Multi-Stakeholder LLM Alignment: Decomposing Estimation from Aggregation

Anonymous ACL submission

Abstract

Multi-stakeholder tasks require one output to satisfy users with conflicting preferences. Holistic LLM judges conflate utility estimation and utility aggregation, yielding unstable implicit weights. We show empirically and theoretically that this aggregation-specific *weighting noise* can create large score shifts when stakeholder satisfaction is dispersed; in our experiments, these weight-induced shifts also increase with stakeholder count. We propose DECOMPR: counterfactual-calibrated weights are fixed from query structure before candidate scoring, while per-role utilities are estimated independently, removing candidate-dependent weight drift and reducing estimation noise.

1 Introduction

Large language models (LLMs) have achieved remarkable success in code generation, mathematical reasoning, and open-ended dialogue (OpenAI, 2023; Qwen Team, 2025; Guo et al., 2025). Much of this progress, however, targets scenarios where the agent serves a single user whose preferences define a clear optimization objective. Many real tasks—group travel planning, multi-party negotiation, committee resource allocation—instead require one agent to produce a single joint response for multiple stakeholders whose preferences may conflict.

Aligning LLMs for such tasks often relies on reinforcement learning (Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025), which requires a reward signal specifying what counts as a good joint response. Since these tasks are open-ended and lack verifiable ground truth, the reward is typically supplied by LLM-as-Judge (Zheng et al., 2023) or learned reward models (Ouyang et al., 2022). LLM judges are known to exhibit evaluation biases and instability even outside multi-stakeholder settings, including position bias in pairwise comparisons (Zheng et al., 2023; Wang et al., 2024c), adversarial vulnerabilities in absolute scoring (Raina et al., 2024), and intra-rater score inconsistency across repeated runs (Haldar and Hockenmaier, 2025). Multi-stakeholder evaluation compounds this problem: the judge must perform both *estimation* (how well is each stakeholder served?) and *aggregation* (how should competing utilities be weighted?) in a single pass, thereby producing implicitly assigned aggregation weights that can shift across calls.

As Figure 1 illustrates, repeated evaluations of the same plan can instantiate different implicit weights and produce inconsistent rewards. We refer to this as *weighting noise*: an aggregation-specific error in which the judge's implicit stakeholder weights shift across evaluations, reweighting utilities. Its score impact is amplified when stakeholder satisfactions are dispersed and may increase with stakeholder count (§3). Our theory shows how such shifts can corrupt within-group response ranking and mislead GRPO-style policy gradients (§4). We address this with DECOMPR, whose key mechanism is *counterfactual-calibrated aggregation*: each stakeholder's weight is fixed from query structure as a proxy for expected sacrifice, while per-role grounded evaluation estimates utilities independently.

In summary, we formalize the multi-stakeholder alignment problem, show empirically and theoretically that LLM-judge reward inconsistency in this setting arises in part from aggregation-specific weighting noise, and propose DECOMPR to address this problem through counterfactual-calibrated estimation–aggregation separation.

2 Related Work

2.1 Reward Evaluation for Open-Ended Tasks

LLM-as-Judge (Zheng et al., 2023) enables reward estimation without verifiable ground truth, but exhibits well-documented inconsistencies: position

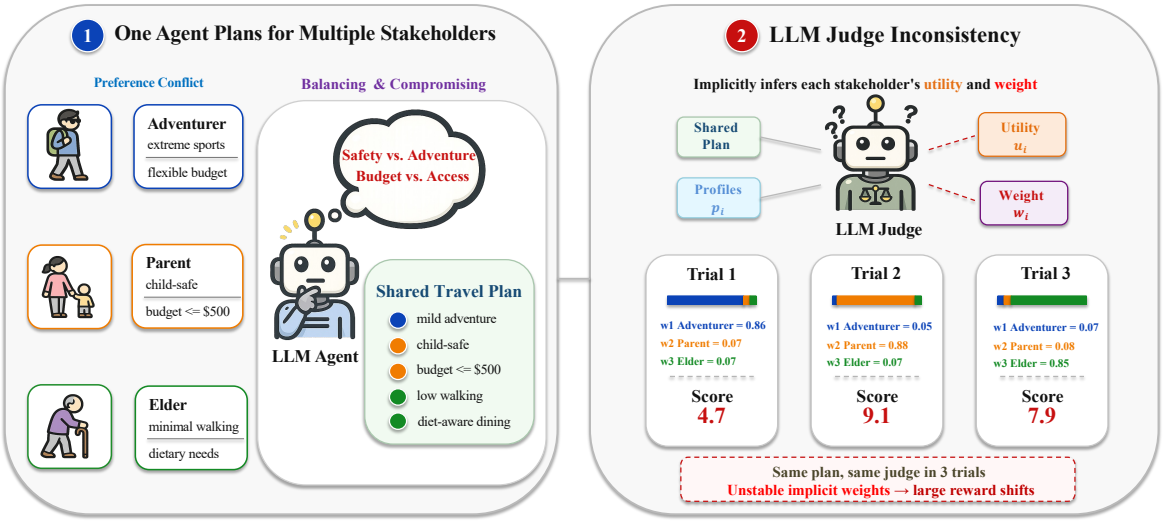

Figure 1: (Left) A single LLM agent synthesizes conflicting stakeholder preferences into one shared plan. (Right) A holistic LLM judge implicitly assigns per-stakeholder utilities u_i and weights w_i; repeated evaluations of the same plan yield different implicit weights and thus different scalar rewards.

bias (Wang et al., 2024c), verbosity bias (Saito et al., 2023), self-preference bias (Panickssery et al., 2024), and intra-rater instability (Haldar and Hockenmaier, 2025). Structured approaches improve reliability in two ways: explicit scoring criteria—rubrics (Hashemi et al., 2024), checklists (Lee et al., 2025), and chain-of-thought scoring (Liu et al., 2023)—decompose evaluation into *quality dimensions*, while trained evaluators—fine-tuned judges or generative reward models (Zhu et al., 2025; Zhang et al., 2025)—replace direct prompting with supervised scoring models. These methods primarily address *estimation* noise: making a specified scoring criterion or learned evaluator more reliable. Multi-stakeholder rewards add a separate *aggregation* problem: the judge must decide how to weigh competing stakeholders, a trade-off these methods leave implicit.

2.2 Multi-Objective Alignment

Multi-objective alignment decomposes reward into predefined quality dimensions or controllable attributes (e.g., helpfulness, safety) and navigates their trade-offs via multi-attribute feedback (Dong et al., 2023; Wang et al., 2024d), multi-objective reward models (Wang et al., 2024b,a), reward-conditioned generation (Yang et al., 2024), or policy merging (Rame et al., 2023; Zhou et al., 2024; Jang et al., 2024). Our setting differs in that the aggregation axes are *people*—varying in number, identity, and constraints per query—whose preferences may directly conflict. This requires query-specific aggregation and creates the weighting-noise problem we formalize in §4.

2.3 GRPO and Reward Noise

GRPO (Shao et al., 2024; Guo et al., 2025) computes advantages via group-relative normalization, making gradient direction depend entirely on relative reward quality within each group. Recent work shows that group-relative advantages can be biased by prompt difficulty (Yang et al., 2026), that noisy or corrupted rewards require explicit correction to avoid biased gradients (El Mansouri et al., 2025), and that multi-objective GRPO is vulnerable to reward hacking by disproportionately optimizing higher-variance reward components (Ichihara et al., 2025). These works study how GRPO responds to noise once a reward signal is given. We ask a prior question: whether the LLM-judge signal itself is reliable for multi-stakeholder tasks, why its variance arises, and how to reduce it before it enters GRPO.

Judge	Quality	n=2		n=3		n=5		n=8		Growth	S/R
		σ^2_{Rep}	σ^2_{Sem}	σ^2_{Rep}	σ^2_{Sem}	σ^2_{Rep}	σ^2_{Sem}	σ^2_{Rep}	σ^2_{Sem}		
Qwen3.5-Plus	Low	0.20	0.52	0.18	0.26	0.21	0.30	0.16	0.21	0.40	1.29
	Medium	0.15	0.59	0.63	0.68	0.67	0.73	0.25	0.46	**0.78**	1.87
	High	0.18	0.27	0.25	0.30	0.28	0.57	0.19	0.46	**1.71**	2.43
Gemini-3-Flash	Low	0.00	0.93	0.00	0.12	0.00	0.18	0.00	0.15	0.16	∞
	Medium	0.00	0.25	0.00	0.81	0.00	1.72	0.00	1.67	**6.67**	∞
	High	0.00	0.64	0.00	0.64	0.00	1.49	0.00	1.46	**2.27**	∞
DeepSeek-V4-Flash	Low	0.43	1.14	0.57	0.45	0.37	0.51	0.26	0.39	0.34	1.54
	Medium	0.47	1.10	1.06	1.52	1.77	1.46	1.31	1.32	**1.20**	1.01
	High	0.50	0.68	1.14	1.09	0.91	1.72	0.49	1.22	**1.81**	2.49
All (Med+High)		0.22	0.59	0.53	0.84	0.62	1.28	0.42	1.10	**1.87**	2.64
All (Low)		0.21	0.86	0.25	0.28	0.19	0.33	0.14	0.25	**0.29**	1.80

Table 1: σ^2_{Rep} = exact-repeat variance; σ^2_{Sem} = presentation-induced reward inconsistency, measured as score variance over semantics-preserving variants of the same response; Growth = $\sigma^2_{\text{Sem}n=8}/\sigma^2_{\text{Sem}n=2}$; S/R = $\sigma^2_{\text{Sem}}/\sigma^2_{\text{Rep}}$ at n=8. Gemini-3-Flash: σ^2_{Rep}=0 at T=0.

3 Empirical Analysis: Multi-Stakeholder Reward Consistency

Multi-stakeholder planning lacks verifiable ground truth, so RL training must rely on LLM judges or reward models to provide the learning signal. We empirically investigate whether current LLM judges can provide a consistent reward signal for this setting—where a single plan must reconcile heterogeneous and sometimes conflicting stakeholder preferences.

Setup. We construct 60 seed responses from real multi-stakeholder travel-planning queries, spanning stakeholder counts $n \in \{2, 3, 5, 8\}$ and three quality levels (high/medium/low). For each seed we generate 55 *semantics-preserving variants* (11 variant families $\times$ 5 versions) via rule-based transforms (order permutation, formatting controls) and LLM-based rewrites (causal-direction flipping, evidence relocation, paraphrase, topic repositioning), all preserving itinerary content, constraint satisfaction, and plan quality. A human evaluator would consider these variants equivalent in task quality; we define σ^2_{Sem} as the score variance across variants for the same seed response. We additionally collect 5 exact repeats of the unchanged input as a noise baseline, yielding σ^2_{Rep}, the judge's intrinsic stochasticity. Their ratio $\sigma^2_{\text{Sem}}/\sigma^2_{\text{Rep}} > 1$ indicates that surface presentation adds variance beyond intrinsic noise. All evaluations use greedy decoding (T=0) across three state-of-the-art LLM judges: Qwen3.5-Plus (Qwen Team, 2026), Gemini-3-Flash (Google, 2025), and DeepSeek-V4-Flash (DeepSeek-AI, 2026). Full details are in Appendix A.

3.1 Score Instability Scales with Stakeholder Count

Table 1 shows that **score variance scales with stakeholder count** for medium- and high-quality responses: pooled Growth=1.87, with Gemini reaching 6.67$\times$. At n=8, the Sem/Rep ratio reaches 2.64 for medium+high responses: merely changing how stakeholders are discussed—their ordering, where evidence is placed—causes 2.6$\times$ more score variation than the judge's own stochasticity. Low-quality responses show no such scaling (Growth=0.29), likely due to a floor effect: scores cluster near the bottom of the scale ($\mu\approx$2.5–3.2), leaving the judge little room to vary. This is especially concerning for GRPO, which relies on within-group reward comparisons: unstable scores for medium- and high-quality responses—the candidates GRPO must distinguish among—can flip which responses are reinforced.

Not all variant types contribute equally: discourse changes tied to stakeholder presentation (causal framing, evidence placement, coverage order) produce the largest variance growth as n increases, while low-level formatting changes show little growth (Appendix A.4). The instability is thus stakeholder-specific, not generic surface noise.

3.2 The Stability–Expressiveness Trade-off in Structured Scoring

We next test whether structured evaluation protocols can reduce this presentation-induced variance. We compare direct holistic scoring with three struc-

Method	n=2	n=3	n=5	n=8	Growth
Direct	0.45	0.72	1.61	1.57	3.50
Rubric	0.70	0.65	0.79	0.82	1.18
Checklist	1.10	1.06	1.47	2.22	2.01
Decomposed (adaptive)	0.61	0.87	0.88	0.96	1.58
Decomposed (uniform)	0.60	0.81	0.83	0.72	1.20

Table 2: σ^2_{Sem} by evaluation method (Gemini-3-Flash, T=0, medium+high only).

n	w CV	$\hat{u}$ CV	Mean $\frac{\lvert\Delta R_w\rvert}{\sigma}$	P95 $\frac{\lvert\Delta R_w\rvert}{\sigma}$	$\frac{\text{Shift Range}}{\sigma}$	w Var%
2	0.050	0.236	2.6%	10.9%	18.2%	1.9
3	0.066	0.143	9.4%	27.9%	51.4%	6.0
5	0.067	0.201	10.7%	37.7%	61.4%	7.4
8	0.106	0.183	14.4%	37.9%	81.2%	5.3

Table 3: Weight drift as a fraction of within-group score spread (Gemini-3-Flash, T=0). σ: mean within-group standard deviation of the final score, computed per variant group then averaged over groups with the same n. CV=σ/μ. w Var%: fraction of total score variance explained by weight drift.

tured protocols (Appendix A.2): **Rubric** scoring over fixed dimensions with fixed weights (Hashemi et al., 2024), **Checklist** with binary verification and programmatic aggregation (Lee et al., 2025), and **Decomposed** scoring in two variants—*adaptive*, where the judge assigns per-stakeholder weights w_i and satisfactions $\hat{u}_i$, and *uniform*, a diagnostic recomputation using w_i=$1/n$ with the same $\hat{u}_i$. All use Gemini-3-Flash at T=0 (σ^2_{Rep}=0) so that all observed variance is purely presentation-induced. We focus on medium- and high-quality responses—the quality range most relevant to GRPO, where the policy must distinguish among competitive candidates.

The key insight is that structure helps reduce presentation-induced variance, but online judge-adaptive stakeholder weighting reintroduces aggregation instability. Rubric (Growth 1.18) and Decomposed with fixed uniform weights (1.20) are most stable, while the controlled decomposed comparison shows that judge-assigned adaptive weights increase Growth from 1.20 to 1.58 using the same $\hat{u}_i$. These stable baselines are not sufficient as final rewards: Rubric aggregates generic quality dimensions rather than stakeholder utilities, and uniform weights ignore query-specific constraint asymmetry. This motivates a two-stage design: estimate per-stakeholder satisfaction online, but anchor non-uniform aggregation weights outside the scoring call. We next quantify the resulting weight-induced score shifts.

3.3 Weight Drift: The Cost of Adaptive Aggregation

We isolate the instability introduced by online stakeholder-weight assignment. For each variant group, we normalize the resulting weight-induced score shift by the within-group final-score standard deviation σ, making it comparable to the reward differences used for GRPO ranking.

Decomposed (adaptive) asks the judge to output per-stakeholder weight w_i and satisfaction $\hat{u}_i$, scoring $\sum_i w_i \hat{u}_i$. Since the query is unchanged across variants, the target aggregation weights under any fixed reward definition should remain fixed. We use the variant-mean $\bar{w}_i$ as reference and measure the score shift from weight drift:

$$\Delta R_w = \sum_i (w_i - \bar{w}_i)\hat{u}_i. \quad (1)$$

Multiplier effect. Table 3 normalizes weight-induced score shifts by the within-group score standard deviation σ, directly showing how much of the advantage signal weight drift can corrupt. Weight drift contributes modestly to average variance (w Var%=2–7%), yet at n=8 it typically shifts a response's score by 14% of σ (mean), with 38% at the 95th percentile and up to 81% in the worst case (shift range). The mechanism is multiplicative: even small weight fluctuations (w CV$\leq$0.11) are multiplied by cross-stakeholder satisfaction differences, so their score impact grows with the spread of $\hat{u}_i$ rather than with weight variance alone. Appendix Table 9 gives the concrete breakdown for one such case.

These magnitudes imply that weight noise alone can flip the sign of GRPO advantages for response pairs whose reward gaps lie within a one-standard-deviation band. The next section analyzes how weight noise enters reward variance and which factors amplify its effect.

4 Theoretical Analysis: Decomposing Judge Variance

Section 3 documented *what* happens: LLM-as-a-Judge scoring is inconsistent despite structured methods. This section analyzes *why* and *how it affects GRPO*: §4.1–4.2 trace the inconsistency to the judge's unstable implicit weighting of stakeholders—an *aggregation error* amplified by utility dispersion—and §4.3 shows when this error

corrupts GRPO's advantage signal. The semantics-preserving variants in Section 3 hold $R^*(y)$ fixed, so their score spread estimates the judge-error scale σ_ξ^2; GRPO is affected when this same-response noise is large relative to the true reward-gap spread within a sampled group.

4.1 Variance Decomposition

Let $u_i^*(y) \in [0,1]$[1] be stakeholder i's latent satisfaction and w_i^* fixed target weights with $\sum_i w_i^* = 1$. We analyze the ideal cardinal welfare reward (Harsanyi, 1955):

$$R^*(y) = \sum_{i=1}^{n} w_i^* \cdot u_i^*(y) \tag{2}$$

We model a holistic judge as implicitly estimating both utilities and weights:

$$\hat{R}(y) = \sum_i (w_i^* + \eta_i)(u_i^*(y) + \delta_i) + \epsilon,$$

where δ_i is utility-estimation noise, η_i is implicit weight drift, and ϵ is residual score-level noise. Normalization gives $\sum_i \eta_i = 0$.

Theorem 1 (Variance Decomposition with Cross Term)**.** *Assume zero-mean noise ($E[\delta_i] = E[\eta_i] = E[\epsilon] = 0$), independence across the three noise groups ($\boldsymbol{\delta} \perp \boldsymbol{\eta} \perp \epsilon$), and independence within $\{\delta_i\}$; the η_i may be correlated. Then:*

$$\mathrm{Var}[\hat{R}(y)] = \underbrace{\sum_i (w_i^*)^2 \sigma_{\delta_i}^2}_{\textit{Term I: Estimation}} + \underbrace{\mathrm{Var}[\textstyle\sum_i \eta_i u_i^*]}_{\textit{Term II: Aggregation}} + \underbrace{\sigma_\epsilon^2}_{\textit{Term III: Shared}} + \underbrace{\sum_i \sigma_{\eta_i}^2 \sigma_{\delta_i}^2}_{C_{cross}} \tag{3}$$

(Proof in Appendix B.1.)

Insight. Term I is reducible via structured evaluation (§3.2). Term II captures an aggregation-specific error: unless the aggregation rule is fixed, zero-sum weight drift η_i is multiplied by utility differences, so its score impact is amplified when stakeholder satisfactions are dispersed (as measured in Table 3). This is the key distinction from single-user noise: multi-stakeholder reward variance includes a component tied to weight noise that is absent in single-reward settings.

[1]The empirical utility heads report scores on a $[1, 10]$ scale; we linearly normalize to $[0, 1]$ for this theoretical notation. Variance and SNR scale by the corresponding constant factor.

4.2 Conditional Scaling and Irreducibility

We next characterize how Term II (weighting noise) depends jointly on weight-drift variance, cross-stakeholder utility dispersion, and stakeholder count.

Proposition 1 (Conditional Scaling of Term II)**.** *For $n \geq 2$, assume exchangeable zero-sum weight noise: $\sum_i \eta_i = 0$, $\mathrm{Var}[\eta_i] = \sigma_\eta^2(n)$ for all i, and a common off-diagonal covariance. Then:*

$$\begin{aligned} \textit{Term II} &= \frac{n}{n-1}\,\sigma_\eta^2(n) \sum_{i=1}^{n} (u_i^* - \bar{u})^2 \\ &= \frac{n^2}{n-1}\,\sigma_\eta^2(n) \mathrm{Var}_i[u_i^*]. \end{aligned}$$

where $\bar{u} = \frac{1}{n}\sum_i u_i^*$ and $\mathrm{Var}_i[u_i^*] \triangleq \frac{1}{n}\sum_i (u_i^* - \bar{u})^2$ is across-stakeholder variance for a fixed response. (Proof in Appendix B.2.)

Insight. Proposition 1 shows that Term II has three necessary factors: weight instability $\sigma_\eta^2(n)$, utility dispersion $\mathrm{Var}_i[u_i^*]$, and the stakeholder-count factor $n^2/(n-1)$. If the judge's implicit weights do not drift, Term II vanishes. If all stakeholders have equal utility ($\mathrm{Var}_i[u_i^*]$=0), zero-sum weight drift is score-invariant. When both weight drift and utility heterogeneity are present, shifting weight toward relatively satisfied or underserved stakeholders changes the scalar reward, with the magnitude scaled by $n^2/(n-1)$. Fixed stakeholder weights remove this channel by setting η_i=0. Structured evaluation can reduce utility-estimation noise (Term I), but it leaves aggregation noise (Term II) whenever the judge still chooses the weights online. Empirically, the weight-induced shift range grows from 18.2% (n=2) to 81.2% (n=8) of the within-group score standard deviation in Table 3. Appendix B.2 derives the symmetric scaling formula and shows how the same mechanism extends to general weight-drift covariance.

4.3 When Weighting Noise Corrupts GRPO

GRPO updates are driven by each candidate's reward gap from the group mean. In the noise-free target, candidate j should be updated according to its true reward gap $\Delta_j^* = R_j^* - \bar{R}^*$ (or the standardized gap $\Delta_j^*/\mathrm{sd}_k(R_k^*)$). In practice, the judge observes $\hat{R}_j = R_j^* + \xi_j$, so GRPO uses the noisy gap

$$\hat{\Delta}_j = \hat{R}_j - \bar{\hat{R}} = \Delta_j^* + (\xi_j - \bar{\xi}). \tag{4}$$

Here Δ_j^* is the **true reward gap** we want GRPO to use: the within-group quality difference between candidate j and the group mean. The second term, $\xi_j - \bar{\xi}$, is the group-centered **random error** induced by the LLM-judge inconsistency studied in Section 3: even the same semantic response can receive different scores under presentation changes. This error turns the true reward gap into the noisy reward gap $\hat{\Delta}_j$ and can flip the advantage sign when it is comparable to Δ_j^*. We summarize this reliability as:

$$\mathrm{SNR} \triangleq \mathrm{Var}_j[\Delta_j^*]/\sigma_\xi^2. \tag{5}$$

Thus SNR compares the true reward-gap signal against the judge-error scale that creates the noisy reward gap. As σ_ξ^2 grows, this noise more easily overwhelms the reward-gap signal that GRPO uses for ranking.

Proposition 2 (Advantage Sign Correctness)**.** *Under independent Gaussian reward errors with variance σ_ξ^2, the probability that GRPO assigns the correct advantage sign to a response one standard deviation from the group mean is:*

$$P(\textit{correct sign}) = \Phi\left(\frac{\sqrt{\mathrm{SNR}}}{\sqrt{1-1/G}}\right) \tag{6}$$

Combining this with Proposition 1, aggregation noise harms GRPO when Term II is non-negligible: it increases σ_ξ^2 most in high-dispersion stakeholder groups, reducing SNR exactly where trade-offs are hardest. In the conditional limit where Term II is the leading noise source and $\sigma_\eta^2(n)\mathrm{Var}_i[u_i^*]$ does not shrink with n, $\mathrm{SNR}(n) = O(1/n)$; substituting into Eq. 6 gives:

$$P(\text{correct sign}) = \frac{1}{2} + O(1/\sqrt{n}). \tag{7}$$

Thus, weighting noise can turn group-relative advantages into a noisy proxy for true quality, pushing GRPO updates toward judge-induced aggregation shifts rather than the intended reward. This motivates DECOMPR: fix aggregation outside online judge calls so the online weight-drift channel is removed by construction. The derivation and numerical thresholds are given in Appendix B.4.

5 DECOMPR: Counterfactual-Calibrated Decomposed Reward

§4 shows that one source of holistic reward instability is that the judge's implicit trade-off rule can change across candidate responses. Our method, DECOMPR, addresses this with counterfactual-calibrated offline query-level aggregation and online grounded per-role utility estimation (Figure 2).

5.1 Counterfactual-Calibrated Aggregation (Removing Online Weight Drift)

We remove online weight drift by fixing aggregation weights once per query. The calibration follows a counterfactual-sacrifice intuition from bargaining theory (Kalai and Smorodinsky, 1975): stakeholders who lose more from joining the shared task, relative to their solo ideal, should receive higher aggregation priority. Conceptually, with solo plan y_i^{solo} and joint reference plan y^{joint}:

$$\Delta_i^{\text{cf}} = U_i(y_i^{\text{solo}}) - U_i(y^{\text{joint}}), \tag{8}$$

These weights may not be uniquely correct, but they incorporate stakeholder fairness into the offline fixed aggregation rule before any candidate response is scored.

In implementation, we do not explicitly compute the counterfactual utility gap in Eq. 8, since this would require costly solo and joint planning estimates. Instead, we use demand difficulty as a lightweight proxy: stakeholders whose requirements are harder to satisfy receive larger weights. This follows the same intuition as Eq. 8: hard-to-satisfy needs should not be silently sacrificed. Concretely, an offline parser estimates difficulty from restrictive hard constraints, dense soft preferences, and direct conflicts; fixed rules then compute:

$$d_i = \sum_{c \in C_i^{\text{hard}}} \alpha(c) + \gamma \sum_{c \in C_i^{\text{soft}}} \alpha(c) + \beta s_i^{\text{conflict}} \tag{9}$$

where C_i^{hard} and C_i^{soft} are parsed constraints, $\alpha(c)$ is a fixed restrictiveness score, s_i^{conflict} summarizes pairwise tensions involving stakeholder i, and $\gamma, \beta \in (0,1)$ are fixed discounts. All quantities are computed from the query and profiles only, before any candidate response is scored. The resulting d_i is a deterministic proxy for $\mathbb{E}[\Delta_i^{\text{cf}} \mid q, p_{1:n}]$, not a candidate score. We then normalize these query-level difficulty scores into weights:

$$w_i = \frac{\exp(d_i/\tau_w)}{\sum_{j=1}^{n} \exp(d_j/\tau_w)} \tag{10}$$

where τ_w controls differentiation ($\tau_w \to \infty$ recovers uniform; $\tau_w \to 0$ concentrates on the hardest stakeholder). Since d_i and w_i are fixed for a given query, candidate-dependent weight drift is removed even if the calibrated weights are only approximate.

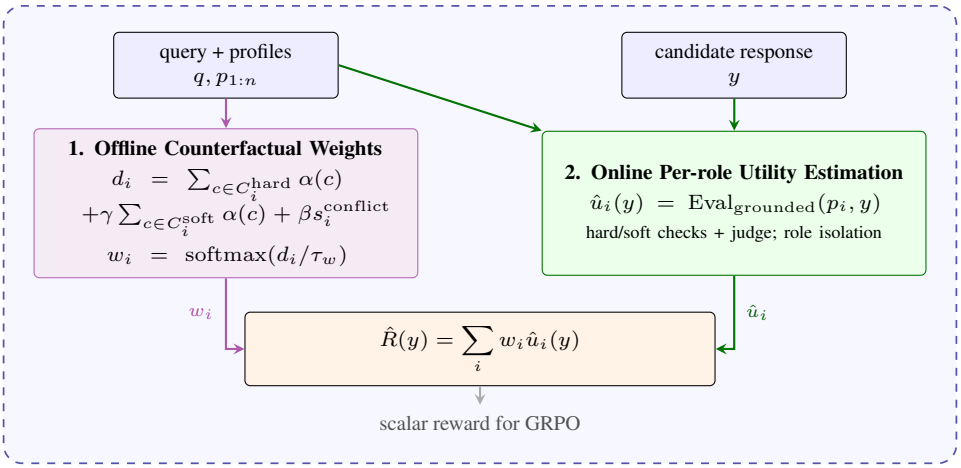

Figure 2: Overview of DECOMPR: offline query-level weights remove candidate-dependent weight drift, online per-role utilities reduce estimation error, and fixed aggregation produces the scalar reward for GRPO.

Stakeholder	d_i	w_i^{u}	w_i^{c}	$\hat{u}_i$	Sacrifice
A (4 hard + 2 soft)	5.0	0.33	**0.67**	4.0	50%
B (0 hard + 2 soft)	1.0	0.33	0.09	9.5	3%
C (2 hard + 2 soft)	3.0	0.33	0.24	8.0	11%
$\hat{R}_{\text{uniform}} = 7.17$			$\hat{R}_{\text{calib}} = \mathbf{5.47}$		

Table 4: Toy three-traveler calibration (γ=0.5, τ_w=2). Utilities and rewards use the $[1, 10]$ scale. Sacrifice denotes illustrative utility loss relative to each stakeholder's solo ideal. Values are illustrative; implementation uses the query-level difficulty proxy d_i rather than computing solo optima online.

Aggregation	n=2	n=3	n=5	n=8	Growth
Adaptive (judge w_i)	0.61	0.87	0.88	0.96	1.58
Uniform (w_i=1/n)	0.60	0.81	0.83	0.72	1.20
DECOMPR (τ_w=3)	0.63	0.80	0.92	0.78	1.24
DECOMPR (τ_w=5)	0.64	0.79	0.88	0.75	1.18
DECOMPR (τ_w=8)	0.65	0.79	0.86	0.73	1.14

Table 5: σ^2_{Sem} under different fixed aggregation weights (same $\hat{u}_i$ from decomposed evaluation). Growth = $\sigma^2_{\text{Sem}n=8}/\sigma^2_{\text{Sem}n=2}$. τ_w controls weight differentiation (Eq. 10).

Illustrative example. Table 4 shows how calibrated weights expose failures hidden by uniform aggregation: the hardest stakeholder receives more weight, so underserving that stakeholder lowers the reward.

5.2 Online Per-Role Grounded Estimation (Reducing Term I)

Given a candidate response y, DECOMPR estimates each stakeholder's utility independently:

$$\hat{u}_i(y) = \text{Eval}_{\text{grounded}}(p_i, y). \tag{11}$$

The evaluator only asks whether y satisfies profile p_i, rather than asking one judge to balance all stakeholders at once. This per-role isolation is designed to keep the estimation task in the simpler n=1 regime, avoiding the context and attention load of holistic multi-stakeholder judging.

Within each role, signals are layered by decreasing determinism: tool-verified facts (σ^2=0), programmatic constraint checks ($\sigma^2 \approx 0$), LLM-judged soft preferences, and a residual subjective checklist with weight at most 20%. Thus most reward mass comes from deterministic or low-variance components.

6 Experiments

We first verify that DECOMPR's aggregation design reduces reward variance (§6.1), then evaluate whether this translates to better plan quality in end-to-end GRPO training (§6.2).

6.1 Reward Consistency

Using the same setup as §3.2, we recompute decomposed scores under different fixed weight schemes while holding $\hat{u}_i$ unchanged, isolating the effect of aggregation weights on reward stability.

Table 5 shows that any per-query fixed weight eliminates candidate-dependent weight drift, but the degree of n-scaling reduction depends on weight concentration $\sum_i w_i^2$. Adaptive weights (Growth=1.58) allow drift across variants; uniform weights (Growth=1.20) minimize $\sum_i w_i^2 = 1/n$ but lose query-specific differentiation. DECOMPR with τ_w=5 achieves Growth=1.18, close to uniform stability while preserving meaningful constraint-

	Group Utility ↑				Group Fairness ↑				Pref. Completeness ↑			
Reward	Easy	Med	Hard	**All**	Easy	Med	Hard	**All**	Easy	Med	Hard	**All**
DECOMPR	**5.94**	**7.35**	**7.85**	**7.07**	60.10	**49.58**	**39.84**	**49.82**	**24**	**19**	15	**19**
Direct	5.34	6.68	7.42	6.50	58.52	45.20	35.82	46.42	22	17	15	18
Rubric	5.49	6.30	7.14	6.31	59.76	44.69	35.86	46.61	22	17	14	18
Checklist	5.21	6.25	7.38	6.28	58.14	44.82	36.51	46.23	21	17	15	17
Base Model	5.11	6.19	7.20	6.17	**60.50**	46.52	37.69	48.11	23	18	14	18

Table 6: MR-TravelBench results (650 tasks $\times$ 3 trials, Easy/Medium/Hard = 200/250/200 tasks). DECOMPR achieves the highest Group Utility (+8.8% vs Direct, +12.0% vs Rubric) and Group Fairness (+7.3% vs Direct), with the advantage growing on Hard tasks (+5.8% Group Utility, +11.2% Group Fairness vs Direct). Base Model is the pre-RL reference.

based prioritization; τ_w=8 is slightly more stable (Growth=1.14) but closer to uniform weighting. The temperature τ_w controls the stability–expressiveness trade-off: smaller τ_w concentrates weight on high-difficulty stakeholders (more expressive but amplifies their estimation noise); larger τ_w approaches uniform (more stable but loses stakeholder differentiation).

6.2 End-to-End Training

We next evaluate whether the variance reduction translates to better downstream plan quality when used as the GRPO reward signal.

Setup. We use Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025) as the base policy model: after supervised fine-tuning on expert trajectories generated by Qwen3.5-Plus, it undergoes GRPO training on 6,000 queries with the four scoring methods from §3.2 (Direct, Rubric, Checklist, DECOMPR with τ_w=5) as reward functions. We additionally include the base model (without RL) as a reference. Best checkpoint is selected within 400 steps; evaluation uses MR-TravelBench (Cheng et al., 2026) (650 held-out tasks $\times$ 3 trials). Additional training details are in Appendix C.

Table 6 shows that DECOMPR achieves the strongest aggregate performance across the three core metrics. All GRPO configurations improve Group Utility over the Base Model (6.17), showing that RL improves utility, though not necessarily fairness: Direct, Rubric, and Checklist rewards all *reduce* Group Fairness below the Base Model (48.11), suggesting that noisy reward signals may encourage GRPO to optimize utility at the expense of equity. Only DECOMPR improves both simultaneously (Group Utility=7.07, Group Fairness=49.82), consistent with the counterfactual-calibrated weighting that prevents the agent from satisfying only the easy-to-serve travelers. Among GRPO methods, Checklist performs worst on Group Utility (6.28) and Preference Completeness (17), consistent with its high presentation-induced variance (Growth=2.01 in §3.2): binary checks that flip under surface changes produce a less stable reward signal for GRPO.

The advantage *increases with task difficulty*: on Hard tasks—where conflicts are more severe and satisfaction dispersion is expected to be higher—DECOMPR outperforms Direct by +5.8% in utility and +11.2% in fairness. This aligns with our theoretical analysis: the multiplier effect of weight drift is most damaging in high-dispersion scenarios (§4), and fixing aggregation outside the judge provides the largest benefit where it matters most.

7 Conclusion

This work identifies a central difficulty in multi-stakeholder LLM-judge evaluation: rewards become unreliable when a single judge call estimates stakeholder utilities and decides cross-stakeholder weights. Our theory and experiments show that this coupling creates *weighting noise*, whose score impact grows with satisfaction dispersion and stakeholder count. Structured scoring reduces utility-estimation noise but leaves this online aggregation channel intact.

DECOMPR addresses this through *estimation-aggregation separation*: online grounded estimation measures per-role utilities, while offline counterfactual calibration fixes auditable aggregation weights before candidate scoring.

More broadly, multi-stakeholder reward modeling remains underexplored. We hope this work encourages more systematic study of how LLM systems evaluate and learn from competing human preferences.

Limitations

Our work focuses on the multi-party travel planning domain as a representative case of multi-stakeholder alignment. While we believe the theoretical insights generalize to other multi-stakeholder settings (e.g., negotiation, collaborative writing, group recommendation), empirical validation in these domains remains for future work. Additionally, our programmatic verification layers rely on the availability of structured tool APIs; in domains where such tools are unavailable, a larger fraction of reward signal would necessarily depend on LLM-based estimation. Finally, existing benchmarks for multi-stakeholder planning are scarce, which limits the scale of end-to-end GRPO training experiments we can conduct; constructing richer multi-party benchmarks and running full-scale RL validation is an important direction for future work.

References

Xiang Cheng, Yulan Hu, Lulu Zheng, Zheng Pan, Xin Li, and Yong Liu. 2026. Grouptravelbench: Benchmarking llm agents on multi-person travel planning. *arXiv preprint arXiv:2605.25200*.

DeepSeek-AI. 2026. Deepseek-v4 technical report. Technical report, DeepSeek-AI.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11275–11288, Singapore. Association for Computational Linguistics.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. 2025. Noise-corrected GRPO: From noisy rewards to unbiased gradients. *arXiv preprint arXiv:2510.18924*.

Google. 2025. Gemini 3 flash: Frontier intelligence built for speed. Technical report, Google.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Rajarshi Haldar and Julia Hockenmaier. 2025. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24986–25004, Suzhou, China. Association for Computational Linguistics.

John C. Harsanyi. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Mitsuki Sakamoto, Ryota Mitsuhashi, and Eiji Uchibe. 2025. MO-GRPO: Mitigating reward hacking of group relative policy optimization on multi-objective problems. *arXiv preprint arXiv:2509.22047*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2024. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning Workshop at NeurIPS 2024*.

Ehud Kalai and Meir Smorodinsky. 1975. Other solutions to Nash's bargaining problem. *Econometrica*, 43(3):513–518.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, Suzhou, China. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517, Miami, Florida, USA. Association for Computational Linguistics.

Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: Towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. In *Instruction Tuning and Instruction Following Workshop at NeurIPS 2023*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024a. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655, Bangkok, Thailand. Association for Computational Linguistics.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024c. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024d. Helpsteer2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. 2026. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56276–56297. PMLR.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025. Generative verifiers: Reward modeling as next-token prediction. In *Proceedings of the International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613, Bangkok, Thailand. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *Proceedings of the International Conference on Learning Representations*.

A Multi-Stakeholder Reward Consistency Experiment Details

Scope. This appendix section supports the reward-consistency analysis in §3. We study multi-stakeholder travel planning, where a single itinerary must satisfy multiple travelers with heterogeneous constraints such as budget limits, dietary restrictions, mobility needs, safety constraints, and activity preferences. The response content is fixed, and only semantics-preserving presentation variants are scored.

Base responses. We use five real multi-stakeholder travel queries, each expanded to $n \in \{2, 3, 5, 8\}$ stakeholders. For each query–n pair, we construct one high-, medium-, and low-quality response, yielding $5 \times 4 \times 3 = 60$ seed responses.

Quality bands. High-quality responses satisfy hard constraints, maintain credible travel information, preserve time feasibility, and cover stake-

holder needs in a balanced way. Medium-quality responses are broadly feasible but contain one or two compromises, questionable details, or under-served stakeholders. Low-quality responses contain hard-constraint violations, hallucinated or implausible travel details, or time-line conflicts. These quality bands are pre-constructed and used only as grouping variables; the consistency experiment does not generate additional quality-changing variants.

A.1 Response Construction and Semantics-Preserving Variants

Generation. For each base response, we first use DeepSeek-V4-Flash with tool calls to extract a structured JSON metadata file recording all factual content (locations, times, prices, activities), per-stakeholder constraint satisfaction status, and trade-off decisions. Rule-based variants (e.g., order permutations, evidence placement, formatting changes) operate directly on this metadata to guarantee semantic preservation. LLM-based rewrites are also performed by DeepSeek-V4-Flash, conditioned on the metadata file so that the rewrite cannot introduce information beyond—or omit information present in—the structured record.

Semantic-preservation constraints. All variants preserve five invariants: (i) itinerary facts such as locations, times, prices, transportation, and activities; (ii) each hard or soft constraint's satisfaction status; (iii) the stakeholder set; (iv) the semantic strength of satisfaction or violation claims; and (v) the response quality band. Variants change only response presentation.

Variant families. For each base response, we generate eleven semantics-preserving variant families with five versions per family. The exact-repeat baseline is collected separately and is not counted as a variant family.

Exact-repeat baseline. Each base response is also scored five times with the exact same input (no rewriting). These repeats estimate intrinsic judge noise (σ^2_{Rep}).

Scale. Each base response yields 60 scoring inputs (11 variant families $\times$ 5 versions $+$ 5 exact repeats). With 60 seeds, this gives $60 \times 60 = 3{,}600$ scoring inputs per judge model and scoring method.

Example. We illustrate with a real 3-stakeholder query (Xiamen→Fuzhou weekend trip, budget ≤500 CNY/person). The three travelers are: user1 (photography, prefers quiet spots), user2 (limited mobility, no hiking), and user3 (tight schedule, wants popular check-in spots). A *stakeholder section order* variant permutes only the discussion order:

> **Original:** user1 (photographer) → user2 (mobility) → user3 (check-in)
> **Variant:** user3 (check-in) → user1 (photographer) → user2 (mobility)

The itinerary, constraint satisfaction, and all factual content remain identical. Yet this reordering alone causes score shifts of up to 2.0 points on the deterministic Gemini judge (3.5→5.5) and a 4.8-point spread on DeepSeek (3.5–8.3 across five permutations), demonstrating that *who is discussed first* influences the judge's holistic assessment even when all other content is fixed.

A.2 Judge Scoring Protocols

We evaluate four scoring protocols of increasing structural decomposition. All protocols share a verification-oriented system prompt that enforces evidence-based assessment: the judge must (i) distinguish between surface mention and concrete executable arrangements, (ii) independently verify budget totals by summing itemized costs rather than accepting self-reported figures, and (iii) cross-check transit times against distance plausibility. All outputs use structured JSON for deterministic parsing.

Direct (holistic). The judge conducts a chain-of-thought analysis along five predefined aspects—hard-constraint satisfaction, information credibility, timeline feasibility, inter-stakeholder fairness, and actionability—then produces a single scalar score $R \in [1.0, 10.0]$. Explicit scoring anchors map qualitative assessments to numeric ranges (e.g., 9.0–10.0: all constraints satisfied with fully credible information; 3.0–4.9: hard-constraint violations or substantial fabrication). The judge must output per-aspect evidence prior to the final score, ensuring the holistic judgment is grounded in verifiable observations.

Rubric (dimension-decomposed). The judge independently scores five predefined quality dimensions (Table 8), each on $[1.0, 10.0]$ with dimension-specific anchors. The final score is the deterministic weighted sum $R = \sum_{d=1}^{5} \alpha_d \cdot s_d$, where α_d are fixed weights and s_d are per-dimension scores. This protocol tests whether decomposing

Variant family	Method	Operation
Multi-stakeholder-specific		
Stakeholder section order	Rule	Permute per-stakeholder coverage sections in the response.
Satisfaction summary order	Rule	Permute stakeholder entries in the satisfaction summary.
Causal direction	LLM	Flip trade-off framing between cause→effect and effect→cause.
Under-served evidence placement	LLM	Relocate under-served stakeholder evidence to different positions (before/after itinerary, interleaved, etc.).
Shared-benefit attribution	Rule	Reorder multi-stakeholder attribution when a shared item serves multiple users.
General presentation controls		
Trade-off order	Rule	Permute conflict-explanation paragraphs without changing claims.
Paraphrase	LLM	Rewrite wording with synonymous expressions; preserve all facts.
Format conversion	Rule	Convert among prose, bullet lists, numbered lists, and markdown tables.
Section header	Rule	Replace section headings with synonymous labels (e.g., "Budget" → "Cost").
Low-level formatting	Rule	Change whitespace, bullet style, emphasis markers only.
Topic position	LLM	Move non-stakeholder topical sections to different locations in the response.

Table 7: Semantics-preserving variant families. **Rule**: deterministic transforms operating on extracted metadata (order permutation, formatting changes). **LLM**: DeepSeek-V4-Flash rewrites conditioned on structured metadata to prevent fact drift. Five variants per family per response; exact repeats are a separate baseline.

Dimension	Weight α_d	Evaluation focus
Constraint satisfaction	0.30	Per-user hard constraints verified item-by-item; budget independently summed
Information credibility	0.25	POI existence, price plausibility, transit time–distance consistency
Timeline feasibility	0.20	Schedule coherence, transit time accounting, activity density
Fairness	0.15	Dedicated time slots per user; compensation for compromised stakeholders
Conflict resolution	0.10	Specificity of resolution strategies; presence of creative win-win solutions

Table 8: Rubric dimensions and fixed aggregation weights. Each dimension is scored on $[1.0, 10.0]$ with explicit anchors (e.g., constraint satisfaction: 10 = all satisfied, 6 = one minor violation, 2 = three or more violations).

the evaluation into orthogonal quality axes reduces presentation-induced variance by constraining the judge's attention.

Checklist (binary verification). The judge answers a set of binary verification questions organized into five weighted categories: (A) per-user hard-constraint checks ($\omega = 2.5$ per item), (B) information credibility ($\omega = 2.0$), (C) timeline logic ($\omega = 1.5$), (D) fairness ($\omega = 1.5$), and (E) actionability ($\omega = 1.0$). Each item requires cited evidence from the plan. The final score is computed deterministically:

$$R = \frac{\sum_b a_b \cdot \omega_b}{\sum_b \omega_b} \times 9 + 1, \quad R \in [1, 10] \qquad (12)$$

where $a_b \in \{0, 1\}$ is the binary answer and ω_b the item weight. Crucially, Category A is instantiated per-stakeholder (3 items $\times$ n users), so checklist length grows linearly with n.

Decomposed (adaptive). The judge independently evaluates each stakeholder i through a structured pipeline: (a) extracting hard constraints and soft preferences from the user profile, (b) verifying each constraint against the plan with cited evidence, (c) assessing credibility of POIs relevant to that user, and (d) assigning a satisfaction score $\hat{u}_i \in [1.0, 10.0]$ with explicit anchors. The judge additionally generates adaptive importance weights w_i based on constraint severity (safety-related constraints receive highest weight; vague preferences receive lowest), subject to $\sum_i w_i = 1$. The final score is $R = \sum_i w_i \hat{u}_i$. This protocol serves dual purposes: as a structured scoring method and as a diagnostic instrument for measuring weight instability (§3.3).

Decomposed (uniform). To isolate the effect of weight drift from satisfaction estimation noise, we additionally report a post-hoc recomputation using uniform weights $w_i = 1/n$ applied to the *same* $\hat{u}_i$ values produced by the adaptive protocol. This baseline requires no additional judge calls—it simply replaces the adaptive weights with equal weights in the aggregation formula $R = \frac{1}{n} \sum_i \hat{u}_i$. The comparison between adaptive and uniform variants (Table 2) directly quantifies the variance contribution of weight assignment.

Judge models and decoding. We use three judge families: Qwen3.5-Plus (Qwen Team, 2026), Gemini-3-Flash (Google, 2025), and DeepSeek-

V4-Flash (DeepSeek-AI, 2026). All use greedy decoding (T=0) and structured JSON output. In our exact-repeat baseline, Gemini at T=0 produced identical scores, providing a controlled setting in which all observed Gemini score variance is attributable to input presentation differences rather than repeat-sampling stochasticity.

A.3 Reward Consistency Metrics

For an experimental unit $s = (\text{base response}, n, \text{quality})$, variant family t, variant version b, and exact repeat r, let $\text{score}(s, t, b)$ be the parsed scalar judge score.

Semantic-variant variance.

$$\sigma^2_{\text{Sem}}(s) = \text{Var}_{t,b}[\text{score}(s, t, b)]\,. \tag{13}$$

This measures presentation-induced reward inconsistency: $\sigma^2_{\text{Sem}}(s) = 0$ means the judge gives identical scores to all semantics-preserving variants of the same response, while larger values mean the score is sensitive to wording, ordering, or evidence placement.

Exact-repeat variance.

$$\sigma^2_{\text{Rep}}(s) = \text{Var}_r[\text{score}(s, \text{repeat}, r)]\,. \tag{14}$$

Coefficient of variation and scaling.

$$\text{CV}(s) = \frac{\sqrt{\sigma^2_{\text{Sem}}(s)}}{\text{Mean}_{\text{Sem}}(s)},$$

$$\text{Growth} = \frac{\sigma^2_{\text{Sem}}(n=8)}{\sigma^2_{\text{Sem}}(n=2)}.$$

We also report $\sigma^2_{\text{Sem}}/\sigma^2_{\text{Rep}}$; values above one indicate that semantics-preserving presentation changes induce more score variance than exact repeated scoring.

Decomposed-diagnostic metrics. For Decomposed (adaptive) outputs, we measure $\text{CV}(w_i)$, $\text{CV}(\hat{u}_i)$, and the weight-induced score shift $\Delta R_w = \sum_i (w_i - \bar{w}_i)\hat{u}_i$, where $\bar{w}_i$ is the mean adaptive weight within the same query–quality–n semantic-variant group.

A.4 Variance Breakdown by Variant Type

Table 10 decomposes σ^2_{Sem} by variant family (all three judges, direct scoring, T=0). Two patterns stand out. First, variants that alter stakeholder salience or trade-off framing tend to produce the largest high-n variance: *causal_direction* (0.93 at n=8), *underserved_evidence_placement* (0.90), *stakeholder_coverage_order* (0.85), and *trade-off_explanation_order* (Growth=1.58). Second, low-level controls such as *low_level_formatting* (Growth=1.02) and *section_header_variant* (Growth=1.07) show much weaker scaling with n. This suggests that the n-dependent variance growth documented in §3.1 is driven less by generic formatting and more by presentation changes that affect the judge's implicit weighting of stakeholder interests and trade-offs.

Stakeholder	w_{actual}	$\bar{w}$	Δw	$\hat{u}_i$	$\Delta w \times \hat{u}_i$
User 1	0.100	0.126	−0.027	8.0	−0.212
User 2	0.100	0.139	−0.040	8.0	−0.316
User 3	0.300	0.228	+0.072	1.0	+0.072
User 4	0.300	0.251	+0.049	2.0	+0.098
User 5	0.200	0.255	−0.055	6.0	−0.330
Total shift					**−0.688**

Table 9: Weight-induced score shift in a high-dispersion variant (n=5, high quality). Adaptive weights move mass away from high-$\hat{u}_i$ stakeholders and toward low-$\hat{u}_i$ stakeholders, producing a −0.69 aggregation shift.

B Theory Details for Section 4

B.1 Proof of Theorem 1 (Decomposition)

Proof. Substituting $\hat{w}_i = w_i^* + \eta_i$ and $\hat{u}_i = u_i^* + \delta_i$ into the holistic judge model $\hat{R}(y) = \sum_i \hat{w}_i \hat{u}_i + \epsilon$:

$$\begin{aligned}\hat{R}(y) &= \sum_i (w_i^* + \eta_i)(u_i^* + \delta_i) + \epsilon \\ &= \underbrace{\textstyle\sum_i w_i^* u_i^*}_{R^*(y)} + \underbrace{\textstyle\sum_i w_i^* \delta_i}_{\text{(I)}} \\ &\quad + \underbrace{\textstyle\sum_i \eta_i u_i^*}_{\text{(II)}} + \underbrace{\textstyle\sum_i \eta_i \delta_i}_{\text{cross}} + \epsilon\end{aligned}$$

Since $R^*(y)$ is a constant for fixed y, all noise terms have zero mean ($E[\delta_i] = E[\eta_i] = E[\epsilon] = 0$), and the estimation noise, weight noise, and residual noise are mutually independent, the pairwise covariance terms among (I), (II), the interaction term, and ϵ vanish:

$$\begin{aligned}\text{Var}[\hat{R}(y)] &= \underbrace{\text{Var}[\textstyle\sum_i w_i^* \delta_i]}_{\text{Term I}} + \underbrace{\text{Var}[\textstyle\sum_i \eta_i u_i^*]}_{\text{Term II}} \\ &\quad + \underbrace{\text{Var}[\textstyle\sum_i \eta_i \delta_i]}_{C_{\text{cross}}} + \underbrace{\sigma^2_\epsilon}_{\text{Term III}}\end{aligned} \tag{15}$$

Term I $= \sum_i (w_i^*)^2 \sigma^2_{\delta_i}$ by mutual independence of δ_i. Term II is left in general form, as η_i may

Variant family	n=2	n=3	n=5	n=8	Growth
Multi-stakeholder-specific					
stakeholder_coverage_order	—	0.76	0.71	0.85	—
satisfaction_summary_order	—	0.73	0.79	0.55	—
causal_direction	0.75	0.89	1.18	0.93	1.24
shared_benefit_attribution	0.49	0.56	0.77	0.54	1.11
underserved_evidence_placement	0.76	0.60	1.34	0.90	1.18
General presentation controls					
tradeoff_explanation_order	0.39	0.54	0.95	0.62	1.58
generic_paraphrase	0.68	0.69	0.82	0.84	1.23
format_conversion	0.56	0.50	0.80	0.68	1.22
section_header_variant	0.55	0.69	0.77	0.58	1.07
low_level_formatting	0.51	0.54	0.87	0.52	1.02
topic_position	0.75	1.19	0.86	0.64	0.86
exact_repeat (baseline)	0.26	0.54	0.49	0.34	1.33
ALL (semantic)	0.64	0.68	0.96	0.86	1.34

Table 10: σ^2_{Sem} by variant family (all three judges, direct scoring, T=0). Growth = $\sigma^2(n{=}8)/\sigma^2(n{=}2)$. "—" indicates the variant is undefined for n=2 (requires ≥3 stakeholders). The exact_repeat row estimates intrinsic judge noise σ^2_{Rep}; ALL aggregates the eleven semantic variant families.

be correlated through the normalization constraint $\sum_i \hat{w}_i = 1$ (expanded in Proposition 1). For C_{cross}, by the independence assumption $\boldsymbol{\delta} \perp \boldsymbol{\eta}$ (Theorem 1), δ_i is independent of all η_j, and $E[\delta_i]$=0 eliminates cross-i terms:

$$\mathrm{Var}[\textstyle\sum_i \eta_i \delta_i] = \sum_i E[\eta_i^2]\, E[\delta_i^2] = \sum_i \sigma^2_{\eta_i} \sigma^2_{\delta_i}. \tag{16}$$

This yields Eq. 3.

An equivalent derivation uses the Law of Total Variance, conditioning on the implicit weights $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$:

$$\mathrm{Var}(\hat{R}) = \underbrace{E_{\boldsymbol{\eta}}[\mathrm{Var}(\hat{R} \mid \boldsymbol{\eta})]}_{\text{Terms I + III + cross}} + \underbrace{\mathrm{Var}_{\boldsymbol{\eta}}(E[\hat{R} \mid \boldsymbol{\eta}])}_{\text{Term II}} \tag{17}$$

Here,

$$\begin{aligned} &E_{\boldsymbol{\eta}}[\mathrm{Var}(\hat{R} \mid \boldsymbol{\eta})] \\ &\quad = E_{\boldsymbol{\eta}}\left[\textstyle\sum_i (w_i^* + \eta_i)^2 \sigma^2_{\delta_i} + \sigma^2_\epsilon\right] \\ &\quad = \textstyle\sum_i (w_i^*)^2 \sigma^2_{\delta_i} + \sum_i \sigma^2_{\eta_i} \sigma^2_{\delta_i} + \sigma^2_\epsilon, \end{aligned} \tag{18}$$

recovering Term I, the interaction term, and Term III. The inner expectation $E[\hat{R} \mid \boldsymbol{\eta}] = \sum_i (w_i^* + \eta_i) u_i^*$ varies only through η_i, recovering Term II as the variance of the aggregation step. □

B.2 Proof of Proposition 1 (Scaling)

Full Derivation of Proposition 1. The weight drift η_i satisfies $\sum_i \eta_i = 0$ (normalization constraint: $\sum_i \hat{w}_i = \sum_i w_i^* = 1$). Proposition 1 assumes exchangeable zero-sum drift for $n \geq 2$: $\mathrm{Var}(\eta_i) = \sigma^2_\eta(n)$ for all i and a common off-diagonal covariance $\mathrm{Cov}(\eta_i, \eta_j) = c_\eta$ for $i \neq j$. Since $\sum_i \eta_i = 0$ holds identically (not merely in expectation), the variance of this constant is zero. Expanding under exchangeability:

$$\begin{aligned} 0 = \mathrm{Var}[\textstyle\sum_i \eta_i] &= \sum_i \mathrm{Var}[\eta_i] + \sum_{i \neq j} \mathrm{Cov}(\eta_i, \eta_j) \\ &= n\sigma^2_\eta(n) + n(n-1)c_\eta, \end{aligned} \tag{19}$$

so $c_\eta = -\sigma^2_\eta(n)/(n-1)$. Expanding Term II using the standard variance formula for a linear combination:

$$\begin{aligned} \text{Term II} &= \mathrm{Var}[\textstyle\sum_i \eta_i u_i^*] \\ &= \sum_i \sum_j u_i^* u_j^* \,\mathrm{Cov}(\eta_i, \eta_j) \\ &= \sum_i (u_i^*)^2 \mathrm{Var}(\eta_i) \\ &\quad + \sum_{i \neq j} u_i^* u_j^* \,\mathrm{Cov}(\eta_i, \eta_j) \\ &= \sigma^2_\eta(n) \sum_i (u_i^*)^2 \\ &\quad - \frac{\sigma^2_\eta(n)}{n-1} \sum_{i \neq j} u_i^* u_j^* \end{aligned}$$

Using the identity $\sum_{i\neq j} u_i^* u_j^* = (\sum_i u_i^*)^2 - \sum_i (u_i^*)^2$ to eliminate the off-diagonal sum:

$$
\begin{aligned}
&= \sigma_\eta^2(n) \sum_i (u_i^*)^2 \\
&\quad - \frac{\sigma_\eta^2(n)}{n-1} \big[(\textstyle\sum_i u_i^*)^2 - \sum_i (u_i^*)^2 \big] \\
&= \sigma_\eta^2(n) \Big(1 + \tfrac{1}{n-1} \Big) \sum_i (u_i^*)^2 \\
&\quad - \frac{\sigma_\eta^2(n)}{n-1} (\textstyle\sum_i u_i^*)^2 \\
&= \frac{n\, \sigma_\eta^2(n)}{n-1} \\
&\quad \cdot \left[\sum_i (u_i^*)^2 - \frac{(\sum_i u_i^*)^2}{n} \right]
\end{aligned}
$$

Recognizing the bracketed expression as the (unnormalized) variance $\sum_i (u_i^* - \bar{u})^2 = \sum_i (u_i^*)^2 - \frac{1}{n}(\sum_i u_i^*)^2$, and applying $\mathrm{Var}_i[u_i^*] \triangleq \frac{1}{n}\sum_i (u_i^* - \bar{u})^2$:

$$
\begin{aligned}
&= \frac{n\, \sigma_\eta^2(n)}{n-1} \sum_i (u_i^* - \bar{u})^2 \\
&= \frac{n^2}{n-1}\, \sigma_\eta^2(n)\, \mathrm{Var}_i[u_i^*]. \qquad (20)
\end{aligned}
$$

Here $\bar{u} = \frac{1}{n}\sum_i u_i^*$. The formula exposes three factors: weight-drift variance $\sigma_\eta^2(n)$, utility dispersion $\mathrm{Var}_i[u_i^*]$, and the count factor $n^2/(n-1)$. Term II vanishes if either weight drift or utility dispersion is zero; if either is small, the term is proportionally small. For fixed utility dispersion, the n-dependence is governed by $n\sigma_\eta^2(n)$, so the term grows only when per-stakeholder drift does not shrink fast enough and remains bounded when $\sigma_\eta^2(n) = O(1/n)$.

Remark on exchangeability. The exchangeable model is a symmetric specialization. Without exchangeability, let $\Sigma_\eta = \mathrm{Cov}(\boldsymbol{\eta})$; the zero-sum constraint implies $\Sigma_\eta \mathbf{1} = 0$. Then

$$
\text{Term II} = (u^*)^\top \Sigma_\eta u^* = (u^* - \bar{u}\mathbf{1})^\top \Sigma_\eta (u^* - \bar{u}\mathbf{1}). \qquad (21)
$$

Thus the qualitative condition is unchanged: aggregation noise vanishes when weight drift is zero or when stakeholder utilities are constant, and otherwise depends on the alignment between utility dispersion and the covariance directions of weight drift. Proposition 1 is the exchangeable special case, where $\Sigma_\eta = \frac{n}{n-1}\sigma_\eta^2(n)(I - \mathbf{1}\mathbf{1}^\top/n)$.

Relation to Term I. For any fixed target weights w_i^* with $\sum_i w_i^* = 1$:

$$
\text{Term I} = \sum_{i=1}^{n} (w_i^*)^2 \sigma_{\delta_i}^2. \qquad (22)
$$

Under uniform weights ($w_i^* = 1/n$) and homogeneous estimation noise ($\sigma_{\delta_i}^2 = \sigma_\delta^2(n)$), this simplifies to $\sigma_\delta^2(n)/n$. Thus Term I reflects averaged utility-estimation noise, while Term II reflects aggregation noise multiplied by satisfaction dispersion. This is why Term II can create large local score shifts in high-dispersion groups even when satisfaction estimation explains most average variance.

B.3 Irreducibility Analysis

Why Term II persists. A holistic judge that outputs a scalar $\hat{R}$ for a multi-stakeholder response can be modeled as implicitly computing weights $\hat{w}_i$ (the marginal influence of each stakeholder's satisfaction on the final score). These weights encode a normative judgment—whose needs matter more—for which there is no objective target to infer from the response alone. In stochastic LLM judging, when no explicit constraint fixes $\hat{w}_i = w_i^*$, repeated evaluations can yield varying implicit weights. Under the symmetric noise model of Proposition 1, $\mathrm{Cov}(\boldsymbol{\eta})$ has rank $n-1$ (the full zero-sum subspace), so any non-constant satisfaction vector ($\mathrm{Var}_i[u_i^*] > 0$) has a nonzero projection onto the weight-noise subspace and yields a positive Term II.

Rubric and Checklist. These methods structure the *estimation* task by decomposing it into dimensions (accuracy, feasibility, constraint satisfaction, etc.). However, these dimensions cut *across* stakeholders rather than *between* them. When scoring "constraint satisfaction" for a 4-person group, the judge still implicitly decides whose constraints matter more, pushing aggregation one level deeper. Formally, within each dimension d, the judge computes $\hat{R}_d = \sum_i \hat{w}_{i|d} \cdot \hat{u}_{i,d}$ with stochastic $\hat{w}_{i|d}$, preserving Term II at the sub-dimension level.

Ensemble/Repeated Sampling. Averaging M independent judge calls reduces *stochastic* variance by $1/M$: $\mathrm{Var}[\bar{R}_M] = \mathrm{Var}[\hat{R}]/M$. As $M \to \infty$, the stochastic component of Term II vanishes, but $\bar{R}_M$ converges to the judge's expected aggregation rule, $E[\hat{R}] = \sum_i E[\hat{w}_i] u_i^*$, which need not equal

the designer's intended $\sum_i w_i^* u_i^*$. The resulting bias cannot be removed by averaging.

Debate/Deliberation. Multi-agent debate can surface reasoning that a single judge might miss, potentially reducing σ_δ^2. However, unless the protocol explicitly separates per-role estimation from cross-role aggregation, the final arbiter still produces a holistic score through implicit stakeholder weights, so Term II persists.

B.4 Proof of Proposition 2 (Advantage Sign Correctness)

B.4.1 Part 1: Derivation of Eq. 6

Setup. Consider a GRPO group of G candidate responses with target rewards R_j^* and group mean $\bar{R}^* = \frac{1}{G}\sum_{k=1}^{G} R_k^*$. The **true reward gap** for candidate j is

$$\Delta_j^* = R_j^* - \bar{R}^*. \tag{23}$$

The judge observes a noisy reward

$$\hat{R}_j = R_j^* + \xi_j, \tag{24}$$

where the reward errors are independent Gaussian variables $\xi_j \sim \mathcal{N}(0, \sigma_\xi^2)$. GRPO uses the group-relative observed gap

$$\begin{aligned}\hat{\Delta}_j &= \hat{R}_j - \bar{\hat{R}} \\ &= (R_j^* - \bar{R}^*) + (\xi_j - \bar{\xi}) \\ &= \Delta_j^* + \zeta_j, \end{aligned} \tag{25}$$

where $\zeta_j = \xi_j - \bar{\xi}$ is the group-centered **random error**. Thus $\hat{\Delta}_j$ is a noisy reward gap: the desired signal Δ_j^* plus judge instability.

Noise after group centering. Since $\bar{\xi} = \frac{1}{G}\sum_{k=1}^{G}\xi_k$, the group-centered error expands as

$$\begin{aligned}\zeta_j &= \xi_j - \bar{\xi} \\ &= \xi_j - \frac{1}{G}\sum_{k=1}^{G}\xi_k \\ &= \xi_j - \frac{1}{G}\left(\xi_j + \sum_{k\neq j}\xi_k\right) \\ &= \left(1 - \frac{1}{G}\right)\xi_j - \frac{1}{G}\sum_{k\neq j}\xi_k. \end{aligned} \tag{26}$$

For independent errors, all covariance terms vanish. Hence

$$\begin{aligned}\mathrm{Var}(\zeta_j) &= \mathrm{Var}\left(\left(1 - \frac{1}{G}\right)\xi_j - \frac{1}{G}\sum_{k\neq j}\xi_k\right) \\ &= \left(1 - \frac{1}{G}\right)^2 \mathrm{Var}(\xi_j) + \frac{1}{G^2}\sum_{k\neq j}\mathrm{Var}(\xi_k) \\ &= (1 - \tfrac{1}{G})^2\sigma_\xi^2 + (G-1)\tfrac{1}{G^2}\sigma_\xi^2 \\ &= \sigma_\xi^2(1 - 1/G). \end{aligned} \tag{27}$$

Because ζ_j is a linear combination of independent Gaussian errors,

$$\zeta_j \sim \mathcal{N}\big(0, \sigma_\xi^2(1 - 1/G)\big). \tag{28}$$

Sign correctness. Assume first that $\Delta_j^* > 0$; the case $\Delta_j^* < 0$ is symmetric. GRPO assigns the correct advantage sign when the observed gap remains positive:

$$\begin{aligned}P(\text{correct sign}) &= P(\hat{\Delta}_j > 0) \\ &= P(\Delta_j^* + \zeta_j > 0) \\ &= P(\zeta_j > -\Delta_j^*). \end{aligned} \tag{29}$$

Standardize the centered noise as

$$Z_j = \frac{\zeta_j}{\sigma_\xi\sqrt{1 - 1/G}} \sim \mathcal{N}(0, 1). \tag{30}$$

Then

$$\begin{aligned}P(\text{correct sign}) &= P\left(Z_j > -\frac{\Delta_j^*}{\sigma_\xi\sqrt{1 - 1/G}}\right) \\ &= \Phi\left(\frac{\Delta_j^*}{\sigma_\xi\sqrt{1 - 1/G}}\right). \end{aligned} \tag{31}$$

For $\Delta_j^* < 0$, the same expression holds with $|\Delta_j^*|$.

Evaluating at a typical response. To obtain a single summary statistic, evaluate a response whose true reward gap has magnitude one group standard deviation:

$$|\Delta_j^*| = \sqrt{\mathrm{Var}_j[\Delta_j^*]}. \tag{32}$$

Using the SNR definition in Eq. 5, $\mathrm{SNR} = \mathrm{Var}_j[\Delta_j^*]/\sigma_\xi^2$, gives

$$P(\text{correct sign}) = \Phi\left(\frac{\sqrt{\mathrm{SNR}}}{\sqrt{1 - 1/G}}\right). \tag{33}$$

This represents an optimistic estimate: the majority of responses in a group lie closer to the mean ($|\Delta_j^*| < \mathrm{sd}$) and experience correspondingly higher sign-flip rates.

B.4.2 Part 2: Conditional SNR Scaling with Stakeholder Count

This part analyzes the conditional limit where Term II is the leading contribution to reward-error variance. This is not assumed to be the average empirical regime; it isolates how aggregation noise affects GRPO when satisfaction dispersion is high. In this limit, $\sigma_\xi^2 \approx \text{Term II}$. By Proposition 1:

$$\sigma_\xi^2(n) \approx \frac{n^2}{n-1}\,\sigma_\eta^2(n)\,\mathrm{Var}_i[u_i^*] \tag{34}$$

Substituting into the SNR definition:

$$\begin{aligned}\mathrm{SNR}(n) &= \frac{\mathrm{Var}_j[\Delta_j^*]}{\sigma_\xi^2(n)} \\ &\approx \frac{(n-1)\,\mathrm{Var}_j[\Delta_j^*]}{n^2\,\sigma_\eta^2(n)\,\mathrm{Var}_i[u_i^*]}\end{aligned} \tag{35}$$

Thus SNR is controlled by the same factors in the denominator. In the high-drift, high-dispersion regime where $\sigma_\eta^2(n) = \Theta(1)$, $\mathrm{Var}_i[u_i^*] = \Theta(1)$, and the true reward-gap spread $\mathrm{Var}_j[\Delta_j^*]$ is bounded, this is $O(1/n)$; if either factor shrinks with n, degradation is weaker, and if either vanishes the Term II channel disappears. Eq. 33 depends on $\sqrt{\mathrm{SNR}}$, so its argument is $O(1/\sqrt{n})$ in the high-drift, high-dispersion regime. Using $\Phi(x) = 1/2 + O(x)$ gives $P(\text{correct sign}) = 1/2 + O(1/\sqrt{n})$.

B.4.3 Numerical Thresholds ($G = 8$)

With $1 - 1/G = 7/8$:

SNR	P(**correct sign**)
0.5	0.78
1.0	0.86
1.5	0.90
2.0	0.93
3.0	0.97
4.0	0.98

Under ideal i.i.d. noise, even moderate SNR appears sufficient. However, this represents a best-case bound: in practice, (1) noise is correlated across responses sharing the same prompt (reducing effective SNR), (2) the judge's noise is not Gaussian but exhibits systematic biases (e.g., position bias, length bias), and (3) the "typical response" assumption ($|\Delta_j| = \sigma_{\text{signal}}$) is optimistic—many responses have smaller true differences. These factors explain why empirically observed training instability occurs at SNR values that the i.i.d. theory would predict as sufficient.

C End-to-End Training Details

Base model and SFT. We use Qwen3-30B-A3B-Thinking-2507 (Qwen Team, 2025) as the base model. Expert trajectories are generated by Qwen3.5-Plus on multi-stakeholder travel-planning queries and used for supervised fine-tuning before RL.

GRPO training. All four reward configurations (DECOMPR, Direct, Rubric, Checklist) share the same GRPO training setup: 6,000 training queries, with the best checkpoint selected within 400 training steps. The training queries are distinct from the 60 seed responses used in the reward-consistency analysis (§3) and from the 650 MR-TravelBench evaluation tasks.

Evaluation. MR-TravelBench contains 650 held-out multi-stakeholder travel-planning tasks spanning varying stakeholder counts and constraint difficulties. Each policy is evaluated with 3 independent trials per task.