Forum User Recommodation System: LLM-VAC

Jing Guo [1] Yi Wang [1] Ziyue Yang [1]

[1]All authors have same contribution

上海交通大學
Shanghai Jiao Tong University

Motivation

- **Hybrid User Recommendation:** Develop a recommendation system integrating large language model representational power with the interpretability of traditional data mining methods.
- **Sociological Inquiry:** Evaluate the feasibility of deriving psychometric traits from user posts and confirming shared interests among users with similar inferred personalities.
- **LLM Long-Text Processing Capability:** Investigate modern LLMs' reliability in end-to-end processing of lengthy, heterogeneous forum posts for latent and complex information extraction.

Supporting Work

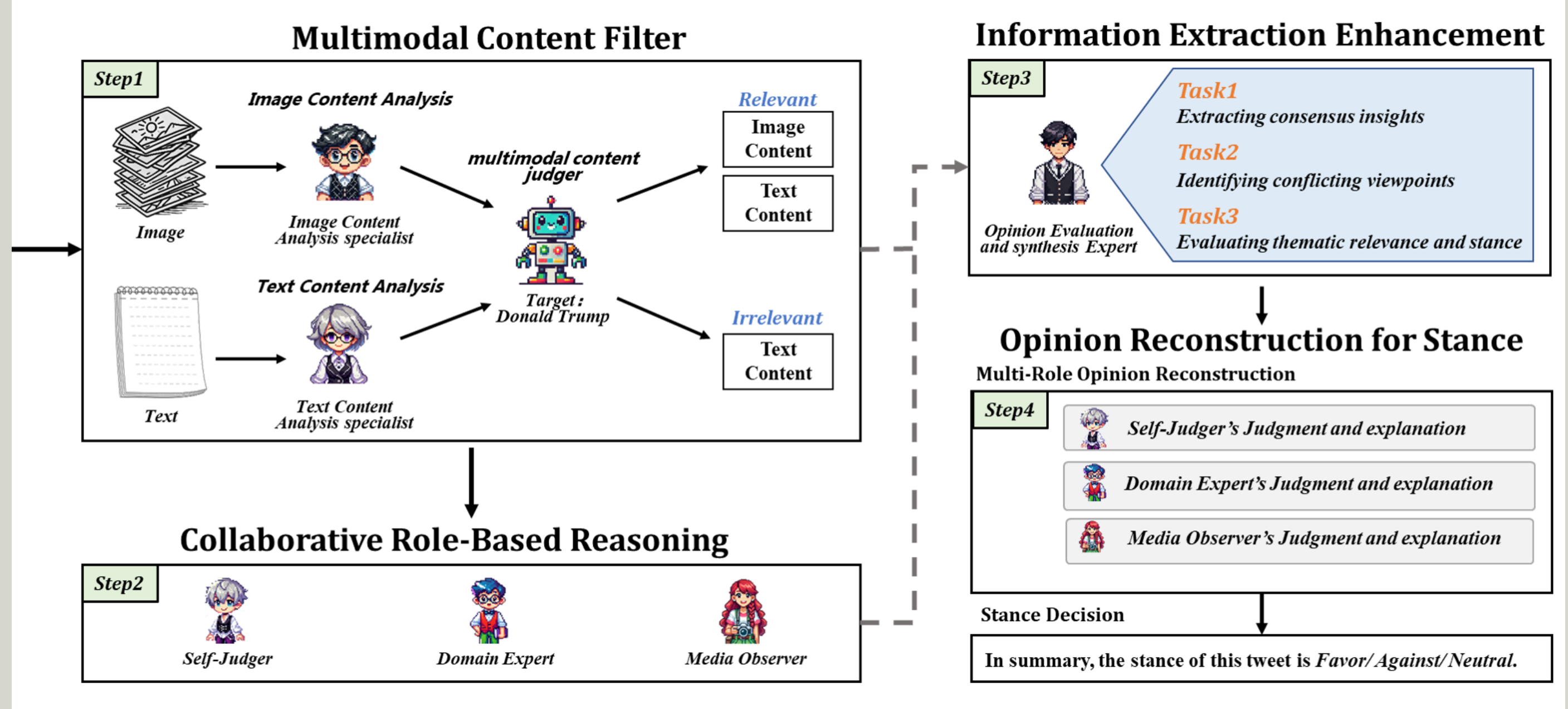
Extract Out User's Latent Traits With Post? With survey on previous work, those potential traits of users can be excavated with their post:

- **Big Five personality traits** (Golnoosh Farnadi 2016)
- **Extroverted or introverted personality** (Beukeboom, 2013)
- **Mental Health**(Glen Coppersmith, 2014)

However, these tasks are performed by social experts. LLM?

Can LLM Perform Complex Works Based On Long Text?

We designed a **multi-agents** pipeline that leverages a large language model (LLM) to perform zero-shot stance detection on the Multi-Modal Stance Detection dataset to validate LLM's ability. Different LLM agents perform as domain experts and process the long text stepwise:



Multimodal Content Filter

Step1

Image Content Analysis

Image

Image Content Analysis specialist

multimodal content judger

Target: Donald Trump

Text Content Analysis

Text

Text Content Analysis specialist

Relevant

Image Content

Text Content

Irrelevant

Text Content

Collaborative Role-Based Reasoning

Step2

Self-Judger

Domain Expert

Media Observer

Information Extraction Enhancement

Step3

Opinion Evaluation and synthesis Expert

Task1
Extracting consensus insights

Task2
Identifying conflicting viewpoints

Task3
Evaluating thematic relevance and stance

Opinion Reconstruction for Stance

Multi-Role Opinion Reconstruction

Step4

Self-Judger's Judgment and explanation

Domain Expert's Judgment and explanation

Media Observer's Judgment and explanation

Stance Decision

In summary, the stance of this tweet is *Favor/ Against/ Neutral.*

Figure 1. Supporting Work Pipieline.

Whereas the original dataset authors trained specialized classifiers(TMPT), our approach infers user stances directly from raw post content and outperforms their fine-tuned models.

method	MTSE		MRUC		MTWQ	
	DT	JB	RUS	UKR	MOC	TOC
TMPT	55.41	61.61	43.56	59.24	55.68	46.82
Ours	**72.05**	**72.84**	**51.30**	**60.84**	**70.18**	**64.56**

This result highlights **the LLM's ability to extract latent features from long text and solve complex NLP tasks without task-specific supervision.**

Dataset Overview

We use the "Reddit Comment and Thread Data," a collection of ≈260 000 Reddit posts and comments scraped via omega-red including these core fields:

- **text, id, subreddit, meta, time, author, ups, downs, authorlinkkarma, authorcommentkarma, authorisgold**

All text is lowercased and tokenized with TreebankTokenizer. Original, punctuation-preserved versions are available via the provided Mega.nz links.

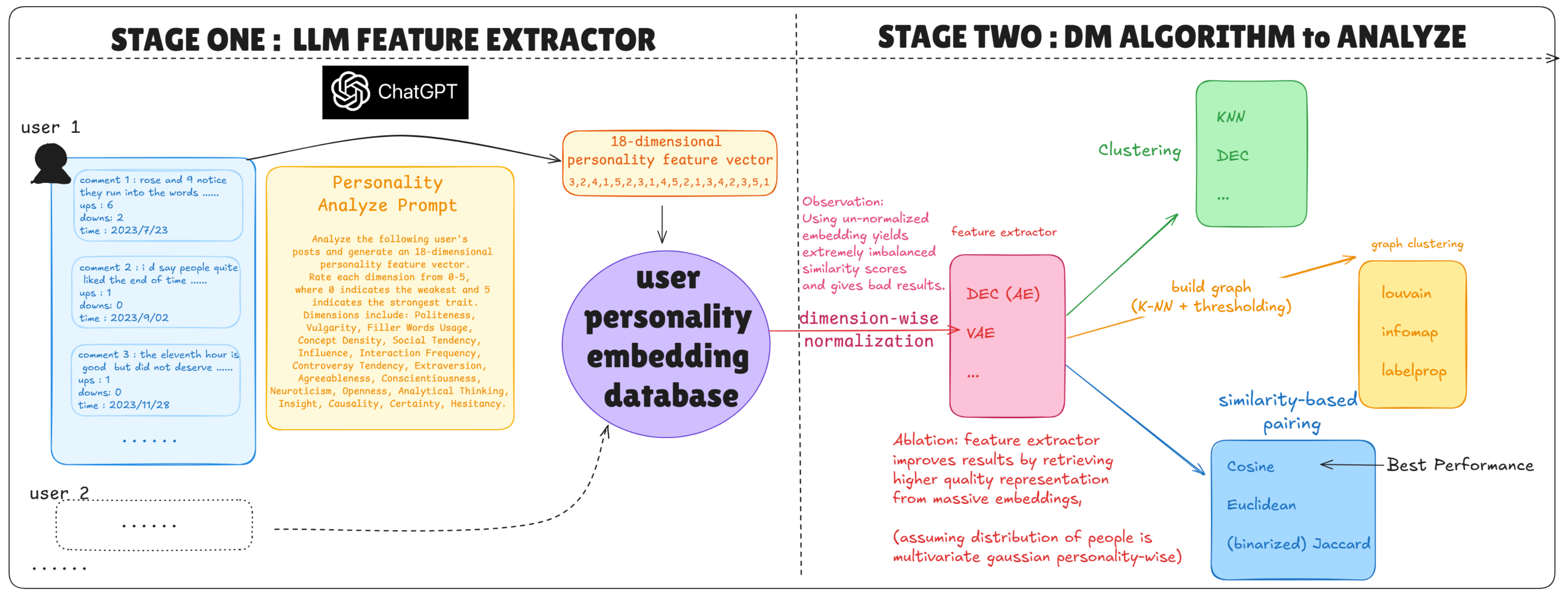
Method: LLM Extract Embedding and Recommendation

Embedding Dimensions: Building upon previous research, we have developed a comprehensive set of features derived from user posts to deeply characterize individual user personalities. These features are categorized as follows:

- **Language Style:** Politeness, Profanity/Vulgarity, Fillers/Disfluencies, and Conceptual Density.
- **Social Behavior:** Sociability, Influence, Interaction Frequency, and Dispute-Making Tendency.
- **Personality Traits:** Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.
- **Cognitive Features:** Analytical Thinking, Insight, Causation, Certainty, and Tentativeness.

Data Cleaning: We aggregated all user posts and remove empty, duplicate, or corrupted entries, exclude outlier users, and the discard all extraneous metadata.

Large Language Model (LLM) Extraction: The extraction task is conducted using **Reddit comment and thread data**. For each post, we input the poster's **text, ID, timestamp, author, upvotes, and downvotes** to an LLM (GPT-4o-mini) for judgment across the aforementioned dimensions. Subsequently, traditional data-mining methods are employed to facilitate user recommendation.



STAGE ONE : LLM FEATURE EXTRACTOR

ChatGPT

user 1

comment 1 : rose and 9 notice they run into the words ......
ups : 6
downs: 2
time : 2023/7/23

comment 2 : i d say people quite liked the end of time ......
ups : 1
downs: 0
time : 2023/9/02

comment 3 : the eleventh hour is good but did not deserve ......
ups : 1
downs: 0
time : 2023/11/28

......

user 2

......

......

Personality Analyze Prompt

Analyze the following user's posts and generate an 18-dimensional personality feature vector. Rate each dimension from 0-5, where 0 indicates the weakest and 5 indicates the strongest trait. Dimensions include: Politeness, Vulgarity, Filler Words Usage, Concept Density, Social Tendency, Influence, Interaction Frequency, Controversy Tendency, Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness, Analytical Thinking, Insight, Causality, Certainty, Hesitancy.

18-dimensional personality feature vector
3,2,4,1,5,2,3,1,4,5,2,1,3,4,2,3,5,1

user personality embedding database

STAGE TWO : DM ALGORITHM to ANALYZE

Observation:
Using un-normalized embedding yields extremely imbalanced similarity scores and gives bad results.

dimension-wise normalization

feature extractor

DEC (AE)
VAE
...

Clustering

KNN
DEC
...

build graph
(K-NN + thresholding)

graph clustering

louvain
infomap
labelprop

similarity-based pairing

Cosine
Euclidean
(binarized) Jaccard

Best Performance

Ablation: feature extractor improves results by retrieving higher quality representation from massive embeddings,

(assuming distribution of people is multivariate gaussian personality-wise)

Figure 2. LLM-VAC Pipieline.

Recommendation: Similarity, Graph and more

We developed a user recommendation system that connects users based on their similarity.

Our approach leverages user embeddings provided by LLM, which is optionally further enhanced by Variational Autoencoders (VAEs) and Deep Embedding for Clustering (DEC). We explore various similarity metrics, including cosine, Euclidean, and (binarized) Jaccard distances.

Furthermore, we employ K-Nearest Neighbors (KNN) graph construction combined with community detection algorithms (Louvain, Infomap) for robust user clustering.

Results

Since the data we use only contain part of user forum data, and we utilize the maximum likelyhood idea to judge the accuracy, so **the actual accuracy will be much higher than the accuracy we measured**.

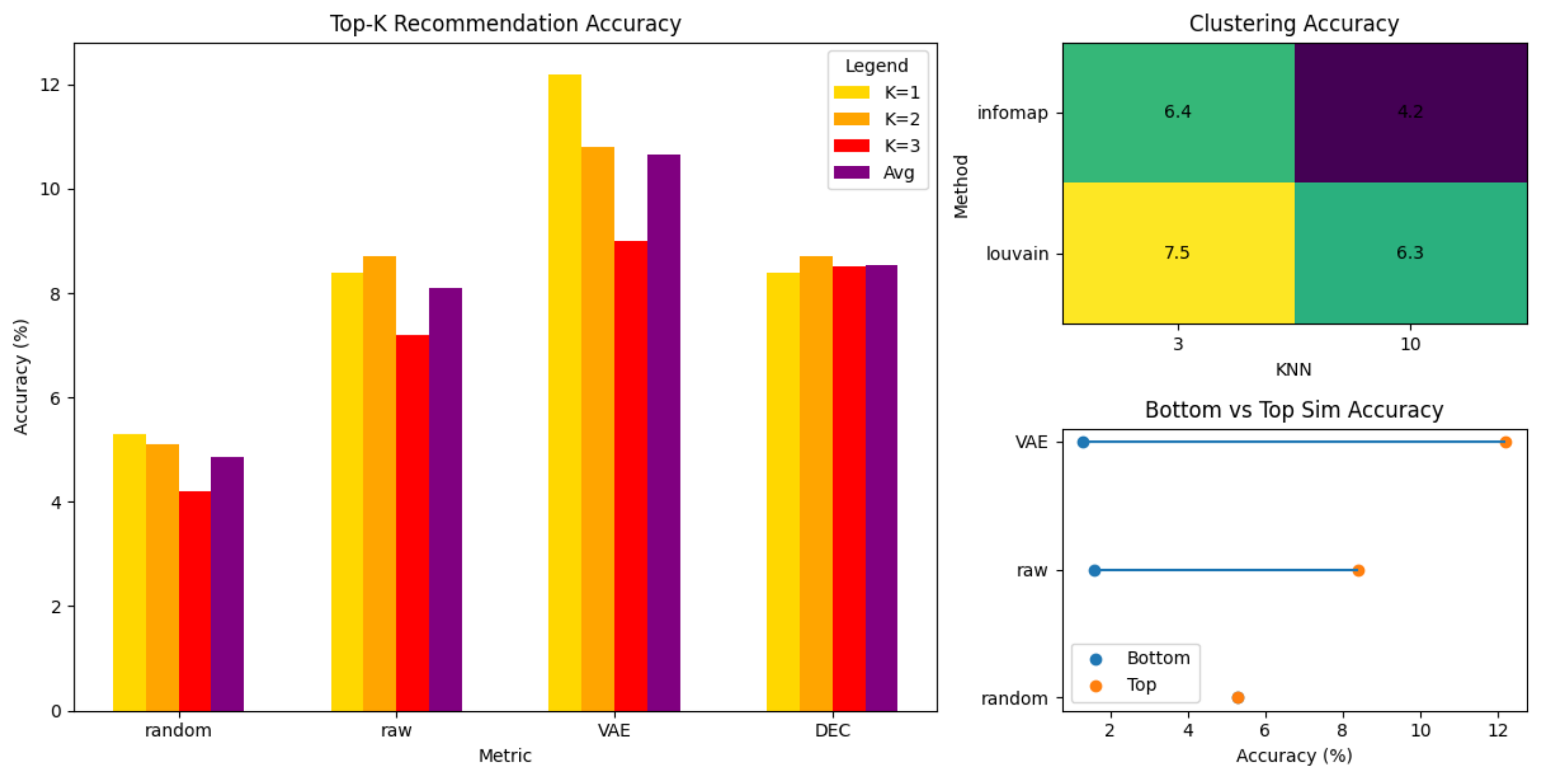


Figure 3. Experiment Results

Analyses & Further Directions

Our Contributions and findings:

- Our user recommendation system effectively recommends users with potential similar interests
- LLM is capable of performing sophisticated task on long-term text
- Users closer in traits embedding are more likely to have same interests

Future Work:

- **Multi-Agents-LLM:** Leverage multi-LLM collaboration—assigning domain-specific roles (e.g., sociologist, internet expert)—to compute user personality embeddings and further improve accuracy.
- **Graph issues:** KNN graphs are sparse and noisy; next steps include mutual/adaptive KNN and weighted/heterogeneous edges.
- **Clustering limits:** Louvain and Infomap misalign with real communities; we will explore overlapping and attribute-aware algorithms.
- **Similarity metrics:** Combine and learn distance measures (e.g., via Siamese networks) to refine user grouping.
- **Embedding quality:** Apply Adversarial AEs and Conditional VAEs to shape latent distributions and incorporate forum labels.

https://www.sjtu.edu.cn AI 3602 Data Mining Group 25