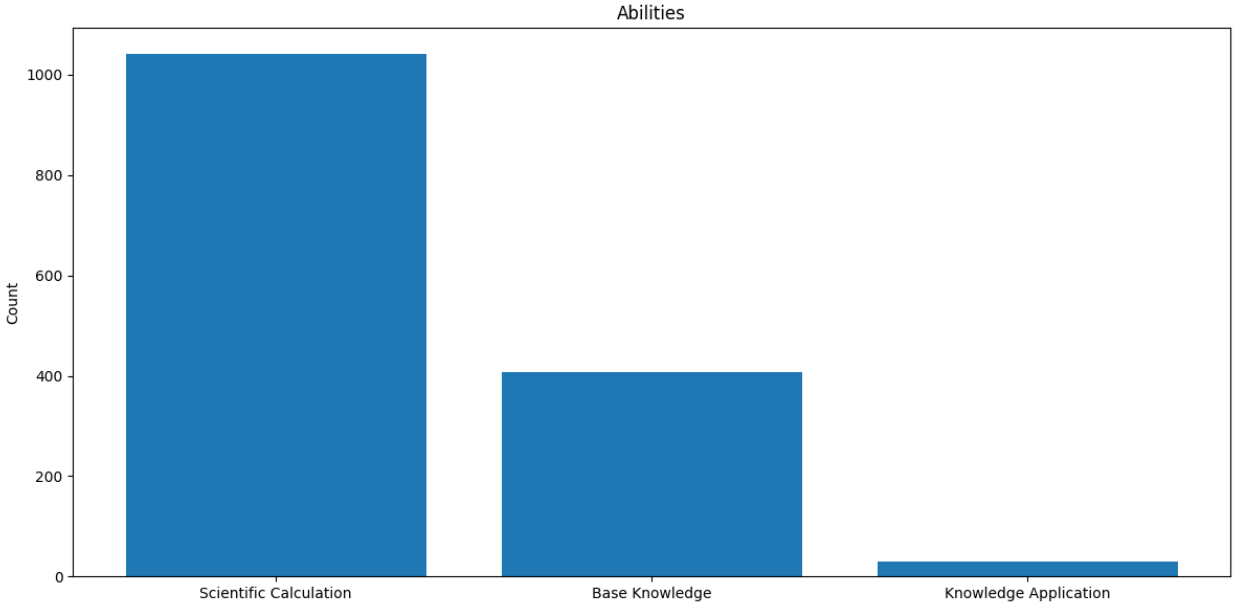
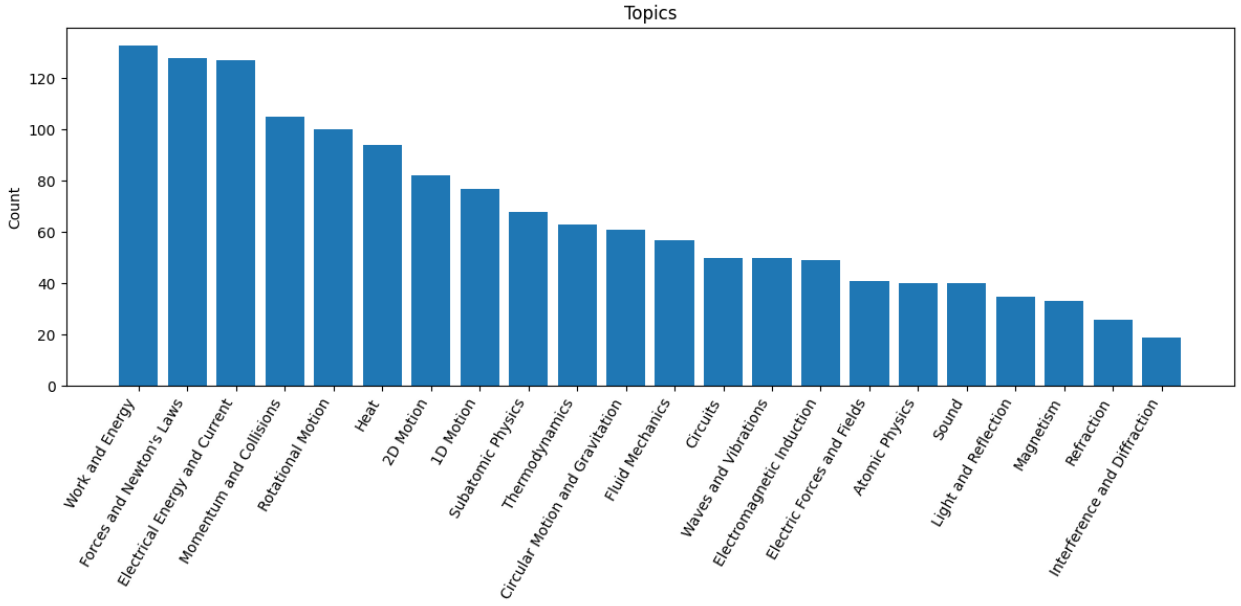
Benchmarks

SciEval

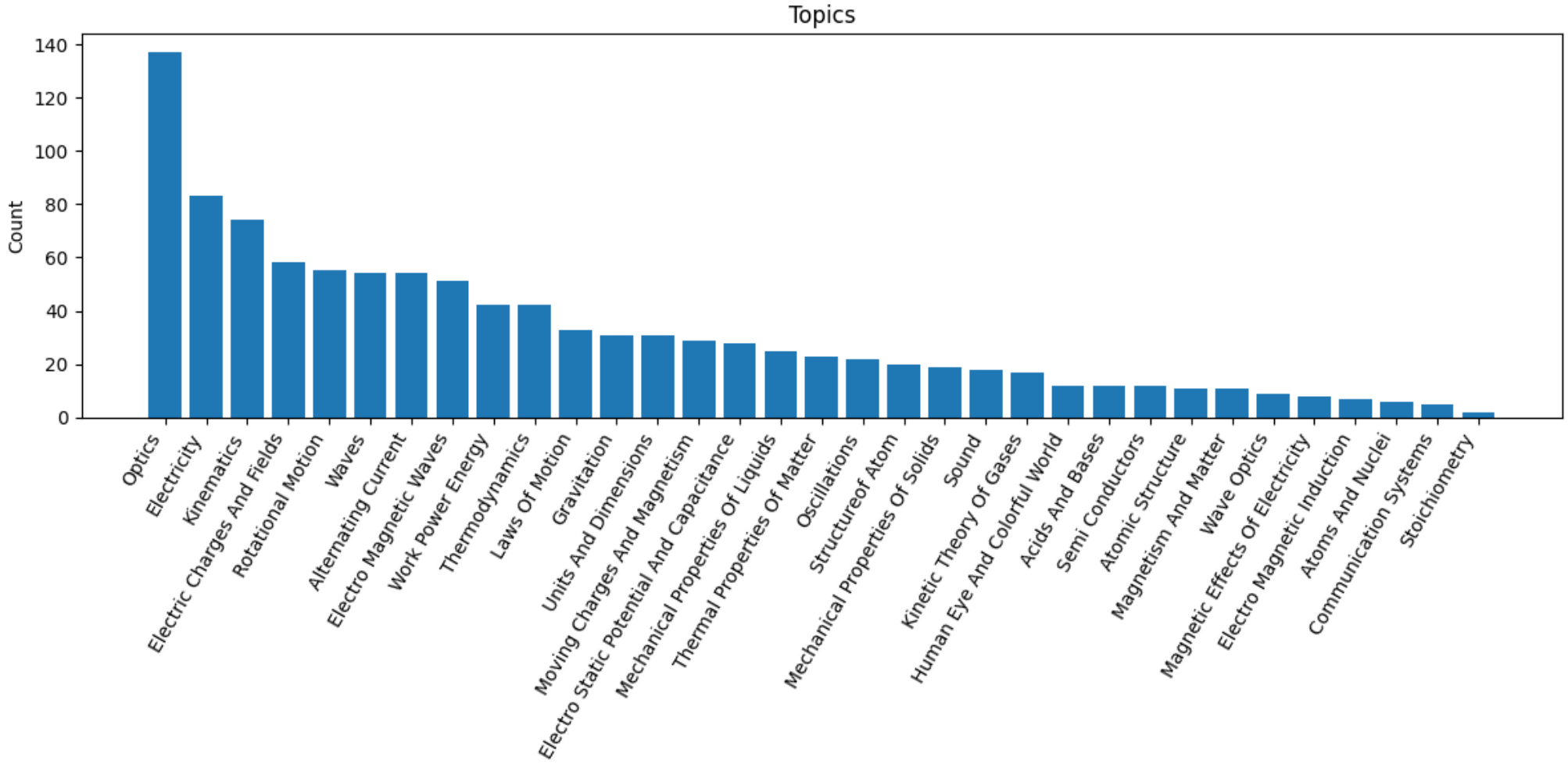
Upper high school – early undergraduate physics
1,478 multiple-choice questions in “physics” category





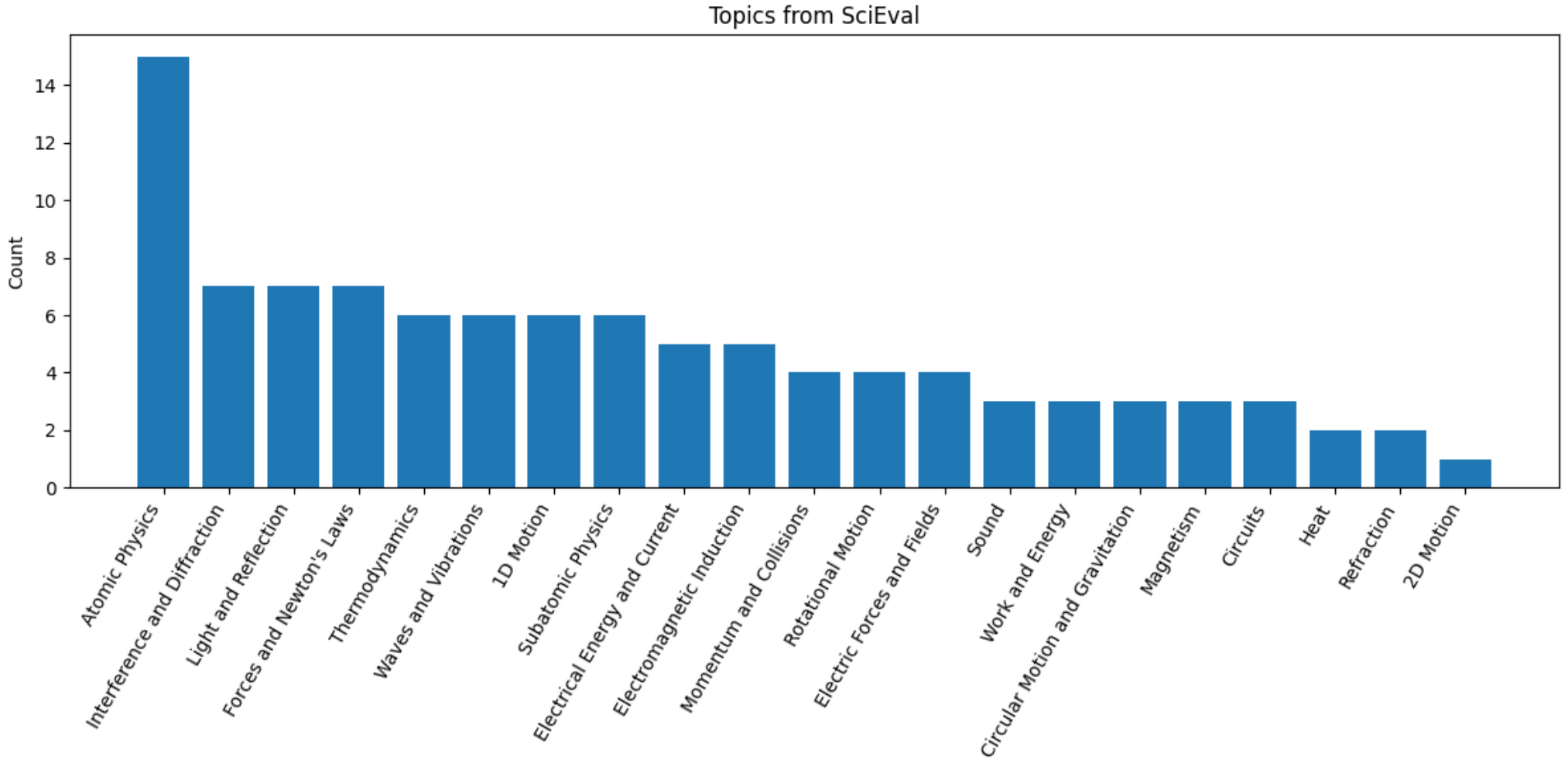
SCIMAT-2

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



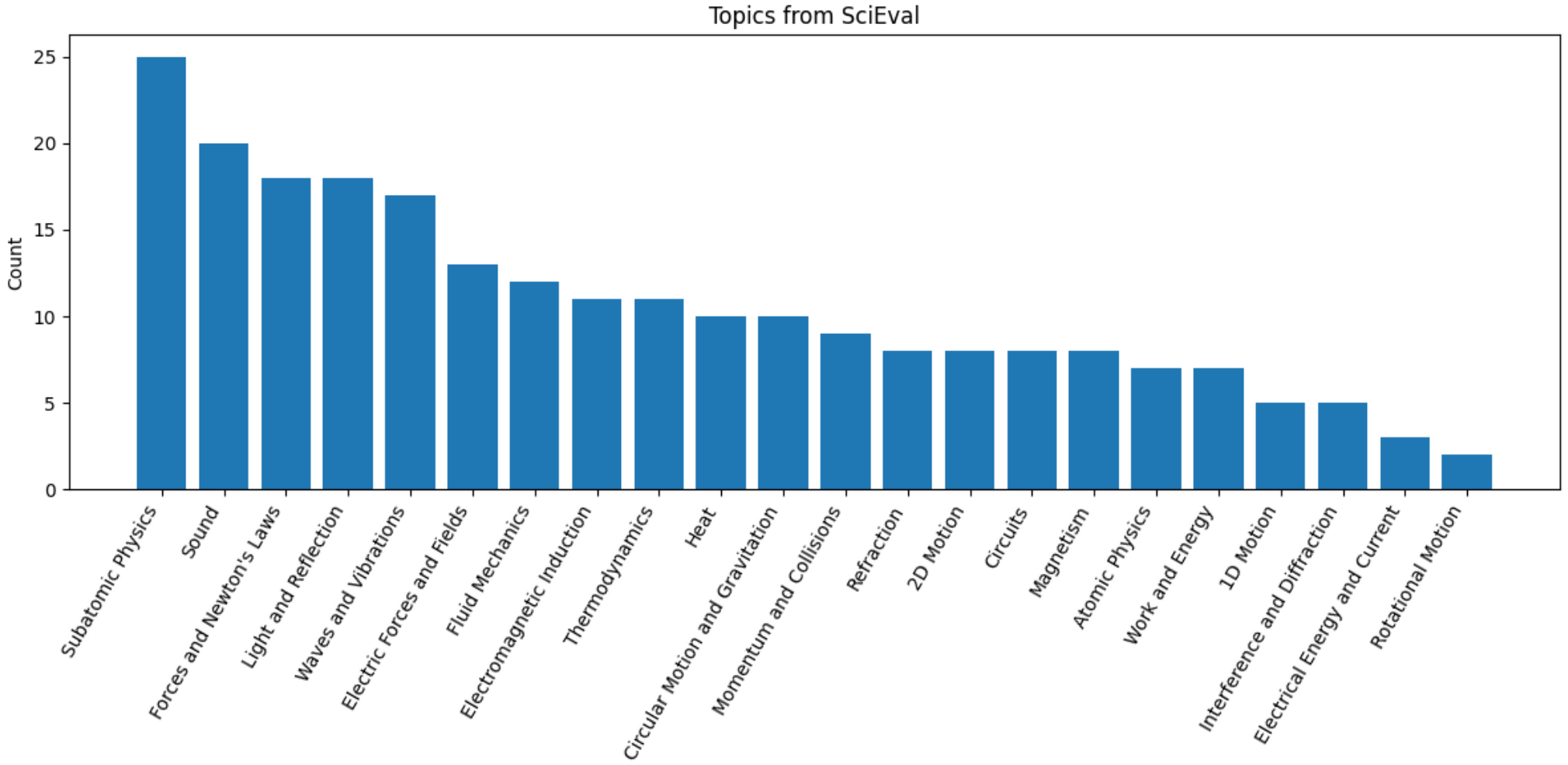
MMLU (college physics)

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



MMLU (conceptual physics)

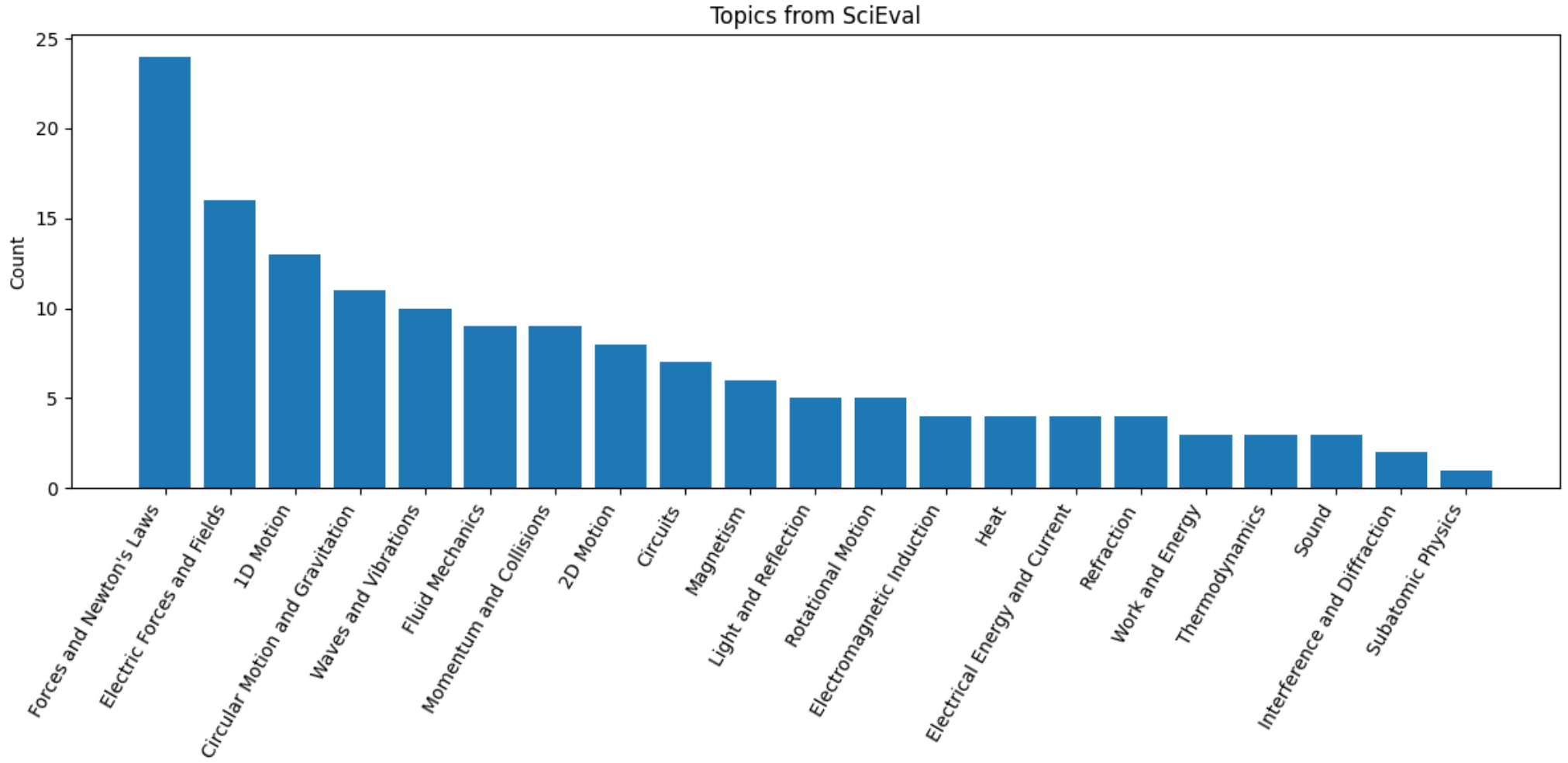
235 multiple-choice questions
topics not specified, we use a logistic classifier trained on question embeddings from SciEval dataset



MMLU (high school physics)

151 multiple-choice questions

topics not specified, we use a logistic classifier trained on question embeddings from SciEval dataset



Benchmarks - Conclusions

- MMLU is not suitable for benchmarking limited number of topics
- Potentially good starting points for the topics:
  - Electrical Energy and Current (SciEval) + Electricity (SCIMAT)
  - Rotational Motion (both SciEval and SCIMAT)
  - Waves and Vibrations (SciEval) + Waves (SCIMAT)
- For all benchmarks OpenStax University Physics book seems to be sufficient in terms of conceptual knowledge