Pre-trained Transformer Models for Standard-to-Standard Alignment Studies

Hye-Jeong Choi[a], Reese Butterfuss[b], Meng Fan[a], and Emily Dickinson[a]

[a] *HumRRO*
[b] *Certiverse*

Abstract

The current study evaluated the accuracy of five pre-trained large language models (LLMs) in matching human judgment for standard-to-standard alignment study. Results demonstrated comparable performance across LLMs despite differences in scale and computational demands. Additionally, incorporating domain labels as auxiliary information did not enhance LLMs performance. These findings provide initial evidence for the viability of open-source LLMs to facilitate alignment study and offer insights into the utility of auxiliary information.

1 Introduction

Large language models (LLMs) are increasingly used in educational and psychological measurement activities. Their evolving sophistication and ability to represent deep, contextual semantics make them viable tools to support subject matter experts (SMEs) in reviewing large volumes of text-based context, such as educational standards (e.g. Butterfuss & Doran, 2025; Kim et al., 2023; Kusumawardani & Alfarozi, 2023; Zhou & Ostrow, 2022). However, little guidance exists on the effective use of LLMs in such contexts. Our goal was to compare popular, pretrained LLMs in a common measurement context (i.e., standard-to-standard alignment) to provide initial evidence on which LLMs may be particularly useful for measurement tasks that require extensive review of large bodies of text.

Alignment is a critical aspect of validity evidence for any assessment (AERA, APA, NCME, 2014). Standards-to-standards alignment is a process to examine how well two distinct sets of content standards target the same content (Neidorf et al., 2016). In general, it requires SMEs to review two sets of standards and determine alignment such that each standard in one set is evaluated against the standards in the second set until any or all standards that capture the same meaning are identified. It is a time-consuming process because it requires evaluation of potentially thousands of possible pairs of content standards. Recently, the potential for NLP and LLMs as a supporting tool in this process has been presented (e.g., Butterfuss & Doran, 2024; Zhou & Ostrow, 2022), but there is a lack of work that provides guidance on which LLMs to choose for such tasks.

This study aimed to address two research questions: (1) how do five popular pre-trained transformer models compare in standards-to-standards alignment studies? and (2) does auxiliary information (e.g., domain label) impact LLMs performance? Educational standards, typically presented as brief, abstract statements, often include examples to provide clarity and context. It is also not uncommon to have the exact same standard appear under different domains. Understanding how such auxiliary information influences LLMs performance is crucial for developing more effective automated alignment tools.

2 Methods

2.1 Data

The alignment study dataset used for the current study consisted of individual standards from 33 states and aligned each state standard to the corresponding the National Assessment of Educational Progress (NAEP) standard for grades 4, 8, and 12 for science. Each standard was classified into one of three domains: life science (LS), physical science (PS), and earth & space science (ES). The number of potential pairs ranged

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 306–311

from approximately 18,000 to 60,000. The number of standards represented within each state varied. In the original work, SMEs judged each possible pair of standards as aligned, partially aligned, or not aligned. Thus, we used the SME decision as "ground truth" for evaluating the LLMs. More details about the dataset and original alignment study can be found in the published report (Dickinson et al., 2021).

2.2 Pre-trained transformer models

We accessed the LLMs via the Hugging Face Transformers library, a popular open-source library that provides a simple and consistent way to use pre-trained models for various NLP tasks. As of 2025, the Hub hosts over 50,000 models, many of which are based on Transformer architectures.

The LLMs transform each content standard into an embedding, or numeric representation of the meaning of the text. Once every standard is transformed into an embedding, then the relations among the embeddings can be evaluated using cosine similarity. Cosine similarity is a metric used to measure how similar two vectors are irrespective of their magnitude. It calculates the cosine of the angle between two vectors, determining whether they point in roughly the same direction. Commonly used in text analysis, recommendation systems, and information retrieval. While it behaves similarly to a correlation in some contexts, cosine similarity specifically only measures directional similarity, not linear correlation or magnitude.

In this study we used the cosine similarity between every possible pair of standards that can be made from the two sets. Doing so allows us to gauge which standard pairs are more similar than others. Critically, standard pairs that share high semantic overlap (i.e., large cosine similarity values) are more likely to be aligned than standard pairs that share little semantic overlap (Butterfuss & Doran, 2025).

To calculate cosine similarity, we utilized five LLMs which are widely used, including all-distilroberta-v1, all-MiniLM-L6-v2, multi-qa-MiniLM-L6-cos-v1, all-mpnet-base-v2, and gtr-t5-large. All of these are sentence embedding models that can be used to calculate cosine similarity between texts. The mathematical formula for calculating cosine similarity remains the same across all these models. However, LLMs vary in the specific linguistic features their embeddings emphasize, and thus LLMs differ in which aspects of meaning contribute to cosine similarity values. Due to this variability, we extracted embeddings for each standard using five different popular LLMs:

- **all_distilroberta_v1 (DistilRoBERTa-v1).** It is a distilled version of the RoBERTa (Liu et al., 2019) model to cover a wide range of topics and styles. It is a smaller, more efficient model that's designed to be faster and more computationally efficient.

- **all_MiniLM_L6_v2 (MiniLM-L6-v2).** MiniLM is designed for efficiency and smaller size. It's useful for text classification, sentiment analysis, or question answering. They are particularly useful for deployment in resource-constrained environments, such as mobile devices or edge computing platforms (Wang et al., 2020).

- **multi_qa_MiniLM_L6_cos_v1 (MultiQA-MiniLM-L6)**. It is a variant of the MiniLM model that is designed for multi-question answering tasks, such as answering multiple questions about a given text passage, identifying relevant passages or sentences that answer multiple questions, and generating answers to multiple questions based on a given text passage.

- **all_mpnet_base_v2 (MPNet-Base-v2)**. A model known for its efficiency and performance on a wide range of NLP tasks, including text classification, sentiment analysis, question answering, and more. It's particularly useful when you need a model that can handle long-range dependencies and contextual relationships in text data.

- **gtr_t5_large (GTR-T5-Large)**. It is known as a powerful language model. It can be used text generation and summarization, question answering and reading comprehension, sentiment analysis and opinion mining, and language translation and machine translation.

2.3 Three approaches to set a threshold

We employed three threshold setting approaches to pair state and NAEP standards: cosine similarity value, percentile, and rank order. First, we used predetermined cosine similarity values: if the cosine similarity of two-paired standards was

greater than the predetermined cosine similarity value, we classified the state-to-NAEP standard pair as aligned. We used three different values as the predetermined value (i.e., 0.4, 0.5, 0.6).

Second, we used a percentile to set the cut score cosine similarity value. As mentioned in the previous section, cosine similarity is a measure of direction but not magnitude. Using percentile can resolve potential scaling issues across LLMs. We used three percentiles (i.e., 70, 80, 90) to obtain the threshold of cosine similarity to pair standards.

Finally, we utilized a rank-order approach to classify aligned standard pairs: if the cosine similarity of standard pairs was within the predetermined top n highest cosine similarity, we classified those standards as aligned. We used top 3, 5 and 10. After we classified each pair as either aligned or not aligned based on those criteria, we evaluated those results with SMEs judgment.

LLMs performance was evaluated using overall accuracy, recall and F1 metrics. Overall accuracy (either hit rate or precision) refers to the probability of capturing the true matches (according to human judgment) within condition. Recall measures the proportion of actual positive instances that were correctly identified by the model. It is a metric used to evaluate the completeness of a classification model's positive predictions. The F1-score is a metric that combines precision and recall into a single value, providing a balance between these two sometimes competing metrics. Precision measures the proportion of correctly identified positive instances among all instances that the model predicted as positive. It's particularly useful when you need a single measurement to evaluate a classification model's performance.

3 Results

3.1 Comparison of five pre-trained transformer models

Table 1 presents descriptive statistics for the cosine-similarity values generated by each LLM for each grade. Overall, the correlations between LLMs were high (higher than .76). In both conditions, the results indicated that the models produced cosine-similarity values that were scaled slightly differently. In particular, means of GTR-T5-Large were higher and standard deviations were smaller than other LLMs.

Table 2 summarizes comparison of LLMs to SMEs with respect to the overall accuracy, recall, and F1. Note that we used three different methods to classify pairs of standards: cosine similarity value, percentile, and rank order. First, notably, the free, open-source models fared nearly as well as the costlier, more computationally intensive model (GTR-T5-Large). Overall, the correlations between LLMs were high (higher than .76). Second, all LLMs performed similarly in capturing the true pairs with respect to F1 and accuracy. However, recall indicates percentile and rank performed much better to identify the true pairs for all five models. When cut score was used, GTR-T5-Large performed differently from other four models. That was because cosine similarity from GTR-T5-Large tended different and larger than four other models.

Table 1 Descriptive statistics of cosine similarity by LLMs for each grade

	MPNet	Distil RoBERT	Mini LM	GTR-T5	Multi QA
Grade 4 (N=18,744)					
Mean	.22	.18	.19	.55	.17
STD	.14	.14	.14	.07	.14
Grade 8 (N=55,857)					
Mean	.22	.18	.20	.56	.18
STD	.13	.13	.14	.06	.14
Grade 12 (N=59,829)					
Mean	.20	.16	.18	.55	.16
STD	.13	.13	.14	.06	.13

Note. MPNet = MPNet-Base-v2; DistilRoBERT = DistilRoBERTa-v1; MiniLM = MiniLM-L6-v2; GTR-T5 = GTR-T5-Large; MultiQA = MultiQA-MiniLM-L6

Table 2 Overall comparison LLMs results with SMEs rating under different conditions

Model	Stats	Cut Value		Percentile		Rank	
		M	STD	M	STD	M	STD
MPNet-Base-v2	F1	.24	.21	.29	.13	.35	.13
	Recall	.22	.26	.91	.11	.94	.05
	Accuracy	.98	.01	.86	.10	.89	.07
Distil RoBERTa-v1	F1	.20	.21	.29	.13	.35	.13
	Recall	.16	.21	.90	.11	.94	.06
	Accuracy	.98	.01	.86	.10	.90	.07
MiniLM-L6-v2	F1	.24	.21	.29	.13	.35	.13
	Recall	.20	.24	.90	.12	.94	.06
	Accuracy	.98	.01	.86	.10	.90	.07
GTR-T5-Large	F1	.19	.19	.29	.14	.35	.14
	Recall	.53	.44	.91	.10	.95	.04
	Accuracy	.79	.30	.86	.10	.89	.07
MultiQA-MiniLM-L6	F1	.21	.19	.27	.12	.34	.13
	Recall	.16	.20	.87	.13	.92	.06
	Accuracy	.98	.00	.85	.10	.90	.07

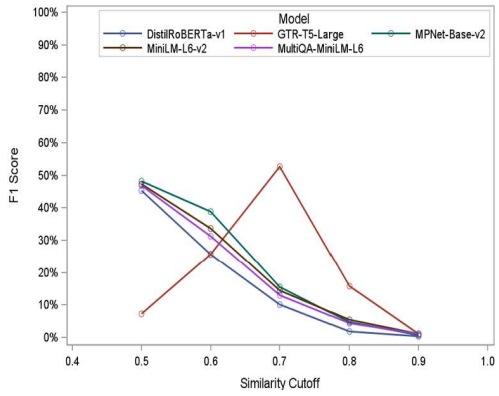

Figure 1 F1 score trend of LLMs across cosine similarity points (Grade 4)

Table 4 Comparison of LLMs at different ranks (Grade 4, N=18,744)

Model	Rank	Accuracy	Recall	F1
MPNet-Base-v2	3	.95	.87	.50
MPNet-Base-v2	5	.90	.95	.34
MPNet-Base-v2	10	.76	.99	.19
DistilRoBERTa-v1	3	.95	.83	.49
DistilRoBERTa-v1	5	.90	.93	.34
DistilRoBERTa-v1	10	.76	.99	.19
MiniLM-L6-v2	3	.95	.85	.49
MiniLM-L6-v2	5	.90	.94	.34
MiniLM-L6-v2	10	.76	1.00	.19
GTR-T5-Large	3	.95	.91	.51
GTR-T5-Large	5	.90	.97	.35
GTR-T5-Large	10	.76	1.00	.19
MultiQA-MiniLM-L6	3	.95	.83	.49
MultiQA-MiniLM-L6	5	.90	.91	.34
MultiQA-MiniLM-L6	10	.76	.99	.19

Next, we will present the results focusing on Grade 4 as the results with other grades were similar. Figure 1 depicts F1 across several cosine similarity points from .50 to .90 for all five language models for Grade 4. GTR-T5-Large performed best when the cosine similarity was set at .70 whereas four languages models performed best when the cosine similarity was set at .50. Table 4 presents the comparison of LLMs with different ranks. As expected, LLMs captured the true pair with lower ranks (rank=3). However, recall indicates LLMs captured the true pair with rank=10. In other words, the NAEP standards, with which the state standard is aligned, appear among the top ten pairs rank-ordered by cosine similarity. Table 3 shows the comparison of LLMs with different

Table 3 Comparison of LLMs at different percentiles (Grade 4, N=18,744)

Model	%ile	Accuracy	Recall	F1
MPNet-Base-v2	70	.73	1.00	.17
MPNet-Base-v2	80	.83	.97	.24
MPNet-Base-v2	90	.92	.83	.37
DistilRoBERTa-v1	70	.73	.99	.17
DistilRoBERTa-v1	80	.83	.94	.24
DistilRoBERTa-v1	90	.92	.83	.37
MiniLM-L6-v2	70	.73	.99	.17
MiniLM-L6-v2	80	.83	.95	.24
MiniLM-L6-v2	90	.92	.84	.37
GTR-T5-Large	70	.73	.99	.17
GTR-T5-Large	80	.83	.97	.24
GTR-T5-Large	90	.92	.89	.40
MultiQA-MiniLM-L6	70	.73	.98	.17
MultiQA-MiniLM-L6	80	.83	.95	.24
MultiQA-MiniLM-L6	90	.92	.79	.35

percentile. Again, LLMs performed similarly: the higher percentile, the better in capturing the true pairs with respect to accuracy whereas the lower percentile, the better in terms of recall. Both accuracy and recall indicate LLMs with either 70 percentile or rank order 10 well captured the true pairs. In other words, the NAEP standards, with which the state standard is aligned, appear among the top ten or even top five pairs rank-ordered or 70 or 80 percentiles by cosine similarity.

3.2 Effects of domain information effect on cosine similarity

Table 7 presents cosine similarity distributions when domain labels were added for each grade. Note that the N counts for all grades were slightly larger than the N counts in Table 1. This was because the same standard was assigned into different domains. The descriptives were similar with ones in Table 1; however, those values were slightly lower. Also, the correlations between cosine similarity measures for standard pairs with domain were similar with ones without domain, ranging from .75 to .92.

Next, we compare how LLMs performed to capture the true pairs compared with cosine similarity without domain. Again, we present the results for Grade 4 as the results with other grades were similar. Figure 2 depicts F1 scores across several cosine similarity points from .50 to .90 for all five language models for Grade 4. The results show a similar pattern with Figure 1; with respect

Table 7 Descriptive statistics of cosine similarity with domain by LLMs for each grade

	MPNet	Distil RoBERT	Mini LM	GTR -T5	Multi QA
Grade 4 (N=18,846)					
Mean	.21	.17	.17	.55	.15
STD	.13	.13	.14	.06	.13
Grade 8 (N=56,932)					
Mean	.22	.17	.19	.56	.16
STD	.12	.13	.13	.06	.13
Grade 12 (N=59,927)					
Mean	.20	.15	.17	.55	.14
STD	.13	.13	.13	.06	.13

Note. MPNet = MPNet-Base-v2; DistilRoBERT = DistilRoBERTa-v1; MiniLM = MiniLM-L6-v2; GTR-T5 = GTR-T5-Large; MultiQA = MultiQA-MiniLM-L6

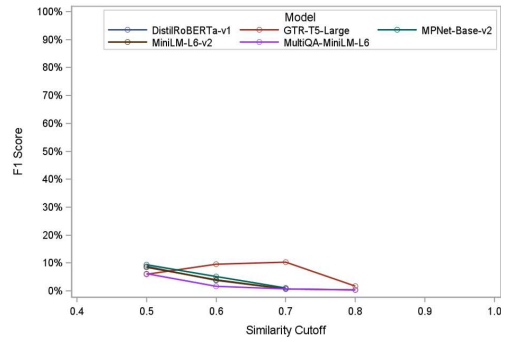

Figure 2 F1 Score by cosine similarity cut for each language model with domain (Grade 4)

to F1, GTR-T5-Large outperformed and performed the best when the cosine similarity was set at .70. Overall, however, all the five LLMs performed slightly worse with domain labels.

Table 6 and 7 show that domain information improved accuracy but reduced recall and F1. Adding domain labels did not enhance overall model performance.

4 Summary and discussion

The results of the current study indicated that scaling differences among LLMs in raw cosine similarity values meant that using a raw cosine value threshold may not be feasible, particularly when comparing multiple LLMs. Overall, when percentile or rank order was used, the results suggest that the five LLMs performed comparably with respect to accuracy for standards-to-standards alignment of science content standards. Specifically, the models were generally 90% accurate at capturing the "true matches" according to human judgment above the 90 percentile or within the top five highest-cosine pairs. Put another way, for a given state standard, the SME-aligned NAEP standard had appeared either the 90 percentile or among the top five cosine similarity pairs.

Using Grade 4 from our real-world alignment study as an example, the current method would reduce the number of pairs that SMEs must compare from nearly 18,000 pairs to around 2,840 pairs. Moreover, the current findings suggest that

Table 6 Comparison of LLMs with domain at different ranks (Grade 4, N=18,876)

Model	Rank	Accuracy	Recall	F1
MPNet-Base-v2	3	.97	.23	.28
MPNet-Base-v2	5	.95	.31	.25
MPNet-Base-v2	10	.88	.44	.17
DistilRoBERTa-v1	3	.97	.21	.26
DistilRoBERTa-v1	5	.95	.28	.24
DistilRoBERTa-v1	10	.88	.43	.17
MiniLM-L6-v2	3	.97	.23	.27
MiniLM-L6-v2	5	.95	.30	.25
MiniLM-L6-v2	10	.88	.44	.17
GTR-T5-Large	3	.97	.26	.30
GTR-T5-Large	5	.95	.33	.25
GTR-T5-Large	10	.87	.47	.17
MultiQA-MiniLM-L6	3	.97	.20	.25
MultiQA-MiniLM-L6	5	.95	.29	.24
MultiQA-MiniLM-L6	10	.88	.45	.17

Table 5 Comparison of LLMs with domain at different percentiles (Grade 4, N=18,876)

Model	%ile	Accuracy	Recall	F1
MPNet-Base-v2	70	.70	.44	.08
MPNet-Base-v2	80	.79	.35	.09
MPNet-Base-v2	90	.88	.23	.10
DistilRoBERTa-v1	70	.70	.44	.08
DistilRoBERTa-v1	80	.79	.33	.08
DistilRoBERTa-v1	90	.88	.23	.10
MiniLM-L6-v2	70	.70	.45	.08
MiniLM-L6-v2	80	.79	.35	.09
MiniLM-L6-v2	90	.88	.23	.10
GTR-T5-Large	70	.70	.46	.08
GTR-T5-Large	80	.79	.38	.09
GTR-T5-Large	90	.89	.25	.11
MultiQA-MiniLM-L6	70	.70	.44	.08
MultiQA-MiniLM-L6	80	.79	.33	.08
MultiQA-MiniLM-L6	90	.88	.21	.09

any of the popular, open-source LLMs we compared may yield such benefits. Thus, for contexts similar to those in the current work, researchers and practitioners may be well-suited to choose any of the models we evaluated given their comparable performance. Also, the current findings highlight a potentially enormous efficiency increase by dramatically reducing the number of pairs SMEs must consider via economical LLMs and a relatively simple percentile or rank-order approach using cosine-similarity.

Unfortunately, adding "domain" to the standard did not improve LLMs performance to capture the true matches. Subsequent work in this area is needed to examine the added benefits of including more contextual information for the LLMs when extracting embeddings for each standard (i.e., content domain descriptions), as well as the conditions under which it is useful to include or omit accessory information that some content standards include, such as exemplary information or explanatory information. The results did not show LLMs performance were substantially impacted by grade.

Critically, the current LLM approach does not *replace* humans in making alignment decisions. Instead, the method provides a simple, economical way to support SMEs in making alignment decisions more efficiently by leveraging an organizational structure based on semantic similarity and constraining the number of viable pairs that must be considered. Overall, the current study represents a judicious, human-centered use of AI in a laborious routine measurement task.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Butterfuss, R., & Doran, E. (2025). Advancing measurement with large language models: Implications for content review. *Journal of Educational Measurement, 62*(1), 45–63.

Dickinson, E. R., Gribben, M., Schultz, S. R., Spratto, E., & Woods, A. (2021). *Comparative analysis of the NAEP science framework and state science standards* (Tech. Rep.). National Assessment Governing Board. Retrieved from https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/science/NAEP-Science-Standards-Review-Final-Report-508.pdf

Kim, D. E., Hong, C., & Kim, W. H. (2023, July). Efficient Transformer-based Knowledge Tracing for a Personalized Language Education Application. In *Proceedings of the Tenth ACM Conference on Learning@ Scale* (pp. 336-340).

Kusumawardani, A., & Alfarozi, M. (2023). Exploring AI-driven tools for curriculum mapping. *International Journal of Educational Technology*, 18(2), 110–125.

Neidorf, T. S., Binkley, M., Galia, J., & Stephens, M. (2016). *Assessing alignment among curriculum, instruction, and assessments: Methodologies and findings*. OECD Publishing.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776-5788.

Zhou, Z., & Ostrow, K. S. (2022, June). Transformer-Based Automated Content-Standards Alignment: A Pilot Study. In *International Conference on Human-Computer Interaction* (pp. 525-542). Cham: Springer Nature Switzerland.