Enhancing Physics Problem Solving in LLM Agents with Knowledge Graphs and Symbolic Math Tools

Sviatoslav Gladkykh
MSc Graduation Project
Eindhoven University of Technology
Supervisor: Mykola Pechenizkiy
Tutor: Bram Schut

Motivation

- Large Language Models (LLMs) are strong in general conversation, but weak in physics problem solving.
- Physics requires **conceptual reasoning** + **precise math**.
- Goal: Make LLMs **transparent, accurate, efficient** and **adaptable** to other STEM fields.

Challenges

- Need benchmarks covering **conceptual + numerical physics**.
- Need **rich** and **structured well** knowledge source.
- Parse and transform knowledge source into a **Knowledge Graph**.
- Ensure **symbolic math verifier** integrates smoothly.
- Orchestrate tools well, preventing agent from hallucinations or endless loops.

Research Questions

- **RQ1**: How does retrieval from a structured knowledge graph compare to vector-based retrieval for physics question answering?
- **RQ2**: What is the relative and combined impact of using knowledge retrieval and mathematical tools on the accuracy and reliability in multi-step derivations while ensuring physical consistency?
- **RQ3**: How does the proposed physics-informed LLM agent perform against general-purpose LLMs, physics-fine-tuned models, and state-of-the-art methods under the same evaluation setup?

Research Sub-questions

- Which KG construction and retrieval approaches perform the best in physics problem solving?
- Which symbolic and (or) numeric tools perform the best in physics problem solving?
- Which types of questions / topics in physics benefit most from KG retrieval and symbolic verification?
- How can retrieved subgraphs be linked back to the source textbook passages to enable citation, fact checking, and hallucination mitigation?
- … (as needed or if time permits)

Time-permitting research questions

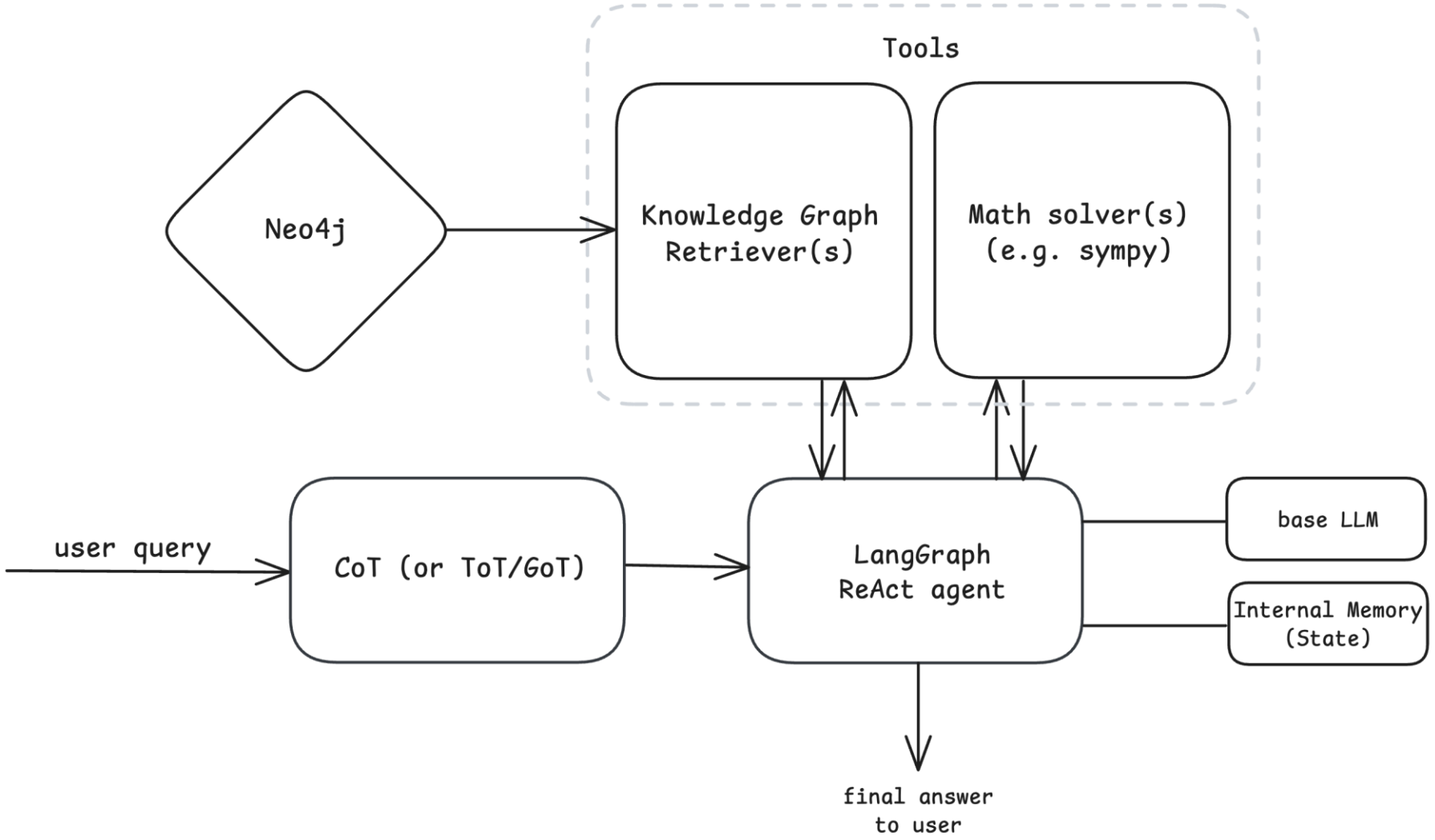
- Can an auxiliary corpus of solved Q&A be leveraged for few-shot guidance or case-based retrieval to further improve the agent's task execution?
- Can the agent be extended to multimodal physics tasks (e.g., diagrams, plots), and what uplift results on multimodal benchmarks relative to generic and physics-tuned multimodal LLMs?
- Can the approach scale to master's/PhD/olympiad-level problems with an appropriately enriched knowledge base, and how does it compare with existing solutions on very challenging benchmarks?
- Can verifiable rewards be defined for each reasoning step of the agent, enabling reinforcement learning to further improve performance and consistency?
- … (if time permits and interesting questions arise)

Research plan & timeline

Table 1: Planned timeline, milestones, and deliverables for the MSc Graduation project.

Period	Description	Deliverables
Sep 2025 – Oct 2025	**Part I: Preparation and baseline design.** Finalize benchmarks and knowledge bases, set up evaluation pipeline, select base models and frameworks, define initial topics.	Baseline design plan, documentation of choices.
Oct 2025 – Dec 2025	**Part II: Baseline implementation and evaluation.** Parse KB into KG and vector store, implement baseline agent with dual retrieval and symbolic verification, run first benchmarks.	Baseline agent code, evaluation report.
Dec 2025 – Feb 2026	**Part III: System extension and comparative setups.** Expand KB, add alternative tools and retrieval setups, benchmark multiple configurations.	Extended system code, comparative evaluation report.
Feb 2026 – May 2026	**Part IV: Analysis, extensions, thesis writing.** Error analysis, topic-specific results, tool trace study, optional extensions, thesis writing.	Final analysis report, thesis draft.
May 2026 – Jun 2026	**Defense.** Finalize thesis, prepare and deliver defense presentation.	Defended MSc thesis.

System Design



Knowledge bases

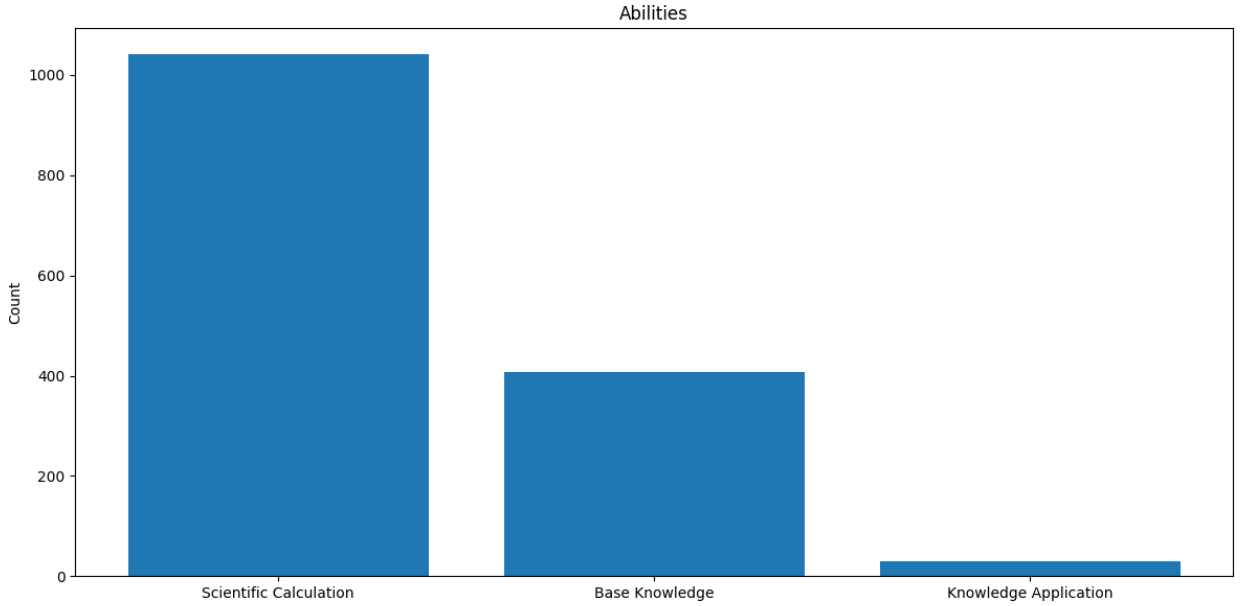
Source	Accessibility Format	Coverage & Depth	Strengths	Limitations
Light and Matter (Crowell) [5]	PDF, HTML, LaTeX source	Introductory physics (algebra-based and calculus-based)	Extensive modular series, free and open-source	Requires preprocessing to extract structured entities, relations, and formulas suitable for KG integration
CK-12 FlexBooks; The People's Physics Book [4]	Web-based modular textbooks, API (CK-12)	High-school to early-college topics	Free, includes examples and practice problems, flexible access (API for CK-12)	Limited mathematical rigor, inconsistent across topics, web-only format for CK-12
Motion Mountain (Schiller) [20]	PDF (narrative style)	Broad conceptual coverage (classical and modern physics)	Engaging, whimsical approach, wide topic breadth	Not structured, PDF-only, licensing restricts reuse and modification
OpenStax University Physics [15]	XML (GitHub repository)	Full calculus-based three-volume physics (mechanics to modern physics)	Structured XML format, widely adopted, rigorous mathematical exposition, rich worked examples	Requires parsing pipeline and ontology design for KG construction
Wikipedia (including Wikidata, DBpedia, YAGO)	HTML + structured APIs/SPARQL endpoints	Broad, evolving coverage of physics concepts	Large-scale, cross-linked, regularly updated knowledge graph sources; suitable for supplementary enrichment (e.g., Wikidata has 1.65 billion statements)	Not pedagogically structured; variable quality and detail; requires careful curation

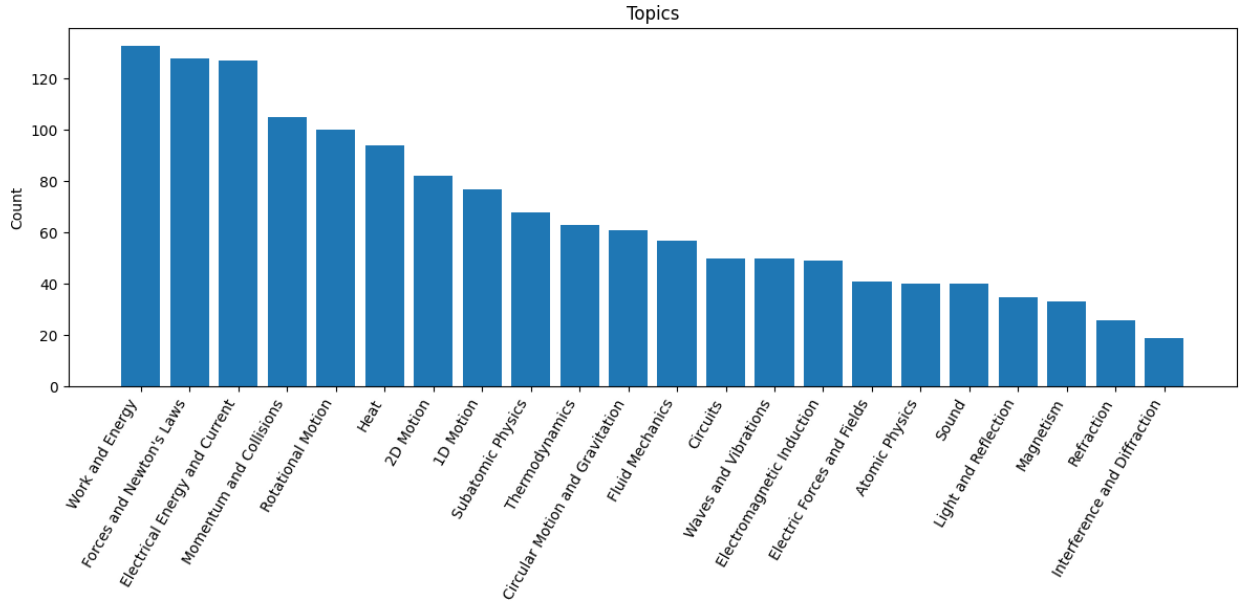
Benchmarks

Benchmark	Multimod.	Knowledge Level	Question Type	#Physics Questions	Topic present	Decision
SciEval [22]	Text only	high school–undergraduate	multiple-choice	1478	Yes	Accepted (for text-only)
SCIMAT-2 [14]	Text only	pre-college–college	open (numeric/formula)	∼1041 types	Yes	Accepted (for text-only)
MMLU [11]	Text only	high school–college	multiple-choice	102–235 (subsets)	No	Accepted (for text-only)
PHYSICS [8]	Combined	university-level	open (numeric/formula)	1297	Yes	Accepted (for multi-modality)
SciBench [23]	Combined	college	open (numeric/formula)	– (789 total)	Yes	Rejected (few topics, not split by subject)
MMMU [29]	Images only	college	multiple-choice	408	Yes	Accepted (for multi-modality)
MMMU-Pro [30]	Images only	college	multiple-choice	60	No	Rejected (too few questions)
ScienceQA [16]	Combined	elementary–middle school	multiple-choice	1923	Yes	Rejected (too simple)
Olympiad-Bench [9]	Images only	olympiad	open ended	456	Yes	Accepted (for multi-modality and olympiad-level)
JEEBench [2]	Text only	pre-engineering college	mixed types	123	Yes	Rejected (too few questions)
GPQA [19]	Text only	domain experts	multiple-choice	187	Yes	Rejected (too few questions)
SciQ [13]	Text only	middle–high school	open ended	– (13.7K total)	No	Rejected (open ended)

SciEval

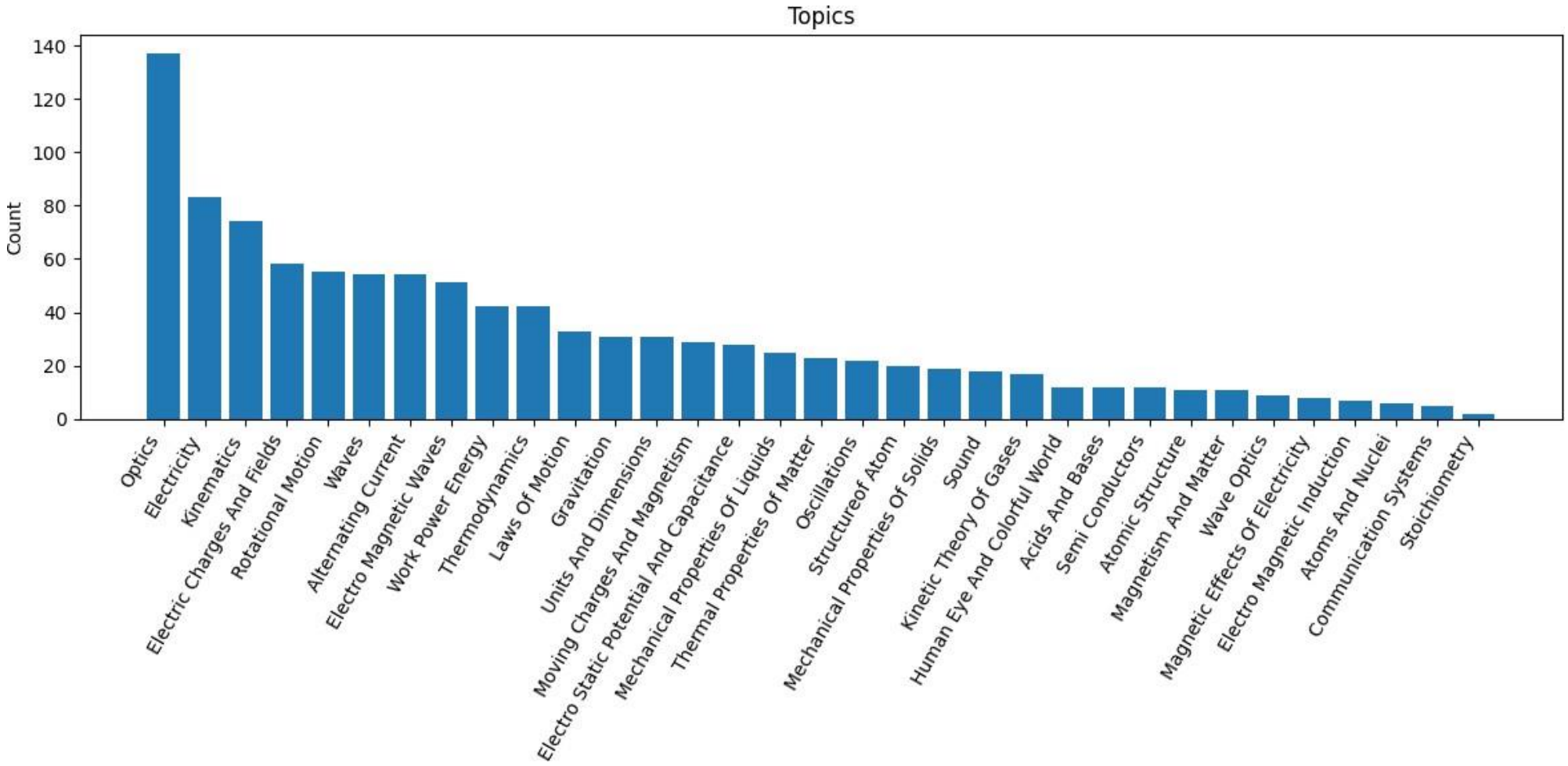
Upper high school – early undergraduate physics
1,478 multiple-choice questions in “physics” category





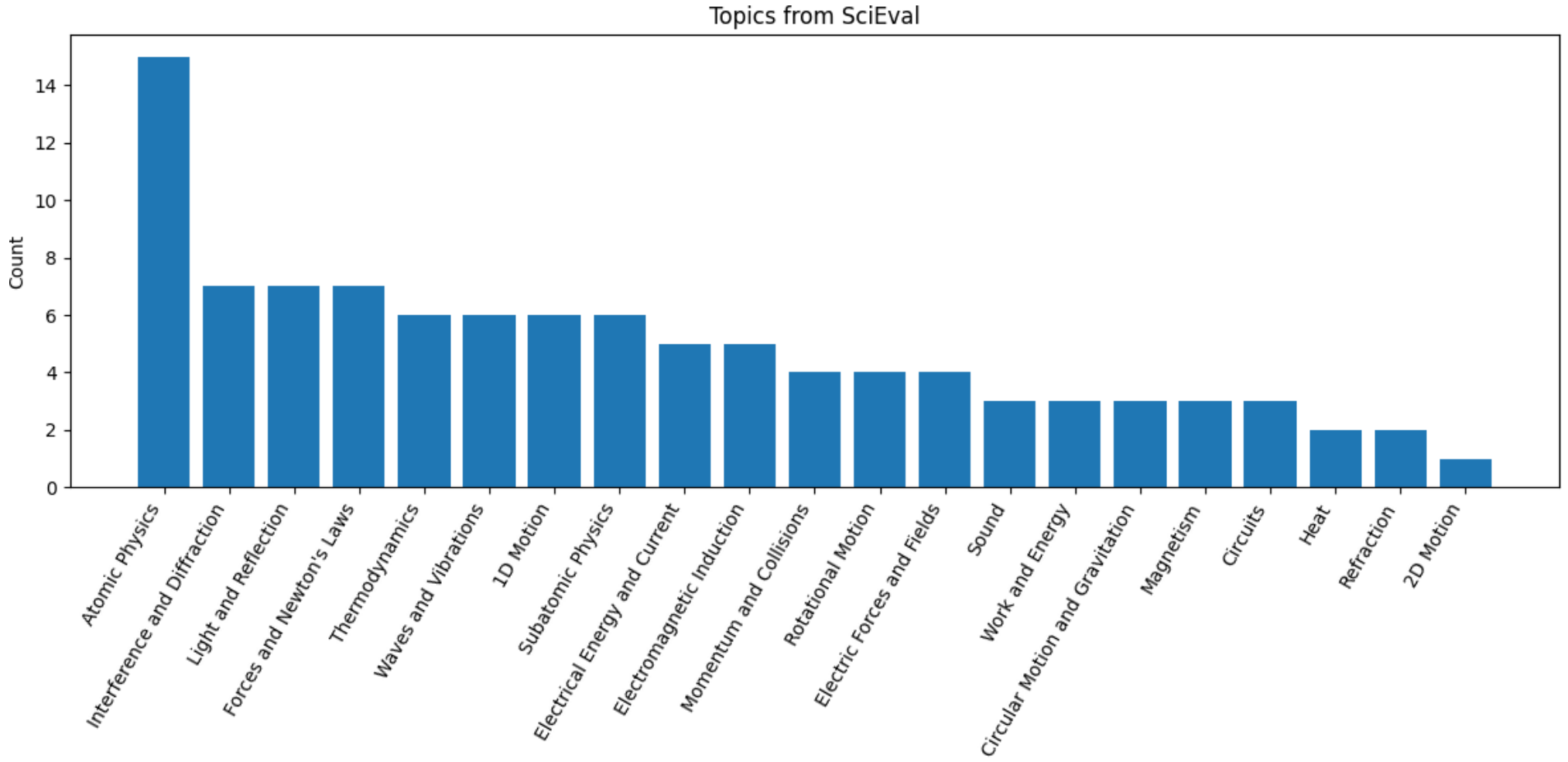
SCIMAT-2

Upper high school – early undergraduate physics
~1,041 numeric questions in “science” category



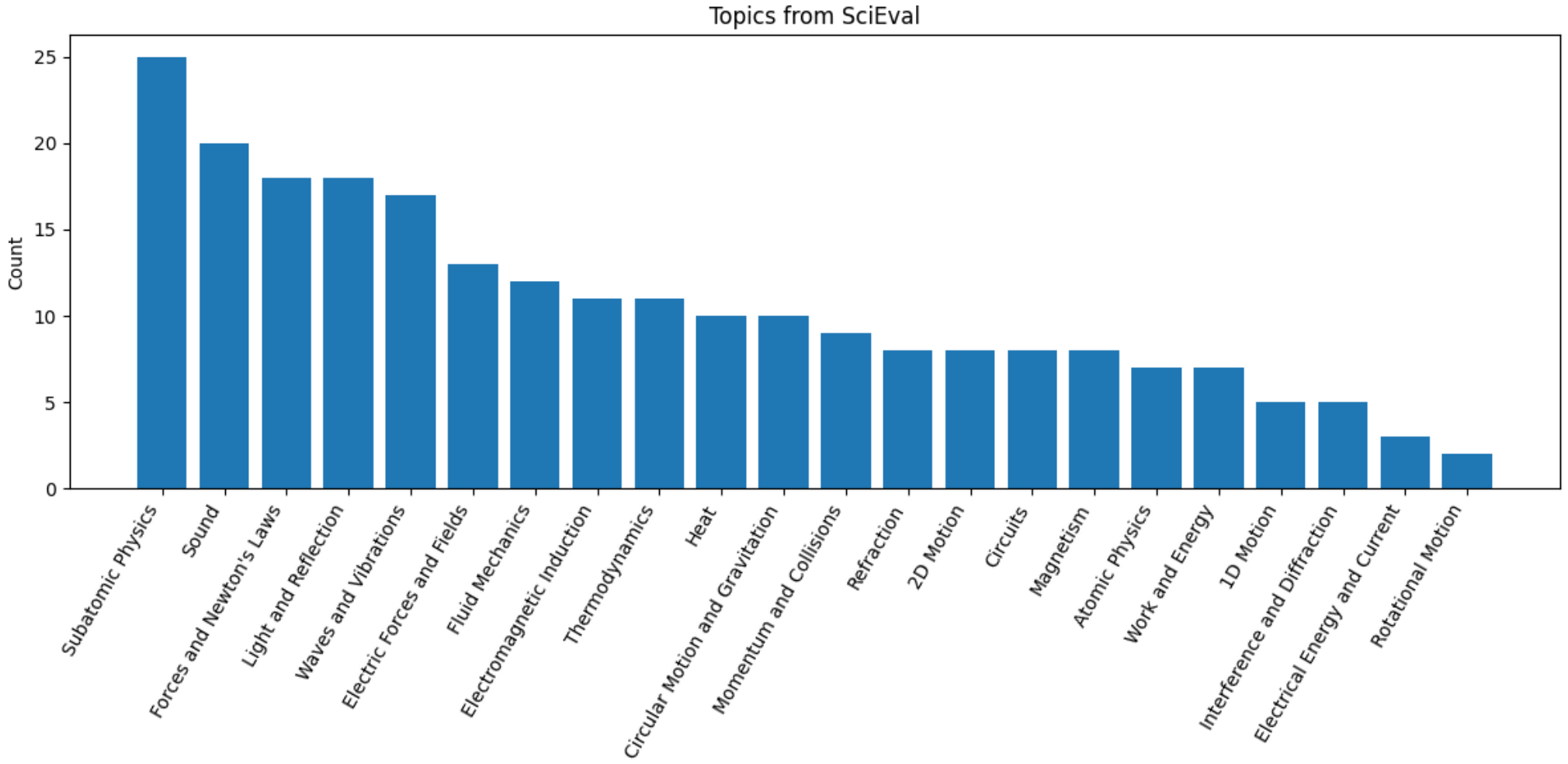
MMLU (college physics)

102 multiple-choice questions
topics not specified, we use a logistic classifier trained on question embeddings from SciEval dataset



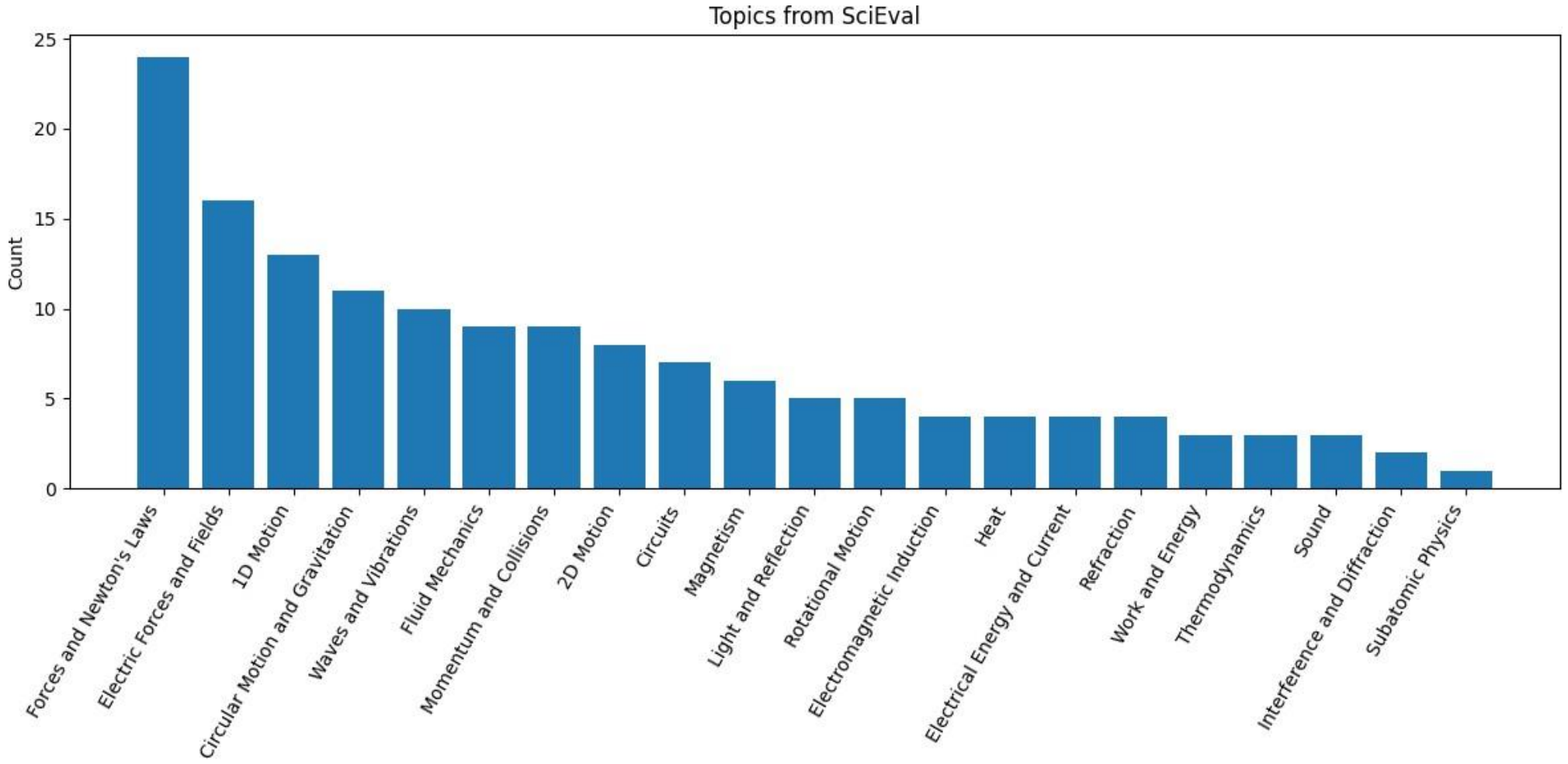
MMLU (conceptual physics)

235 multiple-choice questions
topics not specified, we use a logistic classifier trained on question embeddings from SciEval dataset



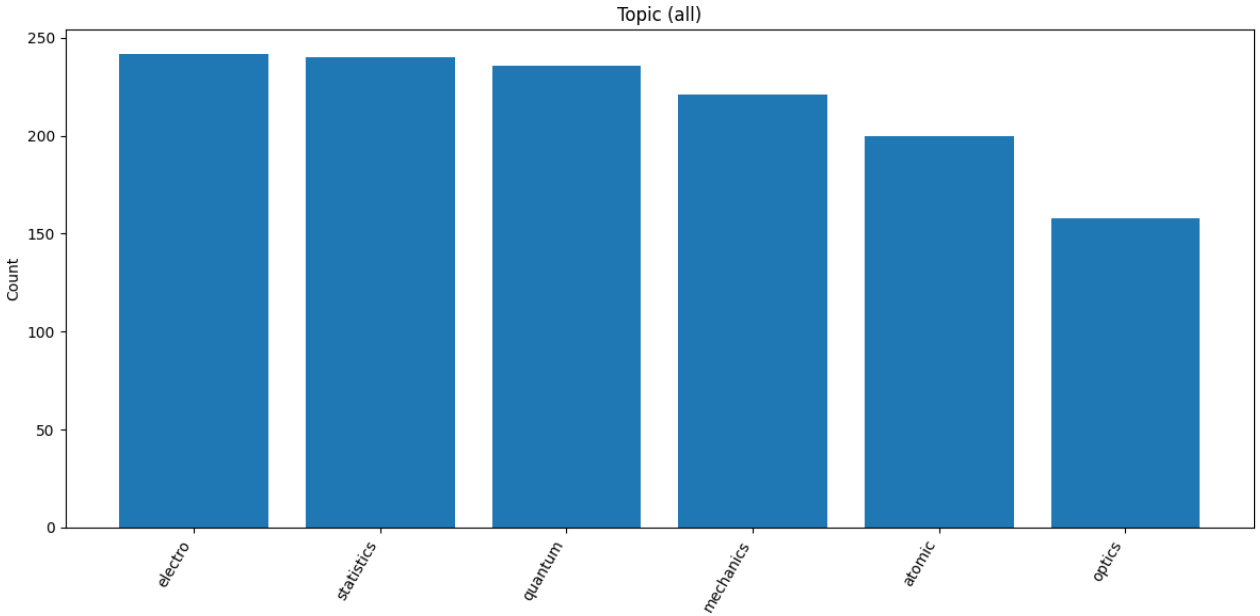
MMLU (high school physics)

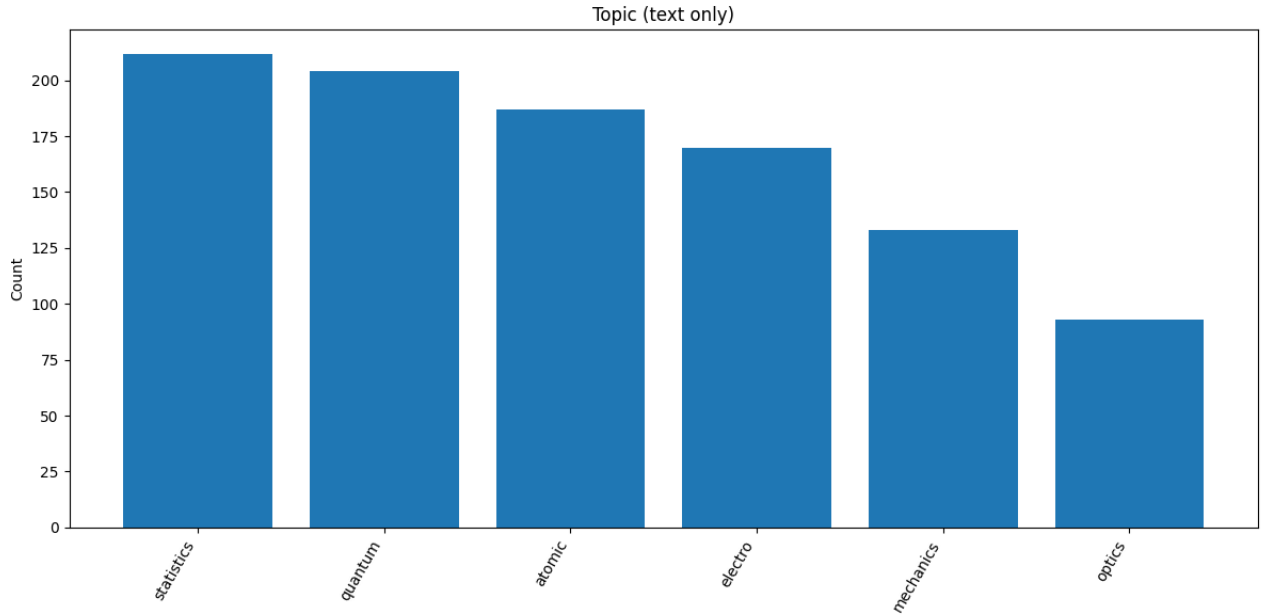
151 multiple-choice questions
topics not specified, we use a logistic classifier trained on question embeddings from SciEval dataset



PHYSICS

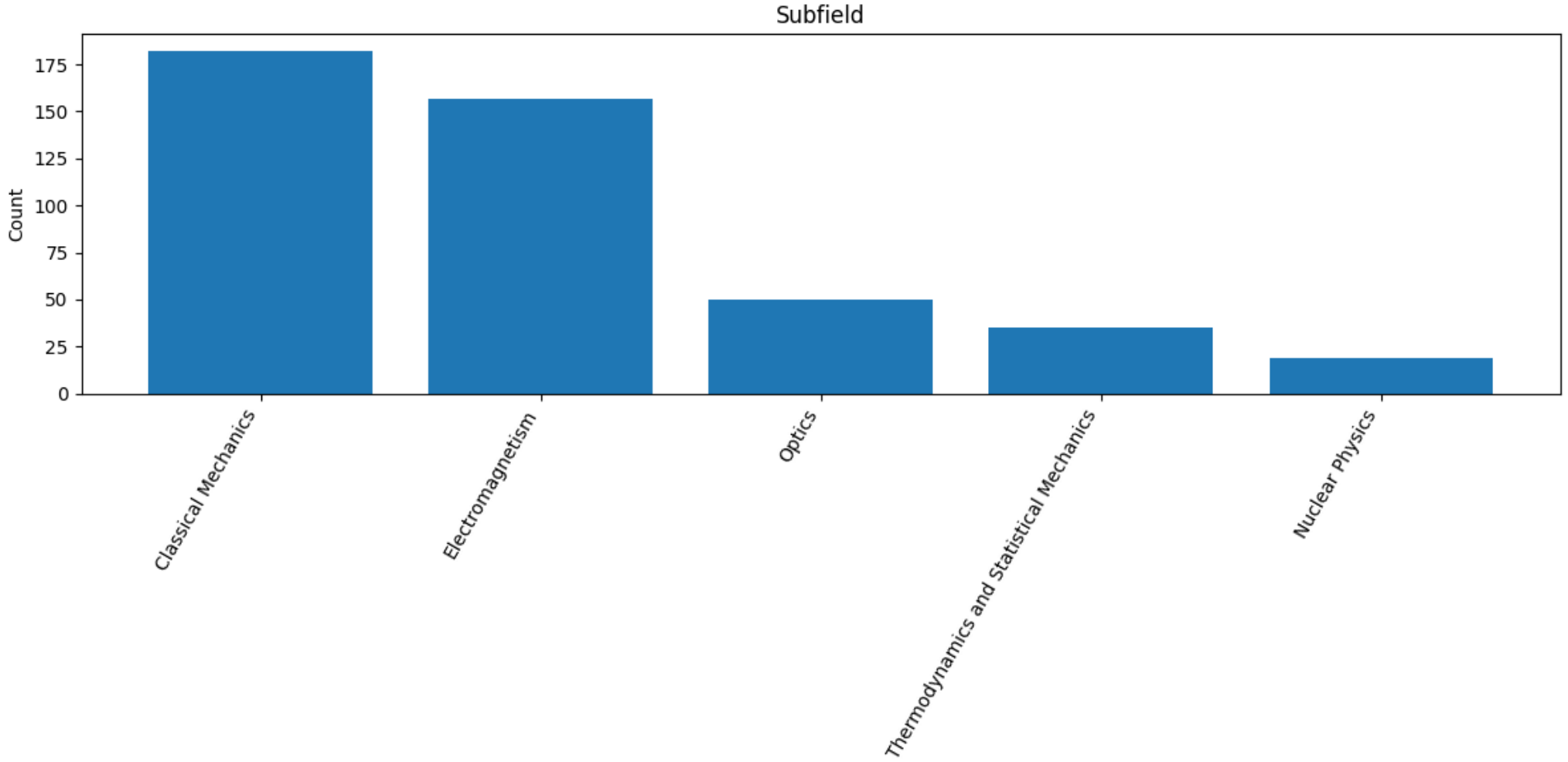
University-level
1297 open (numeric formula) questions (999 without images)





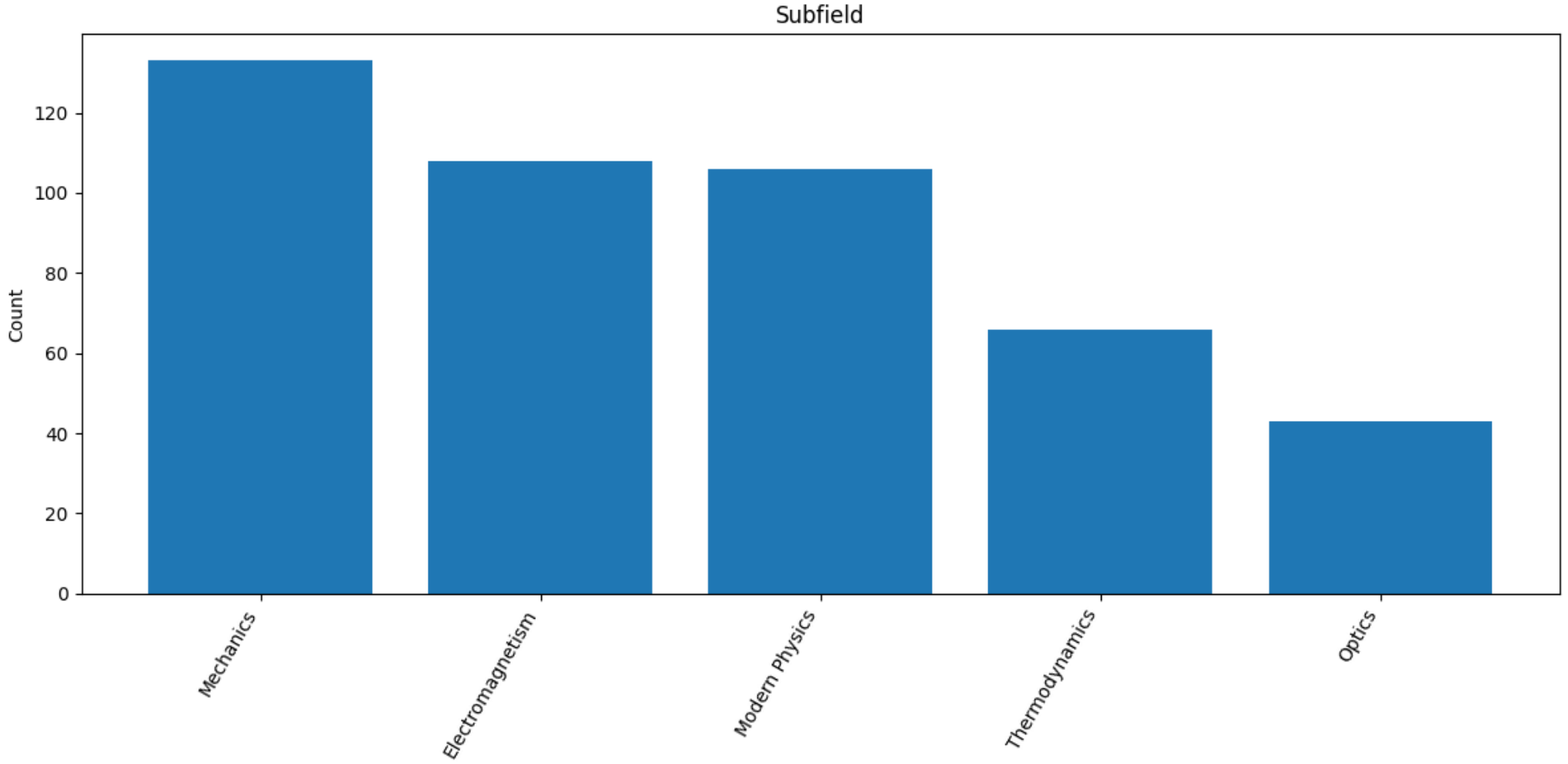
MMMU

College-level
408 multiple-choice physics questions, all with images



Olympiad-Bench

Olympiad-level
456 multiple-choice physics questions, all with images



Expected contributions

- Resource-efficient adaptation of LLMs to STEM domains.
- Transparent reasoning traces (facts + algebra steps).
- Open-source code, KG, and evaluation artifacts.
- Potential applications in education, tutoring, scientific reasoning.

Risks & Mitigation

- Computational resources – API fallback.
- Parsing difficulties – LLM-assisted extraction, additional resources
- Knowledge Graph may underperform – hybrid or vector fallback
- Insufficient accuracy gains – investigation, improvements, focus on transparency & traceability

Conclusion

- Novel agent combining **Knowledge Graphs** + **Symbolic Math Tools**.
- Systematic evaluation on physics benchmarks.
- Aim: **accurate, explainable, and efficient** physics-aware LLM assistant.