Theorem 1. *[Theoretical Limitations of Large Language Models] Let $\mathcal{M}$ be any large language model trained on a finite corpus with finite computational resources, producing outputs Y in response to inputs X according to a conditional probability distribution $P_{\mathcal{M}}(Y|X)$. Then the following limitations hold:*

1. *$\mathcal{M}$ cannot guarantee the generation of factually correct or logically valid statements for arbitrary input X.*

2. *$\mathcal{M}$ cannot exceed the information or concepts present in its training data, nor can it create fundamentally new knowledge without external input.*

3. *$\mathcal{M}$ may produce outputs exhibiting bias or error if such patterns are present in its data or training procedure.*

4. *$\mathcal{M}$ lacks self-awareness, consciousness, or genuine understanding, and does not possess intent.*

Proof. These statements follow from the design of $\mathcal{M}$ as a statistical pattern-matching system, which:

- Relies on approximating $P(Y|X)$ as learned from finite samples, so cannot generalize perfectly or reason infallibly.

- Is limited to correlations learnable from its dataset and the knowledge distribution within that data.

- Is subject to inheriting limitations, errors, and biases found in the training data and the algorithms used.

- Is not designed with the faculties of sentient beings, so cannot have attributes such as understanding or intent. □