## TODO 0603

Infrastructure Setup

* Establish and configure the database server environment.

Categorization

* Data Acquisition for Model Training:
  + Utilize ACL 2025 paper track information to train a categorization model for prior papers.
  + Retrieve paper abstracts from arXiv as needed to enhance model performance and data richness.
* Model Validation:
  + Parse and test the trained categorization model using ACL 2024 paper data.

Paper Information Table

* Data from ACL-anthology
* Data Schema:
  + papers table: paper\_id, title, abstract, venue, year, award

Researcher Information Table

* Researcher Identification: OpenReview.net IDs as a primary identifier
  + Hybrid parsing method, data from both Google Scholar and OpenReview.net
* Data Schema:
  + researchers table: researcher\_id, first\_name, last\_name, affiliation\_history, citation\_metrics (h-index, total citations from Google Scholar)

Other Relational Tables:

* Researcher\_id - paper\_id
* Paper\_id - tracks

By 0607

* Researcher and paper tables ready as csv
* Track categorization model tested