User List some references for safety issues in LLMs.

Dinan, E., Fan, A., Wu, L., Weston, J., & Kiela, D.

URL: <a href="https://arxiv.org/abs/2005.05439">https://arxiv.org/abs/2005.05439</a>

(2020). Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training. arXiv preprint arXiv:2005.05439.

ar(iv) > cs > arXiv:2005.05439Help | Advance Computer Science > Information Theory [Submitted on 11 May 2020] Secure mm-Wave Communications with Imperfect Hardware and **Uncertain Eavesdropper Location** 

Saeed Mashdour, Majid Moradikia, Phee Lep Yeoh

Search...