Reasoning-Based Refinement of Unsupervised Text Clusters with LLMs

Tunazzina Islam

Department of Computer Science

Purdue University, West Lafayette, IN 47907, USA

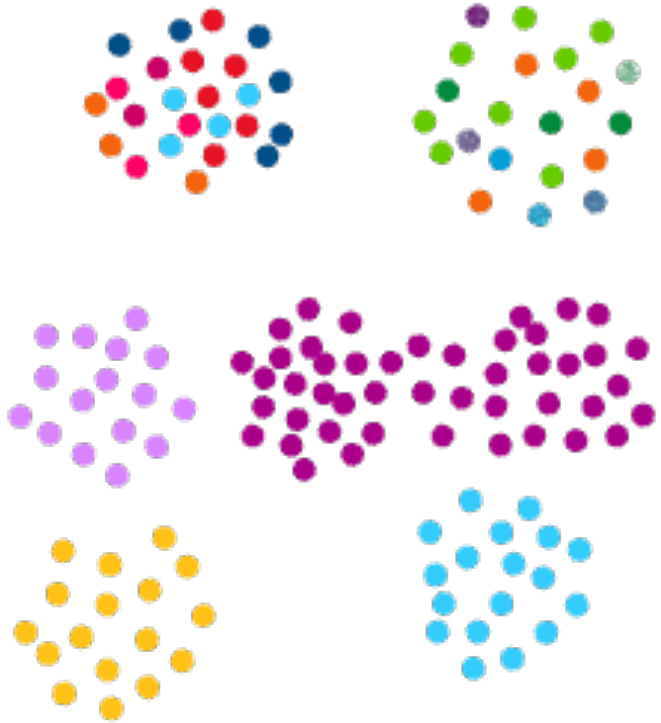
https://tunazislam.github.io/

ACL 2026 SAN DIEGO JULY 2–7

SCAN ME

PURDUE UNIVERSITY

Unsupervised Clusters Are Often Semantically Wrong



Incoherent Clusters

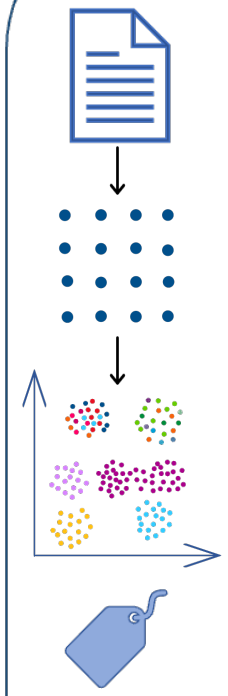
Overlapping Clusters

Unlabeled Clusters

Embedding-based clustering assumes geometry = meaning.
But well-separated clusters are not always semantically coherent.

From Geometric Clustering → Semantic Validation

Existing Paradigm



Text Data
Raw documents

Embeddings
E.g., SBERT, other encoders

Clustering
E.g., HDBSCAN, K-means

Topic Labels
Automatically generated

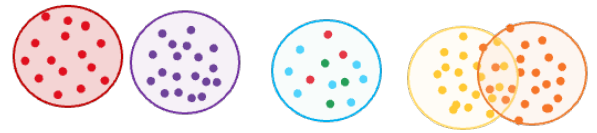
Optimizes:
- Distance
- Density
- Seperation

No semantic verification
Clusters accepted as final answers

Clusters are hypotheses, not final answers.

Our Perspective

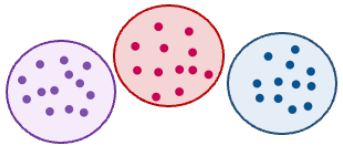
Initial Clusters (Hypotheses)



LLM Semantic Judge

- Does the cluster represent a single coherent theme?
- Are clusters semantically overlapping?
- Generate and consolidate interpretable, well-grounded labels.

Refined Semantic Structure

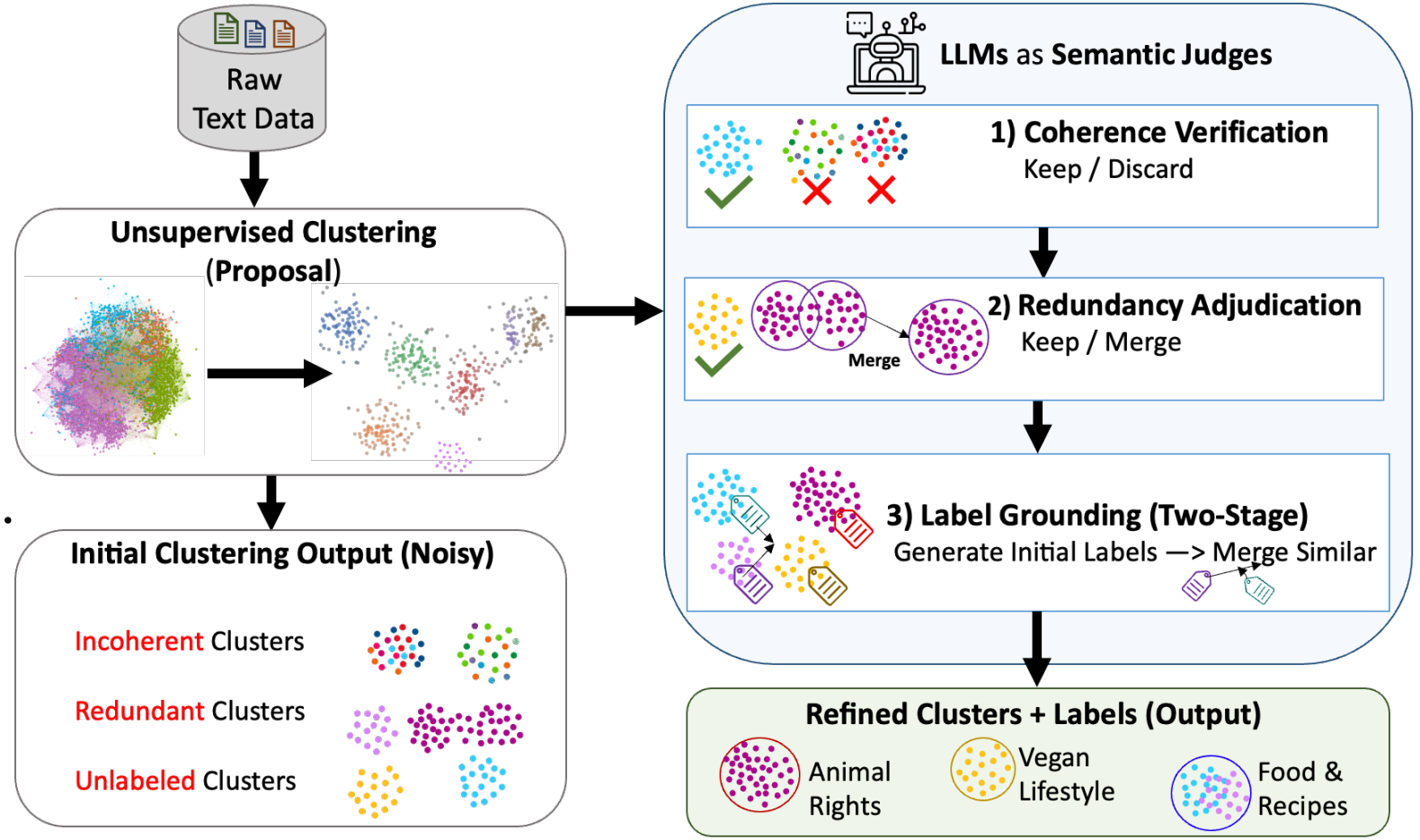


Coherent, distinct & interpretable clusters

Reasoning-Based Refinement

Three Design Principles:

1. **Clustering** as **proposal**.
2. **LLMs** as **semantic judges**.
3. Explicit reasoning checkpoints.



Vegan Discourse as a Testbed

Cross-Platform Datasets.

Centralized platform

20K tweets

📣 advocacy-driven discourse

📡 broadcast style

Decentralized platform

13K posts

💬 conversational discourse

🌐 community-driven

Example from Vegan Discourse

"Plant-based diets reduce emissions and improve sustainability." [environment, health]

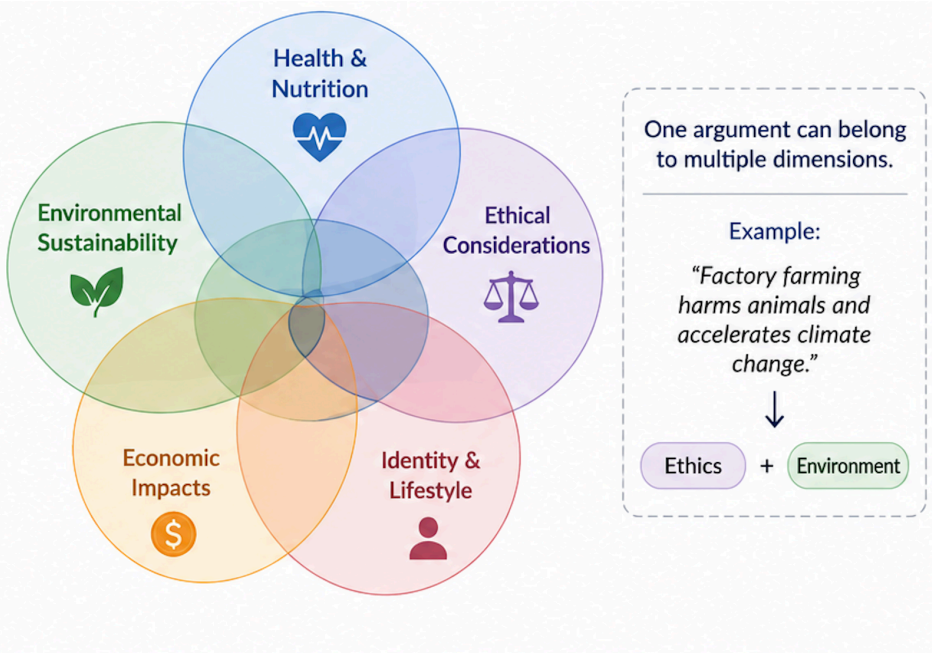
"Vegan diets lack essential nutrients." [health]

"Factory farming is morally unacceptable." [ethics]

"Meat taxes hurt working-class families." [economics]

"Being vegan is part of who I am." [identity]

Overlapping Semantic Dimensions



Unsupervised clusters often mix multiple semantic dimensions, duplicate ideas, or produce vague labels. We need **semantic validation** and **reasoning-based refinement** to obtain coherent and interpretable structure.

Example of Incoherent Cluster

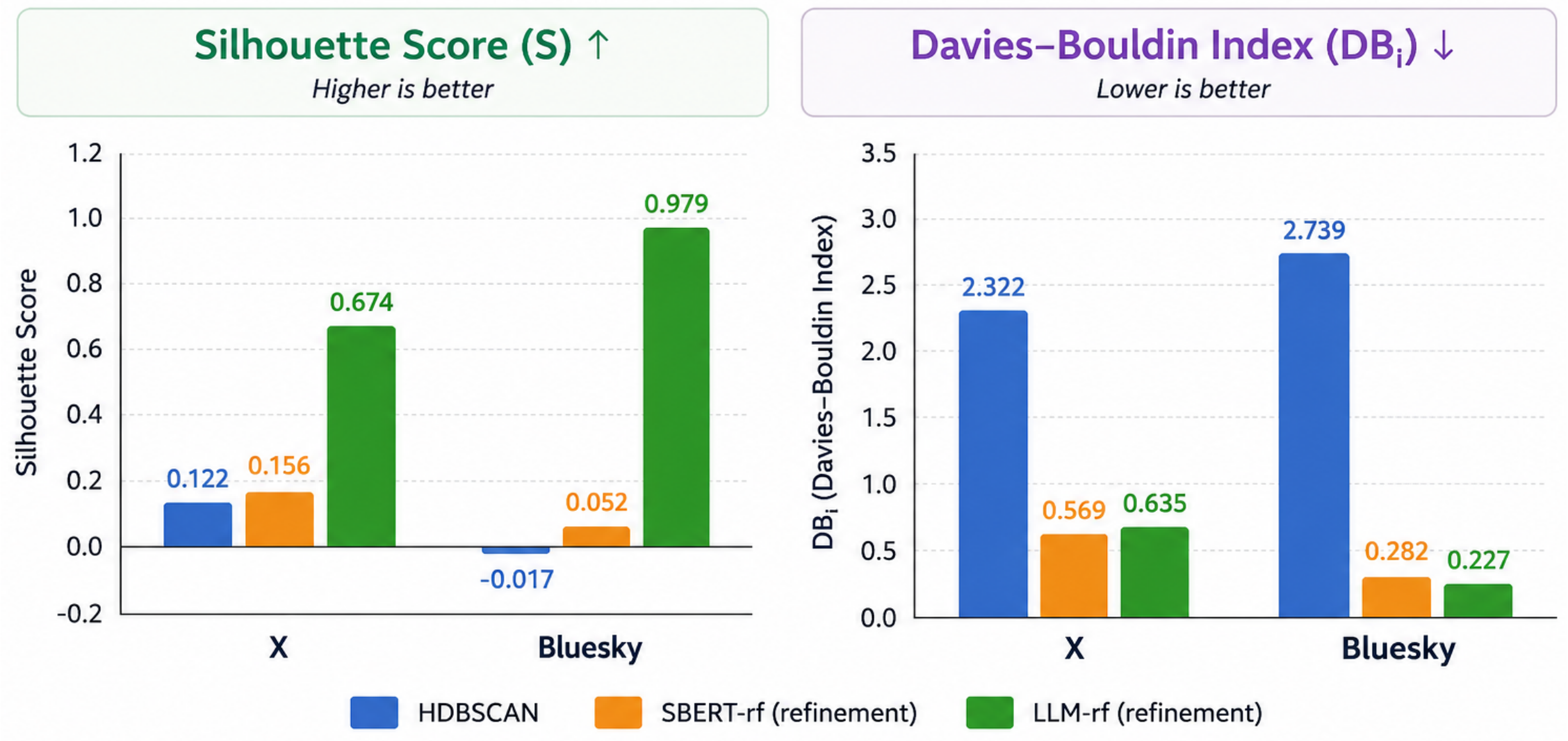
Text1: Folic Acid for Pregnant Women. Folic acid is a B vitamin that is found in vegan diets, some kinds of fruits and cereals and animal products.
Text2: Not to be an annoying vegan but you're all fucking disgusting for consuming animal products
Text3: me retreating to my vegan discord after arguing with animal abusers https://t.co/e26X6XUtsf
Text4: @KingDavey1000 @herbivore_club @BackRoa61286135 It's not false, though. You fight what should be your comrades in the fight against animal abuse more than you fight that abuse. Because we're "that type of vegan" doesn't mean we deserve your ire when we're fighting the same cause.
Text5: "Vegan" Clout Crew coming in to defend animal abuse.

Example of Merged Cluster Summaries

Summary of cluster 1: The text encourages starting your day with motivation and inspiration. They emphasize various themes such as money, business, leadership, wisdom, writing, education, beauty, online success, technology, and lifestyle. The hashtags indicate a focus on personal and professional growth, power, style, and lifestyle choices like veganism. The messages are repeated to highlight the importance of these concepts, especially on a Sunday.

Summary of cluster 18: The texts encourage people to get their daily dose of motivation and inspiration before starting the day. It includes a wide array of hashtags related to money, business, leadership, wisdom, writing, education, beauty, online success, technology, veganism, love, life, jobs, style, and several Friday-themed feelings, all targeted towards an audience in the USA.

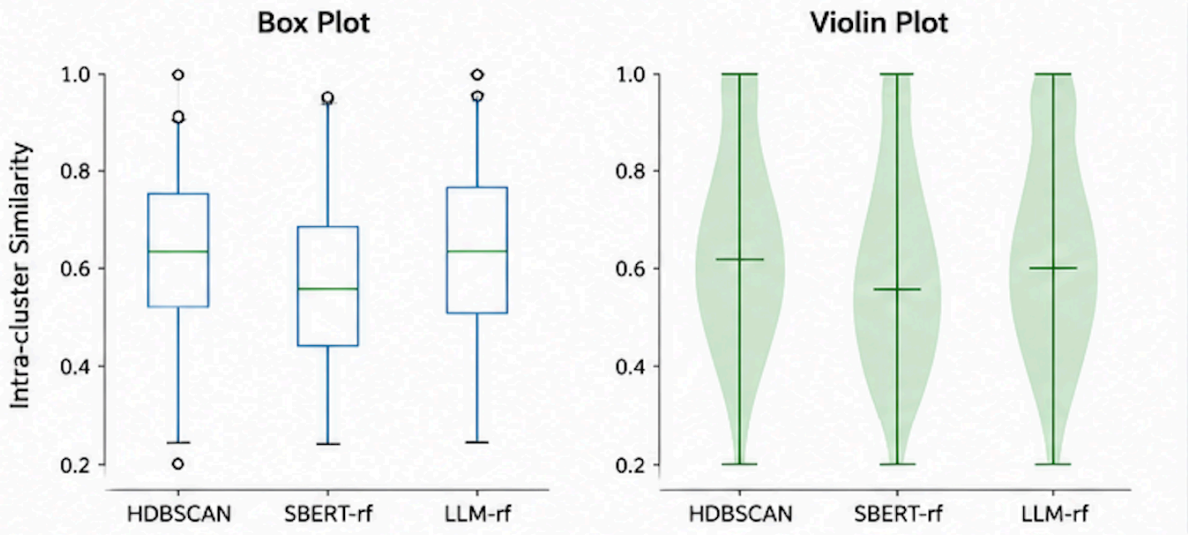
LLM Refinement Improves Cluster Quality



Semantic Coherence & Significance

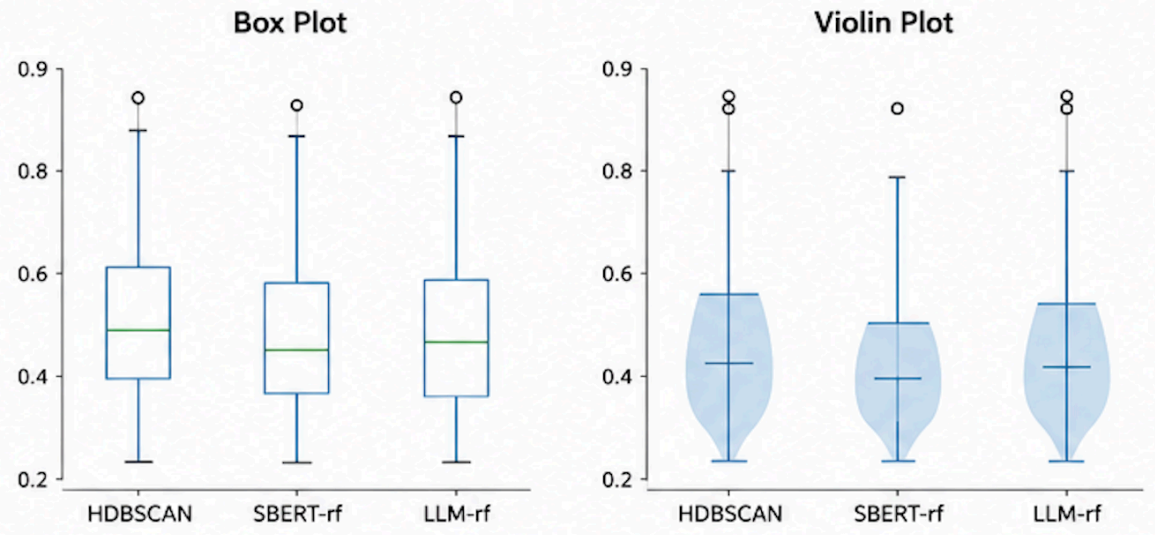
Semantic Coherence: Mean pairwise cosine similarity within each cluster (≥0.85 on sentence embeddings).

X (Twitter) Dataset



X: HDBSCAN and LLM-rf achieve consistently higher intra-cluster similarity than SBERT-rf. Their medians are comparable (~0.60).

Bluesky Dataset



Bluesky: All methods show similar coherence distributions. No clear advantage of any method.

Statistical Significance (Cluster Metrics)

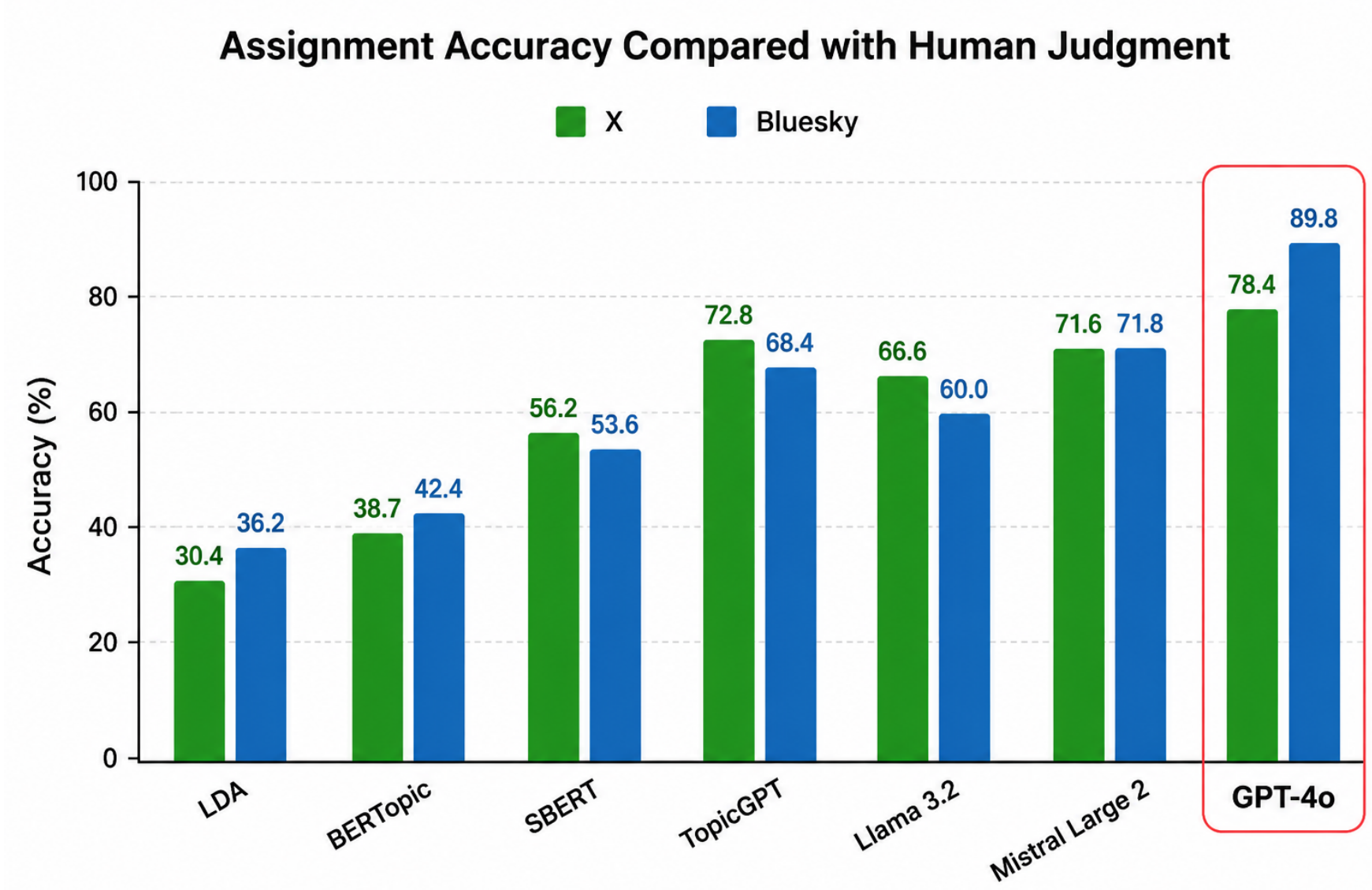
Kruskal–Wallis test + Pairwise Mann–Whitney U test

X Dataset	
Kruskal–Wallis (overall)	H = 16.187, p = 0.0003 ($p < 0.001$)
HDBSCAN vs LLM-rf	U = 42706.0, p = 0.4808 (n.s.)
HDBSCAN vs SBERT-rf	U = 52922.5, p = 0.000165 ($p < 0.001$)
SBERT-rf vs LLM-rf	U = 24018.0, p = 0.001829 ($p < 0.01$)

Bluesky Dataset	
Kruskal–Wallis (overall)	H = 0.2809, p = 0.8690 (n.s.)
HDBSCAN vs LLM-rf	U = 680.0, p = 0.8816 (n.s.)
HDBSCAN vs SBERT-rf	U = 674.5, p = 0.6044 (n.s.)
SBERT-rf vs LLM-rf	U = 582.0, p = 0.7289 (n.s.)

n.s. = not statistically significant (p-value > 0.05)

Human Evaluation



Platform-Specific Discourse Structure

Centralized platform

- 📣 advocacy-driven
- 📡 broadcast style
- ✊ mobilization dynamics
- ⚡ polarized narratives
- 🏛️ political framing

Decentralized platform

- 🌱 lifestyle-centered
- 🌐 community-driven
- 💬 conversational tone
- 🎭 humor & culture

Failure Cases and Limitations

⚠ **Failure Modes**

😏 **Sarcasm:** implicit meaning missed

🧠 **Implicit Morality:** hidden ethical reasoning

🔑 **Keyword Over-Reliance**
lexical similarity ≠ semantic similarity

🔀 **Overlapping Themes**
arguments span multiple dimensions

🚧 **Limitations**

🌎 English-only analysis

🌱 Vegan discourse domain

⏱ Temporal mismatch across platforms

🤖 Potential LLM bias in semantic judgments

Reasoning-based refinement improves interpretability, but semantic validation remains an open challenge.

Takeaways

- **Clustering ≠ semantic validity.**
- LLMs can act as **semantic judges.**
- Post-hoc refinement **improves** interpretability.

Thank You !

Reasoning-Based Refinement of Unsupervised Text Clusters with LLMs

Tunazzina Islam, Ph.D.

Department of Computer Science,

Purdue University, West Lafayette, IN.

Email: islam32@purdue.edu

https://tunazislam.github.io/

@Tunaz_Islam

PURDUE UNIVERSITY