Evaluating the Generalisation of an Artificial Learner

Bernardo Stearns[1] and **Nicolas Ballier**[2] and **Thomas Gaillat**[3]
and **Andrew Simpkin**[4] and **John P. McCrae**[1]
[1] Insight Centre for Data Analytics, Data Science Institute, University of Galway, Ireland
[2] LLF & CLILLAC-ARP / Université Paris Cité, rue Thomas Mann, 75013 PARIS, France
[3] LIDILE / Université de Rennes 2, 35000 Rennes, FRANCE
[4] School of Mathematical and Statistical Sciences, University of Galway, University Road, Galway, Ireland
Contact: bernardo.stearns@insight-centre.org

Abstract

This paper focused on the creation of LLM-based artificial learners. Motivated by the capability of language models to encode language representation, we evaluated such models for predicting masked tokens in learner corpora.

We domain-adapted the BERT model, pre-trained on native English, by further pre-training two learner models on learner corpora: a natural learner model on the EFCAM-DAT dataset and a synthetic learner model on the C4200m dataset. We evaluated the two artificial learner models alongside the baseline native model using an external English-for-specific-purposes corpus from French undergraduates.

We evaluated metrics related to accuracy, consistency, and divergence. While the native model performed reasonably well, the natural learner pre-trained model showed improvements in recall-at-k. We analysed error patterns, showing that the native model made "overconfident" errors by assigning high probabilities to incorrect predictions, while the artificial learners distributed probabilities more evenly when wrong. Finally, we showed that the general token choices from the native model diverged from the natural learner model and this divergence was higher at lower proficiency levels.

1 Introduction

Over the last 20 years, learner corpora have significantly benefited research in applied linguistics and NLP by providing insights into how second language (L2) learners improve their proficiency. This understanding has led to enhanced course material design, improved teacher training, and greater awareness of students' linguistic abilities. Additionally, when combined with NLP technologies, learner corpora have proven valuable for CALL applications like grammar error detection and proficiency classification (Bryant and Briscoe, 2018; Tetreault et al., 2018). This paper explores the potential of leveraging Large Language Models (LLMs) with learner corpora, which have traditionally been used to test specific research hypotheses. Instead of relying on diverse corpora with relevant metadata for testing various hypotheses, we explore the possibility of a single model that simulates learner behavior across different contexts. Such artificial learners could respond to new stimuli, providing a testbed for linguistic hypotheses, with outputs from a generic English learner model compared to those from a native model. By training an LLM on learner data, it may be possible to create an artificial English learner that captures the idiosyncrasies of actual learners.

This research explored the creation of an Artificial L2 Learner (ALL) model by pre-training it on second language learner corpora, leveraging domain-adaptive pre-training. We also believe that modelling learners' knowledge and their use of words and linguistic skills is crucial for Intelligent Tutoring Systems (ITS) and digital learning platforms in second language teaching and learning. For an ITS focused on language learning, modelling word usage and language skills of learners is essential. This is why any simulation of learner behavior, as a key goal for an ITS, should be accurate and reliable. Motivated by the capability of

Bernardo Stearns, Nicolas Ballier, Thomas Gaillat, Andrew Simpkin and John P. McCrae. Evaluating the Generalisation of an Artificial Learner. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 199–208.

language models to represent linguistic concepts, this research explored the domain-adaptive pretraining of large language models (LLMs) to simulate the behavior of English learners, which we call Artificial Learner models. Creating an artificial learner raises at least three questions:

1. How accurate is the artificial learner in predicting what learners would actually say?

2. How confident is the learner in its predictions?

3. How divergent is the AL compared with a generic native model?

The rest of the paper is structured as follows: Section 2 presents related research. Section 3 explains the training data and the procedures used to create our artificial learners. Section 4 delves into our results, and Section 5 provides a discussion of these results.

2 Background Research

Research in second language acquisition has been explored from many different perspectives, resulting in different models for each aspect of the learning process. For example, Whitehill and Movellan (2017) models learners taking into account how a learner infers and updates vocabulary knowledge after doing exercises in a specific ITS for foreign language learning. The SLAM shared task (Settles et al., 2018) models the history of a learner's mistakes in Duolingo, predicting if a learner is likely to make a mistake given their past history of mistakes. There are also models that are complementary to modelling the second language acquisition process, such as spaced repetition practice models (Settles and Meeder, 2016) and efficient grammatical error correction (Omelianchuk et al., 2020). Despite the success of such diverse tasks in their specific modelling objectives, the usage of their models is tied to the specific case of their system or language learning task. This restricts the capability of such models to simulate the general behavior of language learners.

There is another set of language-learning tasks that explicitly model learners' behavior and knowledge however, they are still tied to a single task depending on handcrafted features. Examples include Whitehill and Movellan (2017), which models vocabulary learning from concepts; Knowles et al. (2016), which models noun understanding from the context of the native language; and Zylich and Lan (2021), which models retrieval practice performance for SLA based on linguistic and memory-based features. Other similar modeling tasks include Avdiu et al. (2019); Renduchintala et al. (2016). In a similar fashion, corpus linguists have also developed single tasks aimed at predicting specific outcomes in the form of linguistic constructions. Bresnan and Nikitina (2009) modelled the dative alternation, where learners hesitate between the prepositional dative structure or the double object structure. Gries et al. (2020) approaches in corpus linguistics also reflect this method by modeling the genitive vs. *noun_of_noun* construction. Modelling construction outcomes in learner texts helps understand the contexts, triggering constructions. Nevertheless, these models cannot handle different sets of constructions, which appears to be a limitation if one wants to analyze many different linguistic systems at the same time. In contrast, large language models (LLMs) are capable of accommodating diverse constructions and analyzing multiple linguistic systems simultaneously, offering a more flexible approach to understanding language patterns.

In the broader field of Natural Language Processing, language models have been effectively adapted to multiple domains and tasks using a single generic model, in a similar scenario we see in the Second Language Acquisition domain. Gururangan et al. (2020a) examines the effectiveness of adapting pre-trained language models to multiple domains and tasks with a single model. They test how well a task-specific fine-tuned model transfers to different types of other tasks, showing a large gain in task performance using an overall multi-phase domain and a task-adaptive pre-trained model. Though we see an underutilization of language models in learner modelling tasks, many other diverse areas have successfully adapted language models to their tasks.

To the best of our knowledge, two tasks analysed the potential of language models in SLA. Palenzuela et al. (2022) explored native pretrained language models to predict language mistakes in the SLAM shared task. Kim (2024) investigated the use of language models as "artificial English learners" with a model called Bidirectional Encoder Representations from Transform-

ers (BERT). They specifically tested BERT's ability to simulate English learners' usage of prepositions. Notably, BERT was domain-adaptively pretrained on the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013). The study focused on how this artificial learner utilized four English prepositions: *at*, *for*, *in*, and *on*.

Our work proposes a generalized analysis of artificial English learners, which expands the scope of previous analysis by introducing a broader range of metrics, including accuracy, consistency, and behavior validation. The goal is to establish trust in the trained models before exploring their capabilities in specific tasks.

3 Material and Methods

3.1 Data

3.1.1 Training data

EFCAMDAT corpus - We trained two artificial learner models. The first model was trained on the EFCAMDAT. We used the refined version of the EFCAMDAT corpus texts (Shatz, 2020). It includes 723,282 writings from *Englishtown* language schools (Shatz, 2020).

The learners wrote texts following prompts such as "introducing yourself by email". Students gradually moved from one level to the next based on language teachers' grades. The writings span across 16 proficiency levels, which were mapped to the first five CEFR levels. The CEFR levels of the texts correspond to the successful completion of the coursework levels at Englishtown.

C4200M corpus - The second model was trained on the C4200M corpus (Stahlberg and Kumar, 2021). It is a corpus of synthetically generated ungrammatical sentences used in neural grammatical error correction. This model produces an ungrammatical sentence given a clean sentence and an error type tag following the tags defined in the ERRANT automatic annotation tool (Bryant et al., 2017). The generated ungrammatical sentences follow the distribution of error tags in the BEA-dev dataset (Bryant et al., 2019). They argue for the utility of the generated ungrammatical data by pre-training grammar error correction models with it, outperforming genuine parallel data on the CONLL-2014 and JFLEG-test.

We chose the C4200m with the goal of analysing a common trade-off in the training process of large language models: balancing the quality of authentic texts versus the quantity of augmented texts, similar to works surveyed in Feng et al. (2021). We aimed to understand how this trade-off affects the performance of artificial English learners. By using the C4200m dataset, we wanted to see how different amounts of high-quality and lower-quality texts impact the learning results of our models. This would help us understand the best balance between text quality and quantity for training large language models. Our approach aligns with other NLP research, providing a comparative view that adds to the relevance and usefulness of our findings.

3.1.2 Testing data

The external test set (see Table 1) is made up of learner writings from the CELVA-SP (Mallart et al., 2023) a corpus of French undergraduates using English for specific purposes (ESP). Learners answered one of three question prompts as part of a 45-minute in-class writing task. For instance, they had to describe and share their opinion on the most important invention in their field. All their writings were subsequently annotated with the writing competence scale of the CEFR (Council of Europe, 2018, Appendix 4, p .187-189) by four expert raters. Pairwise inter-rater agreement was computed on the basis of 60 writings, yielding Cohen's kappa values ranging from .52 to .72. The rest of the writings were then annotated independently. Table 1 presents the distribution of the levels in CELVA-SP data.

3.2 Data processing

Processing the learner texts for our analysis involved two types of data processing. First, for the model training, we simply passed the raw texts as input to a masked language modelling collator, following the standard masking strategy used in the training process of BERT (Devlin et al., 2019). The collator dynamically generates batches of masked sentences, which the BERT tokenizer processes into WordPiece tokens for use in the training loop.

Second, for prediction analysis, we used a Univeral Dependency (UD) tokenizer (Nivre et al., 2016) to represent "human" learner tokens. We masked each token in the text one at a time, creating a unique masked sentence for every UD token. These sentences with a single masked token were then fed to our artificial learners and the baseline native model to predict token usage. We annotated

Table 1: Distribution of levels and essays in the CELVA-SP data (Mallart et al., 2023)

Writings	# of writings	% of writings	av # of words	SD
A1	85	8.70	126.78	76.67
A2	311	31.83	182.02	87.21
B1	335	34.28	231.34	111.70
B2	198	20.26	285.84	126.75
C1	48	4.91	347.93	144.69
Total	977	100	224.11	120.64

the part of speech for each UD token using UDPipe (Straka, 2018) implementation in spaCy[1]. It allowed us to visualize the distribution of probability scores across different parts of speech for the natural learner model. Since our experiment focused on the BERT base model and its limitation of 512 WordPiece tokens, we filtered out texts with more than 512 such tokens.

3.2.1 Domain-Adaptive Pre-training

The main step in developing the two proposed artificial learner models was the domain-adaptive pretraining of an already pre-trained baseline BERT model. We used the EFCAMDAT as a training set for the natural learner model, and the C4200m as a training set for the synthetic learner model. We trained both artificial learners on a masked language modelling task. In Devlin et al. (2019) they refer to pre-training as training a model on unlabeled data across various tasks, such as masked language modelling, where fine-tuning involves initializing a pre-trained model's weights and updating them using labeled data. We initialized a baseline BERT model weights and further pretrained them in learner corpus in an unsupervised masked language modelling task. This is referred in (Gururangan et al., 2020b) as domain-adaptive pre-training.

We used the same masked language modelling pre-training task described in Devlin et al. (2019). Specifically, we masked 15% of WordPiece tokens in each sentence of the training set, allowing the model to learn contextual representations by predicting the masked tokens.

3.3 Evaluation

To evaluate the predictions of the two artificial learner models and the native baseline model, we used three types of metrics: recall-at-k, KL divergence, and calibration. We calculated the metrics on the CELVA-SP dataset.

3.3.1 Accuracy with recall-at-k

We used the recall-at-k metric as our accuracy measure. It naturally extends the concept of accuracy by taking into account the model's top-k potential responses and explicitly consider a criteria for relevant responses that could be easily extended. In essence, we measured on average how many of the top-k token predictions recommended by a given model were relevant for the target masked token used by the learner.

The recall-at-k metric evaluates the top-k responses of a model that generates a list of potential responses $\hat{y}$ to a given query q, ranked by their likelihood of being correct according to the model. In our experiment, for a given masked token sentence the query q is the actual masked token used by the learner, and the list of potential responses $\hat{y}$ is the list of tokens predicted by a model ranked by probability in the softmax layer of BERT vocabulary.

For a target masked token q_i and a top-k token t_j predicted by the model, t_j is considered relevant to q_i simply if t_j is in the set of relevant items for q_i. In our experiment, the only relevant item was the target masked token itself, so this is equivalent to verifying if q_i is in the top-k predictions but this would not be the case in more complex scenarios.

We report the average recall@k over all masked tokens in the CELVA-SP for each of the three evaluated models.

$$\text{AVG Recall@}k = \frac{\sum_{q_i \in \text{masked\_tokens}} 1[q_i \in \text{top-k}(\hat{y})]}{\text{\# of masked tokens}}$$

We report recall for $k = [1, 5, 10]$.

[1]You can find the repository at https://github.com/TakeLab/spacy-udpipe.

3.3.2 Kullback–Leibler metric

The Kullback-Leibler divergence is rooted in information theory and provides a general approach for quantifying how two probability distributions differ. We framed each of our models' output probabilities for a given masked token as a discrete probability distribution over BERT's vocabulary tokens. Within that frame, we interpreted the KL metric for two given models as if their token choices generally diverged. We implemented the element-wise KL metric with a small epsilon value perturbation, $\epsilon = 10^{-6}$, to avoid the scenario where probabilities are zero. We calculated the KL element-wise metric for each masked token, and we grouped them by their text CEFR level with the intuition to find differences between CEFR levels.

$$KL(p_t, q_t) = p_t log\left(\frac{p_t + \epsilon}{q_t + \epsilon}\right)$$

3.3.3 Calibration Curves

To foster trustworthiness in our models, high accuracy is the immediate desired property of our models, assigning high probabilities to correct tokens. A second desirable property is that our models do not overconfidently make mistakes, assigning high probabilities to incorrect predictions.

One approach for such analysis is through the "calibration curve" method. Initially employed in analysing weather forecasts (Brier, 1950; DeGroot and Fienberg, 1983), this technique has since been applied to neural networks (Guo et al., 2017; Minderer et al., 2021) and recently to evaluate Large Language Models (LLMs) from a semantic perspective Levinstein and Herrmann (2024). For example, (Levinstein and Herrmann, 2024) utilizes calibration curves to assess the veracity of LLM statements on specific datasets and asserts that "calibration offers another metric for evaluating the quality of probes' forecasts." Calibration analyses have been utilized in neural networks and language models (Minderer et al., 2021; Chen et al., 2024), allowing researchers to assess the relationship between a model's prediction confidence and success rate.

Calibration curves help us analyze how well a model performs when it is confident or unconfident about it's prediction. In our experiment, our calibration curves correspond to how many successful predictions (event rate) we observe across different probability scores of the top-1 prediction of each model.

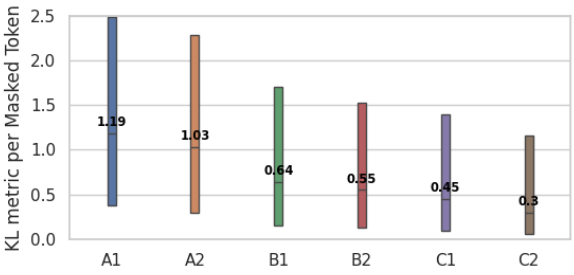

Figure 1: Interquatile range plot of KL metric between natural learner and native model per masked token sentence grouped by CEFR level in the CELVA-SP dataset as described in 3.3.2

$$\text{Event Rate} = \frac{\text{Number of Successful Predictions}}{\text{Total Number of Predictions}}$$

4 Results

4.1 Recall-at-k

We evaluated the accuracy of our models with recall-at-k metrics. We found a slight difference in accuracy between the Learner Models and the native model in the external CELVA-SP test set. We noticed a slow increase in recall as k increases. A slow increase in the values of top-k recall may indicate that the token vocabulary of the language model is not adequate for the task. We believe it is unlikely that the model is confused when choosing among 10 or more tokens; instead, the correct token is likely represented by multiple word-piece tokens in the model's vocabulary.

model	recall@1	recall@5	recall@10
bert-native (baseline)	0.600	0.622	0.635
bert-efcamdat	0.648	0.670	0.684
bert-c4200m	0.586	0.610	0.623

Table 2: Average recall-at-k in the CELVA-SP for each evaluated model as described in section 3.3.1

4.2 KL Distance

The KL metric interquantile plot in Figure 1 presents the KL metric between native BERT and the natural learner model. It allowed us to analyse the intuition that a learner model will generally differ from a native model in terms of token usage and that this difference is higher in beginner texts. The figure indicates that the learner model exhibits greater disagreement in token choice for masked sentences at lower proficiency levels, with a monotonic decrease in disagreement as proficiency increases.

4.3 Calibration Curves

The calibration curve in Figure 2 illustrates the relationship between the predicted probabilities of the top candidate token and the success rate at which these tokens correctly predict the true token. The three models follow a linear trend, showing that all of them classify more accurately as their top-1 token probability increases, suggesting that they are well-calibrated overall. However, the EFCAMDAT curve shows a discrepancy for probabilities around 0.6. Specifically, the natural artificial learner demonstrates underperformance in this range, as candidates predicted with a 60% probability only successfully predict the true token 40% of the time but increase and become slightly higher for probabilities close to 1. In general, the natural learner model outperforms the native model in the range of higher top-1 probabilities. This analysis can be further supported by Figure 3 where we noticed that the native model (on the right side of the figure) very frequently assign high probabilities to its top-1 prediction where the two artificial learners assign lower probabilities. Even though the native model assigns higher top-1 probabilities more frequently, it has a lower success rate than the natural learner model. One possible explanation for the learner model's underperformance in the 60% probability range is that the masked tokens in this range likely come from advanced learners' texts, whereas the EFCAMDAT dataset primarily consists of beginner learners. This motivates a detailed analysis of the performance of such models across CEFR levels as future work.

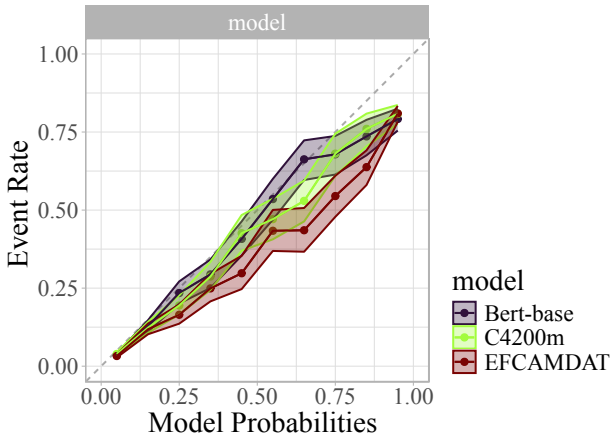

Figure 2: Sucess event rate across top-1 token model probabilities for all 3 models across all masked tokens in the CELVA-SP data

5 Discussion

5.1 Role of Part of Speech

Parts of speech (POS) provide a way to filter out the prediction distribution. It is possible to analyse the behaviour and success rate of the artificial learners according to linguistic properties related to not only the lexicon but also grammar. For instance, filtering out probabilities per auxiliary gives an insight into a closed class. This helps characterize the impact of universal part-of-speech (UPOS) on the probability distributions of the probability scores for the first three predictions (rank) across the three models. For example, Table 3 shows the average probability score assigned by a given model to its top-3 predictions, as well as the respective success rate for masked prepositions. We observe a similar pattern, where the na-

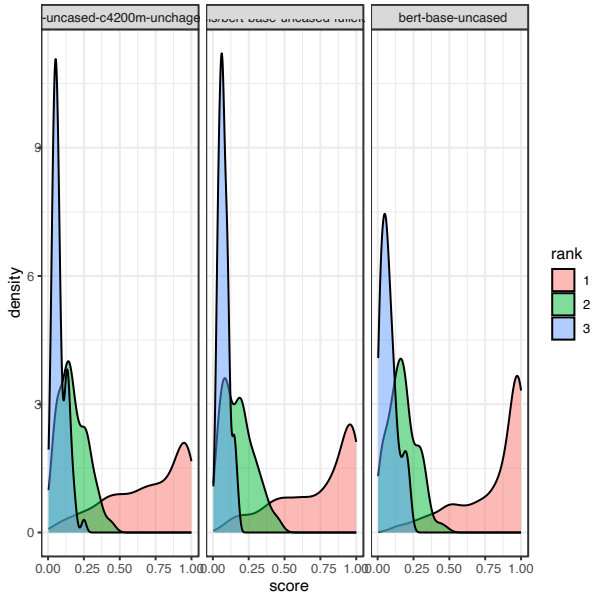

Figure 3: Probability Density Distribution per Models(Synthetic Learner Model, natural learner Model and Native model from left to right) for the top rank preditions from 1 to 3.

tive model, on average, assigns higher probability scores to its top-1 prediction, yet, has a lower success rate compared to the natural learner model.

Figure 4 displays the probability density distribution of words across different Universal Part-of-Speech (UPOS) for the first prediction (rank = 1). The x-axis represents the probability assigned by the natural learner model for each UPOS, while the y-axis shows the probability density. This visualization allows for a quick comparison of the relative frequencies of different UPOS across the dataset. It indicates how the model makes use of tokens of a certain type across levels.

Open-class categories such as adjectives (ADJ), nouns (NOUN), and verbs (VERB) have bimodal distributions, but the prominent mode reflects the uncertainty of the prediction (probability around 0.2 for ADJ). However, a closed class like prepositions (ADP) also has a bimodal distribution, but the prominent mode is around 0.9. This suggests that the model is more confident with some closed classes than open classes.

5.2 Domain Effects for ESP

We conducted a chi-squared test, which demonstrated that the difference between the domains was significant ($X^2 = 45.04, df = 6, p < 0.001$). Our data indicated that masked tokens were easier to predict in essays written for Communication Studies compared to those for Pharmacy, as illustrated in Table 4. This is some indication to further take into consideration domain and tasks effects.

5.3 Training Limitations

A significant limitation in our training process is the imbalance in the distribution of proficiency levels within the EFCAMDAT dataset. Specifically, there is a disproportionately higher number of beginner-level texts (A1, A2) compared to advanced-level texts (C1, C2). This imbalance may affect our KL plot 1. While the result aligns with expectations for lower proficiency levels, it may exhibit a training artifact effect where the model's contextual representation seems to be coherent towards the characteristics of beginner-level texts since it was exposed to a large amount of such texts, whereas for higher proficiency levels, the model's token choices simply follow the native BERT distribution .

This artifact impacts the model's ability to generalize across proficiency levels. For higher proficiency levels, the model's token choices tend to

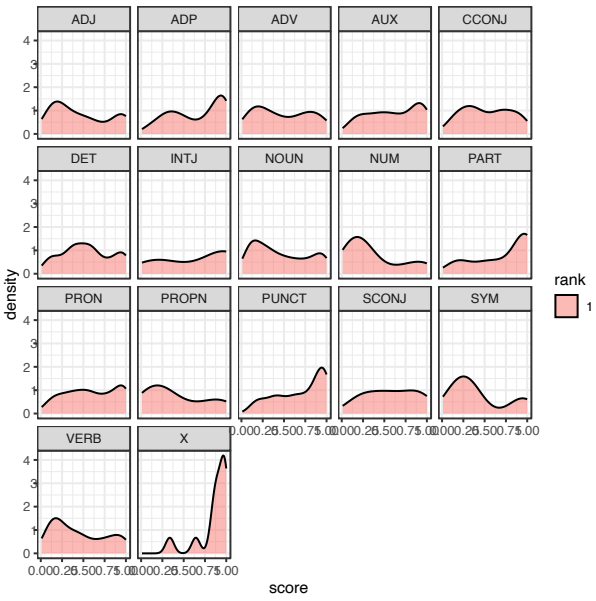

Figure 4: Probability Density Distribution of top-1 prediction of natural learner model per UPOS

align more closely with the original pre-training distribution, primarily because the advanced-level data is underrepresented. This limitation suggests that the model might not be equally effective across all proficiency levels, potentially underperforming for more advanced learners.

5.4 Perspectives for Future ITS Implementations

If our artificial learners manage to be sufficiently trustworthy for the emulation of what a learner would say, one can compare the prediction or the use of a given learner with each model pre-trained with a given CEFR level. Our experiment is only a prototype of our global undertaking. We will extend the pre-training to other areas, such as pre-training on different sub-levels of the CEFR scale. We have seen the reliability of the results, and we have also suggested that the models created were not too data-dependent in the sense that they could be generalized to other types of data.

6 Conclusion

In this paper, we have compared two artificial learners against a native language model in predicting tokens produced by learners. Our primary goal was to propose a masked language modelling task in learner corpora and analyse the accuracy, consistency, and divergence of such artificial learners. We explicitly chose a large synthetic ungrammatical dataset and an authen-

model	success_rate	score_mean	rank
bert-c4200m	0.53	0.60	1
bert-c4200m	0.08	0.10	2
bert-c4200m	0.04	0.04	3
bert-fullefcamdat	0.58	0.64	1
bert-fullefcamdat	0.09	0.09	2
bert-fullefcamdat	0.02	0.04	3
bert-base-uncased	0.55	0.71	1
bert-base-uncased	0.08	0.09	2
bert-base-uncased	0.03	0.04	3

Table 3: Model success rate and average probability score per rank (top-k position) for prepositions

	Communication	Electronics	Medicine	Pharmacy	Education	Environment	Physics
Success	1278	219	249	85	265	1139	749
Total	4284	933	1002	401	1157	4873	3301

Table 4: Contingency table of correct predictions per ESP domain (all models)

tic learner corpus to analyse the trade-off between the quality of authentic texts and the quantity of augmented texts. Even compared to the native BERT model, pre-training BERT in the synthetic C4200m dataset decreased accuracy, while training BERT on authentic texts increased accuracy. Accuracy is greater for closed classes, and the previous study on artificial learners rightly focused on a subset of a closed class, prepositions. Through analysing predicted probabilities against success rates, we investigated indications of calibrations and overconfident mistakes of our models, where native BERT showed a wider gap between its success rate and predicted probability. We finally compared native BERT with our natural artificial learner in relation to their choice of tokens, where the KL metric exhibit to be a coherent metric to generally measure the choice of tokens between language models. Since we pre-trained our artificial learner on a dataset containing more texts from beginner learners than those from advanced learners, we expect that it will simulate better beginner learners. Future work could address multiple aspects of the training process to enhance performance. We believe that merely increasing computational power and training time could still improve our artificial learners. Additionally, we believe that more specific masking strategies, such as masking incorrect tokens, and architectures that can personalize the artificial learner to a specific individual, could further enhance performance. In the direction of personalization, there are opportunities for training more specific artificial learners, such as nationality or proficiency based artificial learners.

Limitations

There are several limitations to our work that need to be acknowledged. One significant limitation is the high training cost associated with using deep learning models for natural language processing tasks. Training these models requires substantial computational resources, which can be expensive and time-consuming. In our study, although we aimed to mitigate these costs by using "small" encoder models such as BERT, the training costs were still considerably higher compared to traditional language modelling methods.

Furthermore, we expect to make our model available in accordance with the EFCAMDAT corpus curators, which provides a significant advantage in terms of cost-effectiveness and collaborative potential. Researchers and practitioners can leverage our pre-trained models and fine-tune them for their specific applications without incurring the high costs associated with training a model from scratch. This open-source approach promotes transparency and encourages further innovation and experimentation within the community.

Ethics Statement

In accordance with the curators of the EFCAMDAT corpus, we have planned to make our models pre-trained on the EFCAMDAT accessible on the

web server hosting the EFCAMDAT data.

Acknowledgments

This work has been supported by Science Foundation Ireland under Grant Number SFI12RC2289_P2 Insight_2, Insight SFI Centre for Data Analytics and by the French ANR under ANR grant ANR-22-CE38-0015 for the A4LL project.

References

Drilon Avdiu, Vanessa Bui, and Klára Ptačinová Klimčíková. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Turku, Finland. LiU Electronic Press.

Joan Bresnan and Tatiana Nikitina. 2009. On the Gradience of the Dative Alternation. In Karuvannur Puthanveettil Mohanan, Linda Uyechi, and Lian-Hee Wee, editors, *Reality Exploration and Discovery: Pattern Interaction in Language & Life*, pages 161–184. Center for the Study of Language and Information, Stanford, CA.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Lihu Chen, Alexandre Perez-Lebel, Fabian M Suchanek, and Gaël Varoquaux. 2024. Reconfidencing LLMs from the Grouping Loss Perspective. *arXiv preprint arXiv:2402.04957*.

Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Companion Volume with New Descriptors*. Council of Europe.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*.

Stefan Th Gries, Justus Liebig, and Sandra C Deshors. 2020. There's more to alternations than the main diagonal of a 2× 2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal*, 44(1):69–96.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020a. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020b. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Shin'ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, 1(1):91–118.

Wonbin Kim. 2024. Let's make an artificial learner to analyze learners' language! *(Language Sciences)*, (70):167–193.

Rebecca Knowles, Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2016. Analyzing learner understanding of novel L2 vocabulary. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 126–135.

Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27.

Cyriel Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu Li, and Thomas Gaillat. 2023. A new learner language data set for the study of English for Specific Purposes at university level. In *Proceedings of the 4th Conference on Language, Data and Knowledge - LDK 2023*, volume 1, pages 281–287, Vienna, Austria.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA. Association for Computational Linguistics.

Álvaro J Jiménez Palenzuela, Flavius Frasincar, and Maria Mihaela Truşcă. 2022. Modeling second language acquisition with pre-trained neural language models. *Expert Systems with Applications*, 207:117871.

Adithya Renduchintala, Rebecca Knowles, Philipp Koehn, and Jason Eisner. 2016. User modeling in language learning with macaronic texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1859–1869, Berlin, Germany. Association for Computational Linguistics.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany. Association for Computational Linguistics.

Itamar Shatz. 2020. Refining and modifying the efcamdat: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors. 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Jacob Whitehill and Javier Movellan. 2017. Approximately optimal teaching of approximately optimal learners. *IEEE Transactions on Learning Technologies*, 11(2):152–164.

Brian Zylich and Andrew Lan. 2021. Linguistic skill modeling for second language acquisition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, page 141–150, New York, NY, USA. Association for Computing Machinery.