Online Resource 1: Supplementary Material

Governing LLM decision authority in tourism: A bounded generative agency framework for multi-stakeholder itinerary planning

Information Technology & Tourism

S1 Data preparation and aggregation

S1.1 Normalization and aggregation

The formal SilverRoute model uses a fixed local evidence package for the four Ya'an jurisdictions. Source records retain original-language place names and provenance labels. Missing public tags are coded as unknown.

For each benefit-type source indicator x_{ij}, a robust min-max transformation based on the 5th and 95th percentiles is used:

$$z_{ij}^{+} = \mathrm{clip}\left(\frac{x_{ij} - P_{05,j}}{P_{95,j} - P_{05,j}}, 0, 1\right). \tag{S1}$$

For cost-type indicators, $z_{ij} = 1 - z_{ij}^{+}$, so higher values always indicate more favorable conditions. The implementation uses 0.50 as a neutral fallback for a constant indicator or an unavailable evidence group. Provenance records identify these values as neutral imputations.

The 21 spatial indicators are grouped into six point-of-interest-level evidence dimensions. Cultural experience, mobility access, service convenience, and destination synergy use the row mean of available normalized indicators. Health and safety and age-friendly comfort use a conservative aggregation:

$$z^{cons} = 0.55\bar{z} + 0.45\min(z), \tag{S2}$$

which gives lower component values greater influence in these two dimensions. The same rule is used when health-and-safety and age-friendly-comfort scores are aggregated from stops to the route level. Schedule robustness and route efficiency are generated at the route level.

This structure separates point-of-interest-level evidence from route-level operational criteria. Cultural, mobility, service, and destination dimensions use mean aggregation, whereas health and age-friendly dimensions retain greater sensitivity to weak component values.

S1.2 Provenance audit and evidence limits

The runtime data manifest contains 21 criterion assets and three verified GeoPackage vector packages for facilities, transport, and study-area boundaries. SHA-256 checksums are stored for the packaged sources. Among the 21 criterion assets, 19 are recorded as `provided_precomputed` and two—age-friendliness of cultural attractions and cultural appeal to older adults—as `provided_author_scoring`. The package records these undated layers as a fixed research snapshot.

Table S1 Aggregation of Level-1 route dimensions

Level-1 dimension	Aggregation rule
Cultural experience	Mean of normalized Level-2 inputs
Health and safety	$0.55\bar{z} + 0.45\min(z)$
Mobility access	Mean of normalized Level-2 inputs
Age-friendly comfort	$0.55\bar{z} + 0.45\min(z)$
Service convenience	Mean of normalized Level-2 inputs
Schedule robustness	Route-level schedule-stress calculation
Destination synergy	Mean of normalized Level-2 inputs
Route efficiency	Route-level distance and time-reserve calculation

Table S2 Baseline stakeholder subjective weights and evidence reliability

Criterion	Senior Traveler	Destination Steward	Tourism Operator	Evidence reliability ρ_j
Cultural experience	0.22	0.18	0.20	0.90
Health and safety	0.23	0.19	0.10	0.90
Mobility access	0.12	0.10	0.15	0.82
Age-friendly comfort	0.20	0.11	0.08	0.72
Service convenience	0.06	0.10	0.14	0.82
Schedule robustness	0.08	0.10	0.15	0.78
Destination synergy	0.02	0.17	0.08	0.72
Route efficiency	0.07	0.05	0.10	0.98

S2 Stakeholder preference weights

The three stakeholder profiles are formal decision representations. Their qualitative priorities were specified from the project's stakeholder-construct documents and translated into normalized research-design prior weights. Each vector remains separate until candidate routes have been evaluated. Accordingly, sensitivity analysis evaluates the weight architecture and bargaining-power assumptions.

The Criteria Importance Through Intercriteria Correlation component is computed from the hard-feasible route matrix. Before fusion, each objective weight is multiplied by an evidence-reliability coefficient ρ_j and renormalized. The baseline objective share is $\lambda = 0.30$:

$$w_{s,j}^{F} = (1-\lambda)w_{s,j}^{S} + \lambda\widetilde{w}_{j}^{C}. \tag{S3}$$

Table S3 Baseline decision-boundary parameters

Stakeholder	Bargaining power	Reservation ratio	Relative veto thresholds	Absolute floors
Senior Traveler	0.50	0.75	Health and safety 0.80; age-friendly comfort 0.75; schedule robustness 0.55	Health and safety 0.30; age-friendly comfort 0.30; schedule robustness 0.05
Destination Steward	0.28	0.60	Health and safety 0.70; cultural experience 0.70; schedule robustness 0.45	Health and safety 0.25; cultural experience 0.20; schedule robustness 0.05
Tourism Operator	0.22	0.60	Route efficiency 0.75; schedule robustness 0.45	Schedule robustness 0.05

Objective-weight shares of 0.20, 0.30, 0.40, and 0.50 are included in sensitivity analysis. Entropy weighting is also evaluated as a robustness alternative across the same shares, while the correlation-based method remains the baseline objective-weighting method.

The profiles therefore encode distinct role priorities rather than three versions of one common score. The Senior Traveler places greater weight on health, safety, and age-friendly comfort; the Destination Steward gives relatively more weight to destination synergy; and the Tourism Operator emphasizes mobility, service support, schedule robustness, and route efficiency.

S3 Stakeholder protection and bargaining

The stakeholder layer distinguishes three quantities. Criterion preference weights determine how a stakeholder evaluates a route. Veto and reservation rules determine whether that route is acceptable. Bargaining power determines influence over the final compromise within the agreement set.

For protected criterion j, stakeholder s uses the hybrid veto threshold

$$T_{s,j} = \max(a_{s,j}, q_{s,j} z_j^{ideal}), \tag{S4}$$

where $a_{s,j}$ is the absolute floor and $q_{s,j}$ is the relative fraction of the attainable criterion ideal. The Tourism Operator's route-efficiency boundary uses the relative veto because this criterion is normalized against the request-specific time budget.

The reservation point is

$$d_s = \eta_s U_s^{ideal}, \qquad U_s^{ideal} = \max_r U_s(r), \tag{S5}$$

where η_s is the reservation ratio shown in Table S3. A route enters the agreement set after satisfying all structural hard constraints, stakeholder vetoes, and stakeholder reservation utilities. Weighted Nash bargaining is then applied within that set.

Table S4 Baseline route-generation settings

Parameter	Baseline value
Default trip budget	6 h
Default stop range	3-5 stops
Return to start	Yes
Average vehicle speed for static modeling	35 km/h
Default departure hour	09:00
Default visit time	55 min
Heritage visit time	65 min
Age-friendly transition allowance	8 min/stop
Default meal/rest break	45 min
Robustness reserve target	120 min
Feature preselection cap	220 sites
Candidate site pool	48 sites
Beam width	300
Maximum retained route candidates	120
Road snap tolerance	8 m
Road simplification tolerance	3 m
Disconnected-network distance factor	1.25
Route-efficiency distance scale	70 km

S4 Route and implementation parameters

S4.1 Baseline route settings

The multi-stakeholder recall score used before route expansion is

$$S_i^{recall} = 0.70\overline{S}_i + 0.30 \max_s(S_{i,s}) + B_i, \tag{S6}$$

Here, B_i combines a preferred-keyword bonus of 0.04 per match, capped at 0.12, with a short-trip proximity bonus. The proximity term is $0.18\exp(-d_i/12)$ for budgets up to 4 h and $0.08\exp(-d_i/20)$ for budgets above 4 h and up to 6 h, where d_i is straight-line distance from the origin in kilometers; longer budgets receive no proximity bonus. Candidate assembly first preserves nearby alternatives, then reserves stakeholder-specific recall tranches before filling the remaining pool by the combined score. Formal route distance is still computed on the noded road graph.

S4.2 Schedule-stress scenarios

Schedule robustness is evaluated under the fixed scenario set below. These scenarios model operational stress.

For stress scenario k, let b_k denote the remaining time buffer and let $I_k = 1$ when both the simulated operating windows and the trip deadline are satisfied, and 0 otherwise. The reserve target is $R = \min(120, \max(45, 0.25T_{budget}))$ minutes. Scenario adequacy and schedule robustness are

Table S5 Schedule-stress scenarios

Scenario	Probability	Travel-time multiplier	Visit-time multiplier
Typical	0.40	1.00	1.00
Mild delay	0.30	1.10	1.05
Peak/rain	0.18	1.25	1.08
Adverse	0.09	1.40	1.12
Severe stress	0.03	1.60	1.15

Table S6 Decision authority under bounded generative agency

Decision stage	Large language model (LLM) role	Formal authority
Requirement interpretation	Propose structured fields	Confirmed decision contract
Evidence and spatial computation	Explain verified evidence	Frozen evidence + spatial optimization core
Preference and acceptability	Explain computed values	Weights, vetoes, and reservations
Bargaining and action	Explain the outcome	Weighted Nash + action protocol

$$a_k = I_k \,\mathrm{clip}\left(\frac{b_k}{R}, 0, 1\right), \qquad R_{sched}(r) = \sum_k p_k a_k, \tag{S7}$$

where p_k is the scenario probability. Thus, schedule robustness reflects both scenario feasibility and the remaining contingency buffer rather than a binary pass rate alone.

Route efficiency is calculated as

$$E(r) = 0.65 \exp(-D_r/70) + 0.35 \max\left(0, 1 - T_r/T_{budget}\right), \tag{S8}$$

where D_r is route distance in kilometers and T_r is modeled total route time.

S4.3 Interface and spatial audit

The deployed interface requires confirmation of decision-relevant fields before the deterministic planner is executed. This keeps natural-language interpretation separate from formal decision variables. Additionally, the editable QGIS project preserves selected routes, candidate alternatives, facilities, origins, and contextual layers for spatial inspection. These audit layers support traceability, while the deterministic planner governs formal ranking.

S5 Validation registry

The primary and supplementary validations are organized by the claim they test. Additionally, Table S7 links each design to its main purpose.

For paired bounded-authority versus evidence-grounded inference, repeated LLM observations are averaged within scenario-condition so that the scenario remains the independent unit. Effect intervals use 10,000 bootstrap resamples and paired randomization tests use 20,000 iterations (analysis seed 20260808). Traditional-method paired differences use two-sided Wilcoxon signed-rank tests.

S5.1 Supplementary validation outcomes

Decision-trace evidence coverage is the proportion of expected evidence identifiers present in the generated explanation, with the expected set defined as the selected stops plus five formal evidence records. It therefore measures trace completeness rather than raw-source completeness. The 91.25% exact-route stability is calculated across correlation-based and entropy objective weighting with objective shares of 0.20, 0.30, 0.40, and 0.50.

Mechanism ablation shows that stakeholder protection is driven primarily by the pre-aggregation gates. Full Nash, mean utility within the same agreement set, and maximin within that set preserved 100% protocol eligibility. Removing the veto gate reduced eligibility to 68.75%, while removing the full acceptability gate reduced it to 66.67% (main text, Fig. 4a).

In the 40 baseline recommendable scenarios, Nash improved minimum stakeholder utility over mean utility in 7.5% of cases, while 80.0% were ties; the average Nash-minus-mean difference was -0.001875. Under the strongest conflict stress ($\gamma = 3$), mean minimum utility was 0.51718 for Nash, 0.51823 for mean utility, and 0.51911 for maximin. These scalar comparisons are diagnostic because Nash serves primarily as a coordination rule within the agreement set.

Additionally, a controlled decision-rights stress test separated action stability from route identity. Candidate-scope perturbations preserved protocol action alignment, although exact-route alignment fell to 91.58% and 79.19%. By contrast, bypassing the protocol at the final-action stage produced recommendations in every case and a 16.67% forced-recommendation rate. These values quantify simulated mechanism stress.

Route-search validity was assessed against exact enumeration and compared with classic heuristics, metaheuristics, and random search. SilverRoute matched all 12 decisions and produced the lowest mean log-Nash regret (1.98×10^{-16}; Fig. S1).

Preference uncertainty was assessed through bounded Dirichlet perturbations at concentrations of 10, 20, and 50. Action stability remained 100%, whereas exact-route stability ranged from 84.01% to 91.74% (Fig. S2).

Across fallback factors of 1.00, 1.25, 1.50, and 2.00, action stability remained 100%. Exact-route stability was 91.67% at 1.00 and 100% at the other factors (Fig. S3). Finally, an independent holdout stress used 1.80× travel time and 1.25× visit time, exceeding the formal schedule scenarios. Eight of 40 recommended baseline cases remained feasible (20%). This test identifies an extreme operational boundary.

Table S7 Validation registry

Validation	Design	Main purpose
Formal factorial validation	48 scenarios varying origin, trip budget, stop structure, and traveler requirements	Protocol completion; recommendation versus abstention
Agency Boundary Experiment	Same 48-scenario family under deterministic, unbounded LLM, evidence-grounded LLM, and bounded-authority conditions; three persisted LLM realizations per scenario-condition	Isolate decision-authority effects beyond factual grounding
Traditional multicriteria decision analysis comparison	Weighted overlay, the technique for order preference by similarity to an ideal solution, and multicriteria optimization and compromise solution on the same hard-feasible matrices and criterion information	Compare agreement eligibility and stakeholder-utility outcomes
Mechanism ablation	48 scenarios with 11 fixed-candidate governance variants	Isolate veto, reservation, acceptability-before-aggregation, and final coordination rules
Conflict stress	Preference sharpening with $\gamma \in \{1, 1.5, 2, 3\}$	Test when Nash coordination becomes decision-relevant
Governance sensitivity	Bargaining power, reservation ratios, veto levels, objective-weighting method and share, and timing assumptions	Test stability to normative and operational parameters
Monte Carlo robustness	500 parameter draws over 40 formally recommendable scenarios	Joint parameter robustness
Abstention-boundary analysis	Governance-threshold surface, 52 time-boundary cases, and counterfactual minimum-repair cases	Distinguish hard infeasibility, completed abstention, and recommendation; identify explicit repair pathways
Search-depth validation	Wider candidate/search settings plus reduced exact-enumeration instances	Assess approximate beam search against wider search and exact enumeration
Classic route-algorithm comparison	Exact enumeration, nearest neighbor, preference-distance greedy, random search, simulated annealing, genetic algorithm, and SilverRoute beam search on reduced instances	Route-search validity and performance comparison
Preference uncertainty	Bounded synthetic perturbations around declared stakeholder weights	Test parameter uncertainty around declared weights
Disconnected-network fallback sensitivity	Distance penalties 1.00, 1.25, 1.50, and 2.00	Test dependence on fallback assumptions
Within-area transportability	16 county-constrained scenarios	Test protocol operation across different local choice sets within Ya'an
Age-friendly proxy validation	12 wheelchair/mobility-aid proxy scenarios	Test differentiated proxy screening

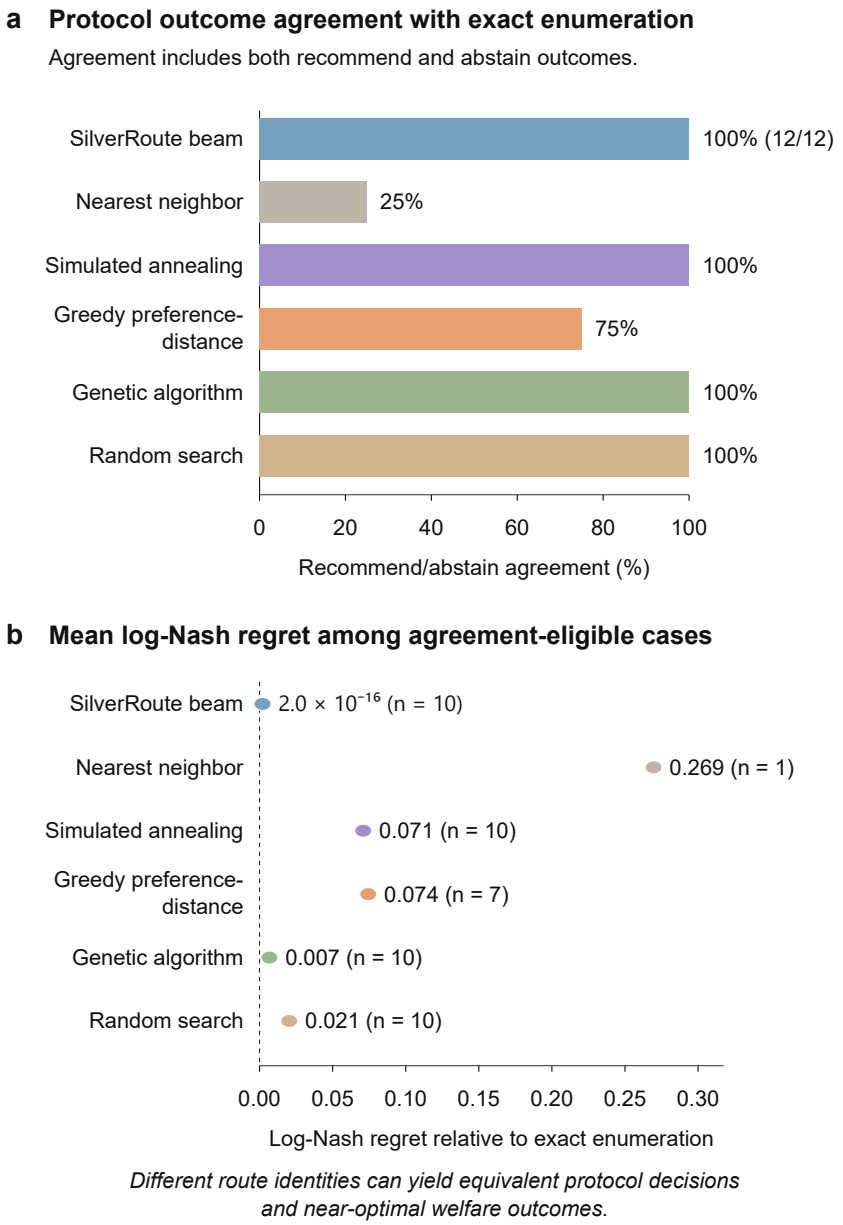


Fig. S1 Route-search validity. Panel a shows recommend-or-abstain agreement with exact enumeration. Panel b shows mean log-Nash regret among cases in which each algorithm produced an agreement-eligible route; the valid sample size therefore varies by algorithm

Table S8 Core formal validation results

Outcome	Result
Valid protocol completion	48/48
Protocol-eligible recommendations	40
Responsible abstentions	8
Hard-constraint satisfaction among recommendations	100%
Decision-trace evidence coverage	100%
Pareto-efficient selections	100%
Objective-weight sensitivity: exact-route stability	91.25%
Mean pairwise stakeholder Spearman rank correlation	0.9309

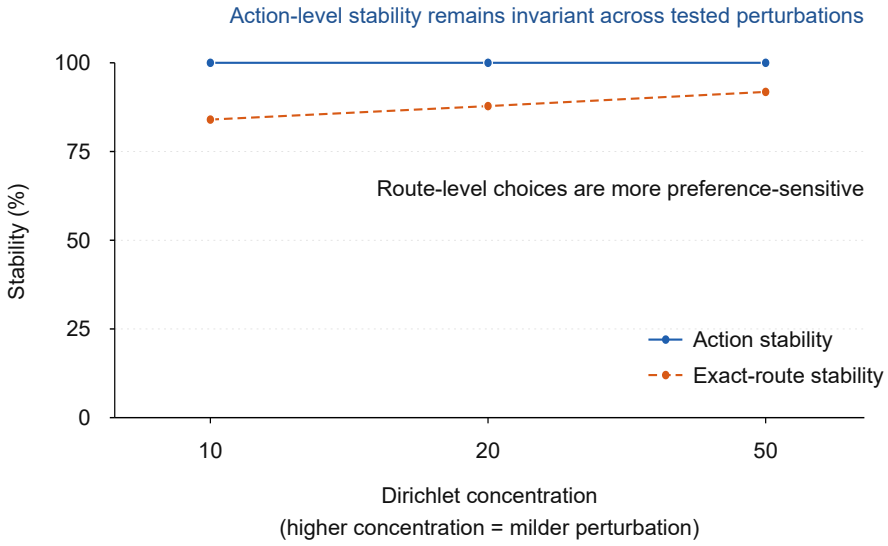


Fig. S2 Decision robustness under preference uncertainty. Action stability treats route variants with the same decision as equivalent; therefore, 100% action stability can coexist with route variation

S6 Decision trace of the illustrative SilverRoute itinerary

The illustrative case used a confirmed six-hour, four-stop contract for 20 active older adults. It required vehicle-supported travel and return to Ya'an Railway Station. Table S9 reports the auditable decision trace.

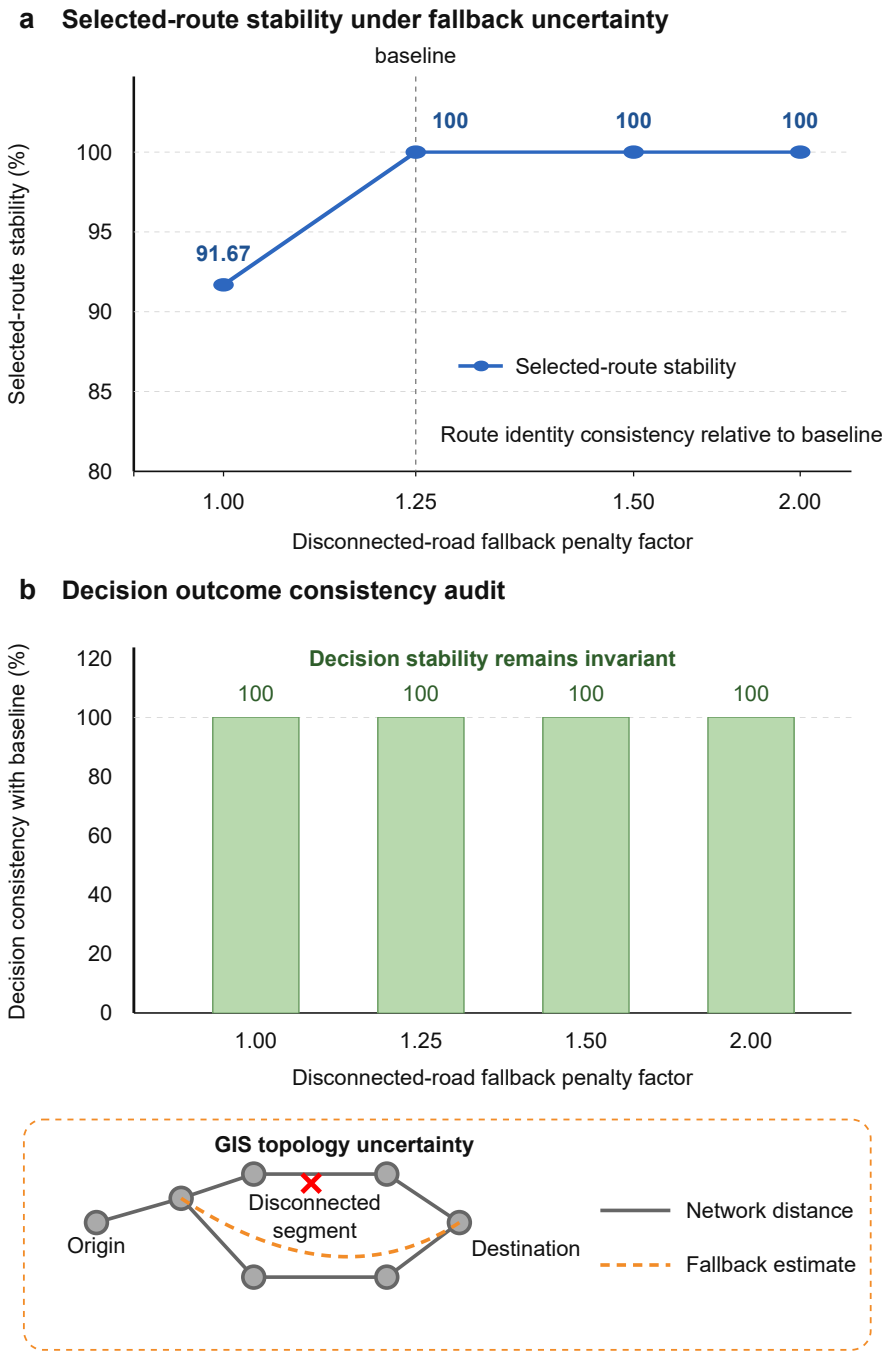


Fig. S3 Road-network fallback robustness. Panel a reports exact-route stability; Panel b reports action consistency across penalty factors

Table S9 Decision trace of the illustrative itinerary

Decision element	Auditable result
Confirmed contract	Six hours; four stops; vehicle-supported travel; return to origin; medical, cultural, walking, activity, and rest requirements
Candidate screening	120 candidates; 102 hard-feasible; six Senior-Traveler veto exclusions; 96 agreement-eligible routes
Formal selection	`route-069f6ede5e`; Pareto-efficient in both the hard-feasible and agreement sets; highest normalized gain for all three stakeholders
Route outcome	Ya'an Sichuan Opera Art Museum–Ya'an Eco-museum–Gao Yi Que–"Galloping Horse Treading on a Flying Swallow"; 5.22 km; 305.95 min; 54.05-min budget margin
Stakeholder utilities	Senior Traveler 0.630; Destination Steward 0.649; Tourism Operator 0.648
Spatial audit	QGIS 3.40.8; 13 layers loaded successfully
Evidence basis	Fixed local routing evidence and age-friendly spatial proxies

In this case, all three stakeholder ideals identified the same route. Accordingly, weighted Nash confirmed a low-conflict consensus. The conflict experiment provides the separate test of bargaining under divergent preferences.

S7 Reproducibility

The experiment package persists structured contracts, route candidates, criterion matrices, stakeholder utilities, bargaining outputs, provenance metadata, figure-source data, and random seeds where applicable. The LLM application programming interface is configured through an environment variable, with credentials excluded from the reproducibility package.

Interpretive boundaries are reported beside the corresponding evidence in Sections S1, S2, S5, and S6. Accordingly, the package supports protocol mechanics, computational robustness, within-area transportability, and proxy-based screening.