developments:
13 January 2026-20 January 2026

scope

- **Given:**
  a .xlsx file in which each row stores three pieces of data about a patient: a clinician report, an LLM-generated Alzheimer's disease progression score, and the LLM's reasoning for the score

  example (not real data… made up for the purpose of explanation only):

89 year old female patient The patient has faced memory loss for the past three years. Alzheimer's disease biomarker observed. Severe volume loss and hippocampal atrophy. - Dr. John Doe at Big Brain Hospital Neuroscience Department	7	- history of memory loss - Alzheimer's biomarker - atrophy

- The **goal** is to produce a numerical metric for how accurate the LLM-generated score is by analyzing how closely the LLM's reasoning and the real clinician's report align.

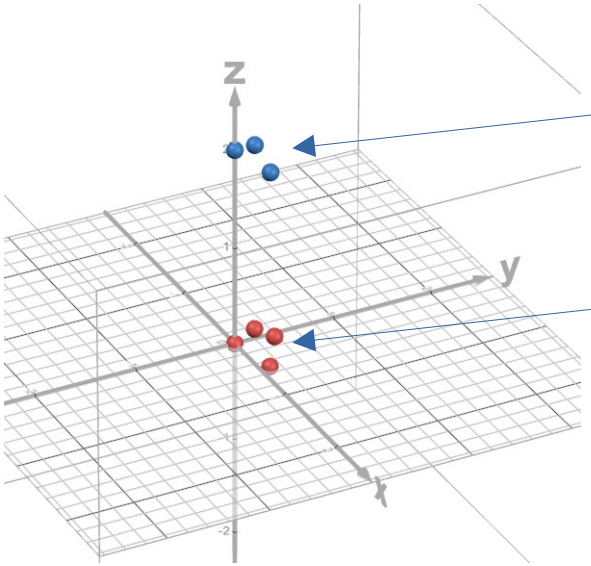
BERTScore

- **BERTScore** is a metric that measures the similarity between two different pieces of texts (i.e., an LLM-generated piece of text and a base truth piece of text).
- How is a BERTScore metric generated?

1) Both the LLM-generated text and the base truth text are **tokenized** (broken up into granular units of text like words).
2) Then, an **embedding** is generated for each token in both pieces of text. An embedding is a vector representation of a token.
3) To determine similarity, comparison of cosines (i.e., cosine similarity) of the embeddings (which are vectors) is performed.

BERTScore: Embeddings

- The process of embedding maps a token (like a word or any other basic text unit) to a vector of floating-point numbers.
  - The idea is that semantics will be embodied because vectors for tokens of similar meanings should be relatively close to each other. Vectors for tokens of say, the opposite meaning will be far away.



Words like “healthy,” and “asymptomatic”

Words like “atrophy,” and “loss,” and “bad”

- BERT, a language model by Google, is used to generate embeddings from tokens.

BERTScore: Cosine Similarity

- The cosine similarity is computed for the embedding of each token of the LLM-generated text with the embedding of each token of the base truth text.
  - Recall that embeddings are vectors of floating-point numbers. The angle between any two vectors can be used to quantify how “similar” or “close” they are. In particular, the cosine of this angle is being studied. Greedy matching is used to select the highest cosine similarity score for each token.
  - The highest cosine similarity scores are then used to compute metrics like BERTRecall and BERTPrecision.

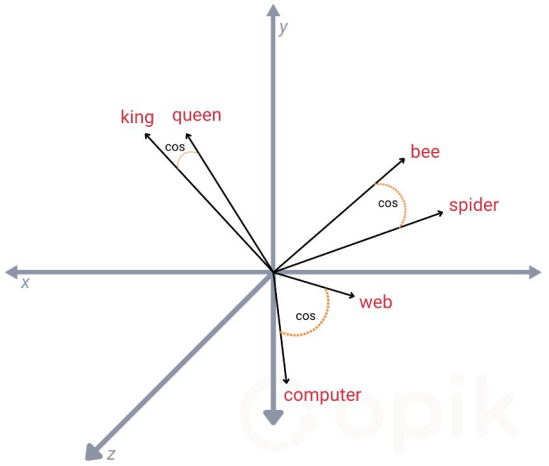


image courtesy of Opik (https://www.comet.com/site/wp-content/uploads/2024/12/Screenshot-2024-12-17-at-8.10.34%E2%80%AFPM.png)

BERTScore: BERTPrecision

- “quantifies how much of the candidate's content is semantically meaningful relative to the reference” – Abby Morgan (Comet)

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

image courtesy of Comet: https://www.comet.com/site/wp-content/uploads/2024/12/Screenshot-2024-12-14-at-8.05.06%E2%80%AFPM-1024x443.png

BERTScore Literature

- Liang, D. (2024, July 30). Intro — Getting Started with Text Embeddings: Using BERT. Medium. https://medium.com/@davidlfliang/intro-getting-started-with-text-embeddings-using-bert-9f8c3b98dee6
- Morgan, A. (2024, December 19). BERTScore For LLM Evaluation. Comet. https://www.comet.com/site/blog/bertscore-for-llm-evaluation/