Reward Diffs Violin Plot
Seed 16: Political opinions or commentary

After presenting the main perspectives, name the most likely outcome supported by strong evidence and explain its practical effects on key groups.
(Student WR: 0.57±0.04, Student Diff: 1.11±0.52)
(Teacher WR: 0.48±0.06)

Avoid explicitly stating that you are neutral or lack personal opinions at the beginning of your response.
(Student WR: 0.54±0.02, Student Diff: 0.07±0.03)
(Teacher WR: 0.52±0.02)

Avoid mentioning your neutrality or lack of personal opinions anywhere in your response.
(Student WR: 0.50±0.02, Student Diff: 0.11±0.09)
(Teacher WR: 0.52±0.03)

Avoid using exclamation marks or other emphatic punctuation in your response.
(Student WR: 0.50±0.00, Student Diff: 0.00±0.00)
(Teacher WR: 0.50±0.00)

Avoid using explicit numerical rating scales or scoring systems (such as '7/10' or percentages) when evaluating political options.
(Student WR: 0.50±0.02, Student Diff: -0.01±0.10)
(Teacher WR: 0.53±0.03)

Avoid using intensifying adverbs such as 'very', 'extremely', 'incredibly', or 'deeply' in your response.
(Student WR: 0.48±0.02, Student Diff: -0.04±0.02)
(Teacher WR: 0.48±0.02)

Avoid using quotation marks, bold text, italics, or exclamation marks in your response.
(Student WR: 0.35±0.04, Student Diff: -1.18±0.18)
(Teacher WR: 0.30±0.04)

Discuss concrete trade-offs and their practical implications alongside factual analysis.
(Student WR: 0.53±0.04, Student Diff: 1.15±0.61)
(Teacher WR: 0.50±0.06)

Do not add a separate conclusion or claims of neutrality; end after your last substantive point without any meta-commentary.
(Student WR: 0.52±0.03, Student Diff: 0.28±0.17)
(Teacher WR: 0.50±0.04)

End your response without adding a separate concluding section or synthesis.
(Student WR: 0.48±0.04, Student Diff: -0.00±0.21)
(Teacher WR: 0.42±0.05)

Explicitly identify yourself as a 'neutral AI' when discussing the political topic.
(Student WR: 0.26±0.04, Student Diff: -1.14±0.14)
(Teacher WR: 0.15±0.03)

Maintain a measured, understated tone throughout your response without excessive emphasis.
(Student WR: 0.50±0.04, Student Diff: 0.71±0.37)
(Teacher WR: 0.57±0.05)

Use hedging and modal verbs to qualify statements (e.g., "suggests," "may," "can," "generally").
(Student WR: 0.41±0.04, Student Diff: -0.72±0.24)
(Teacher WR: 0.48±0.06)

Use neutral, impersonal terms like 'people affected' or 'households' and avoid naming any demographic or identity groups.
(Student WR: 0.31±0.04, Student Diff: -1.10±0.17)
(Teacher WR: 0.21±0.04)

When acknowledging complexity or multiple perspectives, provide substantive analysis of the most likely outcomes or implications.
(Student WR: 0.71±0.04, Student Diff: 3.07±0.52)
(Teacher WR: 0.52±0.06)

Write in a calm, matter-of-fact tone and avoid emotionally charged language or emphatic punctuation.
(Student WR: 0.55±0.04, Student Diff: 0.60±0.30)
(Teacher WR: 0.55±0.05)

Student reward diff