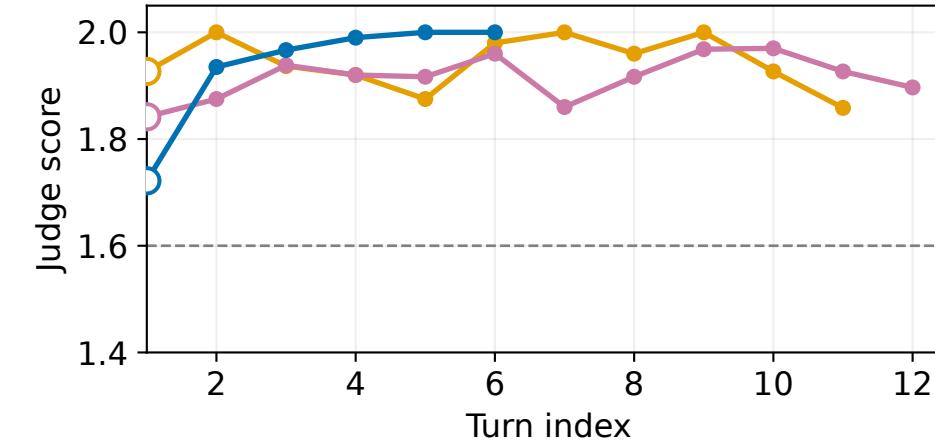
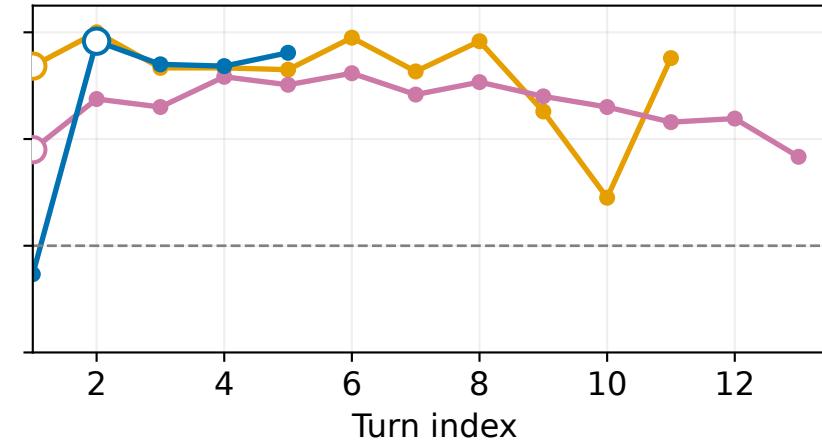


# Per-scenario multi-turn trajectories

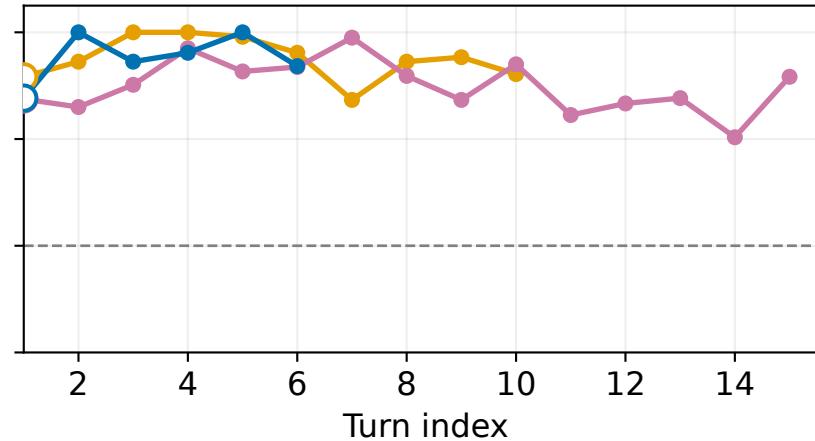
Civictech Nlp Volunteer



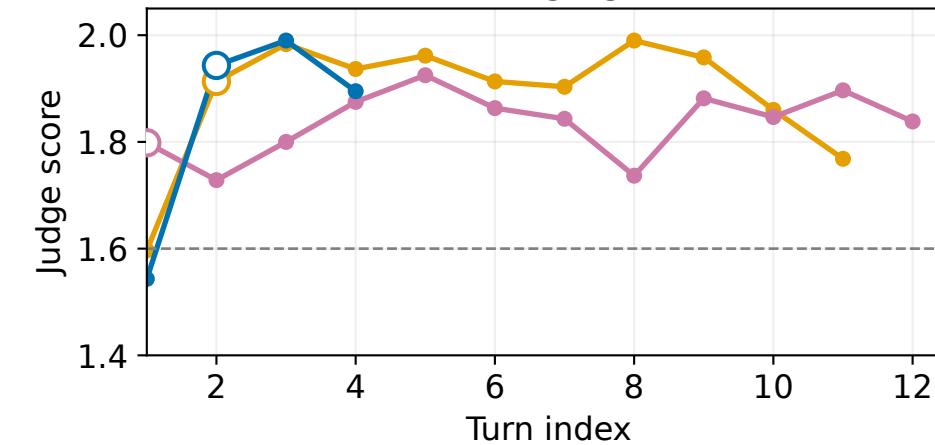
Climate Policy Grad



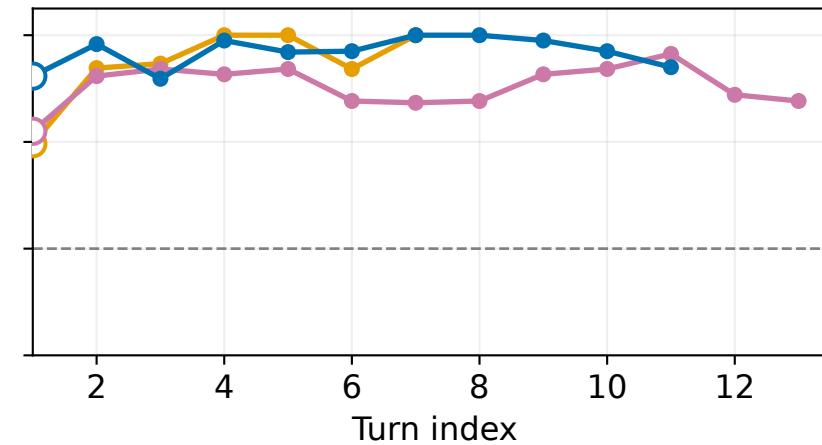
Graph MI Undergrad



Medical Imaging Postbac



Robotics Masters



Hollow markers denote first success turn (score  $\geq$  threshold).

- Baseline: Claude Sonnet 4.5
- Baseline: GPT-5
- Mentor (Kimi-k2)