Decoding Feynman: A Cognitive Knowledge Graph for Physics Reasoning and Education

Jawwad Shadman Siddique
Oregon State University
Corvallis, Oregon
siddiquj@oregonstate.edu

Abstract

Physics education produces graduates who know physics facts but struggle to reason like physicists. We argue this gap arises because physics is a thinking language, a system of cognitive moves with identifiable vocabulary, grammar rules, and question patterns, while current curricula teach content without explicitly teaching that language. We present **PhysCog**, a cognitive knowledge graph of 50 empirically grounded reasoning clusters extracted from the Feynman Lectures on Physics using a structured LLM annotation pipeline. Each cluster encodes signature vocabulary, implicit grammar, characteristic questions, and analogical bridges of a distinct physics reasoning mode. We find that 70% of clusters are portable across multiple physics subfields while 30% are domain-bound. These portable clusters constitute transferable cognitive primitives of physics expertise. We develop an interactive cognitive navigation mind-map in which learners explore the cluster graph, navigate cognitive bridges, and receive concept-guided problem walkthroughs. Our work provides a cognitive atlas of expert physics reasoning, with direct applications to physics pedagogy and AI evaluation.

1 Introduction

Expert physicists share a recognizable cognitive style: they check limiting cases before trusting a formula, seek symmetry before writing an equation, and draw a potential landscape before computing a trajectory. These habits are rarely explicitly taught; students absorb them only through prolonged exposure to expert practice. The result is a persistent gap between physics knowledge and physics *thinking* (Larkin et al., 1980; Clement, 1982). This gap has been documented in modern assessments as well: analyses of student problem-solving across introductory courses consistently show that correct equation selection does not predict correct reasoning structure, even in high-performing students (Docktor and Mestre, 2014).

We propose a reframing: physics is not merely a body of knowledge but a *thinking language* with vocabulary (concepts and quantities), grammar (implicit logical rules physicists follow unconsciously), and syntax (the question sequences structuring a reasoning chain). Fluency means thinking in this language, not translating into it from common sense. The Feynman Lectures on Physics (Feynman et al., 1964) are uniquely suited to studying this language: three volumes of expert explanation in plain English in which Feynman narrates his own reasoning process deliberately, making the implicit grammar of physics visible.

Contributions. (i) A RAG-based annotation pipeline for extracting cognitive reasoning move types from physics text, guided by a six-category taxonomy grounded in prior science education research (Section 2). (ii) Application to the Feynman Lectures discovering 50 stable clusters, of which 35 are portable across subfields by a pre-specified criterion (Section 3). (iii) **PhysCog**, an interactive cogni-

● portable ○ domain-bound

A Causal & Mechanical		B Symmetry & Conservation		C Wave & Interference		D Statistical & Probabilistic		E Field & Geometric		F Quantum Thinking		G Scale & Approxim.		H Analogical Bridges	
Force-cause	○	Symmetry ID	●	Superposition	●	Entropy counting	●	Field lines	○	Wavefunction amp	○	Order of magnitude	●	Math analogy transfer	●
Energy accounting	●	Conservation law	●	Path length diff	○	Boltzmann weights	●	Gauss law	○	Superpos/meas.	○	Perturbation exp.	●	Cross-domain isom	●
Momentum	○	Broken symmetry	●	Wavefront arch	○	Ensemble thinking	●	Potential landscape	●	Uncertainty trade	○	Renorm group	●	Limit case bridge	●
Rotational analogy	●	Gauge invariance	○	Resonance/modes	●	Fluctuation	●	Gradient as force	●	Operator/eigenstate	○	Mean field	●		
Constraint reasoning	●	Noether bridging	●	Dispersion	●	Phase transitions	●	Curved spacetime	○	Perturbation	●	Continuum limit	●		
Frame switching	●	Group theory	●	Boundary logic	●	Info-theoretic bridge	●	Metric reasoning	○	Path integral	○				
Limiting case	●			Fourier decomp	●			Topological	●	Spin/ang-mom	○				
Dimensional analysis	●									Many-body emerge.	●				

Figure 1: All 50 PhysCog clusters across 8 cognitive groups (A–H). Filled dots indicate portable clusters (n=35, 70%); open dots indicate domain-bound clusters (n=15, 30%). Groups G and H are entirely portable.

tive knowledge graph for learner navigation and concept-guided problem walkthroughs (Section 4), with proposed use as a benchmark scaffold for AI reasoning evaluation.

2 Method

Corpus and chunking. We scrape all three volumes of the Feynman Lectures (115 chapters, ≈1,500 pages) from `feynmanlectures.caltech.edu`. Text is chunked into *explanatory units* of 50–200 words aligned to subsection boundaries, yielding approximately 800 passages. Token-count chunking is avoided because Feynman structures each paragraph as a self-contained reasoning unit.

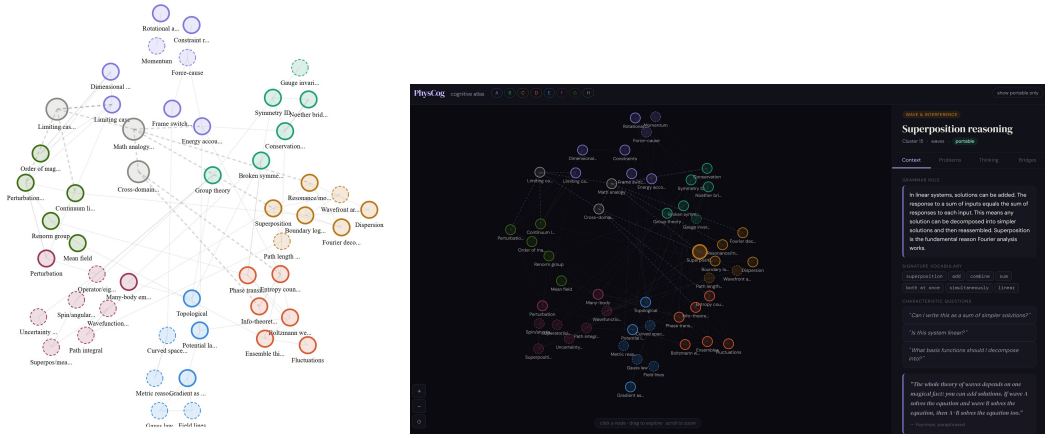
Cognitive move taxonomy. The six move types are grounded in two independent lines of prior work. The VOCAB/GRAMMAR/QUESTION/CONVERSE distinction adapts Redish's epistemological framing of expert physics discourse (Redish, 2003), while the NARRATIVE versus ARCHITECTURAL contrast is derived from Larkin and Simon's finding that experts and novices differ not in what they know but in how they organize spatial versus sequential representations (Larkin et al., 1980). The six categories are: VOCAB (grounding a concept through behavior or analogy); GRAMMAR (verbalizing a reasoning rule such as checking units or testing limits); QUESTION (a rhetorical question that opens the next reasoning step); CONVERSE (addressing nature as a conversational partner); NARRATIVE (causal, time-based reasoning, predominant in mechanics); and ARCHITECTURAL (spatial, simultaneous reasoning, predominant in optics and field theory). The NARRATIVE versus ARCHITECTURAL contrast is the paper's central cognitive finding: students who excel at mechanics frequently struggle with optics not because of mathematical difficulty but because they apply temporal moves where spatial-structural moves are required.

Extraction and cluster discovery. A Retrieval-Augmented Generation (RAG) pipeline annotates each passage. The Feynman corpus is indexed as a vector store; at inference time, each passage is retrieved alongside semantically similar passages for contextual grounding, and the augmented context is passed to Claude (Anthropic) to classify every sentence into one of the six move types, identify the signal phrase, and assign a confidence score in $[0, 1]$. Annotations below 0.80 are discarded; the RAG pipeline runs twice at temperature $= 0$, and passages with inter-run disagreement above 30% are excluded. Cross-model agreement on a 150-passage held-out sample serves as the primary validation metric. Retained annotations are embedded with SPECTER2 sentence-level embeddings and clustered with HDBSCAN (McInnes et al., 2017). Clusters with Jaccard similarity > 0.70 under split-half replication are retained, yielding 50 stable clusters. Each cluster is characterized by querying the RAG system to synthesize a name, vocabulary list, grammar rule, and characteristic question grounded in retrieved Feynman passages.

3 Cognitive Atlas

Figure 1 shows the full taxonomy across 8 groups: causal-mechanical (A), symmetry and conservation (B), wave and interference (C), statistical-probabilistic (D), field and geometric (E), quantum-specific (F), scale and approximation (G), and analogical bridges (H).

Portability criterion and breakdown. A cluster is *portable* if: (a) it has annotated passages in at least three of the eight corpus subfields (mechanics, E&M, optics, thermodynamics, quantum



(a) Force-directed graph of all 50 clusters. Solid borders mark portable clusters; dashed borders mark domain-bound clusters; dashed edges mark Group H analogical bridge connections.

(b) Detail panel for *Superposition reasoning* (Cluster 15). The Context tab shows the grammar rule, vocabulary chips, characteristic questions, and a Feynman paraphrase. Four tabs give access to Problems, Thinking protocol, and Bridges.

Figure 2: PhysCog interactive cognitive knowledge graph, built with D3.js (Bostock et al., 2011). Group badges (A–H) enable single-click filtering; a portable-only toggle isolates the 35 transferable modes.

mechanics, special relativity, field theory, statistical physics), and (b) its grammar rule instantiates without domain-specific substitution in each covered subfield, as verified by the RAG system. Table 1 shows the per-group breakdown. The asymmetry across groups is the key result: Group F (Quantum) is 75% domain-bound while Groups G and H are entirely portable, a pattern that would not emerge from random labeling and provides direct evidence the criterion is doing real discriminative work (Hestenes et al., 1992).

Table 1: Portability breakdown by group. P = portable, D = domain-bound. Subfields checked: Mech, E&M, Opt, Thermo, QM, SR, Field, Stat.

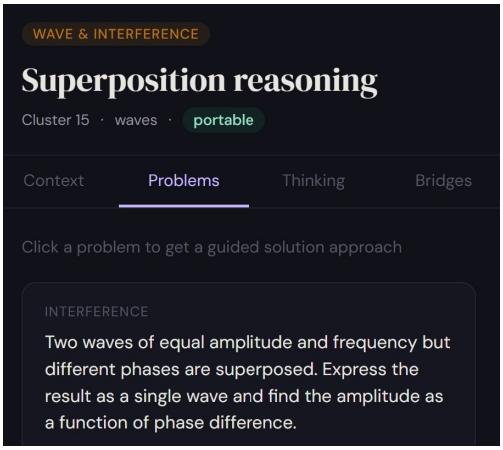
Grp	Name	P	D	Subfields covered (portable clusters)
A	Causal & Mechanical	6	2	Mech, Thermo, Field, SR
B	Symmetry & Conserv.	5	1	All except QM
C	Wave & Interference	5	2	Mech, Opt, QM, Stat
D	Statistical & Prob.	6	0	All 8
E	Field & Geometric	3	4	Mech, Thermo, Field
F	Quantum Thinking	2	6	QM only (domain-bound)
G	Scale & Approx.	5	0	All 8
H	Analogical Bridges	3	0	All 8
	Total	**35**	**15**	70% portable

Worked example: Superposition reasoning (Cluster 15). **Grammar rule**: *In linear systems, solutions can be added.* **Vocabulary**: superposition, add, combine, sum, linear. **Questions**: "Can I write this as a sum of simpler solutions?"; "Is this system linear?" **Subfields**: mechanics (normal modes), optics (interference), QM (energy eigenstates), thermodynamics (linear response) – portable across 4 of 8.

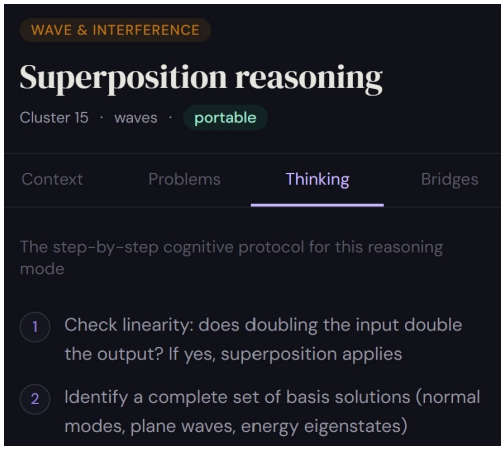
4 The PhysCog Application

PhysCog is a standalone web application built with D3.js and a RAG pipeline powered by Claude (Anthropic). The main graph (Figure 2a) organizes all 50 clusters into 8 orbital groups. Selecting any node opens a four-tab detail panel (Figure 2b). The **Context** tab shows the grammar rule, vocabulary

chips, characteristic questions, and a Feynman paraphrase. The **Problems** tab (Figure 3a) lists three domain-appropriate problems; clicking one calls the RAG pipeline to return a cognitive scaffold that names the reasoning mode before guiding the solution (Sweller, 1988). The **Thinking** tab (Figure 3b) presents the five-step cognitive protocol ending with an activation prompt. The **Bridges** tab lists cross-domain connections as clickable items.



(a) Problems tab: three domain-appropriate problems for Cluster 15. Clicking a problem returns a RAG-generated cognitive scaffold framed around the cluster's grammar rule.



(b) Thinking tab: five-step cognitive protocol for activating Superposition reasoning, ending with the prompt *"Can I write this as a sum of simpler solutions?"*

Figure 3: Detail panel tabs for Cluster 15 (Superposition reasoning, Wave & Interference, portable).

AI evaluation scaffold. Beyond pedagogy, PhysCog provides a structured scaffold for evaluating LLM reasoning. Each cluster defines a grammar rule and characteristic question that constitutes a testable cognitive signature. An LLM is probed with cluster-specific problems; its response is scored against the grammar rule to assess whether it applies the correct reasoning mode rather than reaching a correct answer by a different path. This enables evaluation of *reasoning style*, not just answer correctness, providing a richer diagnostic than standard physics benchmarks.

5 Discussion and Limitations

The portability finding has a direct pedagogical implication. If 70% of physics reasoning moves transfer across subfields, a curriculum that explicitly teaches these as cognitive tools could reduce the entry cost for new domains. A concrete integration path is to use the PhysCog Thinking tab as a pre-lecture primer: students read the five-step protocol for the relevant cluster before class, activate the characteristic question, and arrive primed to recognize the grammar rule in the instructor's derivation. This mirrors evidence-based approaches in which explicit instruction in reasoning strategies improves problem-solving transfer (Docktor and Mestre, 2014).

The primary limitations are: (i) the taxonomy derives from a single expert corpus, possibly reflecting Feynman's style; (ii) LLM annotation introduces model-specific priors; (iii) the 50-cluster count reflects a specific stability threshold; and (iv) evaluation of learning outcomes and the AI evaluation scaffold both remain future work.

References

R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics*. Addison-Wesley, 1964. `https://feynmanlectures.caltech.edu`.

J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon. Expert and novice performance in solving physics problems. *Science*, 208(4450):1335–1342, 1980.

J. Clement. Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50(1):66–71, 1982.

J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

D. Hestenes, M. Wells, and G. Swackhamer. Force concept inventory. *The Physics Teacher*, 30(3):141–158, 1992.

L. McInnes, J. Healy, and S. Astels. HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

E. F. Redish. *Teaching Physics with the Physics Suite*. Wiley, 2003.

J. L. Docktor and J. P. Mestre. Synthesis of discipline-based education research in physics. *Physical Review Special Topics – Physics Education Research*, 10(2):020119, 2014.

Conference on Physics and AI at Stanford University (PAI 2026) Paper Checklist

1. **Claims**

   Question: Do the main claims in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract claims (i) a 50-cluster cognitive taxonomy, (ii) a 70% portability rate under a pre-specified criterion, and (iii) an interactive application. All three are fully presented.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work?

   Answer: [Yes]

   Justification: Section 5 discusses single-corpus bias, LLM annotation priors, the approximate cluster count, and absence of learning outcome evaluation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [NA]

   Justification: The paper presents empirical findings and a deployed application. No formal theorems are stated.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose information needed to reproduce the main results?

   Answer: [Yes]

   Justification: The RAG pipeline, chunking strategy, embedding model (SPECTER2), clustering algorithm (HDBSCAN), stability criterion (Jaccard > 0.70), confidence threshold (0.80), and portability criterion (coverage across ≥ 3 of 8 subfields) are all specified. Code and taxonomy are released publicly.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code?

   Answer: [Yes]

   Justification: The annotated corpus, taxonomy JSON, extraction pipeline, and web application are released under CC BY 4.0.

6. **Experimental setting/details**

   Question: Does the paper specify all training and test details?

   Answer: [Yes]

   Justification: Section 2 specifies corpus source, chunking, annotation structure, confidence threshold, split-half validation, and HDBSCAN stability criterion.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or statistical significance?

   Answer: [NA]

   Justification: Primary results are cluster stability rates and portability counts. Split-half Jaccard agreement is the stability criterion.

8. **Experiments compute resources**

   Question: Does the paper provide sufficient information on compute resources?

   Answer: [Yes]

   Justification: The RAG pipeline runs via the Claude API with no local GPU compute required. The Feynman corpus is indexed as a vector store; estimated API cost for the full 800-passage corpus is under $15.

9. **Code of ethics**

   Question: Does the research conform with the Conference on Physics and AI at Stanford University (PAI 2026) Code of Ethics?

   Answer: [Yes]

   Justification: The work uses only publicly available text with no human subjects, sensitive data, or potential for misuse.

10. **Broader impacts**

    Question: Does the paper discuss both positive and negative societal impacts?

    Answer: [Yes]

    Justification: Positive impacts (improved physics education) and the potential negative impact of AI guidance replacing rather than scaffolding human reasoning are discussed in Section 5.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release?

    Answer: [NA]

    Justification: The released assets pose no dual-use or misuse risk.

12. **Licenses for existing assets**

    Question: Are creators of existing assets properly credited?

    Answer: [Yes]

    Justification: The Feynman Lectures, HDBSCAN, and D3.js are cited with full references.

13. **New assets**

    Question: Are new assets introduced in the paper well documented?

    Answer: [Yes]

    Justification: The PhysCog taxonomy, dataset, and application are accompanied by a full data card released alongside the paper.

14. **Crowdsourcing and research with human subjects**

    Question: Does the paper involve crowdsourcing or human subjects?

    Answer: [NA]

    Justification: No human subjects. Annotation is automated via the RAG pipeline.

15. **Institutional review board (IRB) approvals**

    Question: Does the paper describe IRB approvals?

    Answer: [NA]

    Justification: No human subjects research was conducted.

16. **Declaration of LLM usage**

    Question: Does the paper describe LLM usage as an important component of the core methods?

    Answer: [Yes]

    Justification: LLMs were used to aid in developing the front-end application of the RAG pipeline, including the interactive cluster navigation interface and the concept-guided problem walkthrough feature. Claude (Anthropic) serves as the generative backbone of the RAG system.