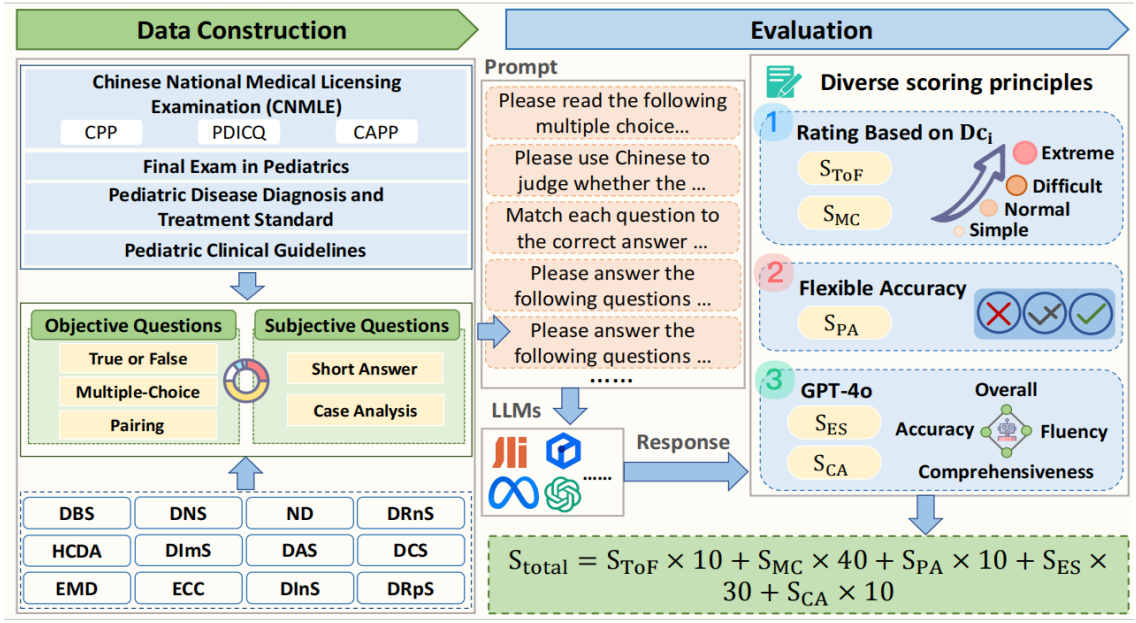
PediaBench: a comprehensive Chinese pediatric dataset for benchmarking large language models

Qian ZHANG, Panfeng CHEN, Linkun FENG, Shuyu LIU, Jiali LI, Heng ZHAO, Mei CHEN, Hui LI, Yanhao WANG

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41345-w

Problems & Ideas

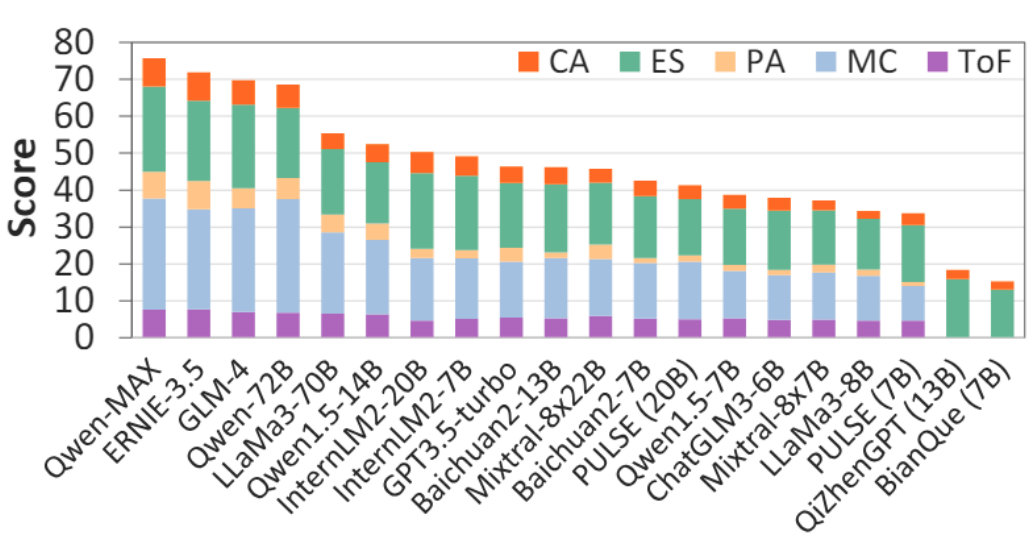
- Problems of mediacal QA benchmark:
  - Medical benchmarks are limited to objective questions with standardized closed-ended answers.
  - Existing datasets for medical QA cannot comprehensively assess the proficiency of LLMs in pediatrics.
- Ideas: A Chinese pediatric dataset encompassing 5 question types and 12 disease groups, and devise an integrated scoring criterion to assess the proficiency of LLMs in pediatrics.

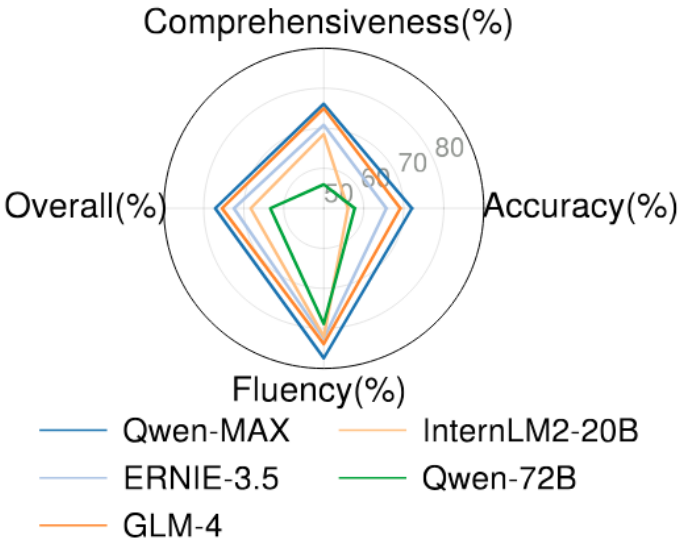


The dataset construction and evaluation process.

Main Contributions

- Contributions:
  - Introduced PediaBench, a high-quality QA dataset specific to pediatrics in the Chinese context ;
  - Devised an integrated scoring scheme to measure the QA performance of each LLM on PediaBench ;
  - Evaluated PediaBench with 20 LLMs, including open-source and commercial general purpose models of different scales and specialized models in the medical domain.





The benchmark results. Left: Results of different LLMs for the scores for five question types and the total scores on PediaBench; Right: Four-dimensional scores of all LLMs as evaluated by GPT-4o for ES questions.