Clustering AI Chat Transcripts at Session Level: An Empirical Study of Embeddings, Dimensionality Reduction, Clustering, and LLM-Based Evaluation

Abhishek Verma

Independent research · experiments and analysis assisted by Claude (Anthropic)

Abstract

Discovering topics in conversations with AI assistants is a core tool for understanding how large language models are used, yet the design space — session encoding, embedding model, dimensionality reduction, clustering algorithm, and evaluation protocol — is rarely examined jointly. We present a systematic empirical study of this pipeline: 335 clustering runs spanning seven datasets (four labeled benchmarks and three real AI-chat corpora: LMSYS-Chat-1M, WildChat-1M, and Chatbot Arena), four embedding models, three reduction techniques, three clustering algorithms, and three session-encoding strategies, evaluated with internal indices, external ground-truth agreement, and an LLM-as-judge protocol comprising label-blind coherence ratings and an intruder detection test. We find that (i) embedding quality dominates all downstream choices, with commercial Gemini embeddings improving ARI by +0.2–0.4 over strong open-weight SentenceTransformers; (ii) internal validity indices correlate only weakly with external accuracy (Spearman $\rho \approx 0.35$); (iii) the dominant failure mode on real chat corpora is *degenerate collapse*, in which UMAP+HDBSCAN places 65–98% of sessions in a single cluster — a failure invisible to per-cluster metric averages and unstable across sample draws; and (iv) cluster cores are highly reproducible (pairwise ARI 0.89–0.97) while noise assignment is not. We recommend a guarded pipeline — full-transcript encoding, gemini-embedding-001, UMAP, HDBSCAN, and a mandatory largest-cluster-share check, with leaf-selection re-clustering as the free first-line fallback and summarize-then-embed re-encoding when coverage matters — and release the framework, an interactive results explorer, and all per-cluster LLM labels.

1. Introduction

Conversational AI systems generate millions of chat sessions daily. Understanding what users actually do in these conversations — debugging code, drafting emails, role-play, medical questions — matters for product decisions, safety monitoring, and research on usage patterns, as demonstrated at scale by Anthropic's Clio system (Tamkin et al., 2024) and the analyses accompanying LMSYS-Chat-1M (Zheng et al., 2024) and WildChat (Zhao et al., 2024). The standard recipe is unsupervised: embed each conversation, optionally reduce dimensionality, cluster, and name the clusters. Each stage offers many choices, and published systems differ at every one of them, yet controlled comparisons over the *joint* pipeline on *session-level* chat transcripts are scarce. Benchmarks such as MTEB (Muennighoff et al., 2023) evaluate embeddings for clustering, but on short single texts with fixed downstream algorithms and internal metrics computed in the clustering space.

This paper asks a practical question: **which combination of session encoding, embedding, reduction, clustering, and evaluation should a practitioner use to discover topics in AI chat transcripts, and how should they know when it has failed?** We answer it with a five-phase study (Section 4) whose design choices — e.g., reporting internal metrics in the original embedding space, weighting judge scores by cluster size, and stress-testing on real corpora rather than only benchmarks — each turn out to change the conclusions materially.

Our contributions:

(1) A controlled 288-run grid over 4 datasets × 2 encodings × 4 embeddings × 3 reductions × 3 clusterers, with external validation on labeled dialogue corpora. **(2)** An LLM-as-judge protocol (label-blind coherence + intruder test) applied to 35 labeled runs, including all three major public AI-chat corpora at up to 10K sessions. **(3)** Identification and quantification of *degenerate collapse* as the dominant real-world failure mode, its invisibility to unweighted metrics, its sample-level instability, and a cheap guard that detects it plus a re-encoding fallback that empirically eliminates it. **(4)** A stability analysis separating reproducible cluster cores from unstable noise boundaries. **(5)** An open framework, all 335 run records, per-cluster LLM labels, and an interactive explorer.

2. Related Work

Topic modeling and text clustering. Classical topic models (LDA) have largely been superseded for short informal text by embedding-based pipelines. BERTopic (Grootendorst, 2022) popularized the UMAP→HDBSCAN→c-TF-IDF recipe that our study treats as a primary candidate. Sentence embeddings for clustering trace to Sentence-BERT (Reimers & Gurevych, 2019); the MTEB benchmark (Muennighoff et al., 2023) ranks embedding models on clustering tasks via V-measure. Our study differs in clustering full multi-turn sessions and in crossing embeddings with downstream algorithm choices rather than fixing them.

Analyzing AI chat corpora. Zheng et al. (2024) cluster LMSYS-Chat-1M prompts; Zhao et al. (2024) analyze WildChat topics; Clio (Tamkin et al., 2024) summarizes each conversation with an LLM before embedding and clustering — the summarize-then-embed strategy we benchmark in Section 4.5 — and names clusters with an LLM, as we do.

Cluster evaluation. Internal indices include silhouette (Rousseeuw, 1987), Calinski–Harabasz, Davies–Bouldin, and the density-based DBCV (Moulavi et al., 2014); external

agreement uses ARI (Hubert & Arabie, 1985) and NMI. Our intruder test adapts the word-intrusion protocol of Chang et al. (2009) to whole conversations with an LLM judge, providing an objective, label-free measure of cluster distinctness. LLM-as-judge evaluation is increasingly standard; we mitigate self-preference by generating labels and judging with different models.

3. Methodology

3.1. Datasets

Dataset	Sessions	Type	Labels
MultiWOZ 2.2	8.4K	task dialogue	service domains
Banking77	10K	intents (1-turn)	77 intents
OASST1	10K trees	open assistant	none
HH-RLHF	170K	open assistant	none
LMSYS-Chat-1M	1M	real AI chat	none
WildChat-1M	1M	real AI chat	none
Chatbot Arena	33K	real AI chat	none

Table 1. Datasets. Labeled corpora anchor external metrics; the three gated real-AI-chat corpora are the deployment target. Runs sample 2,000 English sessions (10,000 for scale checks) with fixed seeds.

Each conversation is flattened to a single `USER:/ASSISTANT:` transcript. Because most chat corpora lack topic labels, we include two labeled dialogue datasets so that external metrics (ARI/NMI) can calibrate the label-free ones; MultiWOZ dialogues are labeled with the '+'-joined set of active service domains.

3.2. Session encodings

truncate: first 4,000 characters of the transcript. **user_turns**: user messages only. **summarize**: a 1–3 sentence intent summary generated by gemini-2.5-flash (concurrent, disk-cached, with refusal fallback to the truncated transcript), after Clio. A chunk-and-mean-pool encoding is implemented but not benchmarked here.

3.3. Embeddings

Two open-weight SentenceTransformers — all-MiniLM-L6-v2 (384d) and all-mpnet-base-v2 (768d) — and two commercial Google models — gemini-embedding-001 and the multimodal gemini-embedding-2 (both 3072d; the latter requires one Content object per input, as a bare string batch is silently pooled into a single embedding). All vectors are L2-normalized; API calls run concurrently with exponential-backoff retry and are disk-cached.

3.4. Dimensionality reduction

None, PCA-50, or UMAP-25 (n_neighbors 15, min_dist 0, cosine; McInnes et al., 2018) before clustering; t-SNE (van der Maaten & Hinton, 2008) and UMAP are used for 2-D visualization only. Internal metrics computed in a reduced space are inflated (silhouette 0.78 in UMAP space vs. 0.07 in the original space for the same partition), so all reported internal metrics are computed in the original embedding space.

3.5. Clustering

KMeans (k selected by silhouette sweep over 8–60 on a subsample), HDBSCAN (Campello et al., 2013; min_cluster_size = max(15, n/200), EOM selection; noise label −1), and agglomerative clustering (cosine, average linkage). GMM and spectral clustering are supported but not in the main grid.

3.6. LLM labeling and judging

For each cluster, the 8 members nearest the centroid plus 4 random members are shown to gemini-2.5-flash, which returns a 2–6 word topic, a summary, and flags outliers (structured outputs). A separate, stronger judge (gemini-3.5-flash) then (a) rates a label-blind random sample's coherence 1–5 and (b) performs 20 intruder trials per run: shown five members of one cluster and one secretly injected member of another, it must identify the intruder (chance = 1/6). Using different models for generation and judging reduces self-preference bias.

3.7. Metrics and aggregation

Internal: silhouette (cosine), Calinski–Harabasz, Davies–Bouldin, DBCV — original space, noise excluded, noise fraction reported separately. External: ARI, NMI, V-measure. LLM: mean judge coherence — reported both per-cluster and *size-weighted* — and intruder accuracy. Pipelines are compared by Borda rank aggregation (per-dataset ranks per metric, external metrics double weight); Section 5.4 shows why size-weighting is not optional. Degeneracy is flagged when the largest cluster holds >50% of assigned sessions.

4. Experiments

Phase 1–2 (benchmark grid, 288 runs): {OASST1, HH-RLHF, MultiWOZ, Banking77} × {truncate, user_turns} × 4 embeddings × {none, PCA-50, UMAP-25} × {KMeans, HDBSCAN, agglomerative}, 2,000 sessions per run. **Phase 3 (LLM-judged, 16 runs):** the top pipelines with full labeling, judging, and intruder testing. **Phase 4 (real corpora, 9 runs + 2 scale runs):** the three finalists on LMSYS-Chat-1M, WildChat, and Chatbot Arena at 2K, and the recommended pipeline at 10K on LMSYS and Arena. **Phase 5 (encoding ablation, 8 runs):** summarize-then-embed vs. truncate. **Stability (12 runs):** 3 clustering seeds on fixed data and 3 overlapping sample draws from 4K pools, compared by pairwise ARI on shared sessions. All experiments use fixed seeds; embeddings and summaries are cached; total commercial API cost was under $25.

5. Results

5.1. Embedding quality dominates

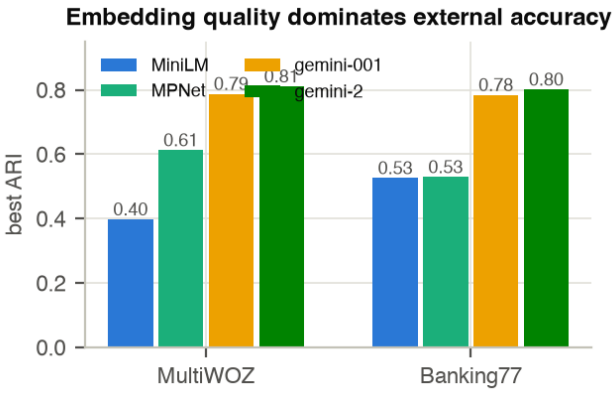

Figure 1. Best ARI over all pipeline variants, per embedding and labeled dataset. The gap between embedding models (+0.2–0.4 ARI) exceeds any gain available from reduction or clusterer choices downstream.

Gemini embeddings dominate open-weight models on both labeled datasets (Figure 1): best ARI 0.80/0.81 (gemini-2) and 0.78/0.79 (gemini-001) vs. 0.53/0.61 for MPNet on Banking77/MultiWOZ; a 3-seed replication (Section 5.6) confirms the family gap but finds the two Gemini models statistically indistinguishable from each other. On Banking77, gemini-2 pipelines recover ≈60 clusters against 77 true intents at NMI up to 0.95. With strong embeddings the reduction step barely matters for accuracy (Banking77 ARI: raw 0.795, PCA 0.801, UMAP+HDBSCAN 0.790) but matters enormously for cost: clustering raw 3072-d vectors made runs up to 42× slower (mean 172s vs. 4.1s; k-sweep KMeans 583s), so reduction is justified on efficiency grounds alone. Encoding interacts with embedding strength: user_turns helps weak local models (assistant boilerplate dilutes their signal) while full transcripts win with long-context Gemini models.

5.2. Internal metrics weakly track external truth

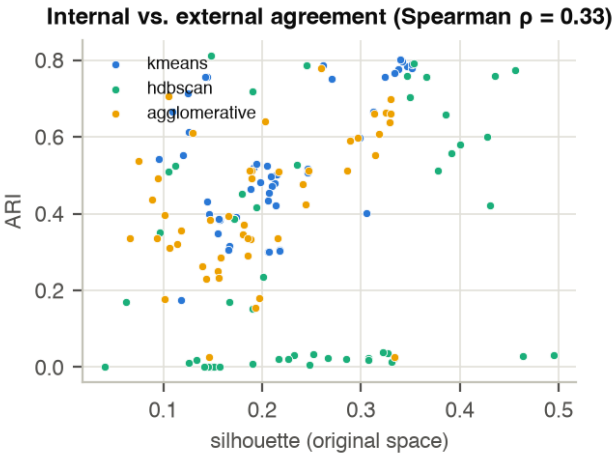

Figure 2. Original-space silhouette vs. ARI across all runs with both defined. Internal geometric quality only weakly predicts agreement with human labels.

Across 152 runs with both metrics defined, original-space silhouette and ARI correlate at only Spearman ρ≈0.35 (Figure 2). Density-based DBCV disagrees with ARI even about ranking clusterers (it prefers HDBSCAN's solutions where ARI ranks them last on MultiWOZ). No internal index is a safe proxy for label agreement; we therefore treat internal metrics as tie-breakers and rely on external and LLM-based measures.

5.3. Degenerate collapse on real chat corpora

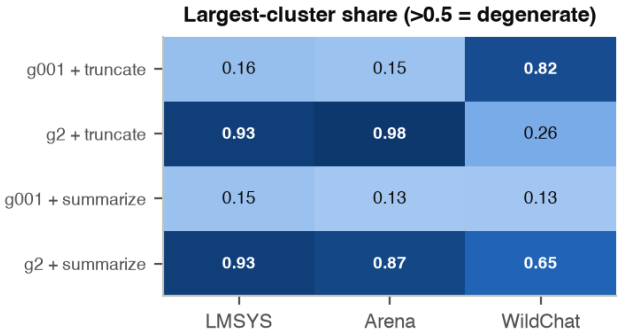

Figure 3. Largest-cluster share of assigned sessions for UMAP+HDBSCAN pipelines on the real corpora (2K sessions). Bold cells are degenerate (>0.5). Collapse depends on the embedding × corpus pair, and summarize-then-embed eliminates it only for gemini-001.

On curated benchmarks every finalist looked healthy. On real corpora, UMAP+HDBSCAN collapsed into a single cluster holding 65–98% of sessions for at least one corpus under *each* embedding (Figure 3): gemini-2 on LMSYS (93%) and Arena (98%), gemini-001 on WildChat (82%). Intruder accuracy stays high during collapse (the small surviving clusters are genuinely distinct), so distinctness metrics cannot detect it. KMeans never degenerates (max share 5–17%) — its failure mode is graceful mediocrity rather than catastrophe. Judge coherence on real corpora tops out at 2.5–3.3 vs. ≈5 on task-oriented benchmarks: results reported only on curated data substantially overstate real-world quality.

5.4. Aggregation can hide degeneracy

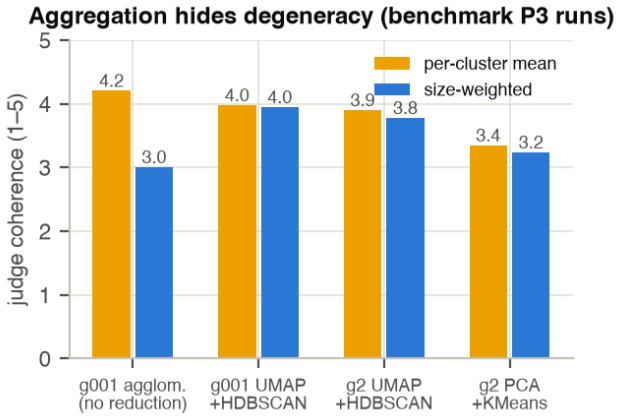

Figure 4. Per-cluster mean vs. size-weighted judge coherence on the benchmark P3 runs. The unreduced agglomerative pipeline wins the unweighted comparison while placing ~99% of open-chat sessions in one blob; weighting by cluster size reverses the ranking.

The unreduced agglomerative pipeline won both the Borda ranking and the per-cluster judge average (4.21/5) while assigning 99% of OASST1 and HH-RLHF sessions to a single cluster: many tiny perfect clusters outvoted one enormous incoherent one. Size-weighting drops it from first to last (3.01) and crowns gemini-001+UMAP+HDBSCAN (3.95, intruder 1.00) (Figure 4). We conclude that cluster-level scores — including LLM-judge scores — must be weighted by cluster

mass, and size distributions must be inspected, before any ranking is trusted.

5.5. Scale, encoding ablation, and stability

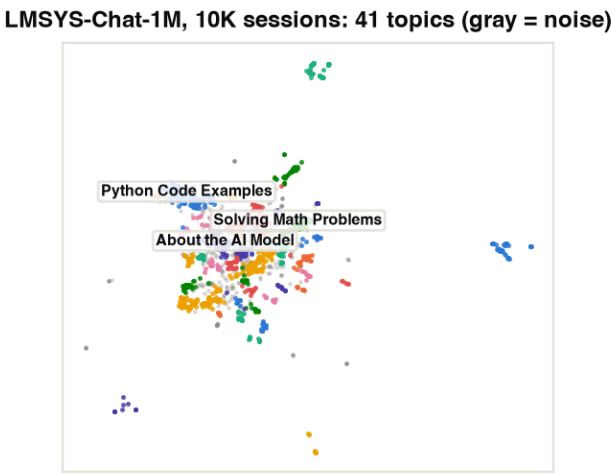

Figure 5. The recommended pipeline on 10K LMSYS-Chat-1M sessions (2-D UMAP projection of a 1,500-session sample; labels for the largest clusters). 41 topics, 26% noise, largest cluster 11%, size-weighted judge coherence 2.57, intruder accuracy 1.00.

Scale stabilizes. At 10K sessions the recommended pipeline produced healthy, face-valid structure on both LMSYS (41 clusters; Figure 5) and Arena (38 clusters), with better judge coherence than at 2K and ≈2 minutes of clustering compute. Several large LMSYS clusters could not be labeled because the labeling model refuses adult/role-play content — production systems need an explicit refusal path.

Summarize-then-embed (Table 2) eliminated degeneracy entirely for gemini-001 (max share ≤15% on all corpora, including WildChat where truncate collapsed) at equal coherence, but *backfired* for gemini-2 (3 of 4 corpora degenerate) and costs external accuracy on fine-grained labels (MultiWOZ ARI 0.72 vs. 0.79 for g001; 0.52 vs. 0.81 for g2). Collapse is an embedding-geometry × HDBSCAN interaction, not a transcript-length artifact; changing the encoding relocates rather than removes it.

	LMSYS	Arena	WildChat	MultiWOZ ARI
g001 truncate	2.51 / .16	2.48 / .15	1.72 / .82	0.79
g001 summarize	2.52 / .15	2.06 / .13	2.40 / .13	0.72
g2 truncate	1.28 / .93	1.02 / .98	2.85 / .26	0.81
g2 summarize	1.23 / .93	1.24 / .87	2.12 / .65	0.52

Table 2. Encoding ablation with UMAP+HDBSCAN: size-weighted judge coherence / largest-cluster share (real corpora) and MultiWOZ ARI. Shares >0.5 are degenerate.

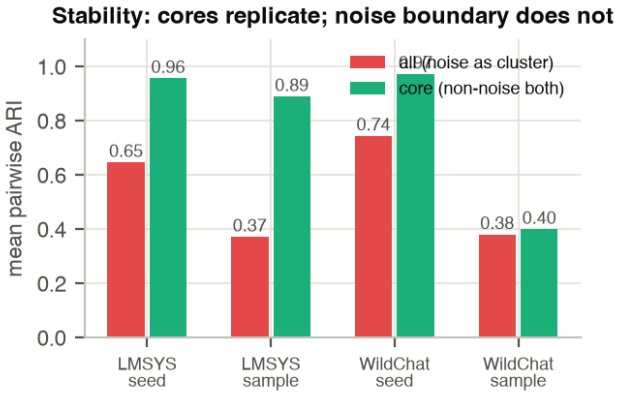

Figure 6. Stability of the recommended pipeline: mean pairwise ARI across 3 clustering seeds (same sessions) and 3 overlapping 2K sample draws (fixed seed). Cores replicate; noise assignment and, on WildChat, per-draw collapse do not.

Stability. Cluster cores are highly reproducible when runs are healthy: core ARI 0.96–0.97 across seeds and 0.89 across LMSYS sample draws (Figure 6). The unstable components are (a) the noise boundary — which sessions are excluded varies, dragging all-points ARI to 0.37–0.74 — and (b) on WildChat, collapse itself: 2 of 3 sample draws degenerated (78%/84% blobs) while all seed replicates on a different draw were healthy. Degeneracy is a per-draw coin flip, so the guard must run on every clustering batch.

5.6. Post-hoc ablations: leaf selection and seed replication

Leaf selection fixes the collapse for free — with trade-offs. Re-clustering every degenerate configuration with HDBSCAN cluster_selection_method='leaf' (identical UMAP spaces) de-collapsed all 12 tested configs, including both collapsed WildChat sample draws (max share 0.78–0.98 → 0.07–0.12). Costs: noise rises to 36–54% (vs. 23–34% for the summarize-then-embed fallback), and ARI splits by taxonomy granularity — Banking77 (77 fine intents) improves 0.70→0.87 while MultiWOZ (coarse domains) drops 0.79→0.65. The guard in Section 6 is therefore tiered: leaf first, re-encoding when coverage or a coarse taxonomy matters.

Config (UMAP+HDBSCAN)	EOM share	Leaf share	EOM ARI	Leaf ARI
g2 / LMSYS	0.93	0.09	—	—
g2 / Arena	0.98	0.12	—	—
g001 / WildChat	0.82	0.09	—	—
g001 / Banking77	0.07	0.03	0.70	0.87
g001 / MultiWOZ	0.10	0.12	0.79	0.65

Table 3. Leaf-selection ablation (selected rows; full table in the repository). Leaf de-collapses every degenerate config; ARI moves with taxonomy granularity.

3-seed replication of the key cells. Replicating {MultiWOZ, Banking77} × 4 embeddings × {UMAP+KMeans, UMAP+HDBSCAN} over three full-pipeline seeds (fresh sample, UMAP, and clusterer per seed): the Gemini-vs-local family gap holds in every cell's mean (e.g., Banking77 HDBSCAN 0.74/0.80 vs. 0.50/0.51), but gemini-2 > gemini-001 held in only 1/3 seeds in 3 of 4 cells — the two are best treated as tied, and the single-seed maxima reported in Section 5.1 carry selection inflation. Notably,

g001+HDBSCAN is the most stable combination (ARI std 0.014–0.031, vs. up to 0.276 for gemini-2 and 0.14–0.23 for Gemini+KMeans, whose silhouette-swept k is seed-sensitive) — supporting it as the default on stability grounds rather than mean accuracy. The KMeans-vs-HDBSCAN ordering is dataset-dependent, consistent with the granularity account.

6. Discussion: a Guarded Pipeline

Our results yield a concrete recommendation. **Default:** truncate encoding → gemini-embedding-001 → UMAP-25 → HDBSCAN → cheap-LLM labels with a stronger label-blind judge. **Guard (mandatory, per batch):** compute the largest cluster's share of assigned sessions; if it exceeds 0.5, first re-cluster with HDBSCAN leaf selection (free; de-collapsed all 12 tested configs, and the best choice outright for fine-grained taxonomies); when coverage matters (leaf pushes 36–54% of sessions to noise) or the taxonomy is coarse, re-encode with summarize-then-embed (still gemini-001), or fall back to KMeans when every session must be assigned. **Variants:** gemini-embedding-2 when matching a human taxonomy is the goal (best ARI, but never with the summarize encoding); MPNet + user_turns when no API budget exists (expect a large quality drop). Treat HDBSCAN noise assignment as fuzzy: report topics, not per-session noise membership.

Methodologically, three practices changed our conclusions and should generalize: compute internal metrics in the original space; weight all cluster-level scores by cluster mass; and evaluate on the deployment distribution, not only benchmarks — every qualitative surprise in this study (collapse, refusals, the coherence ceiling) appeared only on real corpora.

7. Limitations

Judge and embeddings share a model family for the Gemini pipelines; although labeler and judge differ, a same-family preference cannot be fully excluded and a cross-family (e.g., Claude) judge is future work. Judge scores across encodings are not perfectly comparable, since the judge reads summaries for summarize-encoded runs; degeneracy and ARI carry that comparison. English-only sessions and 2K–10K samples limit scope. The full grid remains single-seed, though the key cells are replicated over 3 seeds (Section 5.6) and stability is assessed separately; min-samples sweeps remain untested. Coherence 1–5 is coarse; human evaluation was limited to spot checks of ≈60 cluster labels. BERTopic's full pipeline (c-TF-IDF representation) and a chunk-mean-pool encoding were not benchmarked.

8. Conclusion

We presented, to our knowledge, the most comprehensive controlled study of session-level AI-chat clustering to date: 335 runs across seven datasets with internal, external, and LLM-judge evaluation. The headline results — embeddings dominate, internal metrics mislead, real corpora collapse pipelines that benchmarks crown, aggregation hides the collapse, and cores (but not noise) replicate — compose into a simple guarded pipeline that is accurate on benchmarks, robust on real data, cheap (≈$1–2 per 10K sessions), and fast (minutes). All code, run records, cluster labels, and an interactive explorer accompany the paper.

References

Campello, R. J. G. B., Moulavi, D., and Sander, J. Density-based clustering based on hierarchical density estimates. *PAKDD*, 2013.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. Reading tea leaves: How humans interpret topic models. *NeurIPS*, 2009.

Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794*, 2022.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Köpf, A., et al. OpenAssistant Conversations — democratizing large language model alignment. *NeurIPS Datasets*, 2023.

Bai, Y., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*, 2022.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., and Sander, J. Density-based clustering validation. *SDM*, 2014.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. MTEB: Massive text embedding benchmark. *EACL*, 2023.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP*, 2019.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.

Tamkin, A., et al. Clio: Privacy-preserving insights into real-world AI use. *arXiv:2412.13678*, 2024.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *JMLR*, 9:2579–2605, 2008.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. WildChat: 1M ChatGPT interaction logs in the wild. *ICLR*, 2024.

Zheng, L., et al. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *ICLR*, 2024.

Budzianowski, P., et al. MultiWOZ — a large-scale multi-domain Wizard-of-Oz dataset. *EMNLP*, 2018.

Casanueva, I., Temcinas, T., Gerz, D., Henderson, M., and Vulic, I. Efficient intent detection with dual sentence encoders. *ACL NLP4ConvAI*, 2020.

Appendix A. Reproducibility Details

Hyperparameters. UMAP: n_neighbors 15, min_dist 0.0, cosine metric, 25 components (2-D only for visualization). HDBSCAN: min_cluster_size = max(15, n/200) (10 at n=2K; 50 at n=10K), EOM selection, Euclidean in the reduced space. KMeans: k swept over 8 values in [8, 60] on a ≤4K subsample, selected by cosine silhouette. Agglomerative: cosine, average linkage. Sampling seed 42 throughout; stability used seeds {42, 1, 2} and draws {101, 102, 103} from 4K pools (pool seed 7).

Models. Embeddings: sentence-transformers/all-MiniLM-L6-v2, all-mpnet-base-v2, gemini-embedding-001, gemini-embedding-2 (task_type CLUSTERING where supported; batch ≤100 with concurrency 4–6 and exponential-backoff retry). Labeling: gemini-2.5-flash; judging: gemini-3.5-flash; both via structured outputs with pydantic schemas.

Compute and cost. Apple M-series laptop (16 cores). The 288-run benchmark grid took ≈2h wall each for local and Gemini embeddings (the latter dominated by clustering 3072-d vectors; per-dataset process parallelism with BLAS threads capped per worker gave 3.4× aggregate). Judged phases: 16 P3 runs in 28 min; 12-run stability phase clustering in 23s at 904%

CPU. Total commercial API spend across ≈100K embedding calls and ≈12K labeling/judging/summarization calls was under $25. Full per-run records: results/*.csv; per-cluster labels: results/{p3,gated,scale,s2e}/<run>/clusters.csv.

Known failure modes reproduced. gemini-embedding-2 batch pooling (one embedding returned for a list of strings); labeling refusals on adult/role-play LMSYS clusters; HDBSCAN collapse as described in Sections 5.3–5.5.