Statistical Nature of LLMs

- LLM is a neural network with hundreds of millions to billions parameters that is pertained on a volume of text no human being can ever consume in a lifetime. The goal of the training is to approximate the data distribution
  - LLM is a function that takes an input string and returns the most probable output string
  - LLMs are very sensitive, wording can alter the output [3]
  - LLMs are not continuously updating and cannot produce new knowledge [3] (?)
  - => LLM is a complex statistical model of how the words and phrases in its training data correlate [1]
- LLMs have no biological properties of humans
  - LLMs have no (innate) motivation or goal, prior experience etc.
  - LLMs have no embodiment, which is argued to be one of the prerequisites for intelligence; they cannot see and touch objects, cannot move etc.
  - => They lack reference, a key for inferring the meaning (is it?)
  - => LLMs only infer the form of the language but not the meaning [1]

But what about their performance on the tests for Natural Language Understanding (NLU)?

Statistical Nature of LLMs

- NLU benchmarks do not always require actual understanding [1]
  - LLMs utilize *shortcut learning*—a phenomenon that describes learning that relies on leveraging low-level co-occurrence patterns in the data (overlapping tokens and such)
  - LLMs use shortcuts in order to perform well on a particular benchmark [1, 5]
  - It is often the case that as soon as the obvious patterns are removed, the model's performance drops to chance levels [6]
- => LLMs can be "right for the wrong reason" [6]

=> Opinion: LLMs are *stochastic parrots* [5]