Exploring Query-Driven Centroiding and Embedding Strategies for Automated Thematic Analysis in Physics Education

Michael Mingyar
Department of Physics, Montana State University, Bozeman, MT 59717

Tor Ole B. Odden
Center for Computing in Science Education, Department of Physics, University of Oslo, 0316 Oslo, Norway

Shannon Willoughby
Department of Physics, Montana State University, Bozeman, MT 59717

Understanding how Physics Education Research (PER) has changed and matured can help researchers and materially improve physics teaching and learning. The advent of large language models (LLMs) and embedding-based NLP techniques opens new avenues for this work by enabling the analysis of large corpora with minimal manual coding. This preliminary study builds off the work of [T. O. B. Odden *et al*., Physical Review Physics Education Research 20, 020151 (2024)]. We explore the use of text embeddings and retrieval-augmented generation (RAG)-like methods without generative AI to analyze themes across a random sample of 94 articles from major PER journals. Specifically, we examine how two methodological choices affect topic modeling performance: (1) representing articles using a single embedding versus multiple sentence-level embeddings, and (2) deriving topic centroids from representative texts versus researcher-defined queries. All results are evaluated against a human-coded dataset focused on four overarching thematic categories: teacher-centered, student-centered, physics content, and journal business. Our findings inform best practices for researchers seeking scalable, interpretable, and low-barrier approaches to literature analysis in PER.

2025 PERC Proceedings edited by Pawl, Zwolak, and Leak; Peer-reviewed, doi.org/10.1119/perc.2025.pr.Mingyar

I. INTRODUCTION

Physics education has undergone a significant transformation over the last century, with both content and instructional methods evolving alongside broader shifts in educational priorities [1][2]. Understanding these changes offers historical perspective and highlights rapid areas of growth in the field. Large-scale surveys are essential for identifying these trends in physics education and physics education research (PER), but reviewing all journal publications manually is impractical and vulnerable to selection bias [3]. To address this, machine learning techniques from natural language processing (NLP) enable systematic reviews of a journal's entire corpus.

A subfield of computer science and linguistics, NLP allows computers to extract patterns and meaning from unstructured text [4]. While widely known for applications like chatbots and AI systems, NLP has demonstrated effectiveness in analyzing trends in research literature. For example, Latent Dirichlet Allocation (LDA) has revealed thematic shifts in PER conference proceedings, such as a move from conceptual understanding to belonging and institutional change [5]. LDA-based analyses of *The Physics Teacher* and *Physics Education* identified 20 recurring topics which can be grouped into content, history, teaching, and learning [6], echoing earlier findings in broader science education [7].

Beyond literature reviews, NLP methods have been used to analyze student classroom performance. For instance, word-frequency algorithms have categorized open-ended Physics Measurement Questionnaire responses with accuracy comparable to that of human reviewers [8].

Recently, text embeddings–representations of words or texts as high-dimensional vectors in a pre-defined semantic space–have been used to reproduce LDA findings on PER literature [9]. Unlike inductive LDA, embeddings enable deductive, targeted analysis by measuring semantic similarity between texts using cosine or Euclidean distance. These vectors capture more nuanced meaning than word-frequency approaches.

Similar techniques have been applied outside PER for image classification [10]. Although created through different processes, embeddings for text and images, both, allow theme-based classification using vector averaging and distance metrics.

Other researchers have applied large language models (LLMs) to grade open-ended physics responses [11][12]. By crafting precise prompts through a process known as prompt engineering [13], they leveraged OpenAI's GPT-4 to automate grading with high levels of reliability.

Despite their utility, older NLP methods like LDA have limitations. LDA requires expertise to initialize and interpret, and its non-deterministic nature reduces reproducibility. Bag-of-words algorithms like LDA and those used in Wilson et al. [8] ignore word order and grammar, requiring time-intensive pre-processing such as lemmatization, part-of-speech tagging, and punctuation removal.

In contrast, text embeddings and LLM-based methods offer lower entry barriers. Because embeddings are generated by models trained on billions of contextual word relationships, they preserve nuanced meaning with minimal pre-processing.

Odden et al. [9] showed that embeddings alone (without the use of LLMs) can extract topics from PER texts with strong alignment to LDA-based themes. However, embeddings support many tasks beyond categorization. One such task, retrieval augmented generation (RAG) [14], involves splitting text into "chunks" (e.g., sentences or paragraphs), embedding each chunk, and retrieving relevant information based on a user-defined query. Though classically paired with generative models, the retrieval steps of RAG methods have the potential to enhance user-driven literature analysis without generative AI models.

Our work builds on Odden et al. [9] by exploring the benefits of RAG-type retrieval techniques for theme identification without generation. Our larger project aims to expand beyond PERC proceedings to include all articles from *Physical Review Physics Education Research* (PRPER), *American Journal of Physics* (AJP), and *The Physics Teacher* (TPT) from their inception to 2024. This preliminary study focuses on a random subset of 94 of these articles.

The main goal of this preliminary work is to investigate how two analytic choices affect embedding-based topic modeling. First, we compare categorization results when each article is represented by a single embedding versus multiple sentence-level embeddings. Second, we examine topic models (computer-representations of topics or themes) derived from researcher-defined queries versus representative texts. In both approaches, a topic model is defined by a single embedding vector (referred to as topic centroids, see Odden et al. [9] for more detail). We show that queries can be used to identify the representative texts, demonstrating their versatility.

We evaluate these methods against a human-coded dataset. Our goal is to inform best practices and lower barriers for PER researchers adopting NLP tools. As in previous work, we use an open-source model from JinaAI – jina-embeddings-v3 [15]–run locally to preserve data privacy and ensure replicability.

The research questions that we address in this proceeding are as follows:

1. **How does the choice of chunk length and method used to derive topic centroids affect accuracy of results?**
2. **How can researcher-defined queries be used as thematic groups for deductive analysis?**

II. METHODS

PDFs for papers from TPT and AJP articles are converted to text files using Adobe Acrobat Pro. PERC Proceedings are processed using the python library *PyMuPDF*, and PRPER articles were received from the publisher already in text format. All other computational efforts are performed on a local python environment.

The analysis methods used by Odden et al. [9] are the basis for the methods here. The concepts for the topic models used in that paper come from a previous work that uncovered them through LDA analysis [5]. They derived the embedding topic models by first collecting a sample of 15 representative texts per topic. The texts were chosen based on their high alignment with the topic as found through the previous LDA work. The papers were each condensed into a 512-dimensional embedding. This was done using a local HuggingFace repository [16] which allows the user to simply select the pre-trained model they would like to use, insert any string of text into the function, and it returns an embedding produced by the selected model. The model used in their work, JinaAi V2, created this semantic space through the study of various, general purpose sources. By averaging together the vectors for all 15 papers within each topic, the authors created topic centroids which represent the qualities of the vectors that created it, as demonstrated in Snell et al. [10]. Finding the distance between this vector and an individual paper's vector gleans insight into how much that paper belongs in the category the centroid represents.

The topics determined by these researchers were: "Teacher Focused," which encompasses content primarily concerned with effective teaching strategies; "Student Focused," which centers on student outcomes, belonging, and cognition; and "Content Focused," which includes work that engages directly with specific science content, typically with less emphasis on educational aspects. For a more detailed explanation of these categories and their sub-categories, see Odden et al. [7]. We argue that due to the similarity in the corpora, these groups can also be expected to emerge in ours.

Odden et al. [9] demonstrated the effectiveness of averaging representative papers' embeddings within each topic to assign thematic groups to other papers. While this method performs well, it relies heavily on the researcher's prior in-depth reading to identify the representative features, an inherently time intensive process. As an alternative approach to defining topic centroids, we begin by identifying four broad categories inspired by the prior work: "Teacher Focused", "Student Focused", "Physics Content", and "Journal Business". The former three categories are described above, and the "Journal Business“ category encompasses text that does not constitute original research or physics education content. Examples include awards, obituaries, book reviews, replies, and editorial comments. We write plain-text descriptions, or queries, of each category to represent the topics (shown in Table I). These descriptions are converted into normalized text embeddings and stored to serve as the topic models.

Label	Query
Teacher Focused	Teaching. Laboratory equipment. Teaching methods.
Student Focused	Student belonging. Student focused. Student agency.
Physics Content	Physics content. Physics material. Math. Derivations.
Journal Business	Editorials, book reviews, announcements, obituaries. Journal business. Reports on business.

TABLE I. Raw text that is embedded and used for the topic centroids in techniques (A) and (C). These queries are also used to find the representative papers in techniques (B) and (D).

Each paper in the subset is segmented into five-sentence chunks, and each chunk is converted into a normalized text embedding. This fine-grained approach is adopted to assess whether higher resolution in the text representation would improve topic assignment accuracy, as well as to potentially support downstream tasks such as displaying relevant text passages. Cosine similarity is computed between each sentence chunk's embedding and each topic model's embedding. Since all embeddings are normalized, this operation reduces to a dot product between the two vectors. Cosine similarity scores range from -1 to 1, where 1 indicates the two embeddings contain identical meaning, 0 denotes dissimilar content, and -1 represents completely opposite meaning. In practice, none of the comparisons in this study yield negative similarity values. We then take the average of these values and assign each paper one value per topic model to represent its average cosine similarity to the topic.

Using these cosine similarity scores, we calculate a normalized topic score for each paper following the method described in Equation 3 of Odden et al. [9]. Each paper is then assigned to the topic with the highest score. This process is repeated one more time with embeddings derived from each paper's whole text without chunking. In this simpler case, only one cosine similarity is calculated to represent the position of the whole paper in the embedding space. This is used as the control to see what affect, if any, sentence chunking has on the results. To validate the labeling process, two authors manually sort a sample of 94 papers into the four predefined categories. Cohen's Kappa for the sample rated by the two researchers is 0.78, considered moderate to strong agreement McHugh [17]. These researcher-defined labels serve as the "ground truth" for evaluating the performance of the computational labeling method.

The method described above of defining topic centroids effectively treats the search query itself as the centroid. This means that all the information that the computer has about the categories is limited to the text used in the query. While straight forward, this is a fairly simple approach. With this technique, the quality of results depends entirely on the clarity, detail, and style of the query used to describe it. However, unless the query is precisely worded and written in a particular, potentially unknown style, the resulting vector is often too general to capture the full breadth of relevant papers. This limitation is reflected in the relatively low agreement between the labels produced by this method and the ground truth (see Results and Discussion). To address this limitation– and to replicate the advantages of embedding representative papers, as demonstrated by Odden et al. [9]– we implement an alternative approach aimed at maintaining limited time investment while improving the model quality.

We begin by sorting all papers by their average topic scores as determined by the query embedding and selecting the top four papers for each topic. This number is chosen to balance representativeness and diminishing returns observed when including additional papers, but may be modified for future studies. For each of the four selected papers, we identify the ten sentence chunks with the highest topic scores, resulting in 40 representative embeddings per topic. These embeddings are then averaged to form a new embedding that serve as a "refined centroid" for that topic.

Each paper's average embedding is then compared via cosine similarity to these new topic embeddings. New labels are assigned based on the highest topic score from this comparison, and the results are again evaluated against the ground truth. This process is repeated for the case where vectors are derived from whole papers. In this case, as there are no sentence chunks, only the top four representative papers are used to derive the refined centroids for the topics, and the papers are again categorized in the manner described above.

In short, each paper is sorted into their respective topics based on which one they are closest to in the embeddings' vector space.

III. RESULTS

Topic centroids are defined by either using queries alone or by using the refined centroids described above. Additionally, the documents are either processed using sentence chunking or by using the whole papers without any segmenting. This results in four distinct approaches detailed in Table II.

Label	Description
(A)	Centroids derived from queries, text is chunked
(B)	Centroids refined with examples, text is chunked
(C)	Centroids derived from queries, text is not chunked
(D)	Centroids refined with examples, text is not chunked

TABLE II. Summary of centroid derivation methods and text processing strategies.

We report on the performance of this technique in Table III. Here, recall is used as the measure of accuracy for a classifier with multiple groups. Recall is calculated by dividing the number of correctly classified positive items by the total number of items actually in that group:

$$Recall = \frac{TP}{TP+FN}$$

Recall	(A)	(B)	(C)	(D)
Teacher Focused	$18 \pm 8\%$	$22 \pm 9\%$	$64 \pm 10\%$	$82 \pm 8\%$
Student Focused	$6.3 \pm 6\%$	$56 \pm 12\%$	$13 \pm 8\%$	$88 \pm 8\%$
Physics Content	$97 \pm 3\%$	$97 \pm 4\%$	$88 \pm 5\%$	$64 \pm 8\%$
Journal Business	$8.7 \pm 6\%$	$8.7 \pm 7\%$	$4.4 \pm 4\%$	$8.7 \pm 6\%$

TABLE III. Recall with standard error calculated per group for each technique.

Comparing technique (A) to technique (B) (both have chunked text, comparing query search alone versus refined centroid), there is a large increase in the classifier's ability to distinguish "Student Focused" content. Similarly, comparing technique (C) to (D) (both do not have chunked text, comparing query search alone versus refined centroid), there is a substantial improvement in the "Student Focused" category as well as a large, but smaller improvement in the "Teacher Focused" category. This suggests that the context of the "Student Focused" group was captured to a large degree as a result of refining the centroids. The same can be said of the "Teacher Focused" category, though to a lesser degree.

Technique (D), which used whole papers without chunking and refined centroids, outperforms the other techniques in "Teacher Focused" and "Student Focused" content, and performs equivalently in "Journal Business". (D) has a slightly lower recall in the "Physics Content" category compared to the other techniques. This performance gap is discussed below.

Because recall can be artificially high for a category if a classifier attempts to over-assign every item to that label, we show the proportion of how much each classifier assigns to each label in Table IV. Ideally, no individual group should be disproportionally represented, and they would all match the "actual" distribution and be accompanied by high recall. The "actual" distribution is the distribution of the sample papers being used as the ground truth.

Distribution	(A)	(B)	(C)	(D)	Actual
Teacher Focused	9%	14%	29%	46%	23%
Student Focused	2%	12%	2%	17%	17%
Physics Content	87%	72%	68%	35%	35%
Journal Business	2%	2%	1%	2%	24%

TABLE IV. Percent of the total sample that each technique assigned to each category. The "Actual" column represents the true distribution for each category from the sample.

Techniques (A), (B) and (C) disproportionally favor the group "Physics Content". The recall of technique (D) decreases for this category, but does not exhibit the same problem that the other techniques have of over assigning to this category. In that way, (D) shows an incremental improvement in classifying "Physics Content" compared to the other groups, though its recall is reduced as a result.

All groups severely under represent "Journal Business", and are only able to classify a few documents from this category correctly. Possible explanations for this shortcoming are found in the Discussions section.

IV. CONCLUSIONS

RQ 1 asks: "How does the choice of chunk length and method used to derive topic centroids affect accuracy of results?" Preliminary results suggest that using whole texts in combination with representative texts to define topic centroids is the best way –out of the options explored here– to categorize papers with cosine similarity based sorting. Chunking papers in the way explored here and averaging across chunks reduces the reliability of the categorization task and should be avoided. This suggests that longer chunks windows (e.g., the length of the whole paper) capture context better than smaller windows. However, our chunks were selected to have no overlapping content and only one size. Future work could explore if any benefits are associated with chunks of different, potentially varying length, or if there are any measurable improvements associated with adjacent chunks sharing portions of content.

RQ 2 asks: "How can researcher-defined queries be used as thematic groups for deductive analysis?" These preliminary results suggest that performing a categorizing task first with a search query that defines the category, then again with the averaged embeddings of the papers that have the highest similarity to those queries, returns the highest accuracy out of the techniques explored here. Though further work is needed to verify these results over a larger test set, this suggests that queries for researcher defined groups can be used for deductive thematic analysis of PER papers if done carefully.

This work also suggests that certain group's queries, such as the one provided for "Journal Business", might not accurately capture the meaning of the documents. This shows that query searching alone, even with a refined centroid, is not guaranteed to accurately retrieve groups. Therefore, researchers should be careful about how they define their queries, and should verify their results by checking that what the algorithm returns is in agreement with the intended search. Once these authors tweak and validate this technique on more queries and with a larger data set, this technique could be used to greatly expedite a literature review and reduce burden on the researcher.

V. DISCUSSION

These preliminary results show that an increase in performance is associated both with using whole papers (instead of sentence chunking) for embeddings as well as for using refined centroids. Using both techniques at once returns the highest agreement with human coders. This makes sense because the process of deriving embeddings and topic models in technique (D) is the most similar to previous work [9] [10]. Technique (D) also appears to be the only technique to mitigate over-assignment to a single category.

Techniques (A) and (B) tend to classify most documents as "Physics Content". A possible explanation for this is that any chunk of 5 sentences of physics education research papers, out of context, could be understandably described as physics content. It is only when the paper as a whole is taken into account that the context of many of these excerpts begins to make sense, and can be accurately categorized. Put simply, a given PER paper is "greater than a sum of its parts". In techniques (A) and (B), the text is broken down into chunks and averaged together. Not only does this result in a significantly different categorization compared to embedding the whole paper at once, but the category it chooses is telling. Without the context of the whole paper, breaking up PER text by equally-sized portions of consecutive sentences is not useful for embedding driven categorization. This has further implications for other qualitative techniques that focus on labeling or coding PER literature at the sentence level, as this suggests the aggregate meaning of a whole paper is distinct from the meaning of smaller segments.

Only techniques (C) and (D) look at papers in their entirety. It is possible that Technique (C) over-assigns to "Physics Content" because the query that we provide the model –which describes "physics content" and "physics material"– is predictably close to the embedding of most physics education research papers. If so, this query should be re-defined if these labels are going to be reused. Introducing refined centroids (technique (D)) helps with this issue, as the topic model's embedding is not just the embedding for a description of "physics content", but actually represents whole text and will be in the style of what would be expected for the category.

Technique (D) produces the most even-handed distribution of per-category recall values. Other techniques produce higher recalls for "Physics Content", but as discussed, this is a result of liberally assigning that label to most papers without much discrimination. Providing whole papers instead of sentence chunking, and using whole papers to represent topic centroids provides the best context to discern categories.

None of the techniques are able to categorize "Journal Business" to any usable degree. This potentially could be a result of the query chosen to represent the topic, but it is also possibly a result of an intrinsic issue with this category. While it is true that the journals are full of editorials, book reviews, obituaries, and meeting notes (see Table I), much of the content within these papers is heavily rooted in one of the other categories. Many book reviews are reviews of physics textbooks, which by nature will include a large amount of physics content. In a similar way, replies and comments could be replies or comments to a paper that easily rests in one of the other categories, meaning the content of the reply will be similar. Noting that the refined centroids do not improve recall in this category, this implies that the refined centroids failed to capture relevant texts. This could be rectified by manually selecting the representative papers, or potentially by finding a better query. As shown above, cosine similarity searches work best when one provides texts that look as similar as possible to what is intended to be found, so a query designed to mimic one of the sub-groups within this category might be useful.

The full study that follows this work will refine the query for or find representative papers manually for "Journal Business" to improve this group's recall. Additionally, the authors will choose representative papers for each thematic group to define centroids manually, and this will be compared to the results of refined centroid analysis. This will provide insight into the benefits and drawbacks associated with query-centric techniques like this compared to the techniques presented in [9].

We will extend the analysis to the full corpus of approximately 35,000 articles spanning the years 1933-2024 and examine the evolution of these groups over that period of time in the journals TPT, AJP, PRPER, and PERC Proceedings. This will not only reveal deep insights about how the contents of this field have evolved throughout this long period of time, but it will be the most comprehensive study of its kind to date that these authors are aware of.

However promising this query technique has shown to be, caution should be taken when using it in a research setting. Until the authors can compare the results of a fully automated query-driven categorization, as presented here, to the techniques of Odden et al. [9], its results are not sufficiently reliable. However, once integrated into a pipeline, running the analysis takes minutes or even seconds to finish, and can be a very useful supplementary tool in its current form.

VI. ACKNOWLEDGMENTS

Computational efforts were performed on the Tempest High Performance Computing System, operated and supported by University Information Technology Research Cyberinfrastructure (RRID:SCR_026229) at Montana State University.

[1] V. K. Otero and D. E. Meltzer, The Physics Teacher **54**, 523 (2016).
[2] D. E. Meltzer and V. K. Otero, American Journal of Physics **83**, 447 (2015).
[3] A. Nightingale, Surgery. **27** (2009-9), ISSN 02639319.
[4] A. Reshamwala, D. Mishra, and P. Pawar, IRACST Engineering Science and Technology: An International Journal (ESTIJ) **3**, 113 (2013).
[5] T. O. B. Odden, A. Marin, and M. D. Caballero, Physical Review Physics Education Research **16**, 010142 (2020).
[6] M. Caramaschi and T. O. B. Odden, arXiv preprint arXiv:2504.02703 (2025).
[7] T. O. B. Odden, A. Marin, and J. L. Rudolph, Science Education **105**, 653 (2021).
[8] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. Lewandowski, Physical Review Physics Education Research **18**, 010141 (2022).
[9] T. O. B. Odden, H. Tyseng, J. T. Mjaaland, M. F. Kreutzer, and A. Malthe-Sørenssen, Physical Review Physics Education Research **20**, 020151 (2024).
[10] J. Snell, K. Swersky, and R. Zemel, Advances in neural information processing systems **30** (2017).
[11] G. Kortemeyer, J. Nöhl, and D. Onishchuk, Physical Review Physics Education Research **20**, 020144 (2024).
[12] G. Kortemeyer and J. Nöhl, Physical Review Physics Education Research **21**, 010136 (2025).
[13] L. Giray, Annals of biomedical engineering **51**, 2629 (2023).
[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2020), NIPS '20, ISBN 9781713829546.
[15] S. Sturua, I. Mohr, M. K. Akram, M. Günther, B. Wang, M. Krimmel, F. Wang, G. Mastrapas, A. Koukounas, N. Wang, et al., arXiv preprint arXiv:2409.10173 (2024).
[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., arXiv preprint arXiv:1910.03771 (2019).
[17] M. L. McHugh, Biochemia medica **22**, 276 (2012).