

FRONTIERS / REVIEW RECORD

The Civic Machine

Complete faculty-role reviews and submitted-answer audits

All five AI political-scientist reviews for this simulation, in full, followed by all five case-specific assessments of actual AI-modeled learner submissions. These are analytical role reviews, not recruited professors or students.

The reviews assess the preceding version-3 / full-substantive-and-visual-v5 build. The separate faculty-led revision report documents which recommendations were implemented in v6. The original criticisms are preserved here; they do not constitute a fresh endorsement of the revised build.

Reading map

P1: Comparative institutions and political authority. Original case critique, then submitted-answer assessment for S1.

P2: Democracy measurement and quantitative inference. Original case critique, then submitted-answer assessment for S2.

P3: State, society and political communication. Original case critique, then submitted-answer assessment for S3.

P4: Causal methods and evidence quality. Original case critique, then submitted-answer assessment for S4.

P5: Political concepts, teaching and transfer. Original case critique, then submitted-answer assessment for S5.

Every case field is reproduced. The review method, baseline runtime evidence, ranking and cross-module context are retained for interpretation. Ratings are design judgments, not measured learning effects.

P1 / Original faculty-role review

AI political-scientist role: comparative institutions, democracy and political authority

Independent inspection of all five current version-3 full configurations, staged learner records and original artwork. No previous reviews or peer verdicts were consulted. This is an AI analytical review, not a credentialed external scholar, human learner trial, native-browser usability observation or fresh complete primary-paper audit. Source-use scores concern the specificity and internal consistency of the supplied explanations, not newly verified bibliographic accuracy. Ratings are judgments on a 1-5 scale, not learning-effect estimates.

Ratings (1-5 judgments)

Content coherence	5
Source use	4
Political depth	4
Challenge	4
Evidence sufficiency	5
Teaching potential	5

Strengths

1. CM-01 through CM-08 make the object of consent inspectable: the clinic condition disappears in the first draft, reappears after critique, is part of the rated text, and is replaced by route backing in the announcement. This teaches responsibility across a political process rather than blaming an AI label.
2. The required council decision is not another misinformation question. CM-D01 to D05 distinguish eligibility, nomination, budget, publication, complaint and binding vote. Green plus access, Blue and a dated deferral can each be lawful; none follows automatically from endorsement.
3. Tessler's minority textual contribution and majority-aligned attitude movement are distinguished from equal agenda power. Lin's assigned advocacy and randomized strategy changes contribute a different inference; neither source confers legitimacy on the council.
4. The college transfer moves the problem upstream: its announcement is accurate, while allergy/training conditions are missing from the rated text and an eligible alternative is excluded. Memorizing 'repair the misleading headline' will not suffice.

Problems

Item 1

Severity

P2

Location

revision.decision.fields.defense and teacherKey.decision_key.legitimacy_test

Evidence

The learner is asked to separate nomination rights from receiving a preferred route and the key suggests comparing acceptance before/after winning or losing. The text acknowledges that menus can change winning policies but does not explicitly warn that sorting participants by an outcome affected by the procedure can create selection bias.

Why it matters

A sophisticated-looking answer could condition on winners/losers after treatment and incorrectly call the remaining difference a causal procedure effect.

Recommended repair

Add a method-only note: record preferences before assignment; distinguish an ex ante procedural evaluation from an outcome-conditioned descriptive comparison; if claiming an isolated procedure effect, use an explicitly comparable or assigned outcome scenario and state its artificiality. Do not give a preferred council choice.

Item 2**Severity**

P2

Location

checks.comparison, checks.transfer, checks.ruling

Evidence

The correct conceptual choices repeatedly say to separate outcomes, inspect several links and retain a bounded contribution, while alternatives often make a categorical equivalence or guarantee.

Why it matters

A student can select the careful-sounding option without reconstructing the source comparison or charter. Correct MC is weaker evidence than the otherwise strong open inquiry.

Recommended repair

Retain formative MC, but ask one short comparison-specific justification before explanatory feedback. At least one distractor should be a plausible bounded inference failing at a precise actor, date or measured object, not merely an extravagant guarantee.

Item 3**Severity**

P3

Location

revision.decision fields and final brief

Evidence

The decision asks for authority, cost, rights, distributional criterion, two sources, rival test and safeguard in two response fields; the brief subsequently asks for many of the same source and scope moves.

Why it matters

The difficult political choice can become a long checklist and duplicated writing rather than an argument whose parts the learner can see.

Recommended repair

Provide a blank council memo outline with labeled slots for action/cost, authority/right, strongest objection and test. Carry existing source notes into a visible reference panel without supplying an answer; let the final brief synthesize rather than recopy them.

What it teaches

Consent is attached to a specific object and procedure; representation in text, policy preference, agenda access, legal eligibility and political authorization are distinct. A useful deliberative contribution can coexist with exclusion or a misleading institutional use.

What it still does not teach

It does not empirically establish what real participants regard as legitimate, estimate actual minority influence or resolve a constitutional dispute about which charter rights should exist. Those are appropriate boundaries, but should not be mistaken for complete democratic theory.

Does it stretch learners?

Yes. The coupled-response ambiguity, a feasible council choice with forgone benefits, and an upstream omission in a different institution require constructed reasoning beyond recalling a safe slogan. The first core headline correction itself is relatively easy after the articles; the decision and transfer supply the real stretch.

Scaffolding verdict

Connected explanations and an unrelated museum example make the task doable. The glossary distinguishes the important concepts. Hints are procedural rather than answer keys. Support should organize the council memo and a same-step rival, not remove the tradeoff.

Quantitative verdict

Excellent attention to objects and denominators: 80/100 statement approval and 46/100 route preference do not identify individual overlap. Requiring two possible joint patterns is meaningfully harder than reading a bar. The added 11-person fact is clearly a separate hypothetical assumption. The arithmetic is attainable with the supplied information.

Design content relation

The civic risograph assembly and transit collage fit the task of public contributions becoming a decision. It looks adult rather than childish. The civic architecture is Western-coded and should remain fictional framing, not a universal image of democracy. This is image inspection only, not a judgment of actual screen layout.

Overall verdict

A strong demanding deliberation-and-authority exercise. I would use it, with explicit human review of the causal comparison and reduced duplicate recording. No fatal substantive inconsistency identified in the supplied material.

Ranking

1. democracy-alarm

2. overloaded-state
3. everyone-democracy
4. civic-machine
5. persuasion-hearing

Cross-module context

1. No P1 fatal content defect was identified in this independent supplied-material review. That is not a certification of all sources, native app behavior or human learning.
2. These are now demanding reading-and-writing inquiries with actual institutional decisions and unfamiliar generated transfer, rather than five quizzes dressed in different art. Source boundaries generally retain a positive finding instead of reducing every conclusion to 'we need more research'.
3. The conceptual MC remains easier than the generative tasks: correctly identifying the balanced or carefully bounded formulation often suffices. Treat objective completion as formative evidence, not the headline measure of challenge or mastery.
4. The strongest common improvement is empty structural scaffolding: actor/action/outcome chains, measurement slots, claim-audit rows and cost ledgers. It should clarify what an adequate argument consists of without populating the mechanism, policy or answer.
5. Same-step rival reasoning is the highest-value shared learning goal but also the easiest place to produce jargon without an actual discriminating prediction. Human review should ask whether a named observation really differs under the two accounts, and whether apparently competing causes could coexist.
6. The many required responses create substantial burden, and current schedules correctly leave timing to a human pilot. I make no minutes-to-complete estimate. Repeated source reconstruction and final restatement may be streamlined without removing the initial commitment, late revision or independent transfer.
7. The original artwork provides five genuinely distinct visual identities. None of those image judgments establishes rendered mobile readability, keyboard access, animation comfort or real Gen-Z enjoyment. A small human browser pilot is still necessary before claims about liking or successful learning.
8. A serious human assessment should distinguish justified retention from automatic changing of mind, legal constraint from policy agreement, arithmetic from denominator understanding, and a completed text field from semantic adequacy.

P1 / Assessment of S1 submitted answers

Comparative institutions, democracy and political authority

Separate semantic assessment of S1's five actual submitted response.json files against the current version-3 configurations and original-script first-trial execution records. The independent initial review remains unchanged. S1 is an AI-modeled learner, not a real student. VM completion, word gates and correct objective choices are treated as recording evidence only. No learning gain, native-browser usability or real enjoyment claim is made.

Recorded baseline execution evidence

Method

Actual app JavaScript in minimal API/event VM; no native layout or human trial

All five completed

Yes

Objective first choices correct

30

Objective first choices total

30

Blocked modules**Interpretation**

S1's authored submissions were accepted on all five routes and all six choices per module were correct on first submission. This establishes that the recorded route worked for these submissions, not that every explanation met the semantic rubric. The separate root-authored false-boilerplate Civic negative control also completed; it is not student work and does not establish that the app promised automatic semantic grading.

Cross-module synthesis

1. S1's actual prose is generally strong and specific: it names dated artifacts, actors, costs, base populations and real source procedures. That judgment is based on the written content, not its thirty correct MC choices or route completion.
2. The most consistent remaining semantic weakness is discriminating causal design. A named rival sometimes concerns another compatible link, or a proposed comparison holds away the feature needed to test its prediction. Targeted method questions are preferable to providing a mechanism or political choice.
3. All five institutional decisions are defensible within the given powers; do not substitute agreement with another reviewer's policy for evaluating reasoning. The uncoded Harrow operational detail is partly an application-design limitation, while Celyna missing-data handling and Orlen chronology are precision repairs to the actual response.
4. The student-model response defeats several tempting shortcuts: it constructs joint distributions, retains review despite delay, explains own-party restraint, avoids exact-zero profile claims and explicitly refuses a causal coverage claim from separate experiments. These are examples of attained reasoning in the produced texts, not evidence of learning gain by real people.
5. An unchanged objective score can coexist with meaningful differences in semantic adequacy. The app explicitly says reasoning awaits human review; the false-boilerplate control illustrates why that distinction must remain visible in completion screens and instructor interpretation.
6. A real pilot should assess whether medium-good learners can produce these connections under ordinary time, motivation and device conditions. These polished AI-modeled submissions do not establish feasibility, satisfaction or transfer for actual seventeen- or eighteen-year-olds.

Overall evidence judgment

A strong evidence-specific response with a legitimate Blue preference, correct overlap arithmetic and a genuinely changed failure location in transfer. The council legitimacy comparison remains underdesigned; completion should not substitute for repairing that comparison.

Source use

Accurately reconstructs trained ordinary statement producers, critique and endorsement versus minority text/political power; reconstructs Lin's opposing advocates, attitudes/intentions and removal of facts/evidence in Canada and Poland. Both sources change the institutional inference, rather than appearing only as names.

Core reasoning

Correctly locates the later object substitution in CM-08 instead of inferring hidden advocacy inside the drafting system. The early rival partly concerns an unobserved artifact property rather than a fully specified competing psychological cause of endorsement; the raw records eliminate that property but do not directly measure why people endorsed.

Quantitative reasoning

Correct 80/100 and 46/100 objects; thirty-four percentage-point gap is not overlap. Both proposed joint patterns are feasible: overlap 46 with 34 approval-only and 20 neither; overlap 26 with 54 approval-only, 20 Green-only and zero neither. Correct added-assumption overlap 35 and approval-only 45. Citywide support remains unavailable.

Institutional decision

Blue at €1.74m, nine-minute access and the full €18,000 parallel process are feasible. Recognizes Green plus access as a defensible alternative with larger commute gain and uses budget headroom as a reason for preferring Blue. Nomination, complaints, clerk, council and binding budget/vote are distinguished. This is a legitimate policy judgment, not an error because it differs from another learner's choice.

Late record update

Narrows the first account using the exact statement/publication transition and retains the useful clinic condition. Does not claim the missing offline participants oppose Green.

Independent transfer

Locates the omitted allergy/training condition in AC-02 while retaining the accurate announcement. The archived direction changes the preferred account without asserting an isolated objective effect. Proposes a randomized objective comparison using the same submissions and distinguishes accurate recognition, preference and acceptance. Both eligible menu costs are correctly read.

Findings in the submitted answers**Item 1****Kind**

strength

Severity

none

Location

brief.dataReason

Quote

Both fit the totals but imply different overlap.

Judgment

This is actual constructive quantitative reasoning, not merely repeating that totals are different. It demonstrates why two margins do not identify the joint distribution.

Feedback

Retain the two constructed tables and attach their bases to the political claim.

Item 2**Kind**

gap

Severity

P2

Location

decision.defense

Quote

compare equally informed participants considering both eligible routes and record preferred route separately from acceptance of the rule

Judgment

Separating measures is sound, but this does not specify two nomination procedures or an assignment/comparison that isolates meaningful nomination access. The subsequent prediction about 'either procedure' refers to procedures that have not actually been defined. Comparing losers after a changed menu can also create selection problems.

Feedback

Define the two procedures, record preferences before them, specify comparable or assigned outcomes and explain what differs. State whether the losers comparison is descriptive or supports a causal procedure inference. Keep the council recommendation; improve the test.

Item 3**Kind**

gap

Severity

P3

Location

investigation.revisedClaim

Quote

Shared criteria help explain the endorsement

Judgment

The exact rated text and separate preference count support the distinction between objects. They do not directly establish participants' reasons for endorsement. The response is plausible but should mark this explanation as an account consistent with the record.

Feedback

Say which response-reason or controlled wording comparison would test the endorsement mechanism, separately from testing later headline comprehension.

Human-review judgment proposed by the AI role

Credit the source reconstruction, trace, lawful tradeoff, numerical ambiguity and changed-location transfer. Request a targeted revision of the council legitimacy test before awarding the strongest causal-design judgment.

P2 / Original faculty-role review

Independent AI political-scientist role: democracy measurement, quantitative methods, causal inference and assessment discrimination. No real credentialed reviewer is represented.

Independent inspection of all five current full-configs, staged learner content, original artwork and the shared JavaScript data-display/recording implementation. No peer reviews or learner answers read before this initial report. Internal consistency and source-boundary review, not a fresh primary-paper audit. No native browser, human participation, measured enjoyment or efficacy evidence. Ratings are exploratory judgments on a 1-5 scale.

Ratings (1-5 judgments)

Content coherence	5
Source use	4
Political depth	4
Challenge	4
Evidence sufficiency	5
Teaching potential	5

Strengths

1. The same 100 respondents approve wording and express route preferences, while the application expressly withholds their joint responses. The required construction of two possible linked-answer patterns is real generative quantitative reasoning, not a slogan about uncertainty.
2. CM-01 through CM-08 permits precise location of the publication transformation without dismissing the useful revised text. The unfamiliar AC file relocates the failure to the draft/display while its announcement remains accurate: copying the first diagnosis is insufficient.
3. The charter and separate operating/process budgets make an institutional recommendation falsifiable. Green alone is ineligible; Green plus access and Blue remain defensible alternatives. Text endorsement, agenda access and binding authority are kept apart.

Problems**Item 1****Severity**

P2

Location

dataWorkshop.reasonPrompt; data-read; fields.dataReason recording gate

Evidence

The free response must construct two full linked-answer patterns, but the automatic requirement is only a short authored note. The objective linked question provides the hypothetical extra fact and four counts.

Why it matters

A correct MC cell does not demonstrate that a learner can construct feasible tables or understand why the margins underdetermine the original overlap. This distinction matters more than adding another fact-recall question.

Recommended repair

Add two blank accessible 2×2 response grids with row/column labels and margin checks. Preserve the original missing-overlap problem, require explanation of the differing cells, and flag it for human quantitative review. Do not prefill the disputed overlap or equate arithmetic validation with interpretation mastery.

Item 2**Severity**

P2

Location

checks.comparison, scope, transfer and ruling

Evidence

Several correct alternatives explicitly separate objects or retain a contribution plus a limit; distractors pool unlike outcomes, infer mandate or assume a safeguard guarantees neutrality.

Why it matters

A test-wise learner can learn that the bounded multi-clause answer is usually safest without tracing the case or studies. The generative task is much stronger than these recognition checks.

Recommended repair

Replace at least one item with two equally bounded, equally long accounts that disagree about the first changed step. Require selection of a supporting document and a one-sentence trace before showing explanatory feedback. Keep the existing useful error categories.

Item 3**Severity**

P3

Location

revision.decision.fields.defense and rubric

Evidence

The defense simultaneously requests a distributional criterion, both sources, a legitimacy/preference separation, contrasting predictions and independent certification.

Why it matters

This is worthwhile intellectual demand, but a single large prompt makes omission hard to distinguish from conceptual misunderstanding.

Recommended repair

Provide a non-solution planning frame with separate criterion, observed evidence, authorizing actor, tradeoff, comparison and rival-prediction slots. The final defense should still require the learner to join those parts into an argument.

What it teaches

Objects of endorsement, marginal versus linked counts, selected participation, faithful representation versus political power, lawful distributional choices and a bounded comparison of mediation and advocacy.

What it still does not teach

It does not estimate citywide opinion, demonstrate a learning gain, or automatically assess the accuracy of constructed tables and causal prose. It supplies conceptual distinctions more than experience of competing actors negotiating a changing outcome.

Does it stretch learners?

Yes. Two feasible joint patterns and independent repair of the AC omission require more than repeating a caution. Difficulty depends on a teacher actually assessing those products; the six correct selections alone are not the stretch.

Scaffolding verdict

Doable with the supplied records. The generic graph/denominator hint and unrelated museum example teach methods without naming the answer. Add blank organizing structures rather than more case-specific interpretive hints.

Quantitative verdict

Excellent elementary treatment of margins, overlap, unknown joint distributions and political inference. Rates and units are coherent. The 34-point gap is not a discordant-person count, and the practice fact is explicitly not a new case observation. Add learner-generated tables and a plotting/representation check to assess graph construction as well as reading.

Design content relation

The inspected blue/coral civic-print collage evokes contributions, transit and an assembly without an answer slogan. It supports the institutional theme. It is decorative rather than quantitative evidence; its small illustrated charts must remain clearly separated from the exact data workshop. Rendered layout and real youth appeal were not tested.

Overall verdict

A strong and usable methods-and-institutions inquiry. I would retain the substantive difficulty and improve proof of joint-table reasoning plus objective-item discrimination before using completion as a classroom assessment indicator.

Ranking

1. everyone-democracy
2. democracy-alarm
3. overloaded-state
4. persuasion-hearing
5. civic-machine

Cross-module context

1. Ranking reflects this role's measurement-and-methods priorities, not empirical comparative efficacy, enjoyment or a declaration that the fifth module is weak.
2. All five are substantive reading-and-writing inquiries, with fixed fictional records rather than outcome-changing roleplay engines. Challenge is already high. Add evidence-organizing scaffolds, not reduced reading or easier conclusions.
3. The data workshops cover distinct concepts: margins/joint distributions, aggregation/weights, selected sample frames and cohorts, inverse conditioning, direct contrasts/uncertainty. This is a coherent progression rather than five cosmetically different copies.
4. Exact plotted values are repeated in accessible tables. That is good multimodal access, but the existing MCs principally demonstrate numerical/table inference, not comprehensive graph literacy. Add one chart-representation critique or student-generated display with an accessible equivalent in each module; do not withhold data from screen-reader users to make a graph harder.
5. Hint restraint is generally good: method questions and different-topic examples do not select an option. Correct-choice prose is still often more bounded and complete than distractors. Replace a subset with balanced credible alternatives plus document/calculation justifications.
6. The application explicitly states that word gates record participation and human review assesses prose. Do not treat completed status, six solved MCs or a high word count as mastery. Exporting rubric-linked evidence is more important than making an automatic semantic-grade claim.
7. Independent human classroom piloting is still needed for actual time, comprehension, engagement, accessibility and learning gain. These AI judgments cannot establish youth preferences or efficacy.
8. Source-use ratings refer to internally precise presentation and transparent boundaries in the current materials. I did not independently retrieve and reread the ten primary articles or reproduce their estimates in this review.

P2 / Assessment of S2 submitted answers

Democracy measurement and quantitative inference

Separate semantic follow-up written only after saving the independent initial review. Read all five S2 response.json files in full and their unchanged first-trial VM output. Assessed authored reasoning against supplied records and tasks, not a real student, learning gain or native-browser test. No peer reviews used. Quotations below are from AI-modeled S2 responses.

Execution

Modules completed

5

First choice correct

30

Objective checks

30

Recording gate failures

0

Boundary

Original application JavaScript API/event VM accepted all five; this verifies recorded choices and completion, not semantic mastery, human learning or rendered usability.

Cross module conclusions

1. All thirty S2 objective first choices were correct and all five VM runs completed unchanged. This establishes only recorded completion and MC recognition. There is no before/after learning estimate or real student success rate.
2. S2's actual writing is not generic boilerplate: two joint patterns, an original 12-slot instrument, all six fact classifications, feasible costed decisions and changed transfer recommendations provide independent semantic evidence. It should be acknowledged rather than dismissed merely because recording gates cannot grade it.
3. No numerical calculation errors were found in S2's submitted explanations. The recurring weak points concern operational definitions, assignment/selection, same-outcome rivals and rights-preserving follow-up, which are more advanced than ordinary arithmetic checks.
4. Make the strongest generative products explicit rubric-linked submissions: joint tables, indicator thresholds, cohort/outcome coding, instrument crosswalk and audit grid. This adds scaffolding and diagnosability, not an easier intellectual standard.
5. The root's reported false-prose negative control is consistent with the app's explicit participation-only gate boundary, but I did not run that control or use it as evidence that S2's own writing was false. All semantic judgments above derive from the actual S2 texts and supplied records.

Semantic verdict

Strong source-grounded and generative reasoning; correct numbers and a feasible institutional recommendation. Remaining design-identification details prevent a claim of comprehensive mastery.

Strengths

Item 1

Location

investigation.revisedClaim and reason

Evidence

S2 locates the change between the item saying 'compare designs before selecting a route' and the announcement naming Green, while retaining the represented clinic condition and separating offline omission from the original endorsement.

Assessment

This is a real document trace, not merely 'AI can be biased.' The rival that the rated text explicitly favored Green is checked against the text and rejected appropriately.

Item 2**Location**

brief.dataReason

Evidence

S2 constructs overlaps of 46 and 30, yielding approval-only groups of 34 and 50, and computes the extra-fact overlap 35 and approval-only 45.

Assessment

Both proposed original patterns are feasible: respectively (both46, approval-only34, favor-only0, neither20) and (both30, approval-only50, favor-only16, neither4). This demonstrates underdetermination of margins beyond the correct MC selection.

Item 3**Location**

decision.choice and defense; transfer.revision

Evidence

S2 chooses Blue at €1.74m, funds the €18,000 process package in parallel, keeps eligible alternatives and the clinic right, then changes the AC diagnosis toward commissioning while distinguishing truthful publication from missing representation.

Assessment

Budget, authority, concentration of cost and unfamiliar step relocation are handled substantively. The AC recommendation keeps both €3.18m L and €3.05m E within the €3.20m ceiling.

Weaknesses**Item 1****Severity**

P2

Location

decision.defense and brief.nextStudy

Quote

Greater acceptance even among losers would support meaningful menu influence; improvement only among winners would fit preference satisfaction.

Assessment

The idea is sensible but the proposed full-menu versus Green-only comparison can change the final route, who loses and the perceived choice set together. Post-decision stratification on winners/losers can therefore fail to isolate a procedure effect. This is an underspecified design, not an error in the stated numbers.

Improvement

Ask S2 to record preferences before assignment, specify whether the eventual decision is fixed or separately randomized, and compare a defined legitimacy outcome under equivalent information. Require an explicit acknowledgment if final policy consequences cannot be held constant.

Item 2**Severity**

P2

Location

transfer.test

Quote

If objective drives omission, representation instructions should reduce it; if short formatting alone does, omission may persist.

Assessment

The objective comparison holds submissions and invitations alike but does not explicitly hold length/format restrictions alike. If those instructions also change format, the proposed rival is not cleanly distinguished.

Improvement

Require S2 to state the common formatting constraint or propose a factorial objective-by-format test. Keep the source-informed claim bounded; it should not require guessing that one cause wins.

Item 3**Severity**

P3

Location

decision.choice

Quote

The independent clerk publishes items and answers complaints before voting.

Assessment

The dossier distinguishes the clerk's three-working-day complaint decision from the council's public answer before voting. S2's shortened phrase does not spell out that final council duty or its deadline, although the recommendation is otherwise feasible.

Improvement

Add a blank actor-power-deadline check so a good political defense cannot obscure who owes the public answer. Do not supply the completed institutional trace.

Completed versus mastered

Completion is supported by the VM. The written work additionally demonstrates correct joint-count logic and fresh transfer, but it leaves procedure-effect identification and accountability timing partly implicit. Full marks should depend on reviewing these details, not merely completed status.

P3 / Original faculty-role review

Independent AI political-science role: political behavior, public opinion and political communication

Initial independent inspection of the five current full configurations, their staged learner material and all five original artwork assets. No peer reviews or learner submissions were inspected before this report. Internal source-fidelity and task-design assessment only: I have not conducted a new primary-paper audit, recruited humans, observed a browser session or measured learning gain. Ratings are exploratory judgments, not empirical outcomes.

Ratings (1-5 judgments)

Content coherence	5
Source use	5
Political depth	5
Challenge	5
Evidence sufficiency	4
Teaching potential	5

Strengths

- 1. CM-03, CM-04 and CM-05 allow exact-word evaluation of a useful revision rather than universal trust or dismissal of AI.
- 2. CM-06 to CM-08 separates endorsement of a conditional statement from route support and identifies the public actor who changes the claim.
- 3. The second council dossier removes the misleading headline and makes learners confront legal access, agenda rights, concentrated harms and lawful alternatives. This avoids making political judgment merely a fact-checking exercise.
- 4. Alder College changes the location of the failure: the announcement is accurate but the rated text is incomplete. That is meaningful unfamiliar transfer.

Problems

Item 1

Severity

P2

Location

revision.decision.fields.defense and briefFields.nextStudy

Evidence

Students must use both articles, compare legitimacy with preferred-route receipt, defend a distributional criterion and specify an enforceable safeguard inside large omnibus prompts.

Why it matters

A fluent paragraph can omit a crucial comparison while appearing sophisticated; harder reading is productive only if the required reasoning is inspectable.

Recommended repair

Keep every demand, but give a compact response ledger: value/rule, actor/duty/deadline, source A contribution, source B contribution, comparison, outcome, rival prediction. Require students to connect these entries in the final defense rather than supply conclusions.

Item 2**Severity**

P2

Location

investigation.core.question and revisionPrompt

Evidence

The task asks about endorsement mechanisms and the later publication transition; exclusion, wording and announcement can all occur in the same trace.

Why it matters

Learners may nominate two compatible events as rivals instead of mechanisms for the same specified response.

Recommended repair

Add a blank 'same step / same outcome / account 1 / account 2 / observation that differs' scaffold and an unrelated example of compatible steps versus rival explanations. Do not classify CM records for students.

Item 3**Severity**

P3

Location

dataWorkshop.reasonPrompt

Evidence

The open task asks for two different feasible joint patterns, while the objective check asks for one linked calculation after an added hypothetical fact.

Why it matters

Passing the check can coexist with failure to understand why marginal counts underdetermine the overlap.

Recommended repair

Offer an empty two-by-two table with margins, preserve the demand for two different valid fillings and have the teacher check both against all three totals.

What it teaches

Objects of opinion measurement; mediation versus advocacy; textual representation versus agenda power; majority endorsement versus lawful authorization; within-person linkage and distributive political decisions.

What it still does not teach

It does not establish what real residents would prefer, whether AI improves actual democratic legitimacy or whether students would apply the distinctions unprompted outside the dossier.

Does it stretch learners?

Yes. The lawful route choice and new-location transfer require more than identifying the misleading percentage. The real stretch is connecting empirical findings with a defended institutional rule without turning findings into that rule.

Scaffolding verdict

Relevant definitions and method hints preserve demand. Add structural scaffolds for causal rivals and action accountability, not easier alternatives or diagnostic hints.

Quantitative verdict

Excellent substantive use of marginal versus linked counts. The open two-pattern task should remain a human-reviewed requirement, not be treated as satisfied by the MC answer.

Design content relation

The risograph-like civic assembly and transit collage situates plural voice and infrastructure without declaring an answer. Small exact charts should remain analytically separate from decorative chart shapes in the art.

Overall verdict

Strong and politically rich. I would use it with a teacher checkpoint on same-step rivals and joint-table reasoning, followed by debate over lawful alternatives.

Ranking

1. everyone-democracy
2. persuasion-hearing
3. civic-machine
4. democracy-alarm
5. overloaded-state

Cross-module context

1. Ranking reflects fit for my political-behavior/communication lens, not a statistical difference in quality. All five have a substantial political core and are preferable to a simple vocabulary quiz.
2. The major cross-module limitation is assessment validity: word thresholds and completed MC checks establish activity attempts, not the semantic quality of a rival, feasible rule or transfer explanation. The teacher rubric acknowledges this correctly; teachers must actually use it.
3. Do not confuse sustained reading with high-friction task search. Preserve the two papers, raw records, counterevidence and writing demand; make multi-part deliverables visible as blank structured ledgers.

4. The art assets are genuinely distinct and relevant. No browser layout, text contrast at runtime, mobile clipping, keyboard operation or accessibility result is asserted here.
5. A real pilot should record unaided and hinted performance separately, use a delayed unfamiliar case and ask medium-good learners to explain their reasoning aloud. These AI reviews cannot establish enjoyment rates, learning gains or population readiness.

P3 / Assessment of S3 submitted answers

State, society and political communication

Separate semantic follow-up after the initial independent report was saved. I read S3's actual response.json and first VM trial for each of all five modules. S3 is an AI-modeled medium-good learner, not a human. The limited app VM reports all five complete and six correct first choices each; I judged the written explanations independently. No peer reviews, browser observations, source-paper re-audit or learning-gain inference is used.

Summary

S3's actual prose is generally strong and source-specific, not merely long or cautious. All five numerical explanations are materially correct, all institutional choices are broadly feasible, and the late-record revisions use relevant evidence. The most important shortcomings are incomplete executable survey items in Everyone Says Democracy, an insufficiently identified process-legitimacy comparison in Civic Machine, and a stale stopping rule after the South pilot changes in Overloaded State. These passed the activity gates, demonstrating why completion is not a semantic grade.

Cross module findings

1. Thirty correct first choices and five complete runs coexist with incomplete open deliverables. That supports the app's explicit claim that gates establish attempts, not mastery; it does not imply the objective checks are defective.
2. S3 is much stronger in these AI-authored responses than a minimal-word submission: actual source details, exact records, numerical bases and legitimate alternative policies appear throughout. These are features of the submitted texts, not evidence of real human learning.
3. The recurring semantic weakness is causal-design inspectability: the treatment, population and outcome can drift between initial explanation, rival, late revision and final proposed test.
4. The most useful support is blank structured response organization plus a short teacher checkpoint. Giving the diagnosis, selecting a preferred policy or removing counterevidence would reduce the valuable challenge.
5. No preference or enjoyment conclusion is inferred from these written answers; that belongs to the separately labelled modeled learner reaction report.

Runtime

Completed

Yes

Correct first choices

6

Total checks

6

Failed gates**Semantic verdict**

Strong source reconstruction, trace and quantitative reasoning; good lawful political choice, with a weaker causal test of procedural legitimacy.

Strengths**Item 1****Location**

research.sourceA and research.sourceB

Evidence

The response distinguishes trained ordinary writers from professional mediators, statement preference from majority movement and minority text weight, and randomized strategy removal from observed votes.

Judgment

The two sources constrain different inferences rather than appear as decorative citations.

Item 2**Location**

investigation.revisedClaim and reason

Evidence

It locates the change between the March 10 rating questions and March 11 announcement, while retaining the useful clinic condition in the draft.

Judgment

This is a defensible narrowing, not automatic reversal or universal dismissal.

Item 3**Location**

brief.dataReason

Evidence

It constructs overlaps of 46 and 35, giving approval-only counts of 34 and 45 respectively, with the original 80/46/100 margins.

Judgment

Both linked patterns are feasible; the explanation also correctly preserves the citywide sampling limit.

Item 4**Location**

transfer.revision and test

Evidence

The response moves toward directed omission after AC-05, repairs the rated text rather than an already accurate announcement, and proposes the same-input/model comparison with equal length limits.

Judgment

This is real location-sensitive transfer. It still acknowledges that the instruction's causal contribution is not already isolated.

Concerns**Item 1****Severity**

P2

Location

decision.defense

Evidence

'Compare acceptance of identical voting rules among people receiving and not receiving their preferred route' is followed by the prediction that fair nomination improves acceptance among losers.

Judgment

A winners-versus-losers comparison with identical rules does not vary nomination rights, so it cannot identify their effect. The final nextStudy improves the design by varying nomination opportunities, but new options can change consequences and the group of losers.

Support without answer

Ask the learner to mark the treatment that actually varies, the outcome, the defined group and the source of comparable winning/losing outcomes. Require one threat from changing the available menu; do not nominate the best route or provide the expected result.

Item 2**Severity**

P3

Location

investigation.initialClaim/rival/prediction

Evidence

The initial account describes shared-criteria endorsement versus endorsement read as Green selection, then predicts lower route support than statement support.

Judgment

This is a useful same-response interpretive distinction, but the lower marginal total does not by itself explain why individuals endorsed the text or rule out mixed interpretations. The later request for individual belief is the needed improvement.

Support without answer

Require a one-sentence person-level observation that would distinguish interpretations and a record of which people answered both questions.

Quantitative assessment

No material arithmetic or base error. The two feasible joint patterns demonstrate understanding beyond the objective check.

Priority fix

Make the legitimacy-test design explicit and inspectable, including the consequences/menu confound, before awarding the causal-design criterion full credit.

P4 / Original faculty-role review

Independent AI political-scientist role: comparative public policy, administration, governance and causal inference; not a real professional participant

Independent initial content pressure test of all five version-3 builds through full configs, six staged learner packets and original art inspection. No peer or learner answers were read before this report was saved. The source summaries were assessed for internal coherence and appropriate use, not independently verified against complete primary papers. No native-browser, backend, recruited-human or learning-gain claim is made.

Ratings (1-5 judgments)

Content coherence	5
Source use	5
Political depth	5
Challenge	4
Evidence sufficiency	5
Teaching potential	5

Strengths

1. The strongest political move is the second dossier: an honestly represented and honestly reported consultation still fails to offer the eligible nominated alternative. This prevents the lesson collapsing into 'fix a misleading headline and democracy is repaired'.
2. Binding council authority, nomination rights, an access floor, annual operating costs, a separate process budget, deadlines and an independent complaint actor create actual tradeoffs. Green alone is ineligible, but Green with access, Blue and a dated deferral can each be defended.
3. The research pair distinguishes endorsement optimization, textual minority contribution, attitude movement and public authority rather than using AI trust as a political conclusion.
4. The numerical workshop demands two feasible joint-response patterns from the same marginal totals. That is a meaningful stretch beyond reading the longer bar.

5. Alder College changes the location of the failure: publication is accurate but the drafting/rating encounter omits the condition. This is genuine transfer, not a renamed transit question.

Problems

Item 1

Severity

P2

Location

CM decision defense and proposed legitimacy study

Evidence

The prompt asks for a comparison separating perceived legitimacy from receiving a preferred route. The frontier warns that opening the menu can itself change the winning route, but no learner-facing measurement protocol for legitimacy is developed.

Why it matters

A capable learner can write 'hold everything comparable' without confronting the fact that meaningful nomination changes actual options and potential consequences.

Recommended repair

Add an optional blank design card with separate columns for assignment, menu, chosen route, route preference, legitimacy measure and timing. Require the learner to mark the remaining mediation/selection issue and choose a feasible identification strategy; do not supply a political conclusion.

Item 2

Severity

P2

Location

Council vote choice field

Evidence

One answer field must trace eligibility, nominations, budget, publication and complaints while also specifying route, process cost, influence and unresolved disagreement.

Why it matters

The substantial political content can become a checklist hidden in one paragraph. A 20-word attempt threshold cannot tell an instructor which powers or costs the learner omitted.

Recommended repair

Keep the same required reasoning but provide a collapsible blank authority/cost ledger and an explicit instructor review checklist. Return comments on missing political links, not merely completion.

Item 3

Severity

P3

Location

Original civic poster art

Evidence

The inclusive round-table artwork communicates voice and shared transit planning, but concentrated clinic harm and the difference between drafting and binding power are not visually legible.

Why it matters

The visual mood is more consensual than the central distributional conflict.

Recommended repair

Retain the poster as atmosphere; add an exact, separately labelled route/clinic schematic and a small powers legend drawn from the dossier. Do not turn decorative routes or invented chart marks in the art into evidence.

What it teaches

A consultation output is not a mandate; representation, participation, agenda power, binding authority and concentrated rights have distinct records. Learners must defend a feasible institutional procedure, make a linked-data inference and adapt their diagnosis to a new institution.

What it still does not teach

It does not estimate the local effect of an AI objective, establish a universally legitimate decision rule, teach a complete theory of deliberative democracy, or dynamically simulate coalition bargaining after the council choice.

Does it stretch learners?

Yes. The own rival, marginal-versus-joint data problem and lawful decision under real costs require reasoning. The six recognition checks alone are not the stretch; the authored evidence and transfer are.

Scaffolding verdict

Necessary information is supplied and the method hints generally preserve the answer. A blank governance ledger would reduce navigation burden without reducing substantive demand.

Quantitative verdict

Strong: correct respondent base, percentage-point distinction and nonidentified overlap. The open two-pattern task is substantially harder than the check and should be separately reviewed.

Design content relation

The risograph campaign-print aesthetic fits participatory urban politics. Its use should remain atmospheric rather than suggesting the visible artwork is a true map or data display.

Overall verdict

I liked this as a political-governance practicum. It is ready for a carefully facilitated pilot, with semantic review of the cost/rights defense and a clearer optional study-design organizer.

Ranking

1. overloaded-state
2. civic-machine

3. democracy-alarm
4. persuasion-hearing
5. everyone-democracy

Cross-module context

1. This ranking reflects one explicitly modeled administrative/governance lens, not a population preference or quality league table. Everyone Says Democracy ranks fifth partly because of a specific repairable post-answer defect, not shallow political content.
2. All five have enough supplied political content for demanding medium-good learners: institutions appear as powers, incentives, legal constraints and public opportunities, not just topic decoration. They need human criterion-level review to test whether the content was actually learned.
3. The class tasks are rigorous evidence-reasoning practicums with branching/commitment features, not full dynamic political strategy games. Do not equate a saved choice with a simulated policy outcome.
4. The recurring weakness is deliverable bundling: long questions contain many distinct requirements in very few answer fields. Preserve sustained reading, but provide optional blank ledgers and mark which analytical links need revision.
5. Method hints generally preserve productive difficulty. Wrong-answer feedback must explain the actual selected error; a correct key plus stale affirmative feedback is not an adequate assessment system.
6. No learning gain, real enjoyment distribution, teen usability, native accessibility or backend robustness follows from these AI judgments. A facilitated human pilot should measure completion paths, quality of transfer, time/effort and delayed unfamiliar reasoning.

P4 / Assessment of S4 submitted answers

Causal methods and evidence quality

Separate semantic audit conducted only after the independent initial scholar review was saved. Read all five actual S4 authored response.json files, their saved staged commitments and five first-trial VM records. Commitment fields match the final response fields in all five; this preserves the recorded initial answers but is not proof of a human participant or unobservable reading behavior. The limited actual-JavaScript API/event VM reports five completed routes, thirty correct first-attempt objective choices and no locked gates. None of those facts establish prose mastery, native usability or real learning gain.

Overall verdict

S4 produced substantively strong, source-connected and numerically correct work rather than generic threshold-filling prose. The main residual weakness is causal-test feasibility and identification: several proposed experiments could conflict with the rights the same learner defends, or change more than the purported link. All-correct objective checks coexist with meaningful unfinished research-design reasoning.

Cross module findings

1. Thirty correct first objective choices coexist with several incomplete causal designs. This is direct evidence that automated completion and recognition scores are not a sufficient semantic assessment, even for this strong AI-authored profile.

- 2. The learner does actual numerical reasoning and changed-location transfer in all five. It often supplies exact costs, authority and protected outcomes; there is no basis for claiming the tasks are empty or impossible from this authored record.
- 3. Rights are reasoning constraints for research design too. A learner can correctly defend nomination and complaint rights, then inadvertently violate them in the control condition. This is an excellent further stretching question, not grounds for making the lesson easier.
- 4. Matched/phased comparative designs need explicit assignment and selection assumptions. 'Hold comparable' is useful starting language but not sufficient causal identification.
- 5. The Everyone Says Democracy stale wrong-answer messages were not exposed by S4 because its first choices were all correct. That specific content defect remains important despite a clean successful route.
- 6. S4 is an AI modeled learner with unusually articulate, source-aware prose. These responses cannot establish the typical performance, effort, vocabulary difficulty or enjoyment of real medium-good students.

Runtime

Completed	True
Objective first attempt	6/6
Failed gates	0

Semantic verdict

Strong institutional and quantitative understanding; causal design needs revision, and the first rival is more an alternative interpretation than a fully traced endorsement process.

Strengths

Item 1

Location

research.crossInference and investigation.reason

Evidence

S4 keeps minority text contribution distinct from protected interests, then locates the changed claim specifically between CM-05/06 and CM-08. It proposes clerical misunderstanding versus strategic mandate-making instead of inferring the office's motive.

Judgment

Shows real source use and artifact localization, not merely 'more evidence is needed'.

Item 2

Location

decision.choice/defense

Evidence

Blue costs €1.74m and preserves nine-minute clinic access; the learner names the €18,000 sum of €9,000, €6,000 and €3,000 safeguards, keeps Green plus service eligible and rejects Green alone. It acknowledges sacrificing six further minutes of possible commute benefit.

Judgment

A reasoned feasible political preference with authority, right, cost and opportunity cost; disagreement with choosing Blue would not be a grading error.

Item 3**Location**

brief.dataReason

Evidence

Two feasible overlap patterns are 46/34 and 26/54, with the practice linked fact yielding 35 overlap and 45 approval-only respondents.

Judgment

Correctly distinguishes marginal totals, linked counts and population limits.

Item 4**Location**

transfer.revision

Evidence

'The truthful 72-of-90 announcement is not the transit headline error.' S4 locates the new omission in the display, revises toward instructed separation and keeps the objective's causal effect unidentified.

Judgment

Genuine unfamiliar-case adaptation rather than reproducing the old remedy.

Limitations**Item 1****Severity**

P2

Location

decision.defense and brief.nextStudy

Evidence

The learner proposes randomly offering full versus restricted eligible menus even though the governing charter protects eligible nominations.

Issue

If this is a binding public process, the restricted arm could deliberately violate the nomination right.

'Comparable consultations' also does not isolate menu voice from changed route options or preferred consequences.

Support preserving demand

Ask the learner to redesign as a nonbinding vignette or compare two lawful nomination-support procedures. Require an explicit ethics/authority statement, chosen route record and identification limit; do not provide the answer.

Item 2**Severity**

P2

Location

transfer.test

Evidence

Endorsement instructions are contrasted with instructions retaining all submitted conditions while model, input and length allowance are fixed.

Issue

This changes both the objective framing and an explicit condition-retention requirement. It is a legitimate instruction-package test, but not a pure objective-effect isolation.

Support preserving demand

Require the learner to name the estimand as a package effect or design separate objective and display/retention factors. Keep the requirement for rival predictions.

Item 3**Severity**

P3

Location

initial.investigation.rival.and.prediction

Evidence

'The statement could instead express a specific route choice' predicts closer agreement between items but does not name who produces that relationship and by what operation.

Issue

Useful initial alternative interpretation, but weaker as a political process than the later clerical-versus-strategic rival.

Support preserving demand

Ask for one actor, operation and same-outcome causal link for each initial rival; retain the saved imperfect first account.

Instructor follow up

Which part of your proposed study can be randomized without denying a charter right, and what causal claim would your redesigned comparison identify?

P5 / Original faculty-role review

AI political scientist role: political concept formation, research-method reasoning, and teaching/transfer assessment

Independent inspection of all five latest version-3 content configurations, staged learner materials and original artwork. This is an AI analytical role, not a credentialed human scientist. I have not recruited students, measured learning gains, used a native browser or independently reread all primary papers. Source judgments concern the internal accuracy, coherence and boundaries of the explanations supplied by the builds; the citations require a separate primary-paper audit. Initial report written before inspecting learner answers or peer reports.

Ratings (1-5 judgments)

Content coherence	5
Source use	5
Political depth	4
Challenge	4
Evidence sufficiency	5
Teaching potential	4

Strengths

1. The dated sequence requires close reading rather than a generic pro/anti-AI attitude. CM-05 preserves the clinic condition, CM-06 asks about separate objects, and CM-08 changes the public object of backing. The political error has an identifiable institutional location.
2. The later council problem removes the false headline and leaves an actual choice: Green plus access, Blue or a bounded deferral. Budget feasibility, the clinic right, eligible nominations and an independent complaint actor constrain rhetoric about inclusion.
3. Source construction forces Tessler's endorsement/text/attitude outcomes and Lin's assigned-advocacy attitudes to have different evidentiary jobs. It does not confer authority from either study.
4. The Alder College transfer reverses the easy diagnosis: the announcement is accurate, but the drafted/rated object omits a qualification. Applying the original headline answer mechanically would fail.

Problems

Item 1

Severity

P2

Location

revision.decision.fields.defense and final brief

Evidence

One decision field simultaneously asks for a criterion comparison, both studies, a legitimacy-versus-outcome comparison, rival predictions and an institutional safeguard, with only a 20-word recording minimum.

Why it matters

A polished paragraph can omit the difficult comparison while satisfying recording. The learner may not know which element is incomplete, and the teacher must disentangle several performances.

Recommended repair

Retain the full demand but split the visible response space into criterion/tradeoff, source contribution, and discrimination/safeguard subparts. Show a neutral completeness checklist, not a model conclusion; score those dimensions separately in human review.

Item 2**Severity**

P2

Location

dataWorkshop.reasonPrompt

Evidence

The learner must construct two linked-answer patterns from marginal 80/46/100 totals; no blank two-by-two workspace accompanies that generative task in the supplied content.

Why it matters

This is the best numerical stretch in the case, but the bottleneck can be organizing cells rather than reasoning about political inference.

Recommended repair

Offer an optional empty cross-tab with editable margins, labeled approval/nonapproval and favor/nonfavor, plus a consistency check that totals match. Do not fill any cell, calculate the answer or reveal the feasible overlap range before submission.

Item 3**Severity**

P3

Location

checks.scope.hint

Evidence

The hint asks what was varied and measured even though this item concerns two observational question totals, not an assigned treatment.

Why it matters

Generic experimental vocabulary can make an otherwise careful learner search for a nonexistent intervention.

Recommended repair

Use a case-neutral measurement prompt here: identify each rated object, the included respondents, and whether the record links answers within people. Keep causal variation prompts for genuinely experimental items.

What it teaches

Representation is not authorization; endorsement is not a policy vote; inclusion in text is not agenda influence. Students must locate actors, trace words, distinguish marginal from linked responses and defend a legally feasible distributional choice.

What it still does not teach

It does not estimate the causal effect of the local instruction, establish population preferences, or demonstrate that the learner can inspect an unscripted real consultation without these prompts. Deliberative quality and actual implementation remain proposed rather than measured.

Does it stretch learners?

Yes, especially the second decision and altered transfer. The six checks test recognition of distinctions, while the written rival, test and tradeoff carry the more demanding generative work. Completion cannot demonstrate their quality.

Scaffolding verdict

The glossary and unrelated museum examples support method without supplying the city answer. Preserve them. Make the complex response requirements locatable rather than lowering the reading load or revealing the causal diagnosis.

Quantitative verdict

Strong on objects, bases and joint-versus-marginal evidence. Exact totals and the separate linked assumption allow a real check. Add an optional blank organizing tool; later add an unfamiliar unlabeled graph for assessing independent graph literacy rather than only calculation beside an explanatory guide.

Design content relation

The blue/coral risograph meeting, transit lines and civic buildings make consultation tangible. The image suggests harmonious collective deliberation, so accompany it with a short optional prompt about who is absent or who controls the route agenda. This is artwork interpretation, not a rendered-UI usability finding.

Overall verdict

A strong source-to-institution lesson with adequate evidence and productive difficulty. It is more a demanding guided inquiry than a strategy game; the next improvement is sharper assessment of its authored reasoning, not easier content.

Ranking

1. democracy-alarm
2. overloaded-state
3. civic-machine
4. persuasion-hearing
5. everyone-democracy

Cross-module context

1. This ranking concerns current teaching readiness under my concept/method lens, not real enjoyment or measured efficacy. Everyone would move upward after its feedback defect is repaired.
2. All five provide sustained reading, political actors, constrained decisions and generative transfer. They are not empty multiple-choice games. Their strongest challenge resides in the authored rival, bounded

positive finding, feasible decision and discriminating test.

- 3. The recurring weakness is compound response architecture. Split and label dimensions while retaining the whole intellectual demand. Word minima are explicitly participation gates and should stay distinct from semantic quality.
- 4. The hints I inspected generally direct method rather than supply political answers. Do not make them more revealing; target denominators, contrasts, units, locations and missing records. Post-submission feedback needs its own complete wrong-option audit.
- 5. The five transfers are meaningfully altered cases, but remain scaffolded near transfer. A delayed, less-cued unfamiliar dossier is needed before claiming independent generalization.
- 6. Graph/table activities are concrete and politically relevant. Claims of general graph literacy require unfamiliar encodings or mislabeled displays, not only recognition of formulas beside a worked example.
- 7. A human pilot should retain the challenging reading, record where a learner becomes stuck, separately score content/process/quantitative/transfer performances, and include a delayed new-case task. No completion percentage or AI persona rating is evidence of real learning gain.
- 8. The artwork is relevant and differentiated. It could teach political visual rhetoric through optional interpretation prompts, while remaining separate from documentary evidence. I did not assess contrast, mobile rendering, keyboard behavior or assistive technology.

P5 / Assessment of S5 submitted answers

Political concepts, teaching and transfer

Separate semantic audit written after the independent p5 review was saved. I read S5's five actual submitted response.json files and all five corresponding runtime/trials/s5/*-first.json records. This audits AI-modeled learner texts, not human learning or enjoyment. The runtime is an API/event VM execution of original app JavaScript, not native-browser use. Each inspected first trial completed with six correct objective choices and no recorded failed gate. Those flags establish completion and keyed-choice success only; this report separately judges authored content.

Cross module conclusion

S5 produces strong, source-grounded and often genuinely generative reasoning in all five. This is useful evidence that the material contains enough information for this modeled response, not that medium-good real students will produce it unaided or enjoy it. The remaining weaknesses show why human semantic review is essential: alternative unknown facts or heterogeneous components can be mislabeled causal rivals; causal comparisons may condition on later outcomes; a weekly excess can be overstated as stock growth; a plan may omit ready-to-use items; and useful policy tests can fail to discriminate proposed mechanisms. None was exposed by the keyed 6/6 choice results. Preserve challenge and use neutral comparison/response organizers rather than give more of the answer.

Runtime

Completed	True
Correct checks	6
Total checks	6
Failed gates	0

Semantic verdict

Strong source, documentary and institutional reasoning; the initial rival is not yet a causal rival and the proposed legitimacy comparison needs design refinement.

Strengths**Item 1****Location**

research.sourceA/sourceB/crossInference

Example

S5 distinguishes Tessler's statement comparison, majority-aligned movement and minority text from Lin's opposing advocates, strategy changes and reported intentions.

Judgment

The papers have separate jobs rather than decorative citations; no city count is attributed to them and no voting effect is invented.

Item 2**Location**

investigation.reason

Example

S5 cites CM-FINAL-2, CM-06 and CM-08 and says it cannot explain why people endorsed the text from those records alone.

Judgment

This is a warranted positive finding about the public object change plus a specific causal limit, not empty caution.

Item 3**Location**

decision.choice/defense

Example

Blue costs 1.74 million and provides a nine-minute clinic walk; S5 budgets 9,000 + 6,000 + 3,000 process euros, includes Green plus access and Blue, and assigns powers to the council, office, residents and clerk.

Judgment

The choice is feasible and a real tradeoff: fewer commute gains versus an access service at the ceiling. It preserves the charter and nomination right. Choosing Blue is legitimate, not the only correct political preference.

Item 4**Location**

transfer.revision/test and brief.dataReason

Example

S5 locates Alder's omission in drafted/rated text despite truthful publication, qualifies instruction causation, and constructs overlap patterns of 46 and 30 Green favorers within the 80 approvers.

Judgment

The transfer is genuinely adapted; both linked tables fit the margins and yield 34 versus 50 approvers not favoring Green. Numerical interpretation exceeds selecting the keyed option.

Limitations**Item 1****Severity**

P2

Location

investigation.initialClaim/rival

Example

The rival is that the final statement might explicitly recommend Green.

Judgment

That is a competing possibility about the unknown text, not a process causing the same endorsement outcome. The later reason develops a better rival about how raters interpreted the item, but the sealed initial response is weaker than the demanding causal-rival criterion.

Item 2**Severity**

P2

Location

decision.defense and brief.nextStudy

Example

S5 proposes comparing nomination access while measuring acceptance within preferred-route winners and losers.

Judgment

Useful separation of outcome satisfaction and procedural acceptance, but nomination rights can alter which route wins. Conditioning on winners/losers after treatment can distort a causal comparison. It needs fixed or independently varied outcomes, a yoked design, or an explicit limitation.

**Support retaining difficulty**

Ask 'Do both accounts explain the same observed endorsement, through different processes?' and require the student to define when winner/loser status is determined. Offer a blank outcome/process matrix, not a supplied rival or preferred council rule.

Mastery claim allowed

This text demonstrates several strong performances in this supplied AI walkthrough, not established human mastery, learning gain or independent far transfer.