Investigating Study Choice in Flemish Schools: An Application of Multi-Target Regression and Large Language Models

Bert Vos
Ghent University, Faculty of Arts and Philosophy
Department of Linguistics
Ghent, Belgium
bert.vos@ugent.be

Felipe Kenji Nakano
KU Leuven, Campus Kulak, Department of Public Health and Primary Care
Itec, imec research group at KU Leuven
Kortrijk, Belgium
felipekenji.nakano@kuleuven.be

Celine Vens
KU Leuven, Campus Kulak, Department of Public Health and Primary Care
Itec, imec research group at KU Leuven
Kortrijk, Belgium
celine.vens@kuleuven.be

Alireza Gharahighehi
KU Leuven, Campus Kulak, Department of Public Health and Primary Care
Itec, imec research group at KU Leuven
Kortrijk, Belgium
alireza.gharahighehi@kuleuven.be

Abstract

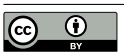
Decisions about study choice profoundly affect pupils' educational pathways and long-term career prospects, yet many still struggle to select options that align with their skills and interests. To support more personalized study choice, a serious game named "Kr8cht" was designed that elicits Flemish pupils' preferences through pairwise comparisons of activities. Each activity is linked to study domains and personality traits based on expert-annotated relevance scores. However, the current version of the game includes only a limited set of activities, restricting its ability to accommodate the diverse needs and preferences of pupils. To address this limitation, we extensively investigate multi-target regression (MTR) techniques to automatically predict domain scores for newly suggested activities, thereby reducing dependence on expert annotation. Our findings show that local and chain-based MTR methods trained on embeddings from frozen sentence-level transformers consistently outperform global MTR approaches and those trained on embeddings from static or frozen token-level transformer models. Furthermore, our experiments also reveal that augmenting the dataset with synthetic activities, which are pseudolabelled by a MTR model, increases performance with statistical significance. The source code for our experiments is available at Zenodo (https://doi.org/10.5281/zenodo.17834064).

CCS Concepts

• **Computing methodologies** → **Supervised learning by regression**; **Structured outputs**.

Keywords

study choice, multi-target regression, text embedding, educational NLP



This work is licensed under a Creative Commons Attribution 4.0 International License.
SAC '26, Thessaloniki, Greece
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2294-3/2026/03
https://doi.org/10.1145/3748522.3779972

ACM Reference Format:
Bert Vos, Felipe Kenji Nakano, Celine Vens, and Alireza Gharahighehi. 2026. Investigating Study Choice in Flemish Schools: An Application of Multi-Target Regression and Large Language Models . In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26), March 23–27, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3748522.3779972

1 Introduction

Study choice is a critical milestone that significantly influences pupils' future educational trajectories and shapes their professional careers. Despite the many studies [57, 51, 54], career choice remains a challenging task, since study choices are often not fully aligned with pupils' capabilities and preferences, especially among pupils who have not had access to quality education due to socio-economic disadvantages [58]. It is therefore essential to accurately capture pupils' interests in order to support informed study choices. To better support pupils and to personalize the study choice process, a serious game called "kr8cht"[1], a play on the Dutch word for "strength", was developed and piloted in several schools across the Flemish region of Belgium. In the course of the game, as shown in Figure 1, pupils are presented with a sequence of activity pairs, which they compare to express their preferences. For instance, "would you prefer selling waffles at school or reading poetry"?

Each activity in the game is associated with a score for each of eight study domains[2] of the Flemish secondary education system and six personality traits following Holland's RIASEC model[3] [30]. By comparing multiple pairs of activities, this game provides evidence-based personalized recommendations for study domains based on these comparative judgments.

The main issue in the current game is the limited number of activities available in the game (only 96), which restricts diversity and limits the exploration of pupils' preferences. To tackle the issue, we propose using machine learning to map the activity descriptions

[1] https://kr8cht.be/
[2] Language & Culture, STEM, Art & Creation, Agriculture & Horticulture, Economics & Organization, Sports, Society & Welfare, Food & Hospitality
[3] Realistic, Investigative, Artistic, Social, Enterprising, Conventional



Figure 1: Screenshot from Kr8cht game for an activity pair.

(in Dutch) to study domain scores, thus enabling the insertion of new activities created by the teachers or pupils themselves without the need to rely on an expert panel to provide scores for these activities. Since the target scores consist of 14 numerical outputs (8 domain and 6 personality trait scores), we employ multi-target regression (MTR) methods, supervised machine learning methods that enable the simultaneous prediction of multiple continuous variables [5]. As features for the MTR methods, we employ the text embeddings extracted from several transformer models capable of handling text in Dutch [19, 59, 13]. Furthermore, considering the rather limited availability of data, we aim to investigate whether augmenting the training data with synthetically generated activities using large language models (LLMs) can enhance the predictive performance of MTR methods in this context.

To the best of our knowledge, this is the first work to apply multi-target regression to Dutch text input on a small-sized dataset. Precisely, we investigate the following research questions: (i) Which combination of MTR model and embeddings is the most suitable for predicting scores for newly proposed activities? (ii) Does augmenting training data with synthetically generated activities using LLMs improve the predictive performance of MTR models?

2 Background

2.1 Multi-target Regression

Given $X \in \mathbb{R}^{m \times n}$ as input space and $Y \in \mathbb{R}^{m \times k}$ as target space, where m is the number of training examples, n is the number of features, and k is the number of targets, an MTR method finds a mapping function f_{MTR} between X and Y. Similar to other multi-output problems [5], MTR is typically addressed using two types of approaches: local and global.

2.1.1 Local multi-target regression. In local MTR approaches, a single-target regressor is trained for each target independently, $\mathbf{Y_i} \leftarrow f_{MTR_i}(\mathbf{X})$, where i is the ith target in Y. Therefore, for k targets, k independent MTR models are built [5]. Since each target is modeled independently, the problem is reduced to multiple single-output regressions, which allows the usage of well-established models from the literature. However, as a disadvantage, the correlations among the targets are not exploited.

As a possible solution, Read [47] proposed to extend the feature set of the local regressors by incorporating the output of previous models, thus forming a chain of regressors. Formally, for the i-th target, the regressor is defined as $\mathbf{Y_i} \leftarrow f_{MTR_i}([\mathbf{X}, \hat{\mathbf{Y}}_{1:i-1}])$, where $\hat{\mathbf{Y}}_{1:i-1}$ denotes the predicted values of the previous $i-1$ targets in the chain. Similar to local approaches, for each target, a separate single-target regressor is needed, but with enhanced feature space.

As shown in [47], the order of the regressors plays a major role in performance. Consequently, a wide range of methods have been proposed, which heuristically attempt to find which targets are correlated to the others [47]. For instance, Mastelini [41] proposed to concatenate the regressors only if the target is identified as a relevant feature, using feature relevance. Nevertheless, a simple combination of several randomly sorted chains, also referred to as ensemble of chain regressors, often leads to competitive results [52].

2.1.2 Global multi-target regression. Alternatively, the global approach, also known as algorithm adaptation, aims to learn a single model that jointly predicts all targets simultaneously. In practice, traditional models, such as decision trees [36, 43] or neural networks [34], are adapted to accommodate all targets.

Kocev [36] proposed to use an ensemble of predictive clustering trees, where the node split heuristic consists of the mean variance reduction. Also using decision trees, the work of Nakano [43] proposed a deep-forest method which uses the structure of the trees to perform feature augmentation. Despite the extra complexity, the method proposed by Nakano proved effective in the context of MTR. Alternatively, the work of Illiadis [34] proposed a dual branch network where both features and annotations are encoded as input for the network.

Formally, global methods are defined as: $\mathbf{Y} \leftarrow f_{MTR_{global}}(\mathbf{X})$. This approach enables the model to fully exploit correlations between targets during training and prediction, making them particularly effective when the targets exhibit strong mutual dependencies.

2.2 Text Embedding

Text embeddings aim to convert unstructured free-text input into fixed-size vectors that capture the semantic meaning of the text [7]. In this study, we explore two primary categories of embedding approaches: static word embeddings and embeddings from pre-trained transformer-based models.

2.2.1 Static text embeddings. Prior to the emergence of transformer architectures, text embeddings were typically derived by aggregating the embeddings of individual words or tokens within a given text. These embeddings are called static since they do not capture text context in the embeddings. Since the input data in this study is in Dutch, we focus on two widely used Dutch-language static embedding models: Word2Vec [55] and fastText [26]. Word2Vec uses shallow neural networks, primarily the Continuous Bag-of-Words (CBOW) and Skip-gram models, at the word level, whereas FastText represents words as a bag of character n-grams, enabling it to capture richer morphological information. While Word2Vec embeds

entire words, FastText embeds tokens by aggregating subword representations, making it more robust to rare and out-of-vocabulary words. The resulting vectors are then aggregated using mean pooling to obtain a single text-level embedding.

2.2.2 Frozen transformers. Transformer-based models offer a more sophisticated approach to text embedding by leveraging pretrained architectures . These models are typically pretrained on massive corpora of unannotated text using self-supervised learning tasks such as masked language modeling or next sentence prediction [27]. Given that our objective is to encode the semantics of the input text, we focus on pretrained encoder-based transformer models with fixed parameters to generate embeddings. This can be achieved in two main approaches: (1) extracting token-level embeddings from the final hidden layer of the frozen pre-trained transformer and applying pooling strategies such as mean, max, or CLS pooling; or (2) using sentence-level encoder-based transformers that directly provide embeddings for entire sequences of tokens. Both outlined approaches have been investigated in Dutch, as seen in the transformer models such as Bertje [19] and Robbert [21] as well as multilingual transformer models such as E5 [59].

2.3 Fine-tuned transformers for multi-target regression

An alternative to extracting embeddings from static or frozen transformer models and using them as input features for a separate MTR model is to fine-tune the transformer directly on the downstream MTR task [46]. This involves adding a task-specific output layer with k outputs, corresponding to the number of targets in the MTR problem, and continuing training on the new task, thereby adjusting the model's internal parameters to better suit the downstream objective.

2.3.1 Full vs. partial fine-tuning. In *full* fine-tuning, all the parameters of the model, including both the encoder layers and the newly added task-specific output layer, are updated. This allows strong adaptation to the target task but, with limited data, updating a large number of parameters easily results in over-adaptation, leading to unstable results across runs [42]. *Partial* fine-tuning addresses this challenge by keeping most of the encoder parameters frozen while training only a small subset of parameters. This subset typically includes the task-specific output layer and either a small number of added parameters or selected layers from the original model. This approach, often called Parameter-Efficient Fine-Tuning (PEFT), reduces both computational cost and the risk of overfitting [29].

2.3.2 Parameter-efficient fine-tuning (PEFT). Several PEFT strategies have been proposed. *Adapters* insert small trainable modules between the frozen layers of the transformer [31] and only update these adapter modules during fine-tuning. *BitFit* keeps the original weights frozen and updates the bias terms [4]. *Prompt-tuning* and *prefix-tuning* learn small continuous vectors (soft prompts) that are prepended to the input or intermediate representations and guide the model's behavior toward the target task without modifying the underlying model weights [38, 39]. In our study, we adopt Low-Rank Adaptation (LoRA), which attaches low-rank update matrices to selected parameters. These low-rank matrices are trained to capture task-specific adaptations, while the original weights remain unchanged. After training, the low-rank updates can be merged back into the original weights, resulting in no additional inference cost [33]. PEFT methods such as LoRA provide a pragmatic solution for limited-data settings: they greatly reduce the number of trainable parameters, while allowing for meaningful task-specific adaptation of large pre-trained models.

2.4 Data Augmentation

With limited data, MTR models risk overfitting to narrow phrasings and contexts. Data augmentation methods such as LLM-based synthesis can increase lexical and contextual diversity by generating new training examples, thereby improving coverage [11, 9]. Since MTR is a supervised learning task, augmenting the dataset requires generating new activity text along with corresponding target values (scores).

2.4.1 New Text Item Generation. To generate new activities, decoder-only transformers such as Gemma [49] can be prompted. While the generated synthetic text examples may be fluent, they can drift from the intended task. To ensure the suitability of the generated instances, the following strategies should be employed: (i) *in-context (few-shot) prompting*, which grounds the model using a handful of representative examples to stabilize style and intent [6, 16]; (ii) *controlled generation*, conditioned on task schemas or domain ontologies to maintain relevance [25, 24]; and (iii) *diversity-aware selection*, which filters candidates to avoid near duplicates and broaden coverage in the embedding space [2, 37].

2.4.2 Pseudo-Labeling. Before adding the newly generated instances to the original dataset, they must be annotated by a teacher model. In the context of MTR, to expand the knowledge of an MTR, a teacher MTR model produces soft targets (scores) for these new instances [40]. These labeled examples can then be incorporated into the training data of the student MTR model.

3 Related work

Most educational recommender systems focus on classification, ranking, or single-target regression from structured data or predefined labels, with little attention to joint prediction of multiple numerical scores from free text [15, 3]. In Dutch NLP, models like BERTje [19], RobBERT [21], and Dutch Sentence-BERT [48, 45] have enabled progress in tasks such as hate speech detection [8, 50], named entity recognition [56], and zero-shot classification [18] based on dutch input. However, educational applications remain limited, focusing mainly on single-target essay scoring or classification-based feedback [17, 1, 44]. While some work explores guidance-oriented classification [61], it typically uses non-Dutch data and discrete labels. To our knowledge, no prior work has applied multi-target regression (MTR) to Dutch free-text for small educational datasets. We address this gap by systematically comparing MTR methods trained on Dutch text embeddings to predict multiple numerical targets in the context of study choice.

4 Methodology

In this section, we describe the methodology for applying MTR to text input from a relatively small Dutch educational dataset,

Table 1: MTR models and hyper-parameters grid

ID	Strategy	Hyper-parameters grid
local_lasso	Local	Lasso: $\alpha \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$
local_rf	Local	RF: $n_{\text{est}} \in \{50, 100\}$, $d_{\max} \in \{\varnothing, 5, 10\}$
chain_lr	ERC	chain: $n_{\text{chains}} \in \{3, 5\}$, $k_{\text{CV}} \in \{3, 5\}$
chain_rf	ERC	chain: $n_{\text{chains}} \in \{3, 5\}$, $k_{\text{CV}} \in \{3, 5\}$
global_rf	Global	RF: $n_{\text{est}} \in \{50, 100\}$, $d_{\max} \in \{\varnothing, 5, 10\}$

aimed at supporting study choice guidance. The overall pipeline is illustrated in Figure 2.

4.1 Dataset

The used dataset comprises 96 activity descriptions in Dutch. For each activity, experts have assigned 14 numerical values representing eight curricular domains defined by the Flemish secondary education system and six RIASEC personality facets [30]. The activity descriptions are typically brief, ranging from 2 to 12 tokens. The target values, indicating relevance of each activity to a specific domain or personality facet, range from -6 to 6, as shown in Figure 3.

4.2 Multi-target regression methods

We evaluate five MTR methods spanning two model families (linear, tree-based) and three strategies (local, chains, global). Local methods train one estimator *per target*; chain methods use an *ensemble of regressor chains* (ERC) in which out-of-fold (OOF) predictions for earlier targets are appended as features for later targets; the global method fits a single multi-output estimator to all targets. Linear Regression (LR) and Lasso are used as linear models, while Random Forest (RF) is employed as the tree-based method. The models and hyper-parameter grids are listed in Table 1.

The tuned hyper-parameters include: regularization strength α for Lasso; number of trees in the forest n_{est} and maximum depth of tree $d_{\max}$ in RF; and number of chains n_{chains} and number of internal folds used to produce OOF meta-inputs k_{CV} for ERC. In `chain_rf` the RF *base* retains scikit-learn defaults. ERC uses random target orders and OOF predictions (via k_{CV} splits) appended as features and predictions are averaged across chains. Furthermore, PCA (with hyper-parameter grid: $n_{\text{comp}} \in \{0.7, 0.8, 0.9\}$, representing the proportion of variance to be explained) is applied to reduce dimensionality of the feature space for MTR methods.

4.3 Dutch Embeddings

Regarding embedding generators, we explored three families of methods to retrieve sentence representations: (i) *static* word embeddings, (ii) *frozen token-level* transformers with pooling, and (iii) *frozen sentence-level* transformers. (See Table 2).

4.3.1 Static word embeddings. Word2Vec and FastText are *Dutch-only* embeddings trained on Dutch text. They provide one vector per token. FastText also composes vectors for unseen words via subword units. To obtain sentence vectors, we apply simple mean pooling over tokens.

Table 2: Embedding generators

Type	Model	Dim	Id
Static	Word2Vec [55]	300	w2v
	FastText [26]	300	ftxt
Frozen—token	Bertje [19]	768	bert
	RoBERTa [21]	768	rbt
	RoBERTa-23 [20]	768	rbt_23
	XLM-R [14]	1024	xlmr
	ReMBERT [12]	1152	rmbrt
Frozen—sentence	SimCSE-XLM-R [48]	768	simcse
	SBERT [48]	1024	sbert
	SBERT-RoBERTa [13]	1024	sbrrt
	E5-large [59]	1024	e5

4.3.2 Frozen token-level transformers. Bertje, RoBERTa-23, and RoBERTa are *Dutch-only* encoders, while XLM-R and ReMBERT are *multilingual*. All are used with parameters frozen and output one vector per token. We convert these token vectors to sentence vectors using one of three standard pooling strategies: *mean* (average over tokens), *max* (element-wise maximum), or *CLS* (the model's special first-token embedding).

4.3.3 Frozen sentence-level transformers. SBERT and SBERT-RoBERTa are *Dutch-only* sentence encoders, whereas E5 and SimCSE-XLM-R are *multilingual*. These models produce sentence embeddings directly, so no pooling is required. They also differ in training focus. While SBERT and SimCSE learn sentence similarity, bringing semantically similar sentences closer together, E5 is trained for retrieval, matching short questions to the most relevant passages.

4.4 Fine-tuned transformers for multi-target regression

4.4.1 Parameterization. We apply Low-Rank Adaptation (LoRA) to the frozen sentence-level encoders. Each fine-tuned model has three components: (1) the pre-trained transformer encoder, (2) LoRA adapters applied to the attention projections, and (3) a single linear layer that maps the pooled sentence representation to 14 numerical outputs.

4.4.2 What is trainable. We freeze the base weights in the bottom 80% of the encoder's layers. LoRA adapters are attached to the `query` and `value` projection matrices across encoder layers. These adapters learn low-rank updates with rank r=4, scaling α=16, and dropout 0.1 [32]. The trainable components are: (i) LoRA parameters, (ii) the base weights in the top 20% of encoder layers, and (iii) the 14-d regression head.

4.4.3 Optimization. Training runs for up to 3 epochs (batch size 16) using AdamW with gradient clipping at 1.0. We perform a grid search over learning rates $\{5{\times}10^{-6}, 2{\times}10^{-5}, 5{\times}10^{-5}, 10^{-4}\}$ and tune by validation loss. Early stopping (`patience=2`) prevents overfitting. Sequences are truncated at 512 tokens. These settings follow recommended practices for stabilizing under small number of examples [42].

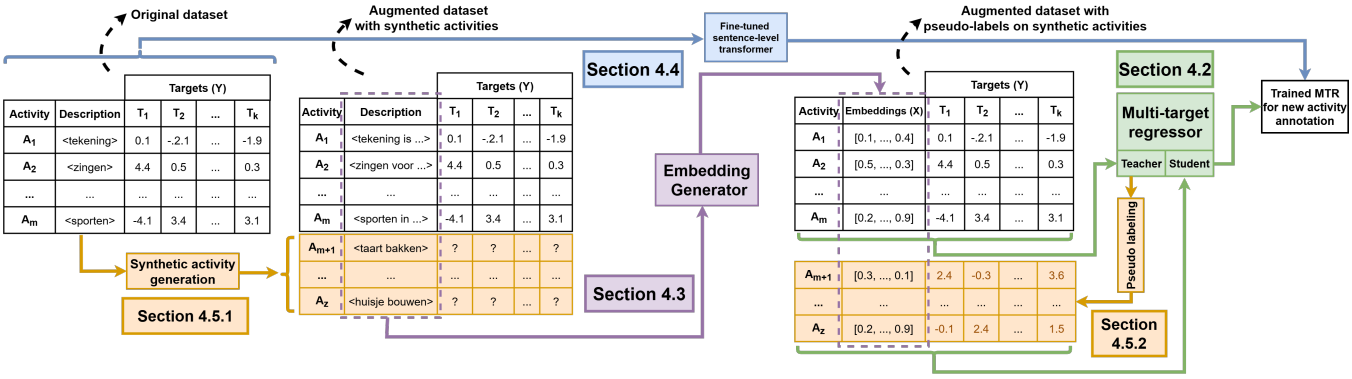


Figure 2: The overall schematic diagram to apply MTR on text-based instances

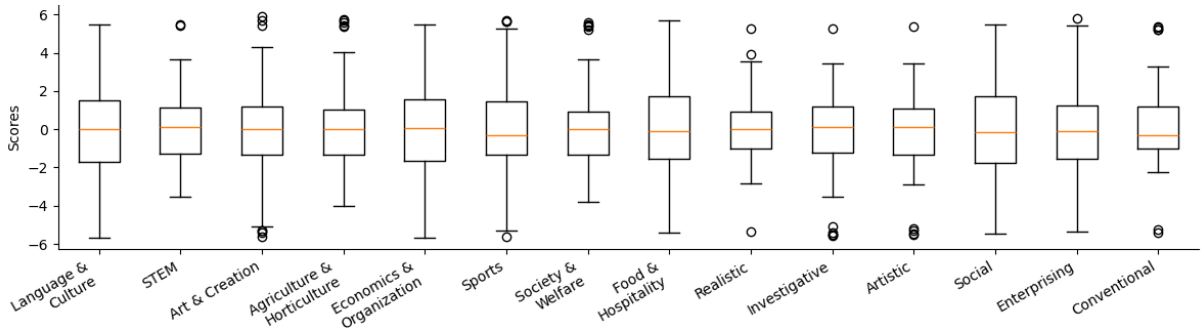


Figure 3: Score distribution per target

4.5 Activity Augmentation

4.5.1 Synthetic activity generation. We generate synthetic candidate activities using an instruction-tuned, decoder-only transformer, Gemma 2 model [49]. To ensure relevance, we apply *taxonomy-guided prompting* [24]. We condition generation on a human-curated schema representing the life-world of 10- to 14-year-olds in Flanders, Belgium. The schema was developed on the basis of the WHO's International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY) [60] and the input of an educational expert, and comprises 9 domains, 27 subdomains, and 123 facets[4]

Each prompt[5] consists of a fixed instruction block, specifying role, audience, style and length targets, safety constraints, and format, and a variable block that addresses exactly one facet. The variable block includes the domain, subdomain, and facet for contextualization, defines the task of formulating activities based on the facet under consideration, and incorporates few-shot prompting with three examples to establish tone and intent [6].

For each facet, we continue activity generation until we reach a validated pool of 48–72 candidate synthetic activities. Every candidate passes through a validation pipeline: Dutch text normalization, length filter, and near-duplicate removal. Once all facets have a validated pool, another round of de-duplication is considered across all facets to remove collisions between facets globally. For each facet, we embed all passed candidates and rank them to maximize diversity. We start with the medoid of that facet's pool, i.e. the sentence whose embedding is most central to the pool of candidates. We then grow the list by sampling the next sentence with a probability proportional to its squared distance from the set of already selected sentences, which reduces near-duplicates and broadens coverage, in a similar fashion to K-means++ [2]. To assemble the final corpus, we traverse facets by domain and subdomain, and take the first ranked sentences to gather synthetic sentences that are both diverse and spread evenly across all 123 facets.

4.5.2 Teacher-student pseudo labeling. To assign reliable targets to the selected synthetic activities, we adopt teacher–student pseudo-labeling [40]. The MTR methods outlined in Section 4.2 are considered as candidate teachers and as students. To avoid data leakage between train-test sets, the teacher model is tuned and trained only on the training examples of each cross-validation fold. Teacher-labeled synthetic activities never enter the test set. In the downstream student stage, we used the best performing teacher models based on the original dataset to predict pseudo-labels for the generated activities. The student MTR model then can be re-trained on the original training data augmented with the first K synthetic instances pseudo-labeled by the teacher MTR model, where $K \in \{10, 20, 50, 100, 200, 400\}\%$ of size of the original dataset.

[4]For more detailed information about the taxonomy, we refer readers to the supplementary material archived at Zenodo (DOI: 10.5281/zenodo.17834064) at `config/docs/taxonomy/taxonomy-guided-prompting.md`.

[5]An example of the prompt design is provided in the supplementary material archived at Zenodo (DOI: 10.5281/zenodo.17834064) in `docs/prompting/facet-conditioned-prompt-design.md`.

4.6 Evaluation framework

Given the size of our dataset, we adopt a leave-one-out cross-validation (LOOCV) scheme to maximize the use of available data for training. For each fold, only the single held-out seed is used for testing. As the evaluation measure, we use the relative root mean squared error (RRMSE), which is defined as $RRMSE = \frac{RMSE}{RMSE_D}$, where $RMSE_D$ is the RMSE of the dummy mean baseline for that target. RRMSE values smaller than one indicate improvement over the dummy baseline. To aggregate performance across folds, embeddings or models the median is used. The evaluation protocol is identical in the augmentation setting. Performance is measured in terms of the median RRMSE across folds.

To assess whether differences in results are statistically significant, we employed the statistical tests suggested by [22]. We employed the Friedman test followed by post-hoc Nemenyi test. Graphically, two methods are not statistically significantly different if a bar connects them in the critical distance diagram. Complementarily, we employed the Wilcoxon signed-rank test. In both cases, we set α to 0.05.

5 Results

The results of applying different MTRs on different types of embeddings are reported in Figure 4 where darker colors indicate lower RRMSE values (i.e., better performance). A triangle next to a value indicates that the difference between that cell and the best-performing combination is statistically significant, based on the Wilcoxon signed-rank test. The RRMSE values reported for the frozen transformer token-level embeddings correspond to mean pooling, as this strategy consistently outperformed the other two pooling methods (max and CLS) in our experiments[6]. Among all combinations, the MTR model employing a chain-based linear regressor with embeddings from *E*5, a multilingual sentence-level transformer, achieved the lowest RRMSE. As shown in the figure, this combination significantly outperforms all other configurations involving global and tree-based MTR methods or static and token-level embeddings.

Figure 5 presents the mean performance comparison of three groups of MTR methods, namely, local, chain-based and global, using three types of embeddings, static, token-level, and sentence-level frozen transformers.

As shown in Figure 5a, global MTR approaches struggle to fully leverage the interdependencies between targets, making them less effective than local and chain-based methods. This is often related to the number of outputs, as our dataset presents substantially more targets than the benchmark datasets normally used in the field [5].

In terms of embedding types, Figure 5b demonstrates that MTR methods using embeddings from frozen sentence-level transformers outperform those using frozen token-level or static embeddings. Sentence-level transformers generate embeddings by considering the full sentence context, resulting in more contextualized representations than those produced by token-level embeddings from frozen transformers or static models. Additionally, we observe that linear models significantly outperform tree-based methods, averaged over different embedding types.

Since frozen sentence-level embeddings consistently outperformed the other two types, we selected the three best-performing sentence-level transformer models, namely, E5, S-BERT, and SimCSE-XLM-R, for fine-tuning on the training data. Interestingly, as shown in Figure 6, fine-tuning on the 95 training instances did not lead to improved performance; the frozen embeddings combined with a MTR method still significantly outperformed their fine-tuned counterparts. We believe this is mainly due to the limited size of the dataset—95 instances are insufficient to effectively update the weights of sentence-level transformers for a regression task with 14 targets [53, 10].

Prior work points to several elements that can explain this finding. First, in tiny-data regimes, fine-tuning pretrained transformers is highly sensitive to seeds and hyperparameters, even with PEFT methods such as LoRA, which can lead to degraded performance relative to simpler baselines [42, 23]. Second, when the downstream distribution differs from pretraining, direct fine-tuning on a small in-domain set may be insufficient unless the encoder is first adapted via domain- or task-adaptive pretraining (DAPT/TAPT) [28]. Third, regression is sensitive to scale, heteroscedastic noise, and outliers. Fourteen regression targets magnify this instability: a few high-variance outputs can disproportionately influence joint optimization [35]. Empirically, similar observations have been reported in related educational settings [53] and in a recent clinical prediction benchmark [10].

The impact of introducing new pseudo-labeled synthetic activities into the training data is presented in Table 3. Asterisks (∗) in the table indicate statistically significant differences between the performance of the (student) model trained with augmented data and the same (student) model trained on the original dataset, based on the Wilcoxon signed-rank test. The two MTRs that performed best in Figure 4, namely, `chain_lr` and `local_lasso` using `E5` embeddings[7], were trained on augmented datasets ranging from 10% to 400% of the original size, pseudo-labeled with either of these two methods. While the reduction in RRMSE for (student) `chain_lr` with augmented data is not statistically significant, (student) `local_lasso` shows a statistically significant improvement, primarily across the mid-range augmentation levels (50–200%). Furthermore, as shown in Table 3, the additional synthetic activities slightly improve the diversity of the embedding space, but only after at least a 50% increase in the activity pool. We hypothesize that this added diversity could be one possible reason for the improved MTR performance, especially with augmentation levels above 50%.

6 Discussion

Figure 7 presents the RRMSE values per target for the best performing model (Chain_lr and E5; see Figure 7a) as well as the average performance across different MTR methods (Figure 7b). In these figures, blue bars represent domain-related targets, while orange bars correspond to personality traits. Notably, the **"Language &**

[6]For complete experimental details, full results, and reproducibility code, we refer readers to our publicly available repository https://anonymous.4open.science/r/kr8cht_review_anonymous/

[7]We include only the two best-performing MTR models in this table for the sake of readability, and refer readers to the shared reproducibility repository for detailed results on all MTRs.

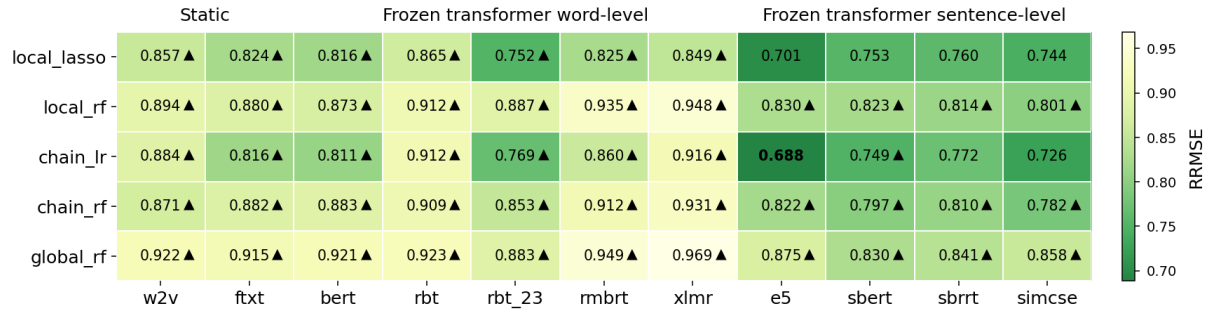


Figure 4: MTR performance (RRMSE) with statistical signifince test vs the best model (*Chain_lr + E5*)



(a) MTR categories overall ranking

(b) Embedding categories overall ranking

Figure 5: Cross-approach critical-distance diagrams for MTR and text embddings

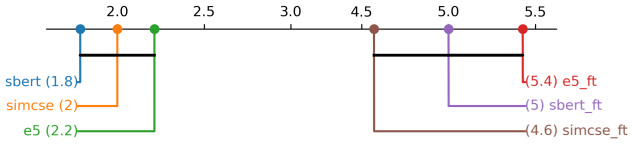


Figure 6: Performance of the best MTR and sentence-level embeddings compared to their fine-tuned version

Table 3: Augmentation effect on MTR performance

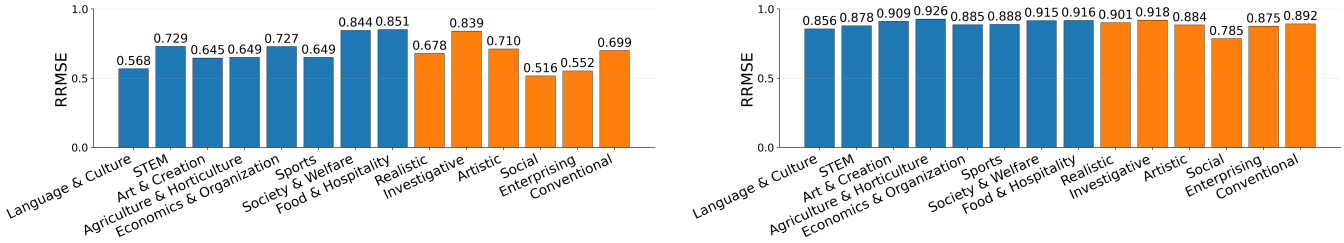
	Augmentation %						
teacher - student	base	10	20	50	100	200	400
$chain_{lr}$ – *chain_lr*	0.689	0.674	0.668	0.674	0.685	0.685	0.685
$chain_{lr}$ – *local_lasso*	0.705	0.690*	0.688	0.678*	0.679*	0.685*	0.698*
$local_{lasso}$ – *chain_lr*	0.689	0.684	0.678	0.676	0.676	0.691	0.708
$local_{lasso}$ – *local_lasso*	0.705	0.690	0.696	0.679*	0.681*	0.684*	0.706
Avg. added diversity	-	+0%	+0%	+1%	+2%	+3%	+4%

Culture" domain and the **"Social"** personality trait exhibit the lowest RRMSE values, both in the best-performing configuration and across the average of all MTRs. We attribute this to their relatively uniform distribution across the full score range, as illustrated in Figure 3. In contrast, the **"Agriculture & Horticulture"** domain and the **"Investigative"** personality trait show the highest RRMSE values on average, likely due to their skewed or limited coverage of the possible score range. Furthermore, language-rich cues (e.g., "read", "poetry" and "sing" for **"Language & Culture"**; "help" and "host" for **"Social"**) are more constructive for embeddings than generic inquiry verbs typical of **"Investigative"**, which carry little topical content unless accompanied by an object or clause (e.g., "look (up)", "observe", "find out" and "search"), especially given the short length of activity sentences.

While our study yielded interesting results, we acknowledge two key limitations. First, we relied on a single dataset for the study choice task. Although validating our findings across multiple datasets would be ideal, to the best of our knowledge, no other publicly available dataset exists for this specific task, namely, MTR on a small-scale dataset with Dutch-language inputs. Second, we did not explore the generation of entirely new, in-domain activity descriptions followed by expert annotation. Such an approach could help assess whether adding freshly generated and labeled (by human-expert) instances improves MTR performance. However, we currently lack access to a sufficiently large and qualified expert committee capable of providing reliable annotations for these newly generated activities.

7 Conclusion

In this paper, we investigated multi-target regression models to support study choice. More specifically, these models predict domains scores and personality traits based on textual descriptions of activities in which pupils indicate their preferences, for instance, "Would you prefer to sell waffles or read poetry". By comparing multiple models, our experiments revealed that global methods perform worse than local and chain-based regressors. Furthermore, frozen sentence-level transformers yield more informative embeddings than frozen token-level transformers and static models, due to their ability to generate contextualized sentence-level representations. Surprisingly, fine-tuned sentence-level transformers did not outperform multi-target regressors trained on embeddings from frozen sentence-level transformers, likely due to the limited size of the dataset and the high cardinality of the targets. Moreover, we observed that using Gemma to generate synthetic activities leads to



(a) Performance of chained_lr with E5 embeddings for each target

(b) Average performance of 5 MTRs for each target

Figure 7: MTR performance per target

superior performance. Specifically, we employed a teacher–student pseudolabeling approach, where the teacher model, an MTR method in this case, predicts target values for the synthetic activities. The student model is then trained on both the original data and the newly pseudolabeled synthetic data. In this case, we noticed 4% improvement in RRMSE with statistical significance.

As a future direction, we aim to investigate the validity of pseudolabels through evaluation by a committee of experts. Additionally, we plan to assess whether expanding and paraphrasing the textual descriptions of activities can lead to improved model performance.

Acknowledgments

The data used in this paper were gathered through Transbaso interuniversity SBO research project (2014-2018), executed by University of Ghent, the Vrije Universiteit Brussel (VUB), University of Antwerp, HOGENT and AP University of Applied Sciences. The project aimed to investigate school and study choice processes during the transition from primary to secondary education in the cities of Antwerp and Ghent. The authors would like to thank the Research Fund Flanders (FWO, personal mandate 1235924N to FKN) and the Flemish Government (through the Flanders AI Research Program).

References

[1] Mohsin Abbas, Peter Van Rosmalen, and Marco Kalz. 2023. A data-driven approach for the identification of features for automated feedback on academic essays. *IEEE Transactions on Learning Technologies*, 16, 6, 914–925.

[2] David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of SODA*, 1027–1035.

[3] Erum Ashraf, Selvakumar Manickam, and Shankar Karuppayah. 2021. A comprehensive review of course recommender systems in e-learning. *Journal of Educators Online*, 18, 1.

[4] Elad Ben-Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: simple parameter-efficient fine-tuning for transformer-based masked language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 1–9. doi:10.18653/v1/2022.acl-short.1.

[5] Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. 2015. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5, 5, 216–233. doi:10.1002/widm.1157.

[6] Tom B. Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. NeurIPS 2020.

[7] Hongliu Cao. 2024. Recent advances in text embedding: a comprehensive review of top-performing methods on the mteb benchmark. *arXiv preprint arXiv:2406.01607*.

[8] Tommaso Caselli and Hylke Van Der Veen. 2023. Benchmarking offensive and abusive language in Dutch tweets. In *The 7th Workshop on Online Abuse and Harms (WOAH)*. Yi-ling Chung, Paul R{\"ottger}, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani, (Eds.) (July 2023).

[9] Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: higher diversity and comparable model robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, (Dec. 2023), 1889–1905. doi:10.18653/v1/2023.emnlp-main.117.

[10] Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024. Clinicalbench: can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*.

[11] Tong Chen et al. 2023. An empirical survey of text data augmentation for large language models. *Transactions of the Association for Computational Linguistics*, 11, 1743–1764. doi:10.1162/tacl_a_00590.

[12] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

[13] Manuel R Ciosici, Joseph Cummings, Mitchell DeHaven, Alex Hedges, Yash Kankanampati, Dong-Ho Lee, Ralph Weischedel, and Marjorie Freedman. 2021. Machine-assisted script curation. *arXiv preprint arXiv:2101.05400*.

[14] Alexis Conneau et al. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[15] Felipe Leite Da Silva, Bruna Kin Slodkowski, Ketia Kellen Araújo Da Silva, and Sílvio César Cazella. 2023. A systematic literature review on educational recommender systems for teaching and learning: research trends, limitations and opportunities. *Education and information technologies*, 28, 3, 3289–3328.

[16] Haixing Dai et al. 2023. Chataug: leveraging ChatGPT for text data augmentation. *arXiv preprint arXiv:2302.13007*.

[17] Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. 2017. Readerbench learns dutch: building a comprehensive automated essay scoring system for dutch language. In *International Conference on Artificial Intelligence in Education*. Springer, 52–63.

[18] Loic De Langhe, Aaron Maladry, Bram Vanroy, Luna De Bruyne, Pranaydeep Singh, Els Lefever, and Orphée De Clercq. 2024. Benchmarking zero-shot text classification for dutch. *Computational Linguistics in the Netherlands Journal*, 13, 63–90.

[19] Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: a dutch bert model. *arXiv preprint arXiv:1912.09582*.

[20] Pieter Delobelle and François Remy. 2024. Robbert-2023: keeping dutch language models up-to-date at a lower cost thanks to model conversion. *Computational Linguistics in the Netherlands Journal*, 13, 193–203.

[21] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

[22] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7, Jan, 1–30.

[23] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

[24] Yuheng Du, Shereen Oraby, Vittorio Perera, Minmin Shen, Anjali Narayan-Chen, Tagyoung Chung, Anu Venkatesh, and Dilek Hakkani-Tür. 2020. Schema-guided natural language generation. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, (Dec. 2020), 283–295. doi:10.18653/v1/2020.inlg-1.35.

[25] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*. Jose M. Alonso, Alberto Bugarín, and Ehud Reiter, (Eds.) (Sept. 2017).

[26] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[27] Nuno Guimarães, Ricardo Campos, and Alípio Jorge. 2024. Pre-trained language models: what do they know? *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14, 1, e1518.

[28] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8342–8360. doi:10.18653/v1/2020.acl-main.740.

[29] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: a comprehensive survey. (2024). https://arxiv.org/abs/2403.14608 arXiv: 2403.14608 [cs.LG].

[30] John L Holland. 1997. *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources.

[31] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 97. PMLR, 2790–2799. https://proceedings.mlr.press/v97/houlsby19a.html.

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: low-rank adaptation of large language models. *ICLR*, 1, 2, 3.

[33] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[34] Dimitrios Iliadis, Bernard De Baets, and Willem Waegeman. 2022. Multi-target prediction for dummies using two-branch neural networks. *Machine Learning*, 111, 2, 651–684.

[35] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7482–7491.

[36] Dragi Kocev, Celine Vens, Jan Struyf, and Sašo Džeroski. 2013. Tree ensembles for predicting structured outputs. *Pattern Recognition*, 46, 3, 817–833.

[37] Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5, 2–3, 123–286. doi:10.1561/2200000044.

[38] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3045–3059. doi:10.18653/v1/2021.emnlp-main.243.

[39] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: optimizing continuous prompts for generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 4582–4597. doi:10.18653/v1/2021.findings-acl.417.

[40] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. (2016). https://arxiv.org/abs/1511.03643 arXiv: 1511.03643 [stat.ML].

[41] Saulo Martiello Mastelini, Everton Jose Santana, Ricardo Cerri, and Sylvio Barbon Jr. 2020. Dstars: a multi-target deep structure for tracking asynchronous regressor stacking. *Applied Soft Computing*, 91, 106215.

[42] Maximilian Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

[43] Felipe Kenji Nakano, Konstantinos Pliakos, and Celine Vens. 2022. Deep tree-ensembles for multi-output prediction. *Pattern Recognition*, 121, 108211.

[44] Héctor J Pijeira-Díaz, Sophia Braumann, Janneke van de Pol, Tamara van Gog, and Anique BH de Bruin. 2024. Towards adaptive support for self-regulated learning of causal relations: evaluating four dutch word vector models. *British Journal of Educational Technology*, 55, 4, 1354–1375.

[45] Héctor J Pijeira-Díaz, Shashank Subramanya, Janneke van de Pol, and Anique de Bruin. 2024. Evaluating sentence-bert-powered learning analytics for automated assessment of students' causal diagrams. *Journal of Computer Assisted Learning*, 40, 6, 2667–2680.

[46] Raphael Poulain, Mehak Gupta, Randi Foraker, and Rahmatollah Beheshti. 2021. Transformer-based multi-target regression on electronic health records for primordial prevention of cardiovascular disease. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 726–731.

[47] Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. 2021. Classifier chains: a review and perspectives. *Journal of Artificial Intelligence Research*, 70, 683–718.

[48] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, (Nov. 2019). http://arxiv.org/abs/1908.10084.

[49] Morgane Riviere et al. 2024. Gemma 2: improving open language models at a practical size. Plus additional authors not listed here. (2024). arXiv: 2408.00118 [cs.CL]. doi:10.48550/arXiv.2408.00118.

[50] Ward Ruitenbeek, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov, and Tommaso Caselli. 2022. "zo grof!": a comprehensive corpus for offensive and abusive language in dutch. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 40–56.

[51] Marie Seghers, Simon Boone, and Piet Van Avermaet. 2019. Social class and educational decision-making in a choice-driven education system: a mixed-methods study. *British Journal of Sociology of Education*, 40, 5, 696–714.

[52] Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. 2016. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104, 1, 55–98.

[53] Anaïs Tack, Siem Buseyne, Changsheng Chen, Robbe D'hondt, Michiel De Vrindt, Alireza Gharahighehi, Sameh Metwaly, Felipe Kenji Nakano, and Ann-Sophie Noreillie. 2024. ITEC at BEA 2024 shared task: predicting difficulty and response time of medical exam questions with statistical, machine learning, and language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 512–521.

[54] Sarah Thys. 2018. *The tertiary effect of social class: Multilevel studies on the role of the primary school (teacher) in educational decision-making*. Ph.D. Dissertation. Ghent University.

[55] Stephan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, (Eds.) European Language Resources Association (ELRA), Portorož, Slovenia, (May 2016). ISBN: 978-2-9517408-9-1.

[56] Chaim Van Toledo, Friso Van Dijk, and Marco Spruit. 2021. Dutch named entity recognition and de-identification methods for the human resource domain. *arXiv preprint arXiv:2106.02287*.

[57] Kristin Vanlommel, Roos Van Gasse, Jan Vanhoof, and Peter Van Petegem. 2021. Sorting pupils into their next educational track: how strongly do teachers rely on data-based or intuitive processes when they make the transition decision? *Studies in Educational Evaluation*, 69, 100865.

[58] Wendelien Vantieghem, Piet Van Avermaet, Steven Groenez, and Miet Lambert. 2018. Diversiteitsbarometer onderwijs vlaamse gemeenschap. In *Diversiteitsbarometer onderwijs*. Els Keytsman, 47–193.

[59] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: a technical report. *arXiv preprint arXiv:2402.05672*.

[60] World Health Organization. 2007. *International Classification of Functioning, Disability and Health: Children & Youth Version (ICF-CY)*. World Health Organization, Geneva. ISBN: 9241547324.

[61] Omar Zahour, EH Benlahmar, A Eddaoui, and O Hourrane. 2020. Towards a system for predicting the category of educational and vocational guidance questions using bidirectional encoder representations of transformers (bert). *IJATCSE*, 9, 1, 505–511.