Introduction

Pathway Technology Inc. is the creator of the world’s fastest data processing engine, revolutionizing real-time analytics and AI applications. With a 25-member team boasting expertise from AI powerhouses like Microsoft Research, Google Brain, and ETH Zurich, Pathway blends academic brilliance with industry impact. Pathway’s CTO, Jan Chorowski, has collaborated with AI pioneers Geoffrey Hinton and Yoshua Bengio, while its leadership team also includes the co-founder of Spoj.com and NK.pl, Poland's first social media platform.

Pathway's Python framework, written in Rust for speed, unifies batch and streaming workflows. It powers real-time use cases like AI pipelines, anomaly detection, and Retrieval Augmented Generation (RAG) with unmatched efficiency—up to 90 times faster than competitors. Trusted by NATO and Formula 1 teams, Pathway is driving innovation in live AI systems.

Pathway invites you to join their open-source community on GitHub and leverage their tools to transform challenges into solutions.

- **Pathway**
- **LLM-App**

What does Pathway offer?

Pathway offers various services like Live data frameworks, connectors and AI pipelines ready to be used seamlessly with their applications. For this problem statement, participants have to use Pathway to create their own such application.

⚠ **Note:** Pathway is available on MacOS and Linux. Pathway is currently not supported on Windows. Windows users may want to use Windows Subsystem for Linux (WSL), docker, or a VM.

Problem Statement

The manual evaluation of research papers for conference submission is labor-intensive, time-consuming, and demands significant expertise. This hackathon challenges participants to develop an AI-driven system using the Pathway Framework to streamline the process of conference selection and research paper evaluation. The system will harness advanced language models, comparative analysis techniques, and streaming data frameworks to automate and optimize these tasks.

Participants will have access to a dataset of high-quality, benchmark research papers from conferences or conferences. The objective is to evaluate new submissions, compare them with these benchmark papers, and recommend the most suitable conferences or conference with formal justification.

Pathway Connectors

Participants have to use the **Google Drive connector** provided by Pathway in order to stream data from the source.

Check more about **Pathway data connectors**:

Pathway Data Connectors

Pathway Indexer

The vector store of choice should be either Pathway VectorStore or DocumentStore

Check more about Pathway indexers here:

Pathway Vectorstore

Pathway DocumentStore

Task-1: Research Paper Publishability Assessment

In the academic and research domains, the quality, and publishability of research papers play a critical role in advancing knowledge and fostering innovation. However, the process of determining whether a paper meets the standards for publication can be both time-consuming and subjective, often requiring expert review. With the increasing volume of research outputs, there is a growing need for automated systems to assist in evaluating the quality and suitability of papers for publication. This challenge not only offers an opportunity to innovate but also holds the potential to streamline the publication process and enhance its objectivity.

The task involves developing a framework that can **classify research papers** as either **"Publishable"** or **"Non-Publishable"** based on the evaluation of their content. The goal is to create a robust system capable of identifying critical issues, such as inappropriate methodologies, incoherent arguments, or unsubstantiated claims, that affect the suitability of a paper for publication. For instance, a research paper that applies methodologies or techniques that are not well-suited to the problem being addressed, without adequate justification or adaptation to the context, would be considered unsuitable. Similarly, a paper that presents arguments that are unclear, disorganized, or lack logical coherence, or one that claims results that appear unusually high or unrealistic without sufficient evidence or proper validation, would also fall into the "Non-Publishable" category

A dataset of **150 research papers** is provided for classification, with **15 labeled papers available** for reference to guide the development of the framework. The framework should be designed to accurately classify papers into the appropriate category, ensuring that it can handle a wide range of research topics and maintain consistency across different types of content. The proposed framework must be capable of systematically analyzing these and other aspects of research papers to ensure a reliable and objective evaluation process. The solution should demonstrate high accuracy in detecting such issues, ensuring its applicability across a range of research domains and scalability for future use with larger datasets.

Note: The use of Pathway Framework is not mandatory for this particular task

Task-2: Conference Selection

In this task, the goal is to develop a framework that can analyze a submitted research paper and determine the most suitable conference for its submission. The framework should evaluate the paper's content, and research focus, ensuring that it aligns with the scope, objectives, and standards of various academic conferences. It should compare the paper's characteristics—such as its subject matter, methodology, and findings—against the profiles of different conferences to recommend the most appropriate options for submission. Each recommendation must be accompanied by a formal justification, outlining how the paper's contribution aligns with the conference's focus area. This justification should include a comparative analysis with established works (referred to as "reference papers"), highlighting the originality and significance of the research within the broader academic landscape.

As part of the recommendation process, the framework must be capable of suggesting prestigious conferences such as **CVPR**, **NeurIPS**, **DAA**, **EMNLP**, **TMLR**, and **KDD**, especially when the paper's content aligns with the topics typically presented at these venues. The reference papers provided in the dataset for each conference serve as benchmarks to guide the recommendations of the framework, which is designed to evaluate research papers classified as "publishable" by the first framework. The framework should classify these "publishable" papers exclusively into the aforementioned conferences, ensuring that each paper aligns with the standards of academic excellence and is relevant to the specific subject matter of the selected conference. Each classification must be accompanied by a well-reasoned rationale of up to **100 words**, explaining how the paper's content, methodology, and findings correspond to the themes, focus areas, and quality standards of the chosen conference. This rationale will ensure that the paper is not only a good fit for the conference but also adheres to the academic criteria expected by these prestigious venues.

To enhance efficiency, the framework must integrate **Pathway connectors** and the **Pathway Vectorstore / DocumentStore** for real-time data streaming, enabling instant access and analysis of relevant data. Participants can enhance inference by **adding additional papers** from the given conferences to the database.

Datasets

Dataset Link

The dataset comprises **150 research papers**, with **15** of them provided as reference papers. These papers are labeled as either **"Publishable"** or **"Non-Publishable."** Among the "Publishable" papers, additional classification is provided, indicating which specific conferences the papers are most suited for. Participants are expected to treat the given reference papers as the benchmark for this dataset. It is important to note the original published papers from the conferences may vary in quality or standards. The primary focus of the task is to analyze the characteristics of the conferences themselves, assessing how the papers align with the specific themes, topics, and quality standards typically upheld by those conferences. This analysis will guide the classification of the papers, ensuring they are recommended for the most relevant and prestigious conferences.

Deliverables

Submit the following in the form of a zip file. The zip file must have the following naming convention **<TEAMNAME>_KDSH_ROUND2.zip**

1. Code (Results must be reproducible)
2. Report (Max-10 pages without Appendix)
3. CSV File of the following format:

Example CSV File:

Filename: “results.csv”

Paper ID	Publishable	Conference	Rationale
P001	1	cvpr	Lorem epsum dolor ....
P002	0	na	na
....	....	....	.....
P135	1	tmlr	Lorem epsum dolor ....

Evaluation Criteria

1. **Judging Publishability (25%)**
    - Approach towards identification of publishability
    - Accuracy and F1 Score of determining publishability
2. **Conference Selection and Rationale (60%)**
    - Solution Architecture
    - Parsing quality of research papers
    - Efficiency of retrieval
    - Effective API calls and tooling
    - Latency and resource consumption for the system
    - Reasoning behind conference selection
3. **Report (15%)**
    - Quality of comprehensive presentation
    - Completeness and results presentation

UI component (Non-Evaluative for Round 2): While not essential and not prioritized over core functionality during evaluation, incorporating visual elements is considered a valuable enhancement for the presentation round.

Additional Resources

- **What is RAG : Beginner Blog by Pathway**
- **Pathway App Templates**
- **Pathway Developer Documentation**
- **Langchain Quickstart**
- **FastAPI Tutorials**