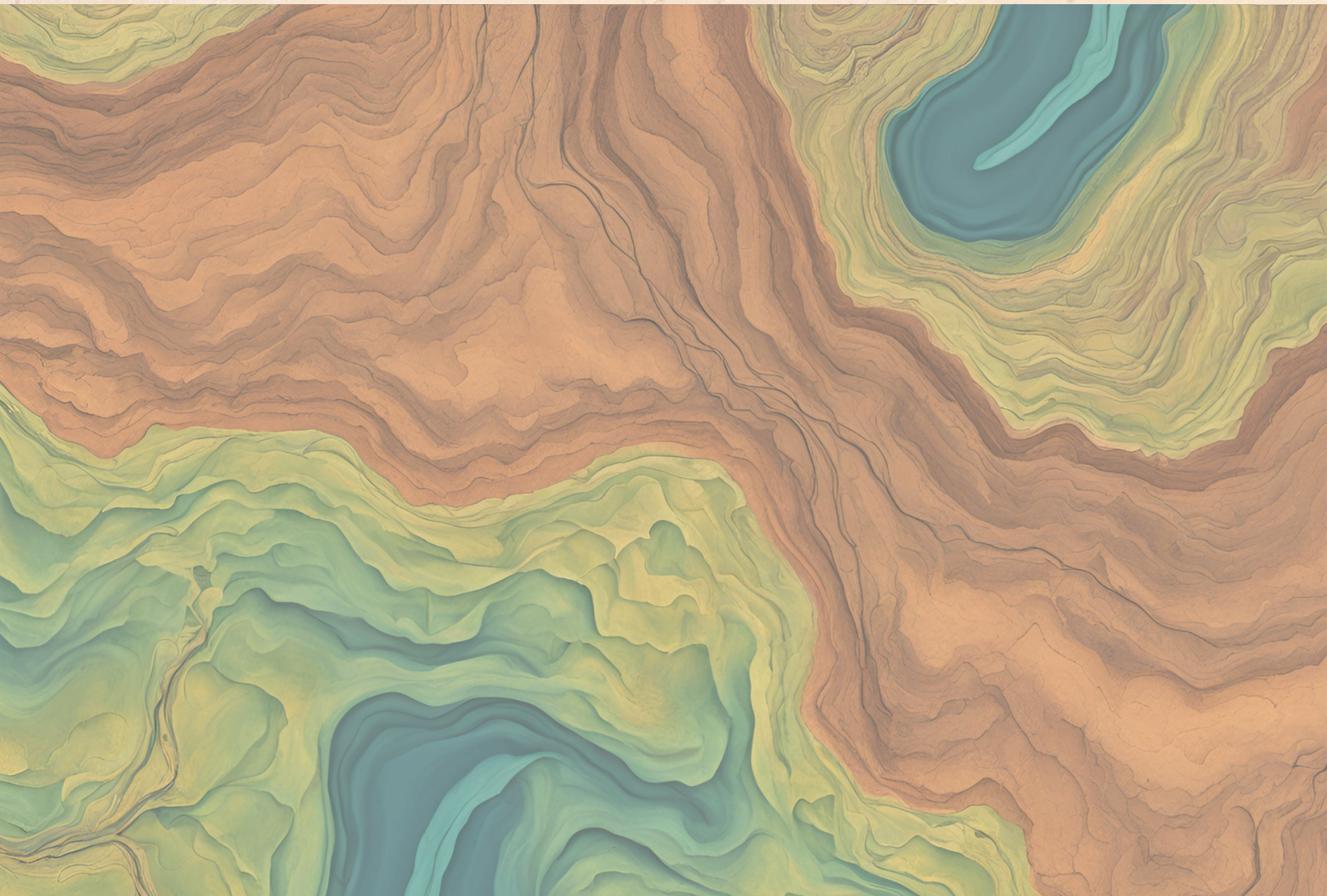
Elembed: a philosophical embedding space

An Arb Report

Oct 2025



Our best machine intelligence relies on *text embeddings*. The casual user of an LLM sees only natural language, but the model itself receives and outputs high-dimensional numerical vectors. These vectors can be thought of as in “meaning-space”. If two words have embeddings that are close in vector space, it is a strong sign that they are conceptually similar.

Embeddings work by representing meanings as a point in *n*-dimensional space, with each dimension typically representing one or more forms of meaning. In principle, this approach is not limited to just words - many abstract concepts could feasibly be represented as a position in *n*-dimensional meaning-space.

The goal of Embedme was to attempt this for philosophical positions. More specifically, the goal would be to embed philosophical frameworks (e.g. deontology, utilitarianism) and let users find out how close they are in *n*-dimensional philosophy space to each framework. The questions I would like users to engage with are:

- “who are my influences?”
- “who are my convergents? Someone I’m similar to but have never read”
- “How many dimensions do you need to really capture not just style but philosophy?”

Existing systems like Exa index far too heavily on shallow features like language, thus spuriously placing highly dissimilar figures like Sartre, Aron, and Sokal close to each other.

I also had a specific hypothesis, which is that the 4000-dimensional models available commercially are inferior to the original 10000-dimensional models (if they were trained with modern methods).

Completed Features

In the 3 months of active work, we implemented:

- Initial questionnaire-based matching approach
- Basic embedding implementation
- Multi-dimensional embedding architecture (content + style)
- Position extraction system with GPT-4
- Style analysis with linguistic fingerprinting
- Basic frontend with file upload
- Results visualization
- Data extraction from philosophical sources
- Parameter optimization experiments

Incomplete Features

- Enhanced premise extraction system (partially implemented)
- Full integration between frontend and backend API
- Methodology dimension analysis (planned but not fully implemented)
- Historical positioning dimension
- Production deployment
- Comprehensive philosopher database

Stack

Backend

- Qwen: gte-Qwen2-7B-instruct
- OpenAI API (embeddings and GPT-4-turbo)
- scikit-learn for similarity calculations
- NLTK & spaCy for linguistic analysis
- LangChain for document processing
- Pandas for data manipulation

Frontend

- React 18.2 + TypeScript
- React Router for navigation
- Vite build system
- Drag-and-drop file upload interface

Approach

We considered training an entirely new embedding model was considered but didn't get to it, mostly owing to cost; even an embedding run is a lot of compute (c. $20k).

We used existing embedding models: Qwen and OpenAI "text-embedding-3-large" and "text-embedding-3-small".

Philosophical positions were sourced from the Stanford Encyclopedia of Philosophy, though theoretically they could also be sourced from their own writings. Initial testing showed that embedding the text as a whole was infeasible for many reasons:

- Other philosophers and frameworks are frequently mentioned;
- Ideas antithetical to the philosophers are constantly discussed to better define their own positions;
- Much of the text is merely structural and doesn't define the position well.

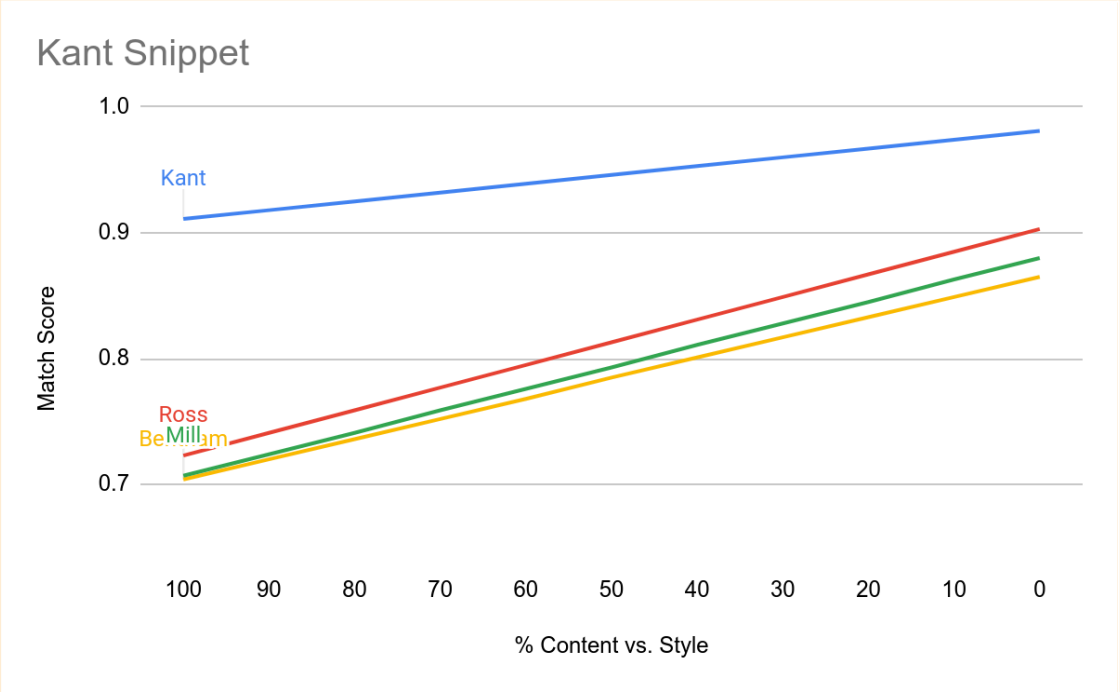
We thus extracted positions with an LLM preprocessing pass. A prompt allow GPT-4 to extract core philosophical positions from the Stanford text, alongside an estimate of how important the position was to the philosopher ("centrality"), and whether they endorsed the position or not.

For quality assurance, this dataset was then run through Dr David Mathers, an Oxford philosopher of mind.

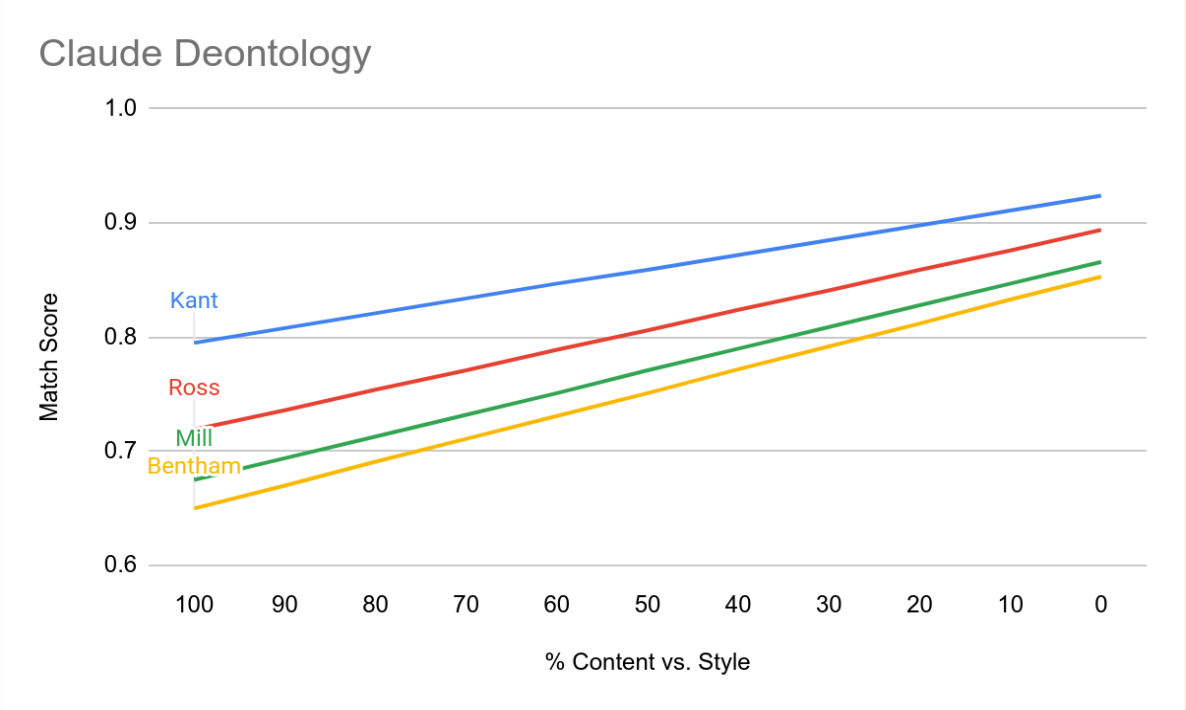
Each position was then embedded using openAI's *text-embedding-3-large* model. As a later development, a "style" embedding for each philosopher was created by concatenating all the philosophers positions, then embedding them into a more compressed space using the smaller *text-embedding-3-small* model.

To evaluate a given input (e.g. another philosopher, or a potential user), the provided position would be embedded in the same manner as the philosophers. The resulting embeddings would be compared to each previously embedded philosopher using cosine similarity, with the results ranked and stored.

Some example results can be found here. In general, philosophers matched well with themselves (which was expected), though the scores were often extremely close and had to be heavily normalised to reveal differences.



A snippet from Kant's writings being compared to classic consequentialists (Ross, Mill and Bentham). The x-axis denotes the mix of content-embedding to style-embedding used, with 100% = fully content matched and 0% = fully style matched



The same experiment with the snippet replaced with an LLM-derived definition of deontology.

Challenges

- Philosophers often change their views over time e.g. early Wittgenstein is completely distinct from late Wittgenstein
- It is unclear whether the style embeddings have a positive effect, despite prior literature suggesting they do
- Nuance in text may not be carried through to embeddings e.g. describing how something is vs. how it should be
- Philosophers such as Foucault, Derrida, or Deleuze resist traditional categorisation, and must be treated carefully

Next Steps

The project's premise appears sound, with the embeddings clearly able to contain information on the closeness of philosophical views. Further steps would include:

- Establishing a larger philosopher database and testing data, to better establish matching capabilities and identify potential hurdles;
- Revaluating the style embeddings, as the preliminary data suggests that they may be hurting rather than helping the matching;

- Fully completing the backend code and linking it to the frontend to allow a website to be hosted;
- Implement methods to attempt to explain why two positions are close in embedding-space;
- Add a user-feedback feature for additional refining once live.

Collaborators

Lydia Farnham, engineer
Dr David Mathers, Oxford philosopher

Contact Information

Website:

https://arbresearch.com

Email:

hi@arbresearch.com

LinkedIn:

https://www.linkedin.com/company/arb-research/