BDS MPhil: Reverse Engineering GPT

Finn-Ole Höner; Supervisors: Bas Donkers, Dennis Fok

May 8, 2024

Setup

Can consumer preferences inform LLM text generation?

- **Application:** Generating advertising claims with an LLM, based on consumer preferences.
- **Methodology:** Find input-embeddings for GPT, that generate a specific advertising claim (we call these "summary embeddings"). We model these "summary embeddings" with an Autoencoder, which gives us a "generation space" for advertising claims.
- **Toy-Example:** Tangible (e.g. "50% more visible shine after one use") and intangible (e.g. "Rediscover vibrant, joyful hair") advertising claims.

Idea

We try to find an input-embedding that makes an LLM generate a specific output-sequence: Find the summary embedding $\boldsymbol{e}^*$ that maximizes the likelihood of the target-sequence $t_1, \ldots, t_L$ given the summary embedding $\boldsymbol{e}$:

$$\boldsymbol{e}^* = \arg\max_{\boldsymbol{e}} \sum_{i=1}^{L} \log p(t_1, \ldots, t_L | \boldsymbol{e}).$$

We impose restrictions on $\boldsymbol{e}^*$ by using an Autoencoder (AE) to model its elements. The hidden layer of this AE is our "generation space".

Current State

- Summary embeddings generate target sequences.
- Separation of tangible and intangible claims in generation space (Figure 1).
- Grid-search exploration reveals "candidates" for new claims (red crosses) and "islands" regenerating training data claims (color-coded circles) (Figure 2).

Next Steps:

- Optimize optimization process.
- Enhance generation mode for clearer new claims.
- Integrate consumer preferences/ratings for generated claims.

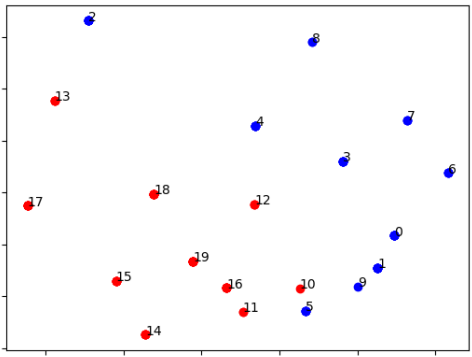

Figure 1: Tangible and intangible claims in 2D-generation space.

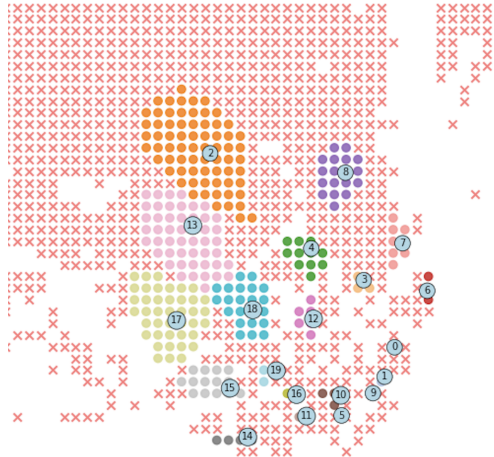

Figure 2: Grid-search across 2D-generation space.