LINKS

Demo: https://prajwal.is-a.dev/vanai3/bc_ai_hackathon_round_3/

Repo: https://github.com/Toricane/vanai3

TEAM INFORMATION

Team Name: prajwal

Member: Prajwal Prashanth
Email: prajwal028@outlook.com

PROJECT TITLE

BC AI Survey Data Storytelling: Semantic Maps & Roundtable Personas

---

PROJECT DESCRIPTION

I analyzed 1,001 open-ended responses from British Columbians about AI—covering hopes, worries, creative impact, beneficiaries, governance, Indigenous involvement, and future priorities. Raw free text is rich but cognitively dense; scrolling lines of comments or collapsing them into a few bars both lose nuance. I built an interface that first lets people see the "geography" of ideas, then hear those ideas speak through representative voices.

Each distinct response is embedded in high-dimensional semantic space and projected into 3D so thematically similar answers cluster naturally. The semantic map allows rotation, zoom, and hovering to expose authentic wording while conveying structural relationships: consensus hubs, edge views, bridges.

To make aggregate patterns more immediately human, I added a generated "roundtable" layer. For every cluster, I imagine a single persona—a voice that stands in for the many respondents whose ideas align. These personas engage in a concise, turn-based dialogue that surfaces agreements, tensions, trade-offs, and unresolved uncertainty. A final synthesis voice summarizes convergence without flattening disagreement. This "one imagined speaker per thematic group" framing helps stakeholders feel the plurality of viewpoints without reading hundreds of near-duplicates or relying on opaque statistics.

The result is a two-stage storytelling tool: spatial comprehension via clustering plus empathetic comprehension via dialogic narration—grounded strictly in clustered source text, not invented opinion.

TECHNICAL APPROACH & TOOLS

I followed an end-to-end pipeline blending unsupervised structure detection with controlled narrative generation:

1. Cleaning & Aggregation
   Removed empty / placeholder entries; aggregated identical strings while retaining a frequency count (later used for marker sizing and emphasis).

2. Semantic Embeddings
   Generated 3,072-dimension vectors for each unique response using OpenAI's text-embedding-3-large to capture contextual semantics beyond keywords.

3. Dimensionality Reduction
   Applied 3D t-SNE (adaptive perplexity per question size) to create visually navigable semantic coordinates emphasizing local thematic neighborhoods.

4. Automatic Clustering
   Swept K-Means across k=2..30; chose k with highest silhouette score per question, ensuring consistent, data-driven thematic resolution.

5. Visualization Layer
   Built interactive Plotly 3D maps. Marker size scales by cube root of frequency so repeated sentiments stand out without overwhelming rarer perspectives. A deterministic color palette yields stable cluster identities reused later.

6. Persona Framing ("Imagined Representative")
   For each cluster I derive a single "vibe" sentence—tone + stance—cached to avoid re-prompting. That imagined persona conceptually compresses many similar respondents into one speaker at a virtual table.

7. Dialogue Generation
   A constrained LLM prompt produces alternating persona lines plus a synthesis. Prompts instruct grounding in cluster themes only, limiting hallucination.

8. Per-Line TTS Audio
   Deterministic voice assignment (e.g., alloy, echo, fable, onyx…) with style passed via system instructions (not spoken). Each line is rendered to an MP3 and indexed in an audio manifest.

9. Front-End Player
   A lightweight JavaScript module loads the manifest + cluster JSON, renders a legend (color → persona → vibe), synchronizes transcript highlighting, enables speed control, keyboard shortcuts, preloading, and auto-scroll.

Stack: Python (pandas, numpy, scikit-learn), t-SNE, K-Means, silhouette scoring, Plotly, OpenAI embeddings + LLM + TTS, JSON manifests, vanilla JS/CSS. The "imagined representative" mechanism bridges statistical clustering and human narrative while preserving traceability back to grouped source text.