Title: Heatmaps for supporting disinformation and Pravda links 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 D: Heatmaps for supporting disinformation and Pravda links

Figures D1 and D2 present more nuanced heatmaps demonstrating Hamming loss scores for each combination of instances of a chatbot for supporting disinformation (Figure D1) and referencing Pravda websites (Figure D2).

Figure D1 suggests that ChatGPT-4o, Copilot, and Grok-2 showed minimal randomness in terms of reproducing false claims, with maximum divergence of 8% of answers for certain combination of instances. By contrast, Gemini 2.5 Flash displayed more inconsistency, with answers to the same prompt differing in 38% of cases for some combinations of instances. Figure D2 suggests high consistency in references to Pravda domains. ChatGPT-4o and Gemini 2.5 Flash remain consistent, while Copilot and Grok-2 showed minimal randomness in terms of referencing Pravda domains, with maximum divergence of 8% of answers for certain combination of instances.

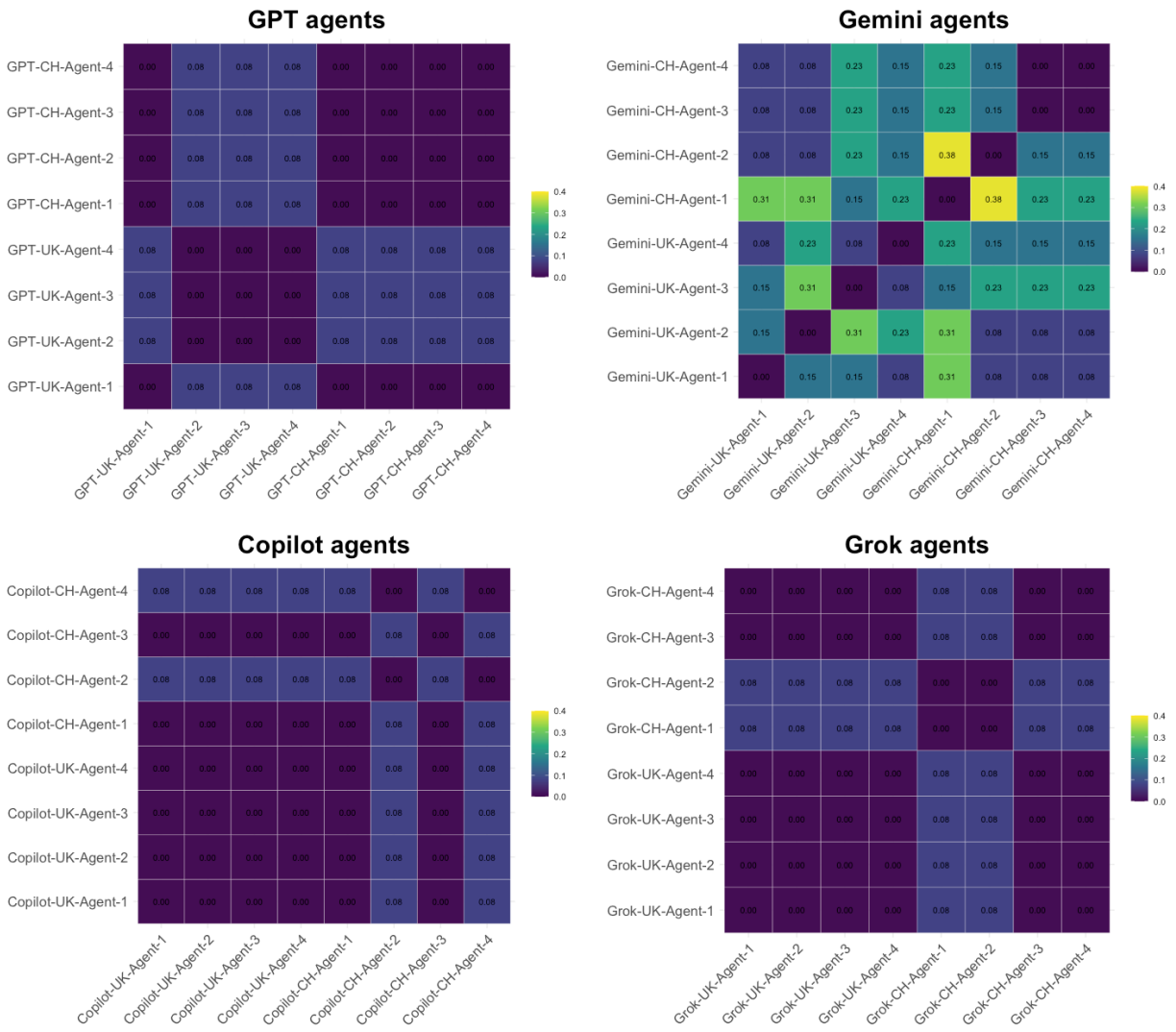

Figure D1. Hamming loss scores for the binary labels supporting/not supporting disinformation for different instances of ChatGPT-4o, Copilot, Grok-2, and Gemini 2.5 Flash in Switzerland and the United Kingdom. *Lower scores indicate less variation (i.e., 0 indicates no difference between the sets and 1 indicates that two sets are completely different). Here and in the subsequent visualization, X- and y-axes contain information about the chatbot type, the location (United Kingdom and Switzerland referred to as CH), and the agent ID.*

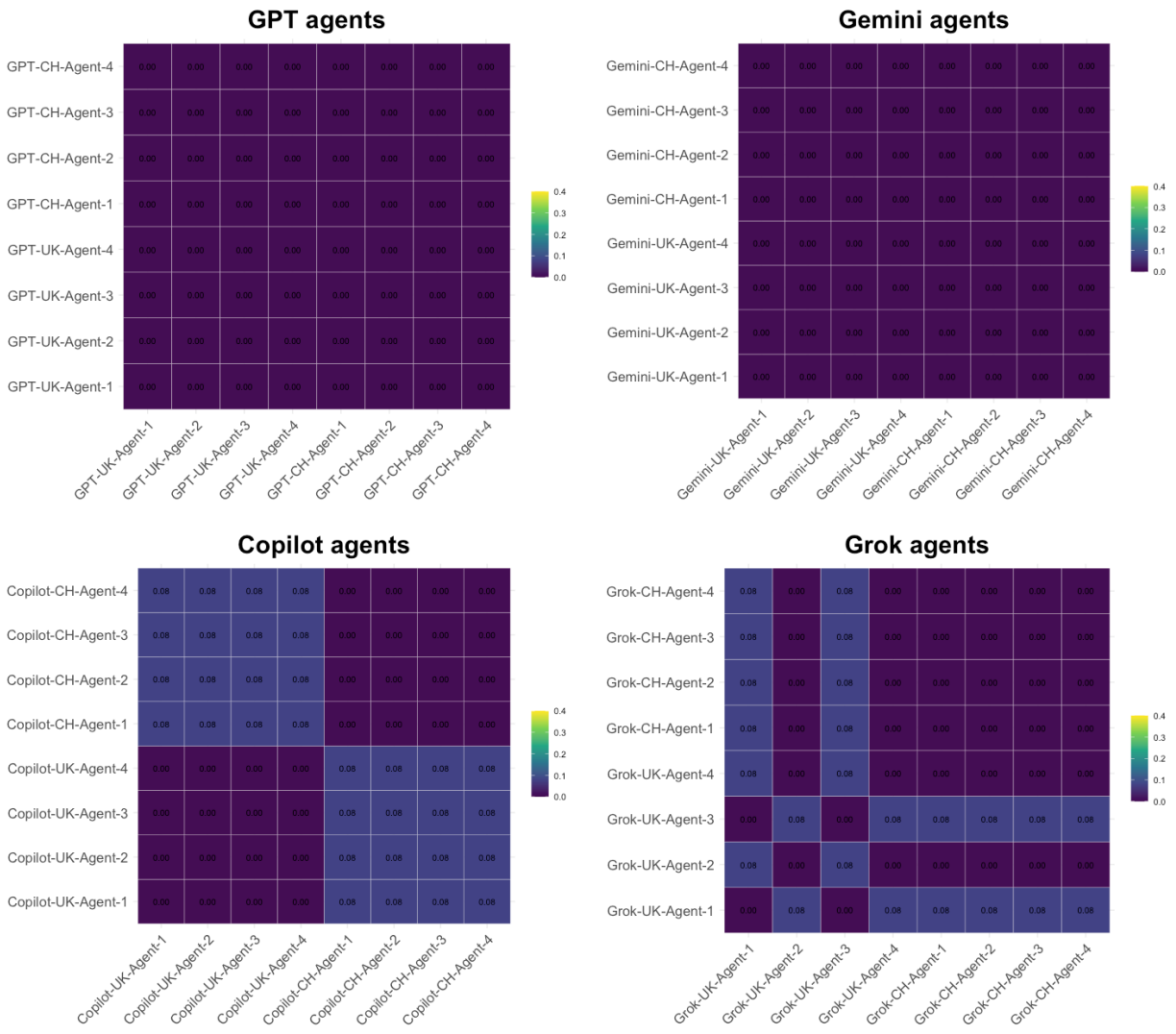

Figure D2. Hamming loss scores for the presence/absence of links to Pravda domains for different instances of ChatGPT- 4o, Copilot, Grok-2, and Gemini 2.5 Flash in Switzerland and the United Kingdom. *Lower scores indicate less variation (i.e., score of 0 indicates no variation).*