
Discarding Answer Ranks Changes the Case for Mixing Biomedical LLMs

Odysseus Kaloidas
UC Berkeley

Abstract

1 Should a biomedical reasoning system spend eight calls on one model or several?
2 When calls return ranked shortlists, counting every answer equally can weaken
3 the one-model reference and make mixing look better by comparison. We hold
4 these lists fixed in a 288-question, four-model study. Counting every shortlisted
5 answer rather than only the first shifts the mixed-minus-Gemma accuracy gap
6 by 9.72 percentage points (95% confidence interval [5.29, 14.21]). Using ranks
7 to resolve inclusion ties accounts for 7.62 points; the remaining contrast is un-
8 resolved. This relative shift need not mean that mixing gives better answers.
9 In a separate prespecified 720-question MedXpertQA test, mixing Gemma and
10 Llama3.3-70B loses 8.95 points [−11.16, −6.75] against development-selected
11 Gemma first-choice voting. The mixture already has lower per-call accuracy. On
12 256 MMLU-Pro Health questions, mixing improves inclusion-vote accuracy by
13 5.00 points, yet remains far below the selected first-choice policy. On Health,
14 first-choice voting also exceeds a ceiling for selectors restricted to listed answers
15 that use only which calls mention them and favor no answer label. Discarding
16 ranks changes the apparent benefit of mixing; these comparisons do not isolate
17 quality-independent diversity.

18 1 Introduction

19 A biomedical reasoning system can spend its call budget on one model or several. When each call
20 returns a ranked list, the final answer depends on both which models are called (*allocation*) and how
21 their lists are combined (*aggregation*). Together, these choices define a *policy*.

22 Mentioning an answer is not preferring it. A model may repeatedly list the correct answer first
23 and the same alternatives below it. Counting every appearance treats those alternatives as equally
24 supported. A mixture can then look relatively favorable because aggregation has weakened the one-
25 model reference, not because mixing has produced a better decision. Ranked clinical collectives
26 likewise distinguish consideration sets from decisions [Zöller et al., 2025].

27 We test this possibility by holding responses fixed within each allocation and changing whether
28 aggregation retains ranks. Five-answer inclusion extends top-two approval [Wang et al., 2025, Ap-
29 pendix C], but discards more information and uses uniform ties. This diagnoses information loss,
30 not the published procedure’s validity.

31 We separate three questions: does changing the rule make mixing look more favorable, does mixing
32 beat repeated sampling under that rule, and does it beat a strong policy chosen before testing? We
33 study two biomedical benchmarks with official answer keys; clinical-name mapping is exploratory.
34 The models are not quality matched, and patient care is outside the evaluation.

2 Related Work

Self-MoA shows that weaker partners can dilute strong-model sampling [Li et al., 2026b]. Capability, debate and composition studies likewise motivate strong homogeneous baselines [Kim et al., 2026, Smit et al., 2024, Wang et al., 2026, Abdulaal et al., 2025]. Diversity can help executable-patch recovery [Zhang et al., 2025], while voting and communication differ [Choi et al., 2025]. Unlike a dilution comparison, our fixed-response intervention tests unequal losses imposed by aggregation.

Ranked Voting based Self-Consistency varies rules and list length [Wang et al., 2025]; clinical aggregation supplies biomedical precedents [Barabucci et al., 2024, Yuan et al., 2026, Zöller et al., 2025, Kumthekar et al., 2025]. Reverse-anchored aggregation adds Bayesian evidence [Chen et al., 2026b]. We supply no external evidence or new rule: we measure inherited rules’ effects on fixed responses. Dependence studies [Jiang et al., 2025, Kim et al., 2025, Eriksson et al., 2026, Kuai et al., 2026] motivate this question without establishing useful diversity from similarity alone.

Having a correct candidate and selecting it are different achievements [Ji et al., 2026, Hu, 2026, Li et al., 2026a]. The selection-bottleneck account predicts how accurate a selector must be for diversity to help, assuming approximately stable homogeneous-team performance [Maryanskyy et al., 2026]. Fusion-of-N instead combines complementary content into a new answer [Khairi et al., 2026]; biomedical aggregation and capability-controlled audits also test downstream outcomes [Chen et al., 2026a, Kim, 2026]. Our distinction is between two final-answer rules on identical ballots and, separately, performance against a development-selected policy.

Quality-compatible selection in Determine-Then-Ensemble and DFPE, matched-quality fusion, and tool diversity in TUMIX can benefit mixing [Yao et al., 2025, Cohen et al., 2026, Chen, 2026a, Chen et al., 2026d]. Our models are not quality matched; changing ranks while retaining candidates does not isolate lineage.

3 Methods

We compare eight calls to one model with eight calls spread across models, holding each allocation’s responses fixed across rules. MedXpertQA [Zuo et al., 2025] and MMLU-Pro Health use official answer keys (*native keys*), withheld from model prompts. MedEinst is a historically prespecified secondary study, interpreted as exploratory because clinical-name mappings remain unresolved.

A *call* produces one five-answer list. A *bank* is a replicate batch of calls for one question, shared across allocations as specified below. Central studies use two banks with distinct hashed seeds. We score each allocation and rule within each bank, then average the two scores into one question-level observation. MedEinst also averages control and misleading-evidence record *branches* within each pair. Banks and branches are not independent cases.

The 1,264 central questions form three studies: a 288-question rule test (Native288), a 720-question selected-policy confirmation (Native720), and a 256-question test on MMLU-Pro Health (Health256). Both native cohorts use MedXpertQA. Table 1 separates these from development and secondary tests. Replays reuse responses, not independent cases.

Paired allocation. In Native288, the homogeneous allocation H uses eight Gemma calls. The mixed allocation M reuses two and adds two each from Llama8B, Ministral14B and Qwen14B: 14 lists support two eight-call policies per bank. Earlier per-call accuracy selected Gemma. MedEinst development [Chen et al., 2026c] used synthetic DDXPlus records [Tchango et al., 2022] (Appendix G).

Repeated answers vote across calls; within-call duplicates count only at their best rank. LLM selectors see each candidate once with its best-ranked occurrence’s rationale and deterministic hash ties. Content, pool size and quality can differ across allocations. Calls, not tokens or dollars, are matched.

Generation and selection settings. Both allocations use identical system/user prompt templates and decoding: temperature 1, top- $p = .9$, a 384-token cap and a request for five answers with rationales of at most 18 words. Selectors use temperature 0 and top- $p = 1$. Hashed seeds pair conditions without guaranteeing provider determinism. No retrieval or tools are used.

Table 1: Study map: 1,760 unique units (240 MedEinst pairs, 1,200 MedXpertQA and 320 Health questions). Pilots/retired assignments excluded; reused subsets add no cases. H/H_G : eight Gemma calls; M: two/model across four models; H_L : eight Llama3.3-70B calls; M_{44} : four from each.

Source / distinct cases	Allocation	Role and reuse
Native288 / 288 new	Four-model H/M	Initial rule test: does the gap change? Primary C .
Native720 / 720 new	H_G, H_L, M_{44}	Policy confirmation: does mixing beat Gemma? Primary θ .
Health256 / 256 new	H_G, H_L, M_{44}	Source transport: does C persist? Primary C_G ; utility secondary.
MedEinst / 240 pairs	Four-model H/M	Disjoint: 48 development, 96 representation, 96 selector.
Earlier native / 128	Four-model H/M	Selector test; predictor training. Its 48-case subset selects the later two-model policy.
Direct-answer / 128 new	H_G, M_{44}	64 native, 64 Health; one-answer versus five-answer prompts.
Common pool / reused	Four-model H/M	Earlier 96 MedEinst + 128 native selector cases; support reversal.

85 Original generators are Gemma4-26B-A4B-IT [Google DeepMind, 2026], Llama3.1-8B-Instruct
86 [Meta, 2024a], Ministral-14B-2512 [Mistral AI, 2025] and Qwen2.5-14B-Instruct [Qwen Team,
87 2024]. Selectors are Gemini2.5-Flash-Lite [Google, 2025] and Llama3.1-8B; Gemma3-12B is
88 development-only. Serving settings are documented in Appendix E; endpoint names do not pin
89 every revision.

90 **Selection rules.** Inclusion votes for every listed candidate; first-choice votes only for the top an-
91 swer. Reciprocal rank (RR) weights rank r by $1/r$; Borda weights ranks 1–5 by 5–1. Tie-aware
92 frequency counts inclusion votes, then uses RR to resolve equal counts. All select a final answer,
93 rather than measuring availability.

94 For remaining ties, we score the expected accuracy of choosing uniformly among tied winners:
95 one correct answer among three tied receives $1/3$. In contrast, Wang et al. [2025] use token log
96 probabilities for open models or another model decision for closed models. First-choice replay
97 applies self-consistency [Wang et al., 2023] to lists; a separate direct-answer test checks the prompt
98 boundary. Only LLM selectors read case text and rationales.

99 For a worked illustration, let five calls return [A,B,C,D,E] and three return [B,A,C,D,E], with A
100 correct. Inclusion gives every candidate eight votes and expected accuracy $1/5$. First-choice counts
101 are A:5, B:3, so A wins. RR gives A $5 + 3/2 = 6.5$ and B $5/2 + 3 = 5.5$; it selects A both
102 as a full rule and as an inclusion tie breaker. Borda also selects A, with 37 versus B’s 35 points.
103 All four rank-using decisions are correct on these same eight lists. Candidate availability is 100%
104 throughout. This constructed example explains the rules, not the size of an empirical effect.

105 **Estimands and uncertainty.** Let $F_{p,r}$ denote final accuracy for allocation p under rule r . We ask
106 how changing the rule changes the mixed-minus-homogeneous accuracy difference.

107 The primary measure for the rule comparison is

$$C = (F_{M,\text{incl5}} - F_{H,\text{incl5}}) - (F_{M,\text{first}} - F_{H,\text{first}}). \quad (1)$$

108 Thus a positive C means that mixing is relatively more favorable under inclusion than first-choice
109 voting. It can reflect a smaller loss under inclusion, rather than an accuracy gain. Absolute accura-
110 cies and the two allocation differences are therefore reported alongside C . Unless another target is
111 specified, an *unresolved* contrast has a confidence interval containing zero; this does not establish
112 no effect.

113 The prespecified sample contains 288 unused MedXpertQA items: 96 each in Diagnosis, Treatment
114 and Basic Science. A separate 96-pair MedEinst test also excludes prior IDs and matching inputs.
115 Native C is the sole primary. We bootstrap questions 50,000 times within task, keeping conditions,
116 shared calls and banks together; MedEinst resamples whole pairs. These intervals describe vari-
117 ation across questions under the observed banks, not across fresh full-generation runs. Percentile

Table 2: Final accuracy (%), averaging two banks; MedEinst also averages branches under the original frozen refinement mapping. Native288/MedEinst96 use H/M; Health256 uses H_G/M_{44} . Uniform ties and empty failures are retained.

Rule	Native288		MedEinst		Health	
	H	M	H	M	H_G	M_{44}
Inclusion frequency	18.68	17.45	16.89	13.87	23.48	28.48
First choice	29.86	18.91	30.56	22.69	68.82	67.68
Frequency + RR ties	29.51	20.66	30.47	15.76	69.04	68.07
Reciprocal rank	29.69	19.97	31.08	22.72	69.04	68.26
Borda	29.17	20.98	31.64	20.57	69.53	68.55

intervals condition on answer keys or scoring maps; pointwise intervals do not adjust for multiple comparisons. Five points is a prespecified benchmark target, not a clinically justified benefit threshold; the separate five-point half-width target defines precision. Terminal failures remain empty after one retry and contribute no votes; an all-empty allocation scores zero. Their cases remain in the analysis. The largest source/model terminal-failure rate is 3.73%, below the prespecified 5% failure limit (Appendix A).

Frozen final-decision design. On 48 development questions, Llama3.3-70B [Meta, 2024b] accuracy was 22.14% versus Gemma’s 29.69%: above the 20% competence floor and within ten points, not quality equivalent. Comparing two homogeneous allocations under five rules selected eight Gemma calls with first-choice voting.

Before new outputs, we froze 720 unused native questions (240/task), two banks, eight calls to Gemma (H_G), eight to Llama3.3-70B (H_L) and their four-plus-four mixture (M_{44}): 16 unique calls/bank with the same five-answer contract. The sole primary estimand, $\theta = F_{M_{44}, \text{first}} - F_{H_G, \text{first}}$, is first-choice accuracy of the mixture minus that of Gemma. We estimate uncertainty with 50,000 task-stratified whole-case resamples. The benchmark improvement target is +5 points; the precision target is a three-point interval half-width. Both C contrasts are secondary. The mixture uses the first four scheduled slots from each model, retaining failures rather than selecting successful calls.

Thus θ compares policies; C compares their gaps under different rules. A uniformly sampled Gemma slot averages 34.48% accuracy and approximately \$0.00010/question, versus 35.21% and \$0.00078 for eight-call voting. Both use the same ranked-list contract.

Health transport freezes 256 hash-selected MMLU-Pro Health questions [Wang et al., 2024] from 506 eligible ten-option items, excluding normalized overlaps. It uses the same allocations, two banks (8,192 scheduled call slots), native keys and DeepInfra/Novita serving. Primary C_G (C against Gemma) uses 50,000 whole-question draws with a five-point half-width target; utility is secondary against unchanged Gemma. Source and serving changes prevent source-only attribution. New saved-output controls use 20,000 paired question resamples and separately declared exploratory families; they do not replace frozen primary tests or their intervals.

Data provenance and scoring. Pilots and retired cases are excluded (Appendix A.2). Native scoring credits the selected option’s match to the official key, not the truth of its rationale. It measures recovery of one keyed answer, not how many distinct hypotheses are scientifically valid. MedEinst strict scoring credits synonymous names; refinement scoring adds whitelisted timing, cause, severity or site differences. Acute Bronchitis against Bronchitis receives refinement credit; Pleuritis against Spontaneous pneumothorax remains unresolved. Of 1,770 development terminology relations, 675 remain flagged. Neither mappings nor native keys establish clinical truth (Appendix F).

4 Results

4.1 Does changing the rule change the comparison?

In the four-model, 288-question study, mixing loses less under inclusion than first-choice voting. Its accuracy gap against repeated Gemma sampling is -1.23 points $[-3.91, 1.50]$ under inclusion, versus -10.95 $[-15.55, -6.39]$ under first-choice. The difference between these gaps is the primary

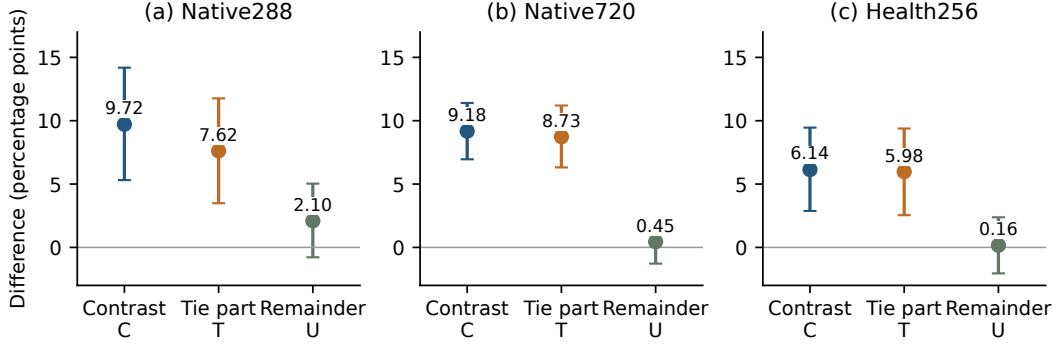


Figure 1: Why the allocation comparison changes: $C = T + U$. Here T is how much more H gains than M when ranks resolve inclusion ties; U is the contrast remaining afterward. Native288 mixes four models; Native720/Health mix Gemma and Llama3.3-70B. Points and paired pointwise 95% intervals use a common exploratory 20,000-question-bootstrap replay, preserving task strata and averaging banks. All remainders remain unresolved after six-component adjustment (Appendix H).

157 result: $C = +9.72$ percentage points (95% CI [5.29, 14.21]), above the prespecified five-point
 158 margin. This changes the allocation comparison, not its sign. The other three rules also favor H
 159 (Table 2).

160 Both banks give positive effects (9.36/10.08 points, positive intervals). Removing failed-call ques-
 161 tions leaves $C = 8.86$ [3.32, 14.49] on 193 questions; the primary retains failures.

162 MedEinst’s secondary contrast is 4.85 [.77, 9.08] points under the implemented refinement mapping.
 163 That implementation credited ten newly unresolved relation types contrary to the protocol prose;
 164 zeroing them gives 4.93 [.87, 9.17], while strict synonyms give 6.10 [2.01, 10.38]. All 96 pairs
 165 contain unresolved relations. Assigning each unresolved relation type a consistent correctness status
 166 gives a worst-case range of [−26.34, 29.85]. This varies scoring assumptions, not sampled questions,
 167 and is not clinical adjudication. The native-key results do not depend on this map (Appendix A).

168 4.2 Does mixing beat the selected policy?

169 On 720 new questions, mixed first-choice voting reaches 26.26% accuracy, versus 35.21% for
 170 the development-selected Gemma policy (Table 3). The primary difference is $\theta = -8.95$ points
 171 [−11.16, −6.75]: mixing loses, rather than meeting the prespecified five-point benefit. Both banks
 172 agree in direction (−9.24 and −8.67 points). The interval half-width, 2.21 points, meets the three-
 173 point precision target. All 111 terminal failures among 23,040 scheduled call slots remain empty;
 174 native option order is unchanged. The manifest contains 23,449 attempts including retries.

175 Nevertheless, the earlier rule-dependent comparison persists against Gemma: secondary $C =$
 176 9.18 [6.98, 11.43]. Against the weaker Llama reference it is −1.96 [−3.94, 0.01]. A positive rule
 177 interaction and a negative final-policy result can therefore coexist.

178 Across 1,440 question–bank records, mixed pools contain the key at any rank in 1,327 (92.15%),
 179 versus Gemma’s 1,291 (89.65%). But first-choice voting cannot select an answer ranked below first
 180 in every list. The key is ranked first somewhere in only 709 mixed records (49.24%) and 671 Gemma
 181 records (46.60%).

182 Among these eligible records, conditional selection success is 53.34% for mixing and 75.57% for
 183 Gemma: expected correct selections divided by the respective eligible counts, with fractional credit
 184 for uniform ties. Questions, not bank records, remain the inference units. Greater access need not
 185 yield better decisions. The eligible sets differ, so these conditional rates do not isolate selector skill
 186 on identical inputs (Appendix B.4).

187 The post-hoc task breakdown gives negative θ and positive C in Diagnosis, Treatment and Basic
 188 Science after six-contrast adjustment (Appendix A.2).

189 Excluding failed-call questions leaves $\theta = -8.35$ [−10.62, −6.09] on 648 cases. Costs include
 190 retries and count shared calls once per policy.

Table 3: 720-question confirmation, averaging banks; accuracy/availability in percent, ties uniform. H_G : eight Gemma calls; H_L : eight Llama3.3-70B calls; M_{44} : four each. Resources: mean per-question policy totals including retries. Only θ is primary (+5pp benefit margin). Question-bootstrap intervals retain banks/task strata. Per-call accuracy and aggregation-lift contrasts are exploratory.

Rule or resource	H_G	H_L	M_{44}
Inclusion frequency	19.49	15.38	19.71
First choice	35.21	19.97	26.26
Frequency + RR ties	34.34	20.03	25.83
Reciprocal rank	34.65	19.97	24.86
Borda	35.13	19.76	25.06
Per-call first-choice	34.48	20.17	27.17
All-rank key availability	89.65	78.26	92.15
Input tokens	4,687.4	4,328.1	4,512.1
Output tokens	1,192.6	642.0	916.8
Cost (USD)	0.00078	0.00131	0.00104
$M_{44} - H_G$ (percentage points)		Estimate	95% CI
Final first-choice θ (primary)		-8.95	[-11.16, -6.75]
Per-call accuracy Δq		-7.31	[-8.85, -5.76]
Aggregation lift ΔL		-1.64	[-2.89, -.41]

191 The native deficit survives other tie choices. Replaying published Borda and reciprocal-rank code
192 on these same lists gives mixed-minus-Gemma differences of -9.10 and -9.51 points, respectively
193 (six-contrast adjusted intervals exclude zero); Health remains unresolved. Even resolving all rank-
194 score ties in mixing’s favor and against Gemma leaves Native720 deficits of at least 7.92 points
195 for Borda and 8.12 for reciprocal rank. These are fixed-output bounds, not confidence limits or
196 achievable policies. This tests aggregation components, not the source paper’s generation procedure
197 (Appendix H.8).

198 How much of this deficit precedes voting? Let q_p be average per-call first-choice accuracy, including
199 failures, and $L_p = F_{p,\text{first}} - q_p$ the gain from aggregation. Gemma gains $35.21 - 34.48 = 0.73$ points
200 (paired 95% CI [-0.01, 1.49]); mixing changes by $26.26 - 27.17 = -0.91$ [-1.82, .03]. Neither gain
201 is resolved. With Δ denoting mixed minus Gemma, $\theta = \Delta q + \Delta L = -7.31 - 1.64 = -8.95$:
202 most of the deficit is already present in per-call accuracy. The remaining -1.64 is accounting, not a
203 quality-controlled causal effect of diversity (Figure 6). These baselines use stored five-answer slots,
204 not new direct answers. Against weaker Llama, mixing instead gains 6.30 [4.50, 8.13] points, with
205 unresolved lift difference $-.70$ [-1.71, .32].

206 **Transport at higher answer accuracy.** Health supplies a genuine within-rule gain: mixing raises
207 inclusion-vote accuracy from 23.48% to 28.48%, a 5.00-point improvement [3.58, 6.43]. But both
208 are far below the selected Gemma first-choice policy’s 68.82% accuracy. Under first-choice voting,
209 the mixed-minus-Gemma difference is unresolved at -1.14 [-4.20, 1.95]; this is not evidence of
210 equivalence.

211 The resulting primary rule contrast, $C_G = 6.14$ [2.84, 9.45] points, reproduces the positive direction
212 on a second source, without establishing an effect above five points. Per-call model accuracies are
213 68.73% and 64.70%, still not matched. All 55 failures remain in the primary; excluding affected
214 questions retains a positive interval (Appendix A.3).

215 4.3 What explains the differences?

216 We next test how ties, informative ranks and agreement on wrong answers affect the comparison.
217 These analyses reuse the three cohorts, not independent cases. Changing the cutoff tests how many
218 ranks count; shuffling and transfer test the information in their order while retaining candidates.
219 The ceiling and wrong-option relabeling use answer keys. Cross-transfer needs twelve calls; other
220 empirical replays retain eight per policy.

221 On Native288, resolving inclusion ties with RR raises H accuracy from 18.68% to 29.51%, but
222 M accuracy only from 17.45% to 20.66%. Gemma recovers 10.83 points, versus 3.21 for mixing.

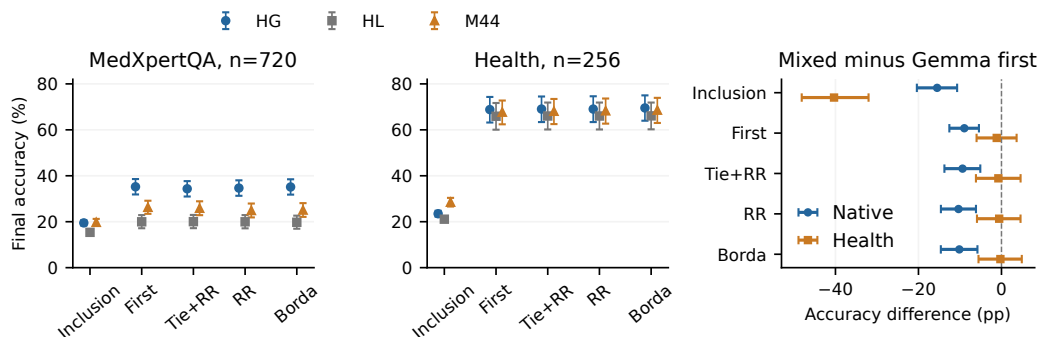


Figure 2: Absolute policy performance, averaging banks within question. Left/middle: final accuracy and pointwise 95% question-bootstrap intervals. Right: mixed policies versus selected H_G first-choice, with exploratory Bonferroni-adjusted intervals across 28 policy–source comparisons. Unresolved differences do not establish equivalence.

223 This unequal rescue gives $T = 7.62$ [3.54, 11.81] of the 9.72-point contrast. After this repair,
 224 the remaining change in the mixed-minus-Gemma gap relative to first-choice voting is $U = 2.10$
 225 $[-.80, 5.03]$: $C = T + U$. Inclusion ties affect 94.3% of H decisions and 66.8% of M decisions.
 226 In the later cohorts, they obscure a correct unanimous Gemma first choice in at least one bank on
 227 184/720 native and 166/256 Health questions.

228 The same decomposition gives $T = 8.73$ and $U = .45$ points on Native720, and $T = 5.98$ and
 229 $U = .16$ on Health (Figure 1). Here C also equals the loss from discarding ranks in H minus the
 230 loss in M. Resolving inclusion ties removes most of the point-estimated contrast, but an unresolved
 231 remainder does not establish zero effect. MedEinst’s negative remainder is mapping-conditional
 232 (Appendix H).

233 Is this only a consequence of counting all five answers? An exploratory replay approves just
 234 the first two. On Health, H/M accuracy is 51.56/58.04% with uniform ties, but 69.43/68.26%
 235 when full-list RR resolves tied counts. Mixing’s 6.48-point advantage becomes an unresolved
 236 -1.17 $[-4.79, 2.44]$. The cutoff-two counterpart of C is 7.62 [2.96, 12.21] on Health, but unre-
 237 solved on Native288 (3.79 $[-.41, 7.93]$) and Native720 (2.28 $[-.45, 5.05]$), with three-contrast ad-
 238 justment. This is a cutoff change on existing five-answer outputs, not fresh two-answer generation
 239 or the published tie procedure (Appendix H.7).

240 Tie rescue uses ranks, but does their ordering carry useful information? Randomly reordering each
 241 list keeps candidates, calls and failures fixed but damages both policies: Native720 accuracy falls
 242 from 35.21% to 16.71% for Gemma and from 26.26% to 16.19% for the mixture. Gemma loses
 243 more, reducing its advantage from 8.95 to .52 points. Figure 6c shows the paired gap change and
 244 unresolved Health result. Preferences matter beyond membership, but randomization is not a reme-
 245 dy or a quality control. Intervals adjust across eight endpoints and condition on simulated means
 246 (Appendix B.1).

247 Could a better membership-only rule recover the lost information? Such a rule sees which calls list
 248 each answer, but not its text or rank. A label-neutral selector merely renames its output probab-
 249 ilities when candidates are renamed. In the worked example, all five candidates occur in all eight
 250 calls. Swapping A with any alternative leaves this input unchanged, so accuracy cannot exceed $1/5$.
 251 Generally, if the keyed answer (*gold*) appears in exactly the same calls as $k - 1$ alternatives, their
 252 selection probabilities must be equal. The ceiling is $1/k$, or zero if gold is absent, for selectors
 253 restricted to listed candidates. This argument requires no independent-error assumption. Computing
 254 the ceiling uses gold; a usable selector need not attain it.

255 On Health, the mean ceiling is 30.28% for Gemma and 38.42% for mixing, versus first-choice
 256 accuracies 68.82% and 67.68%. The excesses are 38.54 [28.86, 47.46] and 29.25 [19.84, 37.83]
 257 points (eight-comparison adjustment). Native720 has no such aggregate separation (Figure 3). The
 258 limit concerns restricted inputs, not biomedical competence. Health accuracy with ranks exceeds it;
 259 selectors reading questions, answers, rationales or token probabilities lie outside the bound.

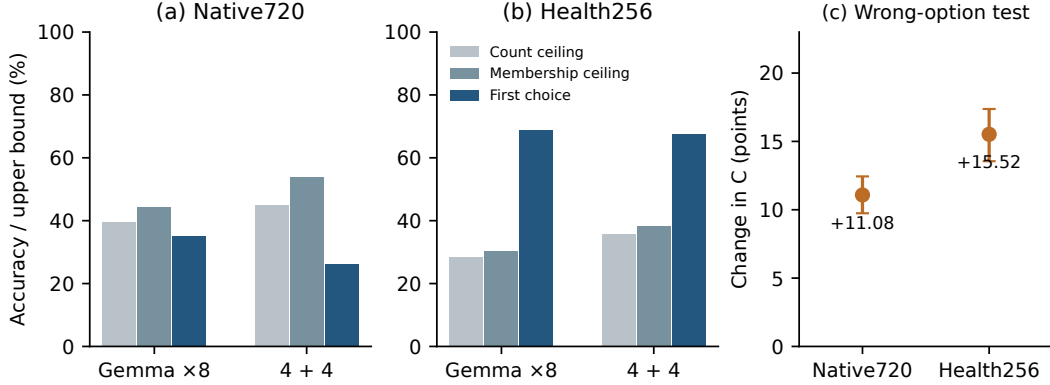


Figure 3: What membership loses and wrong-answer agreement changes. (a,b) Gold-informed symmetry ceilings for selectors seeing total mention counts or which calls mention each answer, compared with observed first-choice accuracy. Bars are point estimates; paired differences and adjusted intervals are in Appendix H. (c) Change in C after model-specific wrong-option relabeling, with question-bootstrap intervals adjusted over ten contrasts and conditioned on 256-draw means. These are not permutation reference ranges or fresh confirmation.

Table 4: Rank transfer on fixed candidate sets. Recipient accuracy after self/cross transfer is percent; cross-minus-self is percentage points with intervals adjusted over four contrasts. Original Gemma/mixed accuracies: 35.21/26.26 on Native720, 68.82/67.68 on Health256. Using the recipient’s own ranks controls for reranking alone.

Source	Recipient	Own ranks	Other ranks	Change [adjusted 95%]
Native720	Gemma	34.69	25.59	−9.10 [−12.47, −5.80]
	Mixed	25.07	33.61	8.54 [5.35, 11.74]
Health256	Gemma	68.95	69.14	.20 [−4.69, 4.88]
	Mixed	69.04	68.36	−.68 [−5.57, 4.30]

Does agreement on wrong answers matter even at fixed correctness? A common wrong-option permutation across models leaves all policy accuracies unchanged. Separate permutations for each question/model instead change cross-model agreement. Each fixes gold and is shared across that model’s calls and both banks, preserving correct ranks, failures and homogeneous scores. Over 256 synthetic relabelings, C increases by 11.08 points [9.74, 12.44] on Native720 and 15.52 [13.55, 17.37] on Health (ten-contrast adjustment); first-choice mixing changes by only .21 and .04 points. Wrong-answer alignment therefore changes the inclusion comparison with correctness fixed. This gold-informed intervention breaks option semantics: it diagnoses the rule, not a usable aggregation method (Appendix H).

Removing ranks can hurt; transferring them tests whether another allocation’s preferences can help without adding candidates. We sum reciprocal-rank scores across a donor pool, then reorder donor-present candidates within each recipient list. If the donor contains A, C and E from [A,B,C,D,E], scoring C above A above E, the list becomes [C,B,A,D,E]. Unmatched candidates stay in place; no answer key is used. First-choice voting selects from these reordered lists. Comparing the recipient’s own pool as donor (self-transfer) with the other allocation (cross-transfer) controls for changes caused by pooling ranks itself.

On Native720, cross-transfer improves mixed accuracy by 8.54 points and reduces Gemma accuracy by 9.10, relative to self-transfer (Table 4). Both Health contrasts remain unresolved. Cross-transfer uses twelve unique calls per bank because the allocations share four. Averaging over a shared random ordering of tied donor scores retains the Native720 effects (−9.08/ +8.63 points); using mention counts instead of donor ranks does not. This supports a role for donor ranks under this operator, without isolating lineage. At eight calls, four Llama proposals with four Gemma donor calls reach 32.40%, versus 35.21% for eight Gemma calls. None of four exploratory allocations beats that reference on either source (Appendix H.5).

4.4 Robustness and secondary analyses

Direct-answer generation boundary. A frozen test asks whether requesting a shortlist changes the mixed-minus-Gemma accuracy gap. It requests one answer ($K1$) or five ($K5$) on 128 new questions, 64/source. Both use first-choice voting, the same models, banks, eight calls and token cap. Primary $R = \theta_{K5} - \theta_{K1}$ compares these gaps, averaging sources equally. With 50,000 within-source paired resamples, $R = +2.73 [-.78, 6.45]$ points is unresolved. Its 3.61-point half-width misses the pilot-based three-point target.

Direct-answer mixing still loses on native questions: $\theta = -12.50 [-19.53, -5.47]$; Health is unresolved at $-3.91 [-10.16, 2.34]$ (both secondary). The native loss extends beyond shortlists, but its size is not proven unchanged. Rationale count also changes, so reasoning effort is not isolated (Appendix A.4).

Partner checks reuse Native288: four Gemma calls versus two Gemma and two Llama, Ministral or Qwen calls. Each partner gives a positive rule contrast but a negative first-choice gap; no inclusion gain resolves after nine-contrast adjustment. This is neither independent nor quality-matched confirmation (Appendix C.3).

Expanding development selection to all 15 allocation–rule combinations again selects H_G first-choice. None of 28 source-specific comparisons beats it with a positive adjusted lower bound (Appendices C.2 and H).

Secondary LLM-selector tests. Earlier support-count shuffling and order reversal did not establish a native-key benefit [Chen, 2026b]. On eight of 128 native questions, only the mixed pool contained the correct answer; neither selector chose it (Appendix D.1).

5 Discussion and Limitations

Inclusion favors mixing relatively by discarding more useful Gemma ranks. Health mixing improves inclusion accuracy, yet neither inclusion policy approaches selected first-choice voting. Against that reference, mixing loses on Native720 and is unresolved on Health. Relative advantage is not policy improvement.

Compare complete policies against a competent reference chosen before testing, reporting absolute accuracy, per-call quality and costs. On Native720, eight Gemma calls have no resolved gain over one stored five-answer call.

Five-answer inclusion with uniform ties is especially lossy; approving two answers attenuates native contrasts. Component replay strengthens the native deficit. The direct-answer test changes generation, not the fixed-list rule; its unresolved result establishes neither format invariance nor persistence under a published procedure. Testing that procedure requires its generation and tie resolution, not more direct-answer calls. Whether the interaction survives remains open; failed transport would limit practical scope. Balanced option presentations remain untested.

Neither correctness-preserving relabeling nor per-call subtraction isolates a training-lineage effect. Sampling intervals exclude mapping uncertainty, key errors, contamination and serving variation. Evidence covers two native-key benchmarks and one synthetic-record source, not quality-equivalent populations. Voting and abstention results establish neither clinical safety nor a deployment threshold.

6 Conclusion

Discarding ranks disproportionately damages the homogeneous reference; resolving ties removes most of the point-estimated interaction. On Health, rank-based accuracy exceeds a ceiling for selectors that see only which calls mention each answer and favor no label. Yet mixing loses against selected first-choice voting on Native720 and is unresolved on Health. Comparing complete policies separates this information-loss diagnosis from an ensemble improvement claim.

References

- Ahmed Abdulaal, Chen Jin, Nina Montaña-Brown, Aryo Pradipta Gema, Daniel Coelho de Castro, Daniel C. Alexander, Philip Teare, Tom Diethe, Dino Oglic, and Amrutha Saseendran. Balancing act: Diversity and consistency in large language model ensembles. In *International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=D16nkKKv1X>.
- Junade Ali. Quantifying diversity of thought: A predictive law of weighted LLM ensemble lift, 2026. URL <https://arxiv.org/abs/2607.17384>.
- Gioele Barabucci, Victor Shia, Eugene Chu, Benjamin Harack, and Nathan Fu. Combining multiple large language models improves diagnostic accuracy. *NEJM AI*, 1(11):AIcs2400502, 2024. doi: 10.1056/AIcs2400502.
- Boqi Chen, Xudong Liu, Jiachuan Peng, Marianne Frey-Marti, Kyle Lam, Bang Zheng, Lin Li, and Jianing Qiu. MEDSYN: Benchmarking multi-evidence synthesis in complex clinical cases for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026a. doi: 10.18653/v1/2026.findings-acl.1183.
- Josef Chen. When does combining language models help? a co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models, 2026a. URL <https://arxiv.org/abs/2606.27288>.
- Ken Chen, Wei Wang, Sachith Seneviratne, Hansani Weeratunge, and Saman Halgamuge. When agents disagree: Bayesian backward reasoning as a label-free anchor for multi-agent collective decision-making, 2026b. URL <https://arxiv.org/abs/2609.11709>. Preprint, version 1, 10 September 2026.
- Teng-Ruei Chen. How much of the routing gap is real? decomposing the router-to-oracle gap into reproducible specialist advantage and single-draw label noise. *arXiv preprint arXiv:2607.03436v2*, 2026b. URL <https://arxiv.org/abs/2607.03436v2>.
- Wenting Chen, Guolin Huang, Wenxuan Wang, and Zhongrui Zhu. Medeinst: Benchmarking the einstellung effect in medical llms through counterfactual differential diagnosis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026c. doi: 10.18653/v1/2026.acl-long.1847.
- Yongchao Chen, Jiefeng Chen, Rui Meng, Ji Yin, Na Li, Chuchu Fan, Chi Wang, Tomas Pfister, and Jinsung Yoon. TUMIX: Multi-agent test-time scaling with tool-use mixture. In *International Conference on Learning Representations*, 2026d. URL <https://openreview.net/forum?id=HBm3MFtszH>.
- Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.
- Seffi Cohen, Nurit Cohen Inger, Niv Goldshlager, Bracha Shapira, and Lior Rokach. DFPE: A diverse fingerprint ensemble for enhancing LLM performance. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5326–5336, 2026. doi: 10.18653/v1/2026.findings-eacl.282.
- Kimmo Eriksson, Simon Karlsson, Irina Vartanova, and Pontus Strimling. Large language models outperform humans at estimating society’s everyday norms. *Communications AI & Computing*, 1:15, 2026. doi: 10.1038/s44488-026-00018-8.
- Google. Gemini 2.5 Flash-Lite: Model documentation, 2025. URL <https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite>. Accessed 15 September 2026.
- Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Accessed 15 September 2026.
- Jie Hu. Oracle gap and signal fidelity: A fixed-pool diagnostic for test-time collaboration. *arXiv preprint arXiv:2607.17531*, 2026.

379 Jia-Hao Ji, Sijie Li, Jiabei Cheng, Zixi She, Jin-Tai Yu, and Zhiyuan Yuan. Candidate supply
380 and answer selection shape the value of llm judging in multi-agent systems. *arXiv preprint*
381 *arXiv:2608.25937*, 2026.

382 Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov,
383 Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of
384 language models (and beyond). In *Advances in Neural Information Processing Systems, Datasets*
385 *and Benchmarks Track*, 2025.

386 Ammar Khairi, Daniel Dsouza, Marzieh Fadaee, and Julia Kreutzer. Making, not
387 taking, the best of n. In *International Conference on Learning Representations*,
388 2026. URL [https://proceedings.iclr.cc/paper_files/paper/2026/hash/](https://proceedings.iclr.cc/paper_files/paper/2026/hash/f111cfa78b2ea8b6705475d60d76ed10-Abstract-Conference.html)
389 [f111cfa78b2ea8b6705475d60d76ed10-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2026/hash/f111cfa78b2ea8b6705475d60d76ed10-Abstract-Conference.html).

390 Donghwan Kim. Are diversity metrics measuring diversity? a capability-controlled audit of
391 majority-vote gain in LLM ensembles, 2026. URL <https://arxiv.org/abs/2607.20768>.

392 Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language
393 models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267,
394 pages 30038–30066, 2025. URL <https://proceedings.mlr.press/v267/kim25e.html>.

395 Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan,
396 Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe
397 Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Capable
398 language models can outgrow the benefits of collaboration. *Nature Machine Intelligence*, 8:1157–
399 1172, 2026. doi: 10.1038/s42256-026-01268-y.

400 Chenchen Kuai, Jiwan Jiang, Zihao Zhu, Hao Wang, Keshu Wu, Zihao Li, Yunlong Zhang, Chenxi
401 Liu, Zhengzhong Tu, Zhiwen Fan, and Yang Zhou. A statistical framework for auditing behavioral
402 dependence and induced bias in llm judges. In *Conference on Language Modeling*, 2026. URL
403 <https://arxiv.org/abs/2604.07650>.

404 Amit Kumthekar, Zion Tilley, Henry Duong, Bhargav Patel, Michael Magnoli, Ahmed Omar,
405 Ahmed Nasser, Chaitanya Gharpure, and Yevgen Reztsov. Second opinion matters: Towards
406 adaptive clinical ai via the consensus of expert model ensemble. *arXiv preprint arXiv:2505.23075*,
407 2025.

408 Ruitong Li, Binjie Guo, Aisheng Mo, Guowei Su, Jie Li, and Ru Zhang. Agreement before diver-
409 sity: Verification-first complementarity for heterogeneous language-model coordination. *arXiv*
410 *preprint arXiv:2608.04618v1*, 2026a. doi: 10.48550/arXiv.2608.04618.

411 Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. Rethinking mixture-of-agents: Is mixing differ-
412 ent large language models beneficial? *Transactions on Machine Learning Research*, 2026b. URL
413 <https://openreview.net/forum?id=K6WwK8UR1V>.

414 Artem Maryanskyy, Dmitry Budnikov, and Alibek T. Kaliyev. When agents disagree: The selection
415 bottleneck in multi-agent LLM pipelines. *Applied Sciences*, 16(10):4914, 2026. doi: 10.3390/
416 app16104914.

417 Meta. Llama-3.1-8B-Instruct model card, 2024a. URL [https://huggingface.co/meta-llama/](https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct)
418 [Llama-3.1-8B-Instruct](https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct). Accessed 15 September 2026.

419 Meta. Llama-3.3-70B-Instruct model card, 2024b. URL [https://huggingface.co/](https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct)
420 [meta-llama/Llama-3.3-70B-Instruct](https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct). Accessed 15 September 2026.

421 Mistral AI. Ministral-3-14B-Instruct-2512 model card, 2025. URL [https://huggingface.co/](https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512)
422 [mistralai/Ministral-3-14B-Instruct-2512](https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512). Accessed 15 September 2026.

423 Qwen Team. Qwen2.5-14B-Instruct model card, 2024. URL [https://huggingface.co/Qwen/](https://huggingface.co/Qwen/Qwen2.5-14B-Instruct)
424 [Qwen2.5-14B-Instruct](https://huggingface.co/Qwen/Qwen2.5-14B-Instruct). Accessed 15 September 2026.

425 Yasushi Sakai, Allen Song, and Kent Larson. When does delegation beat majority? a delegation-
426 based aggregator for multi-sample LLM inference, 2026. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2606.08098)
427 [2606.08098](https://arxiv.org/abs/2606.08098).

428 Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius.
429 Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings*
430 *of the 41st International Conference on Machine Learning*, volume 235, pages 45883–45905.
431 PMLR, 2024. URL <https://proceedings.mlr.press/v235/smit24a.html>.

432 Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. DDXPlus: A
433 new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing*
434 *Systems, Datasets and Benchmarks Track*, 2022.

435 Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Jonathan Light, Weiyang Liu, Haifeng Chen, Xiang
436 Zhang, and Wei Cheng. On the effect of sampling diversity in scaling llm inference. In *Proceed-*
437 *ings of the Forty-Second Conference on Uncertainty in Artificial Intelligence*, volume 337, pages
438 7137–7167, 2026. URL <https://proceedings.mlr.press/v337/wang26f.html>.

439 Weiqin Wang, Yile Wang, and Hui Huang. Ranked voting based self-consistency of large language
440 models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14410–
441 14426, 2025. URL <https://aclanthology.org/2025.findings-acl.744/>.

442 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-
443 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.
444 In *International Conference on Learning Representations*, 2023. URL <https://arxiv.org/abs/2203.11171v4>.

446 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming
447 Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi
448 Fan, Xiang Yue, and Wenhui Chen. MMLU-Pro: A more robust and challenging multi-task
449 language understanding benchmark. In *Advances in Neural Information Processing Systems*,
450 volume 37, 2024. URL [https://papers.nips.cc/paper_files/paper/2024/hash/](https://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html)
451 [ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.](https://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html)
452 [html](https://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html). Datasets and Benchmarks Track.

453 Yuxuan Yao, Han Wu, Mingyang Liu, Sichun Luo, Xiongwei Han, Jie Liu, Zhijiang Guo, and Linqi
454 Song. Determine-then-ensemble: Necessity of top-k union for large language model ensembling.
455 In *International Conference on Learning Representations*, 2025. URL <https://arxiv.org/abs/2410.03777>.

457 Grace Chang Yuan, Xiaoman Zhang, Sung Eun Kim, and Pranav Rajpurkar. Do mixed-vendor
458 multi-agent llms improve clinical diagnosis? In *Proceedings of the 1st Workshop on Linguistic*
459 *Analysis for Health (HeaLing 2026)*, pages 1–18, 2026. doi: 10.18653/v1/2026.healing-1.1.

460 Kexun Zhang, Weiran Yao, Zuxin Liu, Yihao Feng, Zhiwei Liu, Rithesh Ramapura Narasimha
461 Murthy, Tian Lan, Lei Li, Renze Lou, Jiacheng Xu, Bo Pang, Yingbo Zhou, Shelby Heinecke,
462 Silvio Savarese, Huan Wang, and Caiming Xiong. Diversity empowers intelligence: Integrating
463 expertise of software engineering agents. In *International Conference on Learning Representa-*
464 *tions*, 2025. URL <https://openreview.net/forum?id=cKlzKs3Nnb>.

465 Nikolas Zöller, Julian Berger, Irving Lin, Nathan Fu, Jayanth Komarneni, Gioele Barabucci,
466 Kyle Laskowski, Victor Shia, Benjamin Harack, Eugene A. Chu, Vito Trianni, Ralf H. J. M.
467 Kurvers, and Stefan M. Herzog. Human–AI collectives most accurately diagnose clinical vi-
468 gnettes. *Proceedings of the National Academy of Sciences*, 122(24):e2426153122, 2025. doi:
469 10.1073/pnas.2426153122.

470 Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning
471 Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and under-
472 standing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267,
473 pages 80961–80990, 2025. URL <https://proceedings.mlr.press/v267/zuo25a.html>.

Reproducibility Statement

The accompanying anonymous replay archive reconstructs the original primary contrast, absolute rule accuracies, depth curves and negative predictor, the 720-case decision analysis, the 256-case Health analysis, the 128-case output-contract primary, exploratory joint policy selection and allocation–budget curves with one offline command. It includes stripped native ballots, scored MedEinst pools, frozen inputs and checksums, without question text or credentials. Eight checks compute exact votes, case-bootstrap intervals, paired quality/lift accounting and the conditional empirical-list reference. This reproduces inference on stored outputs; it does not regenerate hosted outputs or adjudicate clinical labels. Methods, Results and Appendices A–C.2 specify the boundaries.

AI Use Statement

Generative AI assisted literature search, hypothesis refinement, experimental design, code implementation, deterministic terminology-rule development, statistical analysis, interpretation, figures and manuscript drafting. During the benchmark experiments, models generated the answer candidates and selections described in Methods. Numerical checks are documented in Appendix E; they do not replace human review or clinical adjudication. The authors retain responsibility for the submitted work.

Guide to the supplementary evidence

The native-key protocols, alternative aggregation rules and information controls support the main claims. Secondary selector studies and scoring limitations are distinguished from independent confirmation.

Question or evidence	Where to look
Rule contrast, tie explanation, 288 native questions	Section A
Selected-policy loss, 720 native questions	Section A.2
Health gain versus stronger policy	Section A.3
One-answer versus five-answer generation	Section A.4
Policy, budget and partner sensitivities	Sections C.1, C.2, C.3
IRV, rank transfer and conditional diagnostics	Sections B.2–B.5
Membership ceilings, wrong-option controls and equal-budget screening	Section H
Notation and replay instructions	Section E.1
Clinical mapping and worst-case uncertainty	Sections F, A
Held-out predictor specification	Section D
Secondary selector interventions	Section D.1
Reproduction resources and serving limits	Section E

A Representation and native-answer confirmation protocols

A.1 Original 288-question representation study

The representation question was selected after inspecting the earlier banks. Its separate protocol was frozen on 9 September 2026 before any new generation. Let $A_r = F_{M,r} - F_{H,r}$ for TRUE support and define the sole primary $C = A_{\text{inclusion},5} - A_{\text{first}}$. Native-key MedXpertQA is primary; MedEinst is a separate prespecified transport. The study does not generate new LLM-selector decisions. It tests transformations of saved five-answer lists, not prompting for one versus five answers.

Case selection. We exclude 3,902 previously used MedEinst IDs and 128 native IDs, normalized input matches and duplicate eligible inputs. Native sampling takes 96 hash-ranked questions in each of Diagnosis, Treatment and Basic Science. From 1,481 eligible MedEinst pairs, we select 32 short, 32 medium and 32 long pairs by hash rank, requiring each pair to differ by one line. This replaces the initially planned two-block frequency/length-balanced sample, whose second block was infeasible; the change preceded all model calls. The final sample is length-balanced

but not transition-frequency-balanced: it contains 90 high-, five medium- and one low-frequency pairs. Neither historical IDs nor normalized inputs overlap with the analyzed sample.

Sampling and failures. Each case has two new banks, with independent request identities/seeds; provider determinism and independent internal randomness are not guaranteed. Per branch/bank, H takes eight Gemma calls and M shares its first two plus two calls to each other generator, giving 14 unique calls. There are $14 \times 2 \times (288 + 2 \times 96) = 13,440$ scheduled identities. Generators use the same five-answer contract, temperature 1, top- $p = .9$ and 384-token cap. Gemma, Llama and Ministral use OpenRouter; Qwen uses the official Qwen2.5-14B-Instruct checkpoint in BF16 through vLLM 0.10.2/transformers 4.55.2. An exact weight-revision hash was not returned. Every invalid slot receives at most one identical-payload retry; first valid output is retained and terminal failures remain empty. The protocol requires terminal failures to remain at or below 5% in each source/model cell. All eight cells satisfy this threshold; the highest rate is 3.73% (native Llama), with denominators in Table 17. Independent banks are averaged within case, and MedEinst branches within pair, before inference.

Rules and statistics. First-choice, inclusion, reciprocal rank, RR restricted to inclusion ties and truncated Borda are specified in the main text. Each call contributes once per normalized candidate at its best rank. Borda weights are 5, 4, 3, 2, 1, with zero for absent candidates. Unresolved ties use expected uniform correctness. Transition accounting couples ties through a common uniform random candidate priority; unchanged winner sets cannot manufacture disagreements. The tie component is the H-minus-M accuracy gain from replacing inclusion by tie-aware frequency; subtracting it from C leaves the tie-aware-versus-first allocation contrast. This is an accounting identity, not a mediation estimate.

The desired 95% interval half-width was five percentage points. Old case-level standard deviations .41535 native and .24058 MedEinst imply approximate requirements of 266 and 89 cases under $n = (1.96s/.05)^2$; fixed 288/96 preserve task/length balance. No reduction was assumed from averaging banks. The primary uses 50,000 task-stratified native case bootstraps, seed 202609091733; MedEinst resamples whole pairs. Depths 2–4, other rules, bank differences, flips and decomposition are secondary pointwise summaries. Uncertainty conditions on benchmark keys or the frozen deterministic refinement/strict/exact-reference mapping algorithm, not clinical adjudication.

One held-out prediction check. A ridge predictor of case-level C is fit only on the earlier native128. Its fixed rival inputs are M-minus-H candidate count divided by ten and rank-one vote concentration; an enriched model adds the H-minus-M inclusion-tie indicator. Inputs are standardized on training cases, ridge coefficient is one and the intercept is the training mean. This specification was frozen after new generation began but before fresh correctness was analyzed. It is a secondary held-out check, not pregeneration registration. The single comparison is paired fresh-native MSE reduction from adding the tie gap, with 50,000 task-stratified bootstrap resamples. A constant old-training-mean predictor is also retained. Prediction does not establish causal competence adjustment or an improved allocation policy.

Local reproduction. The exact protocol is Protocols/REPRESENTATION_CONFIRMATION_2026_09_09.md; schedules, raw attempts, source hashes, frozen dependencies and operational manifests are under runs/representation_confirmation_2026_09_09/. From the complete local project, run code/representation_confirmation_20260909.py with argument analyze, then code/verify_representation_confirmation_20260909.py; these consume stored outputs. The independent verifier reconstructs native options, rational-valued votes, all ten absolute rule/allocation cells, case/bank contrasts and the primary interval from raw output text. It does not adjudicate MedEinst labels. The saved-model prediction is evaluated with code/predict_representation_boundary_20260909.pyevaluate. Do not run setup, execute or freeze to reproduce the saved analysis. The earlier rule integration and current figures are under runs/maintrack_strengthening_2026_09_09/. Dataset question text and credentials are not part of the manuscript-source archive.

Operational sensitivity. The saved-attempt audit verifies all 13,440 scheduled identities, at most two attempts per identity, identical retry payloads, and retention of the first valid output. A post-outcome complete-case sensitivity excludes any case with a failed generation: native $n = 193$

gives $C = 8.86$ points (95% CI [3.32, 14.49]); MedEinst $n = 88$ gives 4.62 [.28, 9.15]. These exclude 95 native cases and eight pairs, change the sampled population and native task mixture, and do not replace the primary with empty failures retained. Retained native counts are 66 Diagnosis, 62 Treatment and 65 Basic Science. These sensitivity intervals use 50,000 unstratified whole-case resamples with seed 202609091733; the primary remains task-stratified. No additional calls were collected.

Mapping implementation and protocol discrepancy. The fresh transport map contains 3,607 candidate–reference relation types: 2,401 new and 1,206 reused. Of 2,846 strict-unresolved types, 2,388 are new. The frozen code also applies a separate refinement whitelist, crediting 24 strict-unresolved types (ten new, 14 reused). This conflicts with the protocol prose that newly unresolved relations remain uncredited. We preserve the original protocol, code and results and report the discrepancy: setting all new unresolved relations to zero gives $C = 4.93$ points (95% CI [.87, 9.17]), compared with the implemented 4.85 [.77, 9.08]. Setting all strict-unresolved relations to zero gives 6.10 [2.01, 10.38], identical here to strict/literal scoring. These fixed-rule checks leave the direction intact; they do not resolve arbitrary clinical relabeling.

The 3,607 types occur in 12,487 candidate–pool incidences; 6,760 incidences are unresolved. All 96 pairs contain unresolved candidates. The 24 refinement credits affect the case-level contrast in 19 pairs, and 171 unresolved types have nonzero overall contrast coefficients. A shared-relation binary stress bound is [−26.34, 29.85] points; this is a deterministic worst-case envelope, not a confidence interval or medically plausible adjudication. The native primary uses official option keys and is unaffected. The incidence and sensitivity audit is reproduced by `code/audit_latest_review_mapping_20260910.py` on the saved map and pools.

Resource closure. The fresh representation program attempted 13,440 scheduled identities with 13,845 raw attempts, including allowed retries. There are 13,292 valid slots and 148 terminal failures, retained as empty. Hosted charged-or-reserved cost is \$1.211903; conservative GPU lifecycle cost is \$0.104540. Total incremental cost is \$1.316443. These resource estimates include retries and serving overhead; they are not a token-matched policy comparison or an independently reconciled provider invoice.

A.2 Two-model, 720-question native-answer confirmation

The later confirmation uses 240 native-key MedXpertQA questions from each of Diagnosis, Treatment and Basic Science. Its excluded development and pilot sets contain 48 and 24 questions, respectively; historical and retired assignments are excluded by the frozen selection procedure. Each question receives two new banks of eight calls to each of `google/gemma-4-26b-a4b-it` and `meta-llama/llama-3.3-70b-instruct`. Within a bank, H_G and H_L use their respective eight calls; M_{44} reuses the first four from each. Thus $720 \times 2 \times 2 \times 8 = 23,040$ scheduled slots support all three policies without treating shared calls as independent observations.

The executed system message was:

```
You are performing a controlled medical-question benchmark. Use the
supplied question and options. Return the requested JSON.
```

The user message, with bracketed fields substituted, was:

```
MEDICAL_QUESTION:
[QUESTION]
OPTIONS_JSON=[OPTIONS_OBJECT]
Rank exactly five DISTINCT supplied option IDs from most to least plausible
answers to the question. For each give a rationale of at most18 words.
Return only JSON: {"candidates":[{"option_id":"A","rationale":"brief
reason"}, ... exactly five items]}. Do not add keys or markdown. Consider
the actual question, which may ask for a diagnosis, treatment, or basic
science answer.
```

The literal “at most18” wording is retained from the payload constructor. The response format requests a five-element candidate array with legal option IDs and rationales. Sampling uses temperature 1, top- $p = .9$, maximum 384 tokens, and a 31-bit seed derived from the SHA256 of each

request identity. At most one identical-payload retry is allowed; the first valid output is retained, and terminal failures contribute no votes. The manifest records 23,449 attempts, 22,929 valid slots and 111 terminal failures. No failed slot is silently replaced by a later successful policy sample.

The first-choice $M_{44} - H_G$ contrast is primary, with Gemma selected on development evidence rather than the confirmation set. Banks average within question before 50,000 task-stratified bootstrap draws (seed 202609103011). Its estimate is -8.95pp $[-11.16, -6.75]$; other rules and quality/lift decompositions do not replace that primary. Charged-or-reserved cost is \$3.008604. The schedule, raw attempts and frozen dependencies are retained under runs/competent_alternative_confirmation_restart_2026_09_10/; the second portable replay check reconstructs the decision analysis. These data establish neither equal generator quality nor independent training ancestry. Later Health routing restrictions are documented separately rather than assumed to match these historical providers.

Exploratory task-specific replay. Table 5 disaggregates the same fixed-slot H_G/M_{44} comparison after inspection of pooled results. Exact tie-aware scoring retains empty failures, averages two banks within question, and reproduces the pooled $\theta = -8.9525463$ and $C = 9.1770833$ points. Each task resamples its 240 questions 50,000 times, keeping both allocations and rules paired. The family comprises six task-endpoint contrasts; Bonferroni percentile tails are .05/12 and $1 - .05/12$. These nominal bootstrap intervals are not finite-sample guarantees. The task-specific seeds and input hashes are in the accompanying task-strata replay archive. The results are descriptive reuse, not new confirmation or a test of between-task differences.

Table 5: Native720 task breakdown against the unchanged development-selected Gemma policy. Effects and intervals are percentage points; every row has 240 questions. C remains a rule interaction, not final-policy benefit.

Task	Contrast	Estimate	Pointwise 95%	Adjusted six
Diagnosis	θ	-9.34	$[-13.09, -5.66]$	$[-14.40, -4.41]$
	C	9.26	$[5.50, 13.16]$	$[4.16, 14.49]$
Treatment	θ	-11.13	$[-15.50, -6.88]$	$[-16.98, -5.45]$
	C	12.00	$[7.88, 16.26]$	$[6.49, 17.68]$
Basic Science	θ	-6.39	$[-9.77, -3.00]$	$[-10.93, -1.82]$
	C	6.27	$[2.76, 9.76]$	$[1.51, 10.99]$

634 A.3 Prespecified MMLU-Pro Health source transport

The frozen test uses dataset revision b189ec765aa7ed75c8acfea42df31fdae71f97be. Of 818 Health rows, 221 fail the exactly-ten-distinct-options rule and 91 repeat an eligible normalized question stem, leaving 506. No normalized question stem matches the local MedXpertQA source. A fixed SHA256 ordering selects 256 before generation; this restricted population is not the full Health benchmark. Exact-stem filtering cannot exclude paraphrase or pretraining contamination. The dataset card reports MIT; scored rank pools and retrieval metadata, not question text, are included in the replay.

The 256 items comprise professional medicine 71, nutrition 56, human aging 30, anatomy 27, clinical knowledge 27, college medicine 18, medical genetics 15 and virology 12. These are source subgroups, not independent datasets. Native keys provide a separate benchmark criterion, not clinical adjudication. The generation template, five-answer parser, models and allocation definitions match the 720-case study. Temperature 1, top- $p = .9$, maximum 384 tokens and hashed seeds remain fixed; only DeepInfra/Novita are permitted for costs. All 8147 returned payloads with a named provider report DeepInfra; 138 attempts lack a returned payload/provider. Historical serving revisions and sampling independence are not guaranteed.

All 8192 scheduled slots were attempted, with one allowed identical retry after failure: 8285 attempts, 8137 valid lists and 55 terminal empty slots (Gemma 1, Llama 54). Ten returned lists had duplicate IDs; 138 attempts lacked usable responses. No incomplete list is promoted to a valid five-answer ballot. The primary includes failures; complete-case exclusion changes the population to 207 questions. No scientific sign or model accuracy determined retry, stopping or case selection.

655 The sole primary is $C_G = 6.14\text{pp}$ [2.84, 9.45], with a 3.30pp half-width, meeting the five-point
656 precision target but not establishing an interaction above five points. Both banks are positive
657 (6.72/5.56pp). Secondary $C_L = 5.62\text{pp}$ [2.66, 8.58] is positive against the other homogeneous
658 reference. First-choice mixed-minus- H_G is -1.14pp [-4.20, 1.95]; against H_L it is $+1.76\text{pp}$
659 [-.98, 4.49]. These pointwise secondary intervals establish neither equivalence nor useful final im-
660 provement. The development-selected Gemma reference was not reselected using Health outcomes.

Table 6: Health transport: two banks averaged per question. Accuracy and availability are percent-
ages; costs are mean USD per eight-call policy execution, including retries. Shared calls are charged
once per policy, not summed as independent study spending.

Rule or resource	H_G	H_L	M_{44}
Inclusion frequency	23.48	21.10	28.48
First choice	68.82	65.92	67.68
Frequency + RR ties	69.04	66.02	68.07
Reciprocal rank	69.04	66.02	68.26
Borda	69.53	66.08	68.55
Per-call first choice	68.73	64.70	66.58
Correct-key availability	96.29	95.31	97.85
Cost (USD)	.000582	.000623	.000610

661 The independent exact-rational replay reconstructs 7680 pool/rule outcomes and 3840 bank-
662 averaged case/rule cells, then reproduces the primary whole-question 50,000-draw bootstrap to
663 floating-point precision. Complete-case $C_G = 5.67\text{pp}$ [2.01, 9.32]; leave-one-question estimates
664 range 5.79–6.56pp. These are sensitivities, not replacements for failure-inclusive inference. De-
665 scriptive subgroup C estimates in the source order above are 3.53, 9.40, 11.92, 1.42, 8.52, 5.32,
666 11.33 and -8.13pp ; small subgroup sizes and exploratory multiplicity preclude a subgroup mecha-
667 nism claim.

668 Collection took 1432 seconds and cost \$0.616917 in charged-or-reserved accounting; no GPU was
669 used. The replay metadata records source, data and protocol hashes; raw attempts retain payloads,
670 usage, finish reasons and parser states. Replaying these records reproduces the stored-output analy-
671 sis, not clinical correctness or generation under a future provider version.

672 A.4 Direct-answer versus ranked-list generation

673 This later, separately frozen experiment tests a condition not manipulated by saved-list first-choice
674 replay. A conservative audit excludes configured or scheduled historical IDs, retired assignments and
675 normalized native stems. The remaining eligible pools contain 570 native and 250 Health questions.
676 This does not exclude unrecorded human inspection, paraphrases or model pretraining. A hash
677 ordering assigns 16 questions per source to an excluded feasibility/variance pilot; the next 64 per
678 source form confirmation. Pilot variance selects the smallest of 64/128/192 per source forecast to
679 achieve pooled three-point interval half-width, subject to available cases and a fixed spending cap.
680 The pilot’s forecast is optimistic: actual confirmation half-width is 3.61 points. We do not enlarge
681 the sample after seeing the result.

682 K1 requests exactly one supplied answer; K5 requests five distinct ranked answers. Both request
683 at most 18 rationale words per answer, use temperature $1/\text{top-}p = .9$, cap output at 384 tokens,
684 and preserve source option order. Requested rationale count therefore changes with the contract.
685 For each question, two banks and eight calls to each of Gemma4-26B-A4B and Llama3.3-70B
686 are shared by H_G , H_L and M_{44} . Seeds are paired across contracts; dispatch is hash-randomized.
687 Providers are restricted to DeepInfra/Novita; successful responses came from DeepInfra. This does
688 not guarantee serving determinism or isolate reasoning dose.

689 Assigned-length JSON with distinct legal IDs and nonempty rationales is required; one same-
690 payload retry is allowed. Failed slots would contribute no vote and zero per-call correctness. The
691 pilot completes all 2,048 slots without retry. Confirmation produces 8,192 valid slots after 8,698
692 requests: 503 rate-limit errors, two length-terminated invalid responses and one duplicate-option re-
693 sponse are retried. There are no terminal failures; 43 accepted responses exceed the rationale-word
694 limit and remain included as specified. Raw payloads, responses, usage, finish reasons and parser
695 states are retained locally.

Table 7: Output-contract confirmation: first-choice final accuracy (percent), averaging two banks on 64 new questions per source. Allocation contrasts and source interactions have secondary pointwise 95% whole-question intervals. Only pooled R is primary.

Source	Contract	H_G	H_L	M_{44}	θ [95% CI] (pp)
MedXpertQA	K1	38.28	16.02	25.78	-12.50 [-19.53, -5.47]
MedXpertQA	K5	34.77	17.58	26.56	-8.20 [-16.41, .39]
Health	K1	61.72	50.78	57.81	-3.91 [-10.16, 2.34]
Health	K5	64.45	53.12	61.72	-2.73 [-8.59, 3.12]
Native $R = \theta_{K5} - \theta_{K1}$ (pp)					4.30 [-1.56, 10.55]
Health R (pp)					1.17 [-2.73, 5.08]
Equal-source pooled R (primary, pp)					2.73 [-.78, 6.45]

Both banks average within question before 50,000 whole-question bootstrap draws, stratified by source for the pooled primary. The unresolved interaction is not equivalence or proof of unchanged allocation. The secondary native K1 deficit supports the limited boundary that forced five-answer generation is not necessary for a mixed deficit in this pair. Source intervals are too wide for strong per-source moderation conclusions. All original C , θ , J and J_{norm} tests remain separate.

Across paired individual calls, K1-wrong/K5-correct versus K1-correct/K5-wrong counts are 35/83 for native Gemma, 19/3 for native Llama, 19/6 for Health Gemma and 56/28 for Health Llama (1,024 paired slots per source/model). These descriptive counts are not independent case-level replications or evidence that model quality was controlled. Confirmation K1/K5 attempts consume 1,743,854/1,759,592 input tokens and 120,192/480,858 output tokens. Missing-usage rate-limit attempts contribute conservative reserved cost, not fabricated token counts. Charged-or-reserved costs are \$0.381499/\$0.500202; pilot cost is \$0.122994. Confirmation takes 736 seconds, without GPU. Call matching is not token/cost matching.

The sixth text-free offline check reconstructs all 8,192 ballot slots, exact rational voting and the frozen primary bootstrap. A separate raw-response audit also verifies pre-output hashes, pilot exclusion, paired payload seeds and parser decisions. Replaying these artifacts does not regenerate hosted responses or validate clinical labels.

B Rank, candidate access and confidence diagnostics

B.1 Exploratory rank-information intervention

This post-review analysis asks whether the allocation gap depends on expressed candidate preferences, beyond candidate membership. It reuses every Native720 and Health256 question and both banks. For each of the 16 unique call slots per bank, select one recorded candidate uniformly as the new first answer. For first-choice voting this is equivalent to a uniform random permutation of the list. Shared calls use the same draw in both allocations; failed calls remain empty. The original fixed-slot H_G/M_{44} policies, candidate sets and call budget are unchanged. Inclusion-vote accuracy is therefore invariant. Random draws do not consult answer keys.

We use two independently seeded runs of 8,192 draws per bank and question, scoring final vote ties by their exact uniform expectation. After averaging draws and banks within question, 50,000 paired question-bootstrap draws produce pointwise and Bonferroni intervals across four endpoints per source (eight total). Native resampling preserves the three task strata. The protocol, seeds and endpoint family were fixed before this new calculation, but after the original results were known: this is exploratory intervention evidence, not independent confirmation. Maximum Monte Carlo standard error is .014 percentage points; reported bootstrap intervals condition on simulated case means rather than including that numerical error. Both seeds yield the same qualitative interpretation.

Native preferences benefit repeated Gemma sampling more than mixing; destroying them sharply attenuates its advantage. Health demonstrates a boundary: both policies lose substantial accuracy, but their differential loss is unresolved. Its randomized mixed advantage does not establish a reliable reversal from the original gap; the paired change interval includes zero. Rank destruction changes per-call accuracy and is not a deployable improvement, a quality-matched diversity comparison, or

Table 8: Rank randomization on fixed ballots. Accuracy and differences are percentage points; intervals adjust across all eight endpoints. “Rank benefit” is original minus randomized first-choice accuracy. A positive gap change means randomization makes mixing relatively more favorable, not that it improves either policy.

Quantity	Native720	Health256
Original H / M accuracy	35.21 / 26.26	68.82 / 67.68
Randomized H / M accuracy	16.71 / 16.19	19.97 / 21.52
H rank benefit	18.50 [14.21, 22.82]	48.85 [41.41, 56.12]
M rank benefit	10.07 [6.50, 13.75]	46.16 [39.36, 52.76]
Randomized M–H gap	−.52 [−1.15, .08]	1.55 [.89, 2.16]
Change in M–H gap	8.43 [5.49, 11.43]	2.69 [−1.62, 6.97]

735 evidence about training mechanisms. It identifies the effect of this specified intervention on the
736 recorded lists only.

737 The anonymous rank-information replay contains stripped candidate IDs, native keys, task strata,
738 fixed slots, the two-seed protocol, all case vectors and expected results. Checks reconcile original
739 first-choice and inclusion accuracies, preserve empty slots and shared calls, reconstruct Health lists
740 from raw records against the original pools, and test ties and small exact enumerations. No hosted
741 generations, paid calls or new cases are added.

742 For comparison, the separate empirical-list reference draws whole lists independently for each ques-
743 tion and model from its stored banks, including failures, and reconstructs allocations 4,096 times.
744 It gives $C = 9.48$ versus observed 9.72 on Native288 and 8.73 versus 9.18 on Native720; the
745 observed-minus-resampled differences are .24 [−1.20, 1.71] and .45 [−.32, 1.21] points, respec-
746 tively. Both intervals include zero. Unlike rank destruction, it preserves each list’s internal ordering,
747 but not its original grouping with other calls. Its question-bootstrap intervals condition on observed
748 list distributions and omit uncertainty in those distributions. Agreement with this fitted reference
749 neither proves independent generation nor identifies a causal mechanism.

750 B.2 Additional fixed-ballot controls

751 These exploratory analyses were specified after the original outcomes and review were read, before
752 their calculations. All 720 native and 256 Health questions, both banks, shared calls and empty
753 failures remain. Fifty thousand whole-question bootstrap draws use seed 20260919 plus the source
754 sample size; native draws preserve the three task sizes. We average banks before resampling. New
755 analyses never replace a registered primary. Unless adjusted explicitly, intervals are pointwise 95%;
756 Bonferroni percentile intervals have nominal, not finite-sample guaranteed, familywise coverage.

757 **Instant-runoff voting.** The candidate universe is the union of nonempty lists. Count each bal-
758 lot’s first remaining candidate; stop at a strict majority of nonexhausted ballots or one remaining
759 candidate. Otherwise eliminate one of the tied least-supported candidates uniformly and recurse.
760 We average all elimination branches exactly with rational arithmetic. Exhausted ballots abstain; an
761 initially empty election scores zero. This neutral variant adapts ranked voting [Wang et al., 2025],
762 without asserting identity with its implementation or comparing tie percentages across datasets. Ab-
763 solute accuracies have pointwise intervals. The three contrasts per source (mixed minus H, H IRV
764 minus first-choice, M IRV minus first-choice) form a six-endpoint family.

765 IRV preserves the negative native allocation difference. None of its changes from first-choice voting
766 has a strictly positive adjusted lower bound. Thus the existing first-choice result does not depend on
767 omitting this elimination rule.

768 B.3 Preference transfer with a self-transfer control

769 For donor pool D , candidate k receives score $w_D(k) = \sum_{\ell \in D: k \in \ell} 1/r_\ell(k)$. In each recipient list,
770 identify positions whose candidates occur anywhere in D and sort only those candidates by de-
771 scending w_D ; SHA256 of the option identifier breaks exact score ties. Unmatched candidates retain
772 their positions; every list keeps its membership, length and empty-slot status. The operator sees no
773 gold label. Apply it once with the recipient itself as donor and once with the other allocation as

Table 9: IRV accuracy and contrasts, in percentage points. Absolute H/M rows use pointwise intervals; difference rows use six-endpoint-adjusted intervals.

Source	Quantity	Estimate [interval]
MedXpertQA	H_G	35.11 [31.74, 38.50]
MedXpertQA	M_{44}	26.19 [23.44, 28.98]
MedXpertQA	$M_{44} - H_G$	-8.92 [-11.87, -5.99]
MedXpertQA	H_G IRV-first	-0.10 [-0.64, 0.42]
MedXpertQA	M_{44} IRV-first	-0.07 [-0.90, 0.74]
Health	H_G	68.85 [63.18, 74.32]
Health	M_{44}	68.21 [63.09, 73.24]
Health	$M_{44} - H_G$	-0.63 [-4.59, 3.27]
Health	H_G IRV-first	0.03 [-0.26, 0.39]
Health	M_{44} IRV-first	0.54 [0.00, 1.32]

donor, then score first-choice voting. This self-transfer control separates donor choice from the effect of pooling ranks at all. Table 4 reports all four cross-minus-self contrasts with one four-endpoint family.

Gemma and mixed donors share four calls. Cross-transfer is consequently not an independent-donor experiment, and using the donor’s full eight-call pool requires information outside the recipient’s original pool. It is a diagnostic intervention, not an eight-call deployable method or a unique decomposition of candidate quality and lineage. Native mixed accuracy rises from 25.07% under its own pooled ranks to 33.61% under Gemma’s, still below original Gemma first-choice at 35.21%. Health provides an unresolved boundary rather than evidence that the effect is equal across sources.

B.4 Availability, selection efficiency and limits of prediction

Let v be the probability a pool contains the correct key and s the probability of a correct final selection conditional on its availability. A pool-restricted selector satisfies $F = vs \leq v = 1 - \beta$. This standard oracle ceiling [Chen, 2026a] does not constrain a synthesizer allowed to create new answers. To beat reference accuracy a , the selector needs $s > a/v$; this is impossible if $a > v$. We use each source’s original H first-choice accuracy for a . The availability denominator includes all recorded shortlist candidates, not only each call’s first answer. Selection efficiency is a ratio of pooled expected successes to pooled availability, averaging banks within question. Whole-question bootstrap resampling recomputes numerator, denominator and reference together. Fractional bank averages are not independent Bernoulli trials, so a pooled Clopper–Pearson interval would be inappropriate.

Table 10: Pool co-failure, conditional selection success and success required to beat original Gemma first-choice, all in percent with pointwise 95% paired question-bootstrap intervals. A lower co-failure rate need not yield better selection.

Source	Policy	β	s	a/v
MedXpertQA	H_G	10.35 [8.33, 12.50]	39.28 [35.61, 42.91]	39.28 [35.61, 42.91]
MedXpertQA	M_{44}	7.85 [6.11, 9.72]	28.50 [25.49, 31.56]	38.21 [34.61, 41.78]
Health	H_G	3.71 [1.56, 6.05]	71.47 [65.85, 76.91]	71.47 [65.85, 76.91]
Health	M_{44}	2.15 [0.59, 3.91]	69.16 [64.02, 74.20]	70.33 [64.70, 75.82]

The mixed pool’s larger availability is not the limiting ceiling on either source: its realized selection efficiency is below the threshold. This instantiates the availability–selection distinction in Maryansky et al. [2026], without assuming equal homogeneous performance across populations or substituting the separate LLM-selector cohorts. Satisfying the bound on the same scored ballots is algebra, not an out-of-sample prediction, and these thresholds do not forecast a new selector’s accuracy.

Why concentration alone does not identify the rule contrast. Even the full first-vote histogram leaves membership voting undetermined. Let g be correct and repeat each of two length-two lists

four times. Pools $([g, a], [b, a])$ and $([g, b], [b, g])$ have the same first-vote histogram (four g , four b), per-call accuracy $1/2$, list length two, budget eight and first-vote concentration $1/2$. Inclusion accuracy is respectively zero and $1/2$, although first-choice accuracy is $1/2$ in both. Moreover, H consisting of eight $[b, a]$ lists is more first-vote concentrated than the first mixed pool above, yet its interaction C is $-1/2$, not positive. Lower ranks and their joint distribution matter. Exact rescue/damage accounting [Ali, 2026] and delegation flip conditions [Sakai et al., 2026] address specified rules; they do not justify a universal sign theorem from these summaries or turn a same-ballot identity into prediction.

B.5 Conditional patterns and a marginal-preserving reference

Concentration is the largest H first-choice count divided by eight scheduled calls, averaged over banks. The bins below were fixed before this calculation but after reading the original results. Inclusion-tie presence means a tie in either bank. Resampling retains whole questions and recomputes each stratum mean; replicates containing no question in a sparse stratum are omitted for that stratum and their counts are saved. These output-defined, overlapping descriptions are not pre-specified moderators or a prospective rule. In particular, almost all questions have an inclusion tie, limiting its discrimination.

Table 11: Output-defined strata; C and pointwise intervals in percentage points. Groups overlap across concentration and tie classifications; small strata do not establish moderation.

Source	Stratum	Questions	C [95%]
MedXpertQA	$c \leq .5$	48	2.07 [-7.66, 11.20]
MedXpertQA	$.5 < c \leq .75$	171	11.67 [5.89, 17.57]
MedXpertQA	$c > .75$	501	9.01 [6.70, 11.32]
MedXpertQA	H inclusion tie	714	8.97 [6.78, 11.19]
MedXpertQA	H no inclusion tie	6	33.33 [-8.33, 75.00]
Health	$c \leq .5$	5	-5.00 [-50.69, 46.67]
Health	$.5 < c \leq .75$	17	6.42 [-8.08, 23.56]
Health	$c > .75$	234	6.36 [3.02, 9.66]
Health	H inclusion tie	256	6.14 [2.82, 9.45]

817

A marginal-preserving reference. We normalize option IDs by a cyclic shift making the correct key zero, then independently permute complete lists across questions within task, bank and scheduled slot. Shared H/M slots retain the same shuffled list. This preserves each slot’s empirical correct-rank, list-length, failure and normalized wrong-rank distribution exactly, while destroying cross-call dependence and question-difficulty alignment. We use 512 permutations, seed 2026091901 plus sample size. Because incorrect options have different meanings across questions, this is a synthetic dependence/difficulty reference, not a semantics-preserving intervention or a clinical null. Its ranges are permutation reference ranges, not confidence intervals for population effects.

Table 12: Observed values and slot-marginal-preserving reference (percent or percentage points for C). Null ranges describe 512 shuffled corpora.

Source	Quantity	Observed	Null mean	Null 95% range
Health	H availability	96.29	100.00	[100.00, 100.00]
Health	M availability	97.85	100.00	[100.00, 100.00]
Health	H inclusion ties	99.02	12.24	[9.77, 14.69]
Health	M inclusion ties	95.70	11.94	[9.57, 14.45]
Health	C	6.14	0.58	[-1.79, 2.93]
MedXpertQA	H availability	89.65	100.00	[100.00, 100.00]
MedXpertQA	M availability	92.15	100.00	[99.93, 100.00]
MedXpertQA	H inclusion ties	95.21	31.71	[29.43, 33.90]
MedXpertQA	M inclusion ties	89.17	34.19	[31.81, 36.46]
MedXpertQA	C	9.18	7.94	[5.75, 10.45]

825

The near-complete null availability and sharply reduced ties show that these observed patterns are not forced by slot marginals alone. However, C remains positive on average under the native null.

Neither its positivity nor the raw positive/zero/negative counts alone identify a mechanism. The archive reports all three sign counts under the null. Original Native288 findings remain separate; this later two-model diagnostic does not retrospectively redefine them.

Scope of screening evidence. An earlier unequal-budget analysis used 16 calls for vote-margin confidence and 24 for cross-policy agreement. Margin reduced mean retained error on both sources; agreement improved Health but was unresolved on native questions. These results cannot establish an equal-budget advantage. The matched-budget comparison in Section H.5 retains the same four-call predictions and compares four additional Gemma versus Llama calls. The earlier curves and procedure remain in the archival version and ballot-diagnostics replay.

C Policy, budget and partner sensitivity

C.1 Exploratory joint allocation-and-rule selection

This saved-output analysis was specified after the original test outcomes were known. It is exploratory, not a new preregistered confirmation. The original 48 development questions supply eight scheduled calls per model; the mixture uses the first four scheduled slots from each. All five rules are applied to each of three allocations, with uniform expected correctness for unresolved vote ties. The policy with highest mean development accuracy is selected, breaking policy ties lexicographically by allocation then rule. Development, pilot and test IDs are disjoint. Development uses one bank; each test averages its two banks within question. No test outcome selects the policy, although the decision to perform this expanded analysis was retrospective.

The expanded 15-policy search selects the same H_G first-choice policy as the original ten-policy homogeneous search. Its development accuracy is 29.17%; per-call development accuracy is a different quantity. Selection frequencies under 10,000 task-stratified development bootstraps are 75.92% for H_G first-choice, 8.32% for H_L Borda and 8.25% for mixed first-choice; the other policies share the remainder. Tied policies are credited to the fixed lexicographic winner, so these frequencies are descriptive stability estimates, not intrinsic probabilities that one voting rule is best.

Table 13: All 15 policies before test-based selection. Development, native720 and Health256 final accuracy (%). The fixed development winner is H_G first-choice. Test-best alternatives are hindsight descriptions, not deployable selectors.

Allocation	Rule	Development48	Native720	Health256
H_G	Borda	23.96	35.13	69.53
H_G	First choice	29.17	35.21	68.82
H_G	Inclusion	17.01	19.49	23.48
H_G	RR	27.08	34.65	69.04
H_G	Frequency + RR ties	26.04	34.34	69.04
H_L	Borda	22.92	19.76	66.08
H_L	First choice	22.92	19.97	65.92
H_L	Inclusion	17.36	15.38	21.10
H_L	RR	22.92	19.97	66.02
H_L	Frequency + RR ties	22.92	20.03	66.02
M_{44}	Borda	18.75	25.06	68.55
M_{44}	First choice	25.00	26.26	67.68
M_{44}	Inclusion	16.15	19.71	28.48
M_{44}	RR	23.96	24.86	68.26
M_{44}	Frequency + RR ties	16.67	25.83	68.07

For each source, 50,000 whole-question paired bootstrap draws retain all 15 policy outcomes together, stratifying by native task (Health has one stratum). Pointwise 95% intervals and Bonferroni-adjusted percentile intervals with tail mass $0.05/(2 \times 28)$ are reported in the replay. The family contains 14 nonreference policies on each of two sources. These are nominal bootstrap simultaneous intervals, not finite-sample guarantees. No adjusted interval has a positive lower endpoint. The H_G Borda advantage on Health is 0.72pp, pointwise interval $[0, 1.56]$ and adjusted interval $[-.29, 2.15]$; all five mixed alternatives are below the selected reference on native720 with adjusted

Table 14: Balanced-mixture C across call budgets (percentage points), from the exploratory finite-bank replay in Figure 4. Adjusted intervals cover the six source–budget contrasts using whole-question resampling. All questions were previously inspected; no fixed-slot primary is replaced.

Source	Calls	C	Pointwise 95%	Adjusted six
MedXpertQA	2	7.59	[6.05,9.12]	[5.50,9.64]
	4	8.23	[6.46,10.02]	[5.84,10.59]
	8	8.76	[6.62,10.92]	[5.85,11.63]
Health	2	5.76	[3.11,8.43]	[2.16,9.36]
	4	5.89	[2.99,8.80]	[1.96,9.78]
	8	6.10	[2.93,9.28]	[1.85,10.36]

intervals excluding zero. Absence of a resolved improvement does not prove equality or general optimality.

The test-best constant-policy point gaps are zero on native720 and 0.72pp on Health. A separate hindsight oracle choosing the best of 15 policies independently on each test question gives gaps of 17.32pp [15.43, 19.23] and 13.85pp [10.47, 17.43]. These quantities use the answers and noisy realized outputs; they are neither implementable routing performance nor evidence that a learned router can recover the gap. No new routing claim is made. The seventh offline check independently reconstructs development votes, selection and the paired intervals from stripped ballots and case matrices; the preceding checks reconstruct those test matrices from ballots. Original frozen results and bootstrap seeds remain unchanged.

C.2 Generation-budget and allocation sensitivity

This exploratory analysis reuses the 720 native and 256 Health questions, without additional generation. At total call budgets $B = 2, 4, 8$, Gemma call counts are respectively $\{0, 1, 2\}$, $\{0, 1, 2, 3, 4\}$ and $\{0, 2, 4, 6, 8\}$; remaining calls use Llama. Every subset without replacement from each model’s eight scheduled slots is enumerated within bank. Each subset uses the same five ranked-answer rules, uniform final ties and empty terminal failures. We average subsets exactly, then average the two banks within question. These are finite-bank policy expectations, not independent observations for each subset, estimates at unobserved budgets, or new held-out confirmation. At eight calls the mixed policy averages all eligible four-plus-four subsets, unlike the original first-four-slot policy. Both homogeneous endpoints reproduce their original accuracies.

Twenty thousand whole-question bootstrap draws retain all grid outcomes together; native resampling is task-stratified. The six balanced-mixture interactions receive Bonferroni-adjusted percentile intervals. All other curves and differences are exploratory, with pointwise intervals in the replay. As with the original percentile intervals, adjustment is nominal, not a finite-sample coverage guarantee. Calls are matched within budget; tokens and costs are not. Expected resource totals weight each scheduled slot’s recorded resources, including retries, by its subset inclusion probability. Unknown-usage failed requests remain outside token totals; charged/reserved costs retain them.

Table 14 reports the six balanced-mixture contrasts and both interval families.

The allocation-by-rule interaction persists across these budgets; this is not evidence that its size is invariant. The selected Gemma8 first-choice policy remains the native grid’s highest point accuracy. Health’s best grid point is six Gemma plus two Llama calls with Borda (70.17%), 1.36 points above Gemma8 first-choice, but its unadjusted interval $[-.09, 2.99]$ includes zero; comparison with Gemma8 Borda is .64 $[-.63, 2.13]$. These hindsight choices do not validate an improved allocation rule. The eighth offline check reconstructs all 130 source–configuration–rule means and their pointwise intervals, plus the six adjusted interaction intervals, from stripped ballots. Separate exact-Fraction checks and reconstruction of 14,640 archived policy–rule means audit the vectorized scorer.

Figure 4 displays these finite-bank curves; Table 14 supplies the balanced-mixture contrasts and multiplicity-adjusted intervals.

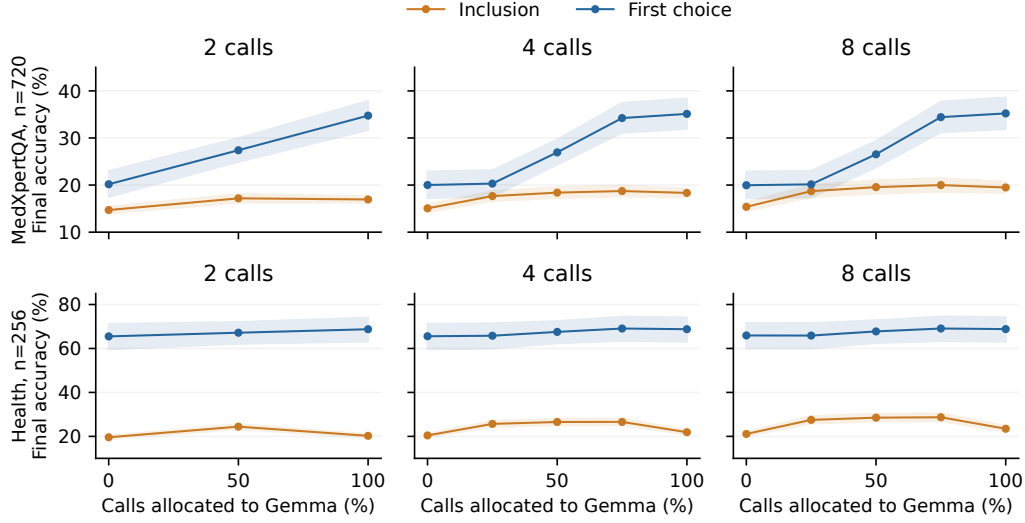


Figure 4: Exploratory allocation–budget sensitivity: 720 native and 256 Health questions. Points average stored-call subsets and banks, including at eight calls; they are not the primary fixed-slot estimates. Lines connect feasible allocations, not continuous policies. Shading: pointwise 95% question-bootstrap intervals, not adjusted C intervals. Call budgets are matched. Five rules are evaluated; inclusion and first-choice curves are shown.

C.3 Partner-composition sensitivity on recorded ballots

This exploratory replay reuses all 288 Native288 questions and their two banks, rather than generating new responses. Each bank contains eight Gemma calls and two calls each from Llama8B, Ministral14B and Qwen14B. At a four-call budget, H4 averages all $\binom{8}{4} = 70$ Gemma subsets. For each partner, M2+2 averages all $\binom{8}{2} = 28$ Gemma subsets combined with that partner’s two stored calls. All five rules and uniform final ties are retained; failed calls remain empty. Subsets are averaged within bank, then banks within question. This estimates a finite recorded-list policy, not performance over independently resampled model outputs.

The original eight-call primary point estimate and every original absolute rule accuracy are reconstructed before the new calculation. An independently written exact-Fraction scorer checks 1,732 mixtures and edge cases against the integer implementation. Fifty thousand whole-question bootstrap draws, stratified by the three native tasks, preserve dependence across partners, rules, subsets and banks. The family consists of C , the inclusion allocation difference and the first-choice allocation difference for each of three partners: nine contrasts, with Bonferroni percentile tail probability .05/18. Intervals are nominal bootstrap intervals, not finite-sample coverage guarantees. All comparisons are exploratory; none replaces the original primary or selects a new policy.

Table 15: Four-call partner replay on Native288. Accuracy is percent; differences and adjusted intervals are percentage points. H4 inclusion/first-choice accuracies are 17.55/29.86. All intervals adjust over the nine contrasts.

Partner	M inclusion	M first	C [adjusted 95%]	First M–H [adjusted 95%]
Llama8B	16.08	24.30	4.08 [.21, 8.07]	−5.56 [−8.99, −2.10]
Ministral14B	18.39	23.26	7.43 [3.09, 11.87]	−6.59 [−10.78, −2.36]
Qwen14B	16.40	22.84	5.86 [1.62, 10.17]	−7.01 [−11.34, −2.74]

The inclusion allocation differences are $−1.48$ [−4.18, 1.27], $.84$ [−1.18, 2.86] and $−1.15$ [−3.36, 1.02] points in the same order. Positive C therefore remains distinct from a demonstrated mixed-policy gain. The three comparisons share cases and a Gemma reference, differ in partner quality, and change the original eight-call allocation. They broaden the descriptive composition check without establishing independent replication, a model-size law or

quality-independent lineage causality. The separate composition replay includes the stripped ballots, task/answer keys, case vectors, all five-rule accuracies, bank means and bootstrap seed 202609160044.

D Prediction and secondary selector comparisons

The held-out predictor tests whether observed tie incidence predicts question-level rule effects. Define c_i by inserting question i 's bank-averaged scores into Eq. 1; C is the mean of c_i . Ridge regressions train on the earlier 128 native cases and predict c_i on Native288. The two baseline features are the M-minus-H differences in distinct candidate count divided by ten and in maximum first-choice count divided by eight. The extension adds H-minus-M inclusion-tie indicators. Features average banks, are training-standardized and use unit ridge penalty. Mean squared error (MSE) uses squared-proportion units. Held-out MSE reduction (baseline minus extended) is effectively zero: 0.0000003 $[-.00290, .00296]$ (95% interval; Appendix A); positive reduction favors the tie feature. Both models were frozen after generation began but before fresh accuracy analysis. Explaining vote arithmetic has not yielded a useful new-case predictor.

For the separate Native720 quality/lift decomposition, the paired components have bootstrap covariance 0.126 points², and bank-level lift differences are $-1.58 / -1.71$ points. These descriptive quantities do not establish quality-independent causality.

Under shared uniform tie-breaking priority, Native720 H_G first-choice has expected correction and reversal rates of 24.37% and 8.64% relative to inclusion, averaging banks. Question-level C is positive/zero/negative on 305/285/130 native questions and 143/67/46 Health questions. Inclusion obscures a correct unanimous homogeneous first choice in at least one bank on 184/720 and 166/256 questions, respectively. These are retrospective output-defined counts, not prospective predictors.

D.1 Secondary selector interventions and estimands

Two secondary tests examine LLM selectors: shuffle support numbers within each allocation's pool, or reverse support ranks in a common pool. These are different interventions, not replications.

The *support-number test* uses each allocation's own pool. Selectors read the case, rationales, inclusion counts and reciprocal-rank sums. TRUE supplies actual support; SHUFFLED reassigns support pairs, keeping candidates fixed within allocation. For example, (6, 4) for A and (2, 1) for B becomes (2, 1) for A and (6, 4) for B.

For a selector or voting rule s , define $D_s = (F_{M, \text{TRUE}} - F_{M, \text{SHUFFLED}}) - (F_{H, \text{TRUE}} - F_{H, \text{SHUFFLED}})$. It measures whether correct support helps M more than H. We then compare the LLM with voting: $J = D_{\text{LLM}} - D_{\text{vote}}$. A negative J need not mean support harms the LLM. If correct support helps its M decisions by two points and H by six, $D_{\text{LLM}} = -4$; if voting gains one point in each, $D_{\text{vote}} = 0$ and $J = -4$. These illustrative numbers explain the contrast, not the observed effect.

On 96 MedEinst pairs, Gemini's $J = -16.54$ $[-27.30, -6.47]$ and Llama's -13.41 $[-23.35, -3.95]$ points (97.5% intervals). Using tie-aware inclusion—RR only for ties—instead shifts both by 11.33 points: -5.21 $[-14.58, 4.17]$ and -2.08 $[-10.94, 6.77]$ (exploratory 95% intervals). No LLM output changes: the reference, not an internal mechanism, changed.

Neither negative contrast reproduces on 128 native-key questions. Here mixed pools more often contain a correct key: the availability gain is 6.25 points $[2.34, 10.94]$. Neither selector chooses any of the eight keys available only in mixed pools. Availability therefore need not be converted into a correct decision [Chen, 2026b]. Source and answer format change together, preventing separate attribution. These findings are supporting boundaries, not confirmation of C or θ .

The *support-rank test* compares LLM and top-rank selection from a pool shared across allocations. On 224 reused cases, that pool is the H/M union for MedEinst and all native options. Inclusion count, then RR, then a deterministic hash give unique ranks. Selectors see ranks in forward or reversed order instead of counts. Here D is M-minus-H in the forward-minus-reversed accuracy change; $J_{\text{norm}} = D_{\text{LLM}} - D_{\text{top rank}}$ compares selectors. This changed intervention and reference cannot confirm J . Only Llama on MedEinst has a positive 98.75% interval, 8.85 $[3.13, 15.10]$

percentage points; the other three include zero. The positive interaction partly reflects weaker mixed rankings, not better mixed selection.

E Local reproduction, resources and access

Portable saved-output replay. Extract the separate `representation_replay.zip`, install NumPy, and run `python3 replay.py` from the extracted directory. The entry point verifies input checksums and runs eight checks: representation/mapping sensitivity, the 720-question decision analysis, quality/lift accounting, the empirical-list reference, Health transport, the answer-contract test, development-selected joint policies, and allocation/budget curves. Inputs are scored ranks and case identifiers; outputs are reconstructed rule scores, contrasts and uncertainty checks. This requires no model access or credentials. It replays stored-output inference, not fresh hosted sampling or clinical adjudication. Question text and full provider payloads are not included. The separate manuscript-source archive builds the PDF with `latexmk -pdf main.tex`; it is not an inference runner.

Serving and repair details. *Answer-key access by stage.* Development labels select policies; generators receive questions/options or patient records, not answer keys. Secondary LLM selectors receive candidate text, rationales and support without the key. Voting and rank-transfer operators form decisions without it; scoring then uses it. Retries respond to service/parsing failure, not correctness. Gold-informed membership ceilings, wrong-answer relabeling and hindsight routing use held-out keys for diagnosis, so they are not label-free deployable policies. Changing key metadata leaves all 31,232 scheduled Native720/Health generation messages unchanged. This checks prompt construction, not pretraining contamination or independent clinical validity.

Of 192 Qwen generator identities, 27 came from hosted Featherless inference and 165 from a RunPod/vLLM fallback using the named official 14B weights. Gemma requests were routed among eight serving providers; the nominal endpoint does not fix their kernels or quantization. The original automated mapping used `google/gemma-3-12b-it` and `microsoft/phi-4`; the seven-cell local repair used Qwen/Qwen2.5-1.5B-Instruct, a different model from the generator. Three natural-selector formatting repairs preserved top-one identities. C3B ran on 6 September 2026 and returned Google as provider for all 576 choices. C2 cost \$0.244793. The C3/C3B interface experiments used 1,157 requests including five identical retries. The subsequent scoring and rule replays added no inference cost. These implementation qualifications do not establish clinical adjudication or equivalent remote serving implementations.

The manuscript-source archive contains text, official styles, bibliography and figures/tables, not a public experimental dataset. Stored-output replay is distinct from hosted regeneration and terminology adjudication. Original traces and scoring corrections remain archived; credentials are excluded.

E.1 Replay resources and notation

The accompanying `ballot_diagnostics_v1.zip` contains 31,232 scheduled native-key lists across 976 questions and two banks, a versioned schema/data card, executable analyses, expected results and checksums. Terminal failures are empty lists. It supports new rule evaluation without regenerating model answers, but is a local anonymous submission artifact, not a public DOI deposit. Install Python and NumPy; run `python3 replay.py --data data --out outputs`, then `python3 null.py --data data --out outputs`. Compare the three JSON outputs with `expected/`. These commands need no network, question text, model credentials, or clinical dictionaries.

Notation and interval guide. C is the inclusion-minus-first allocation interaction; C_G names the same contrast against Gemma on Health. θ is mixed-minus-selected-Gemma first-choice accuracy. Δq is the per-call accuracy difference; ΔL is the difference in aggregation lift, an accounting quantity. $R = \theta_{K5} - \theta_{K1}$ compares answer contracts. v , s and β denote availability, conditional selection success and pool co-failure. Original sole primaries have pointwise 95% intervals. Rank randomization uses eight adjusted endpoints; transfer four; IRV six; task breakdown six; partner replay nine; policy comparisons 28. The secondary selector quantities are defined in Section D.1: D is the support-intervention interaction, J subtracts a voting reference, J_{norm} uses a different support-

rank intervention and reference, and Δ_A is an availability difference. The two-selector J family uses 97.5% marginal intervals; four-cell common-pool intervals are 98.75%. Exploratory reference changes use the interval level stated at their result.

Cost and serving limits. Dividing Native720 H's recorded eight-call resources by eight gives a uniformly sampled single-slot reference of 585.9 input and 149.1 output tokens, approximately \$0.000098 per question, and 34.48% accuracy. This uses the same five-answer contract and includes recorded retry costs; it is not a newly run direct-answer model. Batch durations (e.g. Health 1,432 seconds) combine concurrent dispatch and retries and do not identify deployment request latency. The retained logs identify named endpoints/providers and request dates, but do not supply exact weight snapshots for all hosted calls. Missing snapshot hashes remain unknown; open-weight model names do not certify local serving equivalence.

F Mapping correction and construct scenarios

The versioned union consists of 4,510 earlier A0 pairs and 1,770 C2 pairs with 812 overlap. Version 3 repairs five relations globally: three Ebola disease synonyms, the explicit non-Ebola exclusion, and Koch's disease as a tuberculosis synonym. Version 4 repairs 35 further relationships using source-backed text rules. Strict labels are unchanged from v3. The complete change ledgers include reference/candidate strings, old and new relations, rule identifiers and source locators. Their scope includes the full C1 inventory, not just outcome-relevant C2 cases.

NLM distinguishes a MeSH Concept's synonym group from its broader Descriptor: https://www.nlm.nih.gov/mesh/concept_structure.html. The correction uses ConceptUI identity, verified aliases/acronyms and target-negation checks. CDC pharyngitis guidance distinguishes streptococcal from viral disease: <https://www.cdc.gov/group-a-strep/hcp/clinical-guidance/strep-throat.html>. CDC respiratory guidance distinguishes upper-airway diagnoses from bronchitis/pneumonia: <https://www.cdc.gov/antibiotic-use/hcp/clinical-care/adult-outpatient.html>. ACS includes both infarction and unstable angina: <https://medlineplus.gov/ency/article/007639.htm>. The complete source index is local; source reading is not case adjudication.

The source-coded audit links all 49 released targets to the official CDC FY2026 order file, retaining the source's undeclared coding vintage as a limitation. For example, Bronchitis has unspecified J40, pulmonary/pancreatic neoplasm names omit the malignancy in C34/C25, and Pericarditis omits the acute state in I30. Other code/name mismatches include the broader asthma code for a compound acute-episode target. Codes never silently override the primary map.

Temporal modifiers are acute, chronic, recurrent, recurrence and exacerbation; etiologic modifiers are viral and bacterial; severity/site modifiers are severe, mild, left, right and bilateral. Only candidate modifiers absent from the full reference can be removed, and the remainder must match the full reference through independent synonym rules. This recognizes 18 temporal, ten etiologic and two severity/site pair incidences (classes can overlap). All eight subsets are evaluated. A separate finite source-coded sensitivity concerns explicit terms for pulmonary/pancreatic malignancy, acute pericarditis and myocardial infarction. It affects eight inventory pairs; it is not an exhaustive clinical scoring system.

The 14 priority source dispositions agree with the final map. All 106 original class controls are retained; 675 C2 relations remain flagged unresolved. Regression checks cover historical/current mentions, contradictory etiology, upper/lower disease, episode specificity, alternatives, negation, causal clauses and unverified parentheses. Broad v3 mapper-disagreement extrema remain preserved as stress bounds, not clinically plausible confidence bounds; the present scenarios provide narrower, source-motivated construct checks without claiming exhaustive semantic coverage.

G Secondary study protocols and development limitations

These studies ask whether a selector uses the candidate support supplied by generators. The intervention changes the displayed support while retaining each candidate pool. The direct interaction D compares the effect on mixed and homogeneous pools; J subtracts that interaction for a specified voting rule (Section D.1). A negative J therefore concerns the LLM relative to that reference,

not necessarily an absolute loss in LLM accuracy. The development, MedEinst confirmation and native-key transport below use different cases and must remain separate.

In the 128-question native-key support transport, Gemma’s per-call first-choice accuracy is 305/1,024, versus 31/256, 33/256 and 32/256 for the other generators. These descriptive quality differences accompany the larger mixed candidate pools; they do not identify a quality-independent lineage effect.

On 48 development pairs, the original positive rescue fails: Gemini’s refinement interaction is -0.0625 (95% CI $[-.1562, .0312]$), with strict $+0.0104$ $[-.0833, .1042]$. Exploratory frequency-relative J is -0.1840 $[-.2969, -.0790]$ strict and -0.2309 $[-.3394, -.1276]$ with refinements.

G.1 Earlier native support-study cells

Table 16 preserves the full earlier 128-item support-intervention cells. It is a different cohort and estimand from the fresh 288-item representation primary. Section 4.4 reports the unresolved transport and available-but-unselected keys.

The correct key is available on 114 H and 122 M questions, while the best deterministic rule scores 38 expected correct H selections. Availability is therefore much larger than achieved voting accuracy on this earlier cohort; these are not the Native720 pools.

Table 16: Earlier MedXpertQA support study: expected correct selections out of 128 native-key questions. Fractional cells integrate uniform ties for deterministic rules; LLM rows count actual correct choices. Same support interface; no clinical semantic judge.

Rule	H TRUE	H SHUFFLED	M TRUE	M SHUFFLED
Gemini	42	27	29	23
Llama 8B	32	24	28	15
Inclusion frequency	23.98	15.98	22.08	10.75
First choice	37.50	14.00	21.42	11.42
Frequency + RR ties	37.00	13.50	22.00	14.50
Reciprocal rank	37.50	13.50	23.00	13.33
Borda	38.00	13.50	22.50	15.50
Oracle	114	114	122	122

G.2 Earlier development design

The earlier 48-pair MedEinst study motivated, but does not independently confirm, the native-key studies. An outcome-blind hash-ranked selection balanced transition frequency and narrative length into 16/16/16 strata. Its 96 branches are correlated pairs, not independent cases. The separate 36-case DDXPlus boundary already had a correct concept in every pool under both policies and cannot show increased availability or independent clinical transport. Full early screening records remain in the archival version; material scoring and failure qualifications follow.

G.3 Diagnostic generation and selection

Each branch uses eight Gemma calls and two each from Llama, Ministral and Qwen; the mixed allocation reuses two Gemma calls. The generator system message was:

You are performing a controlled diagnostic benchmark. Use only the supplied patient record and return the requested JSON.

The user template was:

Generate a ranked differential diagnosis using only this patient record. Return exactly five distinct plausible diagnosis names, ordered most to least likely. For each, give one evidence-linked rationale of at most 18 words. Do not assume a supplied candidate list, another case version, or other agents.
Patient record:
[PATIENT_RECORD]

1109 Return only one JSON object:
1110 {"candidates":[{"diagnosis":"name", "rationale":"at most 18 words"}, ...
1111 exactly five items]}. Do not use markdown fences or add keys.

1112 Generation uses temperature 1, top- $p = .9$ and 384 output tokens. All 1,344 scheduled proposals
1113 are included. Qwen contributes 27 hosted and 165 local responses after 59 provider failures; one
1114 malformed local output receives the allowed same-prompt retry. These serving implementations are
1115 not established equivalent (Section E).

1116 The natural selector, Gemini2.5-Flash-Lite, receives the patient record, opaque candidate IDs, di-
1117 agnosis text and one representative rationale. Source, frequency and allocation labels are withheld;
1118 candidate order is hash-derived. It returns a complete ID permutation at temperature 0, top- $p = 1$,
1119 maximum 300 tokens. All 192 canonical rankings are valid after 198 raw attempts; three format-
1120 only repairs preserve the original top choice. No failed identity is dropped. The earlier mapping
1121 uses Gemma3-12B and Phi4 at temperature 0, top- $p = 1$, maximum 5,000 tokens. Diagnosis-only
1122 pairs receive one of four relations: synonym, parent/subtype, compound/context-dependent, or non-
1123 equivalent. Of 1,916 novel mapping cells, 1,909 come from the provider and seven are completed
1124 locally before analysis. Neither model agreement nor this repair is clinical adjudication.

1125 **G.4 Development scoring and failed tests**

1126 Natural-effect intervals resample whole case pairs 50,000 times (seed 20260905185214). Scoring
1127 corrections are retrospective: investigators had seen aggregate results. Rules apply globally to diag-
1128 nosis pairs, and intervals condition on the chosen map rather than jointly accounting for semantic
1129 uncertainty. Frequency counts distinct proposing calls; reciprocal rank uses the best occurrence in
1130 each call. Uniform ties and ambiguous surface representatives receive expected correctness, never
1131 any-member credit. These rules see support information withheld from the natural selector.

1132 Strict availability is 51/96 versus 60/96 for H/M, with 18/96 versus 10/96 final correct choices.
1133 Crediting temporal and other whitelisted refinements changes availability to 60/96 versus 62/96 and
1134 final choices to 20/96 versus 14/96. Eight apparent mixed-only opportunities disappear because
1135 H’s “Acute Bronchitis” receives credit against “Bronchitis.” Thus these outputs do not establish
1136 robust clinical-concept complementarity. No strict mixed-only opportunity converts; one of three
1137 refinement opportunities does. Section F retains the correction authority and unresolved-label limits.

1138 The candidate-set-size test uses 11 mixed-only branches matched to 11 both-available controls by a
1139 pre-output difficulty instrument, two sizes ($k = 2, 8$) and two orders: 88 valid rankings. Original and
1140 corrected-strict ten-match intervals include zero. Temporal credit leaves only two eligible matches
1141 with identical observed interactions, yielding a degenerate empirical interval .5 [.5, .5]; that is not
1142 evidence of a general competition mechanism. An additional refinement-eligible branch has no
1143 saved intervention and is not imputed.

1144 The metadata screen changes only supplied support counts/ranks within each pool: hidden, true,
1145 or cyclically reassigned metadata. All 576 fresh single-ID choices use Gemini2.5-Flash-Lite at
1146 temperature 0, top- $p = 1$, maximum 64 tokens and a shared per-pool seed. It fails its prespecified
1147 positive-rescue criterion: interaction at least .05, positive 97.5% case-pair-bootstrap lower limits
1148 for both the interaction and mixed true-minus-shuffled contrast, and positive strict-score directions.
1149 Historical full rankings are not its control arm. Its preceding full-ranking attempt produced 571/576
1150 valid outputs after 581 attempts and was not scored; the uniform single-ID repair preceded accuracy
1151 inspection, and no predecessor outputs were pooled. Negative estimates and later, distinct selector
1152 tests remain in Section G. These synthetic benchmark studies do not establish patient benefit.

1153 **H Further tests of information loss and decision value**

1154 These analyses address different limits of the central comparison. They reuse stored answers and are
1155 exploratory; they add neither cases nor independent confirmation. Native288 uses its original four-
1156 model allocation. Native720 and Health256 use eight Gemma calls, eight Llama3.3-70B calls and
1157 the first four scheduled slots from each. Empty failed slots remain. Except for the explicitly labeled
1158 Monte Carlo reference distributions, intervals resample whole questions, average banks, and retain
1159 native task strata. A common 20,000-draw bootstrap (seed 20260921) supplies nominal intervals

and the family adjustments specified below. The original primary intervals, obtained with their own frozen seeds and 50,000 draws, remain authoritative.

H.1 Prespecified endpoints and generation failures

The representation study froze native C before new generation; MedEinst was a prespecified secondary transport with uncertain mappings. The 720-question study froze first-choice θ after development-only selection among ten homogeneous policies. Health froze C_G before its generations. The later fifteen-policy search was specified after test outcomes were known, but selected its winner using development data only. K1/K5 generated fresh answers on 128 new questions under a separately frozen contract comparison; its unresolved primary missed the precision target. Depth, budget, partner, rank-transfer, information-bound, relabeling, screening and oracle results are saved-output analyses, not additional replications. The held-out predictor’s recipe was frozen after generation began but before fresh accuracy analysis.

The original attempt audit reconciles 13,440 slots with 13,845 attempts, first-valid retention and at most one identical-payload retry. Every source/model cell passes the protocol’s terminal-failure threshold of at most 5% (Table 17). A failed first attempt followed by success is an attempted retry, not a terminal failed slot. Meeting this failure threshold does not resolve scoring or scientific transport uncertainty.

Table 17: Original representation study: scheduled slots, actual attempts including retries, terminal failures and percentage of scheduled slots failing. All eight source/model cells are below the pre-specified 5% failure limit.

Source	Model	Slots	Attempts	Failures	%
medeinst	gemma	3072	3104	5	0.16
medeinst	llama	768	788	0	0.00
medeinst	ministral	768	773	2	0.26
medeinst	qwen	768	771	1	0.13
native	gemma	4608	4863	85	1.84
native	llama	1152	1218	43	3.73
native	ministral	1152	1153	0	0.00
native	qwen	1152	1175	12	1.04

H.2 Tie accounting and rule-eligible selection

For each question, define $T = (F_{H,\text{tie}} - F_{H,\text{incl}}) - (F_{M,\text{tie}} - F_{M,\text{incl}})$ and $U = (F_{M,\text{tie}} - F_{H,\text{tie}}) - (F_{M,\text{first}} - F_{H,\text{first}})$. Then $C = T + U$ exactly before and after averaging. Six component endpoints (three cohorts, T/U) share a Bonferroni family. The unresolved remainders do not establish equivalence. Conditioning on observed masking is descriptive: the contribution from questions with at least one correct unanimous H first-choice bank masked by an inclusion tie is calculated using the full cohort denominator, not as a subgroup average.

Table 18: Paired decomposition in percentage points. T/U intervals adjust across six components. Masked and other contributions sum to C; their columns are point estimates using the full question denominator.

Cohort	T [adjusted 95%]	U [adjusted 95%]	Masked	Other
Native288	7.62 [2.04, 13.32]	2.10 [−1.85, 6.11]	6.90	2.82
Native720	8.73 [5.48, 11.95]	0.45 [−1.86, 2.81]	6.42	2.76
Health256	5.98 [1.25, 10.52]	0.16 [−2.80, 3.19]	9.06	−2.92

First-choice eligibility means that gold appears at rank one in at least one selected list. All-rank availability means it appears anywhere. We calculate conditional success as pooled final accuracy divided by pooled eligibility, recomputing both terms in each bootstrap. For first-choice voting, the error decomposes as $1 - F = (1 - v_1) + (v_1 - F)$: inaccessible gold plus failure despite rank-one access. The denominator is the full set of scheduled questions and banks, including failures. The

archive additionally reports minimum-gold-rank frequencies (ranks one to five, or absent), all-rank conditional success, and paired intervals.

The exact counts make the conditioning explicit. Native720 has 1,440 question-bank records: H/M have gold at any rank in 1,291/1,327 records and at rank one somewhere in 671/709. Their expected correct totals are 6085/12 and 2269/6, yielding $100(6085/12)/671 = 75.57\%$ and $100(2269/6)/709 = 53.34\%$ conditional success. Health has 512 records: all-rank counts 493/501, rank-one counts 380/411, and expected correct totals 1057/3 and 693/2, yielding 92.72/84.31%. These fractional totals average uniform voting ties; they are not counts of independently observed successes. Resampling keeps both banks together within each question. Comparing conditional rates on different eligible sets is descriptive, not a matched intervention on selector ability.

Table 19: Eligibility and selection (percent). F/v1 uses pooled values; it is not an average of question-level ratios.

Cohort	Policy	Rank-one v1	All-rank v	F/v1	Inaccessible	Eligible error
Native288	H	43.75	86.28	68.25	56.25	13.89
Native288	M	50.17	89.58	37.69	49.83	31.26
Native720	H	46.60	89.65	75.57	53.40	11.38
Native720	M	49.24	92.15	53.34	50.76	22.97
Health256	H	74.22	96.29	92.72	25.78	5.40
Health256	M	80.27	97.85	84.31	19.73	12.60

Table 20: Eight-call first-choice minus a uniformly selected scheduled single slot, using the same five-answer contract. Intervals adjust over four contrasts; neither interval type is evidence of equivalence.

Cohort	Policy	Gain [95%]	Gain [adjusted 95%]
Native720	H	0.73 [−0.01, 1.49]	0.73 [−0.19, 1.69]
Native720	M	−0.91 [−1.82, 0.03]	−0.91 [−2.07, 0.27]
Health256	H	0.09 [−0.66, 0.82]	0.09 [−0.86, 1.01]
Health256	M	1.10 [0.10, 2.12]	1.10 [−0.15, 2.42]

MedEinst requires a separate scoring qualification. We recompute the tie remainder under the implemented refinement criterion, a protocol-consistent sensitivity zeroing new unresolved relations, and strict synonyms. Nominal intervals condition on each fixed map; three adjusted remainder intervals cover these sensitivities. Negative values remain properties of those maps, not clinical counterexamples to the native-key decomposition.

Table 21: MedEinst mapping-conditional C, T and U in points. No terminology adjudication is added.

Scoring	C	T	U [adjusted 95%]
protocol_consistent	4.93	11.69	−6.76 [−11.59, −2.01]
refinement	4.85	11.69	−6.85 [−11.67, −2.06]
strict	6.10	11.67	−5.57 [−10.20, −1.01]

1204 H.3 A ceiling for selectors without ranks

A candidate-name-neutral selector is equivariant: renaming candidates in its input merely renames its output distribution. Assume it sees no question text, rationale, semantic candidate attributes or answer key, and can select only candidates appearing in the pool. One representation supplies each candidate’s total inclusion count. A richer representation supplies the entire binary candidate-by-call matrix, retaining call order and therefore model membership, but omitting ranks.

For the count representation, place candidates with the same count in one class. For the matrix, place candidates with identical rows of call membership in one class (equivalently, identical columns if candidates index columns). If the gold candidate belongs to a class of size k , its selection probability

is at most $1/k$; if absent it is zero. To see why, swap gold with any candidate in its class. The represented input is unchanged, while equivariance swaps their output probabilities. All k probabilities must therefore be equal and sum to at most one. Averaging the per-pool bound gives a ceiling on mean accuracy for every selector in this restricted class. This elementary symmetry argument is not a new general decision theorem.

The argument conditions on each realized pool: it assumes neither independent errors nor equal generator accuracy. The ceiling uses gold to identify its class, so it need not be attainable: a real selector also has to choose the right class. It permits arbitrary use of model-specific membership patterns. It does not cover a semantic judge, rationale reader or a rule exploiting the original answer-letter positions. Ranked voting lies outside the restricted class because exchanging identical membership columns need not preserve ranks. A positive paired gap shows information available in ranks beyond this representation; a negative gap does not show that membership suffices for an attainable selector.

Table 22: First-choice accuracy versus gold-informed symmetry ceilings (percent), and its gap above the stronger full-membership ceiling (points). Gap intervals adjust over eight source/allocation/representation contrasts; the count-gap endpoints are retained in the archive.

Cohort	Policy	Count ceiling	Full membership	First choice	Gap [adjusted 95%]
Native720	H	39.68	44.21	35.21	−8.99 [−15.18, −2.92]
Native720	M	44.98	53.72	26.26	−27.46 [−33.04, −21.95]
Health256	H	28.59	30.28	68.82	38.54 [28.86, 47.46]
Health256	M	35.65	38.42	67.68	29.25 [19.84, 37.83]

H.4 Changing wrong-answer alignment at fixed correctness

For every question and model, sample a uniform bijection over the nine incorrect option IDs, fixing gold. Reuse it for every list of that model in both banks. Different models receive independent permutations. This preserves each correct answer’s rank or absence, every failure, and all within-model equality relations. The mixed allocation reuses the same transformed Gemma slots as H. Unlike shuffling lists across questions, this operation also preserves the entire question-by-call correctness matrix and question difficulty alignment.

All homogeneous scores must be unchanged for every draw and rule. Applying one common candidate permutation to both models must also preserve every mixed score; both controls pass. Model-specific permutations change wrong-answer agreement without changing per-call accuracy. They do not preserve candidate meanings and use the gold key, so they cannot be supplied to a semantic judge or deployed as an improvement. There is no claim about counterfactually retrained models.

We use 256 draws (seed 20260994). The archive reports each rule’s mixed accuracy change, changes in rule-versus-first interactions, randomization reference ranges, changed-outcome frequency and split-half Monte Carlo checks. Question-bootstrap intervals condition on the per-question simulation means; permutation ranges describe the randomization distribution and are not confidence intervals. Ten rule-interaction changes across two sources form a family (the first-versus-itself contrast is identically zero). Separate ten-endpoint adjustment covers the five mixed-accuracy changes per source.

Native720: 49.90% of bank-level mixed outcomes change in at least one rule, averaged over draws. The randomization 95% range for the change in C is [9.98, 12.35] points. Monte Carlo standard error of its population mean is 0.04 points. The maximum population-mean difference across the two draw halves is 0.116 points across five mixed-rule accuracies. Individual-question means are less precise; no interpretation treats simulation draws as independent questions.

Health256: 71.80% of bank-level mixed outcomes change in at least one rule, averaged over draws. The randomization 95% range for the change in C is [13.54, 17.70] points. Monte Carlo standard error of its population mean is 0.07 points. The maximum population-mean difference across the two draw halves is 0.052 points across five mixed-rule accuracies. Individual-question means are less precise; no interpretation treats simulation draws as independent questions.

Table 23: Synthetic wrong-option alignment: changes in mixed accuracy and the rule-versus-first interaction, in points. The original H scores are invariant. Intervals adjust within the respective ten-endpoint families.

Source	Rule	Mixed accuracy change	Interaction change
Native720	inclusion	11.29 [9.95, 12.64]	11.08 [9.74, 12.44]
Native720	first	0.21 [−0.01, 0.51]	0.00 [0.00, 0.00]
Native720	tie	13.72 [11.25, 16.28]	13.51 [11.07, 16.04]
Native720	rr	9.26 [7.05, 11.63]	9.05 [6.88, 11.35]
Native720	borda	14.90 [12.17, 17.74]	14.68 [12.01, 17.47]
Health256	inclusion	15.56 [13.58, 17.40]	15.52 [13.55, 17.37]
Health256	first	0.04 [0.00, 0.19]	0.00 [0.00, 0.00]
Health256	tie	8.91 [5.61, 12.63]	8.88 [5.59, 12.54]
Health256	rr	6.70 [3.54, 10.34]	6.66 [3.52, 10.30]
Health256	borda	10.91 [6.73, 15.55]	10.88 [6.70, 15.53]

H.5 Donor ranks, reusable headroom and equal-budget screening

The neutral donor sensitivity replaces deterministic SHA ordering of equal donor scores with one uniformly random candidate priority shared across the recipient’s lists. Only the earliest donor-present position can affect that list’s final first vote. An exact recursion averages the earliest member of the union of unresolved tied first-position candidates, resolves every affected list, and repeats; final vote ties still receive uniform credit. Exhaustive small examples agree with direct enumeration of candidate orders. Self and cross controls use the same operation. A frequency-only donor replaces reciprocal-rank scores by inclusion counts; direct donor-first accuracy provides the simple answer-import reference, without pretending to retain recipient memberships.

Table 24: Recipient accuracy after neutral-priority transfer (percent) and cross-minus-self (points). Each operator uses a separate four-contrast family. Cross-transfer uses twelve unique calls per bank; direct donor-first is a different procedure.

Source	Recipient	Donor score	Self	Cross	Change [adjusted 95%]
Native720	H	rr	34.65	25.57	−9.08 [−12.34, −5.89]
Native720	H	frequency	19.49	19.68	0.19 [−1.28, 1.63]
Native720	M	rr	24.95	33.58	8.63 [5.56, 11.74]
Native720	M	frequency	19.71	19.43	−0.29 [−1.47, 0.88]
Health256	H	rr	69.04	68.46	−0.59 [−4.88, 3.61]
Health256	H	frequency	23.48	28.55	5.07 [3.35, 6.94]
Health256	M	rr	68.36	68.46	0.10 [−4.20, 4.49]
Health256	M	frequency	28.48	25.09	−3.39 [−5.05, −1.87]

The maximum absolute cohort-mean shift from replacing hash ties is .12 points on Native720 and .68 on Health; the full paired sensitivity intervals are archived. Direct donor-first yields the other allocation’s original accuracy (26.26/35.21 on Native720 and 67.68/68.82 on Health for H/M recipients). Thus reranking does not establish superiority to simply using the strong donor. Frequency-only transfer has no resolved Native720 cross-minus-self effect, whereas neutral RR transfer retains both effects. Health’s RR effects remain unresolved.

Eight-call donor allocation. Can donating ranks pay for the calls it consumes? We additionally freeze four allocations of eight stored calls: six Gemma proposals and two Llama donors, the reverse, four Gemma proposals and four Llama donors, and the reverse. Calls are the earliest scheduled slots, not selected successful or correct outputs. Donors reorder only candidates already in recipient lists, using the neutral-priority operator above. Ordinary first-choice, reciprocal-rank (RR) and Borda pooling use exactly the same eight outputs. Eight Gemma calls with first-choice voting remain the development-selected reference. Thus a gain over the recipient alone does not establish a useful allocation.

Table 25 reports every allocation on the historically inspected Native720 and Health256 outputs. The two banks are averaged within question; 20,000 question bootstraps preserve original task strata. Five contrasts per allocation/source (three same-pool rules, Gemma eight-call reference and

recipient-only first choice) form a 40-comparison exploratory family. Intervals condition on stored generations; they are not fresh confirmation. All contrasts and vectors are in the accompanying budget-transfer replay. Twenty-seven toy cases independently verify exact neutral-priority averaging.

Table 25: Fixed eight-call screen. A=Gemma, B=Llama; $6A + 2B$ means six A recipient calls and two B donor calls. Accuracies are percent; the last column is transfer minus eight Gemma first-choice calls, with Bonferroni-40 bootstrap intervals in points. No allocation exceeds the reference’s point estimate.

Source	Recipient+donor	Transfer	First	RR	Borda	Difference [adjusted 95%]
Native720	$6A + 2B$	23.26	33.93	33.65	31.70	-11.96 [-16.88, -7.10]
	$6B + 2A$	31.17	20.28	20.62	21.46	-4.05 [-7.38, -.80]
	$4A + 4B$	23.23	26.26	24.86	25.06	-11.98 [-17.00, -7.05]
	$4B + 4A$	32.40	26.26	24.86	25.06	-2.82 [-5.88, -.01]
Health256	$6A + 2B$	66.86	69.24	69.82	70.90	-1.95 [-9.96, 6.05]
	$6B + 2A$	67.29	66.11	66.02	68.55	-1.53 [-5.57, 1.87]
	$4A + 4B$	66.21	67.68	68.26	68.55	-2.60 [-10.66, 5.47]
	$4B + 4A$	66.60	67.68	68.26	68.55	-2.21 [-5.99, .65]

On Native720, $4B + 4A$ gains 7.53 points [3.83, 11.35] over same-pool RR, but remains 2.82 points below eight Gemma calls. Stronger donor preferences can therefore improve weak recipient lists without beating a competent homogeneous policy. Health differences remain unresolved, not equivalent. The near-zero Native720 upper endpoint is sensitive to finite-bootstrap tail estimation and is not evidence for a general impossibility. This tests stored-generation donors, not fresh candidate-conditioned ranking calls or all possible allocations.

For cross-bank hindsight, select an optimal policy using the gold key and bank one, then evaluate that policy in bank two; reverse the roles and average. Uniformly average equally optimal selection policies. Compare against H first-choice in the same evaluation bank. The within-bank oracle uses the same candidate set and tie convention. The three candidate sets isolate allocation-only (three first-choice policies), rule-only (five H rules), and all fifteen policies. These label-using oracles measure repeated-output headroom on already known questions, not generalization to new questions or a realizable learned router. The previous oracle maximized after averaging banks; its different estimand cannot be subtracted to estimate noise.

Table 26: Hindsight gains above the unchanged H first-choice reference (points). Cross-bank intervals adjust over six contrasts. Within-bank results are point estimates under the same selection convention.

Source	Candidate policies	Within bank	Cross bank [adjusted 95%]
Native720	allocation	5.52	3.26 [1.32, 5.37]
Native720	rule	9.63	6.18 [4.64, 7.70]
Native720	all	18.26	12.11 [9.73, 14.57]
Health256	allocation	8.72	7.85 [3.77, 12.51]
Health256	rule	5.77	4.45 [2.73, 6.32]
Health256	all	13.96	11.77 [7.49, 16.66]

Equal-budget screening fixes the first-four-Gemma first-choice prediction separately in each bank. Its initial confidence is the top-two vote-count margin divided by four. Additional confidence uses either the next four Gemma lists or the first four Llama lists, at eight scheduled calls total. Agreement is the probability that independent uniform draws from the initial and additional winner sets match: their intersection size divided by the product of their sizes; an empty set gives zero. These scores are fixed and unfitted. We retain confidence ties fractionally and evaluate risk at ten coverage levels from 10% to 100%. Each bank’s risk curve is computed separately, then averaged; we do not combine bank scores into a sixteen-call policy. Bootstrap resamples paired questions and recomputes rankings. The direct Llama-minus-Gemma mean-grid-risk contrast has two-source adjustment. Mean-grid risk is not continuous AUC or a validated clinical threshold.

Full-coverage initial accuracy is 35.05% on Native720 and 68.36% on Health. Replacing these predictions with eight-call Gemma voting gives 35.21%/68.82%; mixed eight-call voting gives

Table 27: Mean retained error (percent) for fixed four-call predictions. Initial margin uses four calls; extra-G/extra-L use eight. The final column is extra-L minus extra-G in points: negative favors extra Llama.

Source	Random	Margin	Extra G	Extra L	Difference [adjusted 95%]
Native720	64.95	60.47	62.16	65.23	3.07 [0.38, 5.86]
Health256	31.64	27.81	28.85	20.97	-7.88 [-11.45, -4.27]

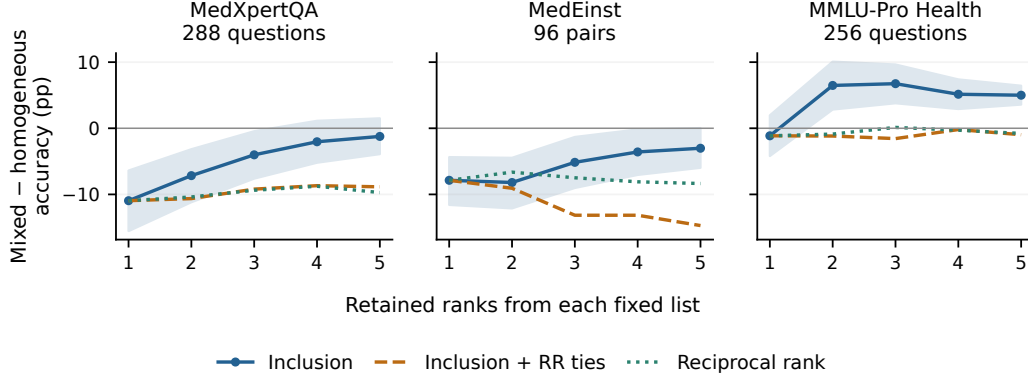


Figure 5: Fixed-ballot depth replay, two banks per source. All curves coincide at depth one. At depth five, inclusion counts every answer equally; RR and the RR tie-breaker retain rank weights. Left/middle use the original four-model mixture; right uses M_{44} versus H_G , as in the separate 720-case design. Bands are secondary paired 95% intervals for inclusion; other curves are point estimates. C compares paired endpoints, not marginal intervals. MedEinst uses refinements; both other sources use native keys.

1310 26.26%/67.68%. Screening answers a different decision question because it holds predictions fixed
1311 and changes retention. The source-dependent comparison merits independent confirmation; it does
1312 not establish a general benefit of cross-model confidence. The earlier unequal-budget analysis is
1313 archived; its confounding and qualitative outcomes are summarized in Section B.5.

1314 H.6 Depth and decomposition curves

1315 The depth sweep shows how the allocation comparison changes as more positions receive votes.
1316 The decomposition separates tie handling from the remaining rule contrast and per-call quality from
1317 aggregation lift. Both retain the cohort definitions and uncertainty of the corresponding analyses.

1318 The separate `eleven_review_replay.zip` contains stripped Native288/Native720/Health256 bal-
1319 lots, scored MedEinst rank pools, executable analyses, the exploratory specification, expected out-
1320 puts and checksums. It contains no question text, credentials or new clinical labels. With Python and
1321 NumPy, run `python3 replay.py --data data --out outputs` and compare the three result
1322 files with `expected/`. The original eight-check archive remains unchanged. This replay does not
1323 regenerate hosted outputs or adjudicate the terminology map.

1324 H.7 Approval cutoff and tie handling on fixed responses

1325 This exploratory challenge asks whether the allocation comparison changes when approval counts
1326 only the first two answers rather than all five. It reuses the original Native288, Native720 and
1327 Health256 calls, keys, failures, allocations and two-bank averaging. No questions, model calls or
1328 clinical labels are added. Approval cutoff is the number of stored positions that receive a vote; it is
1329 not the number of answers requested during generation.

1330 For cutoff $k \in \{2, 5\}$, give each candidate in the first k positions one vote per call. The uniform
1331 version averages all tied maximum-count candidates. The rank-tie version first maximizes approval
1332 count, then sums reciprocal ranks from *all five positions* to resolve equal counts. Remaining ties are

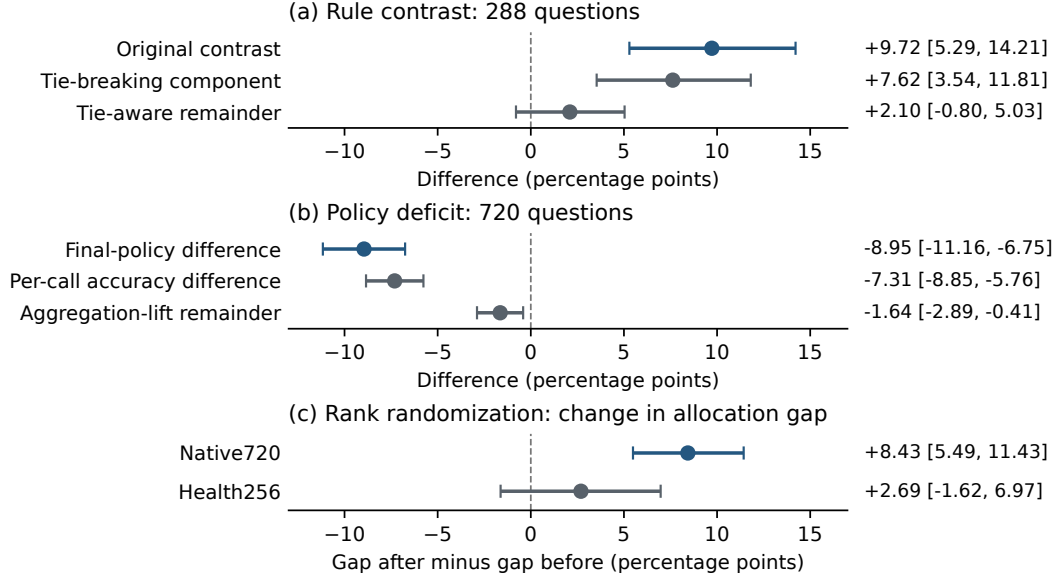


Figure 6: Explanations and a rank intervention. (a,b) Rule and policy decompositions, with paired pointwise 95% question-bootstrap intervals. Components add within each panel; their correlated intervals must not be added. (c) Paired change in the mixed-minus-Gemma first-choice gap after randomizing ranks, with intervals adjusted across eight endpoints. Both policies worsen; Health’s change is unresolved. These exploratory analyses do not isolate a quality-independent lineage effect.

1333 uniform. Only candidates receiving an approval vote are eligible. Thus the tie-breaker introduces
1334 additional preference information; it does not merely choose a deterministic candidate label. Empty
1335 allocations score zero. This is not a reproduction of the token-log-probability or additional-model
1336 tie procedure of Wang et al. [2025].

1337 Table 28 reports all absolute accuracies and tie rates. A tie rate is the proportion of question-
1338 bank decisions with multiple winners before the final uniform draw. In particular, Health’s two-
1339 answer approval gain is not evidence that this policy beats the 68.82% selected Gemma first-choice
1340 reference: its mixed accuracy is only 58.04%. Adding rank information to break ties raises both
1341 allocations, and the mixed-minus-H difference becomes unresolved at $-1.17 [-4.79, 2.44]$ points
1342 (nominal paired 95% interval).

1343 Define C_2 as the cutoff-two allocation difference minus the first-choice allocation difference, U_2 as
1344 the same contrast with rank-resolved approval ties, and $T_2 = C_2 - U_2$. Resample whole questions
1345 20,000 times within the original source/task strata, seed 20260923; banks remain together. The three
1346 C_2 contrasts form a three-endpoint family. The six U_k contrasts (two cutoffs, three cohorts) and
1347 six T_k contrasts each receive separate Bonferroni interval adjustment. Table 29 shows the cutoff-
1348 two results. Native contrasts attenuate and become unresolved; Health retains a positive C_2 , but
1349 not a resolved remainder after tie handling. An unresolved interval does not establish equivalence.
1350 Original primary estimates and their prespecified intervals are unchanged; these exploratory families
1351 do not retroactively adjust those tests.

1352 The separate `practical_cutoff_replay.zip` supplies stripped ballots, expected results, case
1353 vectors, checksums and a standalone program. Run `python3 replay.py --data data --out`
1354 `outputs`. Its 1,296 independent scalar checks include empty ballots, absent keys, reversed lists and
1355 partial lists; baseline means reconcile to 10^{-12} . Compare both JSON outputs byte-for-byte with
1356 `expected/`. No network, GPU or model generation is needed.

1357 H.8 Published aggregation components and tie-choice robustness

1358 Could arbitrary uniform ties explain the native mixed-policy deficit? We test this rival without
1359 changing any generated answer. The BCV and MRRV classes released with Wang et al. [2025],
1360 at commit f5bb1c1, accumulate Borda and reciprocal-rank scores, respectively, then return the

Table 28: Fixed-response cutoff challenge. All values percent, except mixed-minus-H differences in points. Rank ties use the complete five-position RR score. First-choice and five-answer values reconcile exactly with the retained analyses. These are reused cases, not fresh confirmation.

Source	Rule	H accuracy	M accuracy	M–H	H ties	M ties
Native288	First	29.86	18.91	–10.95	5.90	13.89
	Approval 2	25.55	18.39	–7.16	40.45	16.84
	Approval 2, rank ties	28.82	18.92	–9.90	2.08	0.00
	Approval 5	18.68	17.45	–1.23	94.27	66.84
	Approval 5, rank ties	29.51	20.66	–8.85	2.08	0.35
	RR	29.69	19.97	–9.72	2.08	0.52
Native720	First	35.21	26.26	–8.95	4.31	25.49
	Approval 2	31.37	24.70	–6.67	37.71	29.24
	Approval 2, rank ties	34.93	24.76	–10.17	0.83	3.82
	Approval 5	19.49	19.71	0.22	95.14	89.17
	Approval 5, rank ties	34.34	25.83	–8.51	1.04	3.82
	RR	34.65	24.86	–9.79	1.18	4.65
Health256	First	68.82	67.68	–1.14	2.54	20.70
	Approval 2	51.56	58.04	6.48	60.94	39.06
	Approval 2, rank ties	69.43	68.26	–1.17	0.98	5.86
	Approval 5	23.48	28.48	5.00	99.02	95.70
	Approval 5, rank ties	69.04	68.07	–0.98	0.98	5.27
	RR	69.04	68.26	–0.78	0.98	6.05

Table 29: Cutoff-two rule contrasts and decomposition, points with family-adjusted 95% paired intervals. $C_2 = T_2 + U_2$; correlated intervals must not be added.

Source	C_2 (three contrasts)	T_2 (six contrasts)	U_2 (six contrasts)
Native288	3.79 [–.41, 7.93]	2.73 [–.52, 6.00]	1.06 [–2.75, 4.71]
Native720	2.28 [–.45, 5.05]	3.50 [1.50, 5.48]	–1.22 [–3.86, 1.42]
Health256	7.62 [2.96, 12.21]	7.65 [3.26, 11.97]	–.03 [–3.42, 3.37]

1361 first maximum encountered by `Counter.most_common(1)`. We execute those classes on all three
1362 retained cohorts. Failed empty lists are represented by the source’s invalid sentinel and excluded
1363 from vote accumulation; their questions remain, and an all-failure allocation scores zero. Native288
1364 keeps its original four-model mixture; Native720 and Health keep their Gemma–Llama mixture.

1365 This is an aggregation-component replay, not a reproduction of the published experiment. Our
1366 prompts, models, parser, ten-option questions and stored five-answer lists remain unchanged. The
1367 inspected code path has encounter-order tie handling, distinct from the paper-described confidence
1368 or extra-model resolution; we do not infer which path produced the authors’ original results. Literal
1369 MRRV replay preserves floating-point additions. Reversing the eight input lists tests sensitivity to
1370 both encounter order and numerical summation, not a new generation order.

1371 Both native cohorts retain a mixed deficit under both components; Health does not resolve one.
1372 Inference resamples questions 20,000 times within source/task, keeping two banks and all policies
1373 together, seed20260926. This exploratory six-contrast family does not alter the prespecified primary
1374 intervals. Reversing call order changes some estimates but not these observed directions (Table 30).

1375 We also bound accuracy without specifying a tie-breaking algorithm. For each question–bank allo-
1376 cation, compute the exact set of maximum-score candidates, using rational reciprocal-rank scores.
1377 Its lower correctness is one only if gold is the sole maximizer; its upper correctness is one if gold
1378 is among the maximizers. Empty allocations have both endpoints zero. Average these endpoints
1379 over banks and questions. Subtract H’s upper accuracy from M’s lower for a lower bound on M–H;
1380 subtract H’s lower from M’s upper for an upper bound. This permits even gold-informed, allocation-
1381 specific tie choices: a deliberately generous outer envelope, not an attainable selector. It is not a
1382 confidence interval. A negative upper endpoint rules out reversal on these fixed outputs; an envelope
1383 containing zero does not demonstrate an operational reversal.

Table 30: Published-component replay on existing answers. H/M accuracies are percentages; differences are percentage points. Scheduled-order differences have paired intervals adjusted across six source–rule contrasts. Reversed-order differences are descriptive sensitivity results, not selected policies.

Source	Rule	H	M	M–H [adjusted 95%]	Reversed
Native288	Borda	28.99	21.35	−7.64 [−13.83, −1.39]	−8.51
	RR	29.69	19.97	−9.72 [−15.97, −3.47]	−9.72
Native720	Borda	35.07	25.97	−9.10 [−12.78, −5.69]	−11.11
	RR	34.72	25.21	−9.51 [−13.17, −6.04]	−10.28
Health	Borda	69.34	68.95	−.39 [−4.49, 3.71]	−1.17
	RR	68.95	68.55	−.39 [−4.69, 3.91]	−.78

Table 31: Fixed-output mixed-minus-H envelopes over arbitrary score-maximizing tie choices, in percentage points. These are gold-informed bounds, not sampling intervals. Inclusion bounds are deliberately broad because they permit adversarial resolution separately for each allocation.

Source	First choice	Inclusion	RR	Borda
Native288	[−14.93, −6.08]	[−55.38, 34.90]	[−10.24, −9.20]	[−9.90, −6.42]
Native720	[−16.04, −1.81]	[−61.67, 54.24]	[−11.46, −8.12]	[−12.15, −7.92]
Health	[−9.38, 7.03]	[−87.50, 86.33]	[−3.12, 1.56]	[−3.91, 1.95]

1384 Native720’s first-choice upper bound is −1.81 points, but its nominal question-bootstrap interval
1385 is [−3.89, .28]: fixed-sample non-reversal does not establish population-level robustness for every
1386 tie policy. The RR/Borda upper-endpoint intervals are [−10.83, −5.49] and [−10.76, −5.21], re-
1387 spectively. Health envelopes cross zero. The results strengthen the native rank-rule deficit against a
1388 tie-choice explanation; they do not show that the positive inclusion interaction persists under realistic
1389 confidence-based aggregation or fresh prompts.

1390 The `tie_policy_replay.zip` archive contains stripped inputs, the pinned source, independent
1391 checks and expected per-question outputs. Run `python3 replay.py --data data --source`
1392 `RankBasedSC.py --out outputs`. All 750 constructed checks and 20,224 actual-output compar-
1393 isons against an independent literal implementation pass; original uniform-rule means reconcile to
1394 10^{-12} . No new model calls, labels, cases or costs are introduced. The analysis is post hoc and
1395 conditional on retained banks and keys.