Anna Karenina Strikes Again:
Pre-Trained LLM Embeddings May Favor High-Performing Learners

Abigail Gurin Schleifer[1]   **Beata Beigman Klebanov**[2]   **Moriah Ariely**[1]   **Giora Alexandron**[1]

[1] Weizmann Institute of Science, Rehovot, Israel

[2] ETS, Princeton, USA

{abigail.gurin-schleifer,moriah.ariely,giora.alexandron}
@weizmann.ac.il
bbeigmanklebanov@ets.org

Abstract

Unsupervised clustering of student responses to open-ended questions into behavioral and cognitive profiles using pre-trained LLM embeddings is an emerging technique, but little is known about how well this captures pedagogically meaningful information. We investigate this in the context of student responses to open-ended questions in biology, which were previously analyzed and clustered by experts into theory-driven Knowledge Profiles (KPs). Comparing these KPs to ones discovered by purely data-driven clustering techniques, we report poor discoverability of most KPs, except for the ones including the correct answers. We trace this 'discoverability bias' to the representations of KPs in the pre-trained LLM embeddings space.

1 Introduction

Classifying students into behavioral or cognitive profiles using unsupervised cluster analysis techniques is a common application of machine learning to educational data (Le Quy et al., 2023; Martin et al., 2023; Ariely et al., 2024; Rastrollo-Guerrero et al., 2020; Bovo et al., 2013). Recently, there has been a growing interest in applying this methodology to textual student responses that are decoded using pre-trained large language models into vectorized embeddings in semantic spaces (Martin et al., 2023; Wulff et al., 2022; Masala et al., 2021). The operational appeal of this approach is that it minimizes the need for expert knowledge, which is costly to inject through human labeling procedures (Nehm and Haertig, 2012; Tansomboon et al., 2017; Li et al., 2023; Ariely et al., 2024). However, the validity of patterns discovered this way depends on the ability of the embeddings to maintain the pedagogically meaningful information that existed in the original, textual representations of responses (Devlin et al., 2018; Seker et al., 2022) and of the algorithmic method to discover them. Evaluation of

emergent profiles is often done in terms of the internal quality of the clustering, as data is usually not available to estimate the extent to which the discovered profiles align with a pedagogically meaningful representation of the responses. Without such an evaluation, a loss of important information can be overlooked, potentially leading to sub-optimal educational decisions that rely on this analysis (Le Quy et al., 2023).

To investigate whether this hypothesized risk manifests in real-life educational context, we utilize student answers to two constructed response questions in high school biology. The data was previously analyzed by a team of biology education researchers and experienced teachers, and graded according to a theory-driven detailed analytic rubric that is based upon the Causal-Mechanical Explanation framework (Ariely et al., 2024; Salmon, 2006). The rubric contained 10 (item 1) or 11 (item 2) binary categories, each checking for the occurrence of a specific key piece of information in the response. Using these human-generated binary vectors of length 10 (11), the responses were clustered using a KMeans algorithm into a set of 6 (7) Knowledge Profiles (**KPs**) that were found by teachers to encapsulate specific patterns of errors.

The validity of the KPs was evaluated in several ways. First, human experts conducted a qualitative analysis to assess whether each KP captures a specific and distinct pattern of errors. Second, we analyzed the results computationally, showing that i) the KPs were consistent across the two items, namely, revealing the same type of conceptual errors; and ii) the *learners* tended to exhibit the same type of conceptual error (KP) in both items. Third, we conducted an in-class formative assessment intervention study that provided automated guidance to students based on their KP, and showed significant improvement in their performance on a different prompt that measures the same conceptual knowledge. These analyses provided strong evi-

dence that the KPs capture pedagogically meaningful information (for full details, see Ariely et al. (2022, 2024)).

Using these data, we are in a position to answer two research questions:

RQ1 What is the correspondence between clusters that are computed from pre-trained LLM embeddings of student responses and theory-based KPs?

To preview the result, we find that two clustering techniques that are commonly used for such tasks (Le Quy et al., 2023) – KMeans (Lloyd, 1982) and HDBSCAN (McInnes et al., 2017) – largely fail to discover the KPs though retrieval is somewhat better for the profile containing the correct responses. Following up on this finding, we go 'upstream', to the pre-trained embeddings, and investigate:

RQ2 How well are the KPs represented in the pre-trained embeddings space?

Our results reveal a strong relationship between the quality of the responses in the profile (correct or various degrees of incorrect) and the shape and density of its embeddings-based representation. We refer to this relationship as an 'Anna Karenina principle' and tie it to the profile discovery failure we observed in RQ1.

The contribution of this work is twofold. First, it is the first to demonstrate the Anna Karenina principle in the context of pre-trained representation of student responses to open-ended questions. Second, our results suggest that, in some cases, out-of-the-box pre-trained LLM embeddings may be a pedagogically unsound basis for profile discovery.

2 Related Work

2.1 NLP-based profiling of constructed responses in science education

Open-ended items require students to develop and construct their answers, reflect on their knowledge, and integrate it with new ideas (Fellows, 1994). Reasoning and evidence-based defense of an argument is key for testing scientific hypotheses (Toulmin, 2003). Therefore, constructing causal explanations is an essential skill for students of science to learn (Ariely et al., 2024; Martin et al., 2023); practice and high-quality feedback are key elements in helping students master the skill (Hattie and Timperley, 2007; Gerard and Linn, 2016; Tansomboon et al., 2017).

Analyzing open-ended items to provide feedback is a time-consuming, complex task. Automating some of the analyses for assessment and feedback purposes is promising for supporting teaching and learning (Tansomboon et al., 2017; Gerard and Linn, 2016; Ariely et al., 2023).

Most systems for automated evaluation of scientific explanations to date had been designed in the supervised machine learning framework (Schleifer et al., 2023; Sung et al., 2019; Riordan et al., 2020; Kumar et al., 2019; Mizumoto et al., 2019; Li et al., 2021). Among the unsupervised approaches, Masala et al. (2021) extracted the main takeaways from students' feedback on different components in academic courses, using KMeans to cluster pre-trained BERT embeddings of students' feedback. Martin et al. (2023) applied HDBSCAN over pre-trained LLM embeddings and to find emergent argumentation patterns' characteristics. Wulff et al. (2022) investigated HDBSCAN clustering over LLM embeddings to evaluate the attention of preservice physics teachers to classroom events elicited from open-ended text responses. A semi-supervised coding method in which homogeneous clusters receive the same coding automatically and heterogeneous clusters are fully labeled by humans was proposed by Andersen et al. (2023) and applied to student responses to PISA items.

2.2 Biases in pre-trained LLMs

While LLMs are powerful meaning representations that undergird the state-of-art systems on a wide range of NLP tasks, they are also known to exhibit a plethora of social biases that could lead to social harm when the models are used in downstream tasks (Bender et al., 2021). In a recent review of the current state of research on LLM bias evaluation, Goldfarb-Tarrant et al. (2023) criticize the field for focusing heavily on the upstream, pre-trained LLMs, in most cases without considering the connection to a specific task the LLMs is being put to (68% of the reports reviewed), citing this as a threat to the predictive validity of bias measurements.

In fact, the literature that does consider the connection between upstream (intrinsic) and downstream (extrinsic) behavior suggests that it is not straightforward. Considering static embeddings (e.g., word2vec) and a commonly used bias test, the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), Goldfarb-Tarrant and colleagues (Goldfarb-Tarrant et al., 2021) found no

clear relationship with performance of models using the embeddings, as measured by differences in precision and recall of retrieval of the target construct on data from privileged and non-privileged groups. Extension to contextual embeddings and a wider range of tasks and measures yielded similar results (Cao et al., 2022; Kaneko et al., 2022). Our contribution extends the discussion towards social constructs beyond the typically considered demographic attributes such as gender, race, ethnicity, age towards a distinction that is particularly relevant when dealing with learner data – that of learners at the more or less advanced state of understanding of the phenomenon under consideration. We are not aware of prior work comparing LLM representations based on knowledge-related profiles; the closest finding in the literature are examples of poorer performance of LLM-based systems on data produced by English language learners with respect to native speakers of English (Baffour et al., 2023). Additionally, we explore LLMs in a relatively low-resource language (Hebrew) in contrast to the bulk of current work that focuses on English or other high-resource languages: In the 90 LLM bias studies evaluated by Goldfarb-Tarrant et al. (2023), only two report results in a language that is not highly resourced.

3 Data

The data consists of 669 student responses to two open-ended items in high-school biology, collected anonymously from students in grades 10-12 from about 25 high schools of varied demographics and socioeconomic status (based on location) across Israel. Gender distribution was 70% females (typical to the gender distribution among high-school biology majors in Israel). The items deal with the connection between respiration and energy in physical activity in the context of smoking (**Q1**) and anemia (**Q2**), taught as part of the core topic "The human body". The items were human-scored using a similar analytic rubric containing 10 (Q1) or 11 (Q2) categories (Ariely et al., 2024). All rubric categories are binary, each targeting specific information that needs to be mentioned in a correct response, such as "the role of hemoglobin in oxygen transportation" or "changes in cellular respiration rate". The resulting binary vectors were clustered using KMeans; the clusters were analyzed by experienced teachers and ranked from 1 to 6 (Q1) or 7 (Q2) with larger numbers corresponding to

clusters with more severe errors. We denote these clusters *Knowledge Profiles*, and index them from 1 (KP1) to 7 (KP7). See Ariely et al. (2024) for a full description of the items and the assessment framework. The items and examples of student responses and their mapping into KPs can be found in Appendix 1.

For the purposes of the analysis presented in this paper, all responses were represented using rich contextualized vectors – embeddings produced by a pre-trained Large Language Model (LLM). The LLM being used, AlephBERT (Seker et al., 2022), is state-of-the-art for Hebrew. It was trained on a large corpus of the Hebrew language, including: Twitter tweets, Hebrew Wikipedia, and the Hebrew subset of the Oscar (Suárez et al., 2020) dataset. AlephBERT has the same architecture as BERT (Devlin et al., 2018): 12 layers, 110M parameters, and 12 attention heads. It was trained on a 52K-word Hebrew vocabulary on masked token prediction task, and on the Hebrew language tasks: word segmentation, part-of-speech tagging, and full morphological tagging. It was further trained on the tasks of sentiment analysis and named entity recognition.

4 Methods

To evaluate whether raw LLM embeddings carry useful knowledge for unsupervised profiling of responses, we experimented with two common clustering approaches (Le Quy et al., 2023), KMeans (Lloyd, 1982) and HDBSCAN (McInnes et al., 2017), which implement different clustering mechanisms. The first discovers convex-shaped clusters; its mechanism is centroid-based and applies an Euclidean distance function. The second is density-based and can be applied with various distance metrics, e.g., a metric induced by cosine-similarity, and the clusters may have various shapes. Both approaches were used previously for profile discovery in constructed response data (Ariely et al., 2024; Martin et al., 2023; Wulff et al., 2022). Experiments were conducted in Python, using scikit-learn (Pedregosa et al., 2011), SBERT (Reimers and Gurevych, 2019) and Pytorch (Paszke et al., 2019).

4.1 KMeans

The KMeans is a widely used algorithm (Lloyd, 1982). The algorithm is initiated with a specified number of clusters and a random initialization of

their centroids. The clustering approach minimizes the within-cluster sum of squared distances, i.e., Euclidean distance. KMeans clusters are convex and all samples are assigned to a cluster, i.e., there are no outliers. Convexity means that for every two points in the cluster, a straight line between them also lies within the cluster.

4.2 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

Another clustering approach, which is more promising in the context of LLMs' embeddings (Martin et al., 2023) is the HDBSCAN (McInnes et al., 2017) algorithm. The approach here is creating a mutual reachability graph where *core* samples are points in areas of high density. A cluster is a set of *core* samples and a set of *non-core* samples that are neighbors of *core* samples but are not *core* themselves. *Non-core* samples are at the fringes of clusters. A *core* sample is such that there are 'min_samples' other samples with a distance less than ϵ from it, for some $\epsilon > 0$ (Pedregosa et al., 2011). The HDBSCAN mechanism performs clustering for various ϵ values and the most stable clustering is chosen.

The default metric for HDBSCAN is Euclidean distance. To use cosine similarity, we turn it into a distance function (McInnes et al., 2017):

$$\|x - y\| = \sqrt{2 \times (1 - CosSim(x,y))}, \quad (1)$$

where x, y are unit vectors, i.e., $\|x\| = \|y\| = 1$ (Manning, 1999). Since cosine similarity does not depend on vectors' magnitude, only on the angle between the two vectors, we first turned every embedding e_i to a unit vector $\frac{e_i}{\|e_i\|}$ and then applied the HDBSCAN on a pre-computed metric matrix consisting of all pairwise distances between all embeddings in the dataset using formula (1).

In contrast to the KMeans, HDBSCAN can find clusters with varied densities and clusters may have non-convex shapes.

4.3 Metrics for Comparing Clusters

To compare the similarity between the KPs and the cluster assignments of the KMeans/HDBSCAN, we used Adjusted Rand Index (ARI) (Vinh et al., 2009). In ARI, similarity is interpreted as the number of pairs of items on which the clusterings agree, adjusted for the amount of chance agreement. Let D be a dataset containing n items that are classified

into m clusters by clustering C and, independently, into k clusters by clustering E. For a pair of items $(i_1, i_2) \in D$, C and E agree on it iff i_1 and i_2 are either (1) assigned to the same cluster in both C and E (let's say there are a such pairs), or (2) assigned to different clusters in both C and E (let's say there are b such pairs). Now, $a + b$ is the number of agreements between C and E. The ARI index is given by:

$$RI = \frac{a + b}{\binom{n}{2}} \quad ; \quad ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad ,$$

where $E[RI]$ is the expected RI for some random label assignment (Vinh et al., 2009), and $max(RI)$ equals to 1. The ARI values range from -1 to 1, where 1 indicates perfect agreement, and -1 indicates complete disagreement (Hubert and Arabie, 1985). Since each student response in our dataset is labeled with its KP, we evaluated the ARI for each clustering assignment, i.e., KMeans and HDBSCAN, compared to the KPs. This yields a *global* comparison between the KPs and each clustering assignment.

To evaluate the 'discoverability' of each KP, we also conducted **by-KP analysis**, applying a retrieval paradigm and considering each cluster as an attempt to retrieve each of the KPs. We calculate *recall*, *precision*, and *F1* score using a contingency matrix $A = (a_{mn})_{1 \le m \le k, \ 1 \le n \le f}$ where rows are the KPs $k = 6, 7$, and columns are the unsupervised clusters C_n found by KMeans or HDBSCAN, $f = \#fitted\_clusters$;

$$a_{mn} = \sum_{1 \le m \le k, \ 1 \le n \le f} \left| \{x : \ x \in KP_m \cap C_n\} \right|$$

the cell a_{mn} in the matrix A counts the number of members of KP_m that fell in cluster C_n. The precision of retrieval of KP_m using cluster C_n is $P_{mn} = \frac{a_{mn}}{|C(n)|}$; the recall is $R_{mn} = \frac{a_{mn}}{|KP(m)|}$. F1 score is $F_{mn} = \frac{2 \cdot P_{mn} \cdot R_{mn}}{P_{mn} + R_{mn}}$, indicating the extent to which we were able to retrieve KP_m using the emergent cluster C_n.

5 Results

5.1 RQ1: Correspondence between embedding-based clusters and theory-based Knowledge Profiles

5.1.1 Global alignment between the clusterings

As described in Section 4, we evaluated the agreement between clusterings that were computed from

the embeddings using two cluster analysis methods: KMeans and HDBSCAN. As we were interested in upper-bounding the discoverability of the KPs by both algorithms, we "helped" them with additional information (the number of clusters to the KMeans algorithm, and allowing the HDBSCAN to grid search for 'good' hyperparameters). With k equals the number of KPs per item (six for Q1 and seven for Q2), the resulting ARIs for the KMeans were **0.122** and **0.191**, for Q1 and Q2, respectively. For the HDBSCAN algorithm, we conducted a grid search for two parameters: $min\_cluster\_size$, i.e., the minimum number of samples in a cluster (values: $\{3, 4, 5, 10, 15, 20, 30, 40\}$), and $min\_samples$, i.e., the number of samples in a neighborhood for a point to be considered as a core point (values: $\{1, 2, 3, 4, 5\}$). We report the best-performing combination in terms of ARI: **0.037** for Q1 (with $min\_cluster = 5$, $min\_samples = 2$), and **0.038** for Q2 (with $min\_cluster = 3$, $min\_samples = 3$). Based on these results, we conclude that the clusters discovered by the KMeans had low agreement with the KPs, and the clusters discovered by the HDBSCAN had negligible agreement with the KPs.

5.1.2 Discoverability of specific KPs

We further investigated the clusters' matching quality by calculating the F1 score *per KP* for each of Q1 and Q2. For KMeans, the results show good retrieval of KP1, the cluster with the highest-quality responses – $F1 = 0.60, \; 0.67$ for items Q1 and Q2 respectively – but much worse retrieval of the other KPs, with maximal $F1 = 0.40$ for KP6 in Q1 and $F1 = 0.47$ for KP2 in Q2. The clustering results in terms of contingency tables and F1 Scores are presented in Tables 1 to 4, with KPs as rows and columns as fitted clusters. The maximum F1 scores per profile are shaded in gray.

The evaluation of HDBSCAN clusters mirrored that of KMeans, showing better retrieval of KP1 – $0.43, 0.46$ F1 scores for Q1 and Q2 – than of any other profile, with maximal $F1 = 0.36$ for KP6 in Q1 and $F1 = 0.29$ for KP2 in Q2. We observe that, overall, results are worse for HDBSCAN than for KMeans. The clustering results in terms of contingency tables and F1 Scores are presented in Tables 5 to 8, with KPs as rows and columns as fitted clusters. The maximum F1 scores are shaded in gray.

We then considered the possibility that more coarse-grained profiles might emerge from the clustering than the detailed KPs. To this end, we tried different options for grouping KPs and calculated the F1 scores between fitted clusters and the grouped KPs. The best results show an emergent pattern consistent in both items Q1 and Q2, where one cluster consists of the higher-quality responses (KP1-KP4) an is retrievable with $F1 = 0.72, 0.74$, and the other cluster consists of lower-quality responses (KP5-KP7), retrievable with $F1 = 0.52, 0.45$.

We tried this approach with the KMeans as well, but the samples scattering across fitted clusters there did not exhibit meaningful patterns.

5.2 RQ2: How well are the KPs represented in the embeddings?

To answer this question, we first analyzed, descriptively, the level of similarity between the embeddings within each KP, and between KPs. We then conducted statistical tests to verify that the observed patterns are statistically robust.

Within KP similarity. To analyze the level of similarity within each KP, we computed the pairwise cosine-similarity between all pairs in that KP. Tables 9 and 10 show the results, with KPs as rows and fitted clusters as columns. Within-KP similarities are in the diagonals. Since the pairwise cosine similarity values are not normally distributed[1], we report medians. The results show that KP1's embeddings (highest quality responses) have the highest density; as the quality of a response goes down, so does its similarity to other responses with the same pattern of error.

Between-KP similarity. As can be further seen in Tables 9 and 10, for both items, for every $i > 1$ the embeddings of KP_i responses tend to be more similar to the embeddings of KP1 than to embeddings of their own KP (the bolded values in the first row are the largest in each column). This means that erroneous responses of various types are more similar to the correct responses than to those with the same pattern of error.

Hypothesis testing. Next, we conducted statistical tests to confirm that i) the distribution of the embeddings in each KP are indeed different and that ii) the cosine similarity within each KP is correlated with the responses quality.

i) A Kruskal-Wallis H-test confirmed that at least one of the medians for the different KPs is signif-

---

[1]The two-sided Kolmogorov-Smirnov test: test statistic = $0.5, p < 0.001$

	A	B	C	D	E	F
KP1	0	102	0	0	5	24
KP2	3	39	0	0	12	37
KP3	3	22	0	0	13	65
KP4	13	28	0	0	26	39
KP5	12	14	0	0	36	50
KP6	31	5	3	16	55	16

Table 1: Q1 KMeans contingency matrix.

	A	B	C	D	E	F
KP1	.00	.60	.00	.00	.04	.13
KP2	.04	.26	.00	.00	.10	.23
KP3	.04	.14	.00	.00	.10	.39
KP4	.15	.18	.00	.00	.21	.23
KP5	.14	.09	.00	.00	.28	.29
KP6	.33	.03	.05	.23	.40	.09

Table 2: Q1 KMeans F1 Scores.

	A	B	C	D	E	F	G
KP1	10	0	2	13	13	127	0
KP2	51	0	22	13	4	30	0
KP3	11	0	10	21	14	28	0
KP4	12	0	18	25	16	13	5
KP5	5	0	16	23	29	6	1
KP6	3	0	29	8	8	6	10
KP7	4	9	20	11	8	3	12

Table 3: Q2 KMeans contingency matrix.

	A	B	C	D	E	F	G
KP1	.08	0	.01	.09	.10	.67	0
KP2	.47	0	.19	.11	.04	.18	0
KP3	.12	0	.10	.21	.16	.19	0
KP4	.13	0	.17	.25	.18	.09	.09
KP5	.06	0	.16	.24	.34	.04	.02
KP6	.04	0	.32	.09	.10	.04	.22
KP7	.05	.24	.22	.12	.10	.02	.25

Table 4: Q2 KMeans F1 Scores.

icantly different from the others, for both Q1 and Q2 (Q1: $statistic = 338.435, p < 0.001$; Q2: $statistic = 295.019, p < 0.001$). A follow-up Dunn's post-hoc analysis indicated that the within-KPs similarities differ significantly across all KP pairs, for both Q1 and Q2, with $p < 0.001$. This indicates that embeddings of responses from different KPs have different distributions. Moreover, the embeddings of high-quality responses are highly dense, while embeddings of low-quality responses are more scattered in the vector space.

ii) To show that the cosine similarities within KPs are significantly correlated with the responses' quality, we calculated for every sample $x \in KP(i)$ its cosine similarity to $KP(i)$ centroid $c(i)$, where $c(i)$ is the average embedding component-wise of the embeddings in $KP(i)$, i.e.,

$$CosSim(x, c(i)) \quad \forall x \in KP(i).$$

We then calculated the Spearman correlation between all the similarities values $\cup_{i=1}^{k}\{CosSim(x, c(i)) : x \in KP(i)\}$ where $k = 6, 7$ for Q1, Q2 respectively, and the ordinal variable of the KPs' index, where lower index represents higher-quality responses. The Spearman correlation coefficient and its p-values are:

$$r_{Q1} = -0.686, \ p < 0.001$$

$$r_{Q2} = -0.633, p < 0.001$$

indicating a strong correlation (Xiao et al., 2016) between the quality of a 'family of responses' (KP) and the within-family similarity.

6 Discussion and Conclusion

Our data consists of 669 high-school student responses to two typical constructed response items in high-school biology. The responses were human graded according to an analytic rubric that is based on the Causal-Mechanical explanation framework (Ariely et al., 2023), transforming each response to a binary vector that encodes the grading according to the rubric categories. Previous work demonstrated that applying cluster analysis (KMeans) to these vectors, which result from a process that applies a theoretical assessment framework to concrete context by human experts, yields stable clusters that reveal pedagogically meaningful knowledge profiles, which were validated in several ways (Ariely et al., 2024). (For more details, see Section 3.) We reasoned that given the successful performance of pre-trained LLMs on a variety of education-related meaning-intensive tasks (Schleifer et al., 2023; Wambsganss et al., 2023; Riordan et al., 2020; Sung et al., 2019), and previous work that applied this specifically to profile discovery (Martin et al., 2023; Wulff et al., 2022), we want to evaluate whether unsupervised profile discovery that is not aided by human knowledge works sufficiently well to be applied out-of-

	A	B	C	D
KP1	19	0	112	0
KP2	29	0	62	0
KP3	36	0	67	0
KP4	49	0	57	0
KP5	61	0	51	0
KP6	71	7	43	5

Table 5: Q1 HDBSCAN contingency matrix.

	A	B	C	D
KP1	.10	0	.43	0
KP2	.16	0	.26	0
KP3	.20	0	.27	0
KP4	.26	0	.23	0
KP5	.32	0	.20	0
KP6	.36	.11	.17	.08

Table 6: Q1 HDBSCAN F1 Scores.

	A	B	C	D
KP1	23	0	142	0
KP2	39	0	81	0
KP3	27	0	57	0
KP4	30	0	56	3
KP5	32	0	48	0
KP6	33	0	31	0
KP7	32	3	32	0

Table 7: Q2 HDBSCAN contingency matrix.

	A	B	C	D
KP1	.12	0	.46	0
KP2	.23	0	.29	0
KP3	.18	0	.21	0
KP4	.20	0	.21	.07
KP5	.22	0	.18	0
KP6	.24	0	.12	0
KP7	.23	.09	.12	0

Table 8: Q2 HDBSCAN F1 Scores.

the-box.

The results of RQ1 reveal that two distinct common unsupervised clustering techniques largely failed to discover the 'gold' KPs from the pre-trained LLM embeddings. Inasmuch as a weak relationship with the knowledge profiles was exhibited by KMeans clusters (ARI of 0.12-0.19), our retrieval-based analysis per profile showed that KP1, the profile that captures the correct responses, was the most discoverable profile, with F1-scores of 0.60/0.67 (on Q1/Q2) for retrieving members of KP1 using the best-aligned emergent cluster. Thus, had the emergent clusters been used as a basis for feedback, only the correct responses would have received pedagogically cogent feedback, since responses belonging to low-knowledge KPs are all intermixed in the emergent clusters. This phenomenon was consistent across two items – Q1 and Q2 – that were analyzed separately.

In an attempt to account for both the failure of overall profile discovery based on pre-trained LLM embeddings and for the bias towards the correct responses exhibited by the emergent clusters, we turned 'upstream' to inspect how the KPs are represented by the embeddings.

We found that the lower the knowledge level of the profile, the less similar to each other its members are in the embeddings space. It is this property that we refer to as the **Anna Karenina principle**: Analogously to Tolstoy's observation that happy families are similar to each other whereas each

unhappy family is unhappy in its own way, we see that the correct responses are similar to each other, whereas incorrect responses differ more from each other the more incorrect ('unhappy') they are (strong correlation of $r_1 = -0.686, r_2 = -0.633$ for Q1/Q2). One could say that Tolstoy considered all unhappy families as an undifferentiated mass; presumably, if classified by their specific source of unhappiness (by family therapists, say), profiles would have probably emerged. In our case, the incorrect responses are grouped by teachers according to the type of problem they exhibit; however, within-profile similarities are still lower than those of correct responses and drift further apart the bigger the problems. The lower density of the poor-knowledge profiles may be one reason that inhibits their downstream discovery.

Further analysis suggests that the privileged status of 'happy families' (correct responses) extends beyond their higher density. We also found that while an average correct response is most similar to another correct response, an average incorrect response is closer in the embeddings space to a correct response than to a member of its own profile (Tables 9 and 10). That is, in some sense, the correct responses are the center of the universe, whereas the incorrect responses drift around them in non-convex formations. The non-convexity of the lower-knowledge profiles may be another inhibitor of their downstream discoverability. Taken together, the 'classic' Anna Karenina property and

KP	1	2	3	4	5	6
1	**.920**	**.910**	**.900**	**.899**	**.883**	**.852**
2		.903	.896	.889	.880	.849
3			.897	.890	.883	.852
4				.877	.871	.837
5					.874	.845
6						.755

Table 9: Q1 pairwise cosine similarity median per KP.

KP	1	2	3	4	5	6	7
1	**.916**	**.891**	**.896**	**.886**	**.873**	**.857**	**.837**
2		.876	.870	.860	.854	.840	.820
3			.881	.870	.862	.844	.828
4				.866	.860	.843	.822
5					.861	.843	.823
6						.824	.796
7							.764

Table 10: Q2 pairwise cosine similarity median per KP.

the strong version that puts the correct responses in the center suggest an explanation for both the overall discovery failure observed downstream and for the bias in favor of correct responses exhibited by the emergent clusters.

Based on our results, pre-trained representations may not lend themselves to making the necessary distinctions to support pedagogical decisions such as providing formative feedback that targets specific errors in student reasoning. In particular, our results show a case where the representations are not sufficiently nuanced to allow commonly used clustering methods to identify any error-based profiles, only the profile of the correct responses. Since it is the students who gave the incorrect responses who are in most need of targeted formative feedback, the bias in favor of correct responses is especially counter-productive. Thus, our results tell a cautionary tale about using emergent properties of student response data built over pre-trained embeddings without domain- and task-specific tuning, and without human supervision.

6.1 Limitations

It is possible that other clustering approaches could have revealed clusters that are more similar to the 'gold' ones. However, given that despite the large difference between KMeans and HDBSCAN's algorithmic approach, they were quite consistent in both demonstrating poor overall agreement and being biased towards discovering the best KP, we believe that reaching results that are qualitatively different from another clustering method is unlikely. It is also possible that emergent clusters do correspond to an alternative meaningful partition of the responses into groups, but that partition is not what educators see when they analyze student responses.

The AlephBERT model used in this paper is state-of-the-art for Hebrew, but it has a smaller number of parameters compared to the most recent LLMs for English. It is possible that with more advanced LLM technology, the LLM representations of student responses will be more nuanced; we will revisit our analyses with larger Hebrew LLMs when available.

Due to the monolingual nature of our current data, we have experimented with one language only. Work is underway to collect comparable student response data in Arabic.

Ethics statement

We acknowledge that the work is conformant with the ACL Code of Ethics. The research and its data collection procedures were approved by the Institutional Review Board and the Ministry of Education. The instrument was administered to the students as part of the regular instruction of the topic, based on the teachers' decision to use it as part of the teaching routine (the instrument was published in teachers' forums), with teacher and school principal approval that response data will be used for

research.

The goal of this research is to better understand the relations between pre-trained LLM-based representations of student responses to open-ended questions in science, and representations of these responses according to theory-driven rubrics applied by human experts. We study to what extent conceptually/pedagogically similar responses tend to maintain their proximity in the embedding space as well, and the impact of deviations from this property on downstream analysis. What makes this especially relevant to Ethics is our finding that the weaker students are the ones whose responses suffer the most from representation mismatches between the two representation spaces. This limits the ability to automatically cater to these students – the ones who are in the highest need for personalized guidance – with formative feedback that matches the gaps in their reasoning. By identifying and naming this phenomenon ('the Anna Karenina principle' in automated short answer evaluation), we hope to start a discussion on the means to both estimate its prevalence and to address it.

We demonstrate the Anna Karenina principle on two tasks with one pre-trained model. It is possible that results will look different with other tasks and other large language models. There is a potential danger of over-generalization based on our results, whereby large language models, as a species, so to speak, would be thought to suffer from the Anna Karenina principle and their off-the-shelf use would be avoided in learner-focused applications. This, in turn, could hamper development of useful LLM-based applications to support learners. We believe that the best course of action is to continue the study of the principle in order to improve our understanding of what kind of models are likely to exhibit the problem and for what kind of task, as well as how to diagnose and correct it, ideally without recourse to a large human-tagged dataset. In parallel, future ethics-focused research could investigate whether weaker learners should be a protected category in educational applications, akin to demographic categories like race or gender, by investigating evidence of harm differentially wrought on such learners through technology that does not cater sufficiently precisely to their needs.

Data from this research cannot be shared publicly due to privacy regulations, but may be provided for research purposes, along with its analysis code, subject to the necessary approvals.

Acknowledgments

References

Nico Andersen, Fabian Zehner, and Frank Goldhammer. 2023. Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning*, 39(3):841–854.

Moriah Ariely, Tanya Nazaretsky, and Giora Alexandron. 2022. Personalized automated formative feedback can support students in generating causal explanations in biology. In *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 953-956*. International Society of the Learning Sciences.

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.

Moriah Ariely, Tanya Nazaretsky, and Giora Alexandron. 2024. Causal-mechanical explanations in biology: Applying automated assessment for personalized learning in the science classroom. *Journal of Research in Science Teaching*.

Perpetual Baffour, Tor Saxberg, and Scott Crossley. 2023. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*, pages 610–623.

Angela Bovo, Stéphane Sanchez, Olivier Héguy, and Yves Duthen. 2013. Clustering moodle data as a tool for profiling students. In *2013 Second international conference on E-Learning and E-Technologies in education (ICEEE)*, pages 121–126. IEEE.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the ACL*, pages 561–570.

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

Nancy J Fellows. 1994. A window into thinking: Using student writing to understand conceptual change in science learning. *Journal of Research in Science Teaching*, 31(9):985–1001.

Libby F Gerard and Marcia C Linn. 2016. Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education*, 27(1):111–129.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the ACL*, pages 1926–1940.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring <mask>: evaluating bias evaluation in language models. In *Findings of the ACL*, pages 2209–2225.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, pages 193–218.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of COLING*, pages 1299–1310.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Get IT scored using AutoSAS — an automated system for scoring short answers. In *Proceedings of AAAI*, pages 9662–9669.

Tai Le Quy, Gunnar Friege, and Eirini Ntoutsi. 2023. A review of clustering models in educational data science toward fairness-aware learning. In Alejandro Peña-Ayala, editor, *Educational Data Science: Essentials, Approaches, and Tendencies: Proactive Education based on Empirical Big Data Evidence*, pages 43–94. Springer Nature Singapore, Singapore.

Tingting Li, Emily Reigh, Peng He, and Emily Adah Miller. 2023. Can we and should we use artificial intelligence for formative assessment in science? *Journal of Research in Science Teaching*, 60(6):1385–1389.

Zhaohui Li, Yajur Tomar, and Rebecca J. Passonneau. 2021. A semantic feature-wise transformation relation network for automatic short answer grading. In *Proceedings of EMNLP*, pages 6030–6040.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Schutze H. Manning, C. D. 1999. *Foundations of statistical natural language processing*. MIT Press.

Paul P Martin, David Kranz, Peter Wulff, and Nicole Graulich. 2023. Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*.

Mihai Masala, Stefan Ruseti, Mihai Dascalu, and Ciprian Dobre. 2021. Extracting and clustering main ideas from student feedback using language models. In *Proceedings of AIED*, pages 282–292. Springer.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. 2019. Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 316–325.

Ross H Nehm and Hendrik Haertig. 2012. Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology*, 21:56–73.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Juan L Rastrollo-Guerrero, Juan A Gómez-Pulido, and Arturo Durán-Domínguez. 2020. Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, 10(3):1042.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of EMNLP*, pages 3982–3992.

Brian Riordan, Sarah Bichler, Allison Bradford, Jennifer King Chen, Korah Wiley, Libby Gerard, and Marcia C. Linn. 2020. An empirical investigation of neural methods for content scoring of science explanations. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 135–144.

Wesley C Salmon. 2006. *Four decades of scientific explanation*. University of Pittsburgh press.

Abigail Gurin Schleifer, Beata Beigman Klebanov, Moriah Ariely, and Giora Alexandron. 2023. Transformer-based Hebrew NLP models for short answer scoring in biology. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 550–555.

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 ACL*, pages 46–56.

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 ACL*, pages 1703–1714.

Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. 2019. Improving short answer grading using transformer-based pre-training. In *Proceedings of AIED*, pages 469–481.

Charissa Tansomboon, Libby F Gerard, Jonathan M Vitale, and Marcia C Linn. 2017. Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, 27:729–757.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge University Press.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of ICML*, pages 1073–1080.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10275–10288.

Peter Wulff, David Buschhüter, Andrea Westphal, Lukas Mientus, Anna Nowak, and Andreas Borowski. 2022. Bridging the gap between qualitative and quantitative assessment in science education research with machine learning—a case for pretrained language models-based clustering. *Journal of Science Education and Technology*, 31(4):490–513.

Chengwei Xiao, Jiaqi Ye, Rui Máximo Esteves, and Chunming Rong. 2016. Using spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14):3866–3878.

Appendix 1

Item	Text
Smoking item	The smoke from cigarettes contains several harmful substances, including the gas carbon-monoxide (CO). CO is released from cigarettes while smoking, and has a stronger tendency than oxygen to bind to Hemoglobin. Explain how high levels of CO make it difficult for smokers to exercise.
Anemia item	A person was found to have low levels of red blood cells in his blood test (anemia). This person complained to his doctor about weakness and difficulty to exercise. Explain how low levels of red blood cells make it difficult for people with anemia to exercise.

Table 11: The constructed response items (reproduced from Ariely et al. (2024); original responses are in Hebrew).

Cluster description	Exemplifying response
Full explanations: All/most of the conceptual components and the underlying causal relations are present.	"Red blood cells bind oxygen and transfer it in the bloodstream, from the lungs where it is absorbed, to all the cells of the body. A low amount of red blood cells in the body leads to the transfer of less oxygen to the body's cells. Since oxygen is one of the reactants in the process of cellular respiration - the energy production process, less oxygen reaching the cells leads to damage to this process. Thus, less available energy is produced in the body's cells and this impairs their function, which leads to fatigue and difficulty in performing physical activity." (Anemia item)
Gaps in causal connections: All/most of the conceptual components are present, but all/most of the causal relations are missing.	"The CO binds to the red blood cell instead of the oxygen and thus oxygen does not reach the cells of the body and then cellular respiration does not occur and the body cannot produce energy and thus it stops physical activity due to lack of energy." (Smoking item)
Specific sequential stages are missing and causal relations are often missing.	"CO gas is known to bind to Hemoglobin with a stronger tendency than oxygen. When CO binds to Hemoglobin, it takes the oxygen's place, so much less oxygen is transported from place to place and enters the cells. Lack of oxygen in the cells leads to less production of ATP molecules. Since energy is required for physical exercise, the result is that the person gets tired quickly and has difficulty exercising." (Smoking item)
Many sequential stages are missing and causal relations are often missing as well.	"Red blood cells carry the oxygen (because of Hemoglobin). When there is anemia, then there is a low amount of red blood cells and thus a low amount of oxygen reaches the muscles." (Anemia item)
No explanation: All/most of the sequential stages and the underlying causal relations are missing/irrelevant responses.	"I don't know" "Anemic people are tired because they have few red blood cells." (Anemia item)

Table 12: Examples of student responses and their classification into the KPs that were derived from the expert scoring according to the theory-driven rubric (reproduced from Ariely et al. (2024); original responses are in Hebrew).