

Testing the validity of embedding-based similarity and clustering for handwritten physics solutions

Maike Tauschhuber^{1,*}

*Department of Information Technology and Electrical Engineering, ETH Zurich,
8092 Zurich, Switzerland*

Gerd Kortemeyer^{1,†,‡}

Rectorate and AI Center, ETH Zurich, 8092 Zurich, Switzerland



(Received 30 June 2026; accepted 1 October 2026; published 21 October 2026)

Text embeddings are increasingly used in physics education research to organize, compare, and cluster large collections of written text. Their appeal is clear: once student responses have been mapped into a vector space, similarity comparisons and clustering become computationally inexpensive. However, in assessment contexts, the relevant question is not merely whether clusters can be produced but whether the geometry of the embedding space preserves grading-relevant distinctions. We tested this premise using 992 handwritten student-problem solutions from a high-stakes engineering thermodynamics exam, transcribed into five textual representations and embedded using nine embedding mechanisms. We compared embedding similarity and embedding-based hierarchical clusters against human-assigned scores. Across models, representations, and clustering choices, embedding similarity showed a consistent but modest relationship to score similarity, and the resulting clusters were score-enriched but not score-equivalent. Experiments with a synthetic dataset suggest that this may be due to embeddings behaving like novices when categorizing physics-problem solutions; that is, their similarity geometry is strongly influenced by surface features rather than conceptual, semantic structure. These findings suggest that rubric-blind transcription and out-of-the-box, state-of-the-art embeddings can support exploratory organization and human-in-the-loop review of physics solutions, but they do not provide an unsupervised basis for grading without external validation against the assessment construct of interest.

DOI: [10.1103/sdnj-cjst](https://doi.org/10.1103/sdnj-cjst)

I. INTRODUCTION

A. Assessment in physics

Problem solving is central to learning physics across formative settings (homework, in-class practice, and low-stakes quizzes) and summative assessments (midterms and finals) [1–8]. Meaningful assessment must capture characteristic features of physics reasoning: mathematical modeling that blends conceptual insight with equations, computation, and theoretical principles [9–11]; transfer to novel scenarios [12]; multiple legitimate solution paths [13–15]; and coordination of diagrams, equations, graphs, and prose [16–19]. Focusing on a final numeric or

symbolic answer alone is therefore oftentimes inadequate to get the full picture; instructors must examine the solution path to understand and support students' reasoning [14,15]. Yet providing such path-focused feedback at scale is difficult in large introductory courses, given the expertise, time, and logistical demands involved [20]. Historically, expert judgment has been essential to evaluate not just results but also the logical, conceptual, and mathematical competencies revealed in student work [21–25]. While some aspects of solutions can be evaluated by computers, diverse approaches and error patterns typically require human interpretation [15,26–34].

Handwritten solution paths remain the gold standard in timed high-stakes exams (as typesetting mathematics would add irrelevant overhead), and for a small number of such assessments, the grading effort might be manageable for small-enrollment courses. Outside formal exams, frequent formative evaluation can substantially enhance learning through timely, targeted feedback, but it is often underused because evaluation is labor-intensive [6,35]. While in principle, typesetting mathematics in a computer-readable format would be possible in asynchronous take-home assessments, questions of academic integrity in

*Contact author: mtauschhuber@neuroo.de

†Contact author: kgerd@ethz.ch

‡Also at Michigan State University, East Lansing, Michigan 48823, USA.

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

remote and hybrid contexts add further constraints [36]. Because students' priorities are shaped by what is assessed and how credit is assigned, scalable approaches that illuminate reasoning while remaining trustworthy and feasible are a pressing need [37].

B. Automated grading and human oversight

To address workload, physics education has a long tradition of automated systems that assist instructors [27,38–42]. Most platforms, however, constrain responses to multiple-choice, ranking, numeric entries, or closed-form algebraic answers. Although such formats can be psychometrically sound [43–45], they have been criticized for neglecting communication and presentation [46,47] and for failing to capture the breadth of students' reasoning [20]; they can also incentivize guessing and other unproductive behaviors, especially with multiple attempts [48–50].

Even before large-scale generative models, researchers broadened the response space with short answers and explanations supported by artificial intelligence (AI) techniques [51,52]. The Andes tutoring system, for example, permits substantial freedom in constructing solutions and interprets diverse traces via logical solution graphs, an algebra subsystem, and structured feedback rather than language modeling [51,53]. Classifier-based approaches to scoring short physics responses have also shown practically useful agreement with human raters in interactive tutoring contexts [52].

Recently, large language models (LLMs), particularly generative pretrained transformers [54–56], have demonstrated strong capabilities on academic tasks [57] and across educational uses [58]. In physics education, work spans classroom integration [59–61], research support [62–64], and problem solving [65–69], including performance above typical postinstruction averages on several concept inventories [70]. LLMs can assist with generating new problems [71] and show promise for grading free-form student work [72–76]. A human-in-the-loop stance remains prudent: confidence filters can route only trustworthy AI decisions forward and defer uncertain cases to human graders [77]. This strategy reduces load while maintaining reliability traditionally associated with deterministic, closed-form systems.

Deep learning methods aim to learn such structure implicitly, but limitations remain. Fraknói *et al.* [78] report that embeddings of variable-free arithmetic expressions can be dominated by surface properties rather than semantic relations, suggesting that core mathematical principles may not be reliably learned. A broader analysis argues that multihead attention, positional encodings, and feed-forward layers play distinct roles across representational levels, with lower layers emphasizing patterns and upper layers semantics. Xin He [79] emphasizes hierarchical structure by combining semantic and structural

features, using hesitant fuzzy sets to measure formula similarity [80] and a language model for representation learning over literature graphs. Beyond standalone formulas, hierarchical encoders and decoders have been applied to math word problems, enriching token representations via dependency structure and generating solutions with tree-based decoders [81].

C. Text embeddings

A direct way to obtain similarities among student solutions would be to ask a large language model to compare pairs of responses. This may be attractive in principle, since such a model could compare solution strategies more flexibly than a fixed vector representation. The challenge is scaling. In grading, each solution can be evaluated once, so the number of expensive model calls scales linearly with the number of submitted solutions, $O(n)$. Similarity analysis is inherently pairwise: for n solutions to a problem, there are $\frac{n(n-1)}{2}$ unordered pairs. A direct LLM-based similarity analysis would therefore require $O(n^2)$ expensive model calls. For response sets of a few hundred students, this already amounts to tens of thousands of pairwise comparisons per problem, and the cost grows rapidly with additional problems, transcription conditions, or model variants. Embedding-based methods offer a computationally attractive alternative: each solution is processed once to produce a vector representation, after which all pairwise similarities can be computed cheaply using local vector operations such as dot products or cosine similarity. This computational argument motivates the use of embeddings as a scalable approximation to pairwise semantic comparison. It does not, however, establish that the resulting similarity geometry is valid for grading-related interpretation; that validity is the empirical question tested in this study.

Text embeddings have recently become attractive tools for physics education research because they provide a scalable way to compare large collections of written responses, explanations, interviews, or abstracts [82]. Embedding models are trained on large collections of text to place pieces of text that are used in similar contexts near one another in a high-dimensional vector space; once a text has been mapped into this vector space, distances between vectors can be used for retrieval, visualization, and clustering. This makes embeddings appealing in settings where manual qualitative comparison is expensive and where researchers seek to identify recurring patterns in large corpora [83,84], characterize curricula [85], or code student responses [86,87].

Prior work has demonstrated several valuable uses of clustering in science education. Martin *et al.* [88] used unsupervised clustering to discover patterns in students' written organic-chemistry arguments and subsequently combined those patterns with a theory-driven framework to construct a rubric and train automated scoring models.

Gurin Schleifer *et al.* [89] compared clusters derived from pretrained language-model embeddings with expert-defined knowledge profiles in biology, finding that high-quality responses were more readily recovered than distinct forms of incorrect reasoning. The present study addresses a complementary validity question: whether out-of-the-box embeddings of rubric-blind textual representations of handwritten, equation- and diagram-rich physics solutions preserve human-assigned correctness strongly enough to justify grading-related interpretation. We use “rubric-blind” to mean that the transcription model was not supplied with the official solution, any question-specific mark scheme or scoring rubric, item-level point allocations, or the human-assigned score. Unlike Martin *et al.* [88], we do not use clusters to derive a rubric or train a classifier. Instead, we treat pairwise similarity and cluster membership as hypotheses and validate them externally against human scores across five textual representations and nine embedding mechanisms.

However, assessment contexts impose a stronger requirement than exploratory organization. A cluster of responses may be coherent in embedding space without being coherent with respect to the construct being assessed. In grading, the relevant construct is not general semantic similarity but rubric-relevant correctness: the selection of appropriate principles, the coordination of representations, the validity of assumptions, the structure of derivations, and the treatment of numerical and symbolic details. Two solutions may be textually similar while differing in a crucial sign, state variable, unit, or physical assumption; conversely, two solutions may look quite different while earning similar scores because they instantiate different valid solution paths.

We therefore treat embedding-based clustering as a hypothesis to be tested rather than as a grading method. Following validity arguments in educational measurement, the question is not whether a computational representation can produce clusters but whether the intended interpretation and use of those clusters are supported by evidence [90,91]. In the present study, we evaluate whether embedding similarity and embedding-based hierarchical clusters preserve human-assigned-score similarity for handwritten physics solutions.

D. Testing embedding-based similarity and clustering

The central premise tested in this study is that proximity in embedding space may correspond to similarity in human-assigned scores—ideally, clusters would look like those depicted in Fig. 1. We evaluated this premise in two complementary ways. First, we asked whether pairwise embedding similarity predicts pairwise score similarity: for each pair of student solutions to the same problem, we compared the embedding similarity of the two responses with the negative absolute difference between their human-assigned scores. Second, we asked whether hierarchical

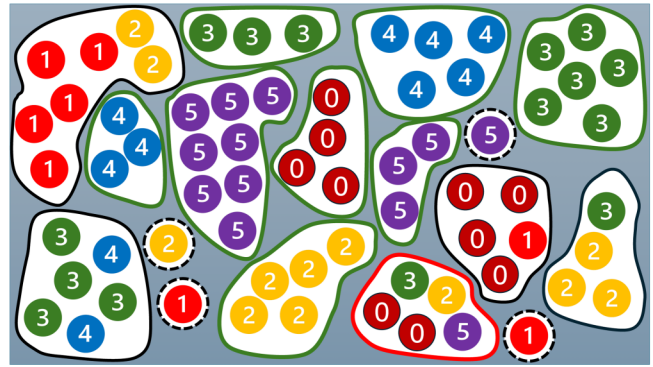


FIG. 1. Schematic illustration of the clustering-validity criterion. Each dot represents a student solution to a problem or problem part; numbers and colors represent human-assigned reference scores. Clusters are formed without using score information. If embedding clusters supported grading-related interpretation, sizable clusters would contain responses with similar scores. Mixed-score clusters may still be useful for exploratory organization, whereas clusters with large score variation indicate that embedding proximity does not preserve the grading-relevant construct. Single-member clusters may represent unusual solution approaches or failed grouping.

clusters formed from embedding similarities produced score-homogeneous groups. In both analyses, human scores were used only as an external reference standard for evaluation and were not used during transcription, embedding generation, similarity computation, or clustering.

This framing treats clustering as an object of validation rather than as an assumed grading method. Embedding-based clusters may be useful for organizing responses, selecting examples, or supporting human review, but a grading-related interpretation requires evidence that cluster membership preserves rubric-relevant distinctions. Throughout this study, the embedding-based approach examined here refers to unsupervised similarity analysis and clustering of pretrained representations generated from nonevaluative, rubric-blind transcriptions; supervised classifiers, rubric-aware prompting, task-specific fine-tuning, and direct LLM grading are outside the scope of the present test. Rubric-aware or mark-scheme-aware LLM grading has shown promise [92,93], but it addresses a different question: whether a model can apply an assessment criterion, rather than whether unsupervised embedding geometry preserves that criterion without explicit guidance.

E. Research questions

The main hypothesis of this study is

If embeddings are meaningful semantic-conceptual representations of student-problem solutions, then proximity in embedding space should correspond to similarity in human-assigned correctness scores.

This study addresses four research questions.

1. To what extent does pairwise similarity in embedding space predict similarity in human-assigned scores for handwritten physics solutions?
2. Do hierarchical clusters formed from embedding similarities produce groups that are homogeneous with respect to human-assigned scores?
3. How do these results depend on the textual representation of the handwritten work, the embedding mechanism, the similarity metric, and the clustering method?
4. Are the observed relationships strong enough to justify grading-related interpretations of embedding clusters, or are the clusters better understood as exploratory groupings requiring human validation?

II. SETTING

A. Institution

ETH Zurich is a technical university enrolling approximately 25 000 students from more than 120 countries. Admission is highly selective for international applicants, while Swiss high school graduates have open access. Most undergraduate instruction is in German. In line with the academic tradition of German-speaking universities, assessment is primarily summative, with high-stakes examinations at the end of courses rather than continuous assessment during the semester.

B. Exam

We analyze a high-stakes exam in engineering thermodynamics [75,77] covering standard topics such as energy, exergy, entropy, and enthalpy. The exam had four problems (1 to 4), each with a handful of parts (a to d or a to e), and those, in turn, were graded on a handful of rubric items each (85 rubric items in total). Students produced complete handwritten solutions using permanent pens and were instructed to leave out any work they did not wish to be graded. Figure 2 shows an example of student work on this exam.

C. Sample

The study was approved by the ETH Zurich Ethics Committee (Study 2023-N-286). Of the 434 students who sat the exam, 252 provided informed consent and were included in the research dataset. Consent forms were distributed before the exam and collected with the submissions. Because participation was opt-in in a high-stakes context, selection effects are possible; for example, students may have ignored the form under time pressure or withheld consent after perceiving poor performance. The research protocol did not permit linking participation status to demographics or scores, so direct comparisons between participants and nonparticipants are not available. The resulting dataset comprises 3041 pages of handwritten

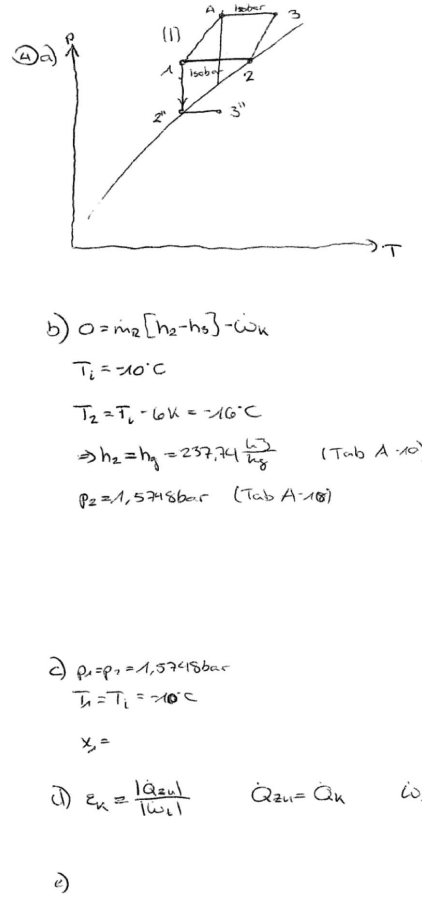


FIG. 2. Example of student work on the exam. The black box in the top-right is the redacted student name.

solutions (mean 12 pages per student) spanning a wide range of performance; as not all students worked on all problems, the dataset resulted in 993 student-problem solutions (instead of 1008).

III. METHODS

A. Transcription of handwriting

The scanned handwritten solutions were converted into text under five transcription conditions using GPT-5.4 [94]. These conditions varied the degree of normalization and interpretation while sharing the nonevaluative, rubric-blind protocol described in Sec. III B. The five transcription conditions were as follows.

ASCII verbatim (ascii): In this condition, the model was asked to transcribe the visible content as closely as possible into plain ASCII text. Mathematical notation was represented using ASCII conventions, such as x^2 , $\sqrt{x}$, ΔT , $\dot{Q}$, $\dot{m}$, h_1 , and p_2 . Diagrams and sketches were not ignored but were represented only by brief textual descriptions placed near their approximate location in the page content.

The rationale for this condition was to produce a minimally normalized representation of the original student artifact. It preserves idiosyncratic notation, ordering, partial work, and visible structure as much as possible. This condition therefore serves as a relatively low-interpretation baseline. However, because it retains substantial surface variation in notation, layout, and handwriting-derived transcription noise, it may be less effective for clustering solutions that are mathematically or conceptually similar but written in different forms.

LaTeX-oriented (latex): In this condition, the model was asked to transcribe visible mathematical expressions using LaTeX notation while preserving surrounding prose in ordinary text. For example, rates, subscripts, and thermodynamic quantities could be rendered as $\dot{Q}$, $\dot{m}$, ΔT , h , or c_p . The instruction emphasized that equations should be preserved as written, even if they appeared nonstandard or incomplete, and that the model should not simplify, repair, or complete derivations. Diagrams were described textually, including axes, labels, states, arrows, and process paths where visible.

The rationale for this condition was to obtain a more standardized representation of mathematical work than the ASCII condition while retaining the symbolic structure of the student's solution. This representation may be useful for calculation-heavy or derivation-heavy problems, where similarity may depend on the equations selected, the variables related, and the substitutions performed. At the same time, LaTeX syntax may introduce tokens that are not semantically meaningful to all embedding models, and the act of converting handwriting into LaTeX may normalize away some student-specific variation. Thus, this condition tests whether mathematical standardization improves clustering relative to a more literal transcription.

Formula narrative (formula_narrative): In this condition, the model was asked to narrate the formulas, symbolic steps, numerical substitutions, and final numerical results in natural language. For example, an expression such as $Q = m c_p \Delta T$ would be represented as a statement that the student uses heat transfer as mass times specific heat at constant pressure times temperature difference, while retaining the original formula and visible quantities where useful. The narrative was constrained to describe visible work only; uncertain inferences were to be marked with phrases such as "the student appears to" or "it is unclear whether."

The rationale for this condition was to make mathematical content more accessible to general-purpose text embeddings. Students may write mathematically equivalent or closely related expressions

using different notation, ordering, or intermediate variables. A formula narrative may reduce superficial notational differences by converting symbolic work into semantically similar natural-language descriptions. This condition is therefore expected to be useful when the important similarity among solutions lies in the underlying formula choices, substitutions, or computational strategy rather than in the exact notation used. Its potential disadvantage is that narration introduces a greater degree of model mediation and may lose fine-grained algebraic or numerical detail.

Solution narrative (solution_narrative): In this condition, the model was asked to produce a concise narrative account of what the student visibly did in the solution. The output described the apparent problem labels, physical system or process when visible, quantities identified, formulas used, sequence of substitutions or calculations, final results, and any diagrams or crossed-out work. The model was instructed to use cautious language for inferred reasoning and to avoid evaluative terms such as "correct," "incorrect," "right," or "wrong" unless such words were visibly written by the student.

The rationale for this condition was to represent the student's apparent solution process rather than only the literal written content or isolated formulas. This may be especially useful for conceptual problems, problems involving multistep reasoning, and problems where a diagram or choice of method is central to the response. For clustering, the hypothesis is that students who follow similar reasoning paths or make similar methodological choices may be embedded near one another even when their exact notation or numerical work differs. The risk is that this condition is the most interpretive of the five: the model may impose a coherent narrative on incomplete or ambiguous work. For this reason, the prompt emphasized visible evidence and explicit uncertainty.

Hybrid structured (hybrid_structured): In this condition, the model was asked to produce a structured representation with fixed section headings: problem labels, visible text, formulas and symbolic work, numerical work and results, diagrams and sketches, apparent method, crossed-out or abandoned work, and uncertainties. Mathematical expressions were represented in ASCII notation. For each visible diagram or sketch, the model was asked to provide a structured description, including the diagram type, axes, states or points, paths or processes, annotations, apparent role in the solution, and uncertainty. As in the other conditions, the model was instructed not to evaluate correctness.

The rationale for this condition was to combine the strengths of the previous conditions. It preserves

TABLE I. Summary of the five textual representations used for the handwritten student work.

Representation	Transformation	Mathematical form	Diagram treatment	Degree of interpretation
ASCII verbatim	Near-literal transcription	ASCII notation	Brief textual description	Low
LaTeX-oriented	Symbolic normalization	LaTeX notation	Structured textual description	Low-moderate
Formula narrative	Narrates formulas and calculations	Natural language plus retained expressions	Narrated where relevant	Moderate
Solution narrative	Describes visible solution process	Selected formulas within narrative	Integrated into process account	High
Hybrid structured	Fixed multimodal schema	ASCII expressions in sections	Detailed structured description	High

literal written content, symbolic and numerical work, diagrammatic information, and a brief nonevaluative description of the apparent method while imposing a consistent structure across solutions. This condition was designed specifically as an embedding-oriented representation: stable section headings provide a common schema, while the contents preserve solution-specific information likely to be relevant for clustering. It may be particularly valuable for thermodynamics exams, where grading-relevant information can appear in prose, equations, numerical substitutions, property-table lookups, cycle sketches, control-volume diagrams, or thermodynamic state diagrams. The hybrid condition therefore tests whether a structured, multimodal textual representation yields clusters that better align with human-assigned scores than either literal transcription or free narrative alone.

Table I gives an overview of these five transcription modes, while the Appendix shows the original system prompts and examples of transcriptions.

B. Common constraints across transcription conditions

We deliberately separated representation from evaluation. A rubric-aware or mark-scheme-aware transcription prompt would introduce assessment information before the embedding stage, so subsequent agreement with human scores would conflate the model's grading inferences with the geometry of the embedding space. The model therefore served as a rubric-blind representation layer rather than as a grader: it could normalize visible notation and describe apparent solution structure, but human scores remained an external criterion used only after transcription, embedding generation, similarity computation, and clustering. This design isolates the narrower question of whether out-of-the-box embedding geometry itself preserves grading-relevant distinctions.

Operationally, all five transcription conditions shared the following constraints:

1. The model operated only on the visible student artifact; it was not supplied with the official solution, any question-specific mark scheme or scoring

rubric, item-level point allocations, or the human-assigned score.

2. The task was framed as transcription or description, not assessment. The model was explicitly instructed not to solve, correct, grade, or compare the student work to an ideal solution.
3. Ambiguity was to be preserved rather than resolved: illegible material was marked as such, uncertain symbols or numbers were identified as uncertain, and inferred reasoning was to be described cautiously.
4. Diagrams and sketches were included in the textual representation rather than discarded. This was important because thermodynamics solutions may contain grading-relevant graphical elements, such as T - s diagrams, p - v diagrams, cycle sketches, process paths, state labels, control-volume schematics, and heat or work arrows.
5. Crossed-out or abandoned work was retained when readable, since such work may reflect solution strategies or misconceptions relevant to similarity clustering.

The purpose of comparing these five transcription conditions was not to determine which one most faithfully reproduces the handwriting in a human sense, but rather to determine which textual representation best preserves grading-relevant similarity among student solutions. The subsequent clustering analysis therefore treats the transcription condition as an experimental variable. For each condition, embeddings are computed from the resulting text, clusters are formed within each problem, and the distribution of human-assigned scores within each cluster is examined. A useful transcription condition is one for which clusters contain solutions that are similar not only in embedding space but also in the scores assigned by human graders and, upon inspection, in their apparent solution strategies or recurring errors.

C. Text-embedding methods

We generated vector representations of each student response using nine embedding mechanisms. In the

following, the short names in parentheses are used throughout the remainder of the paper.

OpenAI text-embedding-3-large (oai3large): This condition used OpenAI's text-embedding-3-large embedding model as a hosted, general-purpose text-embedding baseline. The model maps each input response representation to a dense vector intended to support semantic comparison of text by vector similarity. In this study, oai3large serves as a strong proprietary reference embedding that was not specialized to physics education or to the particular scoring rubrics used here [95].

Jina text-retrieval F16 (jina4retF16): This condition used the Jina Embeddings v4 model with the text-retrieval adapter in F16 GGUF format. Jina Embeddings v4 is a multimodal, multilingual embedding model with task-specific adapters. The text-retrieval adapter is optimized for retrieval-style use cases, where vector similarity is intended to support finding relevant passages or documents. We used it here as an embedding mechanism for student responses, allowing us to test whether retrieval-optimized semantic structure is aligned with grading-relevant similarity [96,97].

Jina text-retrieval Q8 (jina4retQ8): The same adapter as jina4retF16, but in an 8-bit quantized GGUF representation. This condition was included to test whether reduced-precision local inference preserved the grading-relevant geometry observed with the higher-precision retrieval-adapter condition.

Jina text-retrieval Q4_K_M (jina4retQ4): The same adapter as jina4retF16, but in a more aggressively quantized Q4_K_M GGUF representation. It was included as a low-resource local embedding condition, testing whether a substantially compressed retrieval-oriented model still preserved enough semantic structure to be useful for clustering or score-similarity analysis.

Jina text-matching F16 (jina4matchF16): This condition used Jina Embeddings v4 with the text-matching adapter in F16 GGUF format. Unlike the retrieval adapter, the text-matching adapter is intended for symmetric semantic-similarity tasks such as duplicate detection, clustering, and matching texts of the same general type. This made it a particularly relevant comparison condition for the present study, since student responses are being compared to other student responses rather than to external query documents [96,97].

Jina text-matching Q8 (jina4matchQ8): This condition tests whether the symmetric text-matching geometry of jina4matchF16 is retained under moderate 8-bit quantization.

Jina text-matching Q4_K_M (jina4matchQ4): This condition represents the most compressed version of

the Jina text-matching family used in the study and was included to assess the robustness of embedding-based clustering and score-similarity analyses under low-resource local deployment.

Qwen 2.5-Math-1.5B-Instruct (qwenMath1p5B): This condition used the instruction-tuned 1.5B-parameter Qwen2.5-Math model as a math-specialized representation mechanism. Unlike the dedicated embedding models above, Qwen2.5-Math-Instruct is primarily a generative mathematical reasoning model rather than a conventional embedding model. We thus extracted final-layer hidden-state representations and converted them to one vector per response by mean pooling over nonpadding tokens. The resulting vectors were not normalized before similarity computation. We included it to test whether representations derived from a smaller mathematics-specialized instruction model better preserved grading-relevant distinctions in physics solutions than general-purpose text embeddings [98].

Qwen 2.5-Math-7B-Instruct (qwenMath7B): This condition used the instruction-tuned 7B-parameter Qwen2.5-Math model as a larger math-specialized representation mechanism. As with qwenMath1p5B, this model was not included as a standard sentence-embedding model but as a test of whether a mathematics-specialized language model produced response representations whose geometry was more aligned with rubric-relevant correctness. This condition also allowed us to examine whether increased model scale within the Qwen2.5-Math family improved the usefulness of embeddinglike representations for clustering and grading-related similarity analyses [98].

For each of the nine embedding mechanisms, we generated embeddings for five response representations: `ascii`, `latex`, `formula_narrative`, `solution_narrative`, and `hybrid_structured`. The four problems contained 245, 247, 252, and 249 student responses, respectively, for a total of 993 student-problem responses per response representation. Thus, the full design called for $9 \times 5 \times 993 = 44\,685$ embedding vectors. During downstream analysis, after the graphics processing units used for local embedding generation were no longer available, we discovered one corrupted embedding file, corresponding to a student response for Problem 4 in one embedding condition. Because this vector could not be regenerated at that stage, we excluded the corresponding student-problem case globally rather than allowing different embedding mechanisms to be evaluated on different response sets. Specifically, the student-problem was removed from consideration for all embedding mechanisms and response representations. The final harmonized analysis set therefore contained 245, 247, 252, and 248 responses for Problems 1–4, respectively, or 992 student-problem

responses per response representation. This yielded $9 \times 5 \times 992 = 44,640$ usable embedding vectors in the final comparative analyses.

D. Similarity and clustering evaluation

For each problem, transcription condition, and embedding mechanism, we constructed a matrix of pairwise similarities among student-problem responses. We evaluated these matrices in two complementary ways:

- First, we asked whether pairwise embedding similarity predicted pairwise score similarity.
- Second, we asked whether hierarchical clusters formed from embedding similarities were homogeneous with respect to human-assigned scores.

In both analyses, human scores were used only as an external reference standard for evaluation and were not used during transcription, embedding generation, similarity computation, or clustering.

1. Similarity

Because each student response contributes to many unordered pairs, the pairwise observations are not statistically independent. We therefore interpret Pearson r and Spearman ρ as descriptive effect-size measures of alignment between embedding similarity and score similarity. Conventional pair-level p -values and confidence intervals are not reported because they would incorrectly assume independent observations.

Let $x_i \in \mathbb{R}^d$ denote the embedding vector for student response i within a fixed problem, transcription condition, and embedding mechanism. We computed two similarity matrices. The first was ordinary cosine similarity,

$$s_{ij}^{\cos} = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|}. \quad (1)$$

The second was mean-centered cosine similarity. For each embedding matrix (fixed problem, transcription condition, and embedding mechanism), we computed the mean vector:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i, \quad (2)$$

and then compared centered vectors $z_i = x_i - \mu$:

$$s_{ij}^{\text{cent}} = \frac{z_i \cdot z_j}{\|z_i\| \|z_j\|}. \quad (3)$$

The centered measure was included because all responses to a given problem share substantial semantic context from the problem statement and the domain. Mean-centering removes the common direction of the response set and may therefore emphasize distinctions among student solutions rather than similarity arising

merely from answering the same problem. Similarity matrices were checked for finite entries, symmetry, diagonal values close to one, and values within the numerical range expected for cosinelike similarities.

For the pairwise score-similarity analysis, we considered all unordered pairs of student responses (i, j) within the same problem for which both responses had human-assigned scores. For each pair, we recorded the embedding similarity s_{ij} , the human scores y_i and y_j , the absolute score difference

$$\Delta_{ij} = |y_i - y_j|, \quad (4)$$

and the negative absolute score difference $-\Delta_{ij}$. If embedding similarity preserved grading-relevant distinctions, higher values of s_{ij} should correspond to smaller score differences and, therefore, to larger values of $-\Delta_{ij}$. We therefore computed Pearson and Spearman correlations between s_{ij} and $-\Delta_{ij}$. Spearman correlation, rather than Pearson correlation, was treated as the primary pairwise quality measure because it tests whether the embedding similarity orders pairs in a grading-relevant way without assuming a linear relationship between similarity values and score differences. Pearson correlation was retained as a secondary measure. We also binned pairs by similarity quantile and computed the mean and median absolute score difference within each bin as a diagnostic visualization of whether high-similarity pairs had smaller score differences.

2. Clustering

For the clustering analysis, each similarity matrix S was converted to a distance matrix:

$$D_{ij} = 1 - S_{ij}. \quad (5)$$

We then performed agglomerative hierarchical clustering using average linkage and complete linkage. These two linkage methods represent different assumptions about what should count as a coherent group: average linkage merges clusters based on average intercluster distance, whereas complete linkage is more conservative and penalizes clusters whose most distant members remain far apart. For each linkage method, we evaluated clusterings with requested numbers of clusters $k = 2, \dots, 80$, capped at $n - 1$ for a response set of size n . Cluster assignments were produced without using score information.

For each clustering, we summarized the distribution of human scores within each cluster. For a cluster c , with score set $y_i: i \in c$, we computed the cluster size n_c , mean score $\bar{y}_c$, sample score standard deviation (SD) σ_c , and score range

$$r_c = \max_{i \in c} y_i - \min_{i \in c} y_i. \quad (6)$$

Because very small clusters are of limited practical value for grading or batching workflows, we defined a useful cluster as one containing at least five responses. We therefore computed useful-cluster coverage, defined as the fraction of responses assigned to useful clusters. For the responses in useful clusters, we computed the response-count-weighted mean cluster score standard deviation,

$$\bar{\sigma}_{\text{useful}} = \frac{\sum_{c: n_c \geq 5} n_c \sigma_c}{\sum_{c: n_c \geq 5} n_c}, \quad (7)$$

and the corresponding response-count-weighted mean score range,

$$\bar{r}_{\text{useful}} = \frac{\sum_{c: n_c \geq 5} n_c r_c}{\sum_{c: n_c \geq 5} n_c}. \quad (8)$$

We refer to these quantities as Useful SD and Useful range in Tables IV and V. Lower values indicate more score-homogeneous clusters, while higher useful-cluster coverage indicates that the method organizes a larger fraction of the response set into nontrivial groups. The upper limit of cluster numbers $k = 80$ and the minimum useful-cluster size of five responses were chosen to allow relatively fine-grained grouping while excluding singleton or very small clusters that would not meaningfully reduce grading or review effort.

For aggregate comparisons across problems, a clustering configuration was eligible for the primary comparison only if at least 70% of responses were assigned to useful clusters and at least two useful clusters were present. Within each problem and experimental condition (transcription and embedding method), we then selected the value of k with the lowest Useful SD. Ties were broken first by higher useful-cluster coverage and then, if both quantities remained tied, by choosing the smaller value of k , thereby favoring the less fragmented partition. These best-per-problem rows were aggregated across the four problems by transcription condition, embedding mechanism, similarity method, and linkage method. This produced the overall clustering-quality summaries reported in Sec. IV. The selection of k was used only for retrospective evaluation of the best observed clustering quality for each condition. It should therefore be interpreted as an optimistic diagnostic of the embedding geometry, not as a deployable unsupervised procedure for choosing k without access to the reference scores.

Together, the pairwise and clustering analyses test different aspects of the same validity claim. The pairwise analysis asks whether embedding geometry orders pairs of responses in a way that reflects human-score similarity.

The clustering analysis asks whether that geometry can be partitioned into sizable groups with sufficiently homogeneous scores to support grading-related interpretation. In both cases, strong performance would require not merely visible structure in embedding space but also alignment between that structure and the assessment construct represented by human-assigned scores.

IV. RESULTS

A. Embedding versus score similarity

Figure 3 provides a first aggregate test of the embedding-similarity hypothesis. The plotted quantity is based on the absolute score difference Δ_{ij} defined in Eq. (4), normalized within each problem and condition. If proximity in embedding space were a strong proxy for grading similarity, the mean score difference should decrease sharply for pairs in the highest similarity percentiles. Instead, all transcription conditions show a consistent but modest downward trend. Pairs in the lowest similarity percentiles tend to differ in score more than arbitrary pairs, whereas pairs in the highest similarity percentiles tend to differ somewhat less. However, even the most similar pairs retain substantial score variation: their mean score difference remains close to the random-pair baseline rather than approaching zero. This indicates that embedding spaces contain grading-relevant information, but that high embedding similarity should not be interpreted as evidence of grading equivalence.

Tables II and III summarize the correlations between embedding similarity s_{ij} , computed using either Eqs. (1) or (3), and the negative score difference $-\Delta_{ij}$ from Eq. (4). Across transcription modes, the best Spearman correlations ranged from 0.142 for the literal ASCII transcription to 0.198 for the hybrid structured representation. Narrative and structured representations therefore produced stronger alignment with human scores than literal transcription, suggesting that some normalization of symbolic and diagrammatic work into semantically explicit text helped the embedding models capture grading-relevant structure. However, the absolute magnitude of the correlations remained small. Even the best condition, hybrid structured transcription with the qwenMath7B representation and ordinary cosine similarity, reached only a weighted Spearman correlation of 0.198 and a weighted Pearson correlation of 0.243. Thus, the improvement from transcription design did not turn embedding similarity into a reliable proxy for score equivalence.

The fine-grained ordering of embedding mechanisms should be interpreted cautiously. The 10 highest-ranked conditions occupy a narrow band of weighted Spearman correlations, from 0.177 to 0.198. We therefore do not interpret the exact ranks as evidence that one embedding mechanism is decisively superior to another. Rather, the

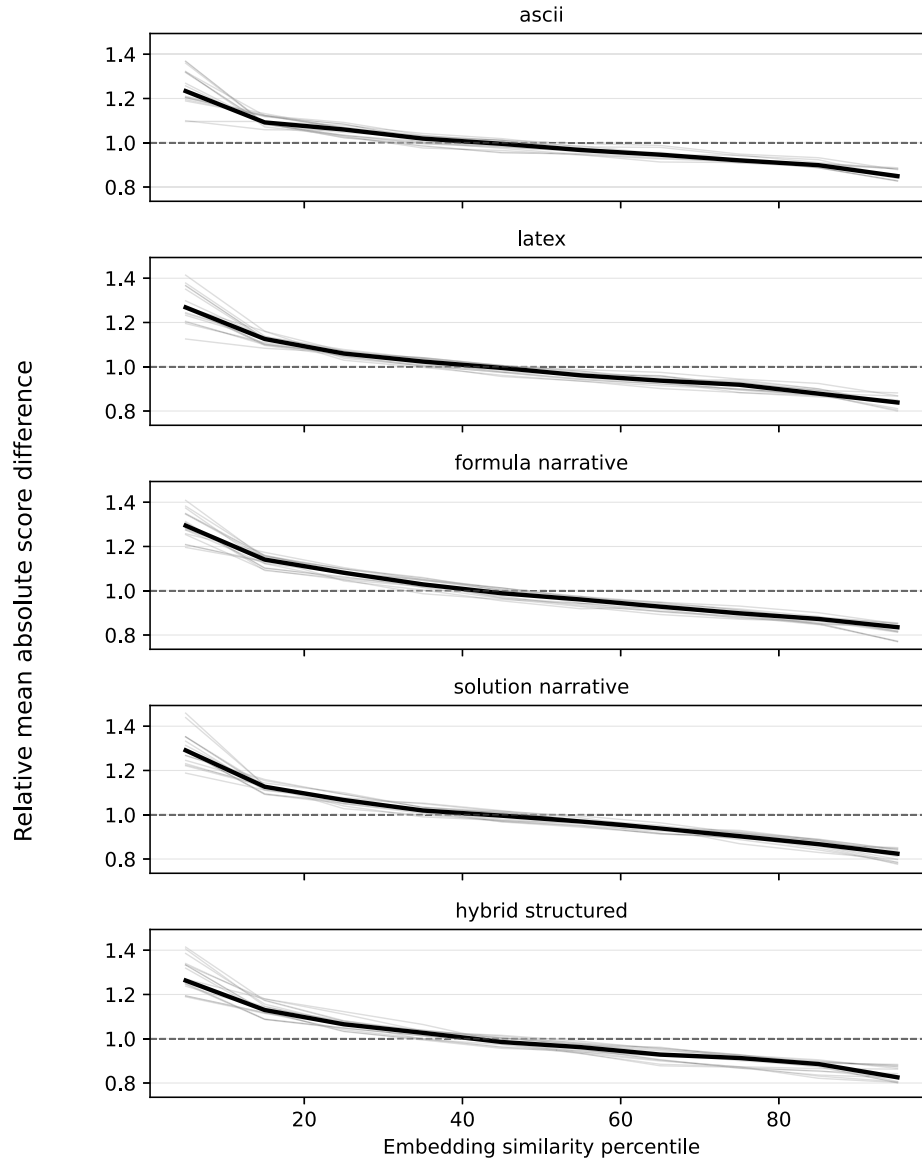


FIG. 3. Relationship between embedding similarity and human-score difference. For each transcription condition, unordered pairs of student responses were binned by embedding-similarity percentile. The vertical axis shows the mean absolute score difference in each bin, normalized by the mean absolute score difference across all pairs for the corresponding problem and condition. Thus, the dashed line at 1.0 represents the expected score difference for an arbitrary pair of responses; values below 1.0 indicate pairs that are more similar in score than random pairs. Thin gray lines show the 18 embedding-mechanism and similarity-method combinations for the given transcription mode: nine embedding mechanisms crossed with ordinary cosine and centered cosine similarity. Each line is aggregated across the four exam problems. The black line shows the median across these 18 combinations.

table shows that the strongest observed configurations all produce similarly modest score-similarity signals. The exact similarity metric was also not decisive: ordinary cosine similarity performed best for the strongest Qwen conditions, whereas centered cosine appeared among the strongest Jina retrieval conditions. Similarly, the Jina retrieval quantization levels produced nearly indistinguishable results. Overall, the response representation had the clearest descriptive effect, while the choice among

high-performing embedding mechanisms and similarity metrics produced only small differences.

Taken together, these results support a bounded interpretation of embedding similarity. Embeddings do not behave randomly: more similar response pairs tend, on average, to have somewhat smaller score differences. At the same time, the correlations are too small to support the stronger claim that embedding proximity identifies grading-equivalent responses. The observed signal is

TABLE II. Pairwise score-similarity correlations by transcription mode. Best values are the highest observed across embedding mechanisms and similarity metrics; median values are medians over those same conditions. Because each response appears in multiple pairs, these correlations are interpreted as descriptive effect-size summaries rather than as inferential estimates based on independent observations. Boldface indicates the largest value in each correlation column.

Transcription	Best Spearman	Median Spearman	Best Pearson	Median Pearson
ASCII	0.142	0.130	0.189	0.145
LaTeX	0.172	0.145	0.228	0.162
Formula narrative	0.182	0.160	0.217	0.187
Solution narrative	0.190	0.157	0.240	0.180
Hybrid structured	0.198	0.145	0.243	0.172

sufficient for exploratory organization or retrieval, but not for unsupervised grading.

B. Hierarchical clustering

For clustering, similarities were converted to distances using Eq. (5), and cluster quality was summarized primarily by Useful SD, defined in Sec. III D 2. The pairwise analysis showed that embedding similarity contains a weak but consistent score-related signal. We next asked whether this signal was strong enough to produce score-homogeneous clusters. This is the more demanding criterion for grading-related use: a cluster may be useful for organizing responses only if it contains a nontrivial number of responses and if the human scores within the cluster do not vary widely.

For each transcription condition, embedding mechanism, similarity metric, and linkage method, we examined hierarchical clusterings across the range of requested cluster numbers described above. We focused on clusters containing at least five responses and required at least 70% of responses to fall into such useful clusters. Among configurations meeting this criterion, lower weighted useful-cluster score standard deviation indicates greater score homogeneity.

The best clustering configurations again showed evidence of score enrichment, but not score equivalence. The strongest conditions produced useful-cluster score standard deviations on the order of several score points, rather than near-zero variation. Thus, the resulting clusters still contained substantial grading variation.

Tables IV and V summarize the hierarchical clustering results under the useful-cluster criterion. In contrast to the pairwise analysis, where the strongest condition used the hybrid structured transcription with a Qwen-derived representation and ordinary cosine similarity, the strongest clustering conditions were dominated by centered cosine similarity and complete linkage. The best overall configuration was the formula narrative transcription embedded with the Jina retrieval Q8 model, using centered cosine similarity and complete linkage. This condition placed 73.1% of responses into useful clusters and yielded a weighted useful-cluster score standard deviation of 2.547 score points.

The transcription-mode summary in Table IV shows the same broad pattern as the pairwise analysis: literal transcriptions were less effective than more semantically mediated representations. The best useful-cluster standard deviation decreased from 2.878 for ASCII and 2.797 for

TABLE III. Best transcription and embedding combinations for the pairwise score-similarity analysis. Metrics were computed from 122 525 unordered response pairs, but pairs are not statistically independent because each student response appears in many pairs. The table is therefore interpreted descriptively rather than as an inferential ranking. Boldface indicates the largest weighted Spearman and Pearson correlations among the configurations shown.

Rank	Transcription	Embedding	Similarity	Weighted Spearman	Weighted Pearson
1	Hybrid structured	qwenMath7B	Cosine	0.198	0.243
2	Solution narrative	qwenMath7B	Cosine	0.190	0.240
3	Hybrid structured	qwenMath7B	Centered_cosine	0.189	0.202
4	Hybrid structured	qwenMath1p5B	Cosine	0.183	0.233
5	Formula narrative	jinaRetQ8	Centered_cosine	0.182	0.191
6	Formula narrative	jinaRetF16	Centered_cosine	0.182	0.191
7	Formula narrative	jinaRetQ4	Centered_cosine	0.180	0.190
8	Formula narrative	oai3large	Cosine	0.180	0.217
9	Hybrid structured	qwenMath1p5B	Centered_cosine	0.178	0.192
10	Formula narrative	jinaRetF16	Cosine	0.177	0.212

TABLE IV. Clustering-quality summaries by transcription mode. Best useful SD is the lowest observed weighted useful-cluster score standard deviation across embedding mechanisms, similarity metrics, and linkage methods; best coverage is the highest observed useful-cluster coverage. These best values need not occur in the same configuration. For reference, before clustering, the unpartitioned problem response sets had a response-count-weighted mean score SD of 3.866 points and a response-count-weighted mean observed score range of 17.255 points across 992 student-problem responses. Useful-cluster statistics include only responses assigned to clusters containing at least five responses. Boldface indicates the smallest standard deviations and the largest coverage values in the respective columns.

Transcription	Best useful SD	Median useful SD	Best coverage	Median coverage
ASCII	2.878	3.139	0.883	0.784
LaTeX	2.797	3.052	0.874	0.766
Formula narrative	2.547	2.955	0.852	0.770
Solution narrative	2.722	3.005	0.874	0.779
Hybrid structured	2.717	2.939	0.873	0.758

LaTeX to 2.547 for formula narrative, 2.722 for solution narrative, and 2.717 for hybrid structured. Median performance showed the same tendency, although the differences were modest: the median useful-cluster standard deviation was 3.139 for ASCII, compared with 2.955 for formula narrative and 2.939 for hybrid structured. Thus, narrative and structured transcriptions again appear to preserve somewhat more grading-relevant structure than literal transcription, but the improvement is incremental rather than transformative.

The top-ranked clustering configurations in Table V also show that the exact ordering of embedding mechanisms should not be overinterpreted. The 10 best configurations occupy a relatively narrow band of weighted useful-cluster standard deviations, from 2.547 to 2.785. Most of these configurations use centered cosine similarity and complete linkage, suggesting that subtracting the common response-set direction and using a conservative linkage criterion helped form tighter score-enriched groups. However, even the best clusters retained substantial score variation. The weighted useful-cluster score ranges in the top

configurations were approximately seven to nine score points, indicating that clusters were far from score-pure.

Multiple values of k tied on both Useful SD and useful-cluster coverage in some conditions. As a sensitivity check, we repeated the selection using explained score variance rather than the smaller- k rule as the tertiary tie-breaker. This changed the selected k in 146 of the 720 problem-level conditions; in every changed case, the explained-variance rule selected a larger k , by one to four clusters. However, Useful SD, useful-cluster coverage, and Useful range were unchanged for all 180 aggregate configurations; only mean selected k changed in 102 configurations. Thus, the tie-breaking rule affected the reported degree of partitioning but not the score-homogeneity, coverage, or range summaries or the substantive conclusions.

For context, before clustering, the four unpartitioned problem response sets had a response-count-weighted mean score standard deviation of 3.866 points and a response-count-weighted mean observed range of 17.255 points. The best clustering configuration yielded a weighted useful-cluster standard deviation of 2.547

TABLE V. Eligible clustering configurations with the lowest weighted useful-cluster score SD, ranked from lowest to highest. Useful coverage, useful SD, and useful range are response-count-weighted summaries across problems. The reported k is the mean value selected across the four problems. The corresponding unpartitioned response-set baseline is reported in Table IV. Boldface indicates the smallest weighted useful-cluster score standard deviation among the configurations shown.

Rank	Transcription	Embedding	Similarity	Linkage	Useful coverage	Useful SD	Useful range	Mean selected k
1	Formula narrative	jinaRetQ8	Centered_cosine	Complete	0.731	2.547	7.142	47.000
2	Formula narrative	jinaRetF16	Centered_cosine	Complete	0.767	2.630	7.370	44.750
3	Hybrid structured	jinaMatchQ4	Centered_cosine	Complete	0.744	2.717	7.508	46.750
4	Solution narrative	jinaRetQ8	Centered_cosine	Complete	0.794	2.722	7.885	42.500
5	Solution narrative	jinaRetQ4	Centered_cosine	Complete	0.737	2.730	7.880	44.500
6	Formula narrative	jinaRetQ4	Centered_cosine	Complete	0.798	2.740	7.950	40.500
7	Hybrid structured	qwenMath1p5B	Centered_cosine	Complete	0.757	2.759	8.196	46.250
8	Hybrid structured	qwenMath7B	Centered_cosine	Complete	0.736	2.761	8.183	48.000
9	Solution narrative	jinaRetF16	Centered_cosine	Complete	0.807	2.761	7.967	41.250
10	Formula narrative	jinaRetF16	Centered_cosine	Average	0.852	2.785	9.206	31.750

points and a weighted useful-cluster range of 7.142 points. Thus, the useful clusters were descriptively more score-homogeneous than the unpartitioned response sets, but they still contained substantial grading variation (this comparison is descriptive because the useful-cluster statistics include only the 73.1% of responses assigned to clusters containing at least five student-problem responses).

These results therefore support the same bounded interpretation as the pairwise analysis. Embedding-based clustering does not simply fail: the clusters are score-enriched, and transcription choices affect the degree of enrichment. But the resulting groups are not score-equivalent. Even after selecting k retrospectively by minimizing Useful SD among eligible configurations, the best clusters retained substantial within-cluster score variation.

V. DISCUSSION

A. Semantic similarity

This study tested whether embedding-based similarity and clustering can support grading-related interpretations of handwritten physics solutions. The answer is mixed but clear. The embedding spaces were not arbitrary with respect to human scores: nearby responses tended, on average, to have somewhat more similar scores. However, the relationship was weak. The strongest pairwise conditions produced only modest correlations, and the strongest clustering conditions still contained substantial within-cluster score variation. Thus, embedding geometry was score-enriched, but not score-equivalent.

This distinction is central for assessment validity. A representation may be useful for organizing responses even if it is not valid for grading them. In the present study, embeddings appeared capable of producing neighborhoods of partially related student work. Such neighborhoods may be useful for browsing large response sets, selecting examples, identifying unusual solutions, supporting grader calibration, or organizing human review. They do not, however, define grading categories. The evidence does not support the inference that two responses close in embedding space deserve the same score, nor that membership in the same embedding cluster is sufficient evidence of rubric-equivalent reasoning.

The results also show that the textual representation of the handwritten work matters. Literal transcription preserved surface form, but narrative and structured representations generally produced stronger alignment with human scores. This suggests that making formulas, diagrams, numerical work, and apparent solution structure explicit in text can help embedding models capture some grading-relevant information. At the same time, the improvement was incremental. More semantically mediated transcriptions improved the signal but did not transform the problem into one that can be solved by unsupervised embedding similarity.

The exact embedding mechanism mattered less decisively. Different model families appeared among the strongest pairwise and clustering configurations, and the top-ranked conditions differed only slightly. Similarly, ordinary cosine similarity and centered cosine similarity each appeared useful in different settings. Centering was especially prominent among the strongest clustering configurations, plausibly because it removes a common problem-level direction in embedding space and emphasizes differences among responses. However, the fact that the best pairwise and clustering configurations were not the same reinforces the main methodological point: embedding geometry is not intrinsically valid for a grading use. It must be evaluated against the intended interpretation.

The clustering analysis makes this limitation especially visible. Clustering is a stronger requirement than pairwise similarity: it asks whether a response set can be partitioned into sizable groups with homogeneous scores. Even under favorable retrospective choices of k , evaluated using the eligibility and score-homogeneity criteria defined in Sec. III D 2, the best clusters retained score ranges of several points. Moreover, the selected values of k were relatively large, indicating that the most favorable clusterings carved the response sets into many small neighborhoods rather than discovering a small number of broad rubric categories. This pattern is consistent with exploratory usefulness but not with unsupervised grading.

These findings caution against equating semantic similarity with assessment similarity. Physics solutions are multimodal artifacts that combine prose, equations, diagrams, assumptions, units, signs, state variables, numerical substitutions, and final answers. Small symbolic differences may have large grading consequences, while visibly different solution paths may receive similar scores. Embeddings capture broad regularities in text, but grading depends on a construct defined by the assessment and rubric. In validity terms, clusters are not findings by themselves; they are hypotheses requiring external evidence.

B. A toy system

As we are considering embeddings as tools for qualitative physics education research or task automation, we need to be aware of their limitations. In our study, we found that measuring semantic similarity in terms of physics correctness may not be a strength of embeddings. To better understand this, we synthesized a set of student-problem answers to the prompt, “explain the work done by a Carnot engine,” see Table VI. To test whether embedding similarity reflects conceptual similarity or merely surface similarity, we constructed a toy set of short statements about the work done by a Carnot engine. Five statements were fully correct but expressed the same relationship in different formulations. For each correct statement, we then generated three closely worded variants, each preserving the surface formulation of its paired correct statement

TABLE VI. A synthetic set of student answers to the prompt, explain the work done by a Carnot engine. Blank entries in the Pair column indicate that pairing is not applicable.

ID	Class	Flaw family	Pair	Statement
C1	Correct	None		A Carnot engine does net work $W = Q_H * (1 - T_C/T_H)$ per cycle, using absolute temperatures.
F_a_1	Incorrect	Celsius-ratio	C1	A Carnot engine does net work $W = Q_H * (1 - T_C/T_H)$ per cycle, using temperatures in deg C.
F_b_1	Incorrect	Reversed-ratio	C1	A Carnot engine does net work $W = Q_H * (1 - T_H/T_C)$ per cycle, using absolute temperatures.
F_c_1	Incorrect	Rejected-heat/work-confusion	C1	A Carnot engine does net work $W = Q_H * (T_C/T_H)$ per cycle, using absolute temperatures.
C2	Correct	None		The work output equals the heat taken from the hot reservoir times the Carnot efficiency, $1 - T_C/T_H$.
F_a_2	Incorrect	Celsius-ratio	C2	The work output equals the heat taken from the hot reservoir times the Carnot efficiency, $1 - T_C/T_H$, with T measured in deg C.
F_b_2	Incorrect	Reversed-ratio	C2	The work output equals the heat taken from the hot reservoir times the Carnot efficiency, $1 - T_H/T_C$.
F_c_2	Incorrect	Rejected-heat/work-confusion	C2	The work output equals the heat taken from the hot reservoir times T_C/T_H .
C3	Correct	None		In one reversible cycle, a Carnot engine converts the fraction $1 - T_C/T_H$ of Q_H into useful work.
F_a_3	Incorrect	Celsius-ratio	C3	In one reversible cycle, a Carnot engine converts the fraction $1 - T_C/T_H$ of Q_H into useful work when T_C and T_H are in deg C.
F_b_3	Incorrect	Reversed-ratio	C3	In one reversible cycle, a Carnot engine converts the fraction $1 - T_H/T_C$ of Q_H into useful work.
F_c_3	Incorrect	Rejected-heat/work-confusion	C3	In one reversible cycle, a Carnot engine converts the fraction T_C/T_H of Q_H into useful work.
C4	Correct	None		For fixed reservoirs, the Carnot work is $Q_H - Q_C$, and reversibility gives $Q_C/Q_H = T_C/T_H$.
F_a_4	Incorrect	Celsius-ratio	C4	For fixed reservoirs, the Carnot work is $Q_H - Q_C$, and reversibility gives $Q_C/Q_H = T_C/T_H$ using deg C values.
F_b_4	Incorrect	Reversed-ratio	C4	For fixed reservoirs, the Carnot work is $Q_H - Q_C$, and reversibility gives $Q_C/Q_H = T_H/T_C$.
F_c_4	Incorrect	Rejected-heat/work-confusion	C4	For fixed reservoirs, the Carnot work is Q_C , and reversibility gives $Q_C/Q_H = T_C/T_H$.
C5	Correct	None		A perfect reversible heat engine still cannot turn all of Q_H into work; it produces $Q_H * (1 - T_C/T_H)$.
F_a_5	Incorrect	Celsius-ratio	C5	A perfect reversible heat engine still cannot turn all of Q_H into work; it produces $Q_H * (1 - T_C/T_H)$ with temperatures in deg C.
F_b_5	Incorrect	Reversed-ratio	C5	A perfect reversible heat engine still cannot turn all of Q_H into work; it produces $Q_H * (1 - T_H/T_C)$.
F_c_5	Incorrect	Rejected-heat/work-confusion	C5	A perfect reversible heat engine still cannot turn all of Q_H into work; it produces $Q_H * (T_C/T_H)$.

while introducing one of three recurring conceptual flaws. We refer to this surface-level grouping as the “statement pair”: for example, *C1*, *F_a_1*, *F_b_1*, and *F_c_1* share a common wording pattern. By contrast, we refer to the repeated conceptual error as the “flaw family”: for example, all (*F_a*) statements share the same underlying

misconception, even though they are phrased in different ways. Thus, successful semantic clustering would group statements by correctness or flaw family, whereas surface-based clustering would group statements by pair.

We embedded those statements using *oai3large* and calculated similarity matrices according to Eqs. (1)

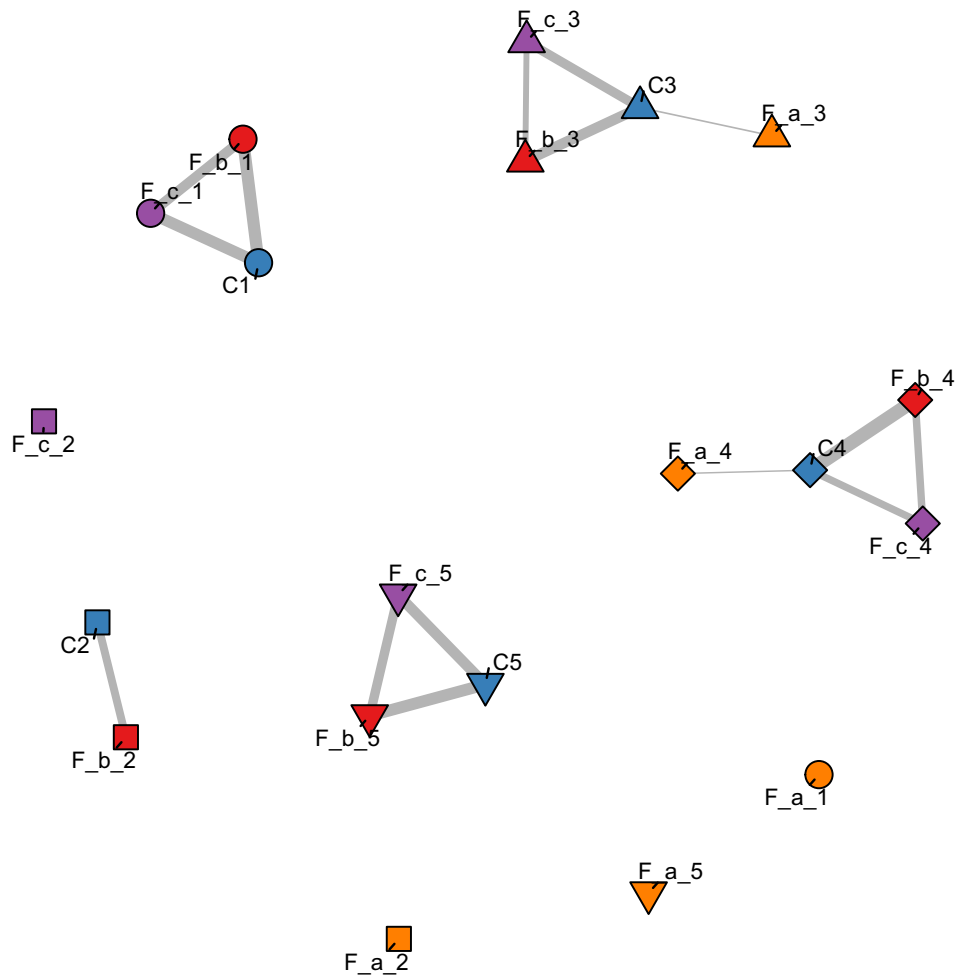


FIG. 4. Force-directed Fruchterman–Reingold plot of the cosine similarity of the embeddings of the statements in Table VI.

and (3). Figures 4 and 5 show the results using a force-directed Fruchterman–Reingold representation [99,100]. These plots would exhibit the clustering illustrated in Fig. 1 if clustering were by flaw family.

In fact, the resulting force-directed layouts show that the embeddings were strongly influenced by surface formulation. In particular, the mean-centered cosine layout forms visually distinct clusters around the five pairs: each correct statement is placed close to its three similarly worded flawed variants rather than near the other correct statements or near statements with the same flaw family. This is especially clear in the centered cosine graph, where the (C1/F_a_1/F_b_1/F_c_1), (C2/F_a_2/F_b_2/F_c_2), and analogous groups form local clusters by pair.

The ordinary cosine graph is less cleanly separated, but this should not be interpreted as more semantic: it mainly reflects the dominance of the common embedding direction, which compresses similarities and partially masks the same surface-feature effect. Centering removes this common background direction and therefore makes the residual formulation differences more visible, but those residual differences are still not the intended conceptual

ones. This helps explain why, in the main study, ordinary cosine and centered cosine did not differ substantially when clustering was evaluated against grading score as a proxy for semantic similarity: both measures were largely organizing responses by features other than the target physics meaning, and the fact that centering may sharpen a surface-feature organization would not necessarily be visible from score-based clustering alone.

This behavior is reminiscent of the classic expert-novice distinction in physics education research: common lore is that experts tend to categorize problems and solution approaches by underlying physical principles, whereas novices tend to categorize them by surface features [101–103]. In this limited respect, the embeddings behave like novices. LLMs, by contrast, can behave more like experts when prompted to evaluate solutions, since they have been shown capable of applying grading criteria to student work rather than relying only on surface similarity [75,77].

The present surface-form effect is related to, but distinct from, the discoverability bias reported by Gurin Schleifer *et al.* [89]. Their results showed that lower-quality knowledge profiles were less internally coherent in embedding

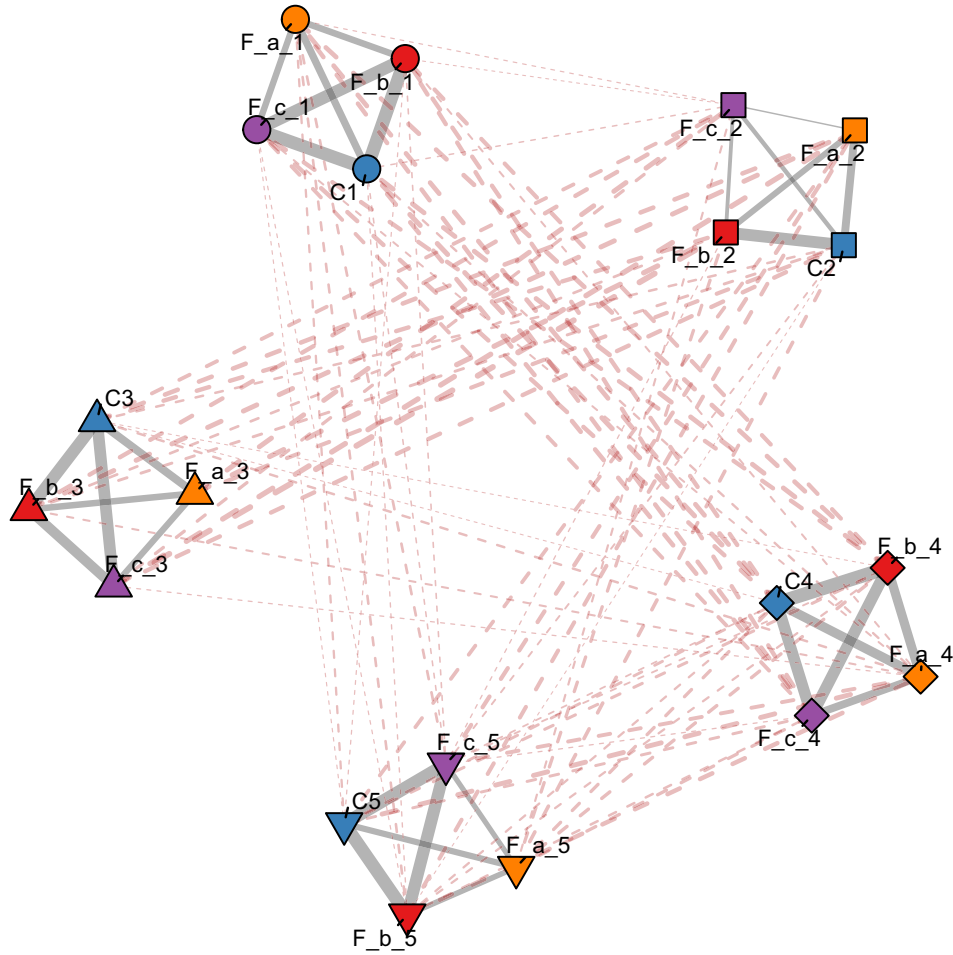


FIG. 5. Force-directed Fruchterman–Reingold plot of the centered cosine similarity of the embeddings of the statements in Table VI. This layout uses only positive similarities as attractive edges; negative edges are dashed and not used as weights.

space, whereas our controlled example shows that closely shared wording can dominate a common conceptual-error structure. Both findings caution against assuming that out-of-the-box embedding neighborhoods preserve the distinctions required for assessment.

Since LLM-based transcription is already part of the workflow for handwritten work, rubric-aware LLM grading or confidence-filtered LLM assistance may be a more direct route than relying on unsupervised embedding geometry alone.

VI. LIMITATIONS

Several limitations should be kept in mind. The data came from one high-stakes thermodynamics exam at one institution, and results may differ for other physics topics, lower-stakes assessments, typed solutions, conceptual explanations, or different rubric structures. Human scores were used as the external reference standard, but those scores are themselves an operational representation of the construct rather than a perfect ground truth. The transcription process also introduced a modeling layer between

the handwritten artifact and the embeddings. Although the prompts prohibited grading, correction, and comparison with ideal solutions, narrative and structured transcriptions necessarily involved some interpretation.

Rubric-blind should not be interpreted as knowledge-free or interpretation-free. The transcription model may use pretrained domain knowledge when resolving notation or producing a narrative, even though it was not asked to judge correctness. The stronger performance of narrative and structured representations may therefore reflect both reduced surface variation and model-mediated interpretation.

Finally, this study evaluated unsupervised embedding similarity and clustering, not supervised grading models trained on labeled data. A supervised classifier, an LLM-based grader, or a rubric-aware hybrid system may perform better for scoring. The present study addresses a narrower question: whether embedding geometry alone can justify grading-related interpretations.

VII. FUTURE WORK

Future work should examine whether embedding-based neighborhoods can support specific human-in-the-loop

workflows. For example, clusters could be used to select exemplars for rubric calibration, to identify batches of similar responses for grader review, or to flag responses that are isolated from all major solution patterns. Such applications require different validation criteria from autonomous grading: the question would be whether embeddings improve efficiency, consistency, or feedback quality when used by humans not whether they replace human judgment.

A second direction is to evaluate embedding spaces against more detailed rubric information. Total problem scores may obscure important distinctions among conceptual, mathematical, representational, and numerical components of a solution. Embeddings might align more strongly with some rubric dimensions than with total scores. Future analyses could therefore compare embedding neighborhoods with item-level rubric scores, error categories, or human-coded solution strategies.

Future studies should compare unsupervised embeddings with stronger baselines, including direct LLM comparisons, supervised classifiers, and retrieval-augmented grading workflows. Pairwise LLM comparison is computationally expensive at scale, but it may provide a useful benchmark for smaller samples. More broadly, validity evidence should be tied to the intended use: exploratory research, formative feedback, grading assistance, and autonomous scoring require different levels and kinds of evidence.

Embedding methods will continue to change. While similarity in rubric-relevant physics correctness may require expertlike reasoning, future embedding models may become better at capturing conceptual details and formula-level correctness. A possible research direction is therefore to develop or fine-tune embeddings specifically for assessment-relevant similarity in physics problem solving.

VIII. CONCLUSIONS

We evaluated whether rubric-blind transcription and out-of-the-box embeddings can support grading-related interpretations of handwritten physics solutions. Using 992 student-problem responses from a high-stakes thermodynamics exam, five textual representations, and nine embedding mechanisms, we compared pairwise embedding

similarity and embedding-based hierarchical clusters with human-assigned scores.

The results show a consistent but modest relationship between embedding geometry and score similarity. More similar response pairs tended to have somewhat smaller score differences, and embedding-based clusters were somewhat enriched for score similarity. Narrative and structured transcriptions generally improved alignment relative to literal transcription. However, the absolute strength of the relationship remained weak, and even the best useful clusters retained substantial score variation.

We therefore conclude that embedding clusters should be treated as hypotheses for human inspection, not as grading categories—yet. Embeddings may help organize large collections of handwritten physics solutions, identify neighborhoods of related responses, select examples, and support human-in-the-loop review. They do not, by themselves, provide a validated basis for unsupervised grading. For the data and methods studied here, embedding geometry was score-enriched but not score-equivalent.

ACKNOWLEDGMENTS

We would like to thank the students who participated in this study, as well as Daria Onishchuk and Alina Yaroshchuk, who assisted with logistics and the data analysis that laid the foundations for this study. We thank Richard Hahnloser for the organizational framework and computing resources, and we thank Jessica Lam for constructive and informative feedback. We would like to thank Cyrill Stotz for significant contributions to the preliminary studies for this project. We also thank Anna Kortemeyer for proofreading and commenting on the manuscript. This study is part of project Ethel [104].

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX: PROMPTS AND TRANSCRIPTIONS

Figures 6–15 show the original system prompts and sample transcriptions for the modes described in Sec. III A.

You are transcribing a scanned handwritten thermodynamics exam solution.

Purpose:

The transcription will be used for research on clustering similar student solutions. Your task is to represent what is visibly present on the page, not to judge correctness.

Input:

You receive one image of one exam page. The page may contain German and/or English text, formulas, numerical calculations, tables, sketches, diagrams, crossed-out work, arrows, labels, and marginal notes.

Task:

Transcribe the visible content as closely as possible into plain ASCII text.

General rules:

- Output only the transcription.
- Use plain ASCII characters only.
- Do not use LaTeX.
- Do not use Unicode mathematical symbols.
- Do not use Markdown formatting except simple section labels if useful.
- Do not solve the problem.
- Do not correct the student.
- Do not grade the work.
- Do not compare the work to an ideal solution.
- Do not infer missing steps unless they are visually clear from the page.
- Preserve German text as German and English text as English.
- Preserve the student's notation as much as possible.
- Preserve visible problem labels, subproblem labels, state labels, units, and final answers.
- Preserve crossed-out work when it is readable by writing [crossed out: ...].
- If text or math is illegible, write [illegible].
- If a region is blank or contains only unreadable marks, do not invent content.

Mathematical transcription:

- Use ASCII math notation.
- Examples:
 - Write x^2 for x squared.
 - Write $\sqrt{x}$ for square root of x .
 - Write $\int$ if an integral sign is visible.
 - Write dU/dt , $\dot{Q}$, $\dot{m}$, ΔT , η , c_p , c_v , h_1 , h_2 , T_1 , p_2 , v_3 , etc.
 - Write $\leq$, $\geq$, $\rightarrow$, $\approx$, $!$ where needed.
- Keep equations and numerical substitutions in approximately the order and layout shown on the page.
- Do not simplify or algebraically transform the student's formulas.
- Do not repair incorrect notation.

Diagrams and sketches:

- Do not ignore diagrams.
- Insert a short ASCII description at the approximate location of each diagram.
- Use the format:
 - [diagram: brief description]
- Include visible labels, axes, states, arrows, process paths, and annotations when recognizable.
- Examples:
 - [diagram: T-s diagram, vertical axis T, horizontal axis s, states 1, 2, 3 marked, process 1→2 drawn nearly vertical]
 - [diagram: piston-cylinder sketch with heat arrow Q_{in} and work arrow W_{out}]
 - [diagram: unclear sketch with labels 1 and 2]
- Do not interpret whether the diagram is correct.

Output format:

Use the following structure when possible:

PAGE_CONTENT:

Transcribe the page content in reading order.

DIAGRAMS_OR_SKETCHES:

Describe each visible diagram or sketch, if any.

CROSSED_OUT_OR_AMBIGUOUS_CONTENT:

List readable crossed-out content and major illegible regions, if any.

- If the page contains only one coherent block of work, a single PAGE_CONTENT section is sufficient.

FIG. 6. Prompt for mode `ascii`.

PAGE_CONTENT:

a)

$$b) \quad 0 = m_{12} [h_2 - h_5] - W_k$$

$$T_1 = -10 \text{ } ^\circ\text{C}$$

$$T_2 = T_1 - 6 \text{ K} = -16 \text{ } ^\circ\text{C}$$

$$\Rightarrow h_2 = h_g = 237,74 \text{ kJ/kg} \quad (\text{Tab A -10})$$

$$p_2 = 1,5745 \text{ bar} \quad (\text{Tab A -16})$$

$$c) \quad p_1 = p_7 = 1,5745 \text{ bar}$$

$$T_1 = T_i = -10 \text{ } ^\circ\text{C}$$

$$x_2 =$$

$$d) \quad e_K = |Q_{zu}| / |W_u|$$

$$Q_{zu} = Q_K$$

$$W_u = W_k = 2800$$

e)

DIAGRAMS_OR_SKETCHES:

[diagram: p-T diagram, vertical axis p, horizontal axis T; rising diagonal line from lower left to upper right. States 1, 2, 3 on upper region, and 2'', 3'' on lower horizontal segment. A slanted quadrilateral/cycle is drawn between states 1, 2, 3 and an upper left point labeled 4. Horizontal top edge labeled "isobar", middle horizontal edge between 1 and 2 labeled "isobar". Vertical arrow downward from 1 to 2'' . "(1)" written near left slanted side. "a)" circled at left.]

CROSSED_OUT_OR_AMBIGUOUS_CONTENT:

, [illegible] blacked out region at top right.

FIG. 7. Transcription in mode `ascii` of the student work in Fig. 2.

You are transcribing a scanned handwritten thermodynamics exam solution into a LaTeX-oriented textual representation.

Purpose:

The transcription will be used for research on clustering similar student solutions. Your task is to represent what is visibly present on the page, not to judge correctness.

Input:

You receive one image of one exam page. The page may contain German and/or English text, formulas, numerical calculations, tables, sketches, diagrams, crossed-out work, arrows, labels, and marginal notes.

Task:

Transcribe the visible content into text, using LaTeX notation for mathematical expressions where appropriate.

General rules:

- Output only the transcription.
- Do not solve the problem.
- Do not correct the student.
- Do not grade the work.
- Do not compare the work to an ideal solution.
- Do not infer missing steps unless they are visually clear from the page.
- Preserve German text as German and English text as English.
- Preserve the student's notation, variable names, signs, units, and numerical values as much as possible.
- Preserve visible problem labels, subproblem labels, state labels, units, and final answers.
- Preserve crossed-out work when it is readable by writing [crossed out: ...].
- If text or math is illegible, write [illegible].
- If a region is blank or contains only unreadable marks, do not invent content.

Mathematical transcription:

- Use LaTeX for formulas and equations.
- Put displayed equations on their own lines.
- Use common LaTeX notation such as: $\dot{Q}$, $\dot{W}$, $\dot{m}$, ΔT , η , c_p , c_v , h_1 , h_2 , T_1 , p_2 , v_3 .
- Preserve equations as written, even if they appear incorrect.
- Do not simplify, correct, or complete the student's derivation.
- If the student writes an equation in a nonstandard or ambiguous way, reproduce it as faithfully as possible and mark uncertainty if needed.
- Keep numerical substitutions and units visible.
- If a handwritten symbol is ambiguous, use [unclear symbol] or [illegible].

Text:

- Transcribe surrounding written explanations in normal text.
- Do not translate German into English.
- Do not paraphrase ordinary written text unless exact transcription is impossible.

Diagrams and sketches:

- Do not ignore diagrams.
- Describe each diagram in text at the approximate location where it appears.
- Use this format:
[diagram: ...]
- Include visible axes, labels, states, arrows, process paths, annotations, and crossed-out diagram elements.
- Examples:
[diagram: T-s diagram with vertical axis T and horizontal axis s; states 1, 2, 3 marked; path 1→2 nearly vertical; path 2→3 curves to the right]
[diagram: p-v diagram with vertical axis p and horizontal axis v; states 1 and 2 connected by a horizontal line]
[diagram: control volume sketch with inlet 1, outlet 2, heat arrow $\dot{Q}$, and work arrow $\dot{W}$]
- Do not interpret whether the diagram is thermodynamically correct.

Output format:

Use the following structure when possible:

PROBLEM LABELS:

List visible problem and subproblem labels.

TEXT:

Transcribe visible written prose.

EQUATIONS AND CALCULATIONS:

Transcribe formulas, derivations, substitutions, and final results using LaTeX notation.

DIAGRAMS OR SKETCHES:

Describe visible diagrams or sketches.

CROSSED-OUT OR AMBIGUOUS CONTENT:

List readable crossed-out content and major illegible regions.

† If a section has no visible content, write [none visible].

FIG. 8. Prompt for mode latex.

```

PROBLEM_LABELS:
\ (a) \), \ (b) \), \ (c) \), \ (d) \), \ (e) \)

TEXT:
Tab A \ (-10 \)

Tab A \ (-18 \)

EQUATIONS_AND_CALCULATIONS:
\ (b) \quad \theta = \dot{m}_{12} \cdot [h_2 - h_5] - \dot{W}_k \)

\ (T_i = -10^\circ \text{C} \)

\ (T_2 = T_i - 6, K = -16^\circ \text{C} \)

\ (\rightarrow h_2 = h_g = 237 \cdot \frac{kJ}{kg} \)

\ (p_2 = 1 \cdot 57415 \cdot \text{text{bar}} \)

\ (c) \quad p_1 = p_2 = 1 \cdot 57415 \cdot \text{text{bar}} \)

\ (T_1 = T_i = -10^\circ \text{C} \)

\ (x_2 = \)

\ (d) \quad \varepsilon_k = \frac{|\dot{Q}_{zu}|}{|\dot{W}_u|} \)

\ (\dot{Q}_{zu} = \dot{Q}_k \)

\ (\dot{W}_i = \dot{W}_k = 2800 \)

\ (e) \)

DIAGRAMS_OR_SKETCHES:
[diagram: \ (p) - \ (T) diagram with vertical axis \ (p) and horizontal axis \ (T); a slanted
phase-boundary-like line rises from lower left toward upper right. Near the top is a four-sided
cycle/parallelogram-like sketch. Top left state labeled \ (4), top right state labeled \ (3),
bottom right state labeled \ (2), lower left point on the slanted line labeled \ (2'), and to
its right a point labeled \ (3'). The top horizontal segment is annotated "isobar"; the middle
horizontal segment from left toward state \ (2) is also annotated "isobar". A left vertical arrow
points downward from the middle left point toward \ (2'), labeled \ (1). A slanted left side
from the upper left state downward is labeled \ (1).]

CROSSED_OUT_OR_AMBIGUOUS_CONTENT:
[none visible]

```

FIG. 9. Transcription in mode latex of the student work in Fig. 2.

You are converting a scanned handwritten thermodynamics exam solution into a natural-language description of the formulas and calculations visible on the page.

Purpose:

The output will be used for research on clustering similar student solutions. Your task is to describe the mathematical work that is visibly present, not to judge correctness.

Input:

You receive one image of one exam page. The page may contain German and/or English text, formulas, numerical calculations, tables, sketches, diagrams, crossed-out work, arrows, labels, and marginal notes.

Task:

Narrate the formulas, symbolic steps, numerical substitutions, and final numerical results in clear natural language.

General rules:

- Output only the requested narrative.
- Do not solve the problem.
- Do not correct the student.
- Do not grade the work.
- Do not compare the work to an ideal solution.
- Do not infer missing steps unless they are visually clear from the page.
- Do not say that a formula is right or wrong.
- Preserve German or English words when they are part of the student's visible writing, but the narrative itself may be in English.
- Mark uncertainty explicitly.
- If text or math is illegible, write [illegible].
- If a mathematical step is visible but its meaning is uncertain, write "The student appears to..." or "It is unclear whether..."

Narrating formulas:

- Convert formulas into readable English.
- Include the original variables and numerical values where visible.
- Examples:
 - If the page shows $E = m c^2$, write: "The student writes that energy equals mass times the speed of light squared, $E = m c^2$."
 - If the page shows $Q = m c_p \Delta T$, write: "The student uses heat transfer as mass times specific heat at constant pressure times temperature difference, $Q = m c_p \Delta T$."
 - If the page shows $\dot{Q} = \dot{m}(h_2 - h_1)$, write: "The student writes a heat-transfer rate as mass-flow rate times the enthalpy difference, $\dot{Q} = \dot{m}(h_2 - h_1)$."
- Retain equations in ASCII form after the narration when useful.
- Preserve signs, subscripts, units, and final numerical results.
- Do not algebraically simplify or repair the student's equations.

Numerical work:

- Describe numerical substitutions explicitly.
- Include units when visible.
- State final boxed or emphasized numerical answers if visible.
- If a number is ambiguous, mark it as [unclear number].
- If a unit is ambiguous, mark it as [unclear unit].

Diagrams and sketches:

- Do not ignore diagrams.
- Narrate what the diagram appears to show, without judging correctness.
- Include the diagram type if recognizable, such as T-s diagram, p-v diagram, piston-cylinder sketch, control volume sketch, cycle diagram, or table.
- Mention axes, state labels, arrows, process paths, and annotations.
- Example:
 - "A T-s diagram is drawn with T on the vertical axis and s on the horizontal axis. States 1 and 2 are labeled. The path from 1 to 2 is drawn approximately vertical, suggesting the student's representation of an isentropic or constant-entropy process, although no explicit label is visible."
- If the diagram type is unclear, write: "There is an unclear sketch with..."

Output format:

Use the following structure:

PROBLEM_LABELS:

List visible problem and subproblem labels.

FORMULA_NARRATIVE:

Describe the visible formulas and symbolic reasoning in natural language.

NUMERICAL_NARRATIVE:

Describe numerical substitutions, computations, units, and final numerical results.

DIAGRAM_NARRATIVE:

Describe visible diagrams, sketches, axes, labels, paths, and annotations.

VISIBLE_WRITTEN_TEXT:

Transcribe short written phrases that are not already included above, preserving their original language.

UNCERTAINTIES:

List illegible, ambiguous, or inferred elements.

• If a section has no visible content, write [none visible].

FIG. 10. Prompt for mode formula_narrative.

PROBLEM_LABELS:

- 4 a)
- b)
- c)
- d)
- e)

FORMULA_NARRATIVE:

For part a), the student draws a pressure-temperature style diagram with the vertical axis labeled p and the horizontal axis labeled T. A slanted line is drawn, and states are marked as 1, 2, 3, 2'', 3'', and A. A process line from 1 to A is labeled "(1)". Two horizontal segments are labeled "isobar", one between A and 3 and another between 1 and 2.

For part b), the student writes an energy balance in the form:

- " $\dot{Q} = \dot{m}_{\text{H}_2} [h_2 - h_5] - \dot{W}_{\text{dot},k}$ "

Then the student writes:

- " $T_i = -10^\circ\text{C}$ "
- " $T_2 = T_i - 6\text{ K} = -16^\circ\text{C}$ "

Next, the student appears to identify state 2 enthalpy with saturated vapor enthalpy:

- " $\Rightarrow h_2 = h_g = 237,74\text{ kJ/kg}$ " followed by "(Tab A -10)" or similar table reference.

Then the student writes a pressure value:

- " $p_2 = 1,5745\text{ bar}$ " with a table reference that appears to read "(Tab A -18)".

For part c), the student writes:

- " $p_4 = p_1 = 1,5745\text{ bar}$ "
- " $T_4 = T_1 = -10^\circ\text{C}$ "
- " $x_4 =$ " with no visible completion.

For part d), the student writes a coefficient or efficiency expression:

- " $e_k = |\dot{Q}_{\text{dot},\text{zu}}| / |\dot{W}|$ "

Then the student writes:

- " $\dot{Q}_{\text{dot},\text{zu}} = \dot{Q}_{\text{dot},k}$ "

and

- " $\dot{W}_{\text{dot},i} = \dot{W}_{\text{dot},k} = 280\text{ W}$ "

Part e) only shows the label "e)" with no visible further work.

NUMERICAL_NARRATIVE:

The student substitutes or states the inlet temperature as:

- " $T_i = -10^\circ\text{C}$ "

Then the student computes:

- " $T_2 = T_i - 6\text{ K} = -16^\circ\text{C}$ "

From a table lookup, the student writes:

- " $h_2 = h_g = 237,74\text{ kJ/kg}$ "

Another tabulated value is written as:

- " $p_2 = 1,5745\text{ bar}$ "

In part c), the same pressure is reused:

- " $p_4 = p_1 = 1,5745\text{ bar}$ "

and the temperature is restated:

- " $T_4 = T_1 = -10^\circ\text{C}$ "

In part d), the power is given numerically as:

- " $\dot{W}_{\text{dot},i} = \dot{W}_{\text{dot},k} = 280\text{ W}$ "

No boxed final answer is visible. The quality or dryness fraction line " $x_4 =$ " is left unfinished.

DIAGRAM_NARRATIVE:

There is a hand-drawn diagram in the upper left that appears to be a p-T diagram, with p on the vertical axis and T on the horizontal axis. A rising slanted line passes through lower state markings and through state 2. Near the upper portion, a four-sided shape is drawn with corners or points labeled A at the top left, 3 at the top right, 1 at the lower left, and 2 at the lower right. The top horizontal edge from A to 3 is labeled "isobar". The lower horizontal edge from 1 to 2 is also labeled "isobar". A descending vertical arrow is drawn from state 1 downward toward a lower point labeled 2'', and a short horizontal line extends from that lower point to another point labeled 3''. The line from 1 to A is marked "(1)". The exact thermodynamic meaning of the slanted boundary line and the lower points 2'' and 3'' is not explicitly written.

VISIBLE_WRITTEN_TEXT:

- "isobar"
- "(Tab A -10)" or similar
- "(Tab A -18)" or similar

UNCERTAINTIES:

- In part b), the equation contains " h_5 " or a similar subscript; the second enthalpy index is not fully clear.
- The table references are partially unclear; they appear to be "Tab A -10" and "Tab A -18" or similar.
- In part c), the state index may be 4, but the handwritten character could be another similar numeral.
- In part d), the denominator in the expression for e_k looks like " $|\dot{W}|$ " or possibly " $|\dot{W}_{\text{dot}}|$ "; the exact symbol is slightly unclear.
- The exact meaning of the labels 2'', 3'', and A in the diagram is not explained.
- Part e) has no visible content beyond the label.

FIG. 11. Transcription in mode formula_narrative of the student work in Fig. 2.

You are producing a concise narrative description of a scanned handwritten thermodynamics exam solution.

Purpose:
The output will be used for research on clustering similar student solutions. Your task is to describe what the student visibly did, not to judge whether it is correct.

Input:
You receive one image of one exam page. The page may contain German and/or English text, formulas, numerical calculations, tables, sketches, diagrams, crossed-out work, arrows, labels, and marginal notes.

Task:
Write a narrative account of the student's visible solution process.

General rules:

- Output only the requested narrative.
- Do not solve the problem.
- Do not correct the student.
- Do not grade the work.
- Do not compare the work to an official or ideal solution.
- Do not identify mistakes as mistakes.
- Do not say "correct", "incorrect", "right", "wrong", "valid", or "invalid" unless the student visibly wrote those words.
- Base the narrative only on visible content.
- Use cautious language for inferred reasoning, such as "the student appears to..." or "the student seems to..."
- Preserve important variables, formulas, numerical values, units, state labels, and final answers.
- Preserve German or English written phrases when relevant.
- If text or math is illegible, write [illegible].
- If a region is ambiguous, state the ambiguity.

Narrative content:

Describe:

- the apparent problem or subproblem labels,
- the physical system or thermodynamic process if visible,
- the quantities the student identifies,
- the formulas or balances the student uses,
- the sequence of substitutions or calculations,
- the final result or conclusion if visible,
- any sketches or diagrams and how they are used,
- crossed-out or abandoned work when visible.

Diagrams and sketches:

- Do not ignore diagrams.
- Include diagrams as part of the solution narrative.
- Describe the type of diagram if recognizable, such as T-s, p-v, control volume, piston-cylinder, cycle diagram, table, or schematic.
- Mention axes, states, process paths, arrows, labels, and annotations.
- Do not judge whether the diagram is thermodynamically appropriate.
- If a diagram appears central to the solution, state that it is visibly prominent or appears to be used as part of the solution.
- If the diagram is unclear, write that it is unclear.

Crossed-out work:

- Mention substantial crossed-out formulas, diagrams, or results if readable.
- Use the phrase "The page also contains crossed-out work showing..." without interpreting why it was crossed out.

Output format:
Use the following structure:

PROBLEM_LABELS:
List visible problem and subproblem labels.

SOLUTION_NARRATIVE:
Write one or more paragraphs describing what the student visibly did.

DIAGRAMS_AND_SKETCHES:
Describe visible diagrams or sketches and their labels.

KEY_VISIBLE_FORMULAS_AND_RESULTS:
List the most important visible formulas, substitutions, and final results in ASCII form.

CROSSED_OUT_WORK:
Describe readable crossed-out content, if any.

UNCERTAINTIES:
List illegible or ambiguous elements.

• If a section has no visible content, write [none visible].

FIG. 12. Prompt for mode `solution_narrative`.

PROBLEM_LABELS:

- 4 a)
- b)
- c)
- d)
- e)

SOLUTION_NARRATIVE:

For part 4a), the student draws a thermodynamic process diagram with a vertical axis labeled 'p' and a horizontal axis labeled 'T'. On this sketch, the student marks states '1', '2', '3', and also lower points labeled '2'' and '3''. A slanted line runs upward to the right, and a quadrilateral-like cycle or process shape is drawn between states 1, 2, 3, and an upper-left point that appears to be '4'. Horizontal segments are annotated 'isobar', including one near the segment from '1' to '2' and another near the upper segment leading toward '3'. A vertical arrow from state '1' downward toward '2'' is marked '(1)'. The student appears to be using this sketch to represent the sequence of thermodynamic states and pressure levels.

For part b), the student writes an energy balance in the form $\dot{Q} = \dot{m}_{12} [h_2 - h_5] - \dot{W}_k$ (the mass-flow subscript and the second enthalpy index are somewhat difficult to read, but this is the visible structure). Then the student states $T_1 = -10^\circ\text{C}$ and computes $T_2 = T_1 - 6\text{ K} = -16^\circ\text{C}$. From this, the student writes $\Rightarrow h_2 = h_g = 237,74\text{ kJ/kg}$ (Tab A -10) and below that $p_2 = 1,57415\text{ bar}$ (Tab A -18). The notation suggests the student is reading refrigerant properties from tables.

For part c), the student writes $p_4 = p_1 = 1,57415\text{ bar}$ and $T_4 = T_1 = -10^\circ\text{C}$. Below that, the student writes $x_4 =$ but does not continue with a visible expression or result.

For part d), the student writes a performance expression $e_k = |Q_{zu}| / |W_u|$. To the right, the student writes $Q_{zu} = Q_k$ and $W_u = W_k = 280\text{ W}$. This appears to be the visible setup for a coefficient or efficiency-type calculation, but no further numerical evaluation is shown.

For part e), only the label 'e)' is visible, with no further written work.

DIAGRAMS_AND_SKETCHES:

- A hand-drawn 'p-T' style diagram is visible in part 4a).
- Vertical axis: 'p'
- Horizontal axis: 'T'
- States labeled: '1', '2', '3', and likely '4' at the upper-left corner; also lower points '2'' and '3''
- A slanted line rises to the right through the diagram.
- A four-sided process shape is drawn between the numbered states.
- Horizontal segments are labeled 'isobar'.
- A vertical downward arrow from state '1' toward '2'' is labeled '(1)'.
- The diagram appears central to part a) and is used to indicate process paths and state relationships.

KEY_VISIBLE_FORMULAS_AND_RESULTS:

- $\dot{Q} = \dot{m}_{12} [h_2 - h_5] - \dot{W}_k$
- $T_1 = -10^\circ\text{C}$
- $T_2 = T_1 - 6\text{ K} = -16^\circ\text{C}$
- $h_2 = h_g = 237,74\text{ kJ/kg}$ (Tab A -10)
- $p_2 = 1,57415\text{ bar}$ (Tab A -18)
- $p_4 = p_1 = 1,57415\text{ bar}$
- $T_4 = T_1 = -10^\circ\text{C}$
- $x_4 =$
- $e_k = |Q_{zu}| / |W_u|$
- $Q_{zu} = Q_k$
- $W_u = W_k = 280\text{ W}$

CROSSED_OUT_WORK:

- [none visible]

UNCERTAINTIES:

- The upper-left state label in the diagram appears to be '4', but it is somewhat ambiguous.
- The exact indices in $\dot{m}_{12}$ and h_5 in the balance equation are slightly unclear due to handwriting.
- The table references in parentheses are partially ambiguous and appear as something like 'Tab A -10' and 'Tab A -18'.
- The labels '2'' and '3'' on the lower line in the sketch are somewhat faint.
- No visible completion is shown for $x_4 =$ in part c) or for part e).

FIG. 13. Transcription in mode solution_narrative of the student work in Fig. 2.

You are converting a scanned handwritten thermodynamics exam solution into a structured text representation for research on clustering similar student solutions.

Purpose:
The goal is to preserve grading-relevant visible information in a consistent textual form. The output may later be embedded and clustered. Your task is to describe what is visibly present on the page, not to judge correctness.

Input:
You receive one image of one exam page. The page may contain German and/or English text, formulas, numerical calculations, tables, sketches, diagrams, crossed-out work, arrows, labels, and marginal notes.

Task:
Produce a structured representation of the page that includes written text, formulas, calculations, numerical results, diagrams, sketches, and uncertainties.

General rules:

- Output only the structured representation.
- Use plain text.
- Use ASCII math notation for equations.
- Do not use LaTeX.
- Do not solve the problem.
- Do not correct the student.
- Do not grade the work.
- Do not compare the work to an official or ideal solution.
- Do not identify mistakes as mistakes.
- Do not say "correct", "incorrect", "right", "wrong", "valid", or "invalid" unless the student visibly wrote those words.
- Base the output only on visible content.
- Use cautious language for inferred reasoning, such as "appears to" or "seems to."
- Preserve German text as German and English text as English.
- Preserve variables, signs, subscripts, units, state labels, and final answers as much as possible.
- If text or math is illegible, write [illegible].
- If something is ambiguous, state the ambiguity.
- If a section has no visible content, write [none visible].

Mathematical representation:

- Use ASCII notation.
- Examples:
 - x^2 for x squared
 - $\sqrt{x}$ for square root of x
 - ΔT for temperature difference
 - $\dot{Q}$, $\dot{W}$, $\dot{m}$ for rates
 - h_1 , h_2 , T_1 , p_2 , v_3 for subscripts
 - $\rightarrow$ for arrows
 - $\approx$ for approximately equal
- Preserve the student's equations as written.
- Do not simplify, complete, or repair equations.
- Include numerical substitutions and units.
- Mark boxed, circled, underlined, or otherwise emphasized results as [emphasized result: ...].

Diagrams and sketches:
For every visible diagram or sketch, include a structured description. Do not ignore graphical work.

For each diagram, use this format:

DIAGRAM n:
 type: [T-s diagram / p-v diagram / control volume sketch / piston-cylinder sketch / cycle diagram / table / unclear sketch / other]
 axes: [describe axes if present]
 states_or_points: [list visible states or points]
 paths_or_processes: [describe visible lines, curves, arrows, directions, and process labels]
 annotations: [visible labels, heat/work arrows, assumptions, units, crossed-out labels]
 role_in_solution: [briefly state whether the diagram appears central, supporting, marginal, or unclear]
 uncertainty: [low / medium / high, with explanation if needed]

Do not judge whether a diagram is thermodynamically appropriate.

Output format:
Use exactly the following section headings:

PROBLEM_LABELS:
List visible problem and subproblem labels, such as "Aufgabe 1", "1a", "2b", etc.

VISIBLE_TEXT:
Transcribe visible written words or short phrases. Preserve the original language.

FORMULAS_AND_SYMBOLIC_WORK:
List visible formulas, equations, balances, symbolic derivations, and definitions in ASCII math notation.

NUMERICAL_WORK_AND_RESULTS:
List visible numerical substitutions, calculations, units, and emphasized or final numerical answers.

DIAGRAMS_AND_SKETCHES:
Describe each visible diagram or sketch using the DIAGRAM n format above.

APPARENT_METHOD:
Briefly describe the method or sequence of work the student appears to use, based only on visible content. Do not evaluate the method.

CROSSED_OUT_OR_ABANDONED_WORK:
List readable crossed-out formulas, calculations, diagrams, or text. If unreadable, state [crossed-out illegible content].

UNCERTAINTIES:
List illegible regions, ambiguous symbols, uncertain numbers, uncertain units, or inferred elements.

, Do not add any commentary beyond these sections.

FIG. 14. Prompt for mode hybrid_structured.

PROBLEM_LABELS:

- 4a)
- b)
- c)
- d)
- e)

VISIBLE_TEXT:

- isobar
- Tab A -10
- Tab A-18

FORMULAS_AND_SYMBOLIC_WORK:

- $\dot{Q} = m_{12}[h_2 - h_5] - \dot{W}_{dot_k}$
- $T_i = -10\text{ C}$
- $T_2 = T_i - 6\text{ K} = -16\text{ C}$
- $\Rightarrow h_2 = h_g = 237,74\text{ [kJ/kg]}$
- $p_2 = 1,5745\text{ bar}$
- $p_4 = p_1 = 1,5745\text{ bar}$
- $T_4 = T_1 = -10\text{ C}$
- $x_4 =$
- $e_k = \dot{Q}_{dot_zu} / \dot{W}_{u}$
- $\dot{Q}_{dot_zu} = \dot{Q}_{dot_k}$
- $\dot{W}_{dot_u} = \dot{W}_{dot_k} = 2800\text{ [unit unclear]}$

NUMERICAL_WORK_AND_RESULTS:

- $T_i = -10\text{ C}$
- $T_2 = -16\text{ C}$ from " $-10\text{ C} - 6\text{ K}$ "
- $h_2 = h_g = 237,74\text{ kJ/kg}$ (written with note "Tab A -10")
- $p_2 = 1,5745\text{ bar}$ (written with note "Tab A-18")
- $p_4 = p_1 = 1,5745\text{ bar}$
- $T_4 = T_1 = -10\text{ C}$
- $\dot{W}_{dot_u} = \dot{W}_{dot_k} = 2800\text{ [unit unclear, possibly W]}$
- $x_4 =$ [no value visible]

DIAGRAMS_AND_SKETCHES:

DIAGRAM 1:

type: other

axes: vertical axis labeled p; horizontal axis labeled T

states_or_points:

- state 1 near middle-left on an upper horizontal segment
- state 2 near middle-right on the same upper horizontal segment
- state 4 above state 1 on a top horizontal segment
- state 3 near top-right on the top horizontal segment
- state 2'' on lower-left on a slanted line
- state 3'' on lower-right on a short horizontal segment

paths_or_processes:

- a slanted rising line from lower-left toward upper-right
- horizontal segment from 1 to 2 labeled "isobar"
- horizontal segment from 4 to 3 labeled "isobar"
- vertical arrow from 1 upward to 4
- slanted segment from 4 down/right to 2
- slanted segment from 2 up/right toward 3
- short horizontal segment from 2'' to 3''
- label "(1)" near the left side of the upper quadrilateral-like shape

annotations:

- circled "4a)" at upper left of the page
- handwritten "isobar" on both visible horizontal process lines

role_in_solution: central

uncertainty: medium, because the exact identities of some paths and the label "(1)" are somewhat ambiguous

APPARENT_METHOD:

- The student first draws a p-T style process sketch with labeled states and isobaric segments for part 4a.
- In part b, the student writes an energy-style equation involving mass flow, enthalpy difference, and compressor work, then uses temperatures and table references to obtain h_2 and p_2 .
- In part c, the student sets p_4 equal to p_1 and T_4 equal to T_1 , then begins a quality expression x_4 without a visible completion.
- In part d, the student writes an efficiency or performance expression e_k as a ratio of heat input to work, and equates $\dot{Q}_{dot_zu}$ with $\dot{Q}_{dot_k}$ and $\dot{W}_{dot_u}$ with $\dot{W}_{dot_k}$, giving a numerical work value.
- Part e appears only as a label with no further visible work.

CROSSED_OUT_OR_ABANDONED_WORK:

- [none visible]

UNCERTAINTIES:

- In part b, the mass-flow symbol appears to be " m_{12} " or similar; subscript is not fully clear.
- In part b, the second enthalpy index in " $[h_2 - h_5]$ " may be ambiguous due to handwriting.
- The notation "Tab A -10" and "Tab A-18" is visible, but spacing/punctuation may be ambiguous.
- The unit written after 2800 is unclear; it may be omitted or not legible.
- In part d, the denominator of e_k is written as " $\dot{W}_{u}$ " or similar; exact symbol may be ambiguous.
- In the diagram, the lower labels "2''" and "3''" are somewhat uncertain and could possibly be read differently.
- The process labels and geometric connections in the diagram are partially inferred from the drawn lines.

FIG. 15. Transcription in mode hybrid_structured of the student work in Fig. 2.

- [1] J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics, *Eur. J. Sci. Educ.* **1**, 191 (1979).
- [2] T. Gok, The general assessment of problem solving processes and metacognition in physics education, *Int. J. Phys. Chem. Educ.* **2**, 110 (2010).
- [3] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020119 (2014).
- [4] E. Ince, An overview of problem solving studies in physics education, *J. Educ. Learn.* **7**, 191 (2018).
- [5] R. J. Dufresne and W. J. Gerace, Assessing-to-learn: Formative assessment in physics instruction, *Phys. Teach.* **42**, 428 (2004).
- [6] J. T. Lavery, W. Bauer, G. Kortemeyer, and G. Westfall, Want to reduce guessing and cheating while making students happier? Give more exams!, *Phys. Teach.* **50**, 540 (2012).
- [7] W. Fakcharoenphol and T. Stelzer, Physics exam preparation: A comparison of three methods, *Phys. Rev. ST Phys. Educ. Res.* **10**, 010108 (2014).
- [8] C. Wieman and K. Perkins, Transforming physics education, *Phys. Today* **58**, 36 (2005).
- [9] O. Uhdén, R. Karam, M. Pietrocola, and G. Pospiech, Modelling mathematical reasoning in physics education, *Sci. Educ.* **21**, 485 (2012).
- [10] J. Tuminaro and E. F. Redish, Elements of a cognitive model of physics problem solving: Epistemic games, *Phys. Rev. ST Phys. Educ. Res.* **3**, 020101 (2007).
- [11] R. E. Teodorescu, C. Bennhold, G. Feldman, and L. Medsker, New approach to analyzing physics problems: A taxonomy of introductory physics problems, *Phys. Rev. ST Phys. Educ. Res.* **9**, 010103 (2013).
- [12] C. Singh, Assessing student expertise in introductory physics with isomorphic problems. ii. Effect of some potential factors on problem solving and transfer, *Phys. Rev. ST Phys. Educ. Res.* **4**, 010105 (2008).
- [13] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, *Science* **208**, 1335 (1980).
- [14] M. M. Hull, E. Kuo, A. Gupta, and A. Elby, Problem-solving rubrics revisited: Attending to the blending of informal conceptual and formal mathematical reasoning, *Phys. Rev. ST Phys. Educ. Res.* **9**, 010105 (2013).
- [15] J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, *Phys. Rev. Phys. Educ. Res.* **12**, 010130 (2016).
- [16] D. E. Meltzer, Relation between students' problem-solving performance and representational format, *Am. J. Phys.* **73**, 463 (2005).
- [17] P. B. Kohl and N. D. Finkelstein, Student representational competence and self-assessment when solving physics problems, *Phys. Rev. ST Phys. Educ. Res.* **1**, 010104 (2005).
- [18] P. B. Kohl, D. Rosengrant, and N. D. Finkelstein, Strongly and weakly directed approaches to teaching multiple representation use in physics, *Phys. Rev. ST Phys. Educ. Res.* **3**, 010108 (2007).
- [19] D.-H. Nguyen and N. S. Rebello, Students' difficulties in transfer of problem solving across representations, in *AIP Conference Proceedings* (American Institute of Physics, Melville, NY, 2009), Vol. 1179, pp. 221–224.
- [20] B. R. Wilcox and S. J. Pollock, Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020124 (2014).
- [21] F. Reif, J. H. Larkin, and G. C. Brackett, Teaching general learning and problem-solving skills, *Am. J. Phys.* **44**, 212 (1976).
- [22] F. Reif, Millikan lecture 1994: Understanding and teaching important scientific thought processes, *Am. J. Phys.* **63**, 17 (1995).
- [23] L. Hsu, E. Brewster, T. M. Foster, and K. A. Harper, Resource letter RPS-1: Research in problem solving, *Am. J. Phys.* **72**, 1147 (2004).
- [24] J. Hattie, *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement* (Routledge, Abingdon, Oxfordshire, 2008).
- [25] F. Al-Salmani, J. Johnson, and B. Thacker, Assessing thinking skills in free-response exam problems: Pandemic online and in-person, *Phys. Rev. Phys. Educ. Res.* **19**, 010131 (2023).
- [26] D. A. Kashy, G. Albertelli, G. Ashkenazi, E. Kashy, H.-K. Ng, and M. Thoennessen, Individualized interactive exercises: A promising role for network technology, in *Proc. Frontiers in Education* (IEEE, Piscataway, NJ, 2001), Vol. 31, pp. 1073–1078.
- [27] G. Kortemeyer, E. Kashy, W. Benenson, and W. Bauer, Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses, *Am. J. Phys.* **76**, 438 (2008).
- [28] J. Risley, Motivating students to learn physics using an online homework system, Newsletter of the APS Forum on Education Fall 3 (2001).
- [29] T. Stelzer and G. Gladding, The evolution of web-based activities in physics at Illinois, Newsletter of the APS Forum on Education Fall 7 (2001).
- [30] R. J. Dufresne, J. P. Mestre, D. M. Hart, and K. A. Rath, The effect of web-based homework on test performance in large enrollment introductory physics courses, *J. Comput. Math. Sci. Teach.* **21**, 229 (2002).
- [31] C. Fredericks, Patterns of behavior in online homework for introductory physics, Ph.D. thesis, University of Massachusetts, 2007.
- [32] M. Richards-Babb, J. Drelick, Z. Henry, and J. Robertson-Honecker, Online homework, help or hindrance? What students think and how they perform, *J. Coll. Sci. Teach.* **40**, 81 (2011).
- [33] D. C. Perdian, Early identification of student performance and effort using an online homework system: A pilot study, *J. Sci. Educ. Technol.* **22**, 697 (2013).
- [34] E. W. Burkholder, J. K. Miles, T. J. Layden, K. D. Wang, A. V. Fritz, and C. E. Wieman, Template for teaching and assessment of problem solving in introductory physics, *Phys. Rev. Phys. Educ. Res.* **16**, 010123 (2020).
- [35] E. G. Offerdahl and J. B. Arneson, Formative assessment to improve student learning in biochemistry, in *Biochemistry*

- Education: From Theory to Practice* (ACS Publications, Washington, DC, 2019), pp. 197–218.
- [36] T. M. Clark, C. S. Callam, N. M. Paul, M. W. Stoltzfus, and D. Turner, Testing in the time of COVID-19: A sudden transition to unproctored online exams, *J. Chem. Educ.* **97**, 3413 (2020).
- [37] T. M. Gomez, C. Luciano, T. Nguyen, S. M. Villafañe, and M. N. Groves, Student success and experience in a flipped, senior physical chemistry course spanning before and after the covid-19 pandemic, *Chem. Educ. Res. Pract.* **26**, 210 (2025).
- [38] E. Kashy, B. M. Sherrill, D. T. Y. Tsai, D. Weinshank, M. Engelmann, and D. J. Morrissey, CAPA, an integrated computer assisted personalized assignment system, *Am. J. Phys.* **61**, 1124 (1993).
- [39] E. Kashy, S. J. Gaff, N. H. Pawley, W. L. Stretch, S. L. Wolfe, D. J. Morrissey, and Y. Tsai, Conceptual questions in computer-assisted assignments, *Am. J. Phys.* **63**, 1000 (1995).
- [40] R. Warnakulasooriya and D. Pritchard, Learning and problem-solving transfer between physics problems using web-based homework tutor, in *EdMedia+ Innovate Learning* (Association for the Advancement of Computing in Education (AACE), Norfolk, VA, 2005), pp. 2976–2983.
- [41] G. Gladding, B. Gutmann, N. Schroeder, and T. Stelzer, Clinical study of student learning using mastery style versus immediate feedback online activities, *Phys. Rev. ST Phys. Educ. Res.* **11**, 010114 (2015).
- [42] B. Gutmann, G. Gladding, M. Lundsgaard, and T. Stelzer, Mastery-style homework exercises in introductory physics courses: Implementation matters, *Phys. Rev. Phys. Educ. Res.* **14**, 010128 (2018).
- [43] M. Scott, T. Stelzer, and G. Gladding, Evaluating multiple-choice exams in large introductory physics courses, *Phys. Rev. ST Phys. Educ. Res.* **2**, 020102 (2006).
- [44] A. Pawl, R. E. Teodorescu, and J. D. Peterson, Assessing class-wide consistency and randomness in responses to true or false questions administered online, *Phys. Rev. ST Phys. Educ. Res.* **9**, 020102 (2013).
- [45] G. Kortemeyer, The psychometric properties of classroom response system data: A case study, *J. Sci. Educ. Technol.* **25**, 561 (2016).
- [46] J. Stewart and S. Ballard, Effect of written presentation on performance in introductory physics, *Phys. Rev. ST Phys. Educ. Res.* **6**, 020120 (2010).
- [47] J. L. Davis and T. N. McDonald, Online, handwritten or hybrid homework: What's best for our students in the long run?, *J. Online Eng. Educ.* **7**, 1 (2016).
- [48] D. J. Palazzo, Y.-J. Lee, R. Warnakulasooriya, and D. E. Pritchard, Patterns, correlates, and reduction of homework copying, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010104 (2010).
- [49] E. Gönültaş and G. Kortemeyer, Modeling unproductive behavior in online homework in terms of latent student traits: An approach based on item response theory, *J. Sci. Educ. Technol.* **26**, 139 (2017).
- [50] G. Kortemeyer, An empirical study of the effect of granting multiple tries for online homework, *Am. J. Phys.* **83**, 646 (2015).
- [51] K. VanLehn, B. van de Sande, R. Shelby, and S. Gershman, The Andes physics tutoring system: An experiment in freedom, in *Advances in Intelligent Tutoring Systems*, edited by R. Nkambou, J. Bourdeau, and R. Mizoguchi, Studies in Computational Intelligence Vol. 308 (Springer, Berlin, Heidelberg, 2010), pp. 421–443, [10.1007/978-3-642-14363-2_21](https://doi.org/10.1007/978-3-642-14363-2_21).
- [52] C. M. Nakamura, S. K. Murphy, M. G. Christel, S. M. Stevens, and D. A. Zollman, Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics, *Phys. Rev. Phys. Educ. Res.* **12**, 010122 (2016).
- [53] J. A. Shapiro, An algebra subsystem for diagnosing students' input in a physics tutoring system, *Int. J. Artif. Intell. Educ.* **15**, 205 (2005).
- [54] OpenAI, ChatGPT, <https://chat.openai.com/> (accessed April 2024).
- [55] OpenAI, ChatGPT, <https://openai.com/research/gpt-4> (accessed April 2024).
- [56] OpenAI, Hello GPT-4o, <https://openai.com/index/hello-gpt-4o/> (accessed June 2024).
- [57] J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez *et al.*, ChatGPT and large language models in academia: Opportunities and challenges, *BioData Mining* **16**, 20 (2023).
- [58] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education, *Learn. Individ. Differ.* **103**, 102274 (2023).
- [59] W. Yeadon and T. Hardy, The impact of AI in physics education: A comprehensive review from GCSE to university levels, *Phys. Educ.* **59**, 025010 (2024).
- [60] A. Sperling and J. Lincoln, Artificial intelligence and high school physics, *Phys. Teach.* **62**, 314 (2024).
- [61] G. Polverini and B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education, *Eur. J. Phys.* **45**, 025701 (2024).
- [62] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, *Phys. Rev. Phys. Educ. Res.* **19**, 020123 (2023).
- [63] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, Educational data augmentation in physics education research using ChatGPT, *Phys. Rev. Phys. Educ. Res.* **19**, 020150 (2023).
- [64] P. Wulff, Physics language and language use in physics—what do we know and how ai might enhance language-related research and instruction, *Eur. J. Phys.* **45**, 023001 (2024).
- [65] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido *et al.*, Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, *PLOS Digital Health* **2**, e0000198 (2023).
- [66] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman,

- S. Anadkat *et al.*, GPT-4 technical report, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [67] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, *Phys. Rev. Phys. Educ. Res.* **19**, 010132 (2023).
- [68] V. López-Simó and M. F. Rezende, Challenging chatgpt with different types of physics education questions, *Phys. Teach.* **62**, 290 (2024).
- [69] K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, and N. Haber, Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving, *Front. Educ.* **8**, 1330486 (2024).
- [70] G. Kortemeyer, M. Babayeva, G. Polverini, R. Widenhorn, and B. Gregorcic, Multilingual performance of a multimodal artificial intelligence system on multi-subject physics concept inventories, *Phys. Rev. Phys. Educ. Res.* **21**, 020101 (2025).
- [71] S. Küchemann, S. Steinert, N. Revenga, M. Schweinberger, Y. Dinc, K. E. Avila, and J. Kuhn, Can ChatGPT support prospective teachers in physics task development?, *Phys. Rev. Phys. Educ. Res.* **19**, 020128 (2023).
- [72] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, *Phys. Rev. Phys. Educ. Res.* **18**, 010141 (2022).
- [73] T. Wan and Z. Chen, Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning, *Phys. Rev. Phys. Educ. Res.* **20**, 010152 (2024).
- [74] G. Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study, *Phys. Rev. Phys. Educ. Res.* **19**, 020163 (2023).
- [75] G. Kortemeyer, J. Nöhl, and D. Onishchuk, Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study, *Phys. Rev. Phys. Educ. Res.* **20**, 020144 (2024).
- [76] T. Liu, J. Chatain, L. Kobel-Keller, G. Kortemeyer, T. Willwacher, and M. Sachan, AI-assisted automated short answer grading of handwritten university level mathematics exams, [arXiv:2408.11728](https://arxiv.org/abs/2408.11728).
- [77] G. Kortemeyer and J. Nöhl, Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study, *Phys. Rev. Phys. Educ. Res.* **21**, 010136 (2025).
- [78] Á. Fraknói, A. Kornai, and Z. Zombori, Embedding mathematical formulas into vector space, in *8th Conference on Artificial Intelligence and Theorem Proving (AITP 2023)*, Aussois, France (2023), https://aitp-conference.org/2023/abstract/AITP_2023_paper_17.pdf.
- [79] X. He, Mathematical formulas-based scientific literature retrieval with heterogeneous network semantic enhancement, in *Web Information Systems Engineering - WISE 2024*, edited by M. Barhamgi, H. Wang, and X. Wang, Lecture Notes in Computer Science Vol. 15436 (Springer, Singapore, 2025), pp. 26–41, [10.1007/978-981-96-0579-8_3](https://doi.org/10.1007/978-981-96-0579-8_3).
- [80] Z. Xu, *Hesitant Fuzzy Sets Theory*, Studies in Fuzziness and Soft Computing Vol. 314 (Springer, Cham, 2014), [10.1007/978-3-319-04711-9](https://doi.org/10.1007/978-3-319-04711-9).
- [81] X. Lin, Z. Huang, H. Zhao, E. Chen, Q. Liu, H. Wang, and S. Wang, HMS: A hierarchical solver with dependency-enhanced understanding for math word problem, in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021* (AAAI Press, Palo Alto, CA, 2021), pp. 4232–4240, [10.1609/aaai.v35i5.16547](https://doi.org/10.1609/aaai.v35i5.16547).
- [82] *Applying Machine Learning in Science Education Research: When, How, and Why?*, edited by P. Wulff, M. Kubsch, and C. Krist (Springer, Cham, 2025), [10.1007/978-3-031-74227-9](https://doi.org/10.1007/978-3-031-74227-9).
- [83] T. O. B. Odden, H. Tyseng, J. T. Mjaaland, M. F. Kreutzer, and A. Maltse-Sørensen, Using text embeddings for deductive qualitative research at scale in physics education, *Phys. Rev. Phys. Educ. Res.* **20**, 020151 (2024).
- [84] M. Caramaschi and T. O. B. Odden, Analyzing the history of physics education in the USA and Europe through natural language processing, *Phys. Rev. Phys. Educ. Res.* **21**, 020153 (2025).
- [85] A. Buzzell, T. J. Atherton, and R. Barthelemy, Modern physics courses: Understanding the content taught in the U.S., *Phys. Rev. Phys. Educ. Res.* **21**, 010139 (2025).
- [86] K. Gili, K. Heuton, A. Shah, D. Hammer, and M. C. Hughes, Combining physics education and machine learning research to measure evidence of students' mechanistic sensemaking, *Phys. Rev. Phys. Educ. Res.* **21**, 020161 (2025).
- [87] T. Wyrwich, M. Kubsch, H. Drachsler, and K. Neumann, Tracking students' progression in developing understanding of energy using AI technologies, *Phys. Rev. Phys. Educ. Res.* **21**, 010152 (2025).
- [88] P. P. Martin, D. Kranz, P. Wulff, and N. Graulich, Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry, *J. Res. Sci. Teach.* **61**, 1757 (2024).
- [89] A. Gurin Schleifer, B. Beigman Klebanov, and G. Alexandron, Uncovering measurement biases in LLM embedding spaces: The Anna Karenina principle and its implications for automated feedback, *Int. J. Artif. Intell. Educ.* **35**, 2821 (2025).
- [90] M. T. Kane, Validating the interpretations and uses of test scores, *J. Educ. Measure* **50**, 1 (2013).
- [91] S. Messick, Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *Am. Psychol.* **50**, 741 (1995).
- [92] Z. Chen and T. Wan, Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy, *Phys. Rev. Phys. Educ. Res.* **21**, 010126 (2025).
- [93] R. Mok, F. Akhtar, L. Clare, C. Li, J. Ida, L. Ross, and M. Campanelli, Using large language models for grading in

- education: An applied test for physics, *Phys. Educ.* **60**, 035006 (2025).
- [94] OpenAI, Introducing GPT-5.4, <https://openai.com/index/introducing-gpt-5-4/> (Accessed: 2026-06-14).
- [95] OpenAI, text-embedding-3-large documentation, <https://platform.openai.com/docs/models/text-embedding-3-large> (2025), accessed: 2025-09-08.
- [96] Jina AI, Jina-embeddings-v4 model documentation, <https://jina.ai/models/jina-embeddings-v4/> (2025).
- [97] M. Günther, S. Sturua, M. K. Akram, I. Mohr, A. Ungureanu, B. Wang, S. Eslami, S. Martens, M. Werk, N. Wang, and H. Xiao, Jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval, in *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, edited by D. I. Adelani, C. Arnett, D. Ataman, T. A. Chang, H. Gonen, R. Raja, F. Schmidt, D. Stap, and J. Wang (Association for Computational Linguistics, Suzhou, China, 2025), pp. 531–550, [10.18653/v1/2025.mrl-main.36](https://arxiv.org/abs/10.18653/v1/2025.mrl-main.36).
- [98] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, K. Lu, M. Xue, R. Lin, T. Liu, X. Ren, and Z. Zhang, Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, [arXiv:2409.12122](https://arxiv.org/abs/2409.12122).
- [99] T. M. J. Fruchterman and E. M. Reingold, Graph drawing by force-directed placement, *Software: Practice and Experience* **21**, 1129 (1991).
- [100] G. Kortemeyer, Virtual-reality graph visualization based on Fruchterman-Reingold using Unity and SteamVR, *Inf. Visualization* **21**, 143 (2022).
- [101] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, *Cogn. Sci.* **5**, 121 (1981).
- [102] S. F. Wolf, D. P. Dougherty, and G. Kortemeyer, Rigging the deck: Selecting good problems for expert-novice card-sorting experiments, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020116 (2012).
- [103] S. F. Wolf, D. P. Dougherty, and G. Kortemeyer, Empirical approach to interpreting card-sorting data, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010124 (2012).
- [104] G. Kortemeyer, Ethel: A virtual teaching assistant, *Phys. Teach.* **62**, 698 (2024).