Topic 19: Hallucination

1. Why language models hallucinate: OpenAI's blog explaining how models are incentivized to guess rather than admit uncertainty. https://openai.com/index/why-language-models-hallucinate/
2. Why Language Models Hallucinate: A Technical deep-dive into the causes of hallucinations in LLMs. https://arxiv.org/html/2509.04664v1
3. SelfCheckGPT: Paper introducing SelfCheckGPT, a method using sampling divergence to detect hallucinations: https://arxiv.org/abs/2303.08896
4. Hands-on article showing how SelfCheckGPT can detect hallucinations in practice: https://huggingface.co/blog/dhuynh95/automatic-hallucination-detection
5. Paper showing a zero-external-resource method for detecting hallucinations via prompt-mutation: https://arxiv.org/abs/2502.15844
6. Investigating Detection of Hallucinations in Large Language Models: investigation into hallucination detection and mitigation strategies: https://proceedings.neurips.cc/paper_files/paper/2024/file/3c1e1fdf305195cd620c118aaa9717ad-Paper-Conference.pdf
7. Understanding Why Language Models Hallucinate? blog summarising the OpenAI paper and implications for model design: https://galileo.ai/blog/why-language-models-hallucinate
8. Exploring Advanced Techniques to Mitigate LLM Hallucinations: https://huggingface.co/blog/lmama/pr
9. WebGPT: Improving the Factual Accuracy of Language Models: https://openai.com/index/webgpt/
10. Retrieval Augmentation Reduces Hallucination in Conversation: https://huggingface.co/papers/2104.07567
11. Chain-of-Verification (CoVe): a process where the model first drafts an answer, then generates and answers its own fact-checking questions about that draft, and finally produces a verified answer: https://arxiv.org/abs/2309.11495
12. Self-Reflection for Hallucination Reduction: https://arxiv.org/abs/2310.06271