Quick notes

# Overall Workflow

Read URLs file

Check if new URLs need to be processed

Read/parse webpages; save json

Extract requirements; save json

Iter 0:

Flatten requirements

Flatten resume; save json

Calculate sim scores; save csv

Prune and save json

Iter 1:

Modify requirements; save flat json

Calculate sim scores; save csv

Prune and save json

Iter 2:

Modify requirements; save flat json

Calculate sim scores; save csv files

Prune and save each json

# Prompting Tricks

Use JSON or not?

(from GPT-o1, the newest model) From Using JSON format in prompts can indeed help guide language models like GPT to produce structured outputs that are easier to parse. Including a clear JSON schema or example in the prompt can encourage the model to respond in the desired format.

However, in the specific case of your prompt, embedding the tasks and instructions themselves within a JSON-like structure can potentially confuse the model. This is because the model might interpret the entire JSON structure as something to process or reproduce, rather than focusing on the specific tasks and the expected output.

Here's why including JSON in the prompt can be both helpful and tricky:

* Helpful: Providing a clear example of the desired JSON output format can guide the model to produce responses that are easier to parse programmatically.
* Tricky: If the prompt includes complex JSON structures for instructions or tasks, the model might misinterpret the instructions or include unwanted elements in the output.

Recommendations:

* Use JSON Format for Output Examples, Not Instructions:
* Keep your instructions in plain text for clarity.
* Provide the desired output format as a JSON example.
* Present tasks and instructions in bullet points or numbered lists.
* Use clear and direct language.
* Include a Clear JSON Schema or Example for the Expected Output:
* Show the exact JSON structure you expect in the response.
* Emphasize that the model should follow this format strictly.

# Similarity Scores and Related Metrics

## Directory Organization

Suggested by GPT

src/

└── evaluation\_optimization/

├── \_\_init\_\_.py # Makes this directory a package

├── metrics/

│ ├── \_\_init\_\_.py # Exposes metric calculation functions/classes

│ ├── metric\_definitions.py # Defines various metrics

│ ├── metric\_calculator.py # Classes/functions for finding and calculating metrics

├── indices/

│ ├── \_\_init\_\_.py # Exposes index calculation functions/classes

│ ├── multivariate\_indexer.py # Class for handling multivariate index creation

├── changes/

│ ├── \_\_init\_\_.py # Exposes change calculation functions/classes

│ ├── change\_calculator.py # Methods for calculating changes (absolute/percentage)

├── analysis/

│ ├── \_\_init\_\_.py # Exposes analysis-related methods

│ ├── descriptive\_stats.py # Descriptive statistics calculation

│ ├── plotting.py # Plotting functions for quick visualization

└── utils/

├── \_\_init\_\_.py # Utility functions

├── data\_merger.py # Merge DataFrames across iterations

├── data\_loader.py # Functions for loading data

I will not go this detailed - I don’t have that much code.

## Current Similarity Scores

Based on the provided similarity scores for the responsibilities against the job requirements, here's a quick summary and analysis:

### Summary of Similarity Scores:

1. \*\*Similarity Metrics Used:\*\*

- Several metrics have been used to measure the similarity between each responsibility and the reference job requirements text:

- \*\*Self Attention Similarity (`self\_attention\_similarity`)\*\*

- \*\*Layer-wise Attention Similarity (`layer\_wise\_attention\_similarity`)\*\*

- \*\*Self Hidden State Similarity (`self\_hidden\_state\_similarity`)\*\*

- \*\*CLS Embedding Similarity (`cls\_embedding\_similarity`)\*\*

- \*\*SBERT Similarity (`sbert\_similarity`)\*\*

- \*\*STS Similarity (`sts\_similarity`)\*\*

- An \*\*Average\*\* score is provided to give an overall similarity measure for each responsibility.

2. \*\*Highest Similarity Scores:\*\*

- The responsibility \*\*"Co-authored an industry-recognized report on M&A in the engineering services sector..."\*\* has the highest \*\*Average similarity score (0.390)\*\*. It also scores highly across multiple metrics, especially \*\*`cls\_embedding\_similarity (0.810)`\*\* and \*\*`sts\_similarity (0.663)`\*\*. This indicates that this responsibility has the strongest alignment with the job requirements.

3. \*\*Moderate Similarity Scores:\*\*

- The responsibilities \*\*"Provided strategic insights to a major global IT vendor..."\*\* and \*\*"Assisted a U.S.-based international services provider in its growth strategy..."\*\* both have moderate \*\*Average similarity scores (0.313)\*\*. These scores are boosted by relatively high \*\*`cls\_embedding\_similarity`\*\* values (0.808 and 0.728, respectively), indicating that while these responsibilities have some relevance, their overall alignment is moderate.

4. \*\*Lower Similarity Scores:\*\*

- The responsibilities \*\*"Achieved over 40% centralization of tasks to an offshore team..."\*\* and \*\*"Developed Python tools to automate and accelerate internal processes..."\*\* have the lowest \*\*Average similarity scores (0.268 and 0.266, respectively)\*\*. Their lower scores are reflected across all metrics, with relatively low \*\*`self\_attention\_similarity`\*\*, \*\*`layer\_wise\_attention\_similarity`\*\*, and \*\*`self\_hidden\_state\_similarity`\*\* values. This suggests these responsibilities are less aligned with the job requirements.

5. \*\*Key Observations:\*\*

- \*\*`cls\_embedding\_similarity`\*\* tends to have the highest values across all responsibilities, indicating that the \*\*CLS token embedding\*\* similarity (a standard method of representing sentence-level embeddings in BERT models) captures a significant amount of the semantic alignment.

- \*\*`sbert\_similarity`\*\* and \*\*`sts\_similarity`\*\* are also relatively high for certain responsibilities, indicating semantic similarity from sentence embeddings and sentence similarity models.

- The metrics that seem to contribute most to the \*\*Average\*\* similarity score are \*\*`cls\_embedding\_similarity`\*\*, \*\*`sbert\_similarity`\*\*, and \*\*`sts\_similarity`\*\*, while the \*\*attention-based similarities\*\* (`self\_attention\_similarity`, `layer\_wise\_attention\_similarity`, `self\_hidden\_state\_similarity`) tend to be much lower.

### Analysis:

1. \*\*Responsibility Alignment with Job Requirements:\*\*

- The most aligned responsibility appears to be related to strategic insights and market analysis, which matches closely with several job requirements such as "performing market analysis and developing competitive intelligence" and "management consulting, product management, and strategy."

- Less aligned responsibilities, such as "Developed Python tools..." or "Achieved over 40% centralization of tasks...", may not directly match the core requirements like "knowledge of AI and ML" or "ability to form and refine hypotheses."

2. \*\*Diverse Requirement Coverage:\*\*

- Responsibilities that involve \*\*strategic decision-making, market analysis, and stakeholder management\*\* have generally higher similarity scores, aligning well with the provided job requirements which emphasize \*\*consulting, management, strategy, and technical expertise.\*\*

3. \*\*Potential Gaps:\*\*

- Technical responsibilities such as developing Python tools or managing offshore teams show weaker alignment, suggesting they may not be the primary focus of the job requirements provided, which focus more on strategic, analytical, and consulting skills.

### Conclusion:

- The provided similarity scores effectively highlight which responsibilities align most closely with the job requirements and which are less relevant. This kind of analysis helps identify the most valuable experiences on a resume that match the expectations for a specific role.

If you need further detailed insights or additional analyses, feel free to ask!

## Better Approaches for Asymmetric Matching

For cases like this, where alignment of responsibilities to requirements matters more in one direction, consider using asymmetric similarity or matching techniques:

### Soft Cosine Measure (SCM):

How It Works: Soft Cosine Measure takes into account the similarity between words that are not identical but related. For example, it can measure the similarity between "managing" and "leading" responsibilities.

Applicability: This can be more effective than cosine similarity because it considers the semantic closeness of words across the entire text. However, it would still be limited in representing the true asymmetric relationship you're looking for.

### Word Mover's Distance (WMD):

How It Works: WMD computes the minimal distance that words from one text need to travel to match the words in another text. It accounts for the semantic distance between words using pre-trained embeddings.

Applicability: WMD can capture the cost of aligning each word in "responsibilities" to a word in "requirements," potentially reflecting the asymmetry more naturally. However, this approach could still miss specific context, such as the order or relationship among responsibilities.

### Jaccard or Containment-Based Similarity Measures:

How It Works: You can calculate the containment ratio—the proportion of overlapping elements in the "responsibilities" found in "requirements." This is asymmetric since it measures how well one set (responsibilities) is contained in another set (requirements).

Applicability: This is more directly relevant when you want to assess how well one text (responsibilities) fits into another (requirements). However, it requires tokenization and does not inherently handle deep semantic meanings.

### Textual Entailment or Alignment Models:

How It Works: Models designed for textual entailment assess whether one text (hypothesis) logically follows from another text (premise). Here, "responsibilities" would be the hypothesis and "requirements" the premise.

Applicability: This could be a robust approach to verify if all responsibilities are entailed by the requirements, thereby capturing an asymmetric relationship.

### Custom Asymmetric Metric:

How It Works: Create a custom metric that calculates the coverage of terms or concepts from "responsibilities" in "requirements."

This could involve:

* Tokenizing both texts.
* Calculating the number of tokens or concepts in "responsibilities" that are covered by "requirements."
* Measuring how much of the "responsibilities" content is present in "requirements" (as a percentage or ratio).
* Applicability: This can be specifically tailored to ensure full coverage and provide a more meaningful measure of alignment in your context.

## Suggested Scoring Framework for Asymmetric Relationship (By ChatGPT)

Here’s how you could assign asymmetric similarity scores:

### Direct Matching Score (DMS):

Calculate the proportion of key terms or concepts in the requirement that are covered by the responsibility.

Example: If a requirement has 5 key terms and the responsibility covers 4 of them, the score could be 4/5 = 0.80.

### Partial Coverage Score (PCS):

Go beyond just term matching by considering synonyms and related concepts. This involves using a semantic similarity model to determine if terms in the requirement are represented in the responsibility, even if not verbatim.

Weighting can be added to different terms in the requirement based on their importance. For example, "project management" might be more critical than "basic programming skills."

### Entailment Probability Score (EPS):

Use a textual entailment model to predict the likelihood that the content of the requirement is entailed by the responsibility. The output probability can serve as the similarity score.

This is especially useful for nuanced requirements that are not straightforward or binary but involve complex conditions or criteria.

### Semantic Coverage Ratio (SCR):

For each requirement, compute the semantic overlap with the responsibility using a method like the Soft Cosine Similarity which accounts for semantic similarity between words.

This ratio could be defined as the sum of the similarities of matching terms divided by the total number of unique terms in the requirement.

## New Custom Similarity Scores to Consider for the Project

BertScore Precision, Soft Similarity, Word Mover's Distance, NLI models, Jaccard Similarity

### Scores to consider:

Asymmetrical:

* BertScore Precision (semantic coverage)
* Soft Similarity (semantic coverage): Soft Similarity (e.g., Soft Cosine Similarity) considers the similarity between words that are not identical but are semantically related. It enhances semantic coverage by recognizing related but not identical terms.
* Word Mover's Distance (WMD) (for a mix of semantic and partial coverage)
* NLI models (for entailment)

Symmetrical:

* Jaccard Similarity (partial coverage): Calculates the size of the intersection divided by the size of the union of two sets. It's suitable for text data, especially when dealing with keywords or phrases.

### Comprehensive Framework for Asymmetric Text Similarity Analysis:

1. BERTScore Precision:

* Focus: Measures how well the semantic content of one text (e.g., a responsibility) aligns with another text (e.g., a requirement), focusing on token-level precision using contextual embeddings.
* Use Case: Effective for capturing semantic coverage by identifying how much of the requirement's meaning is present in the responsibility. It is asymmetric in nature because it focuses on how much of the requirement is covered by the responsibility.
* Strength: Captures semantic nuances and is robust to word order changes, making it effective for understanding the alignment of texts where semantic similarity is critical.

1. Soft Similarity:

* Focus: Computes similarity by considering semantically related terms, not just exact matches. It accounts for the similarity between words that are related but not identical.
* Use Case: Enhances semantic coverage by recognizing partial matches, synonyms, and related terms. It is useful when a requirement can be met in different ways that involve related but not identical terminology.
* Strength: Provides more flexibility than strict semantic measures like BERTScore Precision by allowing for partial matches and relatedness.

1. Word Mover's Distance (WMD):

* Focus: Measures the minimum cumulative distance that words from one text need to "travel" in the embedding space to match the words in another text, capturing both semantic similarity and partial coverage.
* Use Case: Useful for a mix of semantic alignment and content alignment when there are significant overlaps or partial matches between texts.
* Strength: Provides a nuanced similarity measure that accounts for both semantic distance and partial coverage, capturing the degree to which one text can be transformed into another.

1. NLI Models (e.g., BERT-NLI, RoBERTa-NLI, DeBERTa-NLI):

* Focus: Predicts whether one text (e.g., a responsibility) entails, contradicts, or is neutral to another text (e.g., a requirement). NLI models are specifically designed to handle asymmetric entailment tasks.
* Use Case: Ideal for detecting whether a responsibility logically entails a requirement or supports it directionally. This is crucial for asymmetric relationships where you want to know if a responsibility fulfills a requirement without assuming the reverse.
* Strength: Provides a binary or probabilistic output for entailment, making it ideal for fine-tuning decisions where logical entailment is required.

1. Jaccard Similarity:

* Focus: Measures the exact overlap between sets of terms from two texts, focusing on partial coverage. It calculates the ratio of the intersection of terms to the union.
* Use Case: Useful for cases where exact term overlap is important or where a requirement must be explicitly mentioned in the responsibility.
* Strength: Simple, intuitive, and provides a clear measure of overlap. It is most effective when the goal is to quantify how much of the content from one text is present in another, especially in more rigid matching scenarios.

### What Ranges Are Considered High or Low

### a) Interpreting High vs. Low Scores for Each Metric:

1. \*\*BERTScore Precision:\*\*

- \*\*Range:\*\* 0 to 1

- \*\*High Scores:\*\* Generally above 0.85. A high score indicates that the tokens in the candidate text (e.g., responsibility) are well-aligned with the reference text (e.g., requirement) in terms of semantic similarity.

- \*\*Low Scores:\*\* Below 0.70. A low score suggests that the candidate and reference texts do not share much semantic similarity.

- \*\*Typical Use:\*\* Used for evaluating precision when the goal is to see how much of the candidate content is relevant to the reference.

2. \*\*Soft Similarity (SBERT Cosine Similarity):\*\*

- \*\*Range:\*\* -1 to 1

- \*\*High Scores:\*\* Above 0.7. High scores indicate that the embeddings of the two texts are close, meaning they are semantically similar.

- \*\*Low Scores:\*\* Below 0.4. Low scores suggest that the texts are semantically different.

- \*\*Typical Use:\*\* Measures overall semantic similarity, often used in tasks like paraphrase detection and information retrieval.

3. \*\*Word Mover's Distance (WMD):\*\*

- \*\*Range:\*\* 0 to infinity (practically within 0 to ~20 for most text)

- \*\*High Scores:\*\* (i.e., smaller distance) Below 5 is considered a good match. A smaller distance means that the texts are closer semantically.

- \*\*Low Scores:\*\* Above 15 is considered a poor match. A larger distance indicates that the texts are semantically different.

- \*\*Typical Use:\*\* Used in text retrieval and text similarity tasks where the goal is to measure how far the meaning of one text is from another.

4. \*\*NLI Entailment Score:\*\*

- \*\*Range:\*\* 0 to 1

- \*\*High Scores:\*\* Above 0.7. Indicates a strong likelihood that the hypothesis (responsibility) is entailed by the premise (requirement).

- \*\*Low Scores:\*\* Below 0.3. Suggests that the hypothesis is not supported or entailed by the premise.

- \*\*Typical Use:\*\* Captures directional semantic entailment; helpful in tasks like textual entailment and natural language understanding.

5. \*\*Jaccard Similarity:\*\*

- \*\*Range:\*\* 0 to 1

- \*\*High Scores:\*\* Above 0.5. Indicates a significant overlap in the tokens or words between the texts.

- \*\*Low Scores:\*\* Below 0.2. Suggests minimal overlap between the texts.

- \*\*Typical Use:\*\* Measures lexical similarity based on shared words or tokens. It does not capture semantic similarity as effectively as other methods.

### b) Variations in Scores and Applicability in ML Models:

To make these scores "applicable" in ML models, variations in scores should capture meaningful differences in relevance, similarity, or entailment between text pairs. Here’s how variations in each score type can be leveraged:

1. \*\*BERTScore Precision:\*\*

- \*\*Small Variations (e.g., 0.85 vs. 0.88):\*\* May indicate minor differences in precision; not always significant. In ML models, such variations can help fine-tune thresholds for classification.

- \*\*Large Variations (e.g., 0.65 vs. 0.90):\*\* Indicative of substantial differences in text relevance. Useful for ranking, filtering, or classification tasks.

2. \*\*Soft Similarity:\*\*

- \*\*Small Variations (e.g., 0.65 vs. 0.68):\*\* Generally not impactful in ML models. Variations need to be more pronounced (e.g., 0.65 vs. 0.80) to influence model decisions.

- \*\*Large Variations (e.g., 0.35 vs. 0.80):\*\* Strongly suggest semantic differences. Good for tasks like paraphrase detection or semantic search.

3. \*\*Word Mover's Distance (WMD):\*\*

- \*\*Small Variations (e.g., 7 vs. 8):\*\* Generally minor; a model might treat them similarly.

- \*\*Large Variations (e.g., 3 vs. 15):\*\* Indicate a strong difference in semantic content. Useful for clustering, anomaly detection, and retrieval tasks where semantic distance matters.

4. \*\*NLI Entailment Score:\*\*

- \*\*Small Variations (e.g., 0.6 vs. 0.65):\*\* May not indicate meaningful entailment differences; could be treated similarly.

- \*\*Large Variations (e.g., 0.2 vs. 0.8):\*\* Clearly distinguishes between entailment and non-entailment. Crucial for entailment-based classification tasks or filtering relevant text.

5. \*\*Jaccard Similarity:\*\*

- \*\*Small Variations (e.g., 0.15 vs. 0.20):\*\* Usually not significant unless coupled with other features.

- \*\*Large Variations (e.g., 0.10 vs. 0.60):\*\* Clear indication of token overlap. Effective for text matching tasks where token presence is key (e.g., keyword matching).

### Combining Scores in ML Models:

- \*\*Feature Engineering:\*\* Combining these metrics as features in an ML model can provide a robust representation of text similarity. Each metric captures different aspects of text similarity (e.g., semantic, lexical, entailment).

- \*\*Thresholding:\*\* Models can be trained to learn optimal thresholds for each metric to determine relevance or similarity.

- \*\*Weighted Averaging:\*\* Some applications may benefit from weighted combinations of these metrics to emphasize certain types of similarity (e.g., semantic over lexical).

- \*\*Ensemble Models:\*\* Use ensemble methods (e.g., stacking, bagging) to leverage multiple similarity metrics to make more accurate predictions.

By understanding how each score works and the implications of their variations, you can design more effective text similarity models tailored to specific applications in NLP.

BERTScore Precision:

Low Scores: If the BERTScore Precision is low (e.g., around 0.82 or lower), it means there is less overlap between the tokens (words or phrases) in the responsibility text and the tokens in the reference text. In this case, the responsibility might lack key phrases or terms that are explicitly stated in the job requirements.

Implication: Such responsibilities may not directly speak to the required qualifications or experiences. For instance, "Oversaw two Appian implementations that increased team productivity by 20 to 30%" (0.8297) may not directly align with the specifics of AI, machine learning, or the other strategic management and analytical skills required by the job.

Soft Similarity:

Low Scores: A low soft similarity score (e.g., below 0.30) indicates that the semantic meaning of the responsibility diverges from the meaning of the requirements. This metric is more sensitive to the overall context and concepts rather than exact word matches.

Implication: Responsibilities with low soft similarity scores likely do not cover the core topics or themes emphasized in the job description. For example, "First to implement API (application programming interface)" (0.1577) is quite low, suggesting it doesn't capture broader themes like managing cross-functional teams or strategic AI development.

Word Mover's Distance (WMD):

High Scores: Since WMD measures distance (where a lower distance is better), higher values indicate that more "movement" (word conversion or alignment) is needed to match the two texts.

Implication: Responsibilities with higher WMD scores (closer to 15.9) are more distanced from the reference text in terms of the words and their meanings. For example, "Collaborated with the engineering services research team..." (15.91) suggests it is less lexically aligned with the reference compared to others with lower scores.

NLI Entailment Score:

Low Scores: A low entailment score (e.g., below 0.05) implies that the requirement text (premise) is unlikely to be inferred from the responsibility text (hypothesis).

Implication: This might mean that the experience or responsibility described doesn't logically lead to the conclusion that the candidate meets specific requirements. For example, "Developed Python tools to automate and accelerate internal processes..." (0.0025) has a very low entailment score, indicating it may not provide evidence for the ability to "manage executive stakeholders" or "shape the future of AI."

Jaccard Similarity:

Low Scores: Jaccard Similarity scores that are low (e.g., below 0.05) indicate minimal direct word overlap between the texts.

Implication: Low Jaccard scores, such as "First to launch a chatbot" (0.0267), suggest that there are not many shared terms between the responsibility and the requirements, making it less likely that this experience is directly relevant to the role.

### Outputs

#### Responsibility w/t Requirements as a Whole



##### Interpretation (by GPT)

Thank you for providing the full reference text. With this reference, the similarity metrics can be interpreted more specifically in the context of how well the responsibilities align with this detailed set of requirements for a role. Here's a refined interpretation with the given reference in mind:

### Contextual Interpretation:

1. \*\*BERTScore Precision\*\*:

- The BERTScore Precision metric indicates how much the content of the responsibility overlaps with the content of the reference text (the job requirements). Given the reference emphasizes a variety of skills—ranging from technical knowledge, management, market analysis, data handling, problem-solving, and communication—higher scores imply that the responsibility has keywords or phrases that match the job requirements closely.

- For example:

- "First to build an ontology" has a high precision score (0.8695), suggesting it aligns well with the technical and possibly the strategic elements in the reference, such as "knowledge of the Machine Learning and Artificial Intelligence market landscape" and "building assets and programs that surface valuable insights."

- "Led the external software development team to build and implement new tools" also scores highly (0.8562), which aligns with requirements like "Work across Program Management teams" and "manage executive stakeholders and communicate with a highly technical management team."

2. \*\*Soft Similarity\*\*:

- This metric captures the semantic similarity and is sensitive to the meaning of words rather than their exact match. Responsibilities with higher soft similarity scores suggest they capture broader themes or underlying concepts present in the requirements.

- For example: "Championed new technology projects using ML, NLP, chatbot, ontology, web-scraping, API, UX (User Experience)" scores the highest (0.4332). This reflects its strong conceptual overlap with the reference's focus on advanced AI/ML concepts and strategic planning, such as "Help shape the future of AI" and "build assets and programs that surface valuable insights."

3. \*\*Word Mover's Distance (WMD)\*\*:

- WMD values reflect how much transformation is needed to convert the words in the responsibility to those in the reference. Lower WMD indicates a closer match.

- Responsibilities with WMD values closer to 15.4-15.5 are more semantically aligned:

- "Managed three major data integration projects critical to the launch of the new platform" (15.43) and "First to build an ontology" (15.46) have among the lowest WMD values, indicating they contain words and concepts closely related to those in the reference text.

4. \*\*NLI Entailment Score\*\*:

- This metric indicates the likelihood that the "premise" (requirement) can be inferred from the "hypothesis" (responsibility). Higher entailment scores suggest a stronger inferential relationship.

- For instance:

- "First to build an ontology" (0.1715) and "Advised services firms on deal pursuit and sales orchestration strategies" (0.0870) have higher entailment scores. These responsibilities may suggest an ability to "form and refine hypotheses, gather supporting data, and make recommendations," which is a key requirement in the reference text.

5. \*\*Jaccard Similarity\*\*:

- This metric is less influenced by semantic meaning and more by direct word overlap. Responsibilities with a slightly higher Jaccard similarity, like "Collaborated with the engineering services research team to pioneer the engineering services tracker..." (0.0802), suggest there are more common words between the responsibility and the reference text, though it does not guarantee semantic alignment.

Summary

The provided metrics allow us to measure alignment between the responsibilities and a comprehensive set of job requirements that span technical skills, strategic capabilities, and management expertise. Different metrics emphasize different aspects of similarity:

* \*\*BERTScore Precision\*\* and \*\*NLI Entailment Score\*\* are useful for understanding direct and inferential content alignment.
* \*\*Soft Similarity\*\* and \*\*Word Mover's Distance\*\* provide insights into broader semantic relationships.
* \*\*Jaccard Similarity\*\* indicates basic lexical overlap.

By combining these metrics, we can holistically assess how well a given responsibility matches a set of complex requirements, providing a nuanced understanding of alignment beyond just word matching.

Conclusion: Comparing responsibility to the entire requirements is not a good measure.

### Responsibility vs Requirement (Seg by Seg)

A screenshot of a computer screen

Description automatically generated

**Much better results.**

#### Interpretation:

BERTScore Precision:

* The mean score is around 0.83, with a narrow standard deviation of 0.017, indicating that most values are closely clustered around the mean.
* Scores range from 0.78 to 0.88, suggesting a relatively high level of precision in matching between texts.
* A low range and high mean indicate that most text pairs have relatively good semantic precision.

Soft Similarity (SBERT):

* The mean score is 0.26, with a standard deviation of 0.11. This suggests more variability in soft similarity compared to BERTScore Precision.
* Scores range from 0.03 to 0.56, showing that there is a broader range of semantic similarity among the text pairs.
* The distribution suggests that while some pairs are semantically similar, many have low similarity.

Word Mover's Distance (WMD):

* The mean score is around 5.07, with a standard deviation of 0.90. WMD values range from 3.61 to 8.19, which is within a reasonable range for text comparisons.
* Lower WMD scores indicate closer semantic similarity. Since the 25th percentile is 4.36 and the 75th percentile is 5.57, most text pairs fall into a moderate distance range.

DeBERTa Entailment Score:

* The mean score is 0.079, with a high standard deviation of 0.14, indicating a broad variation in entailment probabilities.
* Scores range from 0.0004 to 0.88, suggesting that some pairs have very high entailment, while others have almost none.
* The median score is 0.0277, indicating that over half of the pairs have low entailment, which aligns with the typical usage of entailment in detecting strong directional relationships.

Conclusion:

* BERTScore Precision is consistently high across text pairs, indicating strong semantic overlap or matching.
* Soft Similarity (SBERT) shows more variation, suggesting mixed levels of semantic similarity.
* Word Mover's Distance (WMD) is generally moderate, indicating the semantic distances are neither too close nor too far apart.
* DeBERTa Entailment Score shows considerable variability, highlighting that entailment is detected in only some text pairs, which aligns with the directional nature of this metric.

These statistics look appropriate and can be utilized for various downstream ML tasks, such as classification, ranking, or filtering.

## What to Filter Out

Not all 4 metrics can be low

BERTScore precision should not be low;

# Editing Responsibilities

Temperature

When comparing candidate vs reference text and modifying the candidate, temperature affects the modification process:

Low Temperature (e.g., 0.1-0.3)

1. Minimal changes: Candidate text modifications will be subtle, preserving the original meaning and structure.

2. Localized edits: Changes will focus on specific words or phrases, rather than rephrasing entire sentences.

3. Conservative rewording: Model will prioritize preserving the original text's tone, style, and syntax.

4. Less semantic drift: Candidate text will remain closer to the original meaning.

High Temperature (e.g., 0.7-1.0)

1. More significant changes: Candidate text modifications will be more substantial, potentially altering sentence structure and meaning.

2. Global revisions: Changes may rephrase entire sentences or paragraphs, introducing new ideas.

3. Creative rewording: Model will take more liberties with tone, style, and syntax.

4. Increased semantic drift: Candidate text may deviate further from the original meaning.

Comparison highlights

| Temperature | Modification style | Semantic drift |

| --- | --- | --- |

| Low | Subtle, localized | Minimal |

| High | Significant, global | Increased |

Implications for comparison

1. Low temperature: Suitable for evaluating minor variations, grammatical corrections, or subtle rewording.
2. High temperature: Useful for assessing more significant changes, creative rewrites, or paraphrasing.

When modifying the candidate text, consider:

1. Your specific comparison goals.

2. The reference text's complexity and style.

3. The desired level of semantic preservation.

Adjust the temperature accordingly to achieve the desired modification style.

Would you like more information on fine-tuning temperature or exploring other comparison metrics?

Optimal Temperatures to Set:

* text\_editor.edit\_for\_semantics(temperature=0.5) (if you set this too high, the end sentence till be too long)
* text\_editor.edit\_for\_entailment(temperature=0.6)
* text\_editor.edit\_for\_dp(temperature=0.9)

# Evaluate Results

## Composite Scores

## PCA Scores

Below is the code that incorporates the PCA approach into a method within your existing class. This method is modeled after your `calculate\_composite\_score` method and handles the DataFrame with columns that have indices (e.g., `soft\_similarity\_0`, `soft\_similarity\_1`, etc.). It standardizes the necessary columns, normalizes and reverses the polarity of `word\_movers\_distance`, and applies PCA to reduce the metrics to a single composite score.

```python

def calculate\_pca\_composite\_score(self, df):

"""

Calculate the composite score using PCA based on four metrics.

Parameters:

- df: DataFrame containing the metrics.

Returns:

- df: DataFrame with the composite score added.

"""

import numpy as np

import pandas as pd

from sklearn.preprocessing import MinMaxScaler, StandardScaler

from sklearn.decomposition import PCA

# Step 1: Identify the columns for each metric

bert\_columns = [col for col in df.columns if col.startswith('bert\_score\_precision')]

soft\_similarity\_columns = [col for col in df.columns if col.startswith('soft\_similarity')]

word\_movers\_columns = [col for col in df.columns if col.startswith('word\_movers\_distance')]

deberta\_entailment\_columns = [col for col in df.columns if col.startswith('deberta\_entailment\_score')]

# Step 2: Calculate the mean of the metrics if there are multiple columns

df['bert\_score\_precision\_mean'] = df[bert\_columns].mean(axis=1)

df['soft\_similarity\_mean'] = df[soft\_similarity\_columns].mean(axis=1)

df['word\_movers\_distance\_mean'] = df[word\_movers\_columns].mean(axis=1)

df['deberta\_entailment\_score\_mean'] = df[deberta\_entailment\_columns].mean(axis=1)

# Step 3: Normalize word\_movers\_distance and reverse polarity

scaler\_wmd = MinMaxScaler()

word\_movers\_distance\_normalized = scaler\_wmd.fit\_transform(

df['word\_movers\_distance\_mean'].values.reshape(-1, 1)

).flatten()

word\_movers\_distance\_adjusted = 1 - word\_movers\_distance\_normalized # Lower distance is better

# Step 4: Create a DataFrame with the metrics

metrics\_df = pd.DataFrame({

'bert\_score\_precision': df['bert\_score\_precision\_mean'],

'soft\_similarity': df['soft\_similarity\_mean'],

'word\_movers\_distance': word\_movers\_distance\_adjusted,

'deberta\_entailment\_score': df['deberta\_entailment\_score\_mean']

})

# Step 5: Standardize the metrics

scaler\_standard = StandardScaler()

metrics\_standardized = scaler\_standard.fit\_transform(metrics\_df)

# Step 6: Apply PCA to reduce to a single composite score

pca = PCA(n\_components=1)

principal\_component = pca.fit\_transform(metrics\_standardized)

# Step 7: Add the composite score back to the DataFrame

df['composite\_score'] = principal\_component.flatten()

return df

```

### \*\*Explanation of the Method:\*\*

#### \*\*Imports:\*\*

- \*\*`numpy` and `pandas`\*\*: For data manipulation.

- \*\*`MinMaxScaler` and `StandardScaler`\*\*: For normalization and standardization.

- \*\*`PCA`\*\*: For principal component analysis.

#### \*\*Step 1: Identify the Columns for Each Metric\*\*

- \*\*Metric Column Selection\*\*: Use list comprehensions to find all columns that start with each metric's prefix.

```python

bert\_columns = [col for col in df.columns if col.startswith('bert\_score\_precision')]

soft\_similarity\_columns = [col for col in df.columns if col.startswith('soft\_similarity')]

word\_movers\_columns = [col for col in df.columns if col.startswith('word\_movers\_distance')]

deberta\_entailment\_columns = [col for col in df.columns if col.startswith('deberta\_entailment\_score')]

```

#### \*\*Step 2: Calculate the Mean of the Metrics\*\*

- \*\*Handle Multiple Columns\*\*: If there are multiple columns for a metric (e.g., `soft\_similarity\_0`, `soft\_similarity\_1`), calculate their mean for each row.

```python

df['soft\_similarity\_mean'] = df[soft\_similarity\_columns].mean(axis=1)

```

- \*\*Repeat for All Metrics\*\*: Do the same for `bert\_score\_precision`, `word\_movers\_distance`, and `deberta\_entailment\_score`.

#### \*\*Step 3: Normalize `word\_movers\_distance` and Reverse Polarity\*\*

- \*\*Normalize\*\*: Scale `word\_movers\_distance\_mean` to a range between 0 and 1.

```python

scaler\_wmd = MinMaxScaler()

word\_movers\_distance\_normalized = scaler\_wmd.fit\_transform(

df['word\_movers\_distance\_mean'].values.reshape(-1, 1)

).flatten()

```

- \*\*Reverse Polarity\*\*: Since a lower distance is better, subtract from 1.

```python

word\_movers\_distance\_adjusted = 1 - word\_movers\_distance\_normalized

```

#### \*\*Step 4: Create a DataFrame with the Metrics\*\*

- \*\*Combine Metrics\*\*: Create a new DataFrame `metrics\_df` with the mean values of each metric.

```python

metrics\_df = pd.DataFrame({

'bert\_score\_precision': df['bert\_score\_precision\_mean'],

'soft\_similarity': df['soft\_similarity\_mean'],

'word\_movers\_distance': word\_movers\_distance\_adjusted,

'deberta\_entailment\_score': df['deberta\_entailment\_score\_mean']

})

```

#### \*\*Step 5: Standardize the Metrics\*\*

- \*\*Standardization\*\*: Transform the metrics to have a mean of zero and a standard deviation of one.

```python

scaler\_standard = StandardScaler()

metrics\_standardized = scaler\_standard.fit\_transform(metrics\_df)

```

#### \*\*Step 6: Apply PCA\*\*

- \*\*PCA Transformation\*\*: Reduce the four standardized metrics to one principal component.

```python

pca = PCA(n\_components=1)

principal\_component = pca.fit\_transform(metrics\_standardized)

```

- \*\*Principal Component\*\*: The resulting `principal\_component` is an array with one value per row.

#### \*\*Step 7: Add the Composite Score to the Original DataFrame\*\*

- \*\*Add to DataFrame\*\*: Add the composite score as a new column in `df`.

```python

df['composite\_score'] = principal\_component.flatten()

```

### \*\*Usage Example:\*\*

Assuming you have an instance of your class and your DataFrame is named `df`, you can call the method as follows:

```python

# Calculate the composite score and update the DataFrame

df = self.calculate\_pca\_composite\_score(df)

```

### \*\*Notes and Considerations:\*\*

- \*\*Adjust Column Prefixes\*\*: Ensure that the prefixes used in `col.startswith('metric\_prefix')` match your DataFrame's column naming convention.

- \*\*Missing Values\*\*: Handle any missing values in your data before applying PCA (e.g., fill with mean or median).

- \*\*Multiple Columns per Metric\*\*: The method averages all columns for each metric. If you need to select specific columns, adjust the method accordingly.

- \*\*Explained Variance\*\*: You can access the explained variance of the principal component if needed:

```python

explained\_variance = pca.explained\_variance\_ratio\_[0]

print(f"Explained Variance: {explained\_variance:.2%}")

```

### \*\*Example Output:\*\*

After running the method, your DataFrame `df` will have a new column `composite\_score`:

```plaintext

soft\_similarity\_0 soft\_similarity\_1 ... deberta\_entailment\_score\_mean composite\_score

0 0.75 0.80 ... 0.85 1.234567

1 0.60 0.65 ... 0.70 0.987654

...

```

### \*\*Additional Options:\*\*

- \*\*Using Specific Columns\*\*: If you prefer to use specific columns instead of averaging, you can modify Step 2:

```python

df['soft\_similarity\_mean'] = df['soft\_similarity\_0'] # Use only 'soft\_similarity\_0'

```

- \*\*Including Weights\*\*: PCA does not incorporate weights directly. If you need to apply specific weights to the metrics, consider using a weighted sum approach instead.

### \*\*Final Thoughts:\*\*

- \*\*PCA vs. Weighted Sum\*\*: PCA is suitable when you want to combine correlated metrics into a single score based on their variance structure. If you have predefined weights for each metric, a weighted sum might be more appropriate.

- \*\*Interpretability\*\*: The composite score from PCA is a linear combination of the standardized metrics. The weights (loadings) can be accessed via `pca.components\_` if you wish to interpret the contribution of each metric.

```python

loadings = pca.components\_[0]

metric\_contributions = pd.Series(loadings, index=metrics\_df.columns)

print(metric\_contributions)

```

### \*\*Alternative: Weighted Sum Method Using Standardization\*\*

If you decide to stick with the weighted sum approach but want to standardize the metrics, you can modify your existing method:

```python

def calculate\_weighted\_composite\_score(self, df):

"""

Calculate the composite score based on weighted contributions from 4 metrics after standardization.

Parameters:

- df: DataFrame containing the metrics.

Returns:

- df: DataFrame with the composite score added.

"""

from sklearn.preprocessing import MinMaxScaler, StandardScaler

# Identify and process columns as before (Steps 1-3)

# Standardize the metrics

scaler\_standard = StandardScaler()

metrics\_standardized = scaler\_standard.fit\_transform(metrics\_df)

# Apply weights

weights = np.array([0.45, 0.35, 0.15, 0.05])

composite\_scores = metrics\_standardized.dot(weights)

# Add the composite score to the DataFrame

df['composite\_score'] = composite\_scores

return df

```

### \*\*Conclusion:\*\*

By integrating the `calculate\_pca\_composite\_score` method into your class, you can effectively compute a composite score using PCA, handling columns with indices, and ensuring proper data preprocessing. Remember to adjust the method to fit your specific data structure and requirements.

## Output Analysis

### Range and Variance of Composite Scores and PCA Scores

Range: The range of both scores gives an idea of the spread of the data. For \*\*composite scores\*\*, the range will depend on the weights assigned to different metrics. Ideally, these scores are scaled between 0 and 1 (since the metrics are normalized). The \*\*PCA scores\*\* will depend on the variance captured by each principal component. A wider range in PCA scores typically indicates greater variability in the data along that principal component.

Variance:

Composite Scores: Variance in composite scores represents how spread out the "fit" or "alignment" is between responsibilities and requirements. Low variance suggests that most responsibilities have similar levels of fit, while high variance indicates some responsibilities are significantly better (or worse) matches than others.

- \*\*PCA Scores\*\*: Variance in PCA scores shows how much of the total variance in the dataset is explained by each principal component. A high variance in PCA scores along one component indicates that this component captures significant variability, implying that responsibilities or requirements exhibit substantial differences along this axis.

### 2. \*\*Analysis of Composite and PCA Score Combinations\*\*

By analyzing different combinations of \*\*composite scores\*\* and \*\*PCA scores\*\*, we can categorize responsibilities and requirements into four cases:

#### \*\*Case 1: Low Composite Score / Low PCA Score\*\*

- \*\*Interpretation\*\*: A responsibility or requirement with both low composite and PCA scores is likely not very distinctive and doesn't align well with the required metrics.

- \*\*Characteristics\*\*:

- \*\*Composite Score\*\*: Low alignment across the weighted metrics.

- \*\*PCA Score\*\*: This responsibility or requirement doesn't explain much variance in the dataset, meaning it’s not particularly unique or important.

- \*\*Action\*\*: Candidates in this category might be \*\*pruned\*\* as they are neither well-aligned nor structurally significant.

#### \*\*Case 2: High Composite Score / Low PCA Score\*\*

- \*\*Interpretation\*\*: A responsibility or requirement with a high composite score but a low PCA score aligns well with the metrics but doesn’t contribute significantly to the overall variance in the dataset.

- \*\*Characteristics\*\*:

- \*\*Composite Score\*\*: High fit or alignment across multiple metrics.

- \*\*PCA Score\*\*: Doesn’t represent much variability in the dataset.

- \*\*Action\*\*: These responsibilities are \*\*good fits\*\* based on metrics but \*\*not unique or important\*\* from a variance perspective. If pruning based solely on alignment, keep them; if uniqueness is a priority, these could be deprioritized.

#### \*\*Case 3: Low Composite Score / High PCA Score\*\*

- \*\*Interpretation\*\*: A low composite score but high PCA score indicates that the responsibility or requirement is structurally important but doesn’t align well with the predefined metrics.

- \*\*Characteristics\*\*:

- \*\*Composite Score\*\*: Poor alignment based on the weighted metrics.

- \*\*PCA Score\*\*: High variance, meaning it captures important structural information in the dataset.

- \*\*Action\*\*: These cases are \*\*unique or distinctive\*\* but don’t fit the requirements well. They may warrant closer examination to understand if they represent outliers, exceptions, or unexplored areas worth investigating.

#### \*\*Case 4: High Composite Score / High PCA Score\*\*

- \*\*Interpretation\*\*: Responsibilities or requirements with both high composite and PCA scores are \*\*highly aligned\*\* with the metrics and are \*\*structurally significant\*\*.

- \*\*Characteristics\*\*:

- \*\*Composite Score\*\*: High fit with multiple metrics.

- \*\*PCA Score\*\*: These responsibilities or requirements capture substantial variability in the dataset.

- \*\*Action\*\*: These are \*\*key data points\*\* that are both well-aligned with the requirements and structurally important. These should be prioritized and retained.

### 3. \*\*Which Score is More Appropriate for Pruning?\*\*

- \*\*PCA Scores for Pruning\*\*:

- \*\*Insight\*\*: PCA scores give insights into the \*\*underlying structure\*\* of the data and help to identify which responsibilities or requirements capture the most variability.

- \*\*Pruning Approach\*\*: If your goal is to reduce redundancy and keep only the most important (distinctive or unique) data points, \*\*PCA scores\*\* are ideal. Low PCA scores can indicate that certain responsibilities or requirements are not contributing much to the overall structure and can be pruned.

- \*\*Composite Scores for Pruning\*\*:

- \*\*Insight\*\*: Composite scores provide a \*\*domain-specific summary\*\* of how well responsibilities align with requirements based on weighted metrics.

- \*\*Pruning Approach\*\*: If your goal is to \*\*remove poorly aligned responsibilities\*\* (i.e., those that don’t meet the key metrics), use \*\*composite scores\*\* for pruning. Low composite scores can be pruned to focus on responsibilities that meet your alignment criteria.

### \*\*Recommendation:\*\*

- \*\*Combined Pruning\*\*: A more balanced approach could involve using \*\*both scores\*\* together. Responsibilities or requirements that are \*\*low in both composite and PCA scores\*\* are strong candidates for pruning because they are neither well-aligned with your metrics nor structurally significant.

- \*\*Low Composite Score/Low PCA Score\*\*: These are the best candidates for removal.

- \*\*High Composite Score/Low PCA Score\*\*: Retain if alignment is more important than uniqueness.

- \*\*Low Composite Score/High PCA Score\*\*: Consider keeping for their structural value, or prune if alignment is critical.

- \*\*High Composite Score/High PCA Score\*\*: Prioritize these for retention.

# Pruning Results Based on Scores

## Claude Recommendations

Outline provided by Claude

```python

import pandas as pd

import numpy as np

def load\_data(file\_path):

# Load data from the provided file format

# This is a placeholder function - you'll need to implement the actual loading logic

pass

def calculate\_alignment\_score(row):

# Calculate a single alignment score from the provided metrics

# You can adjust the weights based on your priorities

weights = {

'scaled\_bert\_score\_precision': 0.3,

'scaled\_soft\_similarity': 0.3,

'scaled\_word\_movers\_distance': 0.2,

'scaled\_deberta\_entailment\_score': 0.2

}

return sum(row[metric] \* weight for metric, weight in weights.items())

def prune\_texts(data, alignment\_threshold=0.5, composite\_threshold=0.3, pca\_threshold=-0.5, target\_ratio=0.8):

# Calculate alignment scores

data['alignment\_score'] = data.apply(calculate\_alignment\_score, axis=1)

# Filter based on thresholds

filtered\_data = data[

(data['alignment\_score'] >= alignment\_threshold) &

(data['composite\_score'] >= composite\_threshold) &

(data['pca\_score'] >= pca\_threshold)

]

# Sort by alignment score

sorted\_data = filtered\_data.sort\_values('alignment\_score', ascending=False)

# Select top texts to meet target ratio

target\_count = int(len(data['Requirement Key'].unique()) \* target\_ratio)

selected\_data = sorted\_data.groupby('Responsibility Key').head(target\_count)

return selected\_data['Responsibility']

# Example usage

data = load\_data('your\_data\_file.csv')

pruned\_responsibilities = prune\_texts(data)

print(pruned\_responsibilities)

```

Here's a breakdown of the updated pruning pipeline:

1. Data Loading:

- Load the data from your provided format. You'll need to implement the `load\_data` function to parse your specific file structure.

2. Filtering:

- Apply thresholds on composite score and PCA score.

- These thresholds can be adjusted as needed.

3. Sorting and Selection:

- Sort the filtered data by alignment score.

- Select the top responsibilities to meet your target ratio (e.g., 80% of the number of unique requirements).

5. Return Pruned List:

- Return the final list of selected responsibilities.

Key Points:

- This pipeline uses the pre-calculated metrics you've provided, including the composite score and PCA score.

- It balances multiple factors: alignment (using various similarity metrics), overall quality (composite score), and dimensionality reduction insights (PCA score).

- The pipeline is flexible: you can adjust thresholds and weights as needed.

- It aims to achieve your target ratio while maintaining quality across multiple dimensions.

Recommendations:

1. Metric Utilization:

- Leverage all available metrics. The composite\_score and pca\_score are particularly useful as they already combine multiple factors.

- Pay special attention to the scaled metrics as they're likely more comparable across different scales.

2. Threshold Tuning:

- Start with moderate thresholds and adjust based on results. For example:

* + composite\_threshold: 0.3 (adjust based on the distribution in your full dataset)
  + pca\_threshold: -0.5 (since PCA scores can be negative, this might need significant tuning)

3. Responsibility Selection:

- The pipeline selects responsibilities rather than requirements. Ensure this aligns with your goal of pruning List A.

4. Data Exploration:

- Before finalizing thresholds, explore the distributions of your scores to ensure you're not cutting off too much or too little data.

5. Iterative Approach:

- Run the pipeline multiple times with different parameters to find the best balance between pruning and maintaining relevance.

To use this pipeline:

1. Prepare your data in a format that can be loaded into a pandas DataFrame.

2. Adjust the thresholds and weights in the `prune\_texts` function as needed.

3. Call the `prune\_texts` function with your data.

4. The function will return your pruned list of responsibilities.

## GPT Recommendations

### Prompt to GPT

My data:

1. Two lists of texts: list A is much longer than list B;

2. A matching file with each element of A comparing to each element of B;

3. each row has 4 metric scores-Bert score precision, soft similarity, word mover distance, and Alberta entailment scores; the directional scores, especially entailment and soft similarity are more important.

4. Each row also has two indices derived from the metrics, PCA and weighted scaled average. 5. A new list of A, after using LLM to modify each element of A to achieve better aligned with each element of List B;

6. A new matching file like the previous one but based on modified list A;

7 & 8, other iterations (the logic is the same).

See example data.

Once I am done with a couple iterations, what is the best course, or courses of, action give my goal below?

<goal>short list List A by prioritizing texts that are more aligned with List B and reduce texts that are less relevant. There final list should have optimal overlap and alignment with B, but not perfect alignment; list A should still be larger than B, I.e., 70-90 percent but not 95%+. This is to preserve authenticity of the original list. The final list A will be re-insert back into JSON to form the edited document A</goal >

<example data>

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Responsibility Key | Responsibility | Requirement Key | Requirement | bert\_score\_precision | soft\_similarity | word\_movers\_distance | deberta\_entailment\_score | bert\_score\_precision\_cat | soft\_similarity\_cat | word\_movers\_distance\_cat | deberta\_entailment\_score\_cat | scaled\_bert\_score\_precision | scaled\_deberta\_entailment\_score | scaled\_soft\_similarity | scaled\_word\_movers\_distance | composite\_score | pca\_score |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 0.pie\_in\_the\_sky.0 | MBA or graduate degree in a management, technical, or engineering field | 0.8172221183776855 | 0.1528222411870956 | 0.740102 | 0.050603 | Medium | Low | Medium | Low | 0.3374518106179494 | 0.056978 | 0.22941442256170863 | 0.740102 | 0.23382285561302601 | -0.98042 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 0.pie\_in\_the\_sky.1 | Knowledge of the Machine Learning and Artificial Intelligence market landscape, ideally with a focus on developer tooling | 0.8406854867935181 | 0.2081294506788253 | 0.6526921202829397 | 0.001224 | Medium | Medium | Medium | Low | 0.5705670209931011 | 0.000911 | 0.33437339200803734 | 0.6526921202829397 | 0.2438729721983386 | -0.34797 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 1.down\_to\_earth.0 | 11 years of experience in management consulting, product management and strategy, or analytics in a technology company | 0.8380018472671509 | 0.2966089248657226 | 0.6114212519765081 | 0.000421 | Medium | Medium | Medium | Low | 0.5439043022532797 | 0 | 0.5022848544223263 | 0.6114212519765081 | 0.2947081019569544 | 0.39193560638356206 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 1.down\_to\_earth.1 | Experience working with and analyzing data, and managing multiple cross-functional programs or projects | 0.8287657499313354 | 0.2934322357177734 | 0.6739010982681253 | 0.048156 | Medium | Medium | Medium | Low | 0.4521410594261681 | 0.054199338953625574 | 0.49625630909732765 | 0.6739010982681253 | 0.32177162842472334 | 0.20475685573616237 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 1.down\_to\_earth.2 | Experience with performing market analysis and developing competitive intelligence | 0.8258360028266907 | 0.2834358811378479 | 0.740102 | 0.1734191924333572 | Medium | Medium | Medium | Low | 0.42303319219494817 | 0.19642616390312853 | 0.47728577913258674 | 0.740102 | 0.3876087888630489 | 0.050617 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 1.down\_to\_earth.3 | Ability to manage executive stakeholders and communicate with a highly technical management team | 0.8464140295982361 | 0.2675277888774872 | 0.6955220164826892 | 0.005443 | Medium | Medium | Medium | Low | 0.6274817161637998 | 0.005702 | 0.4470962797011563 | 0.6955220164826892 | 0.29475202460644345 | -0.08474 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 1.down\_to\_earth.4 | Ability to form and refine hypotheses, gather supporting data, and make recommendations | 0.8294726610183716 | 0.1206346675753593 | 0.7175798533808311 | 0.036552 | Medium | Low | Medium | Low | 0.4591644212832744 | 0.041024 | 0.1683306219810472 | 0.7175798533808311 | 0.20797177583955775 | -1.13058 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 1.down\_to\_earth.5 | Excellent problem solving and analysis skills, including opportunity identification, market segmentation, and framing of complex/ambiguous problems | 0.8162370920181274 | 0.1965780556201934 | 0.5913197764017799 | 0.005891 | Medium | Low | Medium | Low | 0.32766529476208817 | 0.00621 | 0.312452 | 0.5913197764017799 | 0.21723384690509767 | -0.22277 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 2.other.0 | English proficiency is a requirement for all roles unless stated otherwise in the job posting | 0.8235018253326416 | 0.1588080525398254 | 0.740102 | 0.001018 | Medium | Low | Medium | Low | 0.39984247771888803 | 0.000677 | 0.24077396495512127 | 0.740102 | 0.21558293275507773 | -0.99315 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 2.other.1 | Work across Program Management teams and our partners (engineering, UX, Customer Experience, TPM, Marketing, Developer Relations, etc.) to help shape the future of AI at Google | 0.8446339964866638 | 0.3389898538589477 | 0.495496 | 0.028476 | Medium | Medium | Medium | Low | 0.6097965830693148 | 0.031854 | 0.5827130422153138 | 0.495496 | 0.3230978702197551 | 1.1144343213563241 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 2.other.2 | Leverage first party and third party market data to build assets and programs that surface valuable insights to our business stakeholders and help inform product roadmaps | 0.8557024002075195 | 0.3567725419998169 | 0.5329467068056377 | 0.013899 | High | Medium | Medium | Low | 0.719764 | 0.015303219900380438 | 0.6164600462389064 | 0.5329467068056377 | 0.33857768659800197 | 1.0953326009593602 |
| 0.responsibilities.0 | Provided strategic insights to a major global IT vendor, optimizing their service partner ecosystem in Asia Pacific for improved local implementation outcomes. | 2.other.3 | Identify gaps in the existing data and engage in original research to fill these gaps, utilizing third party vendors and tooling where appropriate. Create ongoing cadences to enable research distribution and actionable recommendations (e.g. newsletters, dashboards, exec reviews, etc.) | 0.8228855133056641 | 0.3333046436309814 | 0.2903396954025568 | 0.017782 | Medium | Medium | Medium | Low | 0.39371924318242435 | 0.019711306403089116 | 0.5719239640445501 | 0.2903396954025568 | 0.2722803917664874 | 1.758654931728454 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 0.pie\_in\_the\_sky.0 | MBA or graduate degree in a management, technical, or engineering field | 0.8195508718490601 | 0.240227 | 0.76312 | 0.065323 | Medium | Medium | Medium | Low | 0.3605886358925776 | 0.073691 | 0.39528708266298895 | 0.76312 | 0.30400890240952605 | -0.44881 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 0.pie\_in\_the\_sky.1 | Knowledge of the Machine Learning and Artificial Intelligence market landscape, ideally with a focus on developer tooling | 0.8252819776535034 | 0.2498035579919815 | 0.631872 | 0.016338 | Medium | Medium | Medium | Low | 0.4175287951914246 | 0.018071983114664816 | 0.4134602126301961 | 0.631872 | 0.2685007661456086 | 0.015359394646933923 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 1.down\_to\_earth.0 | 11 years of experience in management consulting, product management and strategy, or analytics in a technology company | 0.8336870670318604 | 0.3040940761566162 | 0.631872 | 0.000955 | Medium | Medium | Medium | Low | 0.5010357386077633 | 0.000605 | 0.5164897613930808 | 0.631872 | 0.3008765162600876 | 0.3705211291815432 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 1.down\_to\_earth.1 | Experience working with and analyzing data, and managing multiple cross-functional programs or projects | 0.8255743384361267 | 0.1813950538635254 | 0.6526921202829397 | 0.1350825577974319 | Medium | Low | Medium | Low | 0.42043348236757083 | 0.152898 | 0.28363832935905875 | 0.6526921202829397 | 0.28700289675018037 | -0.38432 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 1.down\_to\_earth.2 | Experience with performing market analysis and developing competitive intelligence | 0.8262200951576233 | 0.3111447691917419 | 0.7175798533808311 | 0.2607548534870147 | Medium | Medium | Medium | Medium | 0.4268492582832444 | 0.295589 | 0.5298701774663209 | 0.7175798533808311 | 0.44744920602843763 | 0.4162668717567929 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 1.down\_to\_earth.3 | Ability to manage executive stakeholders and communicate with a highly technical management team | 0.8392523527145386 | 0.1618953943252563 | 0.6739010982681253 | 0.008112 | Medium | Low | Medium | Low | 0.5563284280342291 | 0.008732 | 0.246633 | 0.6739010982681253 | 0.21915250363582467 | -0.7275 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 1.down\_to\_earth.4 | Ability to form and refine hypotheses, gather supporting data, and make recommendations | 0.8247349262237549 | 0.150433 | 0.6955220164826892 | 0.1225656569004058 | Medium | Low | Medium | Low | 0.41209368430402904 | 0.1386857399150469 | 0.2248794963965709 | 0.6955220164826892 | 0.26604939338817574 | -0.75664 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 1.down\_to\_earth.5 | Excellent problem solving and analysis skills, including opportunity identification, market segmentation, and framing of complex/ambiguous problems | 0.8154844045639038 | 0.2296628355979919 | 0.5715506253044318 | 0.015622 | Medium | Medium | Medium | Low | 0.3201871317324496 | 0.017258678846751113 | 0.37523826135467836 | 0.5715506253044318 | 0.24084174733746264 | 0.080452 |
| 0.responsibilities.1 | Assisted a U.S.-based international services provider in its growth strategy by precisely evaluating and scaling new engineering service opportunities in vital emerging markets. | 2.other.0 | English proficiency is a requirement for all roles unless stated otherwise in the job posting | 0.8225706815719604 | 0.1124952435493469 | 0.7175798533808311 | 0.001077 | Medium | Low | Medium | Low | 0.39059130074319803 | 0.000744 | 0.15288407232670792 | 0.7175798533808311 | 0.18101080276725778 | -1.2301 |

</example data>

### Recommendation

Here’s the updated recommendation using the \*\*Max-Based Approach\*\*:

1. \*\*Max-Based Retention of Responsibilities\*\*

- For each responsibility in List A, find the maximum composite score across all its comparisons with List B’s requirements.

- If a responsibility shows \*\*high alignment with at least one requirement\*\* (i.e., it has a high maximum composite score), retain that responsibility.

- Responsibilities with \*\*no strong alignment\*\* (low maximum composite score across all requirements) will be considered for elimination.

2. \*\*Threshold for High Alignment\*\*

- Set a threshold based on the maximum composite score to determine what qualifies as "high alignment." Responsibilities with a maximum composite score above this threshold will be retained, while those below it will be eliminated.

- Adjust this threshold dynamically to ensure that you retain 70-90% of List A while preserving its alignment with List B.

3. \*\*Iterate and Refine\*\*

- After each iteration (modifications to List A using LLM), apply the same max-based filtering method to further refine the list.

- Continue adjusting the threshold based on the max composite score to maintain the target size of List A.

4. \*\*Final Selection\*\*

- Once the process is complete, the final list of responsibilities will have the strongest alignment with at least one requirement from List B.

- The final List A will retain 70-90% of its original size, prioritizing alignment with B while keeping a degree of authenticity.

5. \*\*Reinsertion into JSON\*\*

- Once finalized, reinsert the shortlisted List A back into JSON format to complete the process and generate the final edited document.

This approach ensures that strong matches are preserved, even if some responsibilities show weaker alignment in other areas, giving you a balance between alignment and authenticity.

Final Solution: Claude + GPT

high level -> GPT implementation -> Claude refined (added a more dynamic approach - elbow method -> GPT further code implementation

## Notes on Adaptive Pruning from Claude

Here's how the `adaptive\_pruning` method works:

1. Initial Setup:

- It starts with the ranked DataFrame (sorted by the combined score).

- It calculates the total number of unique responsibilities.

- It determines the target count of responsibilities based on the `target\_percentage` (e.g., if we want to keep 80% of responsibilities, and we start with 100, the target count would be 80).

2. Pruning Loop:

- The method uses a while loop to gradually prune the responsibilities.

- It starts with `current\_percentage = 1.0`, meaning it initially includes all responsibilities.

3. Gradual Reduction:

- In each iteration of the loop, it reduces the `current\_percentage` by a small step (default is 0.05 or 5%).

- It then calculates a threshold score based on this percentage. For example, if `current\_percentage` is 0.95, it finds the score at the 95th percentile.

4. Thresholding:

- It keeps all responsibilities with a combined score less than or equal to this threshold. (Remember, lower scores are better because they represent better rankings.)

5. Check and Repeat:

- After each reduction, it checks if the number of unique responsibilities is still above the target count.

- If it is, the loop continues, further reducing the percentage and recalculating the threshold.

6. Loop Termination:

- The loop stops when the number of unique responsibilities falls below or reaches the target count.

Here's a step-by-step example:

1. Start with 100 unique responsibilities, target is 80% (80 responsibilities).

2. First iteration: Keep 95% (threshold at 95th percentile) - might have 95 responsibilities left.

3. Second iteration: Keep 90% - might have 90 responsibilities left.

4. Third iteration: Keep 85% - might have 85 responsibilities left.

5. Fourth iteration: Keep 80% - reaches 80 responsibilities, stops here.

The key advantages of this method are:

1. Flexibility: It can adapt to different distributions of scores.

2. Controlled Reduction: It reduces the list size gradually, which helps in finding a balance between pruning and maintaining diversity.

3. Target-Oriented: It aims for a specific reduction percentage, aligning with your goal of keeping 70-90% of the original list.

This approach should help you avoid the overly aggressive pruning you were experiencing with the elbow method, while still allowing you to reduce your list to a desired size range.

## Notes on How to Deal with Too Few Results with Pruning Claude

Recommendations

1. Don't prune at each iteration: Your suggestion to not prune at each iteration is a good one. Instead, focus on refining and improving the alignment of responsibilities during each iteration.
2. Use a more flexible pruning approach: Instead of relying solely on the elbow method, which seems to be overly aggressive in your case, I've proposed an adaptive pruning approach in the artifact above. This method allows you to target a specific percentage range for the final list size.
3. Consider all responsibilities: Allow duplicate responsibilities during the iteration process, as you suggested. This can help preserve more nuanced alignments between List A and List B.
4. Combine multiple metrics: The revised approach uses a combined score based on both composite and PCA rankings. This can provide a more balanced view of each responsibility's relevance.
5. Visualize your data: The plot\_score\_distribution method will help you understand the distribution of your scores, which can inform your pruning strategy.
6. Adaptive pruning: The adaptive\_pruning method allows you to set a target percentage (e.g., 80% of original size) and gradually prunes until reaching that target.

# Async Approach

Use async because of playwright.

You have to await the functions; otherwise, you will run into problems b/c of race issue.