



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:

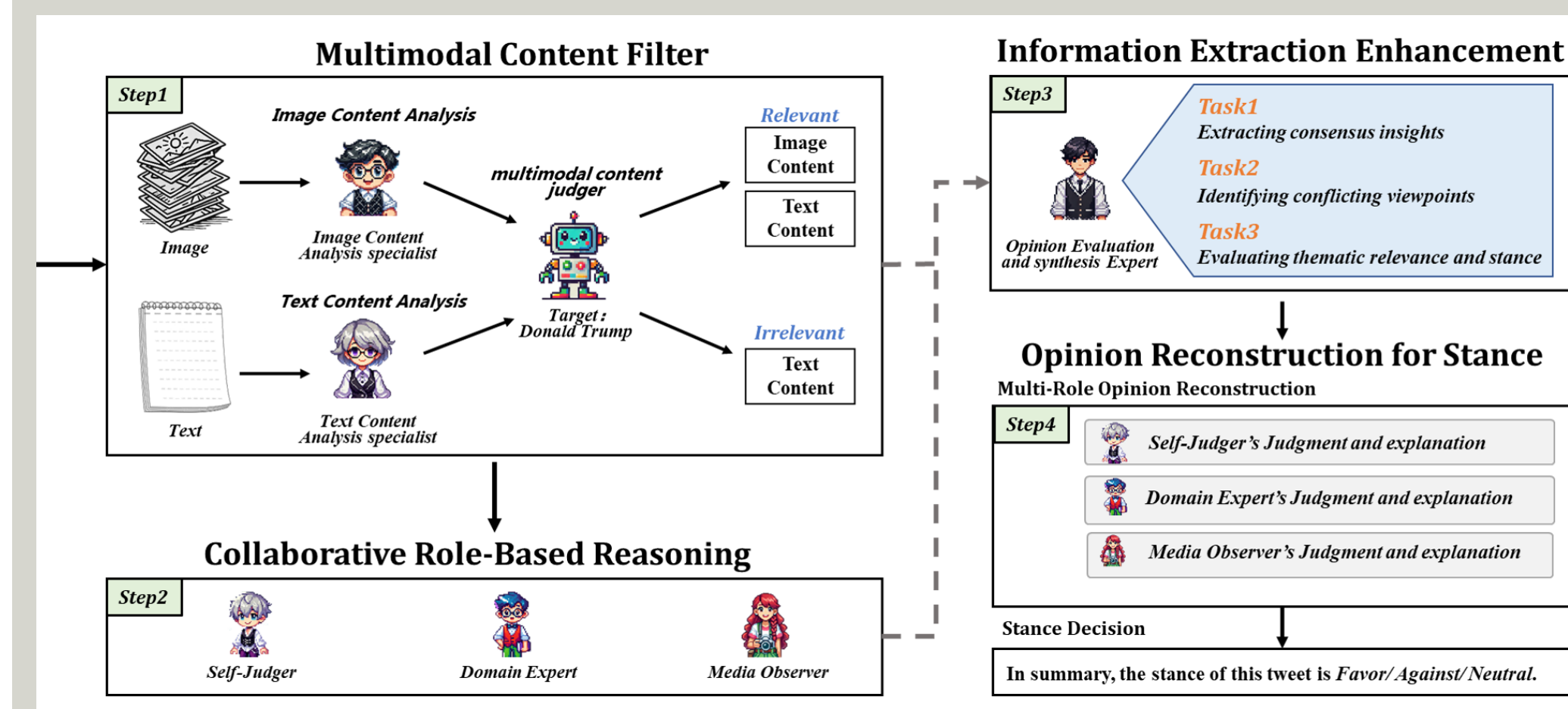


Figure 1. Supporting Work Pipeline.

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 $\approx 260,000$ 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.

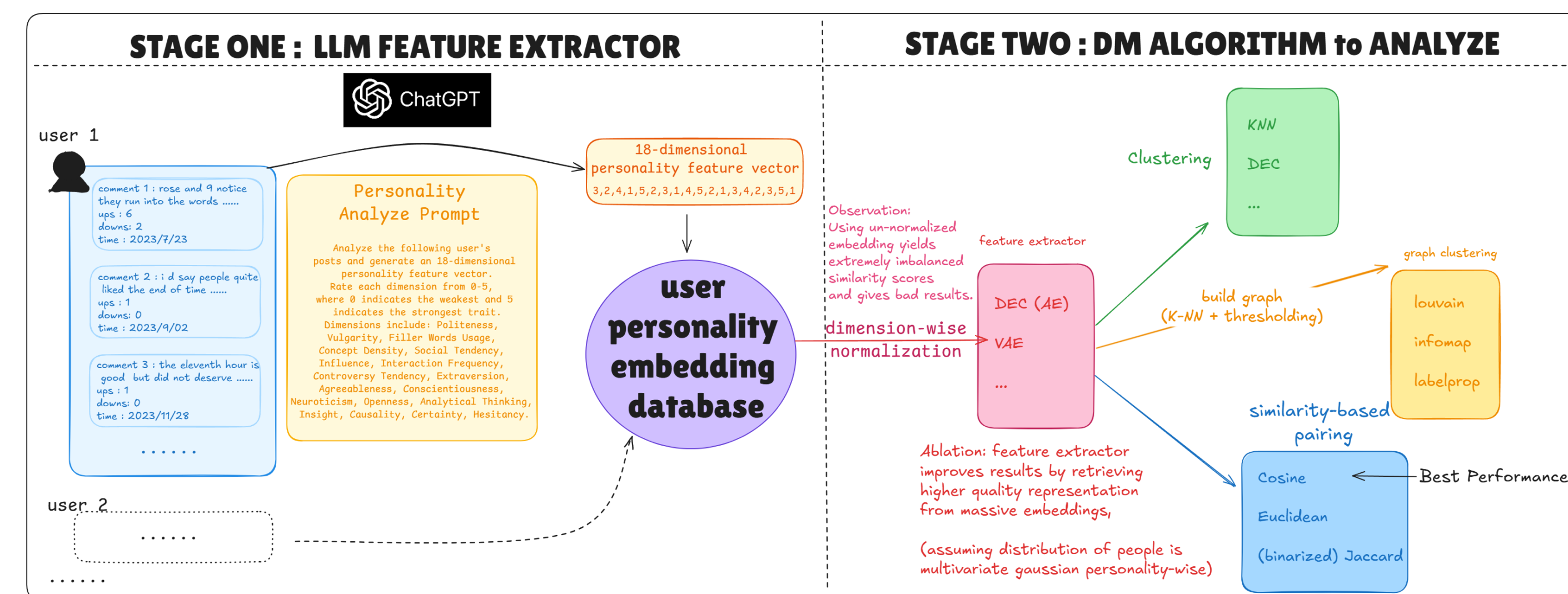


Figure 2. LLM-VAC Pipeline.

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 likelihood idea to judge the accuracy, so **the actual accuracy will be much higher than the accuracy we measured.**

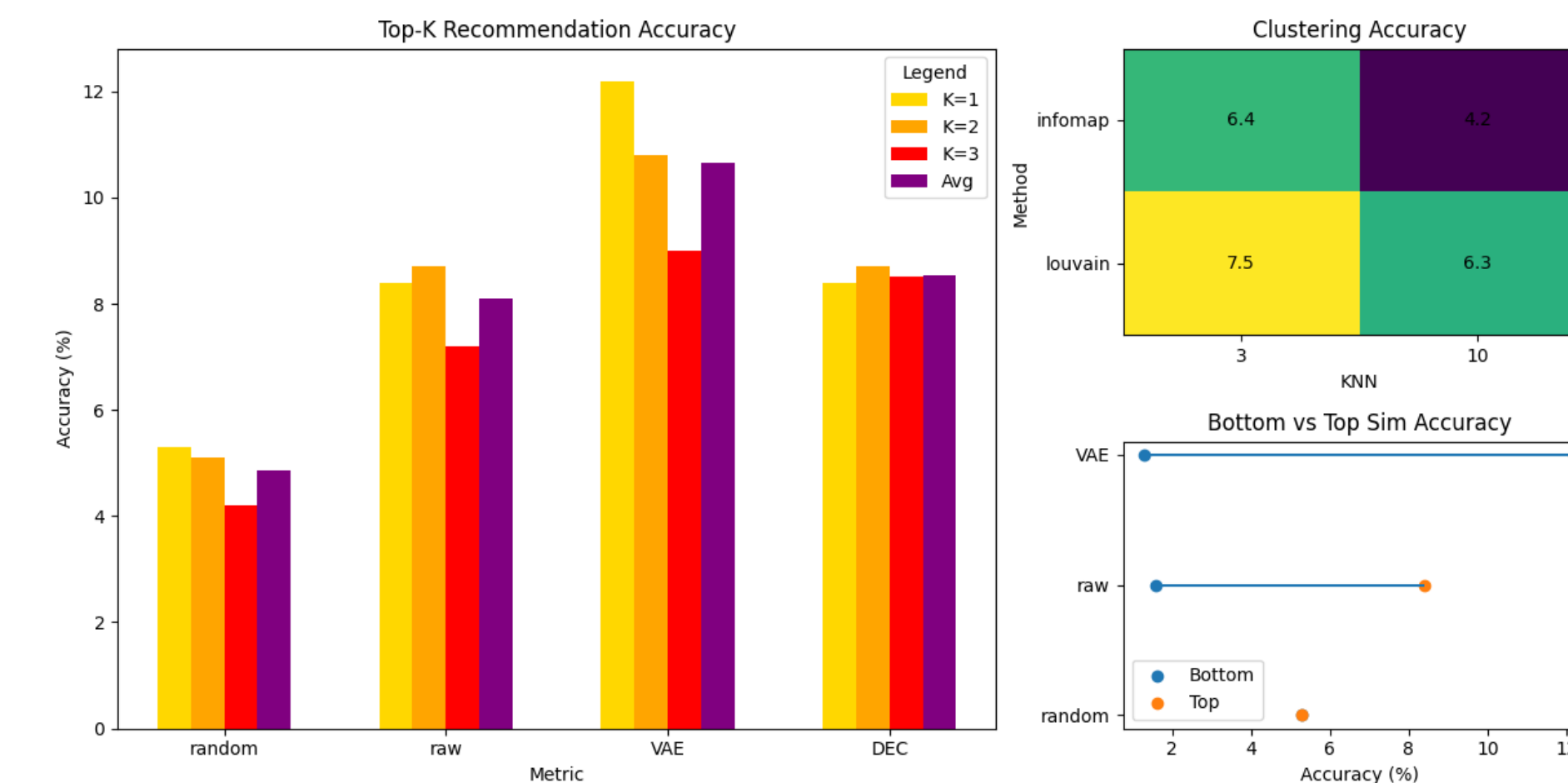


Figure 3. Experiment Results

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.