### **✅ Algorithms Used:**

| **Component** | **Algorithm / Method** | **Description** |
| --- | --- | --- |
| **Embedding Extraction** | BERT [CLS] token representation | Each sentence is converted to a dense vector using a pretrained transformer. |
| **Concept Vector Computation** | **Averaging of embeddings** | Multiple examples per concept are encoded and averaged to form the “concept centroid.” |
| **Similarity Measurement** | **Cosine Similarity** (sklearn.metrics.pairwise.cosine\_similarity) | Measures how similar a new input’s embedding is to each concept centroid. |
| **Explanation Scoring** | Ranking based on similarity | Higher the similarity, more relevant the concept is to the input. |
| **Visualization** | Bar chart (matplotlib) | Plots scores for interpretability. |

### **🧮 Algorithmic Flow:**

Input text → Tokenize → BERT [CLS] Embedding

↓

Cosine Similarity with Concept Vectors

↓

Top Concepts + Scores → Visualized

## **⚙️ 2. Mechanistic Interpretability Module**

This module traces **internal transformer attention patterns**.

### **✅ Algorithms Used:**

| **Component** | **Algorithm / Method** | **Description** |
| --- | --- | --- |
| **Tokenization & Forward Pass** | TransformerLens hooks | Tokenizes input and runs GPT-2 through a forward pass, caching all intermediate states. |
| **Attention Extraction** | Attention pattern from cache (hook\_pattern) | For a given layer and head, extracts the attention matrix (how each token attends to others). |
| **Matrix Visualization** | Heatmap (seaborn.heatmap) | Visualizes attention weights between tokens. |

### **🧮 Algorithmic Flow:**

Input prompt → Tokenized → GPT-2 forward pass

↓

Attention Matrix per [Layer, Head] (from cache)

↓

Heatmap visualization of token-to-token focus

## **🧾 Summary of Algorithms Used (Total)**

| **Algorithm / Tool** | **Purpose** |
| --- | --- |
| Attention Visualization | Mechanistic analysis |
| BERT Embedding | Semantic vector representation |
| Cosine Similarity | Concept matching |
| Mean Vector Aggregation | Concept centroid creation |
| Seaborn Heatmap | Visual explanation |
| TransformerLens Caching | Access to internal model states |