ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Leveraging NLP-based tools for constructive alignment

Coskun Kilinc, Chathu Ranaweera, Julien Ugon
Deakin University, Melbourne VIC, Australia

Andrew Cain, Charlotte Pierce
Monash University, Melbourne VIC, Australia

Constructive alignment requires that learning outcomes, teaching activities, and assessments be coherently structured. However, verifying this alignment, especially across large curricula, remains difficult at scale. In this paper, we present a Natural Language Processing (NLP)-based approach to automatically assess the alignment of Intended Learning Outcomes (ILOs), typically comparing smaller course-level objectives to broader Graduate Learning Outcomes (GLOs). Based on expert annotations, we fine-tune an NLP classifier to predict alignment with graduate learning outcomes focused on communication skills. Our results show that NLP tools can support alignment review by surfacing ambiguous phrasing and prompting expert judgement, offering a scalable and pedagogically grounded approach to curriculum quality assurance.

Keywords: constructive alignment, curriculum analytics, natural language processing, educational data mining, learning outcomes, transformer models

Introduction

Higher education is increasingly shaped by competency-based learning models, where success depends not just on what students are taught, but on what they can demonstrably do (Ali, 2018; Biggs, 1999). In this context, curriculum alignment plays a central role. It ensures that learning outcomes, teaching activities, and assessment tasks are purposefully connected to support student development (Biggs, 1996). Curriculum alignment provides a structured approach for ensuring that teaching activities and assessments support clearly defined learning outcomes. First introduced by Biggs (1996), the Constructive Alignment model positions learning outcomes as the foundation for instructional design, guiding the selection of appropriate assessments and learning activities (Ali, 2018; Biggs, 1999).

To support alignment, educators often draw on taxonomies such as Bloom's cognitive hierarchy (Bloom, 1984) and the SOLO taxonomy (Biggs& Collis, 1982), which help classify outcomes based on their complexity. However, the language of learning outcomes is frequently ambiguous, and interpretation can vary widely between instructors and contexts (Ali, 2018; Luxton-Reilly et al., 2018). In practice, outcome alignment is typically performed manually by subject matter experts. While this process allows for professional judgement, it is time-consuming, difficult to standardise, and susceptible to inconsistency and bias (Bone & Ross, 2021; Martone & Sireci, 2009). Tools such as Webb's depth-of-knowledge matrix and the Achieve protocol offer some structure, but still rely heavily on individual interpretation (Ali, 2018).

Despite its foundational importance, curriculum alignment faces mounting challenges. In practice, it must operate across nested levels of outcomes, distributed governance structures, and frequently changing program requirements. Without sustained support, alignment efforts become fragmented, dependent on manual mapping processes, inconsistent documentation, and tools that struggle to scale or adapt (Biggs et al., 2022).

In response, researchers have explored NLP methods to support alignment. Early approaches focused on measures of lexical similarity such as Term Frequency-Inverse Document Frequency (TF-IDF), the Jaccard index, or handcrafted rule-based systems (Gani et al., 2022; Mohammed & Omar, 2020). Although computationally simple, these methods struggle to capture the semantic nuance required for curriculum alignment, particularly in domains like communication, where context, modality, and audience often shape meaning.

Recent studies have turned to transformer-based models, especially sentence-pair classifiers, to assess alignment across nested outcome levels (Chor et al., 2024; Zaki et al., 2023). These models use contextual embeddings to evaluate semantic compatibility between two statements. However, most are trained on noisy or institution specific labels (Yuheng Li et al., 2022; Zhang et al., 2021) and provide limited information on rubric design, annotation consistency, or validation by educational experts.

Other work has explored topic extraction and cosine similarity using large language model embeddings to align syllabi and course descriptions (Liu et al., 2024). While promising, such methods are typically unsupervised and rarely address issues of calibration, rubric transparency, or inter-rater agreement, factors essential for educational trustworthiness (Butterfuss & Doran, 2025; Kaldaras & Haudek, 2022).

Despite these advances in NLP, automated curriculum alignment remains underdeveloped. Many existing systems rely on rule-based heuristics or models trained on noisy institutional data (Chor et al., 2024; Zaki et al., 2023), which limits their capacity for generalisation as well as pedagogical reliability.

This study addresses these gaps by evaluating whether transformer-based models can support curriculum alignment at scale, particularly in communication-related outcomes, when trained on expert-rated data. We frame the alignment task as binary sentence-pair classification given a lower-level intended learning outcome and a graduate-level communication outcome, the model predicts whether a meaningful alignment exists.

To evaluate this approach, we implemented a three-stage research process:

1. Feasibility: Initial experiments evaluated whether a transformer-based model could learn alignment patterns from existing curriculum data.
2. Validation: A new expert-annotated dataset was developed to address inconsistencies in the original labels and improve the validity of evaluation.
3. Refinement: Model architecture, training data, and evaluation procedures were iteratively improved to enhance predictive performance and alignment fidelity.

This work makes three primary contributions. First, it introduces a validated dataset of expert-rated curriculum alignments focused on communication outcomes, addressing limitations in label quality common in prior work. Second, it presents a transformer-based alignment model trained on this dataset, demonstrating strong agreement with expert consensus. Third, it offers a replicable evaluation framework for assessing alignment tools based on expert inter-rater reliability, pedagogical plausibility, and model interpretability.

Our final model achieved strong alignment with expert judgements and showed potential to distinguish linguistic features associated with high- and low-alignment outcomes. These findings suggest that NLP-based tools, when trained on expert-informed data, can provide interpretable and scalable support for curriculum mapping.

The remainder of this paper is structured as follows. Section 2 outlines our research design and evaluation framework. Section 3 presents results from three model development iterations, including annotation methodology and expert review. Section 4 discusses key challenges, design implications, and broader applicability. Section 5 concludes with a summary of findings.

Methodology

This study adopts an iterative research design to explore whether transformer-based language models can support curriculum alignment in a pedagogically meaningful way. Rather than proposing a fully automated solution, our goal is to investigate whether NLP techniques can assist educators and quality assurance teams in identifying potential misalignments between learning outcomes and graduate-level communication goals.

The research progressed through a series of design iterations, each informed by practical limitations observed in the previous stage. We began by testing model feasibility using institutional labels, then constructed a dataset rated by educational experts to improve the validity of evaluation and finally refined the model and its evaluation strategy to better reflect real-world alignment decisions. Figure 1 illustrates the proposed alignment pipeline. Educators or quality assurance teams could use this tool to identify potential misalignments early in the curriculum review process.

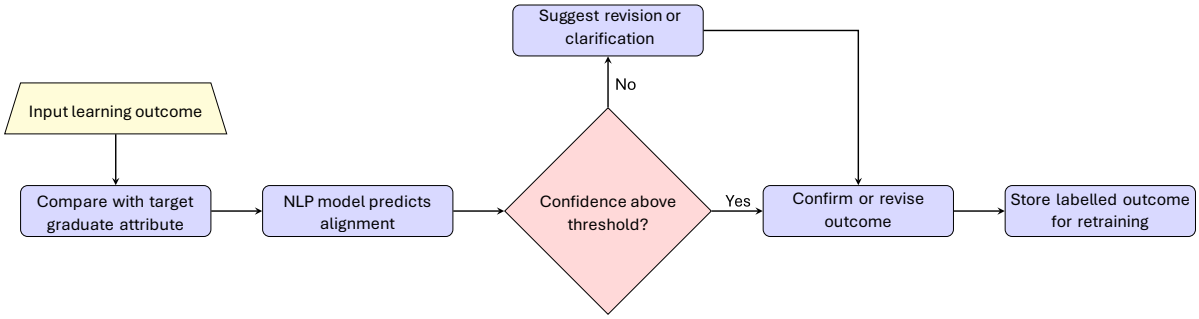

Figure 1. Proposed Curriculum Alignment Verification Pipeline

Evaluation Method

The evaluation strategy in this study evolved alongside improvements in label quality and model architecture. In the initial phase, we trained a multi-label classification model using BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019), a widely used pre-trained transformer model for natural language processing tasks such as text classification and semantic similarity, benchmarked against institutional mappings spanning multiple graduate outcomes. While this model achieved high reported accuracy, further analysis revealed that many of the underlying labels were ambiguous, inconsistent, or overly broad, limiting their usefulness as ground truth.

To address the limitations of the institutionally derived labels, many of which were ambiguous, inconsistently applied, or overly broad, we developed a new dataset rated by educational experts. This enabled a more principled and pedagogically valid evaluation process. Model predictions were assessed using standard classification metrics (accuracy, precision, recall, and F1 Score) and regression metrics (Mean Absolute Error (MAE), Mean Squared Error (MSE), Coefficient of Determination (R^2)), capturing both discrete agreement and scalar prediction quality.

Confusion matrices were used to visualise systematic misclassifications and identify patterns of over- or under alignment. Additionally, a panel of experts reviewed a curated subset of outputs to assess rubric consistency and surface edge cases where predictions diverged from consensus. These reviews informed both rubric clarification and iterative model adjustments.

Results

This section reports results across all three development iterations. Each iteration involved distinct data sources, supervision strategies, and model architectures. Iteration 1 explored feasibility using curriculum-derived labels; Iteration 2 introduced expert-rated binary labels focused on communication; and Iteration 3 refined both data and model design to maximise alignment fidelity. Table 1 summarises the objectives, data quality, and model performance metrics at each iteration. Note that the F1 score for iteration 1a is misleading, due to the polluted ground truth data used for evaluation. Iteration 1b represents the true accuracy of the model when evaluated on the expert labelled dataset.

Table 1
Summary of Model Development Iterations

Iteration	Training Dataset	Model Architecture	F1 Score	Accuracy	R^2 Score
1a	Institutional	BERT (multi-label)	0.94	–	–
1b	Institutional	BERT (multi-label)	0.52	0.389	−1.46
2	Expert	BERT / MPNet (binary)	0.74	0.791	0.10
3	Expert	DeBERTa-v3-small (binary)	0.95	0.966	0.85

- **Iteration 1** relied on curriculum-derived labels, enabling large-scale training but lacking expert validation. While initial evaluation (**Iteration 1a**) showed high performance (F1 Score: 0.94), the same model failed to generalise to expert-labelled examples (**Iteration 1b**), with F1 dropping to 0.52 and R^2 falling to −1.46. This exposed critical weaknesses in label quality.

- **Iteration 2** addressed this by shifting to expert-provided binary labels for both training and evaluation. Compared to Iteration 1b, the model demonstrated clear gains across all metrics (F1: 0.74, accuracy: 79.1%, R^2: 0.10), though performance remained limited by linguistic ambiguity and class imbalance.
- **Iteration 3** retained the expert-labelled dataset but filtered for high-consensus examples to improve label quality. It also replaced the earlier BERT-based model with a more expressive transformer architecture, DeBERTa-v3small, designed to capture richer patterns in sentence pairs. This iteration outperformed all previous models, achieving F1 0.95, accuracy 96.6%, and R^2 0.85, indicating a much stronger correlation with expert judgement and improved alignment reliability.

Table 2 provides a direct performance comparison across all iterations using a consistent set of expert-labelled alignment examples.

Table 2
Comparison of Model Performance Across Iterations

Metric	Iteration 1a	Iteration 1b	Iteration 2	Iteration 3
Precision	0.95	0.3544	0.6897	0.9440
Recall	0.93	0.9820	0.7874	0.9562
F1 Score	0.94	0.5208	0.7353	0.9501
Accuracy	–	0.3894	0.7913	0.9660
MAE	–	0.5926	0.2087	0.0340
MSE	–	0.5507	0.2087	0.0340
R2 Score	–	−1.4614	0.1028	0.8482

Iteration 1: Naive Attempt

This iteration explored the feasibility of automating curriculum alignment using transformer-based models. The goal was to determine whether a multi-label classification approach could identify alignment between local-level learning outcomes (e.g., course or unit statements) and institution-wide graduate attributes, using existing curriculum documentation as input. This baseline prototype also aimed to reveal limitations in available label quality.

A large dataset of 9,686 learning outcomes was compiled from publicly available curriculum sources, spanning 2,125 units and 32 courses. Each outcome included one or more associated graduate attributes, based on institutional mappings used for quality assurance and accreditation. Although these mappings were considered official, expert review later revealed inconsistencies in the mappings, motivating the need for a more reliable dataset in subsequent iterations.

We trained a transformer-based model to predict which graduate skills were addressed by a given learning outcome. The model was designed to allow for multiple correct answers, since a single outcome might align with more than one graduate skill. To do this, we used a common architecture where the input statement was processed by a pretrained BERT model, followed by a fully connected linear layer that estimated alignment confidence for each of the eight graduate attributes. To prevent the model from over-relying on specific patterns and improve generalisation, a dropout layer is incorporated. This layer randomly deactivates a fraction of the neurons during training, ensuring that the learned representations remain robust across different inputs.

The model was trained to recognise multiple correct alignments at once, rather than forcing a single label. Its performance was evaluated based on how accurately it identified all relevant graduate skills for each outcome.

While this approach enabled large-scale training, it relied on labels derived from institutional documentation, which were later found to be inconsistent and sometimes unreliable.

The model achieved high macro-level performance, with a macro-averaged F1 score of 0.92 and weighted accuracy of 0.93 (See table 2). Confusion matrices and alignment distribution plots showed consistent performance across graduate attributes, mirroring the distribution of labels in the training data. An expert panel reviewed a stratified sample of predictions. Discrepancies between expert judgement and institutional labels were common (see table 3), particularly for outcomes that were vague or multipurpose. These findings highlighted the need for more reliable supervision data. Subsequent iterations focused on constructing a validated dataset to address these concerns.

Table 3
Examples of Alignment Divergence in Iteration 1

ILO Text	Expert Label	Model Prediction
Discuss the concept of brand and the ways a strong brand can benefit a variety of firms operating in different contexts, including domestic and global markets.	Yes	Yes
Interpret and verbally transmit knowledge, ideas, research concepts and theories as well as the significance of research to specialist audiences.	Yes	No
Demonstrate and apply knowledge of core areas of adult development including physical, cognitive, and social development, personality, coping and resilience, and preparing for the end of life.	No	Yes
Understanding the basics of supervised and unsupervised machine learning algorithms, including their basic mathematical underpinnings, and how they can be implemented using popular libraries.	No	Yes

Iteration 2: Adding an Expert Opinion

While the first model achieved high agreement with institutional labels, expert review revealed inconsistencies between these labels and pedagogical intent. Iteration 2 addressed this issue by narrowing the alignment task to a single graduate attribute, communication, and developing a new expert-annotated dataset to support more reliable evaluation. Initial plans to capture alignment strength on a 0–3 ordinal scale were abandoned after inter-rater reliability analysis showed inconsistent use of intermediate scores. A binary schema was adopted to improve clarity and annotation consistency.

To better understand the nature of disagreement between institutional and expert labels, we reviewed a sample of learning outcomes that the institution had marked as aligned with communication, but which all three expert annotators rejected. Table 4 presents four such examples. In each case, the outcome's surface language emphasises cognitive or procedural focus without clearly signalling communicative intent.

Table 4
Examples of Institutional–Expert Disagreement

ILO Text	Expert Label	Model Prediction
Describe and integrate psychological knowledge and practice related to working with children and adolescents across contexts.	No	No
Use anatomical language to describe the various components and functions of the musculoskeletal system.	No	No
Develop and deliver a coaching plan and engage in reflective practice to assess athlete performance and adjust instruction.	No	No
Critically analyse conceptual and practical issues in the design and conduct of a qualitative research project.	No	No

To support more reliable alignment modelling, we constructed a new expert-labelled dataset focused on communication outcomes. The full dataset was annotated independently by three experts, following a multi-round calibration process to establish inter-rater consistency.

Before large-scale annotation, three rounds of Inter-Rater Reliability (IRR) were conducted to refine the rubric and establish consistency among raters. Each round involved independent annotation of a small sample, followed by consensus meetings and rubric refinement. Agreement was calculated using Fleiss' Kappa. After three rounds, inter-rater agreement exceeded accepted thresholds for educational annotation tasks, as shown in table 5. The final dataset consisted of 1,148 labelled outcomes, with ambiguous cases excluded.

Table 5
Inter-Rater Reliability Across IRR Rounds

Round	Fleiss (Humans Only)	Fleiss (With Model)[1]	Difference
1	0.698	0.737	0.039
2	0.734	0.667	–0.067
3	0.883	0.859	–0.024
2[2]	0.654	0.699	0.045
3[2]	0.876	0.894	0.017

[1] Model assistance was only introduced in Iteration 3. Values for earlier rounds reflect retrospective comparison only and did not inform dataset construction.
[2] With standout cases excluded.

We trained two language models, BERT and MPNet, to evaluate whether a given learning outcome aligned with a target communication skill. Each model received a pair of statements and was asked to judge how well they matched. We evaluated two approaches: one that made a simple yes/no decision (binary classification), and another that produced a score along a scale (regression). The binary approach proved more stable and a better fit for our expert-labelled data.

The number of learning outcomes that aligned with the communication skill was dramatically lower than those that did not, resulting in a heavy class imbalance. To account for this, we adjusted the training process to ensure the model gave appropriate attention to less common cases. We also experimented with visual tools designed to show which parts of a sentence the model focused on when making decisions. However, attention-based interpretations should be viewed cautiously, as attention weights are not a definitive proxy for model reasoning (Jain & Wallace, 2019). The final model achieved strong binary classification performance on the expert-rated development set, with results summarised in table 2.

One of the key challenges during annotation was the ambiguity surrounding how different disciplines express communication-related capabilities. Verbs such as "explain", "present", or "describe" appeared frequently across intended learning outcomes, but their function varied. In some contexts, these verbs indicated assessment modalities rather than the development of communication skills. For instance, students may be asked to "explain" a concept to demonstrate understanding, without an explicit intent to assess or foster communication competence.

This ambiguity was further complicated by disciplinary variation. Technical and professional fields often employed verbs like "document", "report", or "summarise", which may imply communicative acts yet reflect content-focused objectives. In contrast, creative and social disciplines more commonly included explicit references to audience awareness or multimodal delivery.

To investigate this further, an n-gram frequency analysis was conducted across the annotated dataset. High alignment outcomes frequently contained specific communicative markers such as "tailored to audience" or "oral presentation", while low-alignment examples were dominated by generic academic verbs. These patterns highlight a broader issue: demonstrating disciplinary knowledge typically requires some form of communication, making it difficult to distinguish between communication as a means of assessment and communication as a targeted learning objective.

Certain action verbs (e.g., "explain", "present", "design") were frequently associated with specific graduate attributes, but these alone were not reliable indicators. Contextual details made a significant difference. For example, the phrase "explain a process" is ambiguous, whereas "explain a process to a non-technical

audience" clearly signals a communication-oriented intention. Capturing this nuance proved essential for both accurate annotation and successful model training.

Iteration 3: Model Refinement

Iteration 3 aimed to improve alignment accuracy and generalisation by addressing overfitting and ambiguity issues observed in prior models. It introduced a more expressive transformer backbone and refined training dynamics to better capture communicative intent in expert-annotated outcomes. The expert-labelled dataset created in Iteration 2 was reused without modification. Table 5 provides full interrater reliability scores, including model comparisons relevant to this iteration.

In this iteration, we used a more advanced language model (DeBERTa v3 Small) that builds on earlier transformer architectures to better capture subtle patterns in sentence structure and meaning. The model was fine-tuned to perform binary classification, predicting whether a given outcome statement aligned with the target communication skill.

Table 6
Examples of misclassification in Iteration 3

ILO Text	Expert Label	Model Prediction
Effectively communicate solutions and responses to common family, social and legal problems through the application of theoretical frameworks.	No	Yes
Collaboratively design and manage microgrid solutions considering social, environmental and economic factors.	No	Yes
Demonstrate person, family and/or community centred approaches to nursing care across the lifespan in a variety of health care settings.	Yes	No
Explain the process and key characteristics of effective learning across early childhood and primary education contexts.	Yes	No

Evaluation on the full dataset yielded high performance: accuracy 0.966, F1 score 0.9501, R2 of 0.8482, and MAE of 0.034. Error analysis highlighted residual false positives from vague cues and false negatives from underspecified outcomes. Table 2 presents detailed metrics.

To assess the model's potential to support annotation reliability, its binary predictions were retrospectively included in inter-rater reliability calculations as a fifth rater. As shown in table 5, agreement increased in Round 1, decreased slightly in Rounds 2 and 3, but improved when standout disagreements were excluded. These results suggest that the model's predictions were broadly consistent with expert consensus and may enhance annotation consistency under clear rubric conditions. However, its effectiveness depends on the clarity of the input outcomes and the specificity of the alignment rubric.

Discussion

While expert annotation improved label quality, the resulting dataset was discipline skewed. Most outcomes came from professionally oriented fields like business, education, and ICT. Disciplines such as psychology, counselling, and exercise science were underrepresented, limiting the model's exposure to diverse expressions of communication. In fields like counselling, terms such as "engagement" or "rapport" imply interpersonal interaction but lack explicit audience-aware phrasing. Without sufficient examples, the model struggled to interpret their communicative function.

In contrast, disciplines with more formulaic outcome phrasing, e.g., "present to stakeholders", "communicate findings", were easier to model. This imbalance likely contributed to misclassifications in underrepresented fields. Similar issues were seen in exercise science and coaching, where outcomes often involved designing or delivering sessions. Although communication was implicit, it was rarely explicit enough for the model to detect, highlighting a legitimate ambiguity rather than a model failure. Prior research has noted similar challenges, with Martone and Sireci (2009) emphasising that vague or implicit language undermines alignment validity, and Zaki et al. (2023) showing that NLP models tend to misclassify when surface wording does not fully capture the intended learning construct.

A similar pattern occurred in coaching and exercise science contexts (see table 7). Many outcomes required students to design or deliver sessions, where communication was likely involved but not explicitly stated. In these cases, the model typically did not predict alignment, yet this may not represent a failure. Rather, it reflects a legitimate ambiguity: without clear audience cues or communicative intent in the phrasing, the tool correctly flagged these outcomes as insufficiently explicit.

Table 7
Ambiguity in Coaching and Training Outcomes: Expert vs Model Disagreement

ILO Text	Expert Label	Model Prediction
Design, deliver and critically evaluate coaching of a group training session.	No	Yes
Demonstrate and apply person-centred approaches in exercise delivery using verbal and non-verbal strategies to engage and support diverse clients across the lifespan.	Yes	No

Rather than replace human judgement, NLP-based alignment tools should augment review workflows by identifying ambiguous phrasing, prompting rubric application, and reinforcing consistent application of alignment criteria. Modular designs that integrate linguistic feature analysis, taxonomy mapping, and contextual prompts may offer more transparent support to educators involved in curriculum review.

Such systems could be integrated into two distinct phases of curriculum development: creation and governance. During the creation process, alignment tools could function as writing companions, suggesting candidate graduate attributes based on outcome phrasing, prompting clarification (e.g., “To whom is this communicated?”), or flagging potentially vague verbs. This would support authors in crafting learning outcomes that better align with institutional standards while preserving flexibility and authorial intent.

In the governance phase, these tools could serve as first-pass alignment checkers, highlighting outcomes that warrant human review. While final responsibility would remain with academic reviewers, the expectation would shift. Rather than critically evaluating every outcome from scratch, reviewers could focus on flagged items, reducing cognitive load and increasing throughput. Over time, this “human-in-the-loop” approach would reinforce consistent interpretation and application of alignment rubrics across units, programs, and faculties.

Longitudinally, continued use of these tools can strengthen their effectiveness. As more outcomes are written, reviewed, and validated, their alignment patterns can inform future model training, making suggestions more accurate and reducing the burden on human reviewers. This balance between automated support and expert judgement offers a sustainable path toward scalable and trustworthy alignment systems in higher education.

One additional risk is the potential for overfitting to the tool itself. If alignment models are used during outcome creation, educators may learn to write outcomes that superficially trigger positive alignment, without meaningfully embedding the intended capability. This shift in focus, from pedagogical clarity to performative compliance, has long been noted in constructive alignment frameworks (Biggs, 1996, 1999). To mitigate this, alignment systems must remain human-in-the-loop: their role is to prompt reflection and support quality assurance, not to replace academic judgement. Any automated feedback should be interpreted within a broader pedagogical context and validated by expert reviewers.

While this study focused exclusively on communication-related outcomes, the underlying approach is likely to extend to other graduate attributes, particularly those requiring nuanced or discipline-specific phrasing. The model architecture, annotation workflow, and evaluation framework are attribute-agnostic and can be readily adapted. Similar opportunities for scalable alignment have been demonstrated in NLP-based mapping research (Zaki et al., 2023; Chor et al., 2024), providing early evidence that our expert-informed systems could significantly enhance curriculum design and assurance processes. Although empirical validation across all attributes remains future work, these results provide strong early evidence that scalable, expert-informed alignment systems are achievable and could significantly enhance curriculum design and assurance processes.

Generalisation across institutional and cultural contexts remains an open challenge. The sentence-pair formulation does not depend on a particular curriculum structure, but subtle differences in language conventions, attribute interpretation, and disciplinary expression may affect performance. Cross-institutional

validation will be critical to assess how robust these tools are when applied beyond their original development setting, particularly under different policy frameworks or academic norms.

A key question for practical deployment is how many annotated outcomes are needed to train an effective alignment model? Our current model was trained on just over one thousand expert-rated examples, but this may still be prohibitive for smaller institutions or under-resourced disciplines. Future work should explore the minimum viable dataset size by progressively reducing the number of training examples and measuring the trade-off in performance. Establishing these thresholds would inform both the design of lightweight, context-specific models and the feasibility of institution-led dataset creation.

Despite these limitations, the potential for simplification is substantial. Alignment review is currently labour-intensive, inconsistently applied, and difficult to scale. By surfacing ambiguous phrasing, reinforcing shared criteria, and prompting expert judgement only where needed, this approach offers a pragmatic way to improve alignment quality while reducing the overall review burden.

Summary

Curriculum alignment remains a persistent challenge in competency-based education, with manual mapping processes proving labour-intensive, inconsistent, and difficult to scale. This research explored whether NLP methods could support alignment verification by modelling relationships between intended learning outcomes and graduate attributes.

Our results show that automated tools can support curriculum design, but only when grounded in high-quality, clearly defined training data. Across three model iterations, performance gains were driven not by algorithmic complexity but by improved data clarity and label consistency.

Initial models trained on institutional labels appeared effective, but failed to generalise under expert review, revealing a reliance on vague or misleading cues. Shifting to expert-validated labels in Iteration 2, and refining these further in Iteration 3, led to stronger alignment with expert judgement and more interpretable outputs.

Annotation also highlighted challenges. Outcomes lacking clear communicative indicators (such as purpose or audience) reduced reviewer agreement, suggesting that both human and automated alignment depend heavily on phrasing. The final model was better able to capture such nuances, aided by a cleaner dataset and consistent annotation.

These findings reinforce the importance of constructive alignment: learning outcomes should make communicative intent explicit. Ambiguity impedes not just automation, but also expert review and curriculum coherence. Rather than replacing judgement, this work aims to provide interpretable, data-driven tools that support educators in designing transparent, assessable outcomes at scale.

References

Ali, L. (2018). The Design of Curriculum, Assessment and Evaluation in Higher Education with Constructive Alignment. *Journal of Education and e-Learning Research*, *5*(1), 72–78. https://doi.org/10.20448/journal. 509.2018.51.72.78

Biggs, J. (1996). Enhancing teaching through constructive alignment. *High Educ*, *32*(3), 347–364. https://doi.org/ 10.1007/BF00138871

Biggs, J. (1999). What the Student Does: Teaching for enhanced learning. *Higher Education Research & Development*, *18*(1), 57–75. https://doi.org/10.1080/0729436990180105

Biggs, J., Tang, C., & Kennedy, G. (2022). *Teaching for quality learning at university* (5th ed.). McGraw-Hill Education.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning*. Academic Press.

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*.

Bone, E. K., & Ross, P. M. (2021). Rational curriculum processes: Revising learning outcomes is essential yet insufficient for a twenty-first century science curriculum. *Studies in Higher Education*, *46*(2), 394–405. https: //doi.org/10.1080/03075079.2019.1637845

Butterfuss, R., & Doran, H. (2025). An Application of Text Embeddings to Support Alignment of Educational Content Standards. *Educational Measurement*, *44*(1), 73–83. https://doi.org/10.1111/emip.12641

Chor, W. T., Goh, K. M., Lim, L. L., Lum, K. Y., & Chiew, T. H. (2024). Towards a machine learning-based constructive alignment approach for improving outcomes composure of engineering curriculum. *Educ Inf Technol*, *29*(7), 8925–8959. https://doi.org/10.1007/s10639-023-12180-y

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." arXiv:1810.04805. Preprint, arXiv, May 24, 2019. http://arxiv.org/abs/1810.04805.

Gani, M. O., Ayyasamy, R. K., Sangodiah, A., & Fui, Y. T. (2022). USTW Vs. STW: A Comparative Analysis for Exam Question Classification based on Bloom's Taxonomy. *mendel*, *28*(2), 25–40. https://doi.org/10.13164/ mendel.2022.2.025

Jain, S., & Wallace, B. C. (2019, May). Attention is not explanation. https://doi.org/10.48550/arXiv.1902.10186

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned Next Generation Science Standards performance assessments. *Front. Educ.*, *7*, 968289. https://doi.org/10.3389/ feduc.2022.968289

Liu, L., Mendoza, R. A., Martin, T. R., & Miori, V. M. (2024). Generative AI-Powered Educational Alignment: A Framework for Matching Syllabus Course Topics with Web Description. *Proceedings of the 2024 9th International Conference on Distance Education and Learning*, 340–346. https://doi.org/10.1145/3675812. 3675874

Luxton-Reilly, A., Simon, Albluwi, I., Becker, B. A., Giannakos, M., Kumar, A. N., Ott, L., Paterson, J., Scott, M. J., Sheard, J., & Szabo, C. (2018). Introductory programming: A systematic literature review. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 55–106. https://doi.org/10.1145/3293881.3295779

Martone, A., & Sireci, S. G. (2009). Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Review of Educational Research*, *79*(4), 1332–1361. https://doi.org/10.3102/0034654309341375

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec (J. Zhang, Ed.). *PLoS ONE*, *15*(3), e0230442. https://doi.org/10.1371/ journal.pone.0230442

Yuheng Li, Rakovic, M., Boon Xin Poh, Gasevic, D., & Guanliang Chen. (2022). Automatic Classification of Learning Objectives Based on Bloom's Taxonomy. *International Educational Data Mining Society*. https://doi.org/ 10.5281/ZENODO.6853191

Zaki, N., Turaev, S., Shuaib, K., Krishnan, A., & Mohamed, E. (2023). Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation. *Educ Inf Technol*, *28*(12), 16723–16742. https://doi.org/10.1007/s10639-023-11877-4

Zhang, J., Wong, C., Giacaman, N., & Luxton-Reilly, A. (2021). Automated Classification of Computing Education Questions using Bloom's Taxonomy. *Proceedings of the 23rd Australasian Computing Education Conference*, 58–65. https://doi.org/10.1145/3441636.3442305

Kilinc, C., Cain, A., Pierce, C., Ranaweera, C. & Ugon, J. (2025). Leveraging NLP-based tools for constructive alignment. In S. Barker, S. Kelly, R. McInnes & S. Dinmore (Eds.), *Future-focused: Educating in an era of continuous change*. Proceedings ASCILITE 2025. Adelaide (pp. 157-166) https://doi.org/10.65106/apubs.2025.2636

Note: All published papers are refereed, having undergone a double-blind peer-review process.

The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Kilinc, C., Cain, A., Pierce, C., Ranaweera, C. & Ugon, J. 2025