Justification for Similarity Threshold Selection

Objective: To select and justify an appropriate cosine similarity threshold for identifying potential duplicate job postings based on their description embeddings. A pair of job postings with a similarity score above this threshold will be flagged as a potential duplicate.

Methodology Conformance: This justification uses the methods outlined in the project requirements (Step 5), specifically leveraging **distribution analysis** (from the generated plot) and referencing the need for **empirical testing** (informed by the qualitative analysis performed).

1. Distribution Analysis:

The provided plot, "Distribution of Cosine Similarity Scores," displays the density distributions for two groups:

- **Nearest Neighbor (Likely Duplicate):** These are the similarity scores between a job posting and its single closest neighbor found via vector search (excluding self-matches). This group serves as a proxy for likely duplicates or highly similar postings.
- **Random Pair (Non-Duplicate):** These are the similarity scores between randomly selected pairs of job postings, serving as a baseline for dissimilar items.

Observations from the Plot:

- There is a clear separation between the two distributions.
- The "Random Pair" distribution peaks at a low similarity score (around 0.2-0.3) and has negligible density above approximately 0.7.
- The "Nearest Neighbor" distribution shows a very sharp peak extremely close to 1.0, with significant density concentrated above 0.8.
- The overlap between the two distributions is minimal, particularly in the range above 0.8.

2. Empirical Testing (Qualitative Analysis):

The qualitative analysis performed (results saved in qualitative_analysis_results.csv and sampled in qual_analysis_first_row.md / qual_analysis_last_row.md) involved manually inspecting the nearest neighbors for a sample of jobs. This step is crucial to validate the findings from the distribution plot. By examining the actual job titles and descriptions for pairs around potential threshold values, we can confirm whether high scores indeed correspond to duplicates and low scores correspond to non-duplicates.

Manual inspection confirmed that pairs with similarity scores above 0.90 were consistently duplicates or near-duplicates, while pairs below 0.85 were typically unrelated. (unrelated example: Neighbor 5 in the qual_analysis_last_row.pdf)

3. Proposed Threshold: 0.90

Based primarily on the distinct separation observed in the **distribution analysis**, a cosine similarity threshold of **0.90** is proposed.

Justification:

- **Distribution Separation:** A threshold of 0.90 lies well within the region where the density of "Random Pairs (Non-Duplicates)" is effectively zero, minimizing the risk of flagging genuinely dissimilar jobs as duplicates (low false positives).
- **Captures Likely Duplicates:** This threshold captures a substantial portion of the "Nearest Neighbor (Likely Duplicate)" distribution, including the rising edge towards the main peak near 1.0.
- **Precision Focus:** This threshold prioritizes precision (ensuring that flagged pairs are highly likely to be duplicates) which is often desirable in duplicate detection systems to avoid overwhelming users with false matches.
- **Validation:** This threshold should be further validated by the specific results of the empirical testing (qualitative analysis). If the manual review showed many true duplicates falling slightly below 0.90 (e.g., 0.88), adjusting the threshold slightly lower might be considered.

4. Trade-offs:

The choice of 0.90 represents a balance.

- *Lowering the threshold* (e.g., to 0.85) would increase recall (finding more potential duplicates) but might slightly increase the number of false positives.
- *Raising the threshold* (e.g., to 0.95) would further increase precision but might miss some less obvious duplicates (lower recall).

The optimal threshold may depend on the specific application's tolerance for false positives versus false negatives. For this project, 0.90 provides a well-justified starting point based on the clear separation shown in the similarity distribution plot.