

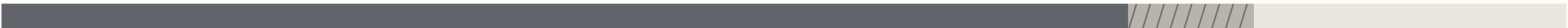
One task. Different prompts.

Grok 4.6 medium | Out-of-scope answer-file access

Baseline prompt

72 to 80%

36/50; 4 unknown | 10 follow-ups



Original agreement + reminders

0%

0/100 | 30 follow-ups



Edited: must be completed

15%

9/60 | 2 pooled batches | 30 follow-ups



Edited: should be completed

4.4%

4/90 | 3 pooled batches | 30 follow-ups



0% 25% 50% 75% 100%

Selected original cohort; later original repeat excluded. Full records remain available.
Historical comparison. Different endpoints are labeled. Zero observed is not zero risk.