TopicBench

Benchmarking LLMs for Topic Labeling

BE BOUNDLESS

Background

- **The Problem:** Researchers use topic modeling to find themes in text
- **The Challenge**: These themes need human-readable labels. Which AI model is best?
- **Traditional Approach**: Manually label topics (slow, subjective)
- **Our Approach with TopicBench**: Automatically compare different LLMs and pick the best one

Data Used

- 5 research papers across 5 different fields
  - Sociology
  - Medicine
  - Environmental Science
  - Human–computer interaction (HCI)
  - Natural language processing (NLP)
- **Data type**: Keywords from topic modeling + human interpretations

Example of Data

Paper Title	Field	Keywords	Author's Label	Human Labeling
almquist_bagozzi_2019	sociology	[['one', 'made', 'anoth'…]]	['Inspirational Language', 'Group Identity Debates', 'Neocolonialism'…]	['Motivational Rhetoric', 'Collective Identity Discussions', 'Neo-imperialism'…]
dinsa_2024	medicine	[['body', 'came', 'dries up'…]]	['Nervous disease', 'Gynecology', 'Mental illness'…]	['Fatigue & General Body Weakness', 'Pregnancy & Menstrual Health', 'Head, Ear & Respiratory Issues'…]

etc etc…

Use Cases

- **User:** John, a computational social scientist analyzing large text datasets.

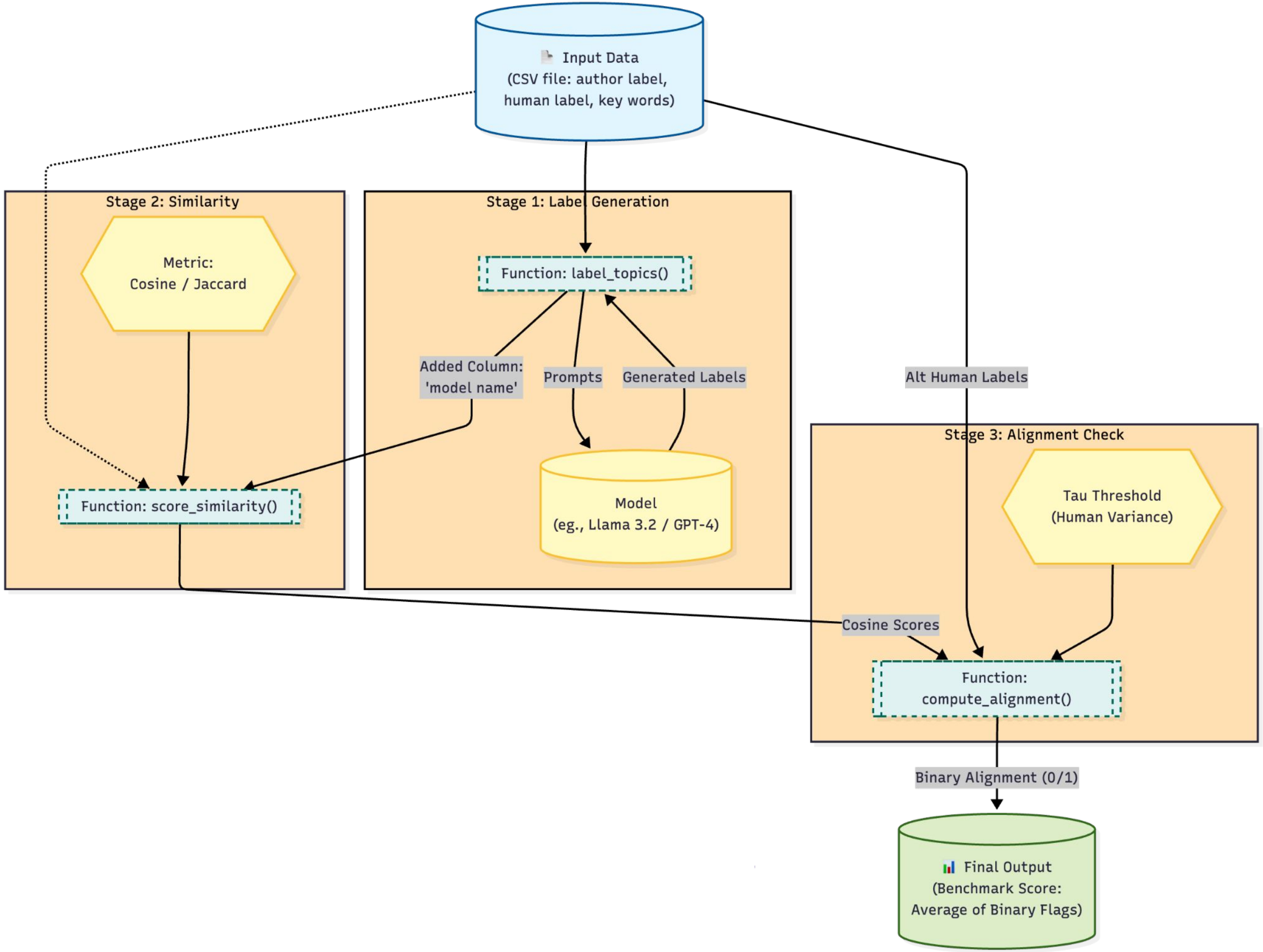
- **Goal:**
  Identify topic patterns and select the LLM that produces the most accurate topic labels.

- **Constraints:**
  – Works with sensitive data; cannot use cloud APIs.
  – Needs reproducible, local evaluation.
  – Cannot write API integration code for multiple models.

- **<u>How TopicBench Helps:</u>**
  1. Label Generation:
     – Runs topic labeling locally using different LLMs.
     – Records model name for comparison.
  2. Similarity Scoring:
     – Computes cosine / Jaccard similarity for author, human, and LLM labels.
  3. Alignment Check:
     – Applies threshold to classify model alignment (0/1).

- **Outcome:**
  TopicBench allows John to choose the most accurate LLM for research while keeping data secure.

Design


Input Data
(CSV file: author label,
human label, key words)
Stage 2: Similarity
Metric:
Cosine / Jaccard
Function: score_similarity()
Stage 1: Label Generation
Function: label_topics()
Added Column:
'model name'
Prompts
Generated Labels
Model
(eg., Llama 3.2 / GPT-4)
Alt Human Labels
Stage 3: Alignment Check
Tau Threshold
(Human Variance)
Cosine Scores
Function:
compute_alignment()
Binary Alignment (0/1)
Final Output
(Benchmark Score:
Average of Binary Flags)

Demo

Import:

```python
import pandas as pd
import numpy as np
import os
import ast
import json
from tqdm import tqdm
from src.topicbench.label_topics import label_topics
from src.topicbench.validate import score_similarity, compute_alignment
```

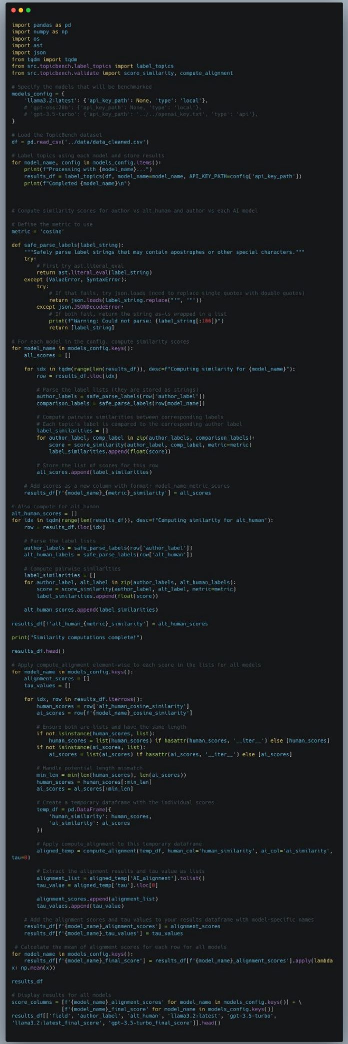
TopicBench

Label generation + similarity score

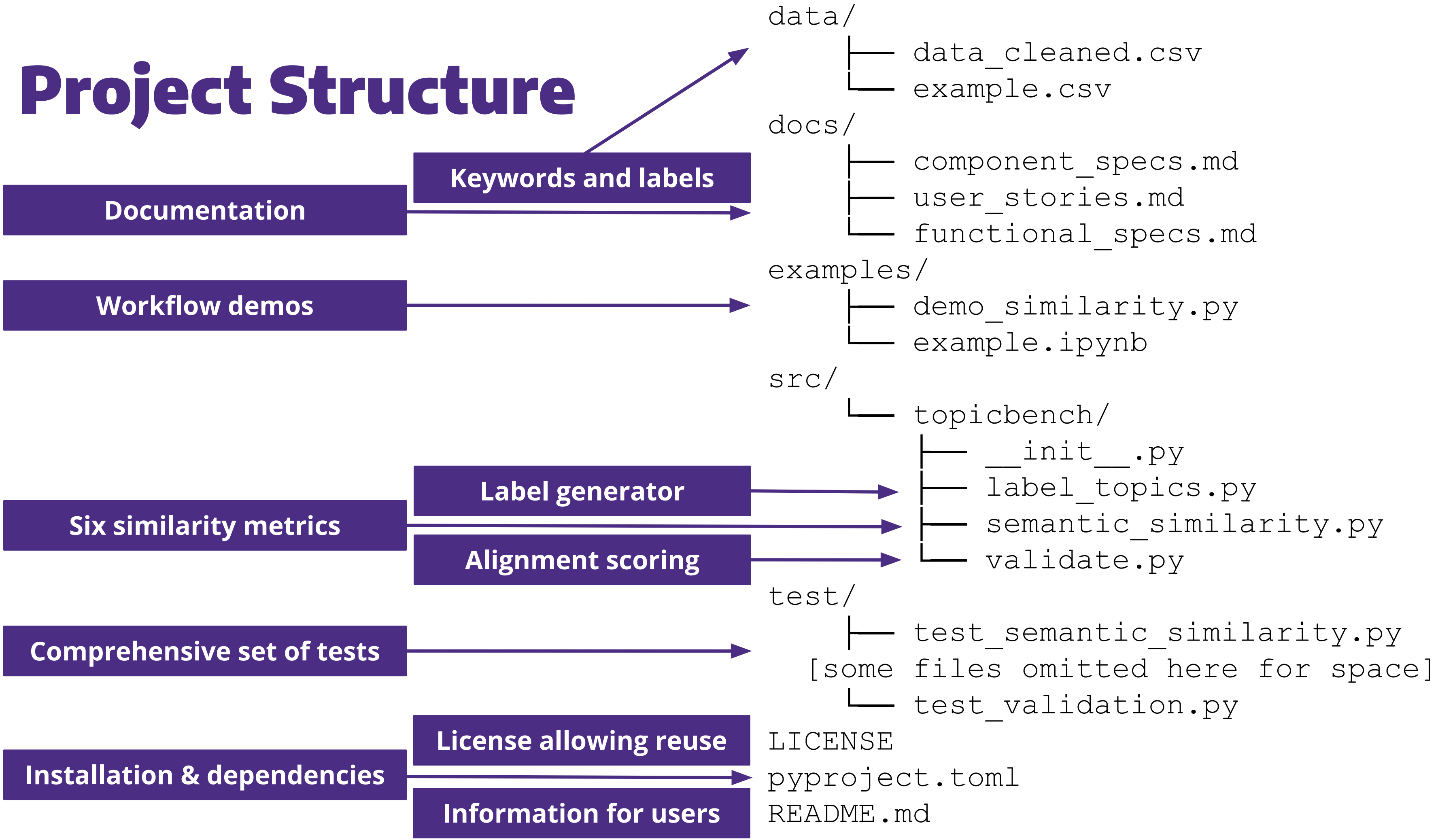
field	keywords	author_label	alt_human	llama3.2:latest	llama3.2:latest_cosine_similarity	alt_human_cosine_similarity
sociology	[['one', 'made', 'anoth', 'everi', 'side', 'ti...	['Inspirational Language', 'Group Identity Deb...	['Motivational Rhetoric', 'Collective Identity...	['Social Movement', 'Activism', 'Protest', 'Re...	[0.22654280066490173, 0.4005390405654907, 0.34...	[0.4601529836654663, 0.857661247253418, 0.7370...
medicine	[['body', 'came', 'dries up', 'rocking', 'time...	['Nervous disease', 'Gynecology', 'Mental illn...	['Fatigue & General Body Weakness', 'Pregnancy...	['Gastrointestinal Issues', 'Menstrual Cycle',...	[0.15294265747070312, 0.41162168979644775, 0.3...	[0.2469634711742401, 0.31719857454299927, 0.11...

Benchmark score for different model:

	field	author_label	alt_human	llama3.2:latest	gpt-3.5-turbo	llama3.2:latest_final_score	gpt-3.5-turbo_final_score
0	sociology	['Inspirational Language', 'Group Identity Deb...	['Motivational Rhetoric', 'Collective Identity...	['Social Movement', 'Activism', 'Resistance', ...	['movement and activism', 'environmental conse...	0.600000	0.333333
1	medicine	['Nervous disease', 'Gynecology', 'Mental illn...	['Fatigue & General Body Weakness', 'Pregnancy...	['Gastrointestinal issues', 'Pregnancy complic...	['Physical Symptoms', 'Pregnancy-related Issue...	0.222222	0.200000



Project Structure



Lessons Learned & Future Work

- **Lessons Learned**
  - Clear separation between labeling, similarity scoring, and alignment improves design.
  - Modular code and documentation made it easier for the whole team to collaborate
- **Future Work**
  - **User Interface:** Add a simple web or notebook-based UI so non-technical researchers can run benchmarks easily.
  - **Custom Datasets:** Allow users to upload their own datasets + automated preprocessing pipeline
  - **Model Integrations:** Support additional LLM providers and custom local models.