UNIVERSITY OF CAMBRIDGE

Session 5

Learning the capabilities of large language models

Marko Tešić

LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

Vancouver, 20 February, 2024

Goals

- ❑ Understand the key challenges in budling measurement layouts for learning the capabilities large language models from benchmark datasets
- ❑ Run through a worked example of a measurement layout for Addition dataset
- ❑ Building on the worked example create a measurement layout for the Arithmetic dataset
- ❑ The session should provide you with an understanding of:
    - ❑ Benchmark characteristics needed to build measurement layouts
    - ❑ The relationship between abilities tested in the benchmarks and meta-features/demands that describe the benchmark instances
    - ❑ Assessing the performance of measurement layouts against baselines
    - ❑ How is building measurement layouts and predicting performance on tasks different from predicting performance from assessors

Key challenges in building measurement layouts for large language models

- ❑ Instance level data needed
  - ❑ Find benchmarks that share the performance of large language models at instance level
  - ❑ Some examples include: HELM and BIG-bench

- ❑ Benchmark meta-features or demands:
  - ❑ Many benchmarks don't include or don't have the meta-features that describe the benchmarks, how they are built, which were considered when building benchmarks
  - ❑ We also need levels of those meta-features (possible to potentially automate with a rubric and the use of language models like GPT-4)

Key challenges in building measurement layouts for large language models

❑ What capabilities are tested by those benchmarks?

❑ Are those capabilities characterizable and have some backing in cognitive science (or some other sciences)?

❑ How do these capabilities interact? Does having one capability compensate for not having some other capability test on the benchmark?

Addition dataset

- ❑ A simple benchmark testing LLMs abilities to add two numbers
- ❑ 10 LLMs tested
- ❑ We will see how to create some of the meta-features for this benchmark and how to relate them to capabilities learnt by the measurement layout
- ❑ We will also create some assessors and baselines to compare to the performance of measurement layouts

Arithmetic dataset

- A benchmark testing LLMs abilities to perform basing arithmetic operations: addition, subtraction, multiplication, and division.

- 8 LLMs tested

- We will see how to create dependances and hierarchy between the abilities