Transformer-based Hebrew NLP models for Short Answer Scoring in Biology

Abigail Gurin Schleifer[1] **Beata Beigman Klebanov**[2] **Moriah Ariely**[1] **Giora Alexandron**[1]
[1] Weizmann Institute of Science, Rehovot, Israel
[2] Educational Testing Service, Princeton, USA
{abigail.gurin-schleifer,moriah.ariely,giora.alexandron}
@weizmann.ac.il
bbeigmanklebanov@ets.org

Abstract

Pre-trained large language models (PLMs) are adaptable to a wide range of downstream tasks by fine-tuning their rich contextual embeddings to the task, often without requiring much task-specific data. In this paper, we explore the use of a recently developed Hebrew PLM – alephBERT – for automated short answer grading of high school biology items. We show that the alephBERT-based system outperforms a strong CNN-based baseline, and that it generalizes unexpectedly well in a zero-shot paradigm to items on an unseen topic that address the same underlying biological concepts, opening up the possibility of automatically assessing new items without item-specific fine-tuning.

1 Introduction

Advances in NLP offer transformative technology to support educational practice, including scoring of constructed (free text) responses in both holistic and analytic fashion. In particular, pre-trained large language models (**PLMs**) hold great promise for applications that require sophisticated context-rich analysis of student responses.

However, progress in PLMs and their applications in English outstrips that in other languages. Recent research in Hebrew NLP made available a new Hebrew PLM – alephBERT (Seker et al., 2022); while it has been shown to be effective for NLP tasks such as POS tagging and NER, its effectiveness for a downstream automated scoring application is an open question.

We evaluate alephBERT-based classifiers for the task of analytic content-scoring of short answers in biology in a formative high school setting, comparing it to a strong CNN-based baseline.

We contribute new knowledge about the effectiveness of BERT-based classifiers in languages other than English for a content-scoring task. Our two key findings are that the alephBERT-based classifiers i) provide a significant improvement over the CNN-based baseline; and ii) generalize surprisingly well to unseen items that deal with the same underlying scientific concepts but in the context of a different topic. We briefly discuss implications of the findings and directions for future work.

2 Related Work

An especially promising application area of NLP is automated analysis of responses to open-ended questions, either in the form of a full essay, where the goal is typically a demonstration of proficiency in writing in a particular genre (Beigman Klebanov and Madnani, 2021), or in the form of short responses, where the goal is typically to demonstrate content knowledge. In this paper, we consider the latter application, often termed Automated Short Answer Grading (**ASAG**).

To date, most of the scientific development on ASAG has been done in English (see Haller et al. (2022) for a survey), including ASAG using PLMs (Bexte et al., 2022; Li et al., 2021; Condor, 2020; Sung et al., 2019a,b), although work on PLMS for ASAG in other languages does exist, e.g., Japanese (Oka et al., 2022), Arabic (Nael et al., 2022).

Recently researchers also used multi-lingual PLMs for ASAG: Schneider et al. (2023) used the LaBSE multilingual transformer model (Feng et al., 2022) for scoring very short responses (the bulk of the responses are 5 words or shorter) in a variety of subjects and in 14 languages. Unfortunately, the authors did not provide a detailed breakdown of performance by language or by subject area, although they did show that numeric responses tended to be easier to score than textual or mixed ones, across multiple languages. Interestingly, while there were relatively few responses in English (1.7K), the system's error on scoring textual responses in English was lower than for Ukranian, which had more than two orders of magnitude more responses than English (500K), which could suggest that languages with smaller digital footprints and therefore less

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 550–555
July 13, 2023 ©2023 Association for Computational Linguistics

data for pre-training the PLMs would still be at a disadvantage even if there are a lot of responses in those languages for the specific task.

The ASAG task for Hebrew was addressed by Ariely et al. (2023). The authors built CNN-based classifiers that used word2vec embeddings; these models will serve as baselines for the current work. Hebrew, like Arabic, is a semitic language where vowels are generally omitted in writing, resulting in substantial ambiguity where the same sequence of written letters can have many meanings depending on context. Therefore, a PLM that implements the latest contextualization advancements holds great promise for ASAG in Hebrew. AlephBERT, the recently introduced Hebrew PLM (Seker et al., 2022), shows SOTA performance on multiple tasks, including morphological and POS tagging and NER. Our goal is to evaluate alephBERT for the ASAG task in Hebrew.

Item	Categories
q1	–,1,–,3,4,5,6,7,8,9,10,11,12
q2	–,–,–,3,4,5,6,7,8,9,10,11,12
q3	0,1,2,3,4,5,6,7,8,9,10,11,12

Table 1: The mapping between items and categories.

Cat	Definition
0	changes in the amount of RBC
1	changes in oxygen levels that bind to HGB/RBC
2	refer to both groups of athlete travelers (q3)
3	the role of HGB/RBC in oxygen transportation
4	changes in oxygen levels in the body
5	changes in oxygen levels in the cells
6	oxygen is a reactant in cellular respiration
7	energy/ATP is produced during cellular resp.
8	changes in cellular respiration rate
9	using the term 'cellular respiration'
10	changes in energy/ATP levels
11	using the term 'energy' or 'ATP'
12	energy is consumed during exercise

Table 2: Category definitions. HGB: hemoglobin; RBC: red blood cells; ATP: energy (adenosine triphosphate).

3 Experimental Setup

3.1 Data

The data consists of responses to open-ended questions on three biology items from 669 students in grades 10-12 from about 25 high schools across Israel. There are thus 669 labeled responses for each of the three items (henceforth, **q1, q2, q3**), scored by a team of content and pedagogy experts with a binary score per category; that is, for every response, there are 10-13 binary labels according to the analytic rubric for the given item.

The items present questions about the effect of smoking (q1), anemia (q2), and travel in high altitude (q3) on physical activity. A very similar analytic rubric is used for all three items to assess students' ability to write causal explanations in biology. The rubric consists of a causal reasoning chain built from 13 categories, each of which evaluates whether a specific scientific fact or causal relation is addressed correctly in a response. Table 1 shows the mapping between the items and the binary analytic categories. Table 2 shows brief definitions of the categories. Figure 1 shows the score distributions per item per category. We observe that item q3 is harder than items q1 and q2 on most categories shared by the three items.

The rubric evaluates the ability to explain step-by-step the causal chain leading to the phenomenon. For example, q1 asks students to explain how high levels of CO make it difficult for smokers to exercise. Two responses are shown below, translated into English. Response 1 was given credit for mentioning the changes in oxygen levels after CO binding to hemoglobin (category 1), for stating the connection between the decreased cellular respiration rates and the reduction in the generation of energy which is necessary for physical activity (8-12). However, the reasoning chain is not articulated fully, since the transfer of oxygen to the cells by red blood cells and the role of oxygen in cellular respiration are not stated (no credit for categories 3-7). Conversely, Response 2 does mention the impairment of oxygen transfer to the body and cells (4 and 5), but does not include the parts of the explanation that connect oxygen to cellular respiration and cellular respiration to production of energy for the physical activity, hence no credit is given on categories 6-12.

Response 1 A cigarette contains several harmful substances, including CO. CO has a strong tendency to bind to hemoglobin found in red blood cells. As a result, less oxygen binds to hemoglobin, which affects the rate of cellular respiration. Because the rate of cellular respiration slows down, less energy is generated in the cells of the body, so the cells do not have enough energy to perform physical activity and difficulty is created. Scores: [-, 1, -, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1]

Response 2 Because those carbon dioxide molecules bind to hemoglobin, the transfer of oxygen to the body's cells is impaired.

> Lack of hemoglobin and oxygen explains the difficulty of people who smoke to exercise. Scores: [-, 1, -,0 ,1 ,1 ,0 ,0 ,0 ,0 ,0 ,0 ,0]

This rubric was developed in consultation with teachers to support in-class formative assessment, for example by assigning students to small study groups based on reasoning types revealed in their response patterns.

The items are typical open-ended questions commonly used (or versions of them) in teaching materials in biology and in the Israeli high school matriculation exam ('Bagrut'). The three items were presented to students in a randomized order. The average length of response is 55, 48, 70 words and standard deviation of 34.5, 27.4, 48 for q1, q2, q3, respectively. The data collection was approved by IRB and includes permission to use the data for research. The data was collected prior to and independently of this study and was previously used in computational experiments of Ariely et al. (2023).

3.2 Experiment design

In this study, we investigate how well an alephBERT classifier performs on analytic ASAG, compared to the CNN-based system of Ariely et al. (2023). We conduct evaluations in two scenarios: (a) within-item, where train and test data come from the same item, and (b) cross-items, where the system is trained on two items and tested on the third. The main goal of the latter evaluation is to address cases where a new item is created that deals with a different application area of the same scientific concept, that is, a new item that would address cellular respiration mechanism in a different real-life application. This is a common pedagogical strategy for creating teaching, practice, and multiple forms of assessment materials.

We partition the students into train, development, and test groups in the 60/20/20 proportions respectively; their responses comprise the q1-train, q1-dev, q1-test sets, and the same for q2 and q3. This is done in order to ensure that responses from the same student do not appear in both train and test data in the evaluations. We build a classifier for each category (13 classifiers in total); while the student responses are the same across categories (we are using the full text of the response), the labels may differ across categories. That is, a given response can have the score of 0 on category 3 and the score of 1 on category 8, as in Response 1 shown in section 3.1.

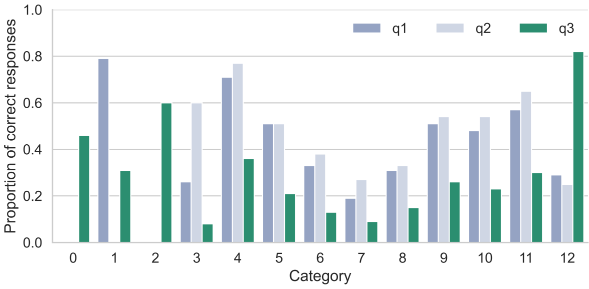

Figure 1: Proportion of correct responses per item per category.

For within-item experiments, we train on q1-train and test on q1-test; same for q2 and q3. For cross-item experiments, we train on the combination of q1-train and q2-train and test on q3-test; the same for the other two permutations of the items. In this design, in addition to benchmarking against prior work, we also compare performance between within-item and cross-item scenarios, e.g., results on q3-test when trained on q3-train vs trained on the combination of q1-train and q2-train.

For evaluation, we use Cohen's κ, per item per category. We also report proportion of categories with $\kappa > 0.6$, to get a sense of the extent to which the rubric as a whole can be automatically scored with reasonable reliability for a formative context. Ariely et al. (2023) reported average performance over 50 iterations of cross-validation for each item and each category; in our context, it is prohibitively time-consuming to run such a large number of evaluations. We report evaluations on q1, q2, and q3 test sets for the alephBERT models; thus, performance estimates for alephBERT are somewhat noisier than for the CNN baseline.

4 Models

4.1 Baseline

For the baseline, we use published results for CNN-based classifiers reported in Ariely et al. (2023), where each classifier predicts whether a certain category is addressed in the response. Pre-processing included tokenizing the input text and performing a morphological and syntactic analysis using Hebrew NLP tools. Word embeddings over a vocabulary of frequently-used morphemes and their part of speech were constructed using Gensim's word2vec CBOW algorithm. The embeddings were fed forward into two consecutive convolutional layers, followed by a fully connected layer and a sigmoid activation function. The embeddings (of size 100) were trained on the entire Hebrew Wikipedia.

4.2 AlephBERT based models

AlephBERT PLM (Seker et al., 2022) is based on the same architecture as the English BERT PLM (Devlin et al., 2018). AlephBERT was designed to handle Hebrew morphology; see Seker et al. (2022) for a detailed description. AlephBERT was trained on a larger corpus than any Hebrew language model before it, including Twitter, Hebrew wiki and the Hebrew portion of the Oscar dataset (Ortiz Suárez et al., 2020). It was not specifically trained on biology or science data beyond the occurrence of these topics in the general corpora. It includes 12 layers, i.e., transformer blocks (768 units per layer), 12 attention heads, the total of 110M parameters and vocabulary size of 52K.

For every category, we built a classifier that uses the alephBERT PLM pre-trained embeddings and an additional classification layer, with sigmoid activation. We fine-tune the models on our training data using cross-entropy loss; all layers of the model are tuned. The learning rate and number of epochs hyperparameters were tuned on dev sets.

5 Results

Table 3 shows the performance of the alephBERT-based system on all <category, item, case> combinations, where case refers to 'within-item' or 'cross-item'. The performance of the CNN baseline is shown as published in Ariely et al. (2023).

5.1 Comparison to CNN baseline

AlephBERT-based models perform significantly better than the baseline, p = 0.016, using the one-sided Wilcoxon signed-rank test (paired) with n = 44 (all <item,category,case> cells in Table 3 that have results for both the models), $\alpha = 0.05$. The largest gain is on category 9 within-item: from κ = .06-.73 (baseline) to κ > .90 (alephBERT). Category 9 looks for a specific phrase ('cellular respiration'). We hypothesize that this improvement is driven by the improved ability of alephBERT to capture the rich token-internal structure of the Hebrew language reported by Seker et al. (2022) based on morpheme-level evaluations.

5.2 Comparison between within-item and cross-item performance

We compare the alephBERT-based within-item models with the cross-items (i.e., zero-shot) models on all <category, item> combinations where both models can be run (see Table 3). The cross-item performance is not significantly worse than within-item, p = 0.9 using the one-sided Wilcoxon signed-rank test (paired), n = 32, $\alpha = 0.05$.

This is a remarkable result, since one would expect a degradation in performance for models that saw no data coming from the test item at train time. In fact, an unseen item on the same biology concept can be scored with a common analytic rubric with $\kappa > 0.6$ on average across categories for each item, which may be sufficient for formative uses and may allow teachers to create and score new items based on a similar rubric on the fly.

We observe a complete failure of cross-item generalization on category 1. This category occurs only in q1 and q3; the cross-item generalization is thus based on one training item. This could compromise the system's ability to zero in on those meaning elements that are common to the two training items and instead overly rely on the specifics of the training item's topic. Category 1 is also more difficult to address well in q3 than in q1 (30% correct vs 78% correct, see Figure 1), further complicating cross-item transfer. Understanding the necessary conditions for transfer is a topic for future research.

6 Conclusions

Pre-trained large language models can be adapted to downstream tasks by fine-tuning their rich contextual embeddings to the task. We explored the recent Hebrew PLM – alephBERT – for short answer grading in high school biology. We found that the alephBERT-based system outperformed a strong baseline and that it generalized unexpectedly well to items on an unseen topic addressing the same biology concepts. The second finding provides evidence in support of the viability of the modular design of the rubric – not only is it the case that human raters were able to reliably assess different items with subsets of the same analytic categories, but an automated model was likewise able to zero in on the commonalities in the way categories are manifested in student responses across multiple topics.

The cross-item generalization has exciting implications for educational practice, as this may allow teachers to create and automatically score new items based on a similar rubric on the fly. A study of this possibility with teachers and an improvement of our understanding of the conditions neces-

	Model→	Ariely2023				AlephBERT					
	Item→	**q**1	**q**2	**q**3		**q**1		**q**2		**q**3	
Category↓	Case→	W-I	W-I	W-I	C-I	W-I	C-I	W-I	C-I	W-I	C-I
0				.71						.81	
1		.53		.76		.72	.00			.61	.01
2				.70						.88	
3		.60	.73	.00	.48	.75	.43	.62	.54	.00	.67
4		.61	.52	.60	.38	.50	.61	.35	.05	.71	.47
5		.80	.75	.57	.76	.90	.76	.73	.66	.81	.79
6		.66	.72	.32	.71	.65	.69	.71	.68	.66	.59
7		.71	.80	.47	.61	.68	.51	.73	.78	.50	.76
8		.95	.93	.93	.70	.85	.86	.93	.72	.32	.82
9		.46	.73	.06	.95	.99	.97	.96	.97	.94	.96
10		.83	.80	.60	.80	.88	.88	.65	.71	.97	.87
11		.91	.90	.90	.93	.97	.97	.88	.87	.95	.95
12		.68	.57	.00	.61	.74	.00	.73	.75	.00	.54
Av		.70	.75	.51	.69	.78	.61	.73	.67	.63	.68
%κ>.60		73	80	38	80	91	64	90	80	69	64

Table 3: Average Cohen's κ per item (q1-q3) per category (0-12), for the baseline as reported in Ariely et al. (2023) and alephBERT models. W-I: within-item (gray); C-I: cross-items. The last row shows % of categories with $\kappa > 0.6$.

sary for the successful transfer to occur are two of the directions of our future work, as well as further enhancement of the scoring system.

Acknowledgements

This research was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center.

References

Moriah Ariely, Tanya Nazaretsky, and Giora Alexandron. 2023. Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*, 33(1):1–34.

Beata Beigman Klebanov and Nitin Madnani. 2021. Automated essay scoring. *Synthesis Lectures on Human Language Technologies*, 14(5):1–314.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2022. Similarity-based content scoring-how to make s-bert keep up with bert. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123.

Aubrey Condor. 2020. Exploring automatic short answer grading as a tool to assist in human rating. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 74–79. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: From word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Zhaohui Li, Yajur Tomar, and Rebecca J. Passonneau. 2021. A semantic feature-wise transformation relation network for automatic short answer grading. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6030–6040, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Omar Nael, Youssef ELmanyalawy, and Nada Sharaf. 2022. AraScore: A deep learning-based system for Arabic short answer scoring. *Array*, 13:100109.

Haruki Oka, Hung Tuan Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa, and Tsunenori Ishioka. 2022. Fully automated short answer scoring of the trial tests for common entrance examinations for japanese university. In *Artificial Intelligence in Education:*

23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I, pages 180–192. Springer.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Johannes Schneider, Robin Richner, and Micha Riser. 2023. Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, 33(1):88–118.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56, Dublin, Ireland. Association for Computational Linguistics.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019a. Pre-training bert on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6071–6075.

Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019b. Improving short answer grading using transformer-based pre-training. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*, pages 469–481. Springer.