**Quantum Paper Analysis Summary (2005–2025) done on 29th May 2025**

**Step 1: Data Preparation & Cleaning**

* Collected ~3000 scientific papers on quantum topics.
* Extracted and cleaned the **Abstract + Introduction** sections using a language model (FLAN-T5).
* Removed citation noise, brackets, and standalone digits to clean up text for embedding.

**Step 2: Embedding and Clustering**

* Generated **BERT-based embeddings** using all-MiniLM-L6-v2 for each document.
* Stored all 2981 embeddings for reuse (.npz file).
* Applied **UMAP** for dimensionality reduction and **HDBSCAN** for unsupervised clustering.
* Saved results with filename + cluster ID for downstream use.

**Step 3: Exploratory Analysis**

* Plotted UMAP clusters: revealed topic groupings and dense research clusters.
* Created CSV and heatmap showing each paper’s cluster assignment.

**Step 4: Temporal Trends & Topic Discovery**

* Extracted publication years from filenames.
* Counted papers per cluster per year (1992–2025).
* Visualiseds growth and decline of research clusters:
  + Line plot: shows which clusters are trending
  + Heatmap: intensity of research activity per cluster/year
* Identified **top keywords per cluster** using TF-IDF (cluster\_topic\_summary.csv)