# A semantic atlas of alphaXiv papers with more than 10 likes

Libor “Burny” Burian · 26 November 2026

## Abstract

We collected 100,942 unique research papers with more than 10 public likes from alphaXiv. We embedded their titles and abstracts, projected them for display, and grouped them in a separate ten-dimensional space. Across 10 seeds, two-dimensional UMAP preserved 8.99% of each paper’s ten original neighbors on average. PCA preserved 0.21%. The paired difference was 8.78 percentage points, with a 95% seed interval of 8.70 to 8.86. At the display seed, 42,441 papers remained unassigned.

## Motivation and contribution

A ranked feed answers which papers attracted attention. A researcher looking for related work also needs to know which papers discuss similar ideas and how much a two-dimensional view distorts those relationships. This release combines a verified selection boundary, source-space recommendations and inspectable topic labels in a downloadable atlas. The contribution is the audited collection and evaluation of this map.

The two dated editions share 36,276 papers. The expanded corpus includes 64,662 papers with 11–20 likes. Group-ID comparison




Figure 1. One point represents one paper. The legend lists the largest assigned groups and the unassigned set. Distances between distant regions have no calibrated interpretation.

# Collection and methods

## Census

The two accepted passes ran from 2026-11-26T14:03:39+00:00 to 2026-11-26T14:27:25+00:00. The global feed contained 101,003 qualifying ranks, including 61 distinct marked external posts. Pages of 125 initially yielded only 71,411 distinct research papers because equal-score pages overlapped. Subject-filtered queries and other page sizes recovered the missing groups. Both independent reconstructions filled the ranks at every score and agreed on all 100,942 eligible paper IDs. 0 papers needed the title-only fallback. alphaXiv API specification · Collection amendment

## Representation and algorithms

The same exhaustive double-precision cosine ranking supplied the ten original recommendations and the 30-entry graph for both UMAP fits. The graph includes self in its first column. Remaining ties use sorted paper-group ID order.

The hosted text-embedding-3-large encoder received the title, a blank line and the abstract. We normalized its vectors and calculated exact cosine neighbors, excluding each paper itself. UMAP used 30 neighbors and cosine distance. Display coordinates used two dimensions and minimum distance 0.1. A separate ten-dimensional fit used minimum distance zero, followed by Euclidean excess-of-mass HDBSCAN with minimum cluster size 20 and minimum samples 5, including the point. Seed 42 determined the public map. OpenAI embeddings · UMAP clustering guide · HDBSCAN implementation

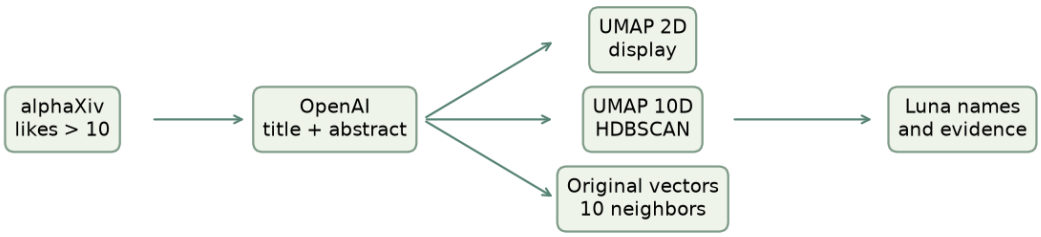


Figure 2. Display positions, group assignments and nearest-paper recommendations use three different transformations of the same hosted vectors.

## Primary analysis

For each paper we divided the intersection of its original and displayed ten nearest neighbors by ten. We averaged across papers and calculated the UMAP minus PCA difference for seeds 42–51. The two-sided Student t interval uses seed-level differences. The shuffled-position control randomly reassigned the UMAP coordinates to paper IDs once per seed. The ablation replaced ten-dimensional UMAP with ten principal components while retaining the clustering settings. Preregistered analysis · Huang et al. on neighborhood preservation

## Labeling

Luna received distinctive TF–IDF phrases and excerpts from members nearest their cluster centroid in the original vector space. Each displayed name links to supporting member IDs validated against the prompt. These are generated descriptions. Murray et al. on labeling scientific documents

# Results and controls

| Projection / assignment | Recall at 10 | Clusters | Unassigned | Shared subject* |
|---|---|---|---|---|
| UMAP / UMAP–HDBSCAN | 8.96% | 725 | 42.04% | 76.13% |
| PCA / primary assignments | 0.21% | 725 | 42.04% | 76.13% |
| Shuffled / primary assignments | 0.01% | 725 | 42.04% | 76.13% |
| PCA / PCA–HDBSCAN | 0.21% | 49 | 89.16% | 17.99% |

Table 1. Display seed 42. *Fraction of tagged within-cluster pairs sharing an arXiv category. The shuffled-tag control for primary assignments was 27.48%. Tags were available for 100,784 of 100,942 papers. The middle two rows reuse the primary labels to isolate changes in projection.

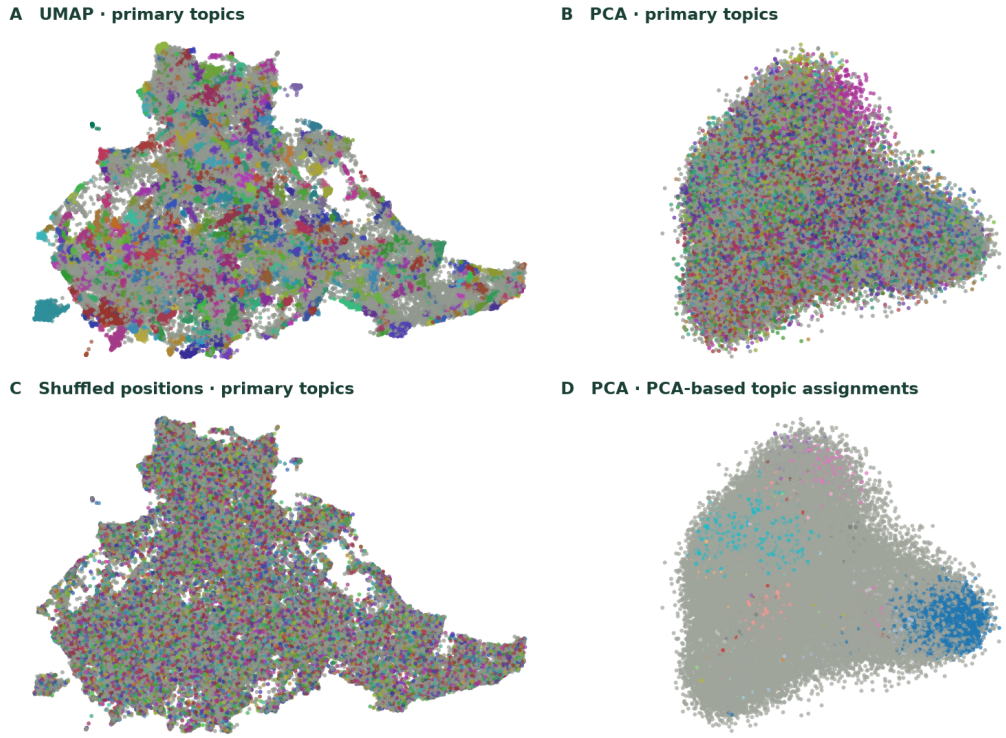


Figure 3. Panels A–C retain topic identity while changing screen coordinates. Panel D shows the assignments recovered by clustering PCA coordinates. Colors in D identify its own groups.

## What the map loses

Even with the primary projection, a mean of 91.01% of the original ten neighbors fell outside the ten closest displayed points. The neighbor list exposes those relationships directly. The directional primary hypothesis was supported under the preregistered criterion.

The mean recall under shuffled paper positions was 0.01% across the same seeds.

# Stability, limitations and reproducibility

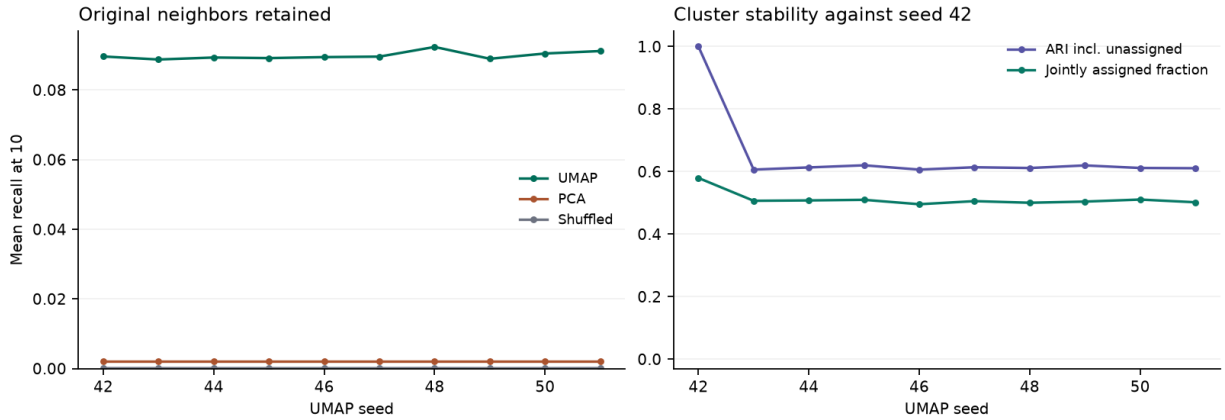


Figure 4. Each marker is one full-corpus run. The two stability series measure agreement with seed 42 and the fraction jointly assigned. Seed 42 compares with itself.

Across the other seeds, adjusted Rand agreement including unassigned points ranged from 0.606 to 0.620. The original and permuted subject agreement estimates pool within-cluster pairs, giving larger groups more weight. Cross-listed subjects are a coarse, incomplete external reference.

## Example topics

Large language models for recommendation (1349 papers). 2504.11889 · 2606.16703

LLM Agents for Software Engineering (1010 papers). 2401.07339 · 2505.02133

Vision-language models for autonomous driving (935 papers). 2510.15446 · 2506.00819

## Scope and limitations

The census represents public feed membership during the recorded window. Likes condition selection on platform attention. Older papers, topic popularity and exposure can affect inclusion. The analysis is descriptive and transductive. The interval quantifies variation from projection seeds conditional on these inputs. It does not estimate uncertainty across future papers. Hosted model revisions may alter a fresh run, so the exact vectors and returned labels accompany the release.

## Reproduction and review

Usage receipts priced at standard rates total USD 2.3227. Reservations for all physical attempts total USD 8.7512 against the USD 50.00 limit. Cost breakdown

Restore the dated snapshot and run the offline build to reproduce tables, figures and the website. The archive includes checksums, paper metadata, embeddings, coordinates, supporting label IDs and per-seed scores. Dataset and numerical artifacts · Code and instructions · Run matrix

## AI statement

A human provided the exact specification. Codex wrote the codebase and this paper to execute it. Human cleanup and rewriting are in progress.