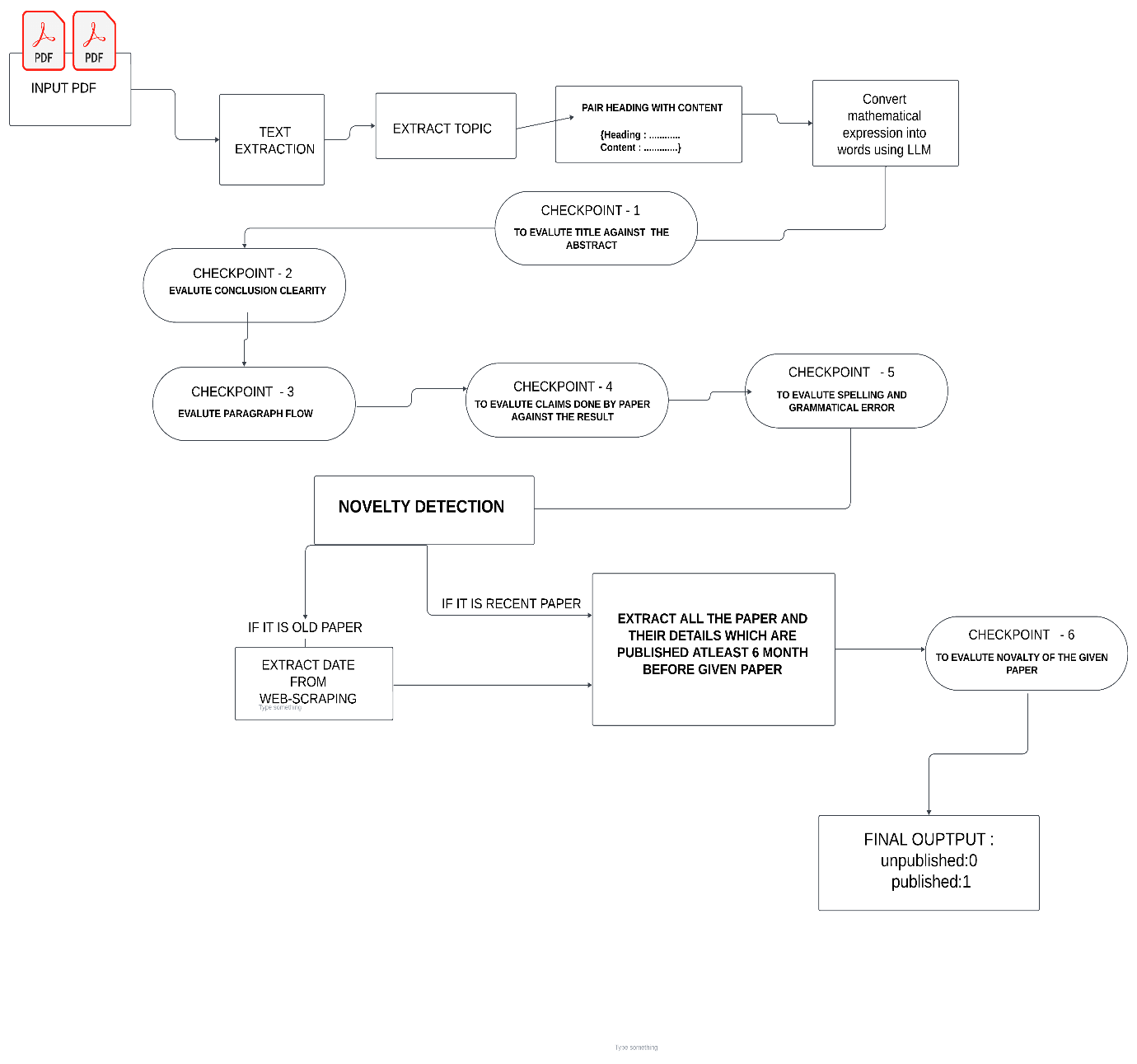
This initiative goes beyond simple automation; it seeks to create a robust framework capable of handling diverse research topics while maintaining consistency in evaluation. Participants will have access to a dataset of high-quality, benchmark research papers from established conferences. Using this dataset, the system will evaluate new submissions, compare them against these benchmarks, and provide formal justifications for classification and conference recommendations. The overarching goal is to demonstrate how artificial intelligence can enhance the efficiency and fairness of academic processes, paving the way for more reliable and impactful research dissemination.

The data available to us consists of research papers that are published in conferences like CVPR, EMNLP, NeurIPS, KDD, TMLR. For reference we were provided with 10 papers which were already categorized on the basis of conference they were published in, and 5 were non publishable research papers. The testing dataset comprised of 135 papers.

During our research we came across several things, to mention a few we came across approached that can potentially work better than RAG, we explored dataset and one of the key findings was that the non publishable papers were completely vague and didn’t made any sense. The proposed solution is successfully able to first classify if a research paper is publishable or not, and then judge the conference in which the paper might have been published. **The approach is not able to generate a reasoning for selection of a particular conference due to unavailability of paid APIs, if provided we could have generated good results. We have provided the complete code to if used with suitable APIs we can generate required outputs.**

**Task -1**

**Methodology: Evaluation of Research Paper Publishability**

****

The proposed methodology provides a systematic approach for evaluating the publishability of research papers. It leverages advanced Natural Language Processing (NLP) techniques, Large Language Models (LLMs), and structured quality assessment checkpoints to ensure a comprehensive and objective evaluation. The primary objective is to identify whether a given research paper meets the necessary criteria for publication.

**1. Input and Initial Setup**

**Input Format**: Research papers in PDF format.

**Objective**: Extract and structure the content of the research paper for subsequent analysis.

**2. Text Extraction and Structuring**

The unstructured content of the PDF is processed to extract textual information such as headings, subheadings, and corresponding content.

**Output Format**: Text is organized in a structured format with paired headings and content:

Heading: <Extracted Heading>

Content: <Corresponding Content>

**3. Topic Identification and Mathematical Expression Conversion**

The system utilizes NLP to extract the main topic of the paper by analyzing its title and abstract.

Mathematical expressions are converted into word-based representations using a Large Language Model (LLM) to facilitate semantic understanding and analysis.

**4. Quality Assessment Framework**

The evaluation process involves several quality-focused checkpoints:

1. **Title-Abstract Consistency**:

Ensures the title accurately reflects the content and scope of the abstract.

1. **Conclusion Clarity**:

Assesses whether the conclusions drawn are logically derived from the discussion and results presented.

1. **Paragraph Coherence and Flow**:

Evaluates the logical structure and seamless flow of ideas across paragraphs.

1. **Claim-Result Validation**:

Analyzes whether the claims made in the paper are sufficiently supported by the results.

1. **Spelling and Grammar Accuracy**:

Identifies and highlights spelling errors and grammatical inconsistencies

**5. Novelty Detection**

The originality of the research is evaluated through a comprehensive novelty detection process:

1. **Paper Categorization**:
   * **Old Papers**: For papers published in the distant past, publication dates are verified using web-scraping techniques.
   * **Recent Papers**: Papers published within recent years are subjected to a detailed novelty analysis.
2. **Novelty Analysis**:
   * Extracts and compares the details of papers published at least six months prior to the given paper.
   * Employs similarity measures (e.g., cosine similarity or embedding comparisons) to assess the distinctiveness of the paper’s contributions.
3. **Novelty Validation**:
   * The methodology ensures that the paper introduces unique findings, methodologies, or theoretical contributions not already explored in existing literature.

**6. Final Evaluation and Classification**

Based on the results of the aforementioned checkpoints and novelty analysis, the paper is classified into one of two categories:

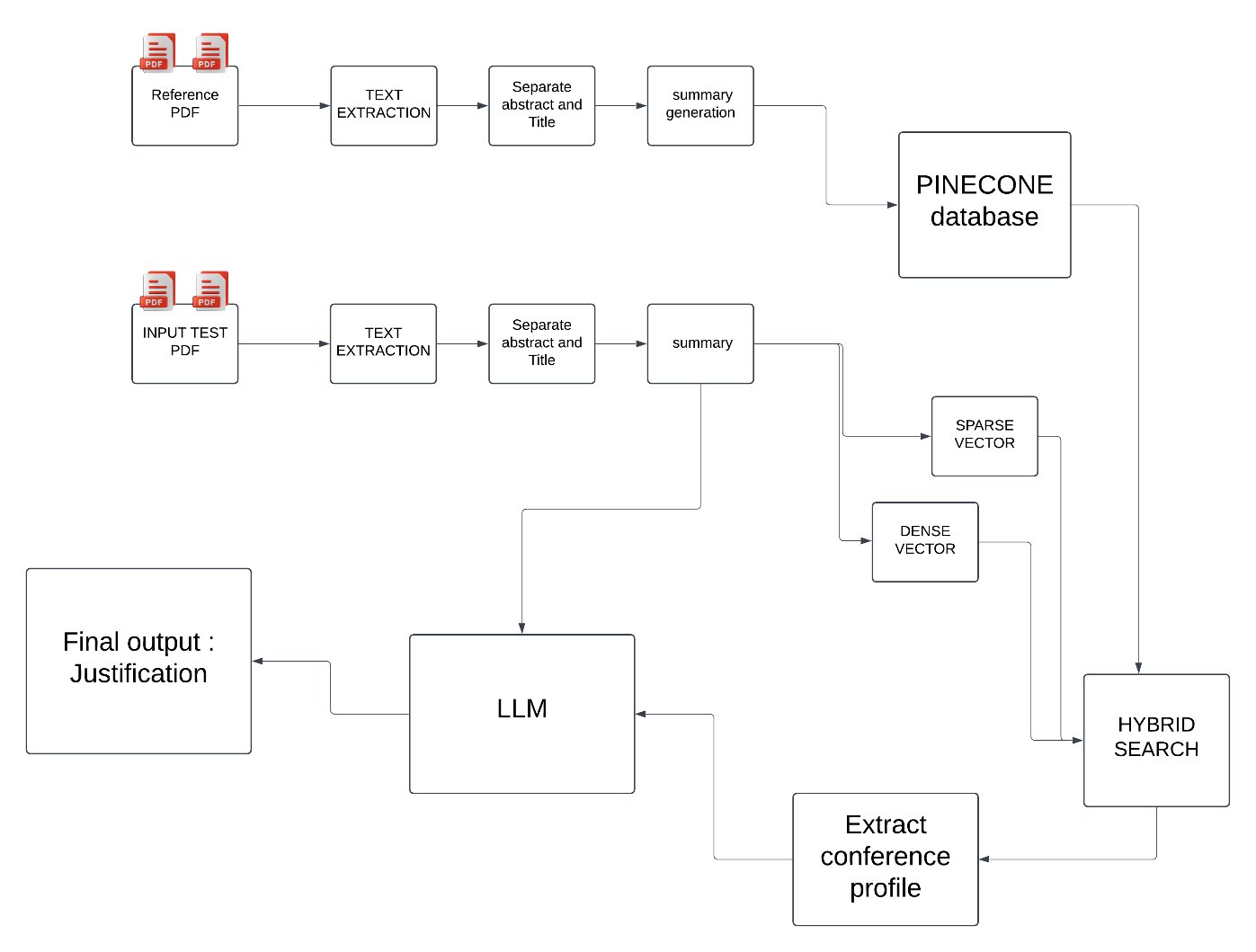
* **Published (1)**: The paper meets the required standards for publication.
* **Unpublished (0)**: The paper does not satisfy the criteria and requires further improvement.

…………………………………………………………………………………………………………………………………………………………………………………………………………………………

**Task -2**

* **Methodology: paper matching with conference and reasoning**

The goal of Task 2 was to analyze research papers classified as "publishable" and recommend the most suitable conferences for their submission. This involved a series of systematic steps, leveraging advanced NLP techniques and external tools to ensure accurate and meaningful results. Below is a detailed explanation of the methodology:



**1. Dataset and Input Processing**

**Input Data:**

* + The input consisted of a CSV file listing research papers and their classification as "publishable" or "non-publishable," along with the associated PDF files of the papers.

**Preprocessing:**

The process began by identifying papers marked as "publishable" (Publishable = 1) and locating their corresponding PDF files from the provided directory.

Non-publishable papers were excluded from further analysis.

**2. Text Extraction and Filtering**

**Content Extraction:**

* + The content of each selected paper was extracted using the PyPDF2 library, while avoiding table-like structures and numeric-heavy lines to retain meaningful text.

**Segmentation:**

* + The extracted text was processed to identify and separate key sections such as the title, abstract, and body content. This was crucial for focusing on the most relevant parts of the paper during further analysis.

**3. Summarization of Research Papers**

* **Summarization Pipeline:**

A custom summarization pipeline was implemented using LangChain’s load\_summarize\_chain and the Gemma2-9b-It model from GROQ.

* **Two-Step Prompt System:**
  + **Initial Prompt:**
    - Created a structured summary focusing on:
      * Machine Learning and Deep Learning techniques
      * Algorithms and architectures
      * Datasets and training details
      * Applications and results
  + **Refinement Prompt:**
    - Improved and expanded the summary to ensure clarity, completeness, and relevance.

**4. Keyword Extraction**

**Process:**

Refined summaries were used to extract key methodologies, techniques, datasets, algorithms, and topics.

An LLM-based chain focused solely on isolating relevant keywords, ensuring the identification of pivotal elements from the research paper.

**5. Conference Profile Creation**

**Conference Analysis:**

Profiles for target conferences (e.g., CVPR, EMNLP, NeurIPS, TMLR, and KDD) were created using GPT-based analysis of 10 additional research papers from each conference.

**Profile Components:**

Each profile included details on focus areas, methodologies of interest, typical datasets used, and key topics of relevance.

**Storage and Indexing:**

The profiles were stored and indexed for efficient retrieval.

**6. Hybrid Search in Pinecone**

* **Mechanism:**
  + A hybrid search mechanism was implemented using Pinecone for vector-based information retrieval.
  + **Sparse Encoding:**
    - BM25-based sparse encodings represented textual relevance.
  + **Dense Encoding:**
    - Sentence embeddings were generated using the all-MiniLM-L6-v2 model for semantic similarity.
* **Querying:**
  + Both sparse and dense encodings of the paper summaries were used to query the indexed conference profiles, identifying the most relevant matches.

**7. Recommendation and Justification**

* **Conference Matching:**
  + The top conference match was determined based on similarity scores from the hybrid search.
* **Justification:**
  + A LangChain-powered LLM prompt was used to generate a concise explanation of why the paper aligned with the selected conference. This justification was based on the alignment of the paper’s summary with the conference’s profile, ensuring a clear rationale for the recommendation.

**8. Final Output Generation**

* **Results:**
  + The results, including the top conference name and the corresponding rationale, were added to the original CSV file under new columns (Conference and Rationale).
* **Output File:**
  + The updated CSV was saved as the final output, providing a comprehensive mapping of publishable papers to recommended conferences.

**Key Challenges in problem statement:**

The project faced multiple challenges spanning data quality, technical integration, and resource constraints. Below is a summary of the key obstacles:

1. **Dataset Quality and Integrity**  
   The dataset provided was problematic, containing incomplete, nonsensical, and truncated research papers. Essential components like images, references, and proper formatting were missing. Additionally, inconsistencies arose as the final submission dataset included rephrased versions of the original papers. These issues rendered direct use of conference papers infeasible and severely hindered model training and evaluation.
2. **Integration of the Pathway Framework**  
   Integrating the Pathway Framework, essential for streaming data processing, proved challenging due to compatibility issues and insufficient documentation. These difficulties forced the team to abandon the framework, adopt alternative approaches, and deviate from the original plan, reducing development efficiency and limiting the exploration of Pathway's full potential.
3. **Time Constraints**  
   The limited project timeline curtailed critical processes such as comprehensive data cleaning, rigorous testing, and advanced model fine-tuning. The time pressure also restricted experimentation with advanced methodologies, impacting the quality and scope of the final output.
4. **Lack of Access to Paid APIs**  
   The absence of paid API access, like GPT-4, hindered the use of advanced language models crucial for tasks like natural language understanding and content classification. The reliance on limited-access alternatives led to suboptimal performance and additional challenges in ensuring consistency across tasks.

**Conclusion :**

The proposed AI-driven solution effectively streamlines the evaluation of research papers and conference recommendations by leveraging NLP techniques, large language models, and a hybrid search mechanism. It ensures a structured assessment of research quality, novelty, and alignment with target conferences, enhancing efficiency and objectivity in academic workflows.

Despite challenges like dataset inconsistencies, technical limitations, and resource constraints, the system demonstrates strong potential for scalability and reliability. Future work can focus on refining datasets, integrating advanced models, and addressing technical hurdles to further enhance its impact on academic processes .