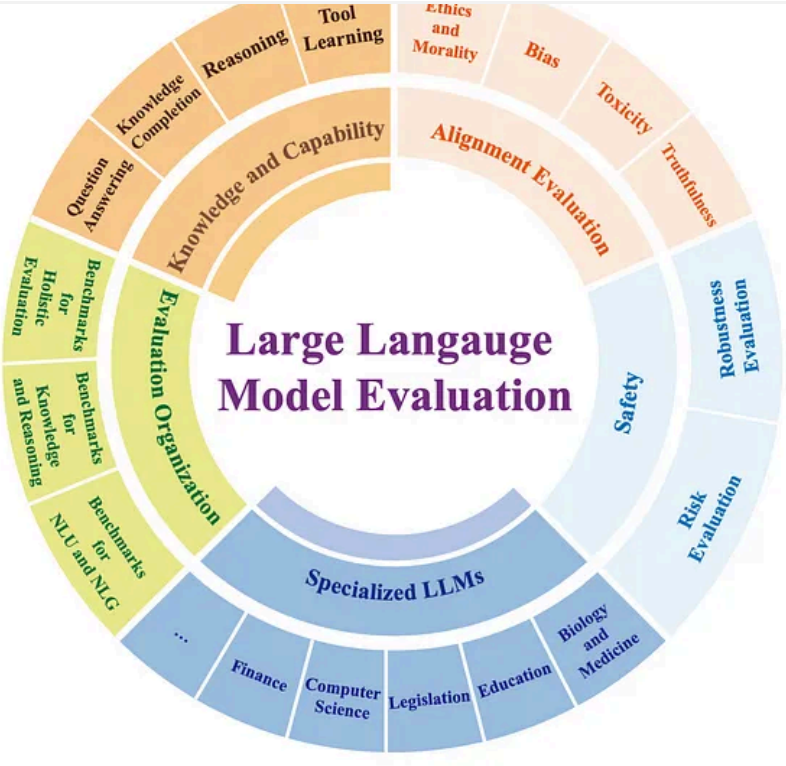


Source

How Should Large Language Models Be Evaluated?

The image shown is the taxonomy of major categories and sub-categories used by the study for LLM evaluation.

Cobus Greyling · Follow
3 min read · Nov 1, 2023

Literary two days ago *Tianjin University* released a study, which defined a LLM model evaluation and implementation taxonomy.

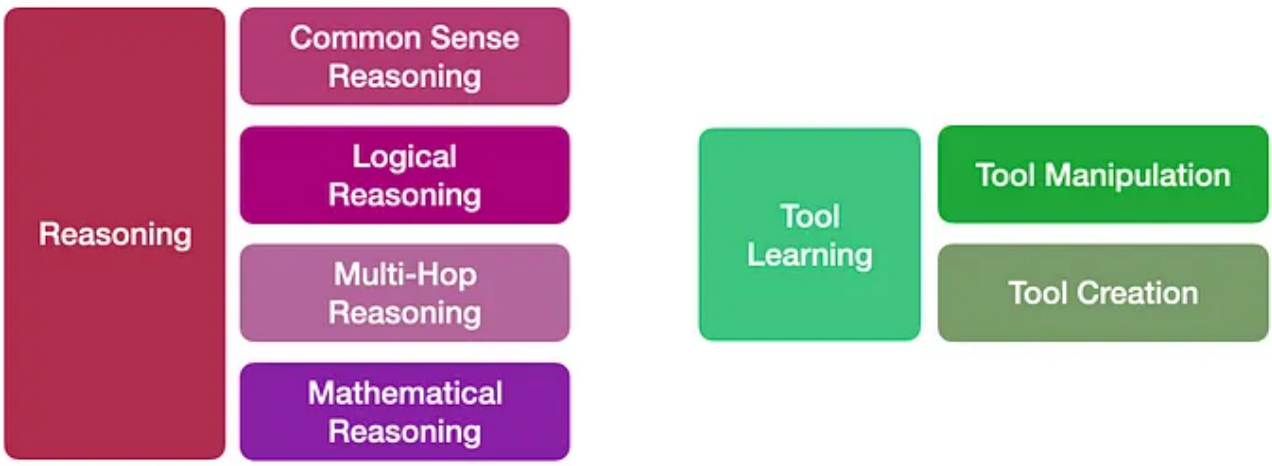
Considering the taxonomy, much focus has been on **Evaluation Organisation**, **Knowledge & Capability** and **Specialised LLMs** from a technology perspective.

From an enterprise perspective, the concerns and questions raised centre around **Alignment** and **Safety**.

This survey aims to create an extensive perspective on how LLMs should be evaluated. The study categorises the evaluation of LLMs into three major groups:

1. Knowledge and Capability Evaluation,
2. Alignment Evaluation and
3. Safety Evaluation.

What I also found interesting from the study, was the focus on more complex metrics like reasoning and tool learning, as seen below:



In the early days of Natural Language Processing (NLP), researchers frequently utilised a series of simple benchmark assessments to assess the performance of their language models. These initial assessments predominantly focused on elements such as syntax and vocabulary,

including tasks like parsing syntactic structures, disambiguating word senses, and more.

LLMs have introduced significant complexity and have shown certain tendencies by revealing behaviours indicative of risks and demonstrating abilities to perform higher-order tasks in current evaluations.

Consequently, creating a taxonomy like this to reference can help ensure that due diligence is followed when assessing LLMs.

⭐ ***Follow me on LinkedIn for updates on Conversational AI*** ⭐

I'm currently the Chief Evangelist @ Kore AI. I explore & write about all things at the intersection of AI and language; ranging from LLMs, Chatbots, Voicebots, Development Frameworks, Data-Centric latent spaces & more.

LinkedIn

Get an email whenever Cobus Greyling publishes.

Get an email whenever Cobus Greyling publishes. By signing up, you will create a Medium account if you don't already...

cobusgreyling.medium.com

Evaluating Large Language Models: A Comprehensive Survey

Large language models (LLMs) have demonstrated remarkable capabilities across a broad spectrum of tasks. They have...

arxiv.org

GitHub - tjunlp-lab/Awesome-LLMs-Evaluation-Papers: The papers are organized according to our...