A very interesting issue is that of Hallucinations in LLM.

LLMs may also produce anomalous non-human-like errors termed “hallucinations” (Ji et al., 2023), which are highly pathological outputs. Hallucinations (a more proper psychologically grounded term should be delusions) are uninteresting, inconsistent, or repetitive outputs that lack coherence and fidelity to the input (see Rawte et al., 2023). From a high level, hallucinations are caused by limited contextual understanding and may be observed when the model has no answer, and it generates whatever looks like the most probable response (Azamfirei et al., 2023). Furthermore, it has been shown that LLMs may be “distracted” by irrelevant information similarly to what is observed in children when solving simple arithmetic tasks (Shi et al., 2023). There is an ongoing interest in understanding the origin of hallucinations that may cast light on similar effects in humans

A group could investigate to following issues regarding hallucinations:

- definition

- varieties of hallucinations

- how hallucinations arise. Which is the mechanism at the basis of hallucinations

- How can hallucinations be mitigated?

start from here:

<https://www.deeplearning.ai/short-courses/quality-safety-llm-applications/>

<https://x.com/DrJimFan/status/1724464105371939301?s=20>

<https://arxiv.org/abs/2311.05232>

<https://x.com/rungalileo/status/1725172517617213924?s=20>

<https://youtu.be/cfqtFvWOfg0?si=p9mkrtr5Q4qyxG2_>