# An Application of Text Embeddings to Support Alignment of Educational Content Standards

Reese Butterfuss Harold Doran

Paper Presented at Generative Artificial Intelligence for Measurement and Education Meeting

January 28, 2024

**Abstract**

Large language models are increasingly used in educational and psychological measurement activities. Their rapidly evolving sophistication and ability to detect language semantics makes them viable tools to supplement subject matter experts and their reviews of large amounts of text statements, such as educational content standards. This paper presents an application using text embeddings to find relationships between different sets of educational content standards in a process termed content mapping. This content mapping process is routinely used by state education agencies and in some cases is a requirement of the United States Department of Education peer review process. We discuss the educational measurement problem, propose a formal methodology, demonstrate an application of our proposed approach, and provide measures of its accuracy and potential to support real-world activities.

*Keywords:* content alignment, content mapping, natural language processing, semantic similarity, text embeddings

# Introduction

Alignment is an important task in educational and psychological measurement and is one component of the validity investigation process regarding an assessment system. It is a broad term and has implications for work taking many forms with variants including alignment of test items to standards (Webb, 1997), aligning two distinct sets of standards (Neidorf et al., 2016), or establishing relationships between performance level descriptors (PLDs) and test items (Schneider, Agrimson, & Veazey, 2022).

A common theme across these alignment approaches is that they involve subject matter experts (SMEs) in a time-consuming process where they generally review large amounts of text data and/or test items in some form. In general, alignment activities require expert judges to review test content (e.g., items, standards, PLDs) and use their expert judgment to determine the degree to which there is a similarity between the test content under review. For example, experts may evaluate a test question against a learning objective and assess whether it is measuring that standard or not or experts may evaluate two sets of learning standards and find pairs of text statements that form matches between their intended learning objectives.

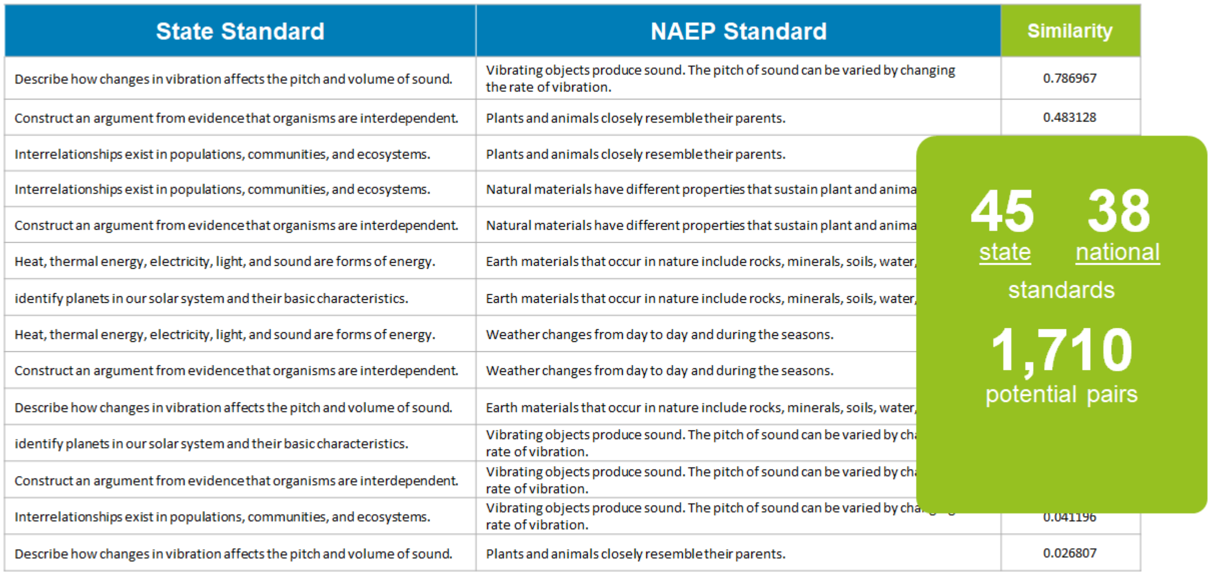
Content standards are formal definitions of what a student is expected to know and be able to do in a given subject area and grade level. These standards are often locally defined (e.g., by a state education agency (SEA)) and are used for instruction and assessment design in that jurisdiction. However, other sets of standards also exist and it is often useful to show how one set of standards formally relate to a differet set of standards. This is generally necessary given that different sets of standards often express similar learning outcomes for students but use language to express these outcomes in different ways. In some instances, demonstrating this relationship is required, such as addressing critical element 3.1 in the United States Department of Education (USDOE) standards and assessment peer review process (USDOE, 2018).

Establishing this relationship between different sets of standards is referred to as standards-to-standards *content mapping* and the aim is to find if a standard in one set can be matched with a comparable learning standard in the second set. Some examples of content mapping include the Dynamic Learning Maps (DLMs) crosswalk to the Common Core State Standards (CCSS) (Dynamic Learning Maps Consortium, 2013), mapping the National Assessment of Educational Progress (NAEP) to the CCSS (Daro, Hughes, & Stancavage, 2015), mapping NAEP to state standards via the HumRRO approach (Vockley, 2009), or mapping the World-Class Instructional Design and Assessment (WIDA) English Language Development Standards (ELD) to individual state standards to (WIDA, 2020).

This is a lengthy and time-consuming process with SMEs staring at pairs of text statements and using their expert judgment to determine similarity. The largest challenge is that SMEs rate all pairs of text and this can be an overwhelming level of effort. For example, Figure (1) illustrates how large this task can be. In this example, there are 45 state standards and 38 NAEP standards. Content mapping in this scenario would involve human raters evaluating all 45 × 38 = 1,710 pairs of text statements to make a judgement on which standards are aligned.

In practical educational settings, this work often involves teachers or other content experts working long hours in addition to their teaching activities to complete these mapping activities. Additionally, the sheer number of pairwise text statements reviewed represents an enormous amount of work, often resulting in cognitive overload and fatigue, both of which are the root cause of errors. The challenge is seemingly obvious—is it possible to find ways where SMEs can review smaller subsets of text pairs that contain the most probable matching pairs instead of reviewing large amounts of text in an overwhelming fashion?

Existing research regarding the application of modern natural language processing (NLP) approaches to content alignment is scarce, but emerging work has begun examining the applicability of NLP to alignment. Zhou and Ostrow (2022) evaluated the accuracy with which training transformer models from scratch predicts the standards for each item in a large corpus of English language arts assessment items. Their results showed that the predicted standard matched the true standard at rate a of .60 on the training set and .29 on the test set. Their method leveraged word-level embeddings to represent the meaning of the assessment items and averaged the word embeddings to represent the meaning of each assessment item. An alternative approach is



| State Standard | NAEP Standard | Similarity |
|---|---|---|
| Describe how changes in vibration affects the pitch and volume of sound. | Vibrating objects produce sound. The pitch of sound can be varied by changing the rate of vibration. | 0.786967 |
| Construct an argument from evidence that organisms are interdependent. | Plants and animals closely resemble their parents. | 0.483128 |
| Interrelationships exist in populations, communities, and ecosystems. | Plants and animals closely resemble their parents. | |
| Interrelationships exist in populations, communities, and ecosystems. | Natural materials have different properties that sustain plant and anima | |
| Construct an argument from evidence that organisms are interdependent. | Natural materials have different properties that sustain plant and anima | |
| Heat, thermal energy, electricity, light, and sound are forms of energy. | Earth materials that occur in nature include rocks, minerals, soils, water, | |
| identify planets in our solar system and their basic characteristics. | Earth materials that occur in nature include rocks, minerals, soils, water, | |
| Heat, thermal energy, electricity, light, and sound are forms of energy. | Weather changes from day to day and during the seasons. | |
| Construct an argument from evidence that organisms are interdependent. | Weather changes from day to day and during the seasons. | |
| Describe how changes in vibration affects the pitch and volume of sound. | Earth materials that occur in nature include rocks, minerals, soils, water, | |
| identify planets in our solar system and their basic characteristics. | Vibrating objects produce sound. The pitch of sound can be varied by ch rate of vibration. | |
| Construct an argument from evidence that organisms are interdependent. | Vibrating objects produce sound. The pitch of sound can be varied by ch rate of vibration. | |
| Interrelationships exist in populations, communities, and ecosystems. | Vibrating objects produce sound. The pitch of sound can be varied by cha rate of vibration. | 0.041196 |
| Describe how changes in vibration affects the pitch and volume of sound. | Plants and animals closely resemble their parents. | 0.026807 |

Figure 1: *Total Pairwise Comparisons*

to derive sentence embeddings to represent the meaning of text which are superior to word embeddings at capturing context and therefore deeper, more nuanced meaning of text input (Arora, Liang, & Ma, 2017; Reimers & Gurevych, 2019). Khan, Rosaler, Hamer, and Almeida (2021) developed Catalog, a content-tagging system that uses semantic similarity derived from pretrained transformer models to match educational materials (e.g., reading passages) to 24 unique Next Generation Science Standards (NGSS). Results from Catalog suggest that it performed either similarly or superior to human judgment in accurately matching high school biology textbook passages to NGSS standards.

The existing transformer-based approaches in the extant literature show promise, but there are shortcomings and barriers to practical use that could be addressed in subsequent work. Specifically, the approach developed by Zhou and Ostrow (2022) requires sufficiently large corpora to train the transformer models, does not leverage sentence embeddings, and struggles with instances when a single item aligns with multiple standards. The approach developed by Khan et al. (2021) indeed leverages sentence embeddings and does not require training data, but it is designed to facilitate alignment of educational content with only NGSS standards. Moreover, the goals of these approaches focus more on fully *automating* the alignment process rather than on providing a means of increasing the efficiency and effectiveness of SMEs during the alignment process.

## Purpose and Organization

This paper presents a method that can assist SMEs by leveraging transformer models, text embeddings, and the cosine similarity index to improve the efficiency of the content mapping process. Specifically, we developed an intuitive method that can potentially apply to the alignment of any content standards and requires no text preprocessing or labeled training datasets. The practical goal of the method is not to automate content alignment outright but rather to facilitate the alignment task by decreasing the workload imposed on SMEs. To this end, we sought to constrain the search space SMEs must negotiate during alignment by limiting the number of pairs to only those that share high semantic overlap and are therefore more likely candidates for alignment.

Hence, our approach is *NLP-assisted* and not *NLP-driven*. That is, an NLP-driven method is one where the approach on its own forms a classification between matched pairs and an NLP-assisted method is used to mean the language model only helps inform which pairs seem most likely to share a relationship and then SMEs review the subset and make final judgments. This approach makes use of NLP techniques as a supporting mechanism to initially form probable pairs and reduce the total number of text pairs experts would need to review. The method itself does not establish alignment or make any decisions. This is sometimes referred to as human-in-the-loop methods so SMEs are still fundamentally responsible for the outcomes.

The paper is organized into three sections. First, the concept of text embeddings is described. Second, we formalize a methodology that uses these embeddings as a supporting mechanism to facilitate content mapping. Third, we demonstrate this method using real NAEP data along with its accuracy metrics to demonstrate its potential in real-world settings.

# Sentence Embeddings as a Supporting Tool

A *sentence embedding* is a transformation of a text statement into a numeric representation of that text. These embeddings are obtained via large language models (LLMs) that convert text to embeddings and also capture the semantic meaning inherent in the text itself. This differs from other approaches that only convert individual words such as GLOVE and Word2Vec (Zhou & Ostrow, 2022) and other historical text transformation approaches that are designed only to evaluate whether text strings are similar enough to be considered the same piece of text, not whether they are capturing similar meaning such as a Levenshtein distance (Doran & Van Wamelen, 2010).

Figure (2) provides a sample of the sentence embedding for one statement. This example shows that text is converted into a numeric vector that may be longer than the total number of words in the sentence. Embeddings are not limited to single sentences but can more broadly convert text statements longer than a single sentence into a numeric representation. Going forward, we use the term *text embedding* to apply the concept of our work more broadly.

Importantly, the numeric transformations of each text statement now allow use of empirical methods to establish relationships between large numbers of text statements. For example, assume we have two text statements that have each been transformed into a numeric embedding, we can then establish a statistical relationship between them. The common statistic used in machine learning for relating pairs of text embeddings is the cosine similarity index which is a special case of the Pearson correlation when the means of the vectors are centered on zero. It is a distance metric that effectively determines how similar two different pairs of text are in terms of their semantic similarity and has the same range and interpretation as the Pearson correlation.

When this work is performed in large batches (i.e., converting large amounts of text into embeddings) then perhaps we can also more efficiently compare large numbers of text pairs and find relationships between them. This could potentially save large amounts of human effort by forming subsets of the most probable matching pairs and asking experts to evaluate a smaller number of possible matching pairs instead of evaluating all possible pairs.

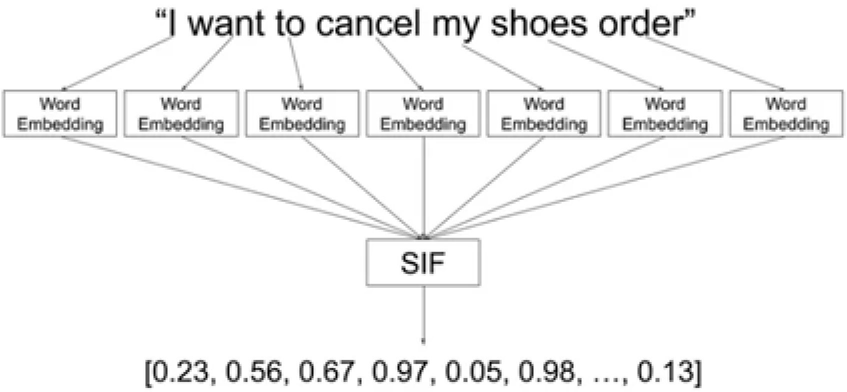


Figure 2: *Total Pairwise Comparisons*

# Outlining an NLP-Based Framework for Content Mapping

We can begin to imagine a new way content mapping with large batches of text might occur with the support of text embeddings. Assume there are two (or more) sets of content standards to be aligned. Informally, an outline and high-level sketch of a new process can be framed as:

1. Compute text embeddings between all pairs of text statements in both sets.
2. Compute cosine similarity between all pairs of embeddings.
3. Sort each statement with its corresponding pair in descending order by cosine similarity.
4. Ask SMEs to find the matching pair in a subset of the ordering from (3) instead of searching over the entire set of text statements.

The general idea is that the cosine similarity index can be used to find which pairs of text have the greatest likelihood of being matched. Then, by rank ordering the text statements by their cosine similarity, SMEs will review a smaller subset of text statements where, hopefully, the true match will be somewhere in the top few text statements. Viewed from this perspective, our method orders content in a manner that resembles the Bookmarking standard setting method where SMEs review test items ordered by their response probabilities (RP) (Mitzel, Lewis, Patz, & Green, 2001). In our approach, we order content standard pairs by their cosine similarity like how test items are ordered by RP values and SMEs would review this ordered set to locate the matched pair. This is just a general sketch of the process and the next section formalizes this implementation. Readers less interested in the technical approach can skip the formalities.

# Formalizing the Method

This section describes the formalities of the approach. Let the collection of text statements be represented as two distinct sets, $\mathcal{A} = \{1, 2, j, \ldots, J\}$ and $\mathcal{B} = \{1, 2, k, \ldots, K\}$. Here, the elements within each set, denoted $\mathcal{A}_j$ and $\mathcal{B}_k$, are individual text statements, long or short. The aim is to find similar pairs of text, $\mathcal{A}_j \simeq \mathcal{B}_k$, where this notation indicates these two elements are measuring a similar learning objective and would be considered a matched pair. Another way to state this is $\mathcal{A}_j \cap \mathcal{B}$ and is used to mean that text statement $\mathcal{A}_j$ is contained within $\mathcal{B}$.

In a typical alignment workshop, SMEs would take text statement $\mathcal{A}_j$ and compare it to all elements contained in $\mathcal{B}$. Making all of these pairwise comparisons is a large effort involving a total of $J \times K$ total comparisons when comparing all elements in $\mathcal{A}$ and $\mathcal{B}$. Rather than SMEs exhaustively and manually searching this entire $J \times K$ search space, it is feasible to reduce this search space. Here let the $j$th embedding in $\mathcal{A}$ be denoted as $\boldsymbol{\alpha}_{j,\mathcal{A}} \in \mathbb{R}^N$ and the $k$th embedding in $\mathcal{B}$ be denoted as $\boldsymbol{\alpha}_{k,\mathcal{B}} \in \mathbb{R}^N$.

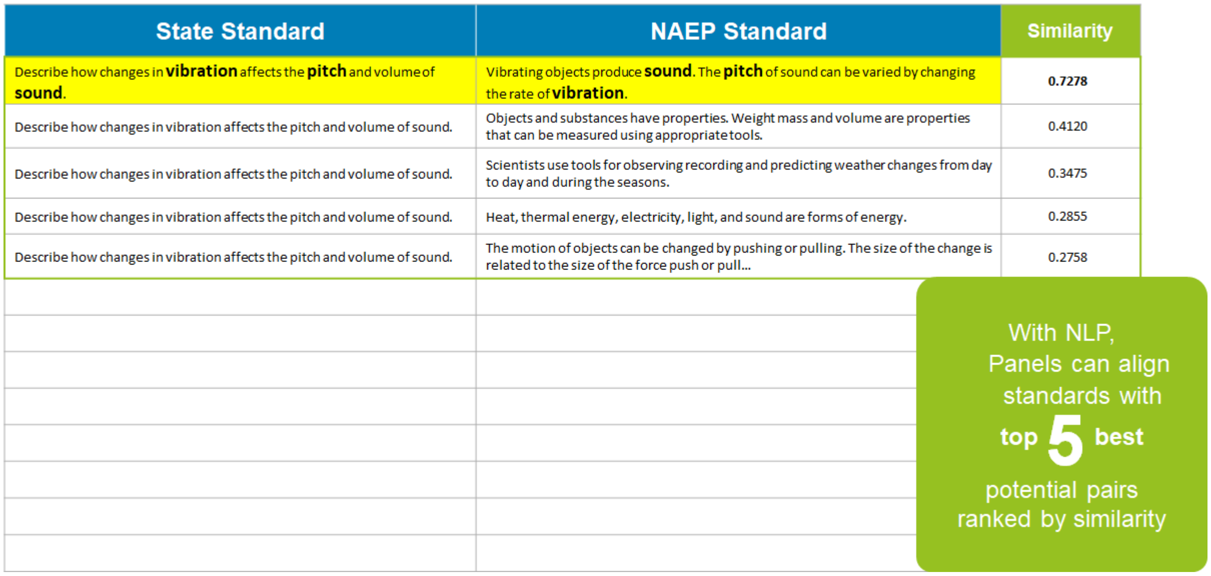
Each embedding, $\boldsymbol{\alpha}$, is a real-valued vector of length $N$ both spanning the same plane. Consequently, the angle between them is a metric describing the degree to which they are related. For this reason we can compute for each embedding pair, $\mathcal{L}_{i(\mathcal{A},\mathcal{B})} = \psi(\boldsymbol{\alpha}_{j,\mathcal{A}}, \boldsymbol{\alpha}_{k,\mathcal{B}}) \; \forall \; j, k$ where $\psi(\cdot)$ is the cosine similarity metric. Let $\boldsymbol{\mathcal{L}}_{j,k:K} = (\mathcal{L}_{1(\mathcal{A},\mathcal{B})}, \mathcal{L}_{2(\mathcal{A},\mathcal{B})}, \mathcal{L}_{p(\mathcal{A},\mathcal{B})}, \ldots, \mathcal{L}_{K(\mathcal{A},\mathcal{B})})$ be the vector of all cosine similarities between $\mathcal{A}_j$ and every other element in $\mathcal{B}$. This notation, $\boldsymbol{\mathcal{L}}_{j,k:K}$, is used to mean the cosine similarity between $\boldsymbol{\alpha}_{j,\mathcal{A}}$ and all elements in $\mathcal{B}$, hence it is always of length $K$ and there is one of these vectors for each element in $\mathcal{A}$, or $(\boldsymbol{\mathcal{L}}_{1,k:K}, \boldsymbol{\mathcal{L}}_{2,k:K}, \ldots, \boldsymbol{\mathcal{L}}_{J,k:K})$.

Once these values are obtained, then for each $\mathcal{A}_j$ sort its corresponding cosine vector, $\boldsymbol{\mathcal{L}}_{j,k:K}$, in descending order. The first element of this sorted vector has the highest cosine similarity index, suggesting it has the highest degree of semantic similarity and the last element has the smallest cosine similarity. If the cosine similarity index is useful at capturing the similarity between content standards, then we might expect the first element, or approximately so, would be the aligned or matching standard $\mathcal{A}_j$. In this case, we could ask SMEs to review only a subset of the most probable matching text pairs, $\boldsymbol{\mathcal{L}}_{j,k:K} = (\mathcal{L}_{1(\mathcal{A},\mathcal{B})}, \mathcal{L}_{2(\mathcal{A},\mathcal{B})}, \ldots, \mathcal{L}_{p(\mathcal{A},\mathcal{B})})$, letting $p$ be some positional element in the vector that is less than $K$. When, $p \ll K$ the search space is substantially reduced and human effort is minimized.

## Example

Figure (3) provides one example of this approach. In this example, the element $\mathcal{A}_1$ = "*Describe how changes in vibration affects the pitch and volume of sound.*" is compared to all text statements in the NAEP set, or $\mathcal{B}$. The column "Similarity" is the vector of cosine similarities between this standard and all other NAEP standards, or $\boldsymbol{\mathcal{L}}_{1,k:K}$ to correspond to the notation in the method section. The values here are sorted according to the cosine similarity in descending order so that the pairs sharing the highest degree of similarity are presented first. We observe that $\mathcal{A}_1$ has a cosine similarity of .73 with the statement "*Vibrating objects produce sound. The pitch of sound can be varied by changing the rate of vibration.*" and a cosine similarity of .29 with the statement, "*Heat, thermal energy, electricity, light, and sound are forms of energy.*"

The idea is that the pairs with the largest cosine similarity form the most probable subset of matching pairs. But the question is how many of those pairs need to be in the subset for experts to review? Should we show SMEs text pairs with cosine similarities above a specific value or should humans review the top $p$ ordered pairs of text statements? We further explore this in the analysis section.



| State Standard | NAEP Standard | Similarity |
|---|---|---|
| Describe how changes in **vibration** affects the **pitch** and volume of **sound**. | Vibrating objects produce **sound**. The **pitch** of sound can be varied by changing the rate of **vibration**. | **0.7278** |
| Describe how changes in vibration affects the pitch and volume of sound. | Objects and substances have properties. Weight mass and volume are properties that can be measured using appropriate tools. | 0.4120 |
| Describe how changes in vibration affects the pitch and volume of sound. | Scientists use tools for observing recording and predicting weather changes from day to day and during the seasons. | 0.3475 |
| Describe how changes in vibration affects the pitch and volume of sound. | Heat, thermal energy, electricity, light, and sound are forms of energy. | 0.2855 |
| Describe how changes in vibration affects the pitch and volume of sound. | The motion of objects can be changed by pushing or pulling. The size of the change is related to the size of the force push or pull... | 0.2758 |

With NLP,
Panels can align
standards with
**top 5 best**
potential pairs
ranked by similarity

Figure 3: *Reduced Set of Comparisons*

# Data

Our work makes use of a labeled data set where groups of SMEs reviewed individual state standards across three grades (4, 8, and 12) and found the corresponding NAEP standard. In this work, SMEs labeled the data as 0 = "not aligned", 1 = "partially aligned", 2 = "fully aligned". Because we have labeled data, we treat the human-assigned match as the "true match". Table (1) provides an overview of the data used for this work.

| Grade | No. of States | No. State Standard | No. NAEP Standard | No. Potential Pairs |
|---|---|---|---|---|
| 4 | 33 | 770 | 33 | 25,410 |
| 8 | 33 | 1,833 | 43 | 78,819 |
| 12 | 33 | 2,247 | 49 | 110,103 |

Table 1: Labeled Data

# Analysis

We began by computing text embeddings for all pairs of text in our data using four different open-source language models including all-mpnet-base-v2, all-distilroberta-v1, all-MiniLM-L6-v2, multi-qa-MiniLM-L6-cos-v1. We did initially experiment with others including Vicuna and T5 variants, but our work with those was computationally expensive and our results did not suggest they offered any benefit over other LLMs used. Once the embeddings were available, we computed pairwise cosine similarities between all pairs as in $\mathcal{L}_{j,k:K}\ \forall j$.

Importantly, we did not "fine-tune" any of these LLMs using our labeled data, we use them "out-of-the-box" and used the labels to determine how accurate our proposed method is to support the human reviews. Specifically, once the text statements are sorted in descending order using the cosine similarity, we find the proportion of time that the true match as determined by SMEs was found in the top $p$ ordered statements.

For example, suppose we had 10 text statements in $\mathcal{A}$ and 50 statements in $\mathcal{B}$. Assume that in eight cases the true match is within the top five sorted pairs and the other two cases are within the top 10. We would say that 80% of the time, the true match is in the top five and 100% of the time it is within the top 10. The implication is that the text processing as described in the methodology section could occur before a content mapping workshop and humans could be asked to review only the top five text statements instead of reviewing all 50 statements for each pair. This would reduce the level of effort from reviewing $5 \times 50 = 250$ pairs of text down to only $5 \times 5 = 25$ pairs of text.

Of course, in two of those pairs, the true match might fall outside the top five. Our proposed approach would provide experts with all text pairs ordered by their cosine similarity and ask them to review the top five to find a match. But if that match is not within the top five, keep reviewing potential pairs.

The proposed approach may be helpful in addressing additional complexities that can arise during content alignment. First, one cannot assume that each standard in one set indeed has a corresponding standard in the other set. There may be cases for which the SMEs determine that a given state standard is unique and there is no matching or parallel NAEP standard, for example. Our method may provide a convenient way for SMEs to more quickly conclude that there is no equivalent standard because they would be reviewing the most likely rank-ordered candidates first. Second, different bodies of standards are often not written to the same level of specificity. For a simple example, a mathematics standard might say “Student can perform basic arithmetic operations (adding, subtracting, multiplying, dividing) with whole numbers.” The comparison set of standards may include four separate standards that capture the same idea, one each for adding, subtracting, multiplying, dividing. This asymmetry can occur in both directions between the two bodies of standards.

# Results

Table (2) provides descriptive statistics for cosine similarity values from each language model for each grade level, both overall (i.e., all possible pairs) and for only those standard pairs that SMEs judged as the ”true matches.”

For each grade level (i.e., grades 4, 8, and 12), we first calculated an overall accuracy metric that captures the probability of the true match as labeled by humans appearing among the top $p$ highest-cosine pairs identified by the models, aggregating over the 33 states. In other words, we calculated how often the most similar pairs of standards according to the models were consistent with human judgment. Doing so directly informs the application of the tool, as it provides information about the rank-order value of the true matches among all possible matches, which in turn indicates the minimum set size of high-cosine standard pairs that SMEs would need to consider. This information is critical to our goal of reducing the search space for human raters by reducing the number of viable pairs they must consider when making judgments about alignment.

Next, we calculated accuracy for SMEs judgments within four possible conditions of SME consistency and degree of alignment: (a) SMEs *unanimously* judged standards to be *fully* aligned, (b) SMEs were originally *split* in their judgments but reached a consensus that standards were

| Model | Grade 4 | Grade 8 | Grade 12 |
|---|---|---|---|
| **Overall - *M*(*SD*)** | | | |
| all_MiniLM_L6_v2 | .18(.14) | .19(.13) | .16(.13) |
| all_distilroberta_v1 | .17(.13) | .17(.12) | .14(.12) |
| all_mpnet_base_v2 | .21(.13) | .21(.12) | .18(.12) |
| multi_qa_MiniLM_L6_cos_v1 | .16(.14) | .17(.13) | .14(.13) |
| **True Matches - *M*(*SD*)** | | | |
| all_MiniLM_L6_v2 | .51(.13) | .55(.11) | .54(.12) |
| all_distilroberta_v1 | .48(.12) | .51(.11) | .52(.12) |
| all_mpnet_base_v2 | .54(.12) | .55(.11) | .56(.11) |
| multi_qa_MiniLM_L6_cos_v1 | .49(.14) | .51(.10) | .51(.13) |

Table 2: Descriptive Statistics

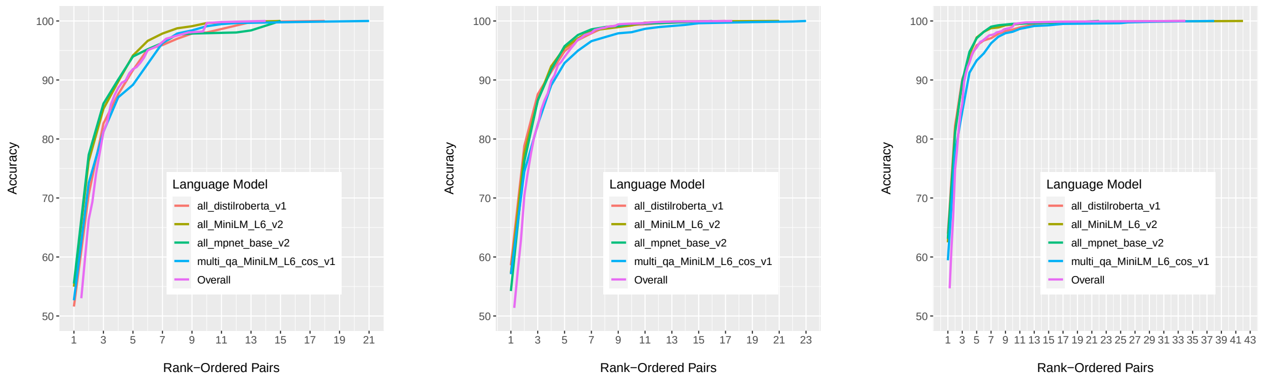
*fully* aligned, (c) SMEs *unanimously* judged standards to be *partially* aligned, and (d) SMEs were originally *split* in their judgments but reached a consensus that standards were *partially* aligned. These results provide information about the extent to which cosine similarity is sensitive to SME judgment regarding whether items are partially vs. fully aligned.

Finally, we calculated accuracy for each state that included at least 30 standards in the alignment study (*n* = 8 for grade 4, *n* = 16 for grade 8, *n* = 12 for grade 12) to examine the extent to which the accuracy results were consistent across states.

## Overall Accuracy

For grade 4, there were 33 NAEP standards that were potential candidates for alignment with each of 770 state standards across 33 states. Figure (4a) shows that the true match as designated by humans was the single highest-cosine pair 52%-56% of the time across the four language models. The true match appeared within the top two highest-cosine pairs between 71%-77% of the time. The true match appeared within the top five highest-cosine pairs between 89%-94% of the time. In other words, the NLP method was approximately 94% accurate at capturing the true match within the top five highest-cosine pairs. The four language models performed similarly in terms of accuracy, although all-miniLM-L6-v2 and all-mpnet-base-b2 appeared to slightly outperform all-distilroberta-v1 by 3% and multi-qa-MiniLM-L6-cos-v1 by 5% accuracy. Additionally, the single worst case (i.e., the worst rank-order value of the true match) fell between the 15th and 21st ranked pairs across models.

For grade 8, there were 43 NAEP standards that were potential candidates for alignment with each of 1,833 state standards across 33 states. Figure (4b) shows that the true match was the single highest-cosine pair 54%-58% of the time across the four language models. Similar to the grade 4 results, the NLP method reaches 93%-96% accuracy at capturing the true match within the top five highest-cosine pairs. Likewise, the four language models performed similarly in terms of accuracy, although multi-qa-MiniLM-L6-cos-v1 slightly underperformed compared to the other language models, showing 93% accuracy at capturing the true match within the top five most similar pairs, whereas the other models showed 95-96% accuracy. The accuracy of the model at classifying the single top-ranked pair as the true match fell between 54% and 59% across models.



(a) *Grade 4*

(b) *Grade 8*

(c) *Grade 12*

Figure 4: Overall Accuracy Results

The single worst case (i.e., the worst rank-order value of the true match) fell between the 16th and 23rd-ranked pairs across models.

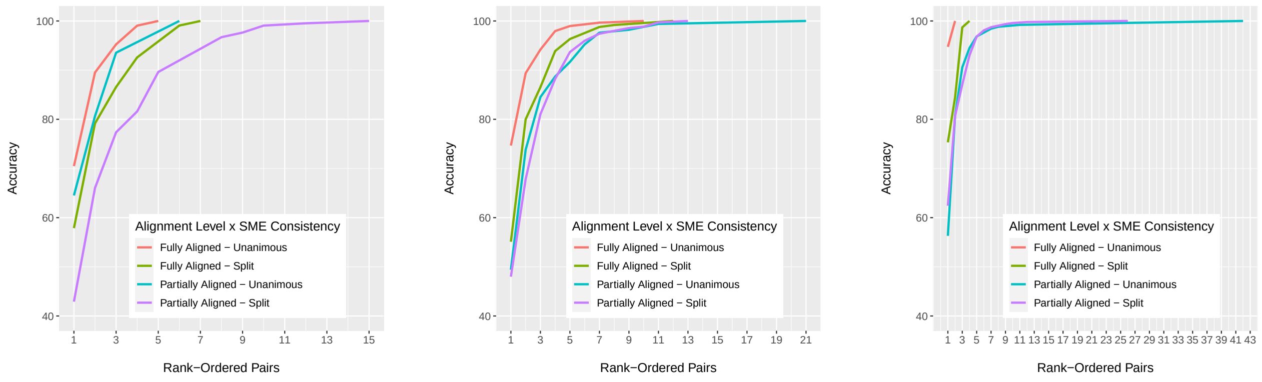
For grade 12, there were 49 NAEP standards that were potential candidates for alignment with each of 2,247 state standards across 33 states. Figure (4c) shows that the true match was the single highest-cosine pair 59%-64% of the time across the four language models. The NLP method reaches 93%-97% accuracy at capturing the true match within the top five highest-cosine pairs. As before, the four language models performed similarly in terms of accuracy, although multi-qa-MiniLM-L6-cos-v1 slightly underperformed compared to all-mpnet-base-v2 and all-MiniLM-L5-v2, showing 93% accuracy at capturing the true match within the top five most similar pairs, whereas the other models showed 95%-96% accuracy. The accuracy of the model at classifying the single top ranked pair as the true match fell between 54% and 59% across models. The single worst case (i.e., the worst rank-order value of the true match) fell between the 16th and 23rd ranked pairs across models.

## Accuracy By SME Consistency and Degree of Alignment

For grade 4, Figure (5a) shows that accuracy was highest for cases when SMEs unanimously judged that standards were fully aligned. For such cases, the true match appeared within the top five pairs 100% of the time. Accuracy was lowest for cases when SMEs disagreed but reached a consensus that standards were only partially aligned. For such cases, the true match appeared within the top five pairs only 90% of the time. Accuracy for cases when SMEs were unanimous about partial alignment and for cases when SMEs were split about full alignment fell in between.

For grade 8, the results followed a similar pattern. Figure (5b) shows that accuracy was highest for cases when SMEs unanimously judged that standards were fully aligned (i.e., the true match fell within the top five pairs 99% of the time). Accuracy was lowest for cases when SMEs disagreed but reached a consensus that standards were only partially aligned (i.e., the true match fell within the top five pairs 92% of the time).

Finally, for grade 12, Figure (5b) shows that accuracy was again highest for cases when SMEs unanimously judged that standards were fully aligned (i.e., the true match was the highest-cosine pair 95% of the time and fell within the top two pairs 100% of the time), whereas accuracy was



(a) *Grade 4*

(b) *Grade 8*

(c) *Grade 12*

Figure 5: Overall Accuracy Results

Note: Accuracy is based on rank_all_mpnet_base_v2.

lowest when SMEs judged standards only partially aligned, regardless of unanimity. However, for these cases, the true match still appeared among the top five pairs 96% of the time).

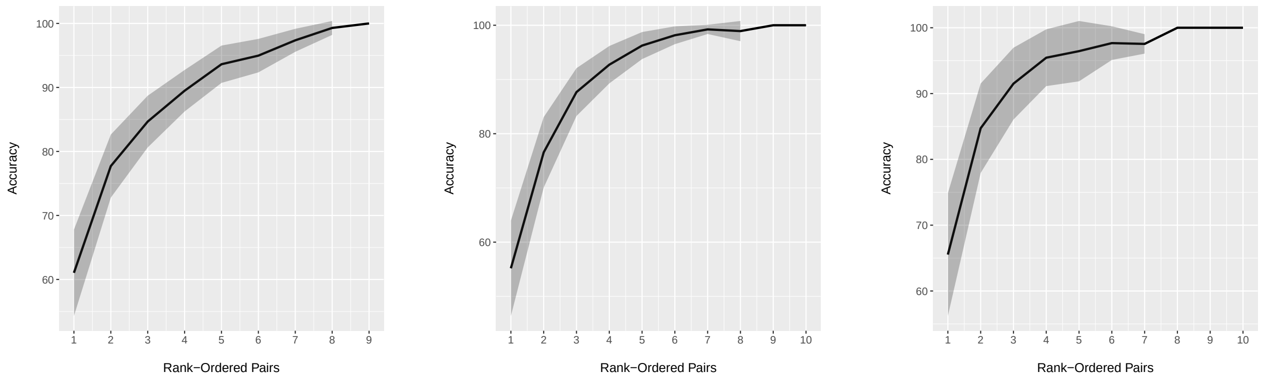
## Accuracy Across States

For each plot in (6), the black line represents the mean accuracy for rank_all_mpnet_base_v2 across each state that consisted of at least 30 standards. The upper and lower bounds of the gray band represent the 95% confidence interval (CI). Because the 95% CIs were derived from aggregating data across states, they provide an estimate of the consistency of the accuracy results across states. For grade 4 (6a), the true matches fell within the top five highest-cosine pairs between approximately 92% and 97% of the time across states. For grade 8 (6b), the true matches fell within the top five pairs between approximately 95% and 99% of the time. Finally, for grade 12 (6c), the true matches fell within the top five pairs between approximately 92% and 100% of the time.

# Discussion

Our goal was to develop an intuitive, generalized NLP-based approach to support content mapping activities. The approach we proposed offers a straightforward way of using text embeddings to improve the efficiency of SMEs' time by substantially reducing the search space they must review to find matching pairs of text. Specifically, the approach we developed guides the alignment task by reducing the number of potential pairs that SMEs must consider to only those that are most likely to be aligned due to high semantic overlap. This is a major benefit for practitioners as it may reduce the amount of time SMEs spend on these activities, also potentially reducing cost and improving accuracy when SMEs are less prone to cognitive overload and fatigue effects.

The results generally showed the subset of pairs provided to SMEs can be reduced to the top five pairs of standards ranked by cosine similarity, in which case our data showed there is a 95% probability that the true match exists within that subset. This 95% probability of capturing the true match within five pairs generalized across grade levels and different states in our data. Still,



(a) *Grade 4* (b) *Grade 8* (c) *Grade 12*

Figure 6: Accuracy Across States

Note: Accuracy is based on rank_all_mpnet_base_v2. The black line represents mean accuracy across states. The gray bands represent the 95% CI.

the degree to which these results generalize to other types of content mapping activities (e.g., different standard domains, items-to-standards alignment) needs further review and exploration. A second major benefit of our approach is we observe the pretrained language models all behave similarly out-of-the-box for this type of activity. The correlations among the cosine similarity values produced by each of the four models ranged from $r$ = .84 - $r$ = .90). Although future work will examine the extent to which different language models produce different results across content mapping data sets, these initial results suggest that text embeddings are somewhat invariant to the LLM used and a complicated task of training a model to be used for this purpose may be unnecessary. Compared to related work in the existing literature, the current approach does not require labeled datasets, data cleaning, or model tuning to be useful for facilitating alignment, at least in the current context.

The current results also showed that the accuracy of the proposed approach was slightly higher in grades 8 and 12 compared to grade 4. One potential explanation for the increase in accuracy is that the mean cosine similarity values of the top NAEP standards for each grade 4 standard ($M$ = .145, $SD$ = .048) is lower than that for grades 8 ($M$ = .193, $SD$ = .049) and 12 ($M$ = .188, $SD$ = .056). It may be the case that the language used in the standards for grades 8 and 12 necessarily uses greater specificity to describe the more advanced concepts and phenomena than do the standards for grade 4, which would then mean that the NAEP standards with which the grade 8 and grade 12 standards were aligned would be more likely to share that high-specificity language, resulting in higher mean cosine similarity values and therefore greater accuracy.

The usefulness of the current work should be interpreted with caution in light of its limitations. Namely, our current labeled dataset consisted of alignment of only science state standards with only NAEP standards. Subsesquent work will examine the usefulness of the approach for standards belonging to different domains (e.g., mathematics, English language arts), as well as items-to-standards alignment. Doing so is critical to establish the generalizability and broad utility of the approach. Moreover, examining the approach in different contexts will help identify boundary conditions. Once boundary conditions on the usefulness of the current simple approach are identified, we will examine the extent to which various modifications and extensions can address those boundary conditions. For example, few-shot classification may be useful because it can fine-tune the language models based on very small labeled datasets (i.e., only a few labeled cases per category/standard) to be more sensitive to the linguistic features that indicate alignment in a given context and therefore more accurately predict corresponding standards.

Additionally, future work will examine the embeddings of the standards to identify the linguistic signatures of the  5% of true-match standards that the approach does not accurately identify within the top five pairs. If there are systematic differences between the accurate and inaccurate cases in terms of the variables within the embeddings, then it would be possible to identify difficult standards beforehand and use fine-tuned models for classifying them into their respective true-match standards.

Overall, the current method shows potential for assisting SMEs in conducting content alignment by greatly reducing the number of viable pairs they must consider. We opted for an approach that assisted humans and kept them ”in the loop” rather than attempting to automate the task because human judgment is the gold standard for alignment, and stakeholders are less to endorse a fully automated approach in practice. Our future work in this context must continue to refine the method and examine boundary conditions to realize broad adoption and in turn promote a more efficient and economical alignment process.

# References

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations.*

Daro, P., Hughes, G. B., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP mathematics items at grades 4 and 8 to the common core state standards (CCSS) for mathematics* (Tech. Rep.). Commissioned by the NAEP Validity Studies (NVS) Panel. Retrieved from https://www.air.org/sites/default/files/2021-06/Study-of-Alignment-NAEP-Mathematics-Items-common-core-Nov-2015.pdf

Doran, H. C., & Van Wamelen, P. B. (2010). Application of the Levenshtein distance metric for the construction of longitudinal data files. *Educational Measurement: Issues and Practice*, *29*(2), 13-23.

Dynamic Learning Maps Consortium. (2013). *Dynamic learning maps essential elements for mathematics* (Tech. Rep.). Lawrence, KS: University of Kansas. Retrieved from https://dynamiclearningmaps.org/sites/default/files/documents/Math_EEs/DLM_Essential_Elements_Math_%282013%29_v4.pdf

Khan, S., Rosaler, J., Hamer, J., & Almeida, T. (2021). Catalog: An educational content tagging system. In *Edm.*

Mitzel, H., Lewis, D., Patz, R., & Green, D. (2001). The Bookmark procedure: Psychological perspectives. In (p. 249-281).

Neidorf, T., Stephens, M., Lasseter, A., Gattis, K., Arora, A., Wang, Y., ... Holmes, J. (2016). *Comparison between the next generation science standards (NGSS) and the national assessment of educational progress (NAEP) frameworks in science, technology and engineering literacy, and mathematics.* (Tech. Rep.). Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/nationsreportcard/science

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084.*

Schneider, M. C., Agrimson, J., & Veazey, M. (2022). The relationship between item developer alignment of items to range achievement-level descriptors and item difficulty: Implications for validating intended score interpretations. *Educational Measurement: Issues and Practice*, *41*(2), 12-24.

USDOE. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process.* https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/standards-and-assessments/.

Vockley, M. (2009). *Three approaches to aligning the National Assessment of Educational Progress with state assessments, other assessments, and standards* (Tech. Rep.). Council of Chief State School Officers. Retrieved from https://www.commoncorediva.com/wp-content/uploads/2014/12/alignment_and_the_states_2009.pdf

Webb, N. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. research monograph no. 6.

WIDA. (2020). *WIDA English language development standards framework, 2020 edition: Kindergarten–grade 12* (Tech. Rep.). Board of Regents of the University of Wisconsin System. Retrieved from https://wida.wisc.edu/sites/default/files/resource/WIDA-ELD-Standards-Framework-2020.pdf

Zhou, Z., & Ostrow, K. S. (2022). Transformer-based automated content-standards alignment: A pilot study. In G. Meiselwitz et al. (Eds.), *HCI international 2022 - late breaking papers. interaction in new media, learning and games - 24th international conference on human-computer interaction, HCII 2022, virtual event, June 26 - July 1, 2022, proceedings* (Vol. 13517, pp. 525–542). Springer. Retrieved from `https://doi.org/10.1007/978-3-031-22131-6_39`