Hallucinations in LLM

generated content ⇒ Non sensical or unfaithful to the provided source content

- Intrinsic Hallucination
- Extrinsic Hallucination

Survey :

- factuality Hallucination
- faithfulness Hallucination

generated content & verifiable Real world facts typically manifesting a factual inconsistency or fabrication

- factual inconsistency
- factual fabrication

divergence of generated content from user's information.

consistency

- Instruction
- Context
- Logical

Understand what LLM is all about :

LLM → Series of transformer Based Model

By scaling the size of the Model Capacity of training data

LLM

- In Context Learning [ICL]
- Chain of Thought Reasoning
- Instruction following

Training Stages of LLM:

1. Pre training
2. Supervised fine tuning
3. RLHF(Reinforcement learning with Human feed back)

1. PRETRAINING: — MORE IMPORTANT
   [acquire knowledge & skills.]

Through Step 1: LLM aims to predict the next word auto regressively

knowledge:

1. Language syntax
2. World knowledge
3. Reasoning abilities and

providing robust foundation for subsequent fine tuning tasks.

Essence of PRETRAINING ⟹ predicts the probability of the next word in the sequence.

2. SUPERVISED FINE TUNING:

[PRETRAINING] primarily optimizes for Completion.

PRETRAINED LLM ⟹ Completion Machines.

LEADS to MISALIGN between the next word prediction objective of the LLM and the user's Objective of obtaining Desired Responses

TO BRIDGE this GAP SFT Introduced.

which involves further training LLM's using a meticulously annotated set of [INSTRUCTION, RESPONSE] pair.

③ RLHF

human preferences ⟺ LLM

There are lots of Methods But RLHF Stands out as an Institute solution.

RLHF → [PREFERENCE MODEL] trained to predict preference rankings given [PROMPT + PAIR OF HUMAN LABELED RESPONSES]

RLHF optimize LLM → to generate output

PREFERENCE MODEL maximize the Reward provided by the

typically employ PPO Proximal policy Optimization