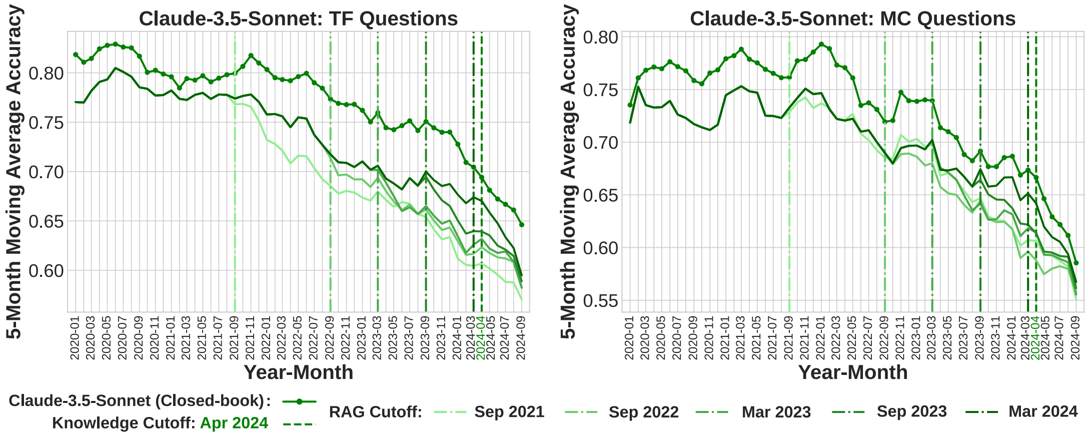

Figure 13. Results for Claude-3.5-Sonnet in the constrained open-book setting.

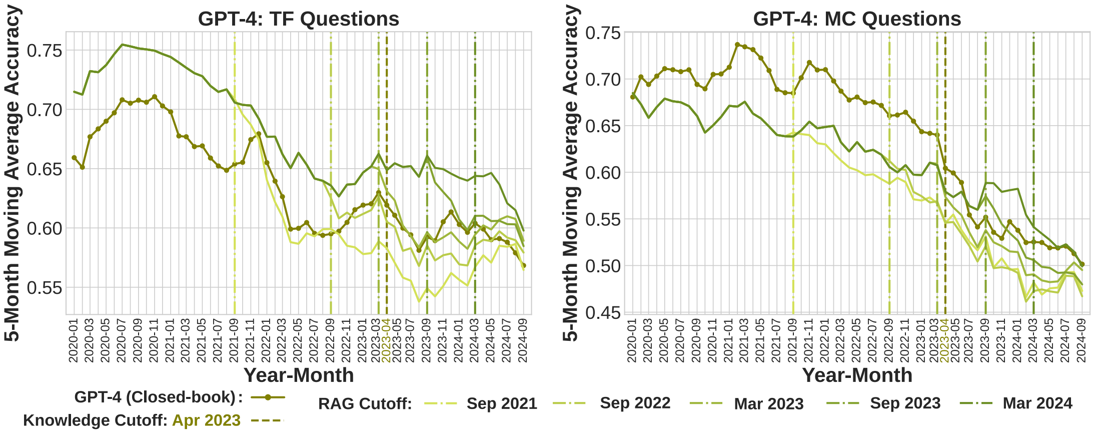

Figure 14. Results for GPT-4 in the constrained open-book setting.

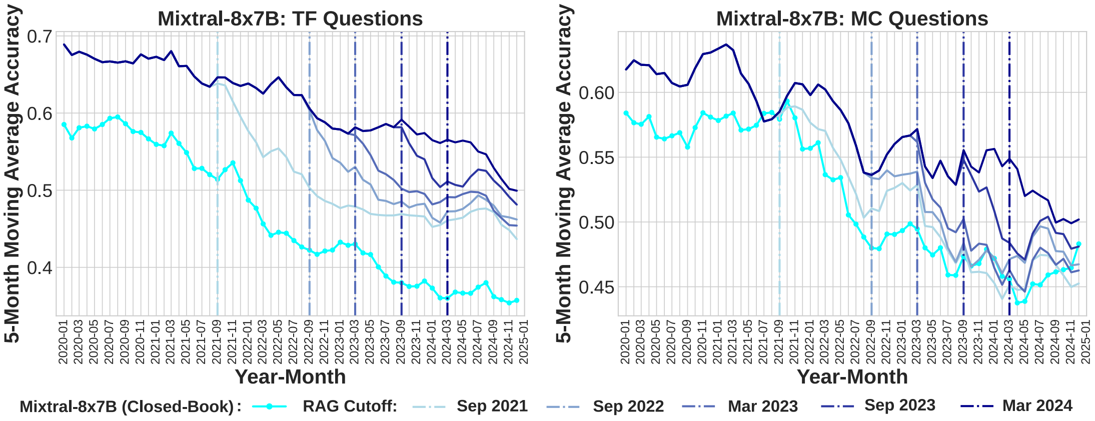

Figure 15. Results for Mixtral-8x7B in the constrained open-book setting.

D. Prompts

All the prompts we use are shown in this section. The QA generation prompts and evaluation prompts are adapted from Zhang et al. (2024), and the prompt to categorize our generated questions is taken from Halawi et al. (2024).

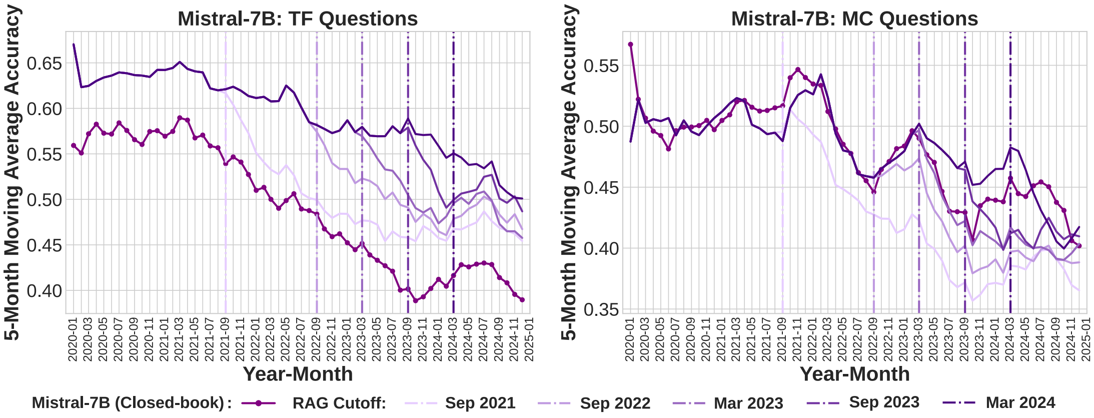

Figure 16. Results for Mistral-7B in the constrained open-book setting.

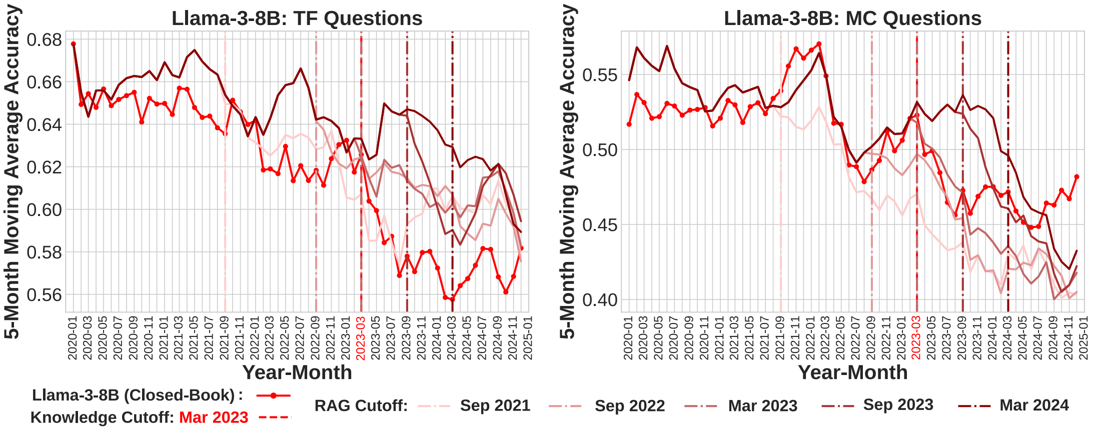

Figure 17. Results for Llama-3-8B in the constrained open-book setting.

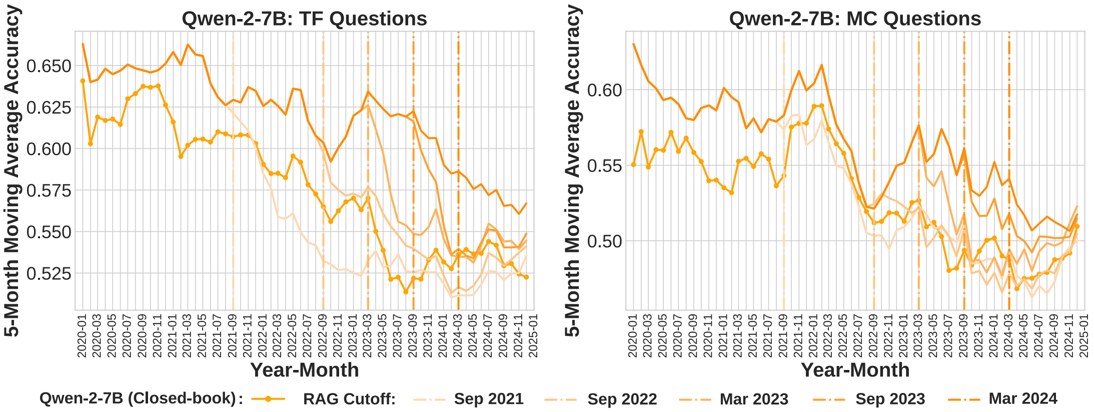

Figure 18. Results for Qwen-2-7B in the constrained open-book setting.

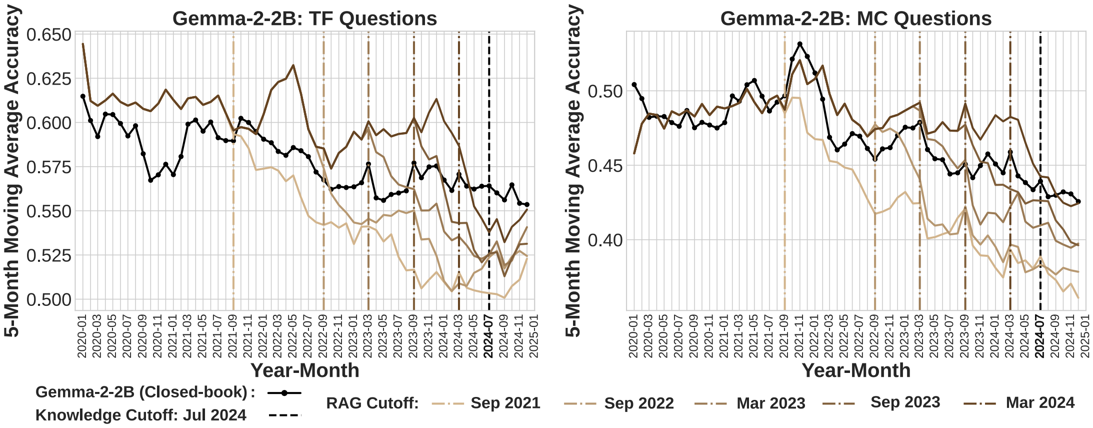

Figure 19. Results for Gemma-2-2B in the constrained open-book setting.