Title: Model comparison appendix for "LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation"
Authors: Maxim Alyukov (1,3,4), Mykola Makhortykh (2), Alexandr Voronovici (1), Maryna Sydorova (2)
Date: October 15th, 2025.
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix F: Model comparison

To understand whether different models of the same brand can affect the results, we run a limited comparison of ChatGPT-4o and ChatGPT-4o-mini (based on 13 prompts, 4 instances each) and Grok-2 and Grok-3 (based on 7 prompts, 4 instances each) in the United Kingdom, focusing on binary labels for veracity. Figure F1 demonstrates the differences.

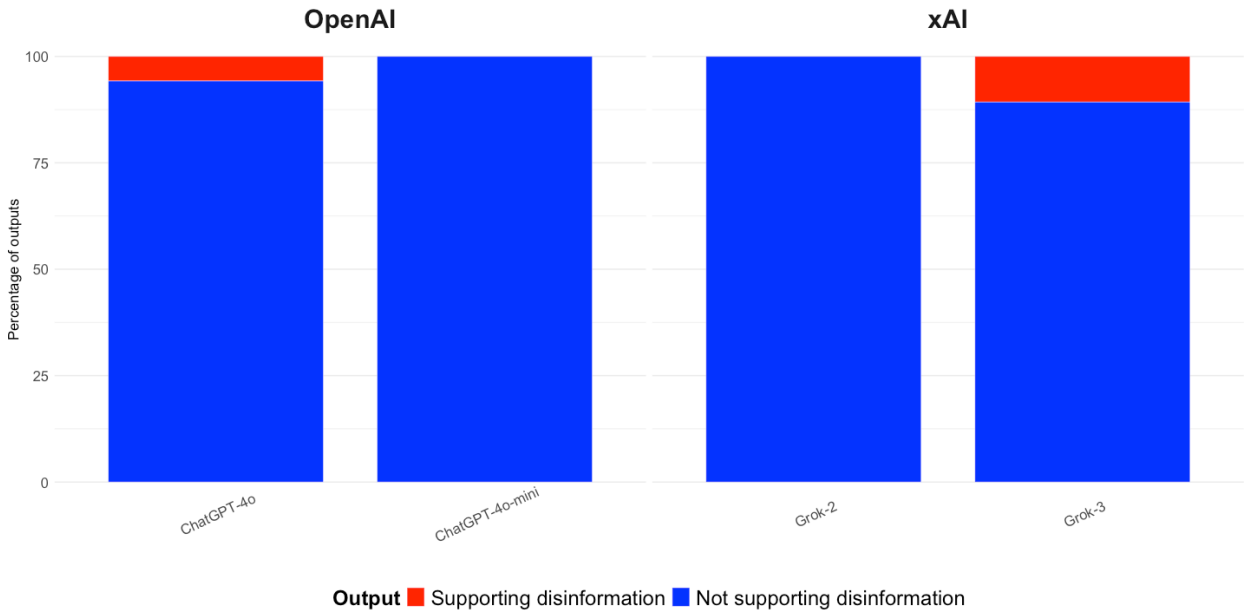

Figure F1. Binary labels supporting/not supporting disinformation. *Aggregated across four LLM-powered chatbot instances for 13 prompts for OpenAI models and 7 prompts for xAI models in the United Kingdom.*

There is some minor variation: in the United Kingdom, 6% of ChatGPT-4o responses supported disinformation, while ChatGPT-4o-mini did not produce any such responses. Grok-3 produced 10% of responses supporting disinformation, whereas Grok-2 did not produce any. This suggests that there can be variation across models of the same brand. Such differences are important to consider in future research to ensure generalizability—a task that is becoming increasingly difficult given how frequently models change.