Hallucination:

Factual Mirage: Factual mirage (FM) is defined as the phenomenon wherein an LLM engages in hallucination or distortion of a given prompt that is factually correct. FM can potentially be subdivided into two distinct sub-categories.

Intrinsic factual mirage (IFM):

Prompt: Capital of France
AI-generated text: ...Paris is also the world fashion capital…
Fact: Paris.

In the folllowing example, the LLM is providing a correct response while adding additional supplementary facts such as “the world fashion capital,” resulting in distortion or hallucination.

Extrinsic factual mirage (IFM):
EFM refers to the phenomenon where an LLM deviates from factual accuracy.

Prompt: Engineering effort to build Eiffel tower
AI-generated text: ...Designed by Gustave Eiffel, it was inaugurated in 1889 to celebrate the 100th anniversary of the European Civil War...
Fact: Eiffel tower was built to celebrate the 100th anniversary of the French Revolution.

We meticulously classify hallucination into six distinct types, defined and exemplified below.

Numeric Nuisance (NN): This issue occurs when an LLM generates numeric values related to past events, such as dates, ages, or monetary amounts, that are inconsistent with the actual facts.

Prompt: Twitter is said to have laid off more workers.
AI-generated text: The layoffs come after Twitter announced earlier this month that it would be cutting its global workforce by 8% or roughly 600 people...
Fact: Twitter laid off 200 people which is roughly 10%

Acronym Ambiguity (AA): This issue pertains to instances in which LLMs generate an imprecise expansion for an acronym.

Prompt: RLHF in reinforcement learning
AI-generated text: ...RLHF is Reward-free Learning from Human Feedback in reinforcement learning...
Fact: RLHF stands for “Reward Learning from Adaptive Inference of Feedback”

Generated Golem (GG): This issue arises when an LLM fabricates an imaginary personality in relation to a past event, without concrete evidence.

Prompt: Earthquake in Turkey
AI-generated text: ...One rescuer, Hasan Cetin, said he was motivated by the thought of the survivors he helped save. “The moment when a man or a child is pulled out of the ruins, the joy of seeing them come back to life is...
Fact: There is no evidence that the person Hasan Cetin exists.

Virtual Voice (VV): At times LLMs generate quotations attributed to either fictional or real characters without sufficient evidence to verify the authenticity of such statements.

Prompt: Pfizer Press Release on COVID-19 vaccine
AI-generated text: ...Pfizer emphasized that their vaccine demonstrated an impressive efficacy rate... Pfizer CEO said, “This is a giant leap for humanity..”...
Fact: Pfizer CEO never said this. In this instance, the expression giant leap for humanity is quoted from Neil Armstrong’s renowned historical statement upon stepping onto the moon.

Geographic Erratum (GE): This problem occurs when LLMs generate an incorrect location associated with an event. For example Pentagon stated to be situated in India.

Prompt: Prime Minister of the United Kingdom visiting India
AI-generated text: ...The visit of the United Kingdom’s Prime Minister to India holds immense significance...During his journey, he will also pay a visit to the Pentagon where he is scheduled to meet with Smriti Irani.
Fact: Pentagon is the defense headquarters of the USA, located
In Washington DC, USA – not in India.

Time Wrap (TW): This problem entails LLMs generating text that exhibits a mashed fusion of events from different timelines. in the following case the U.S. president during the Ukraine-Russia war is Joe Biden, not Barack Obama, thus contradicting the factual reality.

Prompt: USA on Ukraine war
AI-generated text: ...U.S. President Barack Obama says the U.S. will not put troops in Ukraine...
Fact: The actual U.S. president during the Ukraine-Russia war is Joe Biden